DESIGNING POLICY OPTIMIZATION ALGORITHMS FOR MULTI-AGENT REINFORCEMENT LEARNING

A Dissertation
Presented to
The Academic Faculty

By

Sihan Zeng

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2023

DESIGNING POLICY OPTIMIZATION ALGORITHMS FOR MULTI-AGENT REINFORCEMENT LEARNING

Thesis committee:

Dr. Justin Romberg, Advisor
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Daniel K. Molzahn
Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Siva Theja Maguluri
Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Thinh T. Doan
The Harry Lynde Bradley Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University

Dr. Guanghui Lan
Industrial and Systems Engineering
Georgia Institute of Technology

Date approved: May 1, 2023

ACKNOWLEDGMENTS

Over my time as a graduate student at Georgia Tech, I am extremely fortunate to have worked and interacted with a wonderful group of mentors, friends, and colleagues, who have profoundly shaped how I conduct research and how I live my life.

First and foremost, I express my deepest gratitude to my advisor, Dr. Justin Romberg. Without his continuous support, encouragement, and intellectual guidance, the completion of this dissertation would have never been possible. It is a privilege to have known and worked with Justin. His empathy, kindness, dedication, and integrity have set an example that I will forever strive to follow.

I would like to thank Drs. Thinh Doan and Kyle Xu for their collaboration and mentorship in academic research, professional growth, and life in general. Enlightening conversations with Thinh sparked my interest in reinforcement learning research and opened up a new world to me. His strong hands-on research guidance in my early graduate school years helped me quickly get into the field. Kyle is one of the first few people I met upon my arrival at Rice University and has been a close friend and tremendous mentor ever since. The collaboration with Kyle on compressive sensing and generative models is a great pleasure.

My internships are truly fun and rewarding experiences. I am grateful to have collaborated with and been advised by Drs. Alyssa Kody, Daniel Molzahn, Kibaek Kim, and Youngdae Kim at Argonne National Lab, Drs. Parisa Hassanzadeh and Sumitra Ganesh at JPMorgan AI Research, and David Kawashima and Dr. Yukikazu Hidaka at Tinder.

I would also like to extend my gratitude to the rest of my thesis proposal and defense committee members Drs. Siva Maguluri, Guanghui Lan, and Ashwin Pananjady, who have provided insightful feedback and discussions throughout the proposal and defense process and whose research in optimization and reinforcement learning has been a source of inspiration for my own works.

I have had many thoughtful discussions and enjoyable chats with Rakshith, Tomer, Jihui,

Cole, Brighton, Namrata, Liangbei, Nauman, Peimeng, Greg, Andrew, and the rest of the Children-of-the-Norm community. I thank my friends, classmates, and collaborators at Georgia Tech, especially Aqeel Anwar, whose work inspired my initial research on multi-task reinforcement learning. The drone simulation platform developed by Aqeel provides a great testbed for my reinforcement learning experiments.

Special thanks go to my friends outside of Georgia Tech, many of whom I have known since childhood, Dan Zhou, Hanwen Zheng, Zhenzhen Qu, Caleb Lu, Tony Chen, Yuqiang Heng, and others, for their prolonged companionship and emotional support.

Finally, words cannot express how grateful I am to my parents, to whom I owe everything. I am also deeply indebted to my partner, Hanqing Sun, who moved multiple times with me and sometimes put her own goals on hold to ensure that I can focus on my research. Their unconditional love and unwavering support have always been my source of strength.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Multi-agent reinforcement learning (RL) studies the sequential decision-making problem in the setting where multiple agents exist in an environment and jointly determine the environment transition. The relationship between the agents can be cooperative, competitive, or mixed depending on how the rewards of the agents are aligned. Compared to single-agent RL, multi-agent RL has unique and complicated structure that has not been fully recognized. The overall objective of the thesis is to enhance the understanding of structure in multi-agent RL in various settings and to build reliable and efficient algorithms that exploit and/or respect the structure.

First, we observe that many data-driven algorithms in RL such as the gradient temporal difference learning and actor-critic algorithms essentially solve a bi-level optimization problem by tracking an artificial auxiliary variable in addition to the decision variable and updating them at different rates. We propose a two-time-scale stochastic gradient descent method under a special type of gradient oracle which abstracts these algorithms and their analysis in a unified framework. We characterize the convergence rates of the two-time-scale gradient algorithm under several structural properties of the objective functions common in RL problems. Targeting single-agent RL problems, this framework builds the mathematical foundation for designing and studying data-driven multi-agent RL algorithms that we will later deal with.

Second, we consider multi-agent RL in the fully cooperative setting where a connected, decentralized network of agents collaborates to solve multiple RL tasks. Our first problem formulation deploys one agent to each task and considers learning a single policy that maximizes the average cumulative return over all tasks. We characterize the key structural differences between multi-task RL and its single-task counterpart, which make multi-task RL a fundamentally more challenging problem. We then extend our formulation by considering maximizing the average return subject to constraints on the return of each task, which forms

a more flexible framework and is potentially more practical for modeling multi-task RL applications in real life. We propose and study decentralized (constrained) policy gradient algorithms for optimizing the objectives in these two formulations and validate our analysis with enlightening numerical simulations.

While the previous chapter studies cooperative agents, we now shift our focus to the case where the agents compete with each other. We study the two-player zero-sum Markov game, which is a special case of competitive multi-agent RL naturally formulated as a nonconvex-nonconcave minimax optimization program, and consider solving it with the simple gradient descent ascent (GDA) algorithm. The non-convexity/non-concavity of the underlying objective function poses significant challenges to the analysis of the GDA algorithm. We introduce strong structure to the Markov game with an entropy regularization. We apply GDA to the regularized objective and propose schemes of adjusting the regularization weight to make the GDA algorithm efficiently converge to the global Nash equilibrium.

The works we have discussed so far treat RL from the perspective of optimization. In the final chapter, we apply RL to solve optimization problems themselves. Specifically, we develop a multi-agent RL based penalty parameter selection method for the alternating current optimal power flow (ACOPF) problem solved via ADMM, with the goal of minimizing the number of iterations until convergence. Our method leads to significantly accelerated ADMM convergence compared to the state-of-the-art hand-designed parameter selection schemes and exhibits superior generalizability.

CHAPTER 1
INTRODUCTION AND BACKGROUND

1.1 Introduction

Fueled by powerful function approximations such as large-scale deep neural networks, reinforcement learning (RL) has been successfully applied to solve real-life problems in a range of applications including game playing [1–3], healthcare [4, 5], robotics [6, 7], and autonomous navigation [8–10]. From a mathematical standpoint, recent advances have shed light on the structure of various RL problems and facilitated the design of more reliable and efficient algorithms. These achievements, however, are primarily made only in the single agent setting. Our understanding remains inadequate for multi-agent RL systems where the agents exhibit complex interactions driven by different rewards. Depending on whether the rewards of the agents are identical, opposite, or mixed, multi-agent RL problems can be categorized into cooperative, competitive, or more complicated settings, each of which has its own unique characteristics. The main thrust of this dissertation is to develop better insight into the structure of some of these multi-agent RL settings, to design multi-agent RL algorithms that leverage/respect the structure, and to apply multi-agent RL to solve meaningful problems in real life.

On the theoretical perspective, we start by proposing a stochastic optimization framework for single-agent RL that lays the mathematical foundation for data-driven multi-agent RL algorithms. This framework unifies a range of existing methods in RL including the actor-critic algorithm and gradient-based temporal difference learning. Providing an abstraction at a higher level of generality also allows us to discover previous unknown algorithms with state-of-the-art convergence rates. In the multi-agent setting, we study the structure and limitations of two specific problems. In the first problem, we consider a group of the agents

connected in a network that have aligned interests. Specifically, each agent is assigned a local RL task and needs to cooperative with each other to learn a unified policy that performs well across all tasks. The second scenario is the two-player zero-sum Markov game, which is a completely competitive setting with one agent maximizing the same discounted cumulative reward that the other agent seeks to minimize. This problem connects to game theory and lays the foundation for understanding more complicated games involving more than two players or general-sum rewards. By taking advantage of the structure and obeying the limitations, we design reliable and efficient algorithms for the two settings and characterize their convergence properties.

On the side of application, we apply RL to enhance the solution of optimization problems. We focus on solving the alternating current optimal power flow (ACOPF) problem with the alternating direction method of multipliers (ADMM) algorithm. Usually formulated as a non-convex quadratically constraint quadratic program (QCQP), the ACOPF problem studies optimally generating power to satisfy network demands subject to constraints dictated by the power system structure. For large power systems, it is known that the convergence of the ADMM algorithm depends heavily on the choice of penalty parameters, and poorly selected penalty parameters can even lead to divergence. As the current practice of choosing these important parameters is usually based on heuristics, we are motivated to develop a RL parameter selection policy with the aim of accelerating the ADMM convergence. We start our problem formulation and algorithm design from a single-agent perspective, but as we leverage the problem structure to simplify the policy our solution exhibits an interesting multi-agent interpretation. Through extensive numerical simulations, we find that our RL-selected penalty parameters result in significantly accelerated ADMM convergence over state-of-the-art human designed methods.

1.2 Related Literature

In the multi-agent RL setting where the agents live in the same environment and jointly determine the state transition, the problem can be viewed as a single-agent RL for each agent if the policies of the other agents are fixed. This is an important observation that allows our work to take advantage of the recent advances in the understanding of single-agent RL. In particular, [11, 12] find that the value function in RL observes the gradient domination condition that upper bounds the optimality gap (measured in objective function value) by a metric on the norm of the gradient. The authors in [13] show that a stronger structure resembling the Polyak-Łojasiewicz (PŁ) condition exists when the value function is properly regularized by the entropy of the policy.

We briefly give the reference to some fundamental results in multi-task multi-agent RL and two-player competitive RL below and note that each individual chapter will have its detailed discussion on related works.

Cooperative Multi-task Multi-Agent RL. Our aim in multi-task multi-agent RL is to find a single policy that maximize the sum of the cumulative rewards across multiple environments. Most existing works on this problem [14, 15] propose sharing the local trajectories/data collected by the agents in each environment to a centralized server where learning takes place. When the data dimension is large, the amount of information required to be communicated could be enormous. In contrast, we propose what is referred to as the federated reinforcement learning later in the literature [16], where the agents exchange the policy parameters in a decentralized communication graph. In applications with a large state representation but a much smaller policy representation, exchanging the policy parameters can be a much more compact and efficient form of communication. Moreover, we observe that a wide range of practical problems do not allow for a centralized communication topology and the agents may only communicate locally with their neighbors [17].

Our work can be considered as a special case of distributed optimization where the local

objective function is the cumulative reward in each environment. In distributed optimization problems where the objective function has strong structural properties like convexity or strong convexity, decentralized gradient descent methods have been shown to enjoy the same complexity as their centralized counterpart up to a factor that captures the connectivity of the agents [18–21].

Two-Player Zero-Sum Markov Game. The two-player zero-sum Markov game can be naturally formulated a non-convex non-concave minimax optimization program. Minimax optimization has been extensively studied in the case where the objective function is convex/concave with respect to at least one variable [22–25]. In the general non-convex non-concave setting, the problem becomes much more challenging due to the lack of strong structure, and the notion of stationarity is even unclear [26]. In [27, 28], non-convex non-concave objective functions obeying a one-sided or two-sided PŁ condition are considered, which the authors utilize to show the last-iterate convergence of GDA.

By exploiting the gradient domination condition of a Markov game with respect to each player, [29] is the first to show that the GDA algorithm provably converges to a Nash equilibrium of the Markov game. However, the convergence of the algorithm is guaranteed only on the averaged iterate rather than the more preferable last iterate. In addition, no explicit convergence rate has been given. Our work fills in this gap by designing a efficient GDA algorithm whose last iterate converges with explicit finite-time worst-case performance guarantees.

1.3 Contribution

As the first main contribution of the thesis, we present a two-time-scale stochastic optimization framework that unifies the analysis of various actor-critic algorithms in RL. Though it is originally proposed for the single-agent setting, the framework and the mathematical tools introduced therein lay the mathematical foundation for understanding and analyzing online sampled-based multi-agent RL algorithms

Second, we discuss structure of multi-agent RL in the cooperative multi-task setting and the two-player competitive setting that is previously unknown. Specifically, we show that many favorable properties of single-agent RL are violated when multiple agents are involved, including the aforementioned gradient domination condition and the existence of a deterministic optimal policy. We observe that there is an abundance of heuristic multi-agent RL algorithms built by intuitively extending single-agent RL methods (such as federated Q learning [30, 31] for the multi-task setting and gradient-descent ascent for two-player Markov games [32, 33]), and we discuss how these algorithms designed without being aware of the multi-agent RL structure can fail to achieve their desired performance.

Our third contribution is to propose and study algorithms that obey and/or exploit the special characteristics of multi-agent RL. In the cooperative multi-task setting, our work draws inspiration from distributed optimization and federated learning but combines them with RL in a way that has not been considered in the previous literature. In the two-player competitive setting, our work introduces a structured regularization that allows the GDA algorithm to provably find the optimal solution with a convergence rate that vastly improves over the best known existing result.

Finally, we use RL to develop an adaptive parameter selection mechanism for the ACOPF problem solved via ADMM, with the goal of minimizing the number of iterations until convergence. As a main contribution, our work is the first to formulate this problem in the language of RL and to develop a novel Q-learning algorithm for training the penalty parameter selection policy. Through extensive numerical simulations, we show that the RL policy can result in significantly accelerated convergence (up to a 59% reduction in the number of iterations compared to existing, curvature-informed penalty parameter selection methods). Furthermore, we show that the policy demonstrates promise for generalizability, performing well under unseen loading schemes as well as under unseen losses of lines and generators (up to a 50% reduction in iterations). Our work thus provides a successful proof-of-concept for using RL for parameter selection in power systems applications.

Organization. This dissertation is based on the published works [34–40]. The organization of the chapters is as follows. In Chapter 2, we discuss the two-time-scale stochastic optimization framework which models single-agent sample-based RL algorithms. In Chapter 3, we study the multi-task multi-agent RL problem. Chapter 4 presents a regularization-based GDA approach to two-player zero-sum Markov games. Chapter 5 applies RL to improve the solution of a power system optimization problem. We conclude and make a few remarks on possible future works in Chapter 6. As the works are mathematical in nature, we present the problem formulation, algorithms, assumptions, and main theoretical results in the main text and defer the analysis to the appendix in Chapters A-C.

CHAPTER 2

TWO-TIME-SCALE STOCHASTIC OPTIMIZATION AND ITS APPLICATIONS IN ACTOR-CRITIC ALGORITHMS

Actor-critic algorithms are an important class of data-driven techniques for policy optimization in reinforcement learning. They can be cast as optimization programs with a special type of stochastic oracle for gradient evaluations. Specifically, the gradient of the optimization variable is computed with the aid of an auxiliary variable under samples generated by a time-varying Markov chain. In this chapter, we present an abstraction of the actor-critic framework for solving general optimization programs with the same type of stochastic oracle. This optimization framework focuses on the single-agent RL setting but builds the mathematical foundation for studying and analyzing data-driven multi-agent RL algorithms.

The main contribution of this work is to characterize the finite-time and finite-sample complexity of the proposed two-time-scale stochastic gradient method under different structural properties of the objective function, namely, strong convexity, the Polyak-Lojasiewicz (PŁ) condition, and general non-convexity. Our abstraction unifies the analysis of actor-critic methods in reinforcement learning; we show how our main results can be specialized to recover the best-known convergence rate for policy optimization under an infinite-horizon average-reward Markov decision process (MDP) and to derive state-of-the-art rates for the online linear-quadratic regulator (LQR) controller and policy optimization using entropy regularization[1].

[1]The presentation in this chapter is partly adapted from [34].

2.1 Introduction

The overall goal of our optimization framework is to solve the program

$$\theta^\star = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \; f(\theta), \tag{2.1}$$

where the gradient of f is accessed through a stochastic oracle $H(\theta, \omega, X)$. The three arguments to H are the decision variable θ, an auxiliary variable $\omega \in \mathbb{R}^r$, and a random variable X drawn over a compact sample space $\mathcal{X}$. For a fixed θ, there is a single setting of the auxiliary variable, which we will denote $\omega^\star(\theta)$, such that H returns an unbiased estimate of the gradient $\nabla f(\theta)$ when X is drawn from a particular distribution μ_θ,

$$\nabla f(\theta) = \mathbb{E}_{X \sim \mu_\theta}[H(\theta, \omega^\star(\theta), X)]. \tag{2.2}$$

For other settings of $\omega \neq \omega^\star(\theta)$ or X drawn from a distribution other than μ_θ, $H(\theta, \omega, X)$ will be (perhaps severely) biased. The mapping from θ to the "optimal" setting of the auxiliary variable $\omega^\star(\theta)$ is implicit; it is determined by solving a nonlinear system of equations defined by another stochastic sampling operator $G : \mathbb{R}^d \times \mathbb{R}^r \times \mathcal{X} \to \mathbb{R}^r$. Given θ, $\omega^\star(\theta)$ is the (we will assume unique) solution to

$$\mathbb{E}_{X \sim \mu_\theta}[G(\theta, \omega^\star(\theta), X)] = 0. \tag{2.3}$$

Combining Equation 2.2 and Equation 2.3, solving Equation 2.1 is equivalent to finding $(\theta^\star, \omega^\star(\theta^\star))$ that satisfies

$$\begin{cases} \mathbb{E}_{X \sim \mu_{\theta^\star}}[H(\theta^\star, \omega^\star(\theta^\star), X)] = 0, \\ \mathbb{E}_{X \sim \mu_{\theta^\star}}[G(\theta^\star, \omega^\star(\theta^\star), X)] = 0. \end{cases} \tag{2.4}$$

In the applications we are interested in, we only have indirect access to the distribution

μ_θ. Instead of parameterizing a distribution directly, θ parameterizes a set of probability transition kernels on $\mathcal{X} \times \mathcal{X}$ through a mapping $\mathcal{P} : \mathcal{X} \times \mathbb{R}^d \to \text{dist}(\mathcal{X})$. Given θ and X, we will assume that we can generate a sample $X' \sim \mathcal{P}(\cdot|X, \theta)$ using one of these kernels. Each of these $\mathcal{P}(\cdot|\cdot, \theta)$ induces a different Markov chain and a different stationary distribution μ_θ, which is what is used in Equation 2.2 and Equation 2.3 above.

This problem structure is motivated by online algorithms for reinforcement learning. In this class of problems, three of which we describe in details in Section 2.4, we are searching for a control policy parameterized by θ that minimizes a long term cost captured by the function f. The gradient for this cost depends on the value function under the policy indexed by θ, which is specified implicitly through the Bellman equation, the analog to Equation 2.3 above. These problems often also have a mechanism for generating samples, either through experiments or simulations, that makes only implicit use of the transition kernel $\mathcal{P}$.

2.2 Related Works

Our work is closely related to the existing literature on two-time-scale stochastic approximation (SA), bi-level and composite optimization, actor-critic algorithms in reinforcement learning, and single-time-scale stochastic optimization algorithms under unbiased or biased (sub)gradients. In this section, we discuss the recent advances in these domains to give context to our contributions.

Two-Time-Scale SA. Two-time-scale SA solves a system of equations similar in form to Equation 2.4, but typically considers the setting where $\mu_\theta = \mu$ is independent of the decision variable θ. The convergence of two-time-scale SA is traditionally established by analyzing an associated ordinary differential equation [41]. Finite-time convergence of two-time-scale SA has been studied in the case where H and G are linear [42–47] and in more general nonlinear settings [48, 49], under either i.i.d. or Markovian samples. In these previous works, the analysis for the nonlinear setting is restricted to the case where H and G are both strongly monotone, while our work studies a wide range of function structures including

strong convexity, the PŁ condition, and general non-convexity.

Bi-Level and Composite Optimization. The optimization objective in our work is closely connected to the bi-level optimization framework [50–52] which solves programs structured as

$$\min_x f_1(x, y^\star(x)) \quad \text{subject to } y^\star(x) \in \operatorname*{argmin}_y f_2(x, y). \tag{2.5}$$

From the first-order optimality condition, it is clear that Equation 2.5 is equivalent to finding a stationary point (x', y') that observes

$$\nabla_x f_1(x', y') = 0, \quad \nabla_y f_2(x', y') = 0.$$

This is a special case of our objective in Equation 2.4 where G is a gradient mapping. However, in RL applications G usually abstracts the Bellman backup operator which is associated with the estimation of the value function. It is well-known that the Bellman backup operator is not the gradient of any function. In this sense, our framework is more general and suitable for modeling algorithms in RL. In addition, and similar to the works in two-time-scale SA discussed above, the analysis in [51] uses a stochastic oracle with a fixed distribution μ, while we solve Equation 2.4 with the distribution of the samples also depending on the decision variable. This is another important generalization as many realistic problems in control and reinforcement learning can only be abstracted as Equation 2.4 with μ_θ being a function of the control variable θ. Making this generalization requires generating decision-variable-dependent samples from a Markov chain whose stationary distribution shifts over iterations, which creates additional challenges in the analysis.

We also note the connection of our objective to stochastic composite optimization [53, 54], which solves optimization problems of the form

$$\min_x g_1(g_2(x)). \tag{2.6}$$

At a first glance, Equation 2.6 reduces to Equation 2.5 by choosing $g_1 = f_1$ and $g_2(x) = (x, y^\star(x))$ where $y^\star(x)$ is the minimizer of $f_2(x, \cdot)$ and therefore seems more general. However, the key assumption in stochastic composite optimization is the differentiability of g_2 (and g_1) and the access to an oracle that returns the stochastic gradient ∇g_2, which is highly unrealistic in reinforcement learning applications where only indirect information about g_2 is available.

Actor-Critic Algorithms. In the RL literature the aim of actor-critic algorithms is also to solve a problem similar to Equation 2.4, where θ and $\omega^\star(\theta)$ are referred to as the actor and critic, respectively; see for example [55–58]. Among these works, only [58] considers an online setting similar to the one studied in this paper. In fact, the algorithm studied in [58] is a special case of our framework with a non-convex objective function. Our analysis recovers the result of [58] while slightly loosening the assumptions — we are able to remove the projection operator used by [58] to limit the growth of the critic parameter.

Single-Time-Scale Stochastic Optimization. When the samples are i.i.d., stochastic gradient/subgradient algorithms are fairly well-understood for smooth (see [59–61] and the references therein) and non-smooth [62–64] functions. In the smooth setting, [36, 65, 66] study various SGD/SA algorithms under samples generated from time-invariant state transition probabilities (we will later refer to this as a time-invariant Markov chain) and show that the convergence rates are only different from that under i.i.d. samples by a logarithmic factor. The key argument used in these works is that the Markovian samples behave similarly to i.i.d. samples on a mildly dilated time scale.

In many policy optimization algorithms in RL, the samples are drawn under the control of the current policy. As the policy gets updated, the transition probabilities shift, resulting in a Markov chain with a time-varying stationary distribution (we will refer to this as the time-varying Markov chain). This setting requires more sophisticated mathematical treatment. The single-variable SA algorithm under time-varying Markovian samples is first analyzed by [67], while our paper is among the first works to extend the analysis to the

scenario where two coupled variables are updated simultaneously.

2.3 Two-Time-Scale Stochastic Gradient Descent Algorithm

In this section, we present our two-time-scale SGD method (formally stated in Algorithm 2.1) for solving Equation 2.4 under the gradient oracle discussed in Section 2.1. In the algorithm, θ_k and ω_k are estimates of $\theta^\star$ and $\omega^\star(\theta^\star)$. The random variables $\{X_k\}$ are generated by a Markov process parameterized by $\{\theta_k\}$ under the transition kernel $\mathcal{P}$, i.e.,

$$X_0 \xrightarrow{\theta_1} X_1 \xrightarrow{\theta_2} X_2 \xrightarrow{\theta_3} \cdots \xrightarrow{\theta_{k-1}} X_{k-1} \xrightarrow{\theta_k} X_k. \tag{2.7}$$

As θ_k changes in every iteration, so do the dynamics of this Markov process that generates the data. At a finite step k, X_k is in general not an i.i.d. sample from the stationary distribution μ_{θ_k}, implying that $H(\theta_k, \omega_k, X_k)$ employed in the update Equation 2.8 is not an unbiased estimate of $\nabla f(\theta_k)$ even if ω_k tracks $\omega^\star(\theta_k)$ perfectly. This sample bias, along with the gap between ω_k and $\omega^\star(\theta_k)$, affects the variables θ_{k+1} and ω_{k+1} of the next iteration and accumulates inaccuracy over time which needs a careful treatment.

The updates use two different step sizes, α_k and β_k. We choose $\alpha_k \ll \beta_k$ as a way to approximate, very roughly, the nested-loop algorithm that runs multiple auxiliary variable updates for each decision variable update. Many small critic updates get replace with a single large one. In other words, the auxiliary variable ω_k is updated at a faster time scale (larger step size) as compared to θ_k (smaller step size).

The ratio β_k/α_k can be interpreted as the time-scale difference. We will see that this ratio needs to be carefully selected based on the structural properties of the function f for the algorithm to achieve the best possible convergence. Table 2.1 provides a brief summary of our main theoretical results, which characterizes the finite-time complexity of Algorithm 2.1 and the corresponding optimal choice of step sizes under different function structures. Table 2.1 also contrasts the convergence of Algorithm 2.1 with the rates of

Algorithm 2.1: Two-Time-Scale Stochastic Gradient Descent

Initialization: the decision variable θ_0, auxiliary variable ω_0, step size sequences $\{\alpha_k\}$ for the decision variable update, $\{\beta_k\}$ for the auxiliary variable update
Observe a initial sample X_0
for $k = 1, 2, 3, ...$ **do**
Decision variable update:

$$\theta_{k+1} = \theta_k - \alpha_k H(\theta_k, \omega_k, X_k) \tag{2.8}$$

Auxiliary variable update:

$$\omega_{k+1} = \omega_k - \beta_k G(\theta_{k+1}, \omega_k, X_k) \tag{2.9}$$

Draw sample

$$X_{k+1} \sim \mathcal{P}(\cdot \mid X_k, \theta_{k+1}) \tag{2.10}$$

end for

standard SGD (which in our context means that the samples are i.i.d. and the auxiliary variable is always exactly accurate). The PŁ condition and general non-convexity cases are particularly interesting as they abstract actor-critic algorithms in RL which we now discuss.

Table 2.1: *Summary of Main Results - Time and Sample Complexity.*

Structural Property	Metric	Rate	Order of α_k, β_k	Standard SGD Rate	Applications
Strong Convexity	$\|\theta_k - \theta^\star\|^2$	$\widetilde{\mathcal{O}}(k^{-\frac{2}{3}})$	$k^{-1}, k^{-\frac{2}{3}}$	$\mathcal{O}(k^{-1})$	Gradient TD Learning
PŁ Condition	$f(\theta_k) - f(\theta^\star)$	$\widetilde{\mathcal{O}}(k^{-\frac{2}{3}})$	$k^{-1}, k^{-\frac{2}{3}}$	$\mathcal{O}(k^{-1})$	Policy Optimization for LQR, Entropy Regularized MDP
Non-convexity	$\|\nabla f(\theta_k)\|^2$	$\widetilde{\mathcal{O}}(k^{-\frac{2}{5}})$	$k^{-\frac{3}{5}}, k^{-\frac{2}{5}}$	$\mathcal{O}(k^{-\frac{1}{2}})$	Policy Optimization for Infinite-Horizon Average-Reward MDP

2.4 Applications to Actor-Critic Algorithms

In this section, we show how our results on two-time-scale optimization apply to a variety of policy evaluation and optimization algorithms in RL. The first three applications can be categorized as actor-critic algorithms for policy optimization. The objectives are non-convex in these applications, but the second and third problems are more structured and observe the PŁ condition. In Subsection 2.4.4, we briefly discuss an application of the framework to two-time-scale gradient-based policy evaluation algorithms where the objective function is strongly convex.

2.4.1 Online Actor-Critic Method for Infinite-Horizon Average-Reward MDPs

We consider the standard infinite-horizon average-reward MDP model $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ denotes the transition probabilities, and $r : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ is the reward. Our aim is to find the policy $\pi_\theta \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, parameterized by $\theta \in \mathbb{R}^d$ (where d may be much smaller than $|\mathcal{S}| \times |\mathcal{A}|$), that maximizes the average cumulative reward

$$\theta^\star = \operatorname*{argmax}_{\theta \in \mathbb{R}^d} J(\theta) \triangleq \lim_{K \to \infty} \frac{1}{K} \mathbb{E}\Big[\sum_{k=0}^{K} r\left(s_k, a_k\right) \Big] = \mathbb{E}_{s \sim \mu_\theta, a \sim \pi_\theta}[r(s, a)], \tag{2.11}$$

where μ_θ denotes the stationary distribution of the states induced by the policy π_θ. Defining the (differential) value function of the policy π_θ

$$V^{\pi_\theta}(\cdot) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \left(r\left(s_k, a_k\right) - J(\theta)\right) \mid s_0 = \cdot \Big],$$

we can use the well-known policy gradient theorem to express the gradient of the objective function in Equation 2.11 as

$$\nabla J(\theta) = \mathbb{E}_{s \sim \mu_\theta(\cdot), a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)} \Big[\left(r(s, a) - J(\theta) + V^{\pi_\theta}(s') - V^{\pi_\theta}(s)\right) \nabla \log \pi_\theta(a \mid s) \Big].$$

Optimizing Equation 2.11 with (stochastic) gradient ascent methods requires evaluating V^{π_θ} and $J(\theta)$ at the current iterate of θ, which are usually unknown and/or expensive to compute exactly. "Actor-critic" algorithms attack this problem on two scales as discussed in the sections above: an actor keeps a running estimate of the policy parameters θ_k, while a critic approximately tracks the differential value function for θ_k to aid the evaluation of the policy gradient.

For problems with large state spaces, it is often necessary to use a low-dimensional parameter $\omega \in \mathbb{R}^m$ to approximate V^{π_θ} where $m \ll |\mathcal{S}|$. In this work, we consider the linear function approximation setting where each state s is encoded by a feature vector $\phi(s) \in \mathbb{R}^m$ and the approximate value function is $\widehat{V}^{\pi_\theta,\psi}(s) = \phi(s)^\top \psi$. Under the assumptions that the Markov chain of the states induced by any policy is uniformly ergodic (equivalent of Assumption 2.5 in this context) and that the feature vectors $\{\phi(s)\}_{s\in\mathcal{S}}$ are linearly independent, it can be shown that a unique optimal pair $(J(\theta), \psi^\star(\theta))$ exists that solves the projected Bellman equation

$$\mathbb{E}_{s\sim\mu_\theta(\cdot),a\sim\pi_\theta(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a)}\begin{bmatrix} J(\theta) - r(s,a) \\ \big(r(s,a) - J(\theta) + \phi(s')^\top\psi^\star(\theta) - \phi(s)^\top\psi^\star(\theta)\big)\,\phi(s) \end{bmatrix} = 0.$$

We use an auxiliary variable $\omega = (\hat{J}, \psi)$ to track the solution to this Bellman equation.

Due to the limit in the representational power of the function approximation, there is an approximation error between V^{π_θ} and $\hat{V}^{\pi_\theta,\psi^\star(\theta)}$ as a function of θ, which we define over the stationary distribution

$$\epsilon_{\text{approx}}(\theta) = \sqrt{\mathbb{E}_{s\sim\mu_\theta}\left[(\phi(s)^\top\psi^\star(\theta) - V^{\pi_\theta}(s))^2\right]}.$$

We assume the existence of a constant $\epsilon_{\text{approx}}^{\max}$ such that $\epsilon_{\text{approx}}(\theta) \leqslant \epsilon_{\text{approx}}^{\max}$ for all $\theta \in \mathbb{R}^d$.

Comparing this problem with Equation 2.2 and Equation 2.3, it is clear that this is a

special case of our optimization framework with $X = (s, a, s')$, $\omega^\star(\theta) = (J(\theta), \psi^\star(\theta))$ and

$$f(\theta) = -J(\theta), \;\; G(\theta, \omega, X) = [\hat{J} - r(s,a), (r(s,a) - \hat{J} + \phi(s')^\top \psi - \phi(s)^\top \psi)\phi(s)^\top]^\top,$$
$$H(\theta, \omega, X) = -(r(s,a) - \hat{J} + \phi(s')^\top \psi - \phi(s)^\top \psi + \varepsilon_{\text{approx}}(\theta))\nabla \log \pi_\theta(a \mid s),$$

where $\varepsilon_{\text{approx}}(\theta)$ is an error in the gradient of the actor carried over from the approximation error of the critic which can be upper bounded by $2\epsilon^{\max}_{\text{approx}}$ in expectation. In this case, the function $-J(\theta)$ is non-convex and our two-time-scale SGD algorithm is guaranteed to find a stationary point of the objective function with rate $\widetilde{\mathcal{O}}(k^{-2/5})$, up to errors proportional to $\epsilon^{\max}_{\text{approx}}$. This rate matches the state-of-the-art result derived in [58]. A subtle improvement of our analysis is that we do not need to perform the projection of the critic parameter onto a compact set that [58] requires in every iteration of the algorithm to guarantee the stability of the critic.

2.4.2 Online Natural Actor-Critic Algorithm for LQR

In this section, we consider the infinite-horizon average-cost LQR problem

$$\begin{aligned} \underset{\{u_k\}}{\text{minimize}} \quad & \lim_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\sum_{k=0}^{T} \big(x_k^\top Q x_k + u_k^\top R u_k\big) \mid x_0 \Big] \\ \text{subject to} \quad & x_{k+1} = Ax_k + Bu_k + w_k, \end{aligned} \tag{2.12}$$

where $x_k \in \mathbb{R}^{d_1}$, $u_k \in \mathbb{R}^{d_2}$ are the state and the control variables, $w_k \sim N(0, \Psi) \in \mathbb{R}^{d_1}$ is time-invariant system noise, $A \in \mathbb{R}^{d_1 \times d_1}$ and $B \in \mathbb{R}^{d_1 \times d_2}$ are the system transition matrices, and $Q \in \mathbb{R}^{d_1 \times d_1}$, $R \in \mathbb{R}^{d_2 \times d_2}$ are positive-definite cost matrices. It is well-known (see, for example, [68, Chap. 3.1]) that the optimal control sequence $\{u_k\}$ that solves Equation 2.12 is a time-invariant linear function of the state

$$u_k^\star = -K^\star x_k, \tag{2.13}$$

where $K^\star \in \mathbb{R}^{d_2 \times d_1}$ is a matrix that depends on the problem parameters A, B, Q, R. This fact will allow us to reformulate the LQR as an optimization program over the feedback gain matrix K. It is also true that optimizing over the set of stochastic controllers

$$u_k = -Kx_k + \sigma\epsilon_k, \quad \epsilon_k \sim \mathbf{N}(0, \sigma^2\mathbf{I}),$$

with $\sigma \geqslant 0$ fixed will in the end yield the same optimal $K^\star$ [69]. In the RL setting considered below, we will optimize over this class of stochastic controller as it encourages exploration. Defining $\Psi_\sigma = \Psi + \sigma^2 BB^\top$, we can re-express the LQR problem as

$$\begin{aligned} \underset{K}{\text{minimize}}\ & J(K) \triangleq \operatorname{trace}(P_K \Psi_\sigma) + \sigma^2 \operatorname{trace}(R) \\ \text{s.t. } & P_K = Q + K^\top R K + (A - BK)^\top P_K (A - BK). \end{aligned} \tag{2.14}$$

Our goal is to solve Equation 2.14 when the system transition matrices A and B are unknown[2] and we take online samples from a single trajectory of states $\{x_k\}$ and control inputs $\{u_k\}$. This problem has been considered recently in [70], and in fact much of our formulation is modeled on this work. The essential difference is that while [70] works in the "batch" setting, where multiple trajectories are drawn for a fixed feedback gain estimate, our algorithm is entirely online.

Given a feedback gain K, we define

$$E_K \triangleq 2\left(R + B^\top P_K B\right) K - 2B^\top P_K A.$$

It turns out that the natural policy gradient of the objective function in Equation 2.14, which we denote by $\widetilde{\nabla} J$, is $\widetilde{\nabla} J(K) = E_K$.

To track E_K when A and B are unknown it suffices to estimate $R + B^\top P_K B$ and $B^\top P_K A$. We define

[2]We do assume, however, that we know the cost matrices Q and R or at least we can compute $x^\top Q x + u^\top R u$ for any x and u.

$$\Omega_K = \begin{pmatrix} \Omega_K^{11} & \Omega_K^{12} \\ \Omega_K^{21} & \Omega_K^{22} \end{pmatrix} = \begin{pmatrix} Q + A^\top P_K A & A^\top P_K B \\ B^\top P_K A & R + B^\top P_K B \end{pmatrix}, \tag{2.15}$$

of which $R + B^\top P_K B$ and $B^\top P_K A$ are sub-matrices. We define the operator $\text{svec}(\cdot)$ as the vectorization of the upper triangular sub-matrix of a symmetric matrix with off-diagonal entries weighted by $\sqrt{2}$, and define $\text{smat}(\cdot)$ as the inverse of $\text{svec}(\cdot)$. We also define $\phi(x, u) = \text{svec}(\left[x^\top, u^\top\right]^\top \left[x^\top, u^\top\right])$ for any $x \in \mathbb{R}^{d_1}, u \in \mathbb{R}^{d_2}$. Then, it can be shown that Ω_K and $J(K)$ jointly satisfy the Bellman equation

$$\mathbb{E}_{x\sim\mu_K, u\sim N(-Kx,\sigma^2 I)}\left[M_{x,u,x',u'}\right] \begin{bmatrix} J(K) \\ \text{svec}(\Omega_K) \end{bmatrix} = \mathbb{E}_{x\sim\mu_K, u\sim N(-Kx,\sigma^2 I)}\left[c_{x,u}\right], \tag{2.16}$$

where the matrix $M_{x,u,x',u'}$ and vector $c_{x,u}$ are

$$M_{x,u,x',u'} = \begin{bmatrix} 1 & 0 \\ \phi(x,u) & \phi(x,u)\left[\phi(x,u) - \phi\left(x',u'\right)\right]^\top \end{bmatrix}, \; c_{x,u} = \begin{bmatrix} x^\top Q x + u^\top R u \\ \left(x^\top Q x + u^\top R u\right)\phi(x,u) \end{bmatrix}.$$

The solution to Equation 2.16 is unique if K is stable with respect to A and B [70]. An auxiliary variable $\hat{\Omega}$ can be introduced to track Ω_K for the decision variable K.

We connect this to our optimization framework by noting that Equation 2.16 corresponds to Equation 2.3 with K, (x, u, x', u'), and $[J(K), \text{svec}(\Omega_K)^\top]^\top$ mirroring θ, X, and $\omega^\star(\theta)$, respectively. The natural gradient oracle in this case is $H(\theta, \omega, X) = 2\hat{\Omega}^{22}K - 2\hat{\Omega}^{21}$, which does not depend on the samples X directly, and the operator G is $G(\theta, \omega, X) = -M_{x,u,x',u'}\omega + c_{x,u}$. A key structure of Equation 2.12 is that the objective function is non-convex but observes the PŁ condition [70], which we formally define later in Assumption 2.8. As a result, applying Algorithm 2.1 to this problem leads to an online actor-critic flavored algorithm that converges with rate $\widetilde{\mathcal{O}}(k^{-2/3})$ under proper assumptions. To our best knowledge, our work is the first to study the online actor-critic method for solving the LQR, and our result vastly improves over the rate $\widetilde{\mathcal{O}}(k^{-1/5})$ of the nested-loop actor-critic algorithm

derived in [70] which also operates under more restrictive assumptions (e.g. sampling from the stationary distribution, boundedness of the iterates).

2.4.3 Online Actor-Critic Method for Regularized MDPs

As a third application of our framework, we study the policy optimization problem for the infinite-horizon discounted-reward MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ where $\gamma \in (0, 1)$ is the discount factor and the rest are defined in the same manner as above in Section Subsection 2.4.1. We restrict our attention to the tabular setting where the parameter θ encodes the policy through the softmax function

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}.$$

To accelerate the convergence of the actor-critic algorithm, we regularize the objective with the policy entropy as proposed by [13]. Specifically, with regularization weight $\tau > 0$, the regularized value function of a policy π is

$$V_\tau^\pi(s) = \mathbb{E}_{a_k \sim \pi(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)} \Big[\sum_{k=0}^{\infty} \gamma^k \big(r(s_k, a_k) - \tau \log \pi(a_k \mid s_k) \big) \mid s_0 = s \Big].$$

Under the initial state distribution $\rho \in \Delta_{\mathcal{S}}$, the expected cumulative reward collected by policy π is $J_\tau(\pi) = \mathbb{E}_{s \sim \rho}[V_\tau^\pi(s)]$. We consider solving the policy optimization problem

$$\max_\pi J_\tau(\pi).$$

Expressing the gradient of the objective with the policy gradient theorem, we have

$$\nabla_\theta J_\tau(\pi_\theta) = \frac{1}{1-\gamma} \mathbb{E}\big[\left(r(s,a) - \tau \log \pi_\theta(a \mid s) + \gamma V_\tau^{\pi_\theta}(s') - V_\tau^{\pi_\theta}(s) \right) \nabla_\theta \log \pi_\theta(a \mid s) \big],$$

where the expectation is taken over $s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot \mid s), s' \sim \mathcal{P}(\cdot \mid s, a)$, and the the

discounted visitation distribution $d_\rho^\pi \in \Delta_{\mathcal{S}}$ is defined as

$$d_\rho^\pi(s) = (1-\gamma)\mathbb{E}_{a_k \sim \pi(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)}[\sum_{k=0}^{\infty} \gamma^k \mathbf{1}(s_k = s) \mid s_0 \sim \rho].$$

To evaluate the gradient, we need to compute the regularized value function $V_\tau^{\pi_\theta}$, which is the solution to the following Bellman equation

$$\mathbb{E}_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a)}\Big[r(s,a) - \tau \log \pi_\theta(a \mid s) + \gamma V_\tau^{\pi_\theta}(s') - V_\tau^{\pi_\theta}(s)\Big] = 0.$$

Interestingly, [71] shows that we can regard d_ρ^π as the stationary distribution under π in an environment with the modified transition probability

$$\widetilde{P}(\cdot \mid s,a) = \gamma P(\cdot \mid s,a) + (1-\gamma)\rho(\cdot).$$

This observation allows us to generate Markovian samples (s,a,s') with $d_\rho^{\pi_\theta} \otimes \pi_\theta \otimes \mathcal{P}$ as the stationary distribution, in an online manner that resembles [72][Algorithm 1].

In the actor-critic framework, we introduce a critic (auxiliary variable) $\widehat{V} \in \mathbb{R}^{|\mathcal{S}|}$ to estimate the solution of the Bellman equation under the current policy iterate. Our optimization framework abstracts this problem by choosing

$$\begin{aligned}
&X = (s,a,s'), \;\; \omega = \widehat{V}, \;\; f(\theta) = -J_\tau(\pi_\theta), \\
&H(\theta,\omega,X) = \frac{1}{1-\gamma}(r(s,a) - \tau \log \pi_\theta(a \mid s) + \gamma \widehat{V}(s') - \widehat{V}(s))\nabla \log \pi_\theta(a \mid s), \\
&G(\theta,\omega,X) = r(s,a) - \tau \log \pi_\theta(a \mid s) + \gamma \widehat{V}(s') - \widehat{V}(s).
\end{aligned}$$

The objective function is non-convex but satisfies the PŁ condition under standard assumptions (see [13][Lemma 15]). Our two-time-scale SGD framework specializes to an online actor-critic algorithm, which by our analysis to be discussed later in Subsection 2.6.2 is guaranteed to find the globally optimal solution of the regularized objective with rate

$\widetilde{\mathcal{O}}(k^{-2/3})$. To our best knowledge, this is the first time such data-driven algorithms are studied for solving an entropy-regularized MDP in the tabular setting. Compared with the result presented in Subsection 2.4.1, the introduction of the entropy regularization leads to an accelerated convergence rate. We note that the gap between the solutions to the regularized and original MDP is proportional to the regularization weight τ [39, 73]. By carefully choosing τ, solving the regularized MDP provides a reliable and efficient way to find the approximate solution of the original unregularized MDP.

2.4.4 Two-Time-Scale Policy Evaluation Algorithms

Our framework also abstracts GTD (gradient temporal-difference), GTD2 , and TDC (temporal difference learning with gradient correction) algorithms [74, 75], which are gradient-based two-time-scale policy evaluation algorithms in RL. They can be viewed as degenerate special cases of our framework where the expectation in Equation 2.4 is taken over a fixed distribution μ that does not depend on θ, and therefore do not require the full capacity of our framework. The objective function in this problem is strongly convex, and our framework under proper assumptions guarantees a convergence rate of $\widetilde{\mathcal{O}}(k^{-2/3})$, which matches the analysis in [76]. As this subject is well-studied, we skip the detailed discussion of the problem formulation and algorithm statement and refer interested readers to [74–76].

2.5 Technical Assumptions

In this section, we present the main technical assumptions important in our later analysis. We first consider the Lipschitz continuity of H and G.

Assumption 2.1. *There exists a constant $L > 0$ such that for all $\theta_1, \theta_2 \in \mathbb{R}^d$, $\omega_1, \omega_2 \in \mathbb{R}^r$, and $X \in \mathcal{X}$*

$$\begin{aligned}
&\|H(\theta_1, \omega_1, X) - H(\theta_2, \omega_2, X)\| \leqslant L\left(\|\theta_1 - \theta_2\| + \|\omega_1 - \omega_2\|\right), \\
&\|G(\theta_1, \omega_1, X) - G(\theta_2, \omega_2, X)\| \leqslant L\left(\|\theta_1 - \theta_2\| + \|\omega_1 - \omega_2\|\right).
\end{aligned} \tag{2.17}$$

We also assume that the objective function f is L-smooth.

Assumption 2.2. *There exists a constant $L > 0$ such that for all $\theta_1, \theta_2 \in \mathbb{R}^d$*

$$\|\nabla f(\theta_1) - \nabla f(\theta_2)\| \leqslant L\|\theta_1 - \theta_2\|. \tag{2.18}$$

Assumption 2.1 and Assumption 2.2 are common in the literature of stochastic approximation [49, 77] and hold in the actor-critic methods discussed in Section 2.4. Next, we assume that the operator $G(\theta, \cdot, X)$ is strongly monotone in expectation at $\omega^\star(\theta)$ (which we have assumed is unique).

Assumption 2.3. *There exists a constant $\lambda > 0$ such that*

$$\langle \mathbb{E}_{X\sim\mu_\theta}[G(\theta, \omega, X)], \omega - \omega^\star(\theta)\rangle \leqslant -\lambda\|\omega - \omega^\star(\theta)\|^2, \quad \forall \theta \in \mathbb{R}^d, \omega \in \mathbb{R}^r.$$

This assumption is often made in the existing literature on two-time-scale stochastic approximation [49, 51] and is a sufficient condition to guarantee the fast convergence of the auxiliary variable iterate. This assumption essentially states that G behaves similarly to the gradient of a strongly convex function in expectation, though it may not even be a gradient mapping. It can be verified that Assumption 2.3 holds in the actor-critic methods discussed in Section 2.4.

In addition, we assume that $\omega^\star(\cdot)$ is Lipschitz continuous with respect to θ.

Assumption 2.4. *There exists a constant $L, B > 0$ such that*

$$\|\omega^\star(\theta) - \omega^\star(\theta')\| \leqslant L\|\theta - \theta'\|, \quad \|\omega^\star(\theta)\| \leqslant B, \quad \forall \theta, \theta' \in \mathbb{R}^d.$$

Given two probability distributions μ_1 and μ_2 over the space $\mathcal{X}$, their total variation (TV)

distance is defined as

$$d_{\mathrm{TV}}(\mu_1, \mu_2) = \frac{1}{2} \sup_{\nu:\mathcal{X}\to[-1,1]} \left| \int \nu d\mu_1 - \int \nu d\mu_2 \right|. \tag{2.19}$$

The definition of the mixing time of a Markov chain $\{X_k\}$ is given as follows.

Definition 2.1. *Consider the Markov chain $\{X_k^\theta\}$ generated according to $X_k^\theta \sim \mathcal{P}(\cdot \mid X_{k-1}^\theta, \theta)$, and let μ_θ be its stationary distribution. For any $\alpha > 0$, the mixing time of the chain $\{X_k^\theta\}$ corresponding to α is defined as*

$$\tau_\theta(\alpha) = \min\{k \in \mathbb{N} : \sup_{X\in\mathcal{X}} d_{TV}(P(X_k^\theta = \cdot \mid X_0^\theta = X), \mu_\theta(\cdot)) \leqslant \alpha\}.$$

The mixing time $\tau_\theta(\alpha)$ essentially measures the time needed for the Markov chain $\{X_k^\theta\}$ to approach its stationary distribution [78]. We next consider the following important assumption that guarantees that the Markov chain induced by any static θ "mixes" geometrically.

Assumption 2.5. *Given any θ, the Markov chain $\{X_k\}$ generated by $\mathcal{P}(\cdot \mid \cdot, \theta)$ has a unique stationary distribution μ_θ and is uniformly geometrically ergodic. In other words, there exist constants $m > 0$ and $\rho \in (0, 1)$ independent of θ such that*

$$\sup_{X\in\mathcal{X}} d_{TV}(P(X_k = \cdot \,|X_0 = X,\, \theta), \mu_\theta(\cdot)) \leqslant m\rho^k \text{ for all } \theta \in \mathbb{R}^d \text{ and } k \geqslant 0.$$

Denoting $\tau(\alpha) \triangleq \sup_{\theta\in\mathbb{R}^d} \tau_\theta(\alpha)$, this assumption implies that there exists a positive constant C depending only on ρ and m such that

$$\tau(\alpha) \leqslant C \log(1/\alpha). \tag{2.20}$$

Assumption 2.5 is again standard in the existing literature [58, 67, 79].

We also consider the following assumption on the ensemble of transition kernels.

Assumption 2.6. *Given two distributions $d, \hat{d}$ over $\mathcal{X}$ and parameters $\theta, \hat{\theta} \in \mathbb{R}^d$, we draw the samples according to $X \sim d, X' \sim \mathcal{P}(\cdot \mid X, \theta)$ and $\hat{X} \sim \hat{d}, \hat{X}' \sim \mathcal{P}(\cdot \mid \hat{X}, \hat{\theta})$. We assume that there exists a constant $L > 0$ such that*

$$d_{TV}(P(X' = \cdot), P(\hat{X}' = \cdot)) \leqslant d_{TV}(d, \hat{d}) + L\|\theta - \hat{\theta}\|. \tag{2.21}$$

In addition, we assume that the stationary distribution is Lipschitz in θ

$$d_{TV}(\mu_\theta, \mu_{\hat{\theta}}) \leqslant L\|\theta - \hat{\theta}\|. \tag{2.22}$$

This assumption amounts to a regularity condition on the transition probability matrix $\mathcal{P}(\cdot \mid \cdot, \theta)$ as a function of θ, and has been shown to hold in the reinforcement learning setting (see, for example, [58, Lemma A1]). Without any loss of generality, we use the same constant L in Assumption 2.1–Assumption 2.6 and assume $B \geqslant 1$. We define

$$D = \max\{L + \max_{X \in \mathcal{X}} \|G(0, 0, X)\|, \|\omega^*(0)\|, B\}, \tag{2.23}$$

which is a finite constant since $\mathcal{X}$ is compact. A simple consequence of Assumption 2.1 is that for all $\theta \in \mathbb{R}^d, \omega \in \mathbb{R}^r, X \in \mathcal{X}$

$$\|G(\theta, \omega, X)\| \leqslant D(\|\theta\| + \|\omega\| + 1), \quad \text{and} \quad \|\omega^\star(\theta)\| \leqslant D(\|\theta\| + 1). \tag{2.24}$$

Finally, we assume the optimal solution set $\{\theta^\star : f(\theta^\star) \leqslant f(\theta),\ \forall \theta \in \mathbb{R}^d\}$ is non-empty.

2.6 Finite-Time and Finite-Sample Complexity of Two-Time-Scale SGD

This section presents the main results of this paper, which are the finite-time and finite-sample convergence of Algorithm 2.1 under three structural properties of the objective function, namely, strong convexity, non-convexity with the PŁ condition, and general non-

convexity. Our results are derived under the assumptions introduced in Section 2.5, which we assume always hold in the rest of this paper.

The convergence of Algorithm 2.1 relies on $\alpha_k, \beta_k \to 0$ with reasonable rates. As mentioned in Section 2.3, α_k needs to be much smaller than β_k to approximate the nested-loop algorithm where multiples auxiliary variable updates are performed for each decision variable update. Therefore, we consider the following choices of step sizes

$$\alpha_k = \frac{\alpha_0}{(k+1)^a}, \quad \beta_k = \frac{\beta_0}{(k+1)^b}, \quad \forall k \geqslant 0, \tag{2.25}$$

where a, b, α_0, β_0 are some constants satisfying $0 < b \leqslant a \leqslant 1$ and $0 < \alpha_0 \leqslant \beta_0$. Given α_k, recall from Definition 2.1 that $\tau(\alpha_k)$ is the mixing time associated with α_k. In the sequel, for convenience we denote $\tau_k \triangleq \tau(\alpha_k)$. Since $\tau_k \leqslant C \log((k+1)^a/\alpha_0)$ (from Equation 2.20), we have $\lim_{k\to\infty} \alpha_k \tau_k^2 = \lim_{k\to\infty} \beta_k \tau_k^2 = 0$. This implies that there exists a positive integer $\mathcal{K}$ such that

$$\beta_{k-\tau_k} \tau_k^2 \leqslant \min\Big\{1, \frac{1}{6LB}, \frac{\lambda}{22C_1 + 32D^2}, \frac{\lambda}{2C_2}, \frac{\lambda^3}{32L^2C_2}\Big\}, \quad \forall k \geqslant \mathcal{K}, \tag{2.26}$$

where C_1 and C_2 are positive constants defined as

$$C_1 = 18D^2 + 20LDB, \; C_2 = (4D^2+1)(4C_1 + 32D^2) + \frac{2L^2B^2}{\lambda} + 2L^2B^2. \tag{2.27}$$

In addition, there exists a constant $c_\tau \in (0,1)$ such that for any $k > \tau_k$ we have

$$\tau_k \leqslant (1-c_\tau)k + (1-c_\tau), \quad \text{and} \quad c_\tau(k+1) \leqslant k - \tau_k + 1 \leqslant k+1. \tag{2.28}$$

We carefully come up with the constants and conditions in Equation 2.26 and Equation 2.27 to prevent an excessively large step size from destroying the stability of the updates. We stress that $\mathcal{K}$ is a constant that only depends on the quantities involved in the step sizes

in Equation 2.25.

2.6.1 Strong Convexity

We consider the following assumption on function f.

Assumption 2.7. *The function f is strongly convex with constant λ*[3]

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle + \frac{\lambda}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{2.29}$$

Theorem 2.1 (Strongly Convex)**.** *Suppose that Assumption 2.7 holds. Let the step size sequences $\{\alpha_k\}$ and $\{\beta_k\}$ satisfy Equation 2.25 with*

$$a = 1, \quad b = 2/3, \quad \alpha_0 \geqslant \frac{4}{\lambda}, \quad \textit{and} \quad \frac{\alpha_0}{\beta_0} \leqslant \frac{1}{2}.$$

Then for all $k \geqslant \mathcal{K}$ where $\mathcal{K}$ is defined in Equation 2.26, we have

$$\begin{aligned}\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \leqslant \frac{\mathcal{K}+1}{k+1}\left(\mathbb{E}\left[\|\theta_{\mathcal{K}} - \theta^\star\|^2\right] + \frac{4L^2\alpha_0}{\lambda^2\beta_0}\mathbb{E}[\|\omega_{\mathcal{K}} - \omega^\star(\theta_{\mathcal{K}})\|^2]\right) \\ + \frac{C^2\log^2((k+1)/\alpha_0)}{3(k+1)^{2/3}}\left((6C_2 + 2B^2 + \frac{8L^2C_2}{\lambda^2}(2\|\theta^\star\|^2+1))\frac{\alpha_0\beta_0}{c_\tau} + \frac{16L^4B^2\alpha_0^3}{\lambda^3\beta_0^2}\right).\end{aligned}$$

Our theorem states that when f is strongly convex the iterates of Algorithm Algorithm 2.1 converge to the optimal solution with rate $\widetilde{\mathcal{O}}(k^{-2/3})$. Comparing with the deterministic gradient descent setting where the convergence rate is linear and the standard SGD setting where the convergence rate is $\mathcal{O}(k^{-1})$, our result reflects the compromise in the convergence rate due to the gradient noise and inaccurate auxiliary variable. Compared with the convergence rate of the two-time-scale SA algorithm for bi-level optimization [51] under i.i.d.

[3]Without any loss of generality, we slightly overload λ, the strong monotonicity constant of the operator G in Assumption 2.3, to denote the strong convexity constant here.

samples, our rate is the same up to a logarithmic factor which naturally arises from the bias caused by the time-varying Markovian samples.

2.6.2 Non-Convexity under PŁ Condition

We also study the convergence of Algorithm 2.1 under the following condition.

Assumption 2.8. *There exists a constant $\lambda > 0$ such that*

$$\frac{1}{2}\|\nabla f(x)\|^2 \geqslant \lambda\left(f(x) - f^\star\right), \quad \forall x \in \mathbb{R}^d.$$

This is known as the PŁ condition and is introduced in [80, 81]. The PŁ condition does not imply convexity, but guarantees the linear convergence of the objective function value when gradient descent is applied to solve a non-convex optimization problem [82], which resembles the convergence rate of gradient descent for strongly convex functions. Recently, this condition has been observed to hold in many important practical problems such as supervised learning with an over-parametrized neural network [83] and the linear quadratic regulator in optimal control [70, 84].

Theorem 2.2 (PŁ Condition)**.** *Suppose the function f satisfies Assumption 2.8. In addition, we assume that the stochastic gradient is bounded, i.e. there exists a constant $B > 0$ such that* [4]

$$\|H(\theta, \omega, X)\| \leqslant B, \quad \forall \theta \in \mathbb{R}^d, \omega \in \mathbb{R}^r, X \in \mathcal{X}. \tag{2.30}$$

Let the step size sequences $\{\alpha_k\}$ and $\{\beta_k\}$ satisfy Equation 2.25 with

$$a = 1, \quad b = 2/3, \quad \alpha_0 \geqslant \max\{1, \frac{2}{\lambda}\}, \quad \textit{and} \quad \frac{\alpha_0}{\beta_0} \leqslant \frac{1}{4}.$$

[4] Again, for the convenience of notation, we use the same constant B as in Assumption 2.4.

Then for all $k \geqslant \mathcal{K}$ where $\mathcal{K}$ is defined in Equation 2.26, we have

$$\mathbb{E}\left[f(\theta_k) - f^\star\right] \leqslant \frac{\mathcal{K}+1}{k+1}\left(\mathbb{E}\left[f(\theta_{\mathcal{K}}) - f^\star\right] + \frac{2L^2\alpha_0}{\lambda\beta_0}\mathbb{E}[\|\omega_{\mathcal{K}} - \omega^\star(\theta_{\mathcal{K}})\|^2]\right) + \frac{2C^2C_3\log^4(k+1)}{3(k+1)^{2/3}},$$

where $C_3 = \frac{150L^2B^3\alpha_0^2}{c_\tau} + \frac{48L^2C_2}{\lambda}(\|\theta_0\|^2 + B^2\alpha_0^2 + 1)\frac{\alpha_0\beta_0}{c_\tau} + \frac{48L^4B^2\alpha_0^3}{\lambda^2\beta_0^2}$.

Under the PŁ condition, we show that $f(\theta_k)$ converges to the optimal function value $f^\star$ with rate $\widetilde{\mathcal{O}}(k^{-2/3})$. This is the same rate as if f is strongly convex. However, in this case the convergence is measured in the function value, whereas under strong convexity the iterates θ_k converge to the unique global minimizer. The convergence rates of deterministic gradient descent and standard SGD under the PŁ condition also match those in the strongly convex case. To our best knowledge, functions exhibiting the PŁ condition have not been studied in the bi-level optimization framework.

2.6.3 Non-Convexity

Finally, we study the case where the objective function f is non-convex and smooth. In general, we cannot find an optimal solution and may only reach a stationary point. Analyzing the convergence without any convexity or PŁ condition is more challenging, and we need to make an additional assumption to ensure stability.

Assumption 2.9. *There exists a constant $L > 0$ such that*

$$\|G(\theta_1, \omega_1, X) - G(\theta_2, \omega_2, X)\| \leqslant L(\|\omega_1 - \omega_2\| + 1), \forall \theta_1, \theta_2 \in \mathbb{R}^d, \omega_1, \omega_2 \in \mathbb{R}^r, X \in \mathcal{X}.$$

We note that this assumption holds in the actor-critic algorithm discussed in Subsection 2.4.1 where G does not depend on θ, as well as in problems where G is bounded in θ.

Theorem 2.3 (Non-convex)**.** *Let the step size $\{\alpha_k\}$ and $\{\beta_k\}$ satisfy Equation 2.25 with*

$a = 3/5$, *and* $b = 2/5$. *Under Assumption 2.9, we have for all* $k \geqslant \mathcal{K}$

$$\min_{t \leqslant k} \mathbb{E}[\|\nabla f(\theta_t)\|^2] \leqslant \frac{4}{\alpha_0 (k+1)^{2/5}} \mathbb{E}[f(\theta_{\mathcal{K}}) - f^\star] + \frac{4L^2}{\beta_0 \lambda (k+1)^{2/5}} \mathbb{E}[\|\omega_{\mathcal{K}} - \omega^\star(\theta_{\mathcal{K}})\|^2] + \left(\frac{25 L^2 B^3 \alpha_0^2}{2 c_\tau} + \frac{2 L^4 B^2 \alpha_0^3}{\lambda \beta_0^2} + \frac{L^2 \alpha_0 \beta_0 C_2}{c_\tau \lambda} \right) \frac{8 \tau_k^2 \log(k+1)}{5 \log(2) c_\tau (k+1)^{2/5}}.$$

Our theorem in the non-convex case shows the convergence of the two-time-scale SGD algorithm to a stationary point of the objective function (measured by the squared norm of the gradient) with rate $\widetilde{\mathcal{O}}(k^{-2/5})$. One may contrast this with the convergence rate of deterministic gradient descent $\mathcal{O}(k^{-1})$ and standard SGD $\mathcal{O}(k^{-1/2})$ to see the cost of the gradient noise and auxiliary variable inaccuracy. Compared with the bi-level optimization algorithm under i.i.d. samples [51], our rate is again the same up to a logarithmic factor due to the time-varying Markovian samples.

2.7 Conclusion

The main contribution of our work in this chapter is to introduce a novel stochastic optimization framework, which allows us to plug-and-play various data-driven algorithms, especially actor-critic algorithms, in RL and control. Specialized to certain RL settings, our two-time-scale SGD algorithm and its analysis recover existing algorithms with their state-of-the-art convergence rates. In some other settings, our two-time-scale SGD algorithm translates to new algorithms that were previously unknown and/or enjoy superior convergence properties. This framework mainly targets single-agent RL problems but lays the mathematical foundation for understanding and analyzing algorithms in the multi-agent settings.

CHAPTER 3

MULTI-AGENT MULTI-TASK REINFORCEMENT LEARNING

The aim of our work in this chapter is to solve a multi-task RL problem using a network of agents. Each task, characterized by a different MDP, is assigned to one agent. Although each agent only makes observations and acts in its own environment, their goal is to collectively learn a policy that performs well across all environments by sharing information with each other. We do not require the state spaces to be the same in each of the environments. In general, the learned policy is a mapping from the union of state spaces to the action space.

Existing approaches to the multi-task RL problem [14, 15, 85] are mostly heuristic in nature and typically use a specific "master/worker" model for agent interaction where worker agents independently collect observations in their respective environments, which are then summarized (perhaps through a gradient computation) and reported to a central master. We are interested in understanding multi-task RL under a more flexible, decentralized communication model where the agents only share information with a small subset of other agents and in developing algorithms with provable convergence guarantees.

To this end, we first present a clean mathematical formulation for multi-task RL over a network of agents. We study the structure of the underlying optimization objective and show how multi-task RL is fundamentally more challenging to solve than its single-task counterpart through two simple yet illustrative examples.

Despite the challenges, framing the problem in the language of distributed optimization allows us to develop a decentralized policy gradient algorithm that finds a single policy effective for each of the tasks. We provide theoretical guarantees for the performance of our decentralized policy gradient algorithm. Specifically, we show that in the tabular setting, the algorithm converges to a stationary point of the global (non-concave) objective. Under a further assumption on the structure of environments' dynamics, the algorithm is guaranteed

to find the global optimality.

We demonstrate the effectiveness of the proposed method using numerical experiments on challenging multi-task RL problems. Our small-scale "Grid World" problems, which can be reliably solved using a complete tabular representation for the policy, demonstrate how the decentralized policy gradient algorithm balances the interests of the agents in different environments. Our experiments for learning to navigate airborne drones in multiple (simulated) environments show that the algorithm can be scaled to real-life problems that require a significant amount of training data and use complicated function approximations (such as neural networks) to parameterize the policy.

Inspired by the numerical simulations, we propose any formulation of multi-task RL under the constrained MDP framework, to control the performance of the policy in a more fine-grained manner. Under the assumption that the MDPs behind environments operate under the same dynamics, we propose a natural policy gradient based algorithm that efficiently and provably converges to the globally optimal policy, both in objective function and in constraint violation. We then extend this algorithm to the sample-based setting where we do not know the transition probability kernel of the environments, by introducing local "critic" variables that track the local value functions. We present the finite-sample complexity of this algorithm[1].

3.1 Related Works

In recent years, multi-task RL has become an emerging topic as a way to scale up RL solutions. This topic has received a surge of interests, and a number of solutions have been proposed for solving this problem, including policy distillation [86, 87], distributed RL algorithms over actors/learner networks [14, 15, 85, 88], and transfer learning [89, 90]. Distributed parallel computing has also been applied to speed up RL algorithms for solving single task problems [91–93].

[1]The presentation in this chapter is partly adapted from [35–38].

Similar to our work, [14, 15] also aim to solve MTRL with policy gradient algorithms in a distributed manner. These works propose sharing the local trajectories/data collected by workers in each environment to a centralized server where learning takes place. When the data dimension is large, the amount of information required to be exchanged could be enormous. In contrast, exchanging the policy parameters could be a more compact and efficient form of communication in applications with a large state representation but a much smaller policy representation. Moreover, we observe that a wide range of practical problems do not allow for a centralized communication topology [17]. Motivated by these observations, we consider a decentralized policy gradient method where the agents only exchange their policy parameters according to a decentralized communication graph. This makes our work fundamentally different from the existing literature. Indeed, our work can be considered as a decentralized and multi-task variant of the policy gradient method studied in [12], where the authors consider a single-task RL.

Other works in meta-learning and transfer learning also essentially aim to achieve MTRL, where these two methods essentially attempt to reduce the resources required to learn a new task by utilizing related existing information; see for example [10, 94, 95]. Our work is fundamentally different from these papers, where we address MTRL by leveraging the collaboration between a number of agents.

We also note some relevant works on decentralized algorithms in multi-agent reinforcement learning (MARL), where a group of agents operate in a common environment and aim to solve a single task [96–105]. The setting in these work is different from ours since we consider multi-task RL, which is more challenging than solving a single task.

3.2 Average-Performance Multi-Task Reinforcement Learning Formulation

A natural formulation for multi-task RL is to find a single policy that "on average" collects the highest cumulative rewards from all environments. Mathematically, we characterize the MDP at agent i by $\mathcal{M}_i = (\mathcal{S}_i, \mathcal{A}, \mathcal{P}_i, r_i, \gamma_i)$ where $\mathcal{S}_i$ is the set of states, $\mathcal{A}$ is the set of

possible actions which has to be common across tasks, $\mathcal{P}_i$ is the transition probabilities that specify the distribution of the next state given the current state and an action, $r_i : \mathcal{S}_i \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma_i \in (0, 1)$ is the discount factor. We denote by $\mathcal{S} = \cup_i \mathcal{S}_i$, where $\mathcal{S}_i$ can share common states. Each agent i maintains a policy $\pi_i : \mathcal{S} \to \Delta_{\mathcal{A}}$ with $\pi_i(a \mid s)$ denoting the probability of selecting action a in state s.

Given a policy π, let V_i^π be its value function in the i-th environment

$$V_i^\pi(s_i) = \mathbb{E}_{a_k \sim \pi(\cdot|s_k), s_{k+1} \sim \mathcal{P}_i(\cdot|s_k, a_k)} \left[\sum_{k=0}^{\infty} \gamma_i^k r_i(s_i^k, a_i^k) \mid s_i^0 = s_i \right]. \tag{3.1}$$

Similarly, we define the Q function and advantage function in the i_{th} environment

$$\begin{aligned} Q_i^\pi(s_i, a_i) &= \mathbb{E}_{a_k \sim \pi(\cdot|s_k), s_{k+1} \sim \mathcal{P}_i(\cdot|s_k, a_k)} \left[\sum_{k=0}^{\infty} \gamma_i^k r(s_i^k, a_i^k) \mid s_i^0 = s_i, a_i^0 = a_i \right], \\ A_i^\pi(s_i, a_i) &= Q_i^\pi(s_i, a_i) - V_i^\pi(s_i). \end{aligned} \tag{3.2}$$

Without loss of generality, we assume that $r_i(s, a) \in [0, 1]$, implying for any policy π and $\forall s \in \mathcal{S}_i, a \in \mathcal{A}$

$$0 \leqslant V_i^\pi(s) \leqslant \frac{1}{1 - \gamma_i}, \quad -\frac{1}{1 - \gamma_i} \leqslant A_i^\pi(s, a) \leqslant \frac{1}{1 - \gamma_i}. \tag{3.3}$$

Let ρ_i be an initial state distribution over $\mathcal{S}_i$. With some abuse of notation we denote the expected cumulative reward associated with this distribution as $V_i^\pi(\rho_i) = \mathbb{E}_{s_i \sim \rho_i} [V_i^\pi(s_i)]$.

To represent the policy, we consider the scenario where each agent maintains $\theta_i \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ and uses the popular softmax parameterization, i.e.

$$\pi_{\theta_i}(a \mid s) = \frac{\exp(\theta_{i\,;\,s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{i\,;\,s,a'})}. \tag{3.4}$$

The goal of the agents is to cooperatively find a parameter $\theta^\star$ that maximizes the total

cumulative discounted rewards

$$\theta^\star \in \underset{\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}}{\operatorname{argmax}} V^{\pi_\theta}(\boldsymbol{\rho}) \triangleq \sum_{i=1}^{N} V_i^{\pi_\theta}(\rho_i), \quad \boldsymbol{\rho} = [\rho_1; \ldots; \rho_N]. \tag{3.5}$$

Treating each environment as an independent RL problem would produce different policies, each maximizing their respective value function, while our focus is to find a single $\theta^\star$ that balances the performance across all environments.

3.3 Structure in Multi-Task Reinforcement Learning

While single task RL is relatively well understood at least in the tabular setting, multi-task RL is more challenging than it appears from Equation 3.5. We discuss two fundamental challenges of multi-task RL that make this problem much more difficult than its single-task counterpart.

Deterministic vs stochastic policies. Under mild assumptions there always exists a deterministic policy that maximizes the objective in single task RL [106]. The value function of the optimal deterministic policy observes the Bellman optimality equation, which motivates the development of value-based methods for policy optimization. In multi-task RL where the tasks operate under different transition probability kernels, there need not be a deterministic optimal policy in general, and hence there may not be a natural analog of the Bellman optimality equation. We illustrate this with a simple GridWorld example.

In the two-task GridWorld problem shown in Figure 3.1, there are two environments with the same state and action spaces. The dynamics and reward functions, however, are different. The two actions, labeled L and R, deterministically move the agents to the left and right, respectively, in all states in Task 1. In Task 2, the effect of L and R is reversed for states S_2 and S_4: applying L (resp. R) in S_2 transitions to S_3 (resp. S_1), while applying L (resp. R) in S_4 transitions to S_5 (resp. S_3). In both environments, the agents stay in states S_1 and S_5 when they reach them. In Task 1 there is a reward of $+1$ for reaching S_1 and a

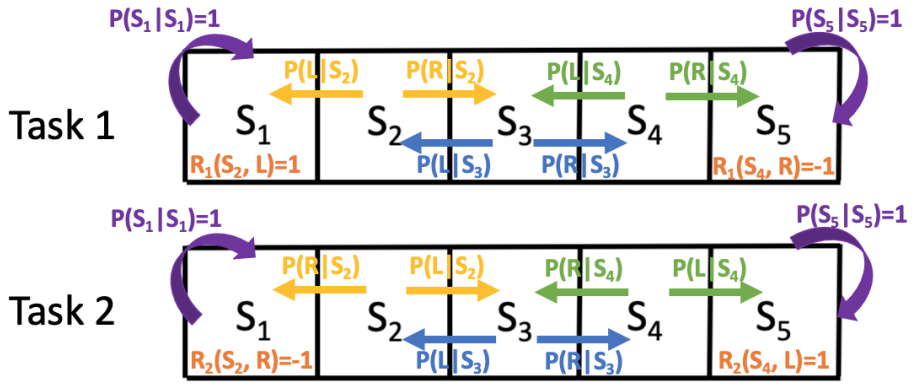


Figure 3.1: Two-Task GridWorld Problem Without a Deterministic Optimal Policy

penalty of -1 for reaching S_5; these rewards are reversed for Task 2.

To find a single policy that maximizes the sum of the cumulative rewards of the two tasks, it is obvious that the optimal policy for state S_2 and S_4 is to always take action L in order to reach the positive reward or to stay away from the negative reward. The only state whose optimal policy remains unclear is S_3. With the detailed computation deferred to Section B.1, we find that the optimal (stochastic) policy $\pi^\star$ is

$$\pi^\star(a|S_3) = \begin{cases} 0.5, & a = L, \\ 0.5, & a = R, \end{cases}$$

which yields $V^{\pi^\star}(S_3) = \frac{2\gamma}{2-\gamma^2}$. By symmetry, the two possible deterministic policies

$$\pi_l(a|S_3) = \begin{cases} 1, & a = L \\ 0, & a = R \end{cases} \text{ and } \pi_r(a|S_3) = \begin{cases} 0, & a = L \\ 1, & a = R \end{cases}$$

produce the same value for state S_3, with $V^{\pi_l}(S_3) = V^{\pi_r}(S_3) = \gamma < V^{\pi^\star}(S_3)$ when $\gamma > 0$. This implies that any deterministic policy is sub-optimal.

As a consequence, RL methods that implicitly rely on the existence of a deterministic optimal policy (e.g., Q learning) cannot solve this type of problems in general. This observation provides motivation for us to study randomized policies and take on a policy

gradient approach.

Gradient domination condition. In single task RL, [12] shows that the objective function, despite being non-concave, satisfies a "gradient domination" condition under the softmax parameterization, which implies that every stationary point is globally optimal. This is important as it guarantees that the policy gradient algorithm can find the globally optimal policy by converging to a stationary point. In the multi-task problem we cannot expect to have this condition in the general setting. The landscape of the multi-task RL objective is so irregular that there could exist multiple stationary points which are not global optima. We illustrate this issue with another simple example.

Let us consider again the 2-task GridWorld problem in Fig.Figure 3.1. Here we make a slight modification to the dynamics of the tasks. In task 1 and task 2, regardless of the action taken in state S_2 and S_4, the transition probability is

$$P_1(s|S_2) = \begin{cases} 1-p, & s = S_1 \\ p, & s = S_3 \end{cases} \qquad P_1(s|S_4) = \begin{cases} 1-p, & s = S_3 \\ p, & s = S_5 \end{cases}$$

$$P_2(s|S_2) = \begin{cases} p, & s = S_1 \\ 1-p, & s = S_3 \end{cases} \qquad P_2(s|S_4) = \begin{cases} p, & s = S_3 \\ 1-p, & s = S_5 \end{cases}$$

for some $0.5 < p \leqslant 1$.

It is obvious that the policy gradients for S_2 and S_4 are always zero as the value function is independent of the policy at these two states. We only have to optimize the policy at S_3.

Under the softmax parameterization, we maintain parameters $\theta_{S_3,L}$ and $\theta_{S_3,R}$ such that

$$\pi_\theta(L|S_3) = \frac{e^{\theta_{S_3,L}}}{e^{\theta_{S_3,L}} + e^{\theta_{S_3,R}}} \text{ and } \pi_\theta(R|S_3) = \frac{e^{\theta_{S_3,R}}}{e^{\theta_{S_3,L}} + e^{\theta_{S_3,R}}}.$$

We consider the case where the agents always start from state S_3. It can be shown that $\theta_{S_3,L} = 1, \theta_{S_3,R} = \infty$ (always taking action R) and $\theta_{S_3,L} = \infty, \theta_{S_3,R} = 1$ (always taking action L) are both stationary points and achieve the global maximum of the objective in

Equation 3.5, while $\theta_{S_3,L} = 1, \theta_{S_3,R} = 1$ (taking action L and R each with probability 0.5) is a sub-optimal stationary point. When gradient based methods are used to optimize Equation 3.5, it could be trapped at the stationary points without finding the global optimality. Later in this chapter, we will dive deeper into the problem and show that the gradient domination condition can be recovered under a restrictive structural assumption.

3.4 Decentralized Policy Gradient Algorithm

In this section, we propose a decentralized variant of the policy gradient algorithm that solves Equation 3.5 in consideration of the aforementioned challenges. Similar to what is observed in the single agent case, the softmax parameterization poises a challenge due to its exponential scaling. To handle the challenge, we use the relative-entropy as a regularization for the objective in Equation 3.5 inspired by [12]. We consider optimizing the modified objective function

$$L^{\lambda}(\theta; \boldsymbol{\rho}) = \sum_{i=1}^{N} L_i^{\lambda}(\theta; \rho_i) = \sum_{i=1}^{N} \left(V_i^{\pi_\theta}(\rho_i) - \lambda \mathrm{RE}\left(\pi_\theta\right)\right),$$

where $\lambda > 0$ is a regularization parameter, and $\mathrm{RE}(\pi_\theta)$ denotes the relative entropy between $U_{\mathcal{A}}$, which is the uniform distribution over $\mathcal{A}$, and π_θ

$$\mathrm{RE}(\pi_\theta) \triangleq \mathbb{E}_{s\sim \mathrm{Unif}_{\mathcal{S}}}\left[D_{\mathrm{KL}}\left(U_{\mathcal{A}}, \pi_\theta(\cdot|s)\right)\right] = -\frac{1}{|\mathcal{S}||\mathcal{A}|}\sum_{a\in\mathcal{A}} \log \pi_\theta(a \mid s) - \log|\mathcal{A}|.$$

We apply gradient ascent to optimize L^{λ} in a decentralized manner, with the updates formally stated in Algorithm 3.1. Each agent can communicate with each other through an undirected and connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where agents i and j can exchange messages if and only if they are connected in $\mathcal{G}$. We denote by $\mathcal{N}_i = \{j \,:\, (i,j) \in \mathcal{E}\}$ the set of agent i's neighbors.

At any time $k \geqslant 0$, agent i first exchanges its iterates with its neighbors $j \in \mathcal{N}_i$ and compute the gradient g_i^k of $L_i^{\lambda}(\theta_i^k; \rho_i)$ only using information from its environment. Agent

Algorithm 3.1: Decentralized Policy Gradient Algorithm (DCPG)

Initialization: Each agent i initializes $\theta_i^0 \in \mathbb{R}^d$, an initial distribution ρ_i, and step sizes $\{\alpha^k\}_{k\in\mathbb{N}}$.

for *k=1,2,3,...* **do**

Each agent i simultaneously implements:

1) Exchange θ_i^k with neighbors $j \in \mathcal{N}_i$

2) Compute the gradient g_i^k of $L_i^\lambda(\theta_i^k; \rho_i)$

3) Policy update:

$$\theta_i^{k+1} = \sum_{j\in\mathcal{N}_i} W_{ij}\theta_j^k + \alpha^k g_i^k. \tag{3.6}$$

end

i updates θ_i by implementing Equation 3.6, where it takes a weighted average of θ_i^k with θ_j^k received from its neighbors $j \in \mathcal{N}_i$, following by a local gradient step. The goal of this weighted average is to achieve a consensus among the agents' parameters, i.e., $\theta_i = \theta_j$, while the local gradient steps are to push this consensus point toward the optimal $\theta^\star$. Here, W_{ij} is some non-negative weight that agent i assigns for θ_j^k. The conditions on W_{ij} to guarantee the convergence of Algorithm Algorithm 3.1 will be specified shortly.

3.5 Convergence Analysis

In this section, our focus is to study the performance of Algorithm 3.1 under the tabular setting, i.e., $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. It is worth recalling that each function V_i^π in Equation 3.5 is in general non-concave. To show the convergence of our algorithm, we first study the case when g_i is exactly ∇L_i^λ, and consider the following assumption on the weight matrix W.

Assumption 3.1. *Let $W = [W_{ij}] \in \mathbb{R}^{N\times N}$ be a doubly stochastic matrix, i.e., $\sum_i W_{ij} = \sum_j W_{ij} = 1$, with $W_{ii} > 0$. Moreover, $W_{ij} > 0$ iff i and j are connected, otherwise $W_{ij} = 0$.*

Assumption 3.1 is fairly standard in the literature of decentralized consensus-based optimization [96, 100]. Given an undirected communication graph, the matrix W satisfying the assumption can be easily generated using the lazy Metropolis method [107]. We denote by σ_2 and σ_N the second largest and the smallest singular values of W. Our first main

result shows that the algorithm converges to a stationary point of Equation 3.5 at the rate $\mathcal{O}(1/\sqrt{K})$ when $\mu_i = \rho_i$.

Theorem 3.1. *We choose the step size of Algorithm 3.1 to be* $\alpha_k = \alpha$ *with* $\alpha \leqslant \frac{1+\sigma_N}{\sum_{i=1}^{N} \frac{16}{(1-\gamma_i)^3} + \frac{4N\lambda}{|\mathcal{S}|}}$. *Then under Assumption 3.1, the iterates* θ_i^k *satisfy* $\forall i = 1, 2, \cdots, N$

$$\min_{k<K} \Big\| \frac{1}{N} \sum_{j=1}^{N} \nabla V_j(\theta_i^k; \rho_j) \Big\|^2 \leqslant \mathcal{O}\Big(\frac{1}{K\alpha} + \frac{\alpha^2}{N(1-\sigma_2)\sum_{j=1}^{N}(1-\gamma_j)^6} + \frac{\lambda^2}{N} \Big). \tag{3.7}$$

First, our upper bound in Equation 3.7 depends quadratically on the inverse of the spectral gap $1 - \sigma_2$, which shows the impact of the graph $\mathcal{G}$ on the convergence of the algorithm. Second, this bound states that under a constant step size the norm of the gradient converges to a ball with radius $\mathcal{O}(\alpha)$ at a rate $\mathcal{O}(1/\sqrt{K})$. As the step size is reduced, we get closer to a stationary point of Equation 3.5. This rate matches the one for single task RL in [12]. However, while we only show the convergence to a stationary point, a global optimality is achieved there. As we have illustrated in Section 3.3, first-order methods can converge to a stationary point which does not have to be globally optimal due to the lack of a gradient domination condition in multi-task RL.

3.6 Achieving Global Optimality

Despite the difficulty of the MTRL problem, we provide a sufficient condition on the structure of the MDPs, under which the gradient domination condition can be recovered and Algorithm 3.1 can find the globally optimal policy.

Assumption 3.2. *Let* π_{θ^*} *be an optimal policy solving Equation 3.5. Then for any* π_θ *and* $\boldsymbol{\mu}$ *we have*

$$\frac{d_{i,\rho_i}^{\pi_{\theta^*}}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)} = \frac{d_{j,\rho_j}^{\pi_{\theta^*}}(s)}{d_{j,\mu_j}^{\pi_\theta}(s)}, \ \forall s : s \in \mathcal{S}_i \cap \mathcal{S}_j, \ \forall i, j \in [N]. \tag{3.8}$$

We know that $d^{\pi_\theta}_{i,\rho_i}(s_i)$ (similarly, $d^{\pi_\theta}_{i,\mu_i}(s_i)$) is the discounted fraction of time that agent i visits state $s_i \in \mathcal{S}_i$ when using ρ_i (similarly, μ_i) as the initial distribution. Qualitatively, this assumption can be interpreted as enforcing that the joint states between the environments are equally explored. Mathematically, this assumption guarantees the objective function Equation 3.5 obeys a kind of gradient domination when each function $V^\pi_i(\rho_i)$ satisfies this condition. We note that Assumption 3.2 holds in the important case where the component tasks share the same state space and transition probability, but differ in their reward functions.

For simplicity, we assume without loss of generality that $\theta^0_i = \theta^0_j, \forall\, i, j$. Let $\alpha^k = \alpha$ satisfying

$$\alpha < \frac{1}{\sum_{i=1}^N \left(\frac{8}{(1-\gamma_i)^3} + \frac{2\lambda}{|\mathcal{S}|}\right)} \min\left\{1 + \sigma_N \;;\; \frac{\lambda N(1-\sigma_2)}{4|\mathcal{S}||\mathcal{A}|\left(2N\lambda + \sum_{i=1}^N \frac{1}{(1-\gamma_i)^2}\right)}\right\}. \tag{3.9}$$

Theorem 3.2. *Suppose that Assumption 3.1 and Assumption 3.2 hold. Given an $\epsilon > 0$, let $\lambda = \epsilon \,/\, 2N\|d^{\pi_{\theta^*}}_{\boldsymbol{\rho}}/\boldsymbol{\mu}\|_\infty$ and α^k satisfy Equation 3.9. Let θ^* be a solution of Equation 3.5. Then $\forall i$, θ^k_i returned by Algorithm 3.1 satisfies*

$$\begin{aligned} &\min_{k<K}\{V(\theta^*;\boldsymbol{\rho}) - V(\theta^k_i;\boldsymbol{\rho})\} \leqslant \epsilon \\ &\textit{if } K \geqslant \mathcal{O}\left(\frac{|S|^2|A|^2\sum_{j=1}^N \frac{1}{(1-\gamma_j)^6}}{(1-\sigma_2)\epsilon^2}\left\|\frac{d^{\pi_{\theta^*}}_{\boldsymbol{\rho}}}{\boldsymbol{\mu}}\right\|^2_\infty\right), \end{aligned} \tag{3.10}$$

where we denote $\left\|\frac{d^{\pi_{\theta^*}}_{\boldsymbol{\rho}}}{\boldsymbol{\mu}}\right\|_\infty = \max\limits_{\substack{s\in\mathcal{S}\\ j:s\in\mathcal{S}_j}} \frac{d^{\pi_{\theta^\star}}_{j,\rho_j}(s)}{(1-\gamma_j)\mu_j(s)}$.

Under Assumption 3.2, Algorithm 3.1 achieves the globally optimal value function with the same rates as the ones in [12], except for a factor $1/(1-\sigma_2)^2$ which captures the impact of communication graph $\mathcal{G}$. Equation 3.10 also shows the impact of the initial distribution $\boldsymbol{\mu}$ on the convergence of the algorithm through the distribution mismatch coefficient. A bad choice of $\boldsymbol{\mu}$ may result in a local optimum (or stationary point) convergence by breaking Assumption 3.2, as we will illustrate by simulation in Subsection 3.7.1.

3.7 Experimental Results

We evaluate the performance of our proposed algorithm on two platforms: GridWorld and drone navigation. We first verify the correctness of our theoretical results by applying the decentralized policy gradient (DCPG) algorithm for solving small-scale GridWorld problems, where each agent uses a tabular policy. We next apply the proposed method to solve the more challenging problem of large-scale drone navigation in simulated 3D environments, where the policy is approximated by neural networks.

General setup. In each simulation, the agents runs a number of episodes of DCPG. In each episode, each agent computes its local gradient by using the Monte-Carlo method. Each agent then communicates with its neighbors over a fixed ring graph (i.e. agent i communicates with agent $i-1$ and $i+1$ for $i = 2, 3, ..., N-1$; agent 1 communicates with agent 2 and N; agent N communicates with agent $N-1$ and 1) and updates its iterates using Equation 3.6. Given the communication graph $\mathcal{G}$, we generate the weight matrix W using the lazy Metropolis method.

3.7.1 GridWorld Problems

We first consider a GridWorld problem in tabular settings, i.e., $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. This is a notable small-scale RL problem, where the agent is placed in a grid of cells. Each cell can be labeled either by the desired goal, an obstacle, or empty. The agent selects an action from the set of 4 actions $\{$up, down, left, right$\}$ to move to the next cell. It then receives a reward of $+1$ if it reaches the desired goal, -1 if it gets into an obstacle, and 0 otherwise. The goal of the agent is to reach a desired position from an arbitrary initial location in a minimum number of steps (or maximize its cumulative rewards).

For multi-task RL settings, we consider a number of different single GridWorld environments of size 10×10, where they are different in the obstacle and goal positions. We assign one agent to each environment, which implements Algorithm 3.1 with the local gradients

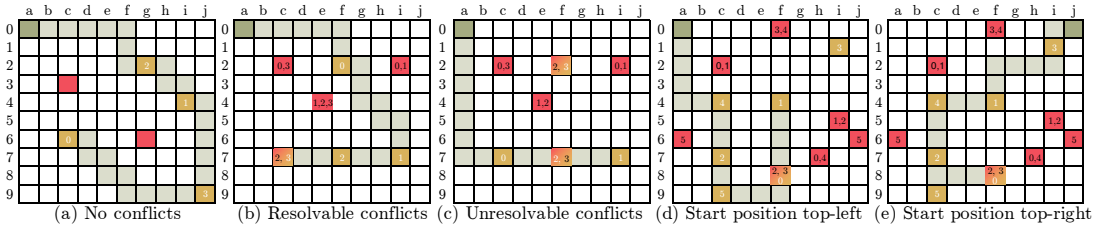


Figure 3.2: Evaluate Learned Policy in Multi-task GridWorld

estimated using a Monte-Carlo approach. The state is the agent's location in the grid. After 1000 training episodes, the agents agree on a unified policy, whose performance is tested in parallel in all environments. The results are presented in Figure 3.2, where we combine all the environments into one grid. In addition, yellow and red cells represent the goal and obstacle, respectively. For each environment, we terminate the test when the agent reaches the goal or hits an obstacle. The light green path is the route which the agent visits in these environments. Since we have a randomized policy, we put the path mostly followed by the agents. Figure 3.2 (a)–(c) consider experiments on four environments, while (d) and (e) are on six environments.

In Figure 3.2(a), we illustrate the performance of the policy when there is no conflict between the environments, i.e., the block of one environment is not the goal of the others and vice versa. In this case, we can see that the algorithm returns an optimal policy which finds all the goals at the environments. Next, we consider the conflict setting in Figure 3.2(b), where one obstacle of environment 2 is the goal of environment 3. Here, the i number in white and black represents the goal and the obstacles of the i-th environment, respectively. Although in this case there is a conflict between the tasks, it is solvable, that is, there is still an optimal path, which the agents eventually find.

We next consider an unsolvable conflict in Figure 3.2(c), where the goal of agent 2 is the obstacle of agent 3 and vice versa. In this case, there does not exists a policy that can always visit all goal positions without running into an obstacle. Instead, the agents need to make a compromise, where they finish three out of the four tasks.

To summarize, the experiments with no conflict and resolvable conflict have dynamics

that allow the optimal value of Equation 3.5 to be the sum of the optimal values of the individual tasks, while the experiment with unresolvable conflict does not. Nevertheless, in all three cases, DCPG successfully finds the optimality of the global objective function in Equation 3.5.

Finally, we illustrate the impact of the initial conditions with the simulations in Figure 3.2(d) and (e). In (d), if the agents start from the top left corner they cannot find the optimal solution. However, when the agents start from the top right corner the algorithms return the gobal optimality as shown in (e). This empirical evidence hints that to achieve the global optimality with the DCPG algorithm, conditions on the initial state distribution like Assumption 3.2 may be necessary.

3.7.2 Drone Navigation

For the drone experiment we use PEDRA, a 3D stimulated drone navigation platform [108]. In this platform, a drone agent is equipped with a front-facing camera, and takes actions to control its flight. The reward received by the drone agent is designed to encourage the drone to stay away from obstacles. We select 4 indoor environments on the PEDRA platform (denoted as Env 0-3), which contain widely different lighting conditions, wall colors, furniture objects, and hallway structures, as shown in Figure 3.3. The performance of a policy is quantified by the mean safe flight (MSF), the average distance travelled by the agent before it collides with any obstacle. This is a standard criterion in evaluating the performance of flying autonomous vehicles [109].

To evaluate the policy learned using Algorithm 3.1 (DCPG), we compare it with the single agent trained independently in each environment. For brevity, we denote by SA-i the single agent trained in environment i. We note that the SAs can be considered as the solutions to the local objective functions, while DCPG optimizes the sum of the local objective functions. Therefore, if trained to the global optimum, each SA provides an upper bound on the performance of the DCPG policy in the respective environment. The aim of

Figure 3.3: Environments used in drone navigation.

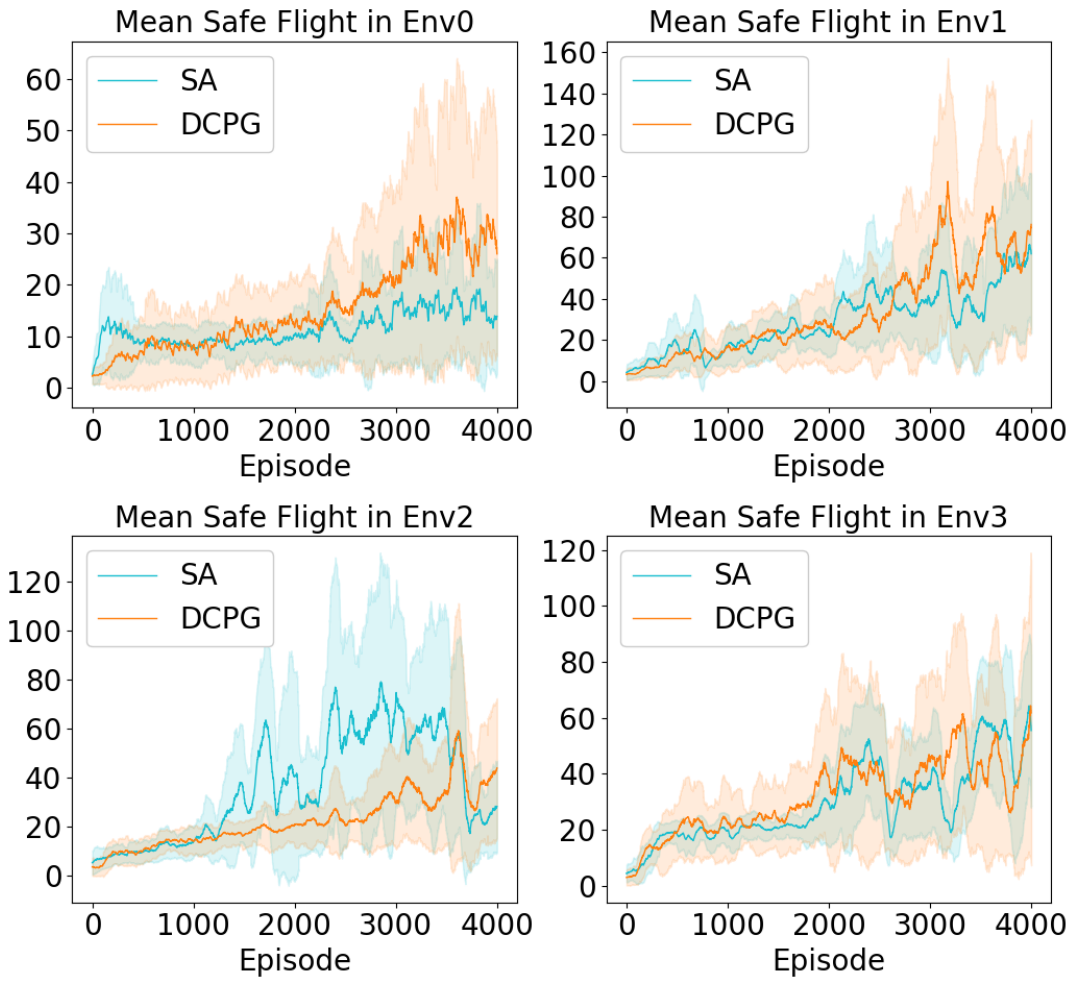


Figure 3.4: MSF During Training (REINFORCE)

the experiments is to show in practical problems where the tasks are highly related, the DCPG policy often performs close to this bound.

To demonstrate the compatibility of our algorithm with a wide range of policy gradient

Table 3.1: MSF of Learned Policy

REINFORCE	Env0	Env1	Env2	Env3	Sum
SA-0	15.9 ± 5.3	4.5 ± 1.2	4.1 ± 1.3	3.6 ± 3.0	28.1
SA-1	3.0 ± 0.2	55.4 ± 29.3	9.7 ± 2.8	8.1 ± 3.8	76.2
SA-2	1.5 ± 0.5	0.8 ± 0.2	21.1 ± 18.3	2.0 ± 0.6	25.4
SA-3	2.3 ± 0.5	0.8 ± 0.2	8.6 ± 2.0	40.1 ± 17.4	51.8
DCPG	**25.2 ± 20.1**	**67.9 ± 35.5**	**40.5 ± 18.0**	**61.8 ± 39.2**	**195.4**
A2C	Env0	Env1	Env2	Env3	Sum
SA-0	21.8 ± 6.5	7.0 ± 0.8	15.1 ± 5.4	14.9 ± 8.2	58.8
SA-1	1.3 ± 0.4	**54.1 ± 20.1**	2.8 ± 0.9	6.4 ± 1.2	59.4
SA-2	1.8 ± 0.7	3.9 ± 0.3	105.2 ± 38.5	9.9 ± 1.3	120.8
SA-3	1.1 ± 0.2	1.4 ± 0.2	15.8 ± 5.0	78.6 ± 25.9	96.9
DCPG	**25.2 ± 7.5**	50.1 ± 24.6	**165.8 ± 64.6**	**159.6 ± 61.0**	**380.7**
PPO	Env0	Env1	Env2	Env3	Sum
SA-0	**28.3 ± 15.5**	11.2 ± 6.3	8.7 ± 5.9	13.5 ± 5.7	61.7
SA-1	1.1 ± 0.6	**75.3 ± 43.2**	1.6 ± 0.4	1.6 ± 0.8	79.6
SA-2	2.5 ± 1.8	3.0 ± 1.1	63.2 ± 36.4	15.6 ± 10.6	84.3
SA-3	1.9 ± 1.6	1.2 ± 0.5	14.3 ± 8.7	139.0 ± 72.5	156.4
DCPG	26.3 ± 10.9	66.7 ± 30.8	**144.0 ± 82.4**	**195.2 ± 92.4**	**432.2**

methods, we conduct three sets of experiments, where we run Algorithm 3.1 with the gradient g_i^k estimated by three popular variants of policy gradient algorithms: REINFORCE, advantage actor-critic (A2C), and proximal policy optimization (PPO). In each case, a 5-layer neural network is used to approximate the policy. We stress that in each set of the experiments, the SAs and DCPG are trained identically, with the only difference being whether the agents communicate their policies.

In Figure 3.4, we show MSF of the DCPG and SA policies in the training phase with the REINFORCE algorithm. In the testing phase, we deploy the policies learned by DCPG and SAs in the four environments and present the results in Table Table 3.1. Across the three sets of experiments, we consistently see the performance difference between DCPG and the SAs. As expected, SA-i only performs well in i-th environment but does not generalize to environment it has not seen. On the other hand, the policy returned by DCPG performs very

well in all environments. Surprisingly, DCPG often performs even better than each SA-i in the i-th environment, which we speculate is due to the benefits of learning common features and representation among the agents.

3.8 Constrained Multi-Task Reinforcement Learning

We observe from Table 3.1 that the DCPG policy, while performing better than the SAs, does not achieve balanced cumulative rewards across the environments. Motivated to control the policy in a more fine-grained manner, we consider another formulation of multi-task RL that allows us to specify the performance upper and lower bounds of the policy in each environment. In the rest of the section, we will focus on the case where all environments have identical state spaces and transition kernels and only differ in the reward functions. This setting satisfies Assumption 3.2 and recovers the gradient domination condition. Under notations introduced in Section 3.2, we denote

$$V_0^\pi(\rho) = \frac{1}{N}\sum_{i=1}^{N} V_i^\pi(\rho)$$

for simplicity. Given local performance upper and lower bounds $\{\ell_i \in \mathbb{R}, u_i \in \mathbb{R}\}_{i=1}^N$, our constrained multi-task RL objective is to solve the following optimization problem

$$\begin{aligned} \pi^\star = \underset{\pi}{\text{argmax}} \quad & V_0^\pi(\rho) \\ \text{subject to} \quad & \ell_i \leqslant V_i^\pi(\rho) \leqslant u_i \quad \forall i = 1, 2, ..., N. \end{aligned} \tag{3.11}$$

It is worth noting that Equation 3.11 obviously recovers the constraint-less formulation in Equation 3.5 by properly choosing ℓ_i, u_i. Again, we considers the tabular setting under the softmax parameterization (Equation 3.4).

Although neither the objective nor the constraint set of Equation 3.11 is convex, it is known from [110][Theorem 3.6] that strong duality holds under the following Slater's

condition.

Assumption 3.3 (Slater's Condition)**.** *There exists a constant* $0 < \xi \leqslant 1$ *and a policy* π *such that* $\ell_i + \xi \leqslant V_i^\pi(\rho) \leqslant u_i - \xi$ *for all* $i = 1, \cdots, N$.

This is a mild and standard assumption in the study of constrained MDPs [38, 111, 112], and states that the constraint set must have at least one interior point.

3.8.1 Algorithm Design

In this section, we develop an algorithm for solving Equation 3.11 and formally present the updates in Algorithm 3.2. As a first step, we form the Lagrangian of Equation 3.11

$$V_L^{\pi,\lambda,\nu}(\rho) = V_0^\pi(\rho) + \sum_{i=1}^{N} \left(\lambda_i \left(V_i^\pi(\rho) - \ell_i\right) - \nu_i \left(V_i^\pi(\rho) - u_i\right)\right), \tag{3.12}$$

where $\lambda = [\lambda_1, \ldots, \lambda_N] \in \mathbb{R}_+^N$ and $\nu = [\nu_1, \ldots, \nu_N] \in \mathbb{R}_+^N$ are the dual variables associated with the lower and upper bound constraints.

The dual function $V_D^{\lambda,\nu}$ is defined as

$$V_D^{\lambda,\nu}(\rho) = \max_\pi V_L^{\pi,\lambda,\nu}(\rho), \tag{3.13}$$

and the dual problem is

$$\lambda^\star, \nu^\star = \operatorname*{argmin}_{\lambda,\nu \in \mathbb{R}_+^N} V_D^{\lambda,\nu}(\rho). \tag{3.14}$$

A consequence of Slater's condition is the boundedness of $\lambda^\star, \nu^\star$.

Lemma 3.1. *Under Assumption 3.3, we have*

$$\|\lambda^\star\|_\infty \leqslant \frac{B_\lambda}{2} \quad \textit{and} \quad \|\nu^\star\|_\infty \leqslant \frac{B_\lambda}{2},$$

where $B_\lambda = \frac{1}{\xi(1-\gamma)}$

The strong duality states

$$V_D^{\lambda^\star,\nu^\star}(\rho) = V_0^{\pi^\star}(\rho), \quad \text{and} \quad \pi^\star, (\lambda^\star, \nu^\star) = \operatorname*{argmax}_{\pi} \operatorname*{argmin}_{\lambda,\nu} V_L^{\pi,\lambda,\nu}, \tag{3.15}$$

where $\pi^\star$, $\lambda^\star$, and $\nu^\star$ are the (not necessarily unique) optimal solutions to Equation 3.11 and Equation 3.14. Motivated by the existence of the strong duality, we take a primal-dual approach to find the saddle point of the minimax objective in Equation 3.15. Specifically, we use $\lambda^k = [\lambda_1^k, \ldots, \lambda_N^k] \in \mathbb{R}^N$ and $\nu^k = [\nu_1^k, \ldots, \nu_N^k] \in \mathbb{R}^N$ to estimate $\lambda^\star$ and $\nu^\star$ and maintain local variables $\theta_i^k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that $\pi_{\theta_i^k}$ tracks $\pi^\star$ at each agent i. We update the variables with gradient descent ascent.

Primal Variable. Carrying out gradient descent ascent requires computing the (natural) gradients of the Lagrangian with respect to the primal and dual variables, which both have closed form expressions. On one side, it is known that the natural gradient of the value function under reward r_i with respect to θ, denoted by $\widetilde{\nabla}_\theta V_i^\pi(\rho)$, is the advantage function scaled by $1/(1-\gamma)$ [113]. This means that for any distribution ρ, we have

$$\begin{aligned}\widetilde{\nabla}_{\theta_{s,a}} V_i^{\pi_\theta,\lambda,\nu}(\rho) &= \frac{1}{1-\gamma}\big(A_0^{\pi_\theta}(s,a) + \sum_{i=1}^N (\lambda_i - \nu_i) A_i^{\pi_\theta}(s,a)\big) \\ &= \sum_{i=1}^N (\frac{1}{N} + \lambda_i - \nu_i) A_i^{\pi_\theta}(s,a). \end{aligned} \tag{3.16}$$

In our decentralized primal variable update in Equation 3.17, each agent essentially moves in the direction of a locally available component of this natural policy gradient, followed by an averaging step that mixes the agents' policy parameters to achieve consensus.

Dual Variable. On the other hand, the gradient of the Lagrangian with respect to the dual variable is

$$\nabla_{\lambda_i} V_L^{\pi,\lambda,\nu}(\rho) = V_i^\pi(\rho) - \ell_i = \sum_{s:\rho(s)>0,a} \rho(s)\pi(a \mid s) Q_i^\pi(s,a) - \ell_i,$$

$$\nabla_{\nu_i} V_L^{\pi,\lambda,\nu}(\rho) = -V_i^{\pi}(\rho) + u_i = -\sum_{s:\rho(s)>0,a} \rho(s)\pi(a \mid s) Q_i^{\pi}(s,a) + u_i.$$

This naturally leads to the update in Equation 3.18, in which the operator $\Pi_{[0,B_\lambda]} : \mathbb{R}^N \to \mathbb{R}^N$ denotes the element-wise projection of a vector to the interval $[0, B_\lambda]$. We use the projection to guarantee the stability of the dual variables and note that the optimal dual variables are in the span of $\Pi_{[0,B_\lambda]}$ according to Lemma 3.1.

Algorithm 3.2: Decentralized Primal-Dual Natural Policy Gradient Algorithm in Tabular Setting

Initialization: Each agent i initializes $\theta_i^0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} = 0$ and dual variables $\lambda_i^0, \nu_i^0 \in \mathbb{R}_+ = 0$

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

  for *Each agent* $i = 1, \cdots, N$ **do**

    1) Exchange θ_i^k with neighbors $j \in \mathcal{N}_i$

    2) Policy update:

$$\begin{aligned} \theta_i^{k+1} &= \sum_{j \in \mathcal{N}_i} W_{ij}\theta_j^k + \alpha\left(\frac{1}{N} + \lambda_i^k - \nu_i^k\right) Q_i^{\pi_{\theta_i^k}} \\ \pi_i^{k+1}(a \mid s) &= \frac{\exp\left(\theta_i^{k+1}(s,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta_i^{k+1}(s,a')\right)} \end{aligned} \tag{3.17}$$

    3) Local dual variable update:

$$\begin{aligned} \lambda_i^{k+1} &= \Pi_{[0,B_\lambda]}\left(\lambda_i^k - \eta\left(V_i^{\pi_{\theta_i^k}}(\rho) - \ell_i\right)\right) \\ \nu_i^{k+1} &= \Pi_{[0,B_\lambda]}\left(\nu_i^k + \eta\left(V_i^{\pi_{\theta_i^k}}(\rho) - u_i\right)\right) \end{aligned} \tag{3.18}$$

  end

end

3.8.2 Finite-Time Convergence

With the detailed proof deferred to the appendix, we now present the finite-time complexity of Algorithm 3.2 in the following theorem, which essentially states that the policy at every local agent converges to the globally optimal policy both in objective function value and constraint violation with rate $\mathcal{O}(K^{-1/2})$.

Theorem 3.3. *Consider the iterates $\{\pi_i^k\}$ obtained from K iterations of Algorithm Algorithm 3.2. Let the step size sequences be*

$$\alpha = \frac{\alpha_0}{K^{1/2}}, \quad \eta = \frac{\eta_0}{K^{1/2}}, \tag{3.19}$$

with $\alpha_0 = \mathcal{O}(\sqrt{1-\sigma_2(W)})$. Then, under Assumption 3.3, we have for any $j = 1, \cdots, N$

$$\max\Big\{\frac{1}{K}\sum_{k=0}^{K-1}(V_0^{\pi^\star}(\rho) - V_0^{\pi_j^k}(\rho)), \frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\Big(\big[\ell_i - V_i^{\pi_j^k}(\rho)\big]_+ + \big[V_i^{\pi_j^k}(\rho) - u_i\big]_+\Big)\Big\} \leqslant \mathcal{O}\Big(\frac{N^{5/4}}{\sqrt{1-\sigma_2}K^{1/2}}\Big).$$

We omit the dependency of the bound on structural constants including $|\mathcal{S}|, |\mathcal{A}|, 1-\gamma$ and note that it is the same as in the centralized single-task constrained MDP setting with a single constraint [112]. The convergence rate scales up with N, which shows the difficulty of the problem as the number of tasks increases. The dependency on $(1-\sigma_2)^{-1/2}$ captures the effect of the network connectivity, which becomes smaller as the communication graph gets denser.

3.9 Conclusion & Future Directions

To conclude and summarize our main contribution in this chapter, we studied two multi-task agent RL formulations and proposed provably convergent algorithms for solving the formulations, under the assumption that the local environments have the same state space and transition probability kernel. In the drone navigation experiments presented in Subsection 3.7.2, we observed that the learned multi-task policy performs even better than each single agent trained in its own environment under a moderate number of training episodes. Conceptually, we attribute this phenomenon to the existence of a common representation which facilitates learning, which is not completely surprising. In the extreme case where all local tasks are exactly identical, learning a joint policy effectively reduces

the noise in the gradient estimates, which could mathematically justify our observation. However, in the drone navigation experiments, the local tasks are related but not identical, and explaining the observation becomes a much more challenging, but still interesting, possible future work. Another future direction from the experimental perspective is to investigate whether the constrained multi-task RL formulation can actually lead to more desirable performance of the learned policy in practical problems.

CHAPTER 4
A DIRECT POLICY OPTIMIZATION APPROACH TO TWO-PLAYER ZERO-SUM MARKOV GAMES

In this chapter, we study the structure in the two-player zero-sum Markov game and leverage it to design a gradient descent ascent (GDA) algorithm that provably and efficiently finds the Nash equilibrium. Despite the fact that Markov games observe a "gradient domination" condition with respect to each player, strong structure such as the convexity does not exist that can be exploited to guarantee the fast convergence of GDA.

Our approach to this challenge is to introduce a structured entropy regularization. The regularized Markov game enjoys a series of favorable properties including the existence and uniqueness of the Nash equilibrium (whose distance from the Nash equilibrium of the unregularized problem can be upper bounded by the regularization weight) and a Polyak-Łojasiewicz (PŁ) flavored condition. Exploiting these properties, we show that the the GDA algorithm can find the unique Nash equilibrium of the regularized Markov game linearly fast. We propose schemes of adjusting the regularization weight properly over time that allows the last iterates of the GDA algorithm to converge to the Nash equilibrium of the original Markov game[1].

4.1 Introduction

The two-player zero-sum Markov game is a special case of competitive multi-agent reinforcement learning where two agents driven by opposite reward functions jointly determine the state transition in an environment. Usually cast as a non-convex non-concave minimax optimization program, this framework finds applications in many practical problems including game playing [114, 115], robotics [116, 117], and robust policy optimization [32].

[1]The presentation in this chapter is partly adapted from [39].

A convenient class of algorithms frequently used to solve multi-agent reinforcement learning problems is the independent learning approach. Independent learning algorithms proceed iteratively with each player taking turns to optimize its own objective while pretending that the policies of the other players are fixed to their current iterates. In the context of two-player zero-sum Markov games, the independent learning algorithm performs GDA, which alternates between the gradient updates of the two agents that seek to maximize and minimize the same value function. Despite the popularity of such algorithms in practice, their theoretical understandings are sparse and do not follow from those in the single-agent case as the environment is not stationary from the eye of any agent. [118] shows that iterates of GDA can possibly diverge or be trapped in limit cycles even in the simplest single-state case when the two players learn with the same rate.

It may be tempting to analyze the two-player zero-sum Markov game by applying the existing theoretical results on minimax optimization. However, as the objective function in a Markov game is not convex or concave, current analytical tools in minimax optimization that require the objective function to be convex/concave at least on one side are inapplicable. Fortunately, the Markov game has its own structure: it exhibits a "gradient domination" condition with respect to each player, which essentially guarantees that every stationary point of the value function is globally optimal. Exploiting this property, [29] builds on the theory of [22] and shows that a two-time-scale GDA algorithm converges to the Nash equilibrium of the Markov game with a complexity that depends polynomially on the specified precision. However, deriving an explicit finite-time convergence rate is still an open problem. In addition, the analysis in [29] does not guarantee the convergence of the last iterate; convergence is shown on the average of all past iterates.

In this chapter, we show that introducing an entropy regularizer into the value function significantly accelerates the convergence of GDA to the Nash equilibrium. By dynamicially adjusting the regularization weight towards zero, we are able to give a finite-time last-iterate convergence guarantee to the Nash equilibrium of the original Markov game. The main

contribution of the work in this chapter is twofold.

First, we show that the entropy-regularized Markov game is highly structured; in particular, it obeys a condition similar to the well-known PŁ condition, which allows linear convergence of GDA to the (unique) equilibrium point of the regularized game with fixed regularization weight. We also show that the distance of the equilibrium point of the regularized game to the equilibrium point of the original game can be bounded in terms of the regularizing weight.

Furthermore, we show that by dynamically driving the regularization weight towards zero, we can solve the original Markov game. We propose two approaches to reduce the regularization weight and study their finite-time convergence. The first approach uses a piecewise constant weight that decays geometrically fast, and its analysis follows as a straightforward consequence of our analysis for the case of fixed regularization weight. To reach a Nash equilibrium of the Markov game up to error ϵ, we find that this approach requires at most $\mathcal{O}(\epsilon^{-3})$ gradient updates, where $\mathcal{O}$ only hides structural constants. The second approach reduces the regularization weight online along with the gradient updates. Through a multi-time-scale analysis, we optimize the regularization weight sequence along with the step size as polynomial functions of k, where k is the iteration index. We show that the last iterate of the GDA algorithm converges to the Nash equilibrium of the original Markov game at a rate of $\mathcal{O}(k^{-1/3})$. Compared with the state-of-the-art analysis of the GDA algorithm without regularization which shows that the convergence rate of the averaged iterates is polynomial in the desired precision and all related parameters, our algorithms enjoy faster last-iterate convergence guarantees.

4.2 Related Works

A Markov game reduces to a standard MDP with respect to one player if the policy of the other player is fixed. This is an important observation that allows our work to exploit the recent advances in the analysis of policy gradient methods for MDPs [13, 113, 119–

121]. Various entropy-based regularizers are introduced in these works that inspire the regularization of this paper. Our particular regularization is also considered by [122], but we discuss and leverage structure in the regularized Markov game that was previously unknown.

As the two-player zero-sum Markov game can be formulated a minimax optimization problem, our work relates to the vast volume of literature in this domain. Minimax optimization has been extensively studied in the case where the objective function is convex/concave with respect to at least one variable [22–25]. In the general non-convex non-concave setting, the problem becomes much more challenging as even the notion of stationarity is unclear [26]. In [27], non-convex non-concave objective functions obeying a one–sided PŁ condition are considered, which the authors use to show the convergence of GDA. [28] analyzes GDA under a two-sided PŁ condition and has a tight connection to our work as the value function of our regularized Markov game also has structure that is similar to, but weaker than, the PŁ condition on two sides.

By exploiting the gradient domination condition of a Markov game with respect to each player, [29] is the first to show that the GDA algorithm provably converges to a Nash equilibrium of a Markov game. A finite-time complexity is not derived in [29], but their analysis and choice of step sizes indicate that the convergence rate is at least worse than $\mathcal{O}(k^{-1/10.5})$. Additionally, [29] does not guarantee the convergence of the last iterate, but rather analyzes the average of all iterates. In contrast, our work provides a finite-time convergence analysis on the last iterate of the GDA algorithm.

While our work treats the Markov game purely from the optimization perspective, we would like to point out another related line of works that consider value-based methods [122–126]. In particular, [123] is among the first works to extend value-based methods from single-agent MDP to two-player Markov games. Since then, the basic techniques for analyzing value-based methods for Markov games are relatively well-known. [124] considers a value iteration algorithm with confidence bounds. In [122], a nested-loop algorithm is designed where the outer loop employs value iteration and the inner loop runs a gradient-

descent-ascent-flavored algorithm to solve a regularized bimatrix game. In comparison, pure policy optimization algorithms are much less understood for Markov games, but this is an important subject to study due to their wide use in practice. In single-agent MDPs, value-based methods and policy optimization methods enjoy comparable convergence guarantees today, and our work aims to narrow the gap between the understanding of these two classes of algorithms in two-player Markov games.

Finally, we note the recent surge of interest in solving two-player games and minimax optimization programs with extragradient or optimistic gradient methods in the cases where vanilla gradient algorithms often cannot be shown to converge [122, 127–132]. These methods typically require multiple gradient evaluations at each iteration and are more complicated to implement. Most related to our work, [122] shows the linear convergence of an extragradient algorithm for solving regularized bilinear matrix games. They also show that a regularized Markov game can be decomposed into a series of regularized matrix games and present a nested-loop extragradient algorithm which solves these games successively and eventually converges to the Nash equilibrium of the regularized Markov game. The regularization weight can then be selected based on the desired precision of the unregularized problem. Although our overall goal of finding the Nash equilibrium of a general Markov game is the same, the manner in which we decompose and analyze the problem is different. Our analysis here is based on GDA applied directly to a general regularized Markov game. We show that for a fixed regularization parameter for a general Markov game, GDA has linear convergence to the modified equilibrium point. We also give a scheduling scheme for adjusting the regularization parameter as the GDA iterations proceed, making them converge to the solution to the original problem.

4.3 Preliminaries

We consider a two-player Markov game characterized by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{B}, \mathcal{P}, \gamma, r)$. Here, $\mathcal{S}$ is the finite state space, $\mathcal{A}$ and $\mathcal{B}$ are the finite action spaces of the two players, $\gamma \in (0, 1)$ is

the discount factor, and $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$ is the reward function. Let $\Delta_{\mathcal{F}}$ denote the probability simplex over a set $\mathcal{F}$, and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta_{\mathcal{S}}$ be the transition probability kernel, with $\mathcal{P}(s' \mid s, a, b)$ specifying the probability of the game transitioning from state s to s' when the first player selects action $a \in \mathcal{A}$ and the second player selects $b \in \mathcal{B}$. The policies of the two players are denoted by $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}$, with $\pi(a \mid s)$, $\phi(b \mid s)$ denoting the probability of selecting action a, b in state s according to π, ϕ. Given a policy pair (π, ϕ), we measure its performance in state $s \in \mathcal{S}$ by the value function

$$V^{\pi,\phi}(s) = \mathbb{E}_{a_k \sim \pi(\cdot|s_k), b_k \sim \phi(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k, b_k)} \Big[\sum\nolimits_{k=0}^{\infty} \gamma^k r\left(s_k, a_k, b_k\right) \mid s_0 = s \Big].$$

Under a fixed initial distribution $\rho \in \Delta_{\mathcal{S}}$, we define the discounted cumulative reward under (π, ϕ)

$$J(\pi, \phi) \triangleq \mathbb{E}_{s_0 \sim \rho}[V^{\pi,\phi}(s_0)],$$

where the dependence on ρ is dropped for simplicity. It is known that the Nash equilibrium always exists in two-player zero-sum Markov games [133], i.e. there exists an optimal policy pair $(\pi^\star, \phi^\star)$ such that

$$\max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \min_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} J(\pi, \phi) = \min_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} J(\pi, \phi) = J(\pi^\star, \phi^\star). \tag{4.1}$$

However, as J is generally non-concave with respect to the policy of the first player and non-convex with respect to that of the second player, direct GDA updates may not find $(\pi^\star, \phi^\star)$ and usually exhibit an oscillation behavior, which we illustrate through numerical simulations in Section 4.6. Our approach to address this issue is to enhance the structure of the Markov game through regularization.

4.3.1 Entropy-Regularized Two-Player Zero-Sum Markov Games

In this section we define the entropy regularization and discuss structure of the regularized objective function and its connection to the original problem. Let the regularizers be

$$\mathcal{H}_\pi(s,\pi,\phi) \triangleq \mathbb{E}_{a_k\sim\pi(\cdot|s_k),b_k\sim\phi(\cdot|s_k),s_{k+1}\sim\mathcal{P}(\cdot|s_k,a_k,b_k)}\Big[\sum\nolimits_{k=0}^{\infty} -\gamma^k \log\pi\left(a_k \mid s_k\right) \mid s_0 = s\Big],$$
$$\mathcal{H}_\phi(s,\pi,\phi) \triangleq \mathbb{E}_{a_k\sim\pi(\cdot|s_k),b_k\sim\phi(\cdot|s_k),s_{k+1}\sim\mathcal{P}(\cdot|s_k,a_k,b_k)}\Big[\sum\nolimits_{k=0}^{\infty} -\gamma^k \log\phi\left(b_k \mid s_k\right) \mid s_0 = s\Big].$$

We define the regularized value function

$$\begin{aligned} V_\tau^{\pi,\phi}(s) &\triangleq V^{\pi,\phi}(s) + \tau\mathcal{H}_\pi(s,\pi,\phi) - \tau\mathcal{H}_\phi(s,\pi,\phi) \\ &= \mathbb{E}_{\pi,\phi,\mathcal{P}}\Big[\sum\nolimits_{k=0}^{\infty}\gamma^k\Big(r\left(s_k,a_k,b_k\right) - \tau\log\pi(a_k \mid s_k) + \tau\log\phi(b_k \mid s_k)\Big) \mid s_0 = s\Big], \end{aligned}$$

where $\tau \geqslant 0$ is a weight parameter. Again under a fixed initial distribution $\rho \in \Delta_{\mathcal{S}}$ we denote $J_\tau(\pi,\phi) \triangleq \mathbb{E}_{s\sim\rho}[V_\tau^{\pi,\phi}(s)]$. The regularized advantage function is

$$A_\tau^{\pi,\phi}(s,a,b) \triangleq r(s,a,b) - \tau\log\pi(a \mid s) + \tau\log\phi(b \mid s) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a,b)}\left[V_\tau^{\pi,\phi}(s')\right] - V_\tau^{\pi,\phi}(s),$$

which later helps us to express the policy gradient.

We use $d_\rho^{\pi,\phi} \in \Delta_{\mathcal{S}}$ to denote the discounted visitation distribution under any policy pair (π,ϕ) and the initial state distribution ρ

$$d_\rho^{\pi,\phi}(s) \triangleq (1-\gamma)\mathbb{E}_{\pi,\phi,\mathcal{P}}\Big[\sum\nolimits_{k=0}^{\infty}\gamma^k\mathbf{1}(s_k = s) \mid s_0 \sim \rho\Big]$$

For sufficient state visitation, we assume that the initial state distribution is bounded away from zero. This is a standard assumption in the entropy-regularized MDP literature [13, 134].

Assumption 4.1. *The initial state distribution ρ is strictly positive for any state, and we*

denote $\rho_{\min} = \min_{s \in \mathcal{S}} \rho(s) > 0$.

When the policy of the first player is fixed to $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, the Markov game reduces to an MDP for the second player with state transition probability $\widetilde{\mathcal{P}}_{\phi}(s' \mid s, b) = \sum_{a \in \mathcal{A}} \mathcal{P}(s' \mid s, a, b)\pi(a \mid s)$ and reward function $\widetilde{r}_{\phi}(s, b) = \sum_{a \in \mathcal{A}} r(s, a, b)\pi(a \mid s)$. A similar argument holds for the first player if the second player's policy is fixed. To denote the operators that map one player's policy to the best response of the other player and the corresponding value function, we define

$$\pi_{\tau}(\phi) \triangleq \operatorname*{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} J_{\tau}(\pi, \phi), \quad \phi_{\tau}(\pi) \triangleq \operatorname*{argmin}_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} J_{\tau}(\pi, \phi),$$
$$g_{\tau}(\pi) \triangleq \min_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} J_{\tau}(\pi, \phi) = J_{\tau}(\pi, \phi_{\tau}(\pi)). \tag{4.2}$$

For any $\tau > 0$, the following lemma bounds the performance difference between optimal and sub-optimal policies and establishes the uniqueness of $\pi_{\tau}(\phi)$ and $\phi_{\tau}(\pi)$. When $\tau = 0$, we use $\pi_0(\phi)$ and $\phi_0(\pi)$ to denote one of the maximizers and minimizers since they may not be unique.

Lemma 4.1 (Performance Difference)**.** *Under Assumption 4.1 and given $\tau > 0$, $\pi_{\tau}(\phi)$ is unique for any $\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}$, and $\phi_{\tau}(\pi)$ is unique for any $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$. Given any min player policy $\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}$,*

$$J_{\tau}(\pi_{\tau}(\phi), \phi) - J_{\tau}(\pi, \phi) \geqslant \frac{\tau \rho_{\min}}{2 \log(2)} \|\pi_{\tau}(\phi) - \pi\|^2, \quad \forall \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}. \tag{4.3}$$

Given any max player policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$,

$$J_{\tau}(\pi, \phi_{\tau}(\pi)) - J_{\tau}(\pi, \phi) \leqslant -\frac{\tau \rho_{\min}}{2 \log(2)} \|\phi_{\tau}(\pi) - \phi\|^2, \quad \forall \phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}. \tag{4.4}$$

The Nash equilibrium of the regularized problem is sometimes referred to as the quantal response equilibrium [135] and is known to exist under any τ. Leveraging Lemma 4.1, we

formally state the conditions guaranteeing its existence and affirm that it is unique.

Lemma 4.2 (Minimax Theorem for Entropy-Regularized Markov Game)**.** *Under Assumption 4.1, for any regularization weight $\tau > 0$, there exists a unique Nash equilibrium policy pair $(\pi_\tau^\star, \phi_\tau^\star)$ such that*

$$\max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \min_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} J_\tau(\pi, \phi) = \min_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} J_\tau(\pi, \phi) = J_\tau(\pi_\tau^\star, \phi_\tau^\star). \tag{4.5}$$

We are only interested in the solution of the regularized Markov game if it gives us knowledge of the original problem in Equation 4.1. In the following lemma, we show that the distance between the Nash equilibrium of the regularized game and that of the original one is bounded by the regularization weight. This is an important condition guaranteeing that we can find an approximate solution to the original Markov game by solving the regularized problem. In addition, this lemma also shows that the same policy pair produces value functions with bounded distance under two regularization weights.

Lemma 4.3. *For any $\tau \geqslant \tau' \geqslant 0$ and policy π,*

$$-(\tau - \tau') \log |\mathcal{B}| \leqslant J_\tau(\pi_\tau^\star, \phi_\tau^\star) - J_{\tau'}(\pi_{\tau'}^\star, \phi_{\tau'}^\star) \leqslant (\tau - \tau') \log |\mathcal{A}|. \tag{4.6}$$

$$-(\tau - \tau') \log |\mathcal{B}| \leqslant g_\tau(\pi) - g_{\tau'}(\pi) = J_\tau(\pi, \phi_\tau(\pi)) - J_{\tau'}(\pi, \phi_{\tau'}(\pi)) \leqslant (\tau - \tau') \log |\mathcal{A}|. \tag{4.7}$$

$$-\frac{\tau - \tau'}{1 - \gamma} \log |\mathcal{B}| \leqslant J_\tau(\pi, \phi) - J_{\tau'}(\pi, \phi) \leqslant \frac{\tau - \tau'}{1 - \gamma} \log |\mathcal{A}|. \tag{4.8}$$

4.3.2 Softmax Parameterization

In this work we use a tabular softmax policy parameterization and maintain two tables $\theta \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$, $\psi \in \mathbb{R}^{\mathcal{S} \times \mathcal{B}}$ that parameterize the policies of the two players according to

$$\pi_\theta(a \mid s) = \frac{\exp\left(\theta(s, a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta(s, a')\right)}, \quad \text{and} \quad \phi_\psi(b \mid s) = \frac{\exp\left(\psi(s, b)\right)}{\sum_{b' \in \mathcal{A}} \exp\left(\psi(s, b')\right)}.$$

The gradients of the regularized value function with respect to the policy parameters admit closed-form expressions

$$\begin{aligned}\frac{\partial J_\tau(\pi_\theta,\phi_\psi)}{\partial\theta(s,a)} &= \frac{1}{1-\gamma}d_\rho^{\pi_\theta,\phi_\psi}(s)\pi_\theta(a\mid s)\sum\nolimits_{b\in\mathcal{B}}\phi_\psi(b\mid s)A_\tau^{\pi_\theta,\phi_\psi}(s,a,b),\\ \frac{\partial J_\tau(\pi_\theta,\phi_\psi)}{\partial\psi(s,b)} &= \frac{1}{1-\gamma}d_\rho^{\pi_\theta,\phi_\psi}(s)\phi_\psi(b\mid s)\sum\nolimits_{a\in\mathcal{A}}\pi_\theta(a\mid s)A_\tau^{\pi_\theta,\phi_\psi}(s,a,b),\end{aligned} \tag{4.9}$$

and computing them exactly requires knowledge of the dynamics of the environment. Note that the gradients of value function and the regularizer are Lipschitz with respect to the policy parameters with constants $L_V = \frac{8}{(1-\gamma)^3}$ and $L_{\mathcal{H}} = \frac{4+8\log|\mathcal{A}|}{(1-\gamma)^3}$. This property is more formally stated and proved in Lemmas C.1 and C.2 of the appendix.

We next present an important property that we will later exploit to study the convergence of the GDA updates to the solution of the regularized Markov game. Under the softmax parameterization, the regularized value function enjoys a gradient domination condition with respect to the policy parameter that resembles the PŁ condition.

Lemma 4.4 (PL-Type Condition)**.** *Under Assumption 4.1, we have for any* $\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ *and* $\psi \in \mathbb{R}^{\mathcal{S}\times\mathcal{B}}$

$$\begin{aligned}\|\nabla_\theta J_\tau(\pi_\theta,\phi_\psi)\|^2 &\geqslant \frac{2(1-\gamma)\tau\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,a}\pi_\theta(a\mid s)\right)^2\left(J_\tau(\pi_\tau(\phi_\psi),\phi_\psi)-J_\tau(\pi_\theta,\phi_\psi)\right),\\ \|\nabla_\psi J_\tau(\pi_\theta,\phi_\psi)\|^2 &\geqslant \frac{2(1-\gamma)\tau\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,b}\phi_\psi(b\mid s)\right)^2\left(J_\tau(\pi_\theta,\phi_\psi)-J_\tau(\pi_\theta,\phi_\tau(\pi_\theta))\right).\end{aligned}$$

The PŁ condition is a tool commonly used in the optimization community to show the linear convergence of the gradient descent algorithm [34, 82]. The condition in Lemma 4.4 is weaker than the common PŁ condition in two aspects. First, our PŁ coefficient is a function of the smallest policy entry. When we seek to bound the gradient of the iterates $\|\nabla_\theta J_\tau(\pi_{\theta_k},\phi_{\psi_k})\|^2$ and $\|\nabla_\psi J_\tau(\pi_{\theta_k},\phi_{\psi_k})\|^2$ later in the analysis, the PŁ coefficients will depend on $\min_{s,a}\pi_{\theta_k}(a\mid s)$ and $\min_{s,b}\phi_{\psi_k}(b\mid s)$, which may not be lower bounded by any positive constant. Second, the coefficients involve τ, which is not a constant but needs to be

carefully chosen to control the error between the regularized problem and the original one.

4.4 Solving Regularized Markov Games

Leveraging the structure introduced in Section 4.3, our first aim is to establish the finite-time convergence of the GDA algorithm to the Nash equilibrium of the regularized Markov game under a fixed regularization weight $\tau > 0$. The GDA algorithm executes the updates

$$\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}), \qquad \psi_{k+1} = \psi_k - \beta_k \nabla_\psi J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k}). \tag{4.10}$$

The convergence bound we will derive reflects a trade-off for the regularization weight τ: when τ is large, we get faster convergence to the Nash equilibrium of the regularized problem, but it is farther away from the Nash equilibrium of the original one. The result in this section will inspire the τ adjustment schemes designed later in the paper to achieve the best possible convergence to the Nash equilibrium of the original unregularized Markov game.

It can be shown that the Nash equilibrium of the regularized Markov game is a pair of completely mixed policies, i.e. $\forall \tau > 0$ there exists $c_\tau > 0$ such that $\min_{s,a} \pi^\star_\tau(a \mid s) \geqslant c_\tau$, and $\min_{s,b} \phi^\star_\tau(b \mid s) \geqslant c_\tau$ [119]. In this work, we further assume the existence of a uniform lower bound on the entries of $(\pi^\star_\tau, \phi^\star_\tau)$ across τ.

Assumption 4.2. *There exists a positive constant c (independent of τ) such that for any $\tau > 0$*

$$\min_{s,a} \pi^\star_\tau(a \mid s) \geqslant c, \quad \min_{s,b} \phi^\star_\tau(b \mid s) \geqslant c.$$

To measure the convergence of the iterates to the Nash equilibrium of the regularized Markov game, we recall the definition of g_τ in Equation 4.2 and define

$$\delta^\pi_k = J_\tau(\pi^\star_\tau, \phi^\star_\tau) - g_\tau(\pi_{\theta_k}), \quad \delta^\phi_k = J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) - g_\tau(\pi_{\theta_k}). \tag{4.11}$$

The convergence metric is asymmetric for two players: the first player is quantified by its performance when the second player takes the most adversarial policy, while the second player is evaluated under the current policy iterate of the first player. We note that δ_k^π and δ_k^ϕ are non-negative, and $\delta_k^\pi = \delta_k^\phi = 0$ implies that $(\pi_{\theta_k}, \phi_{\psi_k})$ is the Nash equilibrium. Under this convergence metric, the following theorem states that the GDA updates in Equation 4.10 solve the regularized Markov game linearly fast. The proofs of the theoretical results of this paper are presented in Section C.1 of the appendix.

Theorem 4.1. *We define* $L = 3L_{\mathcal{H}} \max\{\tau, 1\}$, $C_1 = \frac{\rho_{\min} c^2}{64 \log(2)}$, *and* $C_2 = \frac{2\sqrt{|\mathcal{S}|}}{\sqrt{(1-\gamma)\rho_{\min} c}}$, *and choose the initial policy parameters to be* $\theta_0 = 0 \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ *and* $\psi_0 = 0 \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{B}|}$ *(the initial policies* π_{θ_0} *and* ϕ_{ψ_0} *are uniform). Let the step sizes of Equation 4.10 be*

$$\alpha_k = \alpha, \quad \beta_k = \beta,$$

with α, β *satisfying*

$$\max\{\alpha, \beta\} \leqslant \frac{1}{L}, \ \frac{\alpha}{\beta} \leqslant \min\{\frac{(1-\gamma)\rho_{\min}^3 c^2 \tau^2}{152 \log(2) |\mathcal{S}| L^2}, 8\}, \ \alpha \leqslant \min\{(L + \frac{C_2 L^2}{\tau})^{-1}, \frac{16|\mathcal{S}|}{(1-\gamma)\rho_{\min}^2 c^2 \tau}\}.$$

If Assumption 4.1 holds and

$$3\delta_0^\pi + \delta_0^\phi \leqslant C_1 \tau, \tag{4.12}$$

then the iterates of Equation 4.10 satisfy for all $k \geqslant 0$

$$3\delta_k^\pi + \delta_k^\phi \leqslant (1 - \frac{(1-\gamma)\alpha\tau\rho_{\min}^2 c^2}{32|\mathcal{S}|})^k (3\delta_0^\pi + \delta_0^\phi).$$

Theorem4.1 establishes the linear convergence of the iterates of Equation 4.10 to the Nash equilibrium of Equation 4.5, provided that the initial condition Equation 4.12 is satisfied. The convergence is faster when τ is large and slower when τ is small. Choosing

τ to be large enough guarantees the initial condition but causes the Nash equilibrium of the regularized Markov game to be distant from that of the original Markov game. This motivates us to make the regularization weight a decaying sequence that starts off large enough to meet the initial condition and becomes smaller over time to narrow the gap between the regularized Markov game and the original one. We discuss two such schemes of reducing the regularization weight in the next section.

4.5 Main Results - Solving the Original Markov Game

This section presents two approaches to adjust the regularization weight that allow the GDA algorithm to converge to the Nash equilibrium of the original Markov game. The first approach uses a piecewise constant weight and results in the nested-loop updates stated in Algorithm 4.1. In the inner loop the regularization weight and step sizes are fixed, and the two players update their policy iterates towards the Nash equilibrium of the regularized Markov game. The outer loop iteration reduces the regularization weight to make the regularized Markov game approach the original one. The regularization weight decays geometrically in the outer loop, i.e. $\tau_{t+1} = \eta\tau_t$, where $\eta \in (0, 1)$ must be carefully balanced. On the one hand, recalling the definition of g_τ in Equation 4.2 and defining

$$\delta_{t,k}^{\pi} = J_{\tau_t}(\pi_{\tau_t}^{\star}, \phi_{\tau_t}^{\star}) - g_{\tau_t}(\pi_{\theta_{t,k}}), \quad \delta_{t,k}^{\phi} = J_{\tau_t}(\pi_{\theta_{t,k}}, \phi_{\psi_{t,k}}) - g_{\tau_t}(\pi_{\theta_{t,k}}),$$

we need η to be large enough that if $\theta_{t,0}$ and $\psi_{t,0}$ observe the initial condition $3\delta_{t,0}^{\pi} + \delta_{t,0}^{\phi} \leqslant C_1\tau_t$, then so do $\theta_{t+1,0}$ and $\psi_{t+1,0}$ in the worst case. On the other hand, an η selected excessively large makes the reduction of τ_t too slow to achieve the best possible convergence rate. Our next theoretical result, as a corollary of Theorem4.1, properly chooses η and K_t and establishes the convergence of Algorithm 4.1 to the Nash equilibrium of the original original problem.

Corollary 4.1. *Suppose that Assumption 4.1-Assumption 4.2 hold and τ_0 is chosen such*

Algorithm 4.1: Nested-Loop Policy Gradient Descent Ascent Algorithm with Piecewise Constant Regularization Weight

Initialize: Policy parameters $\theta_{0,0} = 0 \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\psi_{0,0} = 0 \in \mathbb{R}^{\mathcal{S}\times\mathcal{B}}$, step size sequences $\{\alpha_t\}$ and $\{\beta_t\}$, an initial regularization parameter τ_0

for $t = 0, 1, \cdots, T$ **do**

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

    1) Max player update:

$$\theta_{t,k+1} = \theta_{t,k} + \alpha_t \nabla_\theta J_\tau(\pi_{\theta_{t,k}}, \phi_{\psi_{t,k}})$$

    2) Min player update:

$$\psi_{t,k+1} = \psi_{t,k} - \beta_t \nabla_\psi J_\tau(\pi_{\theta_{t,k+1}}, \phi_{\psi_{t,k}})$$

  end

  Set initial policies for next outer loop iteration $\theta_{t+1,0} = \theta_{t,K_t}$, $\psi_{t+1,0} = \psi_{t,K_t}$

  Reduce regularization weight $\tau_{t+1} = \eta\tau_t$ and properly adjust α_t, β_t

end

that $3\delta_{0,0}^{\pi} + \delta_{0,0}^{\phi} \leqslant C_1{\tau_0}^2$. *We choose* $\eta = \frac{C_1+2L_\delta}{2C_1+2L_\delta}$, *where* $L_\delta = 4\log|\mathcal{A}| + 3\log|\mathcal{B}| + \frac{\log|\mathcal{B}|}{1-\gamma}$ *and* C_1 *is defined in Theorem4.1. Then, under proper choices of* α_t *and* β_t, *the iterates of Algorithm 4.1 converge to a point such that*

$$J(\pi^\star, \phi^\star) - g_0(\pi_{\theta_{T,0}}) \leqslant \epsilon \quad \textit{and} \quad J(\pi_{\theta_{T,0}}, \phi_{\psi_{T,0}}) - g_0(\pi_{\theta_{T,0}}) \leqslant \epsilon \tag{4.13}$$

in at most $T = \mathcal{O}(\log(\epsilon^{-1}))$ *outer loop iterations. The total number of gradient updates required is* $\sum_{t=0}^{T} K_t = \mathcal{O}(\epsilon^{-3})$.

Corollary 4.1 guarantees that $(\pi_{\theta_T}, \phi_{\psi_T})$ converge to an ϵ-approximate Nash equilibrium of the original Markov game in $T = \mathcal{O}(\epsilon^{-3})$ gradient steps. In order to achieve this rate, K_t has to be adjusted along with τ_t: we need $K_t = \mathcal{O}(\tau_t^{-3})$ when τ_t becomes smaller than 1. The varying number of inner loop iterations may cause inconvenience for practical implementation. To address this issue, we next propose another scheme of adjusting the regularization weight that is carried out online along with the update of the policy iterates.

Presented in Algorithm 4.2, the second approach is a single-loop algorithm that reduces

[2]This inequality is guaranteed to hold with a large enough τ_0 if π_{θ_0} and ϕ_{ψ_0} are initialized to be uniform.

Algorithm 4.2: Policy Gradient Descent Ascent Algorithm with Diminishing Regularization Weight

Initialize: Policy parameters $\theta_0 = 0 \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\psi_0 = 0 \in \mathbb{R}^{\mathcal{S}\times\mathcal{B}}$, step size sequences $\{\alpha_k\}$ and $\{\beta_k\}$, regularization parameter sequence $\{\tau_k\}$

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

1) Max player update:

$$\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})$$

2) Min player update:

$$\psi_{k+1} = \psi_k - \beta_k \nabla_\psi J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_k})$$

end

the regularization weight as a polynomial function of the iteration k. We define the auxiliary convergence metrics

$$\delta_k^\pi = J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) - g_{\tau_k}(\pi_{\theta_k}), \quad \delta_k^\phi = J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) - g_{\tau_k}(\pi_{\theta_k}),$$

which measure the convergence of $(\pi_{\theta_k}, \phi_{\psi_k})$ to the Nash equilibrium of the Markov game regularized with weight τ_k. To judge the performance of the iterates in the original Markov game, we are ultimately interested in bounding $J(\pi^\star, \phi^\star) - g_0(\pi_{\theta_k})$ and $J(\pi_{\theta_k}, \phi_{\psi_k}) - g_0(\pi_{\theta_k})$. Thanks to Lemma 4.3, we can quantify how fast δ_k^π and δ_k^ϕ approach these desired quantities as τ_k decays to 0. Under an initial condition on δ_k^π and δ_k^ϕ, we now establish the convergence rate of Algorithm 4.2 to $(\pi^\star, \phi^\star)$ of Equation 4.1 through a multi-time-scale analysis.

Theorem 4.2. *Let the step sizes and regularization parameter be*

$$\alpha_k = \frac{\alpha_0}{(k+h)^{2/3}}, \quad \beta_k = \beta_0, \quad \tau_k = \frac{\tau_0}{(k+h)^{1/3}},$$

with α_0, β_0, τ_0, and $h \geqslant 1$ satisfying a system of inequalities discussed in details in the analysis. Under Assumption 4.1-Assumption 4.2, the iterates of Algorithm 4.2 satisfy for all

$k \geqslant 0$

$$J(\pi^\star, \phi^\star) - g_0(\pi_{\theta_k}) \leqslant \frac{C_1\tau_0 + 3(\log|\mathcal{A}| + \log|\mathcal{B}|)\tau_0}{3(k+h)^{1/3}}, \tag{4.14}$$

$$J(\pi_{\theta_k}, \phi_{\psi_k}) - g_0(\pi_{\theta_k}) \leqslant \frac{(1-\gamma)C_1\tau_0 + (\log|\mathcal{A}| + \log|\mathcal{B}|)\tau_0}{(1-\gamma)(k+h)^{1/3}}, \tag{4.15}$$

where the constant C_1 is defined in Theorem4.1.

Theorem4.2 states that the last iterate of Algorithm 4.2 converges to an $\mathcal{O}(k^{-1/3})$-approximate Nash equilibrium of the original Markov game in k iterations. This translates to the same sample complexity as Algorithm 4.1 derived in Corollary 4.1. Compared with Algorithm 4.1, reducing τ_k online along with the gradient updates in a single loop simplifies the algorithm and makes tracking the regularization weight, step sizes, and policy iterates simpler and more convenient. We note that the techniques in [29] may be used to analyze the finite-time performance of GDA for Markov games and lead to a convergence rate at least worse than $\mathcal{O}(k^{-1/10.5})$, which we improve over.

Remark 4.1. *Assumption 4.2 is a restrictive assumption that does not seem necessary but rather arises as an artifact of the current analysis. When we apply the weaker PL-type condition (Lemma 4.4) in the analysis, the entries of the iterates $\pi_{\theta_k}, \phi_{\psi_k}$ need to be uniformly lower bounded, which is difficult to establish using the game structure. We come up with an innovative induction approach to quantify the connection between $\min_{s,a}\pi_{\theta_k}(a \mid s), \min_{s,b}\phi_{\psi_k}(b \mid s)$ and the optimal gap $\delta_k^\pi, \delta_k^\phi$. This approach allows us to transform the uniform lower bound requirement on $\pi_{\theta_k}, \phi_{\psi_k}$ to that on the Nash equilibrium, leading to Assumption 4.2.*

A Markov game is said to be completely mixed if every Nash equilibrium of the game consists of a pair of completely mixed policies, i.e. $\min_{s,a}\pi^\star(a \mid s) > 0, \min_{s,b}\phi^\star(b \mid s) > 0$ for any Nash equilibrium $(\pi^\star, \phi^\star)$ of the Markov game (if more than one exists). Assumption 4.2 intuitively seems no stronger than requiring the original Markov game to be completely mixed. If the original Markov game has at least one completely mixed

Nash equilibrium, the Nash equilibrium of the regularized Markov game should also be completely mixed even when the regularization weight is small, since the regularization encourages the solution to be more uniform. The reward function that results in completely mixed Markov games is well studied in [136–138].

4.6 Numerical Simulations

In this section, we numerically verify the convergence of Algorithm 4.2 on small-scale synthetic Markov games. Our aim is to confirm that the algorithm indeed converges rather than to visualize the exact convergence rate, as achieving the theoretical rate derived in Theorem4.2 requires very careful selection of all involved parameters. Considering an environment with $|\mathcal{S}| = 2$ and $|\mathcal{A}| = |\mathcal{B}| = 2$, we first choose the reward and transition probability kernel such that the Markov game is completely mixed[3].

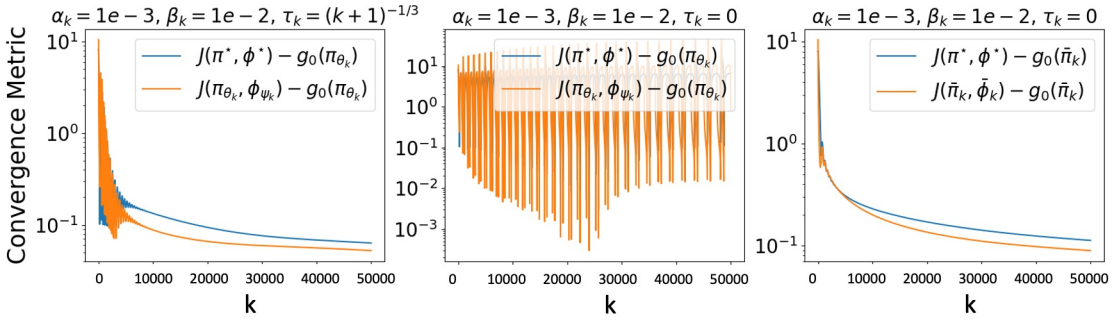


Figure 4.1: Convergence of GDA for a Completely Mixed Markov game

We run Algorithm 4.2 for 50000 iterations with $\alpha_k = 10^{-3}$, $\beta_k = 10^{-2}$, $\tau_k = (k+1)^{-1/3}$, and measure the convergence of π_k and ϕ_k by metrics considered in Equation 4.14 and Equation 4.15 of Theorem4.2. As shown in the first plot of Figure 4.1, the last iterate exhibits an initial oscillation behavior but converge smoothly after 10000 iterations. In comparison, we visualize the convergence of the last iterate and averaged iterate of the

[3]To create a completely mixed game with $|\mathcal{A}| = |\mathcal{B}| = 2$, we simply need to choose the reward function such that $r(s, \cdot, \cdot)$ as a 2x2 matrix is diagonal dominant or sub-diagonal dominant for any state $s \in \mathcal{S}$, and we can use an arbitrary transition probability kernel. The exact choice of the reward function and transition kernel as well as the Nash equilibrium of this Markov game are presented in Section C.3 of the appendix.

GDA algorithm without any regularization (second and third plots of Figure 4.1), where the average is computed with equal weights as $\bar{\pi}_k = \frac{1}{k+1}\sum_{t=0}^{k}\pi_{\theta_t}, \bar{\phi}_k = \frac{1}{k+1}\sum_{t=0}^{k}\phi_{\psi_t}$. The existing theoretical results in this case guarantee the convergence of the averaged iterate but not the last iterate [29]. According to our simulations, the last iterate indeed does not converge, while the averaged iterate does, but at a slower rate than the convergence of the last iterate of the GDA algorithm under the decaying regularization.

The theoretical results derived in this paper rely on Assumption 4.2. To investigate whether this assumption is truly necessary, we also apply Algorithm 4.2 to a Markov game that has a deterministic Nash equilibrium and does not observe Assumption 4.2[4]. As illustrated in Figure 4.2, the experiment shows that Algorithm Algorithm 4.2 still converges correctly to $(\pi^\star, \phi^\star)$ of Equation 4.1. This observation suggests that Assumption 4.2 may be an artifact of the current analysis and motivates for us to investigate ways to remove/relax this assumption in the future. We note that the pure GDA approach without regularization also has a last-iterate convergence and does not exhibit the oscillation behavior observed in Figure 4.1, since the gradients of both players never change signs regardless of the policy of the opponent in this Markov game.

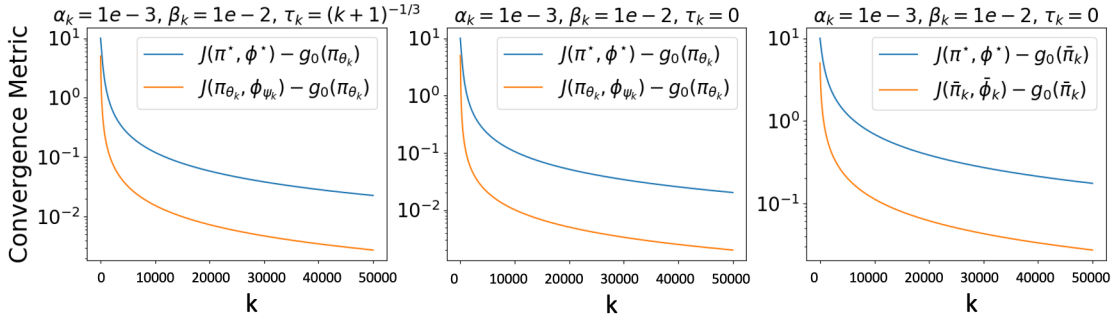


Figure 4.2: Convergence of GDA for a Deterministic Markov game

[4]The detailed description of the game is again deferred to Section C.3 of the appendix.

4.7 Future Directions

Our current work on Markov games relies on the Nash equilibrium being a pair of completely mixed policies. Numerical simulations suggest that our proposed algorithm converges efficiently in Markov games that do not satisfy this assumption. However, significant challenges are present in removing or relaxing this assumption, and we leave it as a possible future direction.

It is also interesting to investigate the extension of the work to the sample-based setting. The gradient of the policy optimization objective in Equation 4.9 depends on the value functions, which can be estimated with a critic variable updated on a faster timescale. Our current analysis for the deterministic gradient setting relies on a connection between the optimality gap $(\delta_k^{\pi},\, \delta_k^{\phi})$ and difference in smallest policy entry $(\min_{s,a} \pi^{\star}(a \mid s) - \min_{s,a} \pi_k(a \mid s), \min_{s,a} \phi^{\star}(b \mid s) - \min_{s,a} \phi_k(b \mid s))$ established in Lemma C.4 and its proof. Showing a similar connection under stochastic errors is the biggest challenge of this extension. It is possible that a convergence with high probability (rather than in expectation) is the correct metric to use to control the aforementioned difference in smallest policy entry in the stochastic setting.

CHAPTER 5
ACCELERATING POWER SYSTEM OPTIMIZATION WITH REINFORCEMENT LEARNING

In this chapter, we apply reinforcement learning to solve a parameter selection problem in power system optimization. In particular, we consider the alternating current optimal power flow (ACOPF) problem, which studies minimizing the cost of generating and transmitting electrical power while satisfying the network demands and obeying physical transmission laws. Formulated as a complicated and highly non-convex optimization program, the ACOPF problem is crucial for the efficient operation of modern power networks and needs to be solved at a high frequency in real time as the network demands and topology change. One of the most successful approaches of solving large-scale ACOPF problems leverages the alternating direction method of multipliers (ADMM) algorithm [139], which efficiently distributes the computation and accelerates the solution.

However, it is known that the convergence behavior of ADMM in this context is highly dependent on the selection of penalty parameters, which are usually chosen heuristically. [140] shows that poorly selected parameters can severely slow down the algorithm convergence or even lead to divergence.

Motivated to develop a more reliable penalty parameter selection scheme, we view the ADMM solving process as stochastic environment and propose learning a parameter selection policy using RL, with the goal of minimizing the number of iterations until convergence. We train our RL policy using deep Q-learning, and show that this policy can result in significantly accelerated convergence (up to a 59% reduction in the number of iterations compared to existing, curvature-informed penalty parameter selection methods). We also show the superior generalizability of our policy, which performs well under unseen loading schemes as well as under unseen losses of lines and generators (up to a 50%

reduction in iterations). Though initially formulated as a single agent RL problem, our solution interestingly turns out to have a multi-agent interpretation[1].

5.1 Related Works

To speed up convergence and reduce the effort of penalty parameter tuning in ADMM, adaptive penalty parameter algorithms have been studied in order to update penalty parameters during the optimization using feedback from the previous iteration. Examples include residual balancing [141], which increases or decreases penalty parameters based on the relative magnitudes of the primal and dual residuals, and methods that use estimates of the local curvature of the dual function to inform updates [142]. Mhanna et al. in [143] demonstrate significantly improved convergence performance for the ACOPF problem using adaptive penalty parameter algorithms over vanilla ADMM with static penalty parameters. However, the techniques in [143] still rely on tuned parameters within the adaptive algorithm, and also require additional logic steps and the computation and storage of gradient information.

Ultimately, these existing adaptive penalty parameter algorithms rely on heuristics, presenting an opportunity for their replacement with machine learning techniques that may have superior performance. In this work, we develop a reinforcement learning (RL) [144] method to train a policy for selecting penalty parameters to accelerate the convergence of an ADMM algorithm for solving ACOPF problems. The ADMM parameter selection task has a sequential decision making structure, as penalty parameters are updated based on feedback from past iterations. RL, as a convenient framework for sequential decision making problems, is a natural fit for this task.

Machine learning techniques have been used to design optimization methods [145, 146]. There are fewer works that develop embedded-ML methods specifically for distributed optimization algorithms. In [147], a recurrent neural network is trained to predict the converged values of variables in ADMM subproblems for DC-OPF. In [148], the authors

[1]The presentation in this chapter is partly adapted from [40].

replace ADMM subproblems with an RL policy that predicts solutions. In [149], the authors learn to solve ADMM subproblems by recasting them as deep neural networks. Recent contemporaneous work [150] trains an RL policy to tune parameters to accelerate ADMM convergence using policy gradient methods; however, they focus on convex QP problems with convergence guarantees and do not specifically consider power systems problems. Moreover, RL methods have shown promise in other power systems applications [151, 152].

5.2 Preliminaries

In this section, we provide a brief overview of the ADMM algorithm, present the ACOPF problem formulation, and describe how the underlying objective for ACOPF can be re-formulated to fit into the ADMM framework and to be solved with ADMM.

5.2.1 Alternating Direction Method of Multipliers

ADMM is designed to solve problems of the form

$$\begin{aligned} \min_{x \in \mathbb{R}^{n_1}, \bar{x} \in \mathbb{R}^{n_2}} \quad & f(x) + g(\bar{x}) \\ \text{s.t.} \quad & Ax + B\bar{x} = c, \end{aligned} \tag{5.1}$$

where $A \in \mathbb{R}^{n_3 \times n_1}$, $B \in \mathbb{R}^{n_3 \times n_2}$, and $c \in \mathbb{R}^{n_3}$, and where $f : \mathbb{R}^{n_1} \to \mathbb{R}$ and $g : \mathbb{R}^{n_2} \to \mathbb{R}$ are closed functions. Only linear equality constraints are present in this formulation, but we note that non-linear and/or inequality constraints can be easily modeled by properly introducing slack variables [153].

Let $y \in \mathbb{R}^{n_3}$ be the vector of Lagrange multipliers used to enforce the constraints. We form the augmented Lagrangian as

$$L_\rho(x, \bar{x}, y) = f(x) + g(\bar{x}) + y^T(Ax + B\bar{x} - c) + \frac{1}{2}(Ax + B\bar{x} - c)^\top \Omega (Ax + B\bar{x} - c).$$

The matrix $\Omega \in \mathbb{R}^{n_3 \times n_3}$ is a diagonal matrix with the diagonal entry defined as $\Omega_{ii} = \rho_i$ for

some scalar $\rho_i > 0$. We refer to ρ_i as the i-th penalty parameter.

The ADMM algorithm essentially uses a blend of dual descent and method of multipliers to find the saddle point of the Lagrangian. Let k be the ADMM iteration counter, where iterates are marked via square brackets in superscript. In each iteration of ADMM, we sequentially update variable x according to Equation 5.2a , variable $\bar{x}$ according to Equation 5.2b, and the Lagrange multipliers y via Equation 5.2c.

$$x^{[k+1]} = \underset{x}{\operatorname{argmin}}\, L_\rho(x, \bar{x}^{[k]}, y^{[k]}) \tag{5.2a}$$

$$\bar{x}^{[k+1]} = \underset{\bar{x}}{\operatorname{argmin}}\, L_\rho(x^{[k+1]}, \bar{x}, y^{[k]}) \tag{5.2b}$$

$$y^{[k+1]} = y^{[k]} + \Omega(Ax^{[k+1]} + B\bar{x}^{[k+1]} - c) \tag{5.2c}$$

The primal residual $r_p^{[k]}$ and dual residual $r_d^{[k]}$, defined as follows, provide a metric of convergence.

$$r_p^{[k]} = Ax^{[k]} + B\bar{x}^{[k]} - c \tag{5.3}$$

$$r_d^{[k]} = 2\Omega A^T B\left(\bar{x}^{[k]} - \bar{x}^{[k-1]}\right). \tag{5.4}$$

The ADMM iterations proceed until the l_2 norms of the primal and dual residuals, which represent the feasibility of the primal and dual problems, meet their convergence thresholds $\epsilon_p > 0$ and $\epsilon_d > 0$, respectively:

$$\left\|r_p^{[k]}\right\|_2 \leqslant \epsilon_p \quad \text{and} \quad \left\|r_d^{[k]}\right\|_2 \leqslant \epsilon_d, \tag{5.5}$$

5.2.2 Alternating Current Optimal Power Flow

Consider a power system represented by an undirected graph $(\mathcal{B}, \mathcal{L})$, where $\mathcal{B}$ and $\mathcal{L}$ denote the collection of nodes and edges. Each node $i \in \mathcal{B}$, also referred to as a bus, has a complex power demand denoted as $d_i = p_i^d + j * q_i^d$ for some $p_i^d, q_i^d \in \mathbb{R}$. The voltage of bus i is $v_i \in \mathbb{C}$,

and we use e_i and f_i to denote the real and imaginary parts, i.e. $v_i = e_i + j * f_i$. We can alternatively represent the voltage in a polar form with $w_i = e_i^2 + f_i^2$ and $\theta_i = \arctan(f_i/e_i)$. A subset of the buses may have a power generator attached, and we use $\mathcal{G} \subseteq \mathcal{B}$ to denote the collection of generators[2]. Each generator bus $i \in \mathcal{G}$ can generate a complex power with a real part $p_i^g \in \mathbb{R}$ and imaginary part $q_i^g \in \mathbb{R}$.

An edge of the graph, also referred to as a branch, represents a transmission line. For a branch from bus i to j, p_{ij} and q_{ij} denote the real and imaginary power flow through the branch in the nominal direction, and p_{ji} and q_{ji} denote the real and imaginary power flow in the reverse direction. We note that p_{ji} and q_{ji} are not simply the negative of p_{ij} and q_{ij}; these quantities are determined from the voltage at bus i and j by solving a system of power flow equations, which corresponds to Equation 5.6d-Equation 5.6g in the optimization problem below.

The objective of the ACOPF problem, presented in Equation 5.6, is to find the most economic operating point of the generators that obeys the physical laws and satisfies the power demand p_i^d, q_i^d at every node i. The generation cost c_i is a quadratic function in the real power output. Equation 5.6b-Equation 5.6c are known as power balance equations and represent the power transmission laws along with Equation 5.6d-Equation 5.6j. Equation 5.6k-Equation 5.6l restrict the magnitude of the power flow between bus i and j. Equation 5.6m-Equation 5.6n represent the limit of the power generators.

$$\min_{p_{g_i}, q_{g_i}, w_i, \theta_i, w_{ij}^R, w_{ij}^I} \sum_{i \in \mathcal{G}} c_i(p_i^g) \tag{5.6a}$$

$$\text{s.t. } p_i^g - p_i^d = g_i^S w_i + \sum_{j \in \mathcal{N}_i} p_{ij}, \qquad \forall i \in \mathcal{B} \tag{5.6b}$$

$$q_i^g - q_i^d = -b_i^S w_i + \sum_{j \in \mathcal{N}_i} q_{ij}, \qquad \forall i \in \mathcal{B} \tag{5.6c}$$

$$p_{ij} = g_{ii} w_i + g_{ij} w_{ij}^R + b_{ij} w_{ij}^I, \qquad \forall (i,j) \in \mathcal{L} \tag{5.6d}$$

[2]We assume that there is at most one generator at each bus to simplify the discussion. In general, multiple generators can be on a bus, and our problem formulation easily extends to such scenarios.

$$q_{ij} = -b_{ii}w_i - b_{ij}w_{ij}^R + g_{ij}w_{ij}^I, \qquad \forall (i,j) \in \mathcal{L} \tag{5.6e}$$

$$p_{ji} = g_{jj}w_j + g_{ji}w_{ij}^R - b_{ji}w_{ij}^I, \qquad \forall (i,j) \in \mathcal{L} \tag{5.6f}$$

$$q_{ji} = -b_{jj}w_j - b_{ji}w_{ij}^R - g_{ji}w_{ij}^I, \qquad \forall (i,j) \in \mathcal{L} \tag{5.6g}$$

$$-2\pi \leqslant \theta_i \leqslant 2\pi, \qquad \forall i \in \mathcal{B} \tag{5.6h}$$

$$(w_{ij}^R)^2 + (w_{ij}^I)^2 = w_i w_j, \qquad \forall (i,j) \in \mathcal{L} \tag{5.6i}$$

$$\theta_i - \theta_j = \arctan(w_{ij}^I / w_{ij}^R), \qquad \forall (i,j) \in \mathcal{L} \tag{5.6j}$$

$$\sqrt{p_{ij}^2 + q_{ij}^2} \leqslant \bar{r}_{ij}, \qquad \forall (i,j) \in \mathcal{L} \tag{5.6k}$$

$$\sqrt{p_{ji}^2 + q_{ji}^2} \leqslant \bar{r}_{ij}, \qquad \forall (i,j) \in \mathcal{L} \tag{5.6l}$$

$$\underline{p}_i^g \leqslant p_i^g \leqslant \overline{p}_i^g, \qquad \forall g_i \in \mathcal{G} \tag{5.6m}$$

$$\underline{q}_i^g \leqslant q_i^g \leqslant \overline{q}_i^g, \qquad \forall g_i \in \mathcal{G} \tag{5.6n}$$

We use $\mathcal{N}_i$ to denote the neighbors of bus i, i.e. $\mathcal{N}_i = \{j \in \mathcal{B} : (i,j) \in \mathcal{L}\}$. The decision variables of this optimization program include $p_i^g, q_i^g, w_i, \theta_i$ and auxiliary variables w_{ij}^R and w_{ij}^I, which are defined to be $\sqrt{w_i w_j}\cos(\theta_i - \theta_j)$ and $\sqrt{w_i w_j}\sin(\theta_i - \theta_j)$. The other quantities in Equation 5.6 are parameters that depend on the structure and physical properties of the power network (see [143] for details).

5.2.3 ACOPF Solved via ADMM

The authors in [143] propose a method to decompose the ACOPF problem Equation 5.6 based on the observation that certain variables can be decoupled by properly duplicating these variables and enforcing a consensus through coupling constraints. Ultimately, Equation 5.6 can be reformulated as the composition of small sub-problems and written in the form of Equation 5.1 with suitable choices of A, B, c, and (non-convex) loss functions f and g.

Recall that the i-th coupling constraint in the ADMM formulation is associated with penalty parameter ρ_i. In [140], improved convergence performance is observed for ACOPF when ρ_i values are assigned based on the type of coupling constraint they are penalizing.

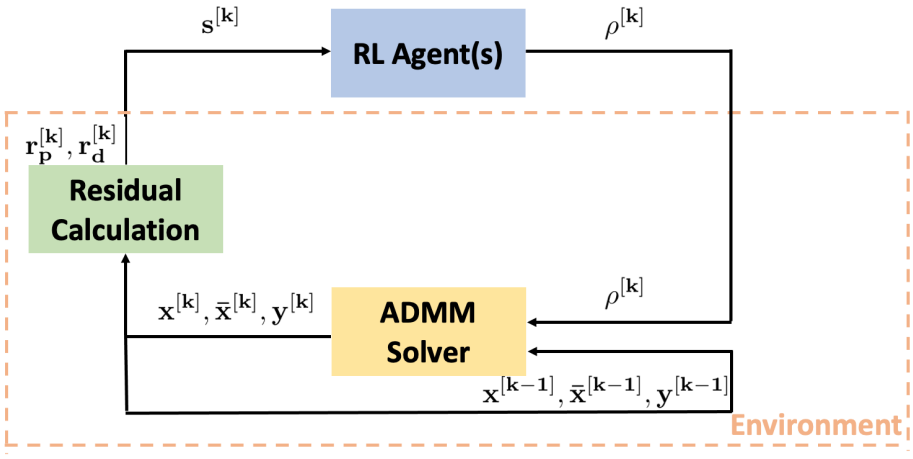


Figure 5.1: Environment (ADMM Solver) and RL Agent Interaction

They categorize the coupling constraints into two different types: constraints that correspond to the real (p) and reactive (q) power flows, and constraints that correspond to voltages (v) and angles (θ). We use n_{pq} and $n_{v\theta}$ to denote the number of the two types of constraints, and define $\mathcal{C}_{pq}$ and $\mathcal{C}_{v\theta}$ to be the index set of power related and voltage related constraints, respectively. We use $\rho_{pq} \in \mathbb{R}^{n_{pq}}$ for the penalty parameters for the p or q coupling constraints and $\rho_{v\theta} \in \mathbb{R}^{n_{v\theta}}$ for the penalty parameters for the v, w, or θ coupling constraints.

5.3 Reinforcement Learning Algorithm Design

While we seek to reduce the number of ADMM iterations until convergence by properly choosing penalty parameters, the goal of an RL agent is to maximize the discounted cumulative reward it collects from the environment. To translate our objective to that of the RL agent, we have to model our ADMM parameter selection problem as a suitable RL problem, which includes identifying the environment and dynamics and making the proper choice of the state space, action space, and reward function.

We regard the ADMM solution process as the RL environment in the following sense. Each iteration of the ADMM algorithm corresponds to one RL iteration. In iteration $k = 0, 1, \ldots$, the agent observes the current state of the ADMM solver $s^{[k]}$. Based on $s^{[k]}$, the agent selects an action $a^{[k]}$, which is simply a choice of $\rho^{[k]}$, the penalty parameter of

the k-th iteration, and receives a reward $R(s^{[k]}, a^{[k]})$, which we will design to reflect the value of the current state to the ADMM convergence. The parameter $\rho^{[k]}$ is then fed back to the ADMM solver for another ADMM iteration. This process is repeated until both the primal and dual residuals from the ADMM solve drop below the thresholds in Equation 5.5. The interaction of the environment and the agent in ADMM solving process is shown in Figure 5.1.

State space $\mathcal{S}$: The state provides an important source of information that should summarize the progress of the ADMM algorithm and include key factors necessary for the agent to make decisions about ρ. In this problem, we naturally expect the primal and dual residuals to contain information about the optimal choice of ρ. To ensure that $s^{[k]}$ sufficiently represents the state of the ADMM solving process, we include the past n-point history of the residuals in $s^{[k]}$, i.e.

$$s^{[k]} = \left[(r_p^{[k-n+1]}, r_d^{[k-n+1]}), \cdots, (r_p^{[k]}, r_d^{[k]})\right] \in \mathbb{R}^{2n \times (n_{pq}+n_{v\theta})}.$$

Action space $\mathcal{A}$: As ρ values are continuous variables, the action space for this problem is continuous, which dictates the use of RL algorithms compatible with continuous action spaces. Nevertheless, in this work we discretize the action space into the collection of 10 values, motivated by the observation that the effective discretization of a continuous action space can sometimes lead to better trained policies [154].

The existing literature suggests that ρ values picked from a certain range result in superior convergence speed. Specifically, [143] considers using two different ρ for the two types of constraints: for constraints related to real and reactive power, $\rho_{pq} = 400$ is used for IEEE 9-bus, 30-bus, and 118-bus systems; for constraints related to voltage, $\rho_{v\theta} = 40000$ is used for IEEE 9-bus and 30-bus systems and $\rho_{v\theta} = 4000$ is used for the 118-bus system. Though this particular choice of the parameters may not be optimal, it suggests a reasonable interval for ρ to provide to the RL agent. We select $[100, 1000]$ as the range of ρ_{pq}, and $[500, 70000]$

for $\rho_{v\theta}$ in the 9-bus and 30-bus systems and $[500, 7000]$ in the 118-bus system, discretized as shown in Table 5.1.

Table 5.1: RL Action Space & Initial ρ Values

ρ Category	Initial Value	Action Space
ρ_{pq}	400	{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000}
$\rho_{v\theta}$ (9-, 30-bus)	40000	{500, 2000, 5000, 10000, 20000, 30000, 40000, 50000, 60000, 70000}
$\rho_{v\theta}$ (118-bus)	4000	{500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 5500, 7000}

Reward function R: The reward function is a crucial signal that affects the behavior of the agent. We have to carefully design the reward function to translate our objective, which is to accelerate ADMM convergence, correctly to the agent. The reward function R should be chosen such that $R(s, a)$ is large if taking action a while in state s leads to fast convergence and small if taking action a while in state s leads to slow convergence. With this in mind, a natural choice of the reward function is a large bonus given only to the convergence state; for instance,

$$R_{\text{conv}}(s^{[k]}, a^{[k]}) = \begin{cases} 200, & \text{if } \left\| r_p^{[k+1]} \right\|_2 \leqslant \epsilon_p \text{ and } \left\| r_d^{[k+1]} \right\|_2 \leqslant \epsilon_d, \\ 0, & \text{else.} \end{cases}$$

Due to the presence of the discount factor $\gamma \in (0, 1)$, the reward received further in the future becomes less valuable. Therefore, to maximize the discounted cumulative reward under this reward function, the agent will aim to reach the convergence state in as few iterations as possible.

Though this design of the reward function encodes our objective, it causes the agent to receive extremely sparse reward signals in the training process. Until the very last iteration, the agent will not receive any useful signal throughout the hundreds or thousands

of iterations that are typically required for ADMM algorithms to converge for moderately sized ACOPF problems. Sparse rewards commonly cause exploration and credit assignment issues in RL and significantly slow down the learning process.

To offer a denser signal to the RL agent, we add the residuals to the reward function. Specifically, the reward received by the agent in state $s^{[k]}$ is proportional to the reduction in $\|r_p^{[k+1]}\|_2$ and $\|r_d^{[k+1]}\|_2$ from $\|r_p^{[k]}\|_2$ and $\|r_d^{[k]}\|_2$:

$$R_{\text{res}}(s^{[k]}, a^{[k]}) = \frac{1}{Z_p}(\|r_p^{[k]}\|_2 - \|r_p^{[k+1]}\|_2) + \frac{1}{Z_d}(\|r_d^{[k]}\|_2 - \|r_d^{[k+1]}\|_2),$$

where Z_p and Z_d are normalizing factors that balance the magnitude difference between the primal and dual residuals. This reward function makes sense, as achieving fast convergence is equivalent to quickly driving the residuals to the thresholds. This reward is non-zero in every ADMM iteration.

While we observe that the combination of R_{conv} and R_{res} works well in this problem, we further innovate the reward function design by taking advantage of the non-counterfactual nature of the environment. We note that in most RL problems, the environment transition is irreversible, that is, once an action $a^{[k]}$ is deployed in state $s^{[k]}$, the environment moves forward to the next state $s^{[k+1]}$, and the consequence of selecting a different action in $s^{[k]}$ is never observable. However, in this problem, the progress of every ADMM iteration can be saved and we can therefore try different actions in the same state and compare their outcomes. This feature of the environment affords more flexibility in the reward design.

In this work, we use a reward function computed with the help of a baseline policy $\tilde{\pi}$. In state $s^{[k]}$, we select the baseline action $\tilde{a}^{[k]} \sim \tilde{\pi}(\cdot \mid s^{[k]})$ and observe the resulting next state $\tilde{s}^{[k+1]}$ including primal and dual residuals $\tilde{r}_p^{[k+1]}$ and $\tilde{r}_d^{[k+1]}$. We note that this baseline action is only used to compute the residuals. We roll back to state $s^{[k]}$ once the residuals are collected. From state $s^{[k]}$, we then deploy the RL policy, making the environment transition to $s^{[k+1]}$ and reveal $r_p^{[k+1]}$ and $r_d^{[k+1]}$. The reward is defined as the relative advantage of the

RL policy over the baseline:

$$R_b(s^{[k]}, a^{[k]}) = \frac{\|\tilde{r}_p^{[k+1]}\|_2 - \|r_p^{[k+1]}\|_2}{\|\tilde{r}_p^{[k+1]}\|_2} + \frac{\|\tilde{r}_d^{[k+1]}\|_2 - \|r_d^{[k+1]}\|_2}{\|\tilde{r}_d^{[k+1]}\|_2}.$$

This reward function essentially aims to achieve the same goal as R_{res}, but can have much smaller variance. We note that $\|r_p^{[k+1]}\|_2 - \|r_p^{[k]}\|_2$ and $\|r_d^{[k+1]}\|_2 - \|r_d^{[k]}\|_2$ can fluctuate across several orders of magnitude through ADMM iterations regardless of the choice of ρ. The reward function R_b effectively removes the impact of the natural fluctuation of the residuals and makes the variance of R_b significantly smaller than that of R_{res}. We emphasize that the sole purpose of the baseline policy is to offset the fluctuation in the norm of the residuals over iterations. Therefore, the baseline policy can be very simple. In the experiments of this work, the baseline policy is to always use $\rho_{pq} = 500$ and $\rho_{v\theta} = 500$. Accordingly, the reward function we choose in this work combines r_{conv} and r_b:

$$R(s^{[k]}, a^{[k]}) = R_{\text{conv}}(s^{[k]}, a^{[k]}) + R_b(s^{[k]}, a^{[k]}).$$

5.3.1 Factorized Entry-wise Policy & Multi-Agent Interpretation

We have discussed the transformation of the ADMM parameter selection problem into a RL problem where the policy selects a vector ρ given the state vector. With the ten possible choices of ρ values for each constraint, the total cardinality of the action space is $10^{n_{pq}+n_{v\theta}}$, which grows exponentially in the number of constraints and quickly becomes computationally intractable. To address this issue, we reduce the action space by simplifying the policy using the structure of ACOPF.

First, we find that the existing heuristic methods of adjusting ρ, which determine ρ_i in an element-wise manner only using the residuals of constraint i, lead to reasonably accelerated convergence rate. This observation suggests that the local information may provide sufficient knowledge for us to (almost) optimally determine the local penalty parameter ρ_i. As a result,

we are motivated to factorize the policy into the product of local policies with significantly reduced action spaces. Specifically, let π_i denote the policy for updating parameter ρ_i and s_i denotes the portion of the state vector s associated with constraint i. We assume that the optimal policy $\pi^\star$ can be factorized as

$$\pi^\star(\rho \mid s) = \prod_{i=1}^{n_{pq}+n_{v\theta}} \pi_i^\star(\rho_i \mid s_i),$$

which means that we can equivalently train smaller policies π_i for each $i = 1, \ldots, (n_{pq} + n_{v\theta})$.

Learning the set of small policies with its size scaling up linearly with the number of constraints, however, can still be prohibitive in computation and memory usage. Therefore, we make one more simplification by restricting all power related constraints to employ the the same policy, i.e. $\pi_i^\star = \pi_{pq}^\star$ for all $i \in \mathcal{C}_{pq}$, and all voltage and angle related constraints to employ the same policy, i.e. $\pi_i^\star = \pi_{v\theta}^\star$ for all $i \in \mathcal{C}_{v\theta}$. This means that the policy can be represented as

$$\pi^\star(a \mid s) = \left(\prod_{i \in \mathcal{C}_{pq}} \pi_{pq}^\star(a_i \mid s_i) \right) \left(\prod_{i \in \mathcal{C}_{v\theta}} \pi_{v\theta}^\star(a_i \mid s_i) \right). \tag{5.7}$$

As a result of this factorization, we only need two small entry-wise policies, each mapping the local state vector $s_i \in \mathbb{R}^{2n}$ to an action from 10 possible choices. The cost of maintaining and updating such policies is fairly small.

Along with advantages in computational tractability, another important benefit of the factorized entry-wise policy lies in its ability to be deployed to ACOPF ADMM problems with different numbers of constraints from the one seen by the RL agent in training. This means that the entry-wise policy pair trained under one power network can be flexibly applied to various other network structures. Later in Section 5.4, we will discuss an important generalization of the learned policy to minor system modifications, where it is necessary for the policy to adapt to a change in the number of constraints.

Interestingly, another interpretation of this factorized policy is that there exist two cooperative agents in the environment with aligned reward functions. In the current learning paradigm, the agents do not communicate with each other and can only learn individually. Suppose we use an iterative learning algorithm to find $\pi^\star_{pq}$ and $\pi^\star_{pq}$, where we use $\pi^{[k]}_{pq}$, $\pi^{[k]}_{v\theta}$ to denote the policy iterates in the k-th iteration. Then, when $\pi^{[k]}_{pq}$ (resp. $\pi^{[k]}_{v\theta}$) is updated, it essentially seeks to find the optimal policy in the environment with the state transition and reward function marginalized over $\pi^{[k-1]}_{v\theta}$ (resp. $\pi^{[k-1]}_{pq}$).

5.3.2 Q Learning Algorithm in ADMM Solver

In Algorithm 5.1, we formally present how we incorporate the RL agent to the ADMM solver. We use deep Q learning to find $\pi^\star_{pq}$, $\pi^\star_{v\theta}$. Specifically, we maintain and update two neural networks parameterized by ψ_{pq} and $\psi_{v\theta}$ to approximate the Q function of $\pi^\star_{pq}$ and $\pi^\star_{v\theta}$, updated as shown in line 12-14. The behavior policy used to generate the samples is the ϵ-greedy policy based on ψ_{pq} and $\psi_{v\theta}$ (line 15), where we select ϵ to be a small constant.

The ADMM solver employs a prescribed ρ vector initially and starts sampling ρ from the behavior policy after n iterations.

5.4 Numerical Experiments

We demonstrate the performance of our RL policy on the 9-bus, 30-bus, and 118-bus IEEE networks in the MATPOWER format [155]. Two additional evaluation tasks are carried out to validate the generalization of the learning performance to the practical scenarios in power system operations. In the first task, the RL policy is evaluated for its effectiveness in unseen load profiles in the original network. This is an important task as the loads of a power system frequently change, requiring the ACOPF problem to be solved repeatedly in an efficient way. The second task tests the RL policy on a slightly modified version of the system by removing generators and/or disconnecting transmission lines. This task is more challenging and also important in practice since we may need to solve ACOPF problems under generator

Algorithm 5.1: Parameter Learning Through Q-Learning in ADMM ACOPF Solver

ADMM initialization: Initial parameters $x^{[0]} \in \mathbb{R}^{n_1}, \bar{x}^{[0]} \in \mathbb{R}^{n_1}, y^{[0]} \in \mathbb{R}^{n_3}, \bar{\rho} \in \mathbb{R}^{n_3}$

RL initialization: Initial Q function parameters $\psi_{pq}^{[0]}$ for pq agent and $\psi_{v\theta}^{[0]}$ for $v\theta$ agent, step size sequence $\alpha^{[k]}$, exploration parameter ϵ, state vector length n

for $k = 0, 1, 2, ...$ **do**

  if $k \geqslant n$ **then**

    Compute residuals $r_d^{[k]}$, $r_p^{[k]}$ from $x^{[k]}, \bar{x}^{[k]}$ and form state vector $s^{[k]} = [(r_p^{[k-n+1]}, r_d^{[k-n+1]}), \cdots, (r_p^{[k]}, r_d^{[k]})]$

    Sample action $a_i^{[k]}$ in an element-wise manner and translate to $\rho^{[k]}$

$$a_i^{[k]} \sim \begin{cases} \widehat{\pi}_{pq}^{[k]}(\cdot \mid s_i^{[k]}), & \text{for } i \in \mathcal{C}_{pq} \\ \widehat{\pi}_{v\theta}^{[k]}(\cdot \mid s_i^{[k]}), & \text{for } i \in \mathcal{C}_{v\theta} \end{cases}$$

  else

    Use the initial ρ value: $\rho^{[k]} = \bar{\rho}$

  end if

  Perform an ADMM update Equation 5.2 with penalty parameter $\rho^{[k]}$

  if $k \geqslant n$ **then**

    Observe $R(s^{[k]}, a^{[k]})$ and $s^{[k+1]}$ and compute the vector Q^{target} such that

$$Q_i^{\text{target}} = \begin{cases} R(s^{[k]}, a^{[k]}) + \max_a Q^{\psi_{pq}^{[k]}}(s_i^{[k+1]}, a), & \text{for } i \in \mathcal{C}_{pq} \\ R(s^{[k]}, a^{[k]}) + \max_a Q^{\psi_{v\theta}^{[k]}}(s_i^{[k+1]}, a), & \text{for } i \in \mathcal{C}_{v\theta} \end{cases}$$

    Compute loss

$$\ell(\psi_{pq}^{[k]}, \psi_{v\theta}^{[k]}) = \sum_{i \in \mathcal{C}_{pq}} \left(Q^{\psi_{pq}^{[k]}}(s_i^{[k]}, a^{[k]}) - Q_i^{\text{target}} \right)^2 + \sum_{i \in \mathcal{C}_{v\theta}} \left(Q^{\psi_{v\theta}^{[k]}}(s_i^{[k]}, a^{[k]}) - Q_i^{\text{target}} \right)^2$$

    Update the Q function parameter

$$\psi_{pq}^{[k+1]} = \psi_{pq}^{[k]} - \alpha^{[k]} \nabla_{\psi_{pq}} \ell(\psi_{pq}^{[k]}, \psi_{v\theta}^{[k]})$$
$$\psi_{v\theta}^{[k+1]} = \psi_{v\theta}^{[k]} - \alpha^{[k]} \nabla_{\psi_{v\theta}} \ell(\psi_{pq}^{[k]}, \psi_{v\theta}^{[k]})$$

    Set the behavior policy to be ϵ-greedy for all s

$$\widehat{\pi}_{pq}^{[k+1]}(a \mid s) = \begin{cases} 1 - \frac{(|\mathcal{A}|-1)\epsilon^{[k]}}{|\mathcal{A}|}, & \text{if } a = \hat{a}_{pq}^{[k+1]}(s) \\ \frac{\epsilon^{[k]}}{|\mathcal{A}|}, & \text{otherwise} \end{cases},$$
$$\widehat{\pi}_{v\theta}^{[k+1]}(a \mid s) = \begin{cases} 1 - \frac{(|\mathcal{A}|-1)\epsilon^{[k]}}{|\mathcal{A}|}, & \text{if } a = \hat{a}_{v\theta}^{[k+1]}(s) \\ \frac{\epsilon^{[k]}}{|\mathcal{A}|}, & \text{otherwise} \end{cases},$$

    where $\hat{a}_{pq}^{[k+1]}(s) = \text{argmax}_a Q^{\psi_{pq}^{[k+1]}}(s, a)$, $\hat{a}_{v\theta}^{[k+1]}(s) = \text{argmax}_a Q^{\psi_{v\theta}^{[k+1]}}(s, a)$.

  end if

  Terminate if ADMM has converged

end for

and line outages.

Two small-sized neural networks of identical structure (4 fully-connected layers with hidden dimension 256) are used to approximate Q_{pq} and $Q_{v\theta}$. The action space has dimension 10, and we choose the number of residual history points $n = 20$. This makes the input and output dimension of the neural network 40 and 10, respectively. We take the initial ρ_{pq} and ρ_{va} to be the values suggested by [143] (provided in Table 5.1). Each test instance is solved from a cold-start in ADMM.

Table 5.2: Performance of RL Policy Under Training Loads (ADMM Iterations)

	[Mhanna 2019]	RL policy	Iteration Reduction
9-bus	879	358	59.3%
30-bus	1400	738	47.3%
118-bus	525	343	34.7%

5.4.1 Performance on Training Scheme

The RL policy is trained under the default loading for 1000 RL episodes, where one episode is a complete ADMM solving process. Compared with the state-of-the-art ρ adjustment scheme in [143] that results in ADMM convergence in 879, 1400, and 525 iterations for 9-bus, 30-bus, and 118-bus systems, the RL policy reduces the number of ADMM iterations by at least 30% (see Table 5.2). To understand the mechanism behind the fast convergence under the RL policy, we show the primal and dual residuals over ADMM iterations under the RL policy and the scheme in [143] for the 9-bus system. While the scheme in [143] leads to frequent fluctuations of the residuals which prolong the ADMM solving process, the RL policy avoids these fluctuations. Although this trend is not as obvious in 30-bus and 118-bus systems, we still observe that the RL policy allows the residuals to drop more smoothly.

5.4.2 Generalization of RL Policy to Varying Loads

We also test the generalization of the RL policy to varying loads. Note that the RL policy has only been trained on the default loads from MATPOWER, not on any other loading schemes.

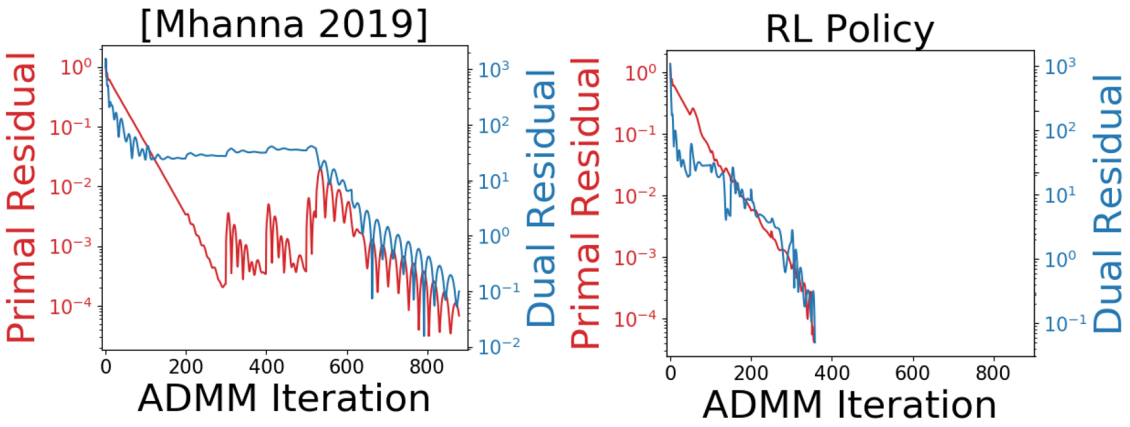


Figure 5.2: Primal and Dual Residuals under RL Policy for 9-bus System

We create a dataset of 50 test instances by randomly perturbing the default loads in the range $[-10\%, 10\%]$ at each bus. We summarize the number of ADMM iterations to convergence in Table 5.3. The RL policy reduces the ADMM iterations by 28% to 50% across test cases compared with the scheme in [143].

Table 5.3: Performance of RL Policy Under Varying Loads (ADMM Iterations)

	ρ selection method				
	[Mhanna 2019]		RL policy		
	mean	std	mean	std	Iteration Reduction
9-bus	813.4	20.4	407	9.9	50.0%
30-bus	1414.3	43.6	772.5	18.9	45.4%
118-bus	486.6	8	346	7.2	28.9%

5.4.3 Generalization of RL Policy to Generator and Line Outages

In practical situations, we may need to solve the ACOPF problem after generator and line outages. It is therefore of interest to investigate the performance of the RL policy in a modified network. In this section, we evaluate the ADMM convergence speed when applied to systems with 1) one generator removed and 2) one line disconnected.[3] Again, we note that the RL policies were trained on the original MATPOWER networks, without considering

[3]We consider all possible generator outage scenarios. Line outages are sampled in a uniformly random manner such that they do not island the network. We exclude line outages that lead to infeasible solutions under the method in [143].

line or generator losses. Table 5.4 and Table 5.5 summarize the performance of the RL policy and its comparison with the state-of-the-art method in [143].

Table 5.4: Performance of RL Policy Under Generator Outages (ADMM Iterations)

	No. of instances	ρ selection method				Iteration Reduction
		[Mhanna 2019]		RL policy		
		mean	std	mean	std	
9-bus	3	856.0	221.4	654.0	119.9	23.6%
30-bus	6	1325.8	404.3	695.8	78.9	47.5%
118-bus	54	483.8	17.7	340.0	8.8	29.7%

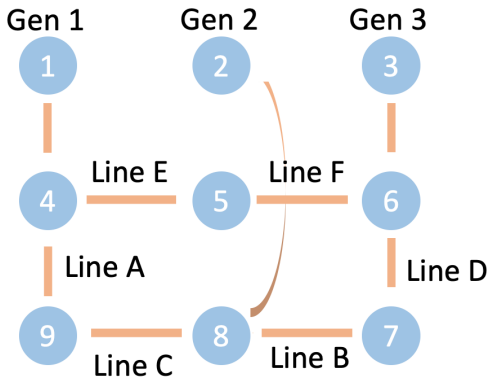
Table 5.5: Performance of RL Policy Under Line Outages (ADMM Iterations)

	No. of instances	ρ selection method				Iteration Reduction
		[Mhanna 2019]		RL policy		
		mean	std	mean	std	
9-bus	6	698.7	218.5	367.3	31.1	47.4%
30-bus	10	1455.5	225.6	800.4	93.2	45.0%
118-bus	50	486.5	6.0	346.1	6.1	28.9%

In the 9-bus system, there are three generator buses and six lines that can be disconnected while avoiding islands. In Figure 5.3, we detail the ADMM convergence under the RL policy for each outage scenario, and note that the proposed method always outperforms [143] by a large margin.

5.4.4 Generalization of RL Policy to Unseen Network Structures

We also performed experiments on the generalization of the RL policy to networks that were not seen during training. For example, one may be interested in training a RL policy for a 9-bus system and deploying it to a 30-bus system. Though our policy factorization described in Subsection 5.3.1 makes it possible to apply the RL policy to an ACOPF problem with a different number of constraints, we experimentally found that policies trained in one network perform poorly in a completely different network. This observation strengthens our belief



Outage	[Mhanna 2019]	RL Policy
Gen 1	555 iters	485 iters
Gen 2	1081 iters	727 iters
Gen 3	932 iters	750 iters
Line A	535 iters	335 iters
Line B	1178 iters	431 iters
Line C	591 iters	360 iters
Line D	604 iters	351 iters
Line E	680 iters	377 iters
Line F	604 iters	350 iters

Figure 5.3: ADMM Convergence with RL Policy for the 9-bus System with Generator and Line Outages

that there may not exist a universally optimal strategy that works for any ADMM problem, and thus supports the need for specialized approaches like the RL policy in this work.

5.5 Future Directions

We conclude by pointing out a few ways to further improve the performance of the RL trained penalty parameter selection method. First, we can improve the training of π_{pq} and $\pi_{v\theta}$ through communications. The current learning paradigm outlined in Algorithm 5.1 updates each of the two policies independently assuming that the other policy is fixed to its previous iterate. As we discussed in Subsection 5.3.1, the two policies can be regarded as two agents that have aligned interest but would like to achieve their interest without explicit cooperation. In multi-agent RL, it is known that in general independent learning does not lead to the optimal policies and communication between the agents may be required. Properly designing the communication between the agents as well as the overall algorithm under the communication is a natural next step of this work.

Another interesting direction is to use information beyond what is available at the local constraint. To factorize the policy according to Equation 5.7, we made the assumption that the optimal choice of ρ_i can be determined solely from residuals available at constraint i, which likely does not hold. The fact that local information is insufficient is reflected in our

observation that the advantage of the RL policy over [143] shrinks as the size of the power network scales up (since the local residuals make up a smaller percentage of the overall information as the network becomes larger). A more reasonable assumption may be that we can determine ρ_i from residuals at constraint i and its one-hop neighboring constraints (constraints that share at least one variable), which leads to a policy factorization of the form

$$\pi^\star(a \mid s) = \left(\prod_{i \in \mathcal{C}_{pq}} \pi^\star_{pq} \left(a_i \mid \{s_j\}_{j \in \mathcal{N}_i} \right) \right) \left(\prod_{i \in \mathcal{C}_{v\theta}} \pi^\star_{v\theta} \left(a_i \mid \{s_j\}_{j \in \mathcal{N}_i} \right) \right),$$

where $\mathcal{N}_i$ denotes the collection of one-hop neighbors of constraint i. A significant challenge of this approach lies in modelling the variable input dimension of the policy as the cardinality of $\mathcal{N}_i$ can be different across i.

CHAPTER 6
CONCLUSION

In this dissertation, we presented a collections of results on single-agent and multi-agent RL from both theory and application perspectives. To summarize, in the first aim, we recognized that a range of data-driven algorithms in RL can be regarded as using two-time-scale stochastic gradient descent to solve a optimization problem with a special type of gradient oracle. We proposed a mathematical algorithmic framework that unifies these algorithms and present the convergence rates of the algorithm for strongly convex, nonconvex, and PŁ objective functions.

In the second aim, we considered multi-agent multi-task RL in the average cumulative reward formulation. We discussed two properties of this multi-task RL problem which make it significantly harder to solve than its single-task counterpart, followed by the introduction and analysis of a decentralized policy gradient algorithm that converges in local and global senses under different assumptions. We then shifted focus to a constrained multi-task RL formulation which allows for the specification of the performance of the policy in each task. We presented a decentralized primal-dual algorithm that provably converges the globally optimally policy, both in objective function value and in constraint violation.

In the third aim, we studied using GDA to find the Nash equilibrium of the two-player zero-sum Markov game, which is notoriously hard to solve with direct optimization methods due to its nonconvex-nonconcave objective function. To bypass the issue and introduce stronger structure into the problem, we regularize the reward function by the policy entropy. The regularized value function exhibits a property that resembles the PŁ condition, which guarantees that GDA converges linearly fast to the Nash equilibrium of the regularized objective. We then designed methods to properly reduce the regularization weight that allows GDA to efficiently converge to the Nash equilibrium of the original unmodified

Markov game.

Finally, in the third aim we applied RL to design a penalty parameter selection policy with the aim of improving the convergence of the ADMM algorithm applied to a power system optimization problem. We showed that the RL policy significantly accelerates the ADMM convergence compared with the state-of-the-art human designed penalty parameter adjustment scheme. In addition, the RL policy exhibits strong promise for generalizability, performing well under unseen loading schemes as well as under unseen line and generator outages.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR RESULTS IN CHAPTER 2

A.1 Analysis Decomposition and Proof of Main Theorem

In this section, we briefly explain the main technical challenge in analyzing Algorithm 2.1, which is the coupling between θ, ω, and the time-varying Markovian samples. Our approach to the challenge is to properly "decouple" the variable updates so that we can handle them individually. Specifically, we first show under time-varying Markovian samples the convergence of the decision variable up to an error in the auxiliary variable (Subsection A.1.1) and the reduction of the auxiliary variable error which hinges on the decision variable convergence (Subsection A.1.2), which essentially form a coupled dynamical system. In Subsection A.1.3, we introduce an important lemma that performs Lyapunov analysis on a coupled dynamical system of two inequalities. This lemma is a unified tool to analyze our algorithm under different function structures and may be of independent interest in the study of the finite-time performance of multiple-time-scale dynamical systems apart from those considered in this paper. Finally, we prove the theorem under the PŁ condition in Subsection A.1.4. Strongly convex and general non-convex functions can be treated with similar analytical techniques, and the full details of their analyses can be found in [34].

When f observes the PŁ condition, we show $\|f(\theta_k) - f^\star\|^2 \to 0$. We frequently employ a few quantities for which we introduce the following shorthand notations

$$\begin{aligned} z_k &\triangleq \omega_k - \omega^\star(\theta_k), \\ \overline{\Delta H_k} &\triangleq H(\theta_k, \omega_k, X_k) - H(\theta_k, \omega^\star(\theta_k), X_k), \\ \Delta H_k &\triangleq H(\theta_k, \omega^\star(\theta_k), X_k) - \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_k, \omega^\star(\theta_k), \hat{X})], \\ \Delta G_k &\triangleq G(\theta_k, \omega_k, X_k) - \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[G(\theta_k, \omega_k, \hat{X})]. \end{aligned} \tag{A.1}$$

We can think of $\overline{\Delta H_k}$ as the bias in the stochastic gradient due to the inaccurate auxiliary variable and ΔH_k and ΔG_k as the errors that the Markovian samples cause to H and G.

A.1.1 Decision Variable Convergence

We derive a recursive formula for the iteration-wise decision variable convergence measured in $\mathbb{E}[f(\theta_k) - f^\star]$. As a first step, we have from the update rule Equation 2.8 and the L-smoothness of f

$$\begin{aligned}
f(\theta_{k+1}) &\leqslant f(\theta_k) + \langle \nabla f(\theta_k), \theta_{k+1} - \theta_k \rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= f(\theta_k) - \alpha_k \langle \nabla f(\theta_k), H(\theta_k, \omega_k, X_k) \rangle + \frac{L\alpha_k^2}{2}\|H(\theta_k, \omega_k, X_k)\|^2 \\
&= f(\theta_k) - \alpha_k \langle \nabla f(\theta_k), \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_k, \omega^\star(\theta_k), \hat{X})] \rangle \\
&\quad - \alpha_k \left\langle \nabla f(\theta_k), \Delta H_k + \overline{\Delta H_k} \right\rangle + \frac{L\alpha_k^2}{2}\|H(\theta_k, \omega_k, X_k)\|^2 \\
&= f(\theta_k) - \alpha_k \|\nabla f(\theta_k)\|^2 - \alpha_k \left\langle \nabla f(\theta_k), \Delta H_k + \overline{\Delta H_k} \right\rangle + \frac{L\alpha_k^2}{2}\|H(\theta_k, \omega_k, X_k)\|^2, \quad \text{(A.2)}
\end{aligned}$$

where the last equality follows from Equation 2.2, i.e. $\nabla f(\theta_k) = \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_k, \omega^\star(\theta_k), \hat{X})]$. A key challenge to overcome is the time-varying Markovian randomness. If the samples were i.i.d. and the auxiliary variables were always solved perfectly, we would have $\mathbb{E}[\Delta H_k] = \mathbb{E}[\overline{\Delta H_k}] = 0$, reducing the problem to the one studied in the standard SGD. In the following lemma, we carefully treat the Markovian noise by leveraging the uniform geometric mixing time of the time-varying Markov chain and the Lipschitz condition of the state transition kernel.

Lemma A.1. *For any $k \geqslant \tau_k$, we have*

$$\mathbb{E}\left[-\langle \nabla f(\theta_k), \Delta H_k \rangle\right] \leqslant 12L^2B^3\tau_k^2\alpha_{k-\tau_k}.$$

We can use the Lipschitz continuity of H to study the error caused by $\overline{\Delta H_k}$ and show that it can be bounded by the sum of $\|\theta_k - \theta^\star\|^2$ and $\|z_k\|^2$. The bound on $\overline{\Delta H_k}$ together with

the result established in Lemma A.1 leads to the following proposition, which states that $\mathbb{E}[f(\theta_k) - f^\star]$ is sufficiently reduced in every iteration if the auxiliary variable error z_k is controlled.

Proposition A.1. *Under Assumption 2.1-Assumption 2.6, we have for all $k \geqslant \mathcal{K}$*

$$\mathbb{E}[f(\theta_{k+1}) - f^\star] \leqslant (1-\lambda\alpha_k)\mathbb{E}[f(\theta_k) - f^\star] + \frac{L^2\alpha_k}{2}\mathbb{E}\left[\|z_k\|^2\right] + \frac{25L^2B^3}{2}\tau_k^2\alpha_k\alpha_{k-\tau_k}.$$

Proof. By the Lipschitz condition of the operator H,

$$\begin{aligned}
-\mathbb{E}[\langle\nabla f(\theta_k), \overline{\Delta H_k}\rangle] &\leqslant \frac{1}{2}\mathbb{E}[\|\nabla f(\theta_k)\|^2 + \|H(\theta_k, \omega_k, X_k) - H(\theta_k, \omega^\star(\theta_k), X_k)\|^2] \\
&\leqslant \frac{1}{2}\mathbb{E}[\|\nabla f(\theta_k)\|^2] + \frac{L^2}{2}\mathbb{E}\left[\|z_k\|^2\right].
\end{aligned}$$

Using this inequality along with Lemma A.1 in Equation A.2, we have for all $k \geqslant \tau_k$

$$\begin{aligned}
\mathbb{E}[f(\theta_{k+1})] &\leqslant \mathbb{E}[f(\theta_k)] - \alpha_k\langle\nabla f(\theta_k), \Delta H_k\rangle - \alpha_k\mathbb{E}[\|\nabla f(\theta_k)\|^2] \\
&\quad - \alpha_k\mathbb{E}\left[\left\langle\nabla f(\theta_k), \overline{\Delta H_k}\right\rangle\right] + \frac{LB^2\alpha_k^2}{2} \\
&\leqslant \mathbb{E}[f(\theta_k)] + 12L^2B^3\tau_k^2\alpha_k\alpha_{k-\tau_k} - \alpha_k\mathbb{E}[\|\nabla f(\theta_k)\|^2] \\
&\quad + \frac{\alpha_k}{2}\mathbb{E}[\|\nabla f(\theta_k)\|^2] + \frac{L^2\alpha_k}{2}\mathbb{E}\left[\|z_k\|^2\right] + \frac{LB^2\alpha_k^2}{2} \\
&\leqslant \mathbb{E}[f(\theta_k)] - \frac{\alpha_k}{2}\mathbb{E}[\|\nabla f(\theta_k)\|^2] + \frac{L^2\alpha_k}{2}\mathbb{E}\left[\|z_k\|^2\right] + \frac{25L^2B^3}{2}\tau_k^2\alpha_k\alpha_{k-\tau_k} \\
&\leqslant \mathbb{E}[f(\theta_k)] - \lambda\alpha_k\mathbb{E}\left[f(\theta_k) - f^\star\right] + \frac{L^2\alpha_k}{2}\mathbb{E}\left[\|z_k\|^2\right] + \frac{25L^2B^3}{2}\tau_k^2\alpha_k\alpha_{k-\tau_k},
\end{aligned}$$

where the last inequality is due to the PŁ condition. Subtracting $f^\star$ from both sides of the inequality leads to the claimed result.

□

A.1.2 Auxiliary Variable Convergence

In this section, we present and analyze the convergence of the auxiliary variable, summarized in the proposition below.

Proposition A.2. *Under Assumption 2.1-Assumption 2.6, we have for all $k \geqslant \mathcal{K}$*

$$\mathbb{E}[\|z_{k+1}\|^2] \leqslant (1-\frac{\lambda\beta_k}{2})\mathbb{E}[\|z_k\|^2]+C_2\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|\theta_k\|^2\right]+\frac{2L^2B^2\alpha_k^2}{\lambda\beta_k}+C_2\tau_k^2\beta_{k-\tau_k}\beta_k.$$

Recall the auxiliary variable error z_k defined in Equation A.1. Proposition A.2 establishes an iteration-wise reduction of this error in expectation in face of the drift of θ_k. To prove this proposition, we introduce Lemma A.2 that bounds the error in the auxiliary variable caused by the Markovian samples. We skip the proof of this lemma due to its similarity to Lemma A.1 and refer interested readers to [34] for the full proof details.

Lemma A.2. *Recall the definition of C_1 in Equation 2.27. For any $k \geqslant \tau_k$, we have*

$$\mathbb{E}[\langle z_k, \Delta G_k\rangle] \leqslant C_1\tau_k^2\beta_{k-\tau_k}\mathbb{E}\left[\|z_{k-\tau_k}\|^2+\|\theta_k\|^2+\|\omega_k\|^2+1\right].$$

Analyzing Proposition A.2 requires properly controlling $\|\omega_k-\omega_{k-\tau_k}\|$, which we handle in the following lemma.

Lemma A.3. *For all $k \geqslant \tau_k$, we have*

$$\|\omega_k-\omega_{k-\tau_k}\| \leqslant 3D\beta_{k-\tau_k}\tau_k\left(\|\omega_k\|+\|\theta_k\|+1\right).$$

Proof. Recall that $z_k = \omega_k - \omega^\star(\theta_k)$. We have from Equation 2.9

$$\begin{aligned}
\|z_{k+1}\|^2 &= \|\omega_k+\beta_k G(\theta_{k+1},\omega_k,X_k)-\omega^\star(\theta_{k+1})\|^2\\
&= \|\left(\omega_k-\omega^\star(\theta_k)\right)+\beta_k G(\theta_{k+1},\omega_k,X_k)+\left(\omega^\star(\theta_k)-\omega^\star(\theta_{k+1})\right)\|^2\\
&\leqslant \|z_k\|^2+2\beta_k\langle z_k, G(\theta_{k+1},\omega_k,X_k)\rangle+2\langle z_k,\omega^\star(\theta_k)-\omega^\star(\theta_{k+1})\rangle
\end{aligned}$$

$$+ 2\beta_k^2 \left\|G(\theta_{k+1}, \omega_k, X_k)\right\|^2 + 2\|\omega^\star(\theta^k) - \omega^\star(\theta^{k+1})\|^2.$$

From the definition of ΔG_k in Equation A.1,

$$\begin{aligned}
\|z_{k+1}\|^2 &\leqslant \|z_k\|^2 + 2\beta_k\langle z_k, G(\theta_{k+1}, \omega_k, X_k) - G(\theta_k, \omega_k, X_k)\rangle + 2\beta_k\langle z_k, G(\theta_k, \omega_k, X_k)\rangle \\
&\quad + 2\langle z_k, \omega^\star(\theta_k) - \omega^\star(\theta_{k+1})\rangle + 2\beta_k^2 \left\|G(\theta_{k+1}, \omega_k, X_k)\right\|^2 + 2\|\omega^\star(\theta_k) - \omega^\star(\theta_{k+1})\|^2 \\
&\leqslant \|z_k\|^2 + 2\beta_k\langle z_k, G(\theta_{k+1}, \omega_k, X_k) - G(\theta_k, \omega_k, X_k)\rangle\rangle \\
&\quad + 2\beta_k\langle z_k, \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[G(\theta_k, \omega_k, \hat{X})]\rangle + 2\beta_k\langle z_k, \Delta G_k\rangle + 2\langle z_k, \omega^\star(\theta_k) - \omega^\star(\theta_{k+1})\rangle \\
&\quad + 2\beta_k^2 \left\|G(\theta_{k+1}, \omega_k, X_k)\right\|^2 + 2L^2B^2\alpha_k^2, \qquad \text{(A.3)}
\end{aligned}$$

where the second inequality applies Assumption 2.1 and Equation 2.8, i.e.,

$$\|\omega^\star(\theta_k) - \omega^\star(\theta_{k+1})\|^2 \leqslant L^2\|\theta_k - \theta_{k+1}\|^2 = L^2\|\alpha_k H(\theta_k, \omega_k, X_k)\|^2 \leqslant L^2B^2\alpha_k^2. \qquad \text{(A.4)}$$

We next analyze each term on the right-hand side of Equation A.3. First, using the relation $\langle 2v_1, v_2\rangle \leqslant c\|v_1\|^2 + \frac{1}{c}\|v_2\|^2$ for any vectors v_1, v_2 and scalar $c > 0$, we bound the second term of Equation A.3

$$\begin{aligned}
&\langle z_k, G(\theta_{k+1}, \omega_k, X_k) - G(\theta_k, \omega_k, X_k)\rangle \leqslant \frac{\lambda}{4}\|z_k\|^2 + \frac{1}{\lambda}\|G(\theta_{k+1}, \omega_k, X_k) - G(\theta_k, \omega_k, X_k)\|^2 \\
&\leqslant \frac{\lambda}{4}\|z_k\|^2 + \frac{L^2}{\lambda}\|\theta_{k+1} - \theta_k\|^2 \leqslant \frac{\lambda}{4}\|z_k\|^2 + \frac{L^2B^2\alpha_k^2}{\lambda}, \qquad \text{(A.5)}
\end{aligned}$$

where the second inequality follows from the Lipschitz continuity of G and the last inequality is due to Equation A.4. Similarly, we consider the fifth term of Equation A.3

$$\begin{aligned}
&2\langle z_k, \omega^\star(\theta_k) - \omega^\star(\theta_{k+1})\rangle \leqslant \frac{\beta_k\lambda}{2}\|z_k\|^2 + \frac{2}{\lambda\beta_k}\|\omega^\star(\theta_k) - \omega^\star(\theta_{k+1})\|^2 \\
&\leqslant \frac{\beta_k\lambda}{2}\|z_k\|^2 + \frac{2L^2}{\lambda\beta_k}\|\theta_k - \theta_{k+1}\|^2 \leqslant \frac{\beta_k\lambda}{2}\|z_k\|^2 + \frac{2L^2B^2\alpha_k^2}{\lambda\beta_k}. \qquad \text{(A.6)}
\end{aligned}$$

Next, using Assumption 2.3 and $z_k = \omega_k - \omega^\star(\theta_k)$ we treat the third term of Equation A.3

$$2\beta_k \langle z_k, \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[G(\theta_k, \omega_k, \hat{X})]\rangle \leqslant -2\lambda\beta_k \|z_k\|^2. \tag{A.7}$$

By Equation 2.24 and Equation 2.8 we have

$$\|G(\theta_{k+1}, \omega_k, X_k)\|^2 \leqslant 2D^2\big(\|\theta_{k+1}\| + \|\omega_k\| + 1\big)^2 \leqslant 2D^2\big(\|\theta_k\| + B\alpha_k + \|\omega_k\| + 1\big)^2. \tag{A.8}$$

Taking the expectation on both sides of Equation A.3 and using Equation A.5–Equation A.8 and Lemma A.2

$$\begin{aligned}
\mathbb{E}[\|z_{k+1}\|^2] &\leqslant \mathbb{E}[\|z_k\|^2] + \frac{\beta_k\lambda}{2}\mathbb{E}[\|z_k\|^2] + \frac{2L^2B^2\beta_k\alpha_k^2}{\lambda} + \frac{\beta_k\lambda}{2}\mathbb{E}[\|z_k\|^2] + \frac{2L^2B^2\alpha_k^2}{\lambda\beta_k} \\
&\quad - 2\lambda\beta_k\mathbb{E}[\|z_k\|^2] + C_1\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|z_{k-\tau_k}\|^2 + \|\theta_k\|^2 + \|\omega_k\|^2 + 1\right] \\
&\quad + 2D^2\beta_k^2\big(\|\theta_k\| + B\alpha_k + \|\omega_k\| + 1\big)^2 + 2L^2B^2\alpha_k^2 \\
&\leqslant (1 - \lambda\beta_k)\mathbb{E}[\|z_k\|^2] + C_1\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|z_{k-\tau_k}\|^2\right] \\
&\quad + (C_1 + 8D^2)\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|\theta_k\|^2 + \|\omega_k\|^2\right] \\
&\quad + \frac{2L^2B^2\alpha_k^2}{\lambda\beta_k} + (C_1 + 32D^2 + \frac{2L^2B^2}{\lambda} + 2L^2B^2)\tau_k^2\beta_{k-\tau_k}\beta_k,
\end{aligned} \tag{A.9}$$

where the last inequality uses $\alpha_k \leqslant \beta_k$ and $B\alpha_k \leqslant 1$. Note that $\|z_{k-\tau_k}\|^2$ obeys

$$\begin{aligned}
\|z_{k-\tau_k}\|^2 &= \|z_k - (\omega_k - \omega_{k-\tau_k}) + (\omega^\star(\theta_k) - \omega^\star(\theta_{k-\tau_k}))\|^2 \\
&\leqslant 3\left(\|z_k\|^2 + \|\omega_k - \omega_{k-\tau_k}\|^2 + \|\omega^\star(\theta_k) - \omega^\star(\theta_{k-\tau_k})\|^2\right) \\
&\leqslant 3\|z_k\|^2 + \frac{9}{4}\left(\|\omega_k\|^2 + \|\theta_k\|^2 + 1\right) + L^2\|\theta_k - \theta_{k-\tau_k}\|^2 \\
&\leqslant 3\left(\|z_k\|^2 + \|\omega_k\|^2 + \|\theta_k\|^2\right) + L^2B^2\tau_k^2\alpha_{k-\tau_k}^2 + \frac{9}{4} \\
&\leqslant 3\left(\|z_k\|^2 + \|\omega_k\|^2 + \|\theta_k\|^2 + 1\right),
\end{aligned}$$

where the second inequality is due to Lemma A.3 and the Lipschitz continuity of $\omega^\star$, and

the last inequality follows from the step size condition $LB\tau_k\alpha_{k-\tau_k} \leqslant \frac{1}{6}$. Substituting the preceding relation into Equation A.9, we have for all $k \geqslant \tau_k$

$$\begin{aligned}
\mathbb{E}[\|z_{k+1}\|^2] &\leqslant (1-\lambda\beta_k)\mathbb{E}[\|z_k\|^2] + 3C_1\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|z_k\|^2\right] \\
&\quad + (4C_1+8D^2)\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|\theta_k\|^2+\|\omega_k\|^2\right] \\
&\quad + \frac{2L^2B^2\alpha_k^2}{\lambda\beta_k} + (4C_1+32D^2+\frac{2L^2B^2}{\lambda}+2L^2B^2)\tau_k^2\beta_{k-\tau_k}\beta_k \\
&\leqslant (1-\lambda\beta_k)\mathbb{E}[\|z_k\|^2] + (11C_1+16D^2)\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|z_k\|^2\right] \\
&\quad + (4D^2+1)(4C_1+8D^2)\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|\theta_k\|^2\right] + \frac{2L^2B^2\alpha_k^2}{\lambda\beta_k} \\
&\quad + ((4D^2+1)(4C_1+32D^2) + \frac{2L^2B^2}{\lambda}+2L^2B^2)\tau_k^2\beta_{k-\tau_k}\beta_k, \qquad \text{(A.10)}
\end{aligned}$$

where in the last inequality we use Equation 2.24 to derive

$$\|\omega_k\|^2 \leqslant 2\|\omega_k-\omega^\star(\theta_k)\|^2+2\|\omega^\star(\theta_k)\|^2 \leqslant 2\|z_k\|^2+2D^2(\|\theta_k\|+1)^2 \leqslant 2\|z_k\|^2+4D^2(\|\theta_k\|^2+1).$$

By the choice of the step size we have $(11C_1+16D^2)\tau_k^2\beta_{k-\tau_k} \leqslant \frac{\lambda}{2}$. Thus, using the constant C_2 defined in Equation 2.27, we know that Equation A.10 implies

$$\mathbb{E}[\|z_{k+1}\|^2] \leqslant (1-\frac{\lambda\beta_k}{2})\mathbb{E}[\|z_k\|^2] + C_2\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|\theta_k\|^2+1\right] + \frac{2L^2B^2\alpha_k^2}{\lambda\beta_k}.$$

□

Propositions A.1 and A.2 show that the convergence of the decision variable and the auxiliary variable forms a coupled dynamical system that evolves under two different rates. In the next section, we introduce a two-time-scale lemma that solves the system.

A.1.3 Two-Time-Scale Lemma

Although we analyze the performance of our algorithm for different types of objective functions and with different convergence metrics, these analyses eventually reduce to the

study of two coupled inequalities. The dynamics of these two inequalities happen on different time scales determined by the two step sizes used in our algorithm. In this section we present a general result, which we call the two-time-scale lemma, that characterizes the behavior of these coupled inequalities.

Lemma A.4. *Let* $\{a_k,\, b_k,\, c_k,\, d_k,\, e_k,\, f_k\}$ *be non-negative sequences satisfying* $\frac{a_{k+1}}{d_{k+1}} \leqslant \frac{a_k}{d_k} < 1$*, for all* $k \geqslant 0$*. Let* $\{x_k\}, \{y_k\}$ *be two non-negative sequences. We consider two settings on their dynamics.*

1. *Suppose that* x_k, y_k *satisfy the following coupled inequalities*

$$x_{k+1} \leqslant (1 - a_k)x_k + b_k y_k + c_k, \quad y_{k+1} \leqslant (1 - d_k)y_k + e_k x_k + f_k. \tag{A.11}$$

In addition, assume that there exists a constant $A \in \mathbb{R}$ *such that*

$$Aa_k - b_k - \frac{Aa_k^2}{d_k} \geqslant 0 \quad \textit{and} \quad \frac{Ae_k}{d_k} \leqslant \frac{1}{2}, \quad \textit{for all } k \geqslant 0. \tag{A.12}$$

Then we have for all $0 \leqslant \tau \leqslant k$

$$x_k \leqslant (x_\tau + \frac{Aa_\tau}{d_\tau} y_\tau) \prod_{t=\tau}^{k-1} (1 - \frac{a_t}{2}) + \sum_{\ell=\tau}^{k-1} \Big(c_\ell + \frac{Aa_\ell f_\ell}{d_\ell} \Big) \prod_{t=\ell+1}^{k-1} (1 - \frac{a_t}{2}).$$

2. *Suppose that* $\{x_k, y_k\}$ *satisfy the following coupled inequalities*

$$x_{k+1} \leqslant (1 + a_k)x_k + b_k y_k + c_k, \quad y_{k+1} \leqslant (1 - d_k)y_k + e_k x_k + f_k. \tag{A.13}$$

$\{u_k\}$ *is a non-negative sequence such that*

$$u_k \leqslant (1 + a_k)x_k - x_{k+1} + b_k y_k + c_k, \tag{A.14}$$

then we have for any $0 \leqslant \tau \leqslant k$

$$\sum_{t=\tau}^{k} u_t \leqslant \Big(1 + \sum_{t=\tau}^{k} (a_t + \frac{b_t e_t}{d_t}) e^{\sum_{t=\tau}^{k} (a_t + \frac{b_t e_t}{d_t})}\Big)\Big(x_\tau + \frac{b_\tau y_\tau}{d_\tau} + \sum_{t=\tau}^{k} (c_t + \frac{b_t f_t}{d_t})\Big).$$

Proof. Case 1) Consider $V_k = x_k + \frac{Aa_k}{d_k} y_k$. From the second equation in Equation A.11,

$$\begin{aligned}\frac{Aa_{k+1}}{d_{k+1}} y_{k+1} &\leqslant \frac{Aa_k}{d_k} y_{k+1} \leqslant \frac{Aa_k}{d_k}\left((1-d_k) y_k + e_k x_k + f_k\right) \\ &= (1-a_k)\frac{Aa_k}{d_k} y_k + (a_k - d_k)\frac{Aa_k}{d_k} y_k + \frac{Aa_k e_k x_k}{d_k} + \frac{Aa_k f_k}{d_k}.\end{aligned}$$

Combining this with the first inequality of Equation A.11 yields

$$\begin{aligned}V_{k+1} &= x_{k+1} + \frac{Aa_{k+1}}{d_{k+1}} y_{k+1} \\ &\leqslant (1-a_k) x_k + b_k y_k + c_k + (1-a_k)\frac{Aa_k}{d_k} y_k + (a_k - d_k)\frac{Aa_k}{d_k} y_k + \frac{Aa_k e_k x_k}{d_k} + \frac{Aa_k f_k}{d_k} \\ &= (1-a_k)\left(x_k + \frac{Aa_k y_k}{d_k}\right) + \left(\frac{Aa_k^2}{d_k} - Aa_k + b_k\right) y_k + c_k + \frac{Aa_k e_k x_k}{d_k} + \frac{Aa_k f_k}{d_k} \\ &\leqslant (1-a_k) V_k + \frac{a_k}{2} x_k + c_k + \frac{Aa_k f_k}{d_k} \leqslant (1 - \frac{a_k}{2}) V_k + c_k + \frac{Aa_k f_k}{d_k},\end{aligned}$$

where the second inequality follows from Equation A.12. Applying this relation recursively,

$$\begin{aligned}x_k \leqslant V_k &\leqslant V_\tau \prod_{t=\tau}^{k-1} (1 - \frac{a_t}{2}) + \sum_{\ell=\tau}^{k-1} \left(c_\ell + \frac{Aa_\ell f_\ell}{d_\ell}\right) \prod_{t=\ell+1}^{k-1} (1 - \frac{a_t}{2}) \\ &\leqslant (x_\tau + \frac{Aa_\tau}{d_\tau} y_\tau) \prod_{t=\tau}^{k-1} (1 - \frac{a_t}{2}) + \sum_{\ell=\tau}^{k-1} \left(c_\ell + \frac{Aa_\ell f_\ell}{d_\ell}\right) \prod_{t=\ell+1}^{k-1} (1 - \frac{a_t}{2}).\end{aligned}$$

Case 2) Re-arranging the second inequality of Equation A.13 and multiplying by $\frac{b_k}{d_k}$,

$$b_k y_k \leqslant \frac{b_k}{d_k} y_k - \frac{b_k}{d_k} y_{k+1} + \frac{b_k e_k x_k}{d_k} + \frac{b_k f_k}{d_k}.$$

Plugging this inequality into the first inequality of Equation A.13 yields

$$\begin{aligned} x_{k+1} &\leqslant (1+a_k)x_k + c_k + \frac{b_k}{d_k}y_k - \frac{b_k}{d_k}y_{k+1} + \frac{b_k e_k x_k}{d_k} + \frac{b_k f_k}{d_k} \\ &\leqslant (1+g_k)x_k + \frac{b_k}{d_k}y_k - \frac{b_k}{d_k}y_{k+1} + c_k + \frac{b_k f_k}{d_k}, \end{aligned} \tag{A.15}$$

where we define $g_k = a_k + \frac{b_k e_k}{d_k}$. Since $1 + c \leqslant \exp(c)$ for any scalar $c > 0$, we have

$$\begin{aligned} x_{k+1} &\leqslant \exp(g_k)x_k + \frac{b_k}{d_k}y_k - \frac{b_k}{d_k}y_{k+1} + c_k + \frac{b_k f_k}{d_k} \\ &\leqslant \exp(\sum_{t=\tau}^{k} g_t)x_\tau + \exp\big(\sum_{t=\tau}^{k} g_t\big)\sum_{t=\tau}^{k}\left(\frac{b_t}{d_t}(y_t - y_{t+1}) + c_t + \frac{b_t f_t}{d_t}\right) \\ &\leqslant \exp(\sum_{t=\tau}^{k} g_t)\Big(x_\tau + \frac{b_\tau y_\tau}{d_\tau} + \sum_{t=\tau}^{k}(c_t + \frac{b_t f_t}{d_t})\Big), \end{aligned} \tag{A.16}$$

where the second inequality applies the first inequality recursively. The inequalities Equation A.15, Equation A.16, and Equation A.14 together imply

$$\begin{aligned} \sum_{t=\tau}^{k} u_t &\leqslant \sum_{t=\tau}^{k}(x_t - x_{t+1}) + \big(\max_{\tau \leqslant t \leqslant k} x_t\big)\sum_{t=\tau}^{k} g_t + \sum_{t=\tau}^{k}\left(\frac{b_t}{d_t}(y_t - y_{t+1}) + c_t + \frac{b_t f_t}{d_t}\right) \\ &\leqslant x_\tau + \sum_{t=\tau}^{k} g_t \exp\big(\sum_{t=\tau}^{k} g_t\big)\Big(x_\tau + \frac{b_\tau y_\tau}{d_\tau} + \sum_{t=\tau}^{k}(c_t + \frac{b_t f_t}{d_t})\Big) + \frac{b_\tau}{d_\tau}y_\tau + \sum_{t=\tau}^{k}(c_t + \frac{b_t f_t}{d_t}) \\ &= \left(1 + \sum_{t=\tau}^{k}(a_t + \frac{b_t e_t}{d_t})\exp(\sum_{t=\tau}^{k}(a_t + \frac{b_t e_t}{d_t}))\right)\left(x_\tau + \frac{b_\tau y_\tau}{d_\tau} + \sum_{t=\tau}^{k}(c_t + \frac{b_t f_t}{d_t})\right). \end{aligned}$$

□

Lemma A.4 studies the behavior of the two interacting sequences $\{x_k\}$ and $\{y_k\}$ that have generic structure. In our analysis, properly selected convergence metrics on θ_k and ω_k evolve as x_k and y_k above, respectively, according to Equation A.11 for strongly convex and PŁ functions and Equation A.13 for non-convex functions, while the sequences $\{a_k,\, b_k,\, c_k,\, d_k,\, e_k,\, f_k\}$ are ratios and products of the step sizes $\{\alpha_k\}$ and $\{\beta_k\}$.

A.1.4 Proof of Main Results

In this section, we present the proof of Theorem 2.2 which considers functions observing the PŁ condition. The analyses of strongly convex and general non-convex functions use similar techniques: one needs to properly select a convergence metric according to the function structure, set up a step-wise decay of the convergence metric like Proposition A.1 which forms a coupled dynamical system with Proposition A.2, and apply the two-time-scale lemma introduced in Subsection A.1.3 to the coupled system.

From the analysis of the auxiliary variable in Proposition A.2, we have for all $k \geqslant \mathcal{K}$

$$\mathbb{E}[\|z_{k+1}\|^2] \leqslant (1-\frac{\lambda\beta_k}{2})\mathbb{E}[\|z_k\|^2]+C_2\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|\theta_k\|^2\right]+\frac{2L^2B^2\alpha_k^2}{\lambda\beta_k}+C_2\tau_k^2\beta_{k-\tau_k}\beta_k.$$

Due to the boundedness of the operator H,

$$\|\theta_k\| \leqslant \|\theta_0\| + \sum_{t=0}^{k-1}\|\theta_{t+1}-\theta_t\| \leqslant \|\theta_0\| + \sum_{t=0}^{k-1}\frac{B\alpha}{t+1} \leqslant \|\theta_0\| + \frac{B\alpha\log(k+1)}{\log(2)},$$

where the last inequality follows from $\sum_{t=0}^{t'}\frac{1}{(t+1)^u} \leqslant \frac{\log(t'+2)}{\log(2)}$ for any $t' \geqslant 0$. This relation implies for any $k \geqslant 0$

$$\|\theta_k\|^2 \leqslant 2\|\theta_0\|^2 + \frac{2B^2\alpha^2\log^2(k+1)}{\log^2(2)} \leqslant 24(\|\theta_0\|^2+B^2\alpha^2)\log^2(k+1).$$

Using this inequality in the bound on $\mathbb{E}[\|z_{k+1}\|^2]$, we have

$$\begin{aligned}\mathbb{E}[\|z_{k+1}\|^2] &\leqslant (1-\frac{\lambda\beta_k}{2})\mathbb{E}[\|z_k\|^2]+C_2\tau_k^2\beta_{k-\tau_k}\beta_k\mathbb{E}\left[\|\theta_k\|^2\right]+\frac{2L^2B^2\alpha_k^2}{\lambda\beta_k}+C_2\tau_k^2\beta_{k-\tau_k}\beta_k \\ &\leqslant (1-\frac{\lambda\beta_k}{2})\mathbb{E}[\|z_k\|^2]+24C_2(\|\theta_0\|^2+B^2\alpha^2+1)\tau_k^2\beta_{k-\tau_k}\beta_k\log^2(k+1)+\frac{2L^2B^2\alpha_k^2}{\lambda\beta_k}.\end{aligned}$$

We can apply Lemma A.4 case 1) to the result of Proposition A.1 and the inequality

above with $\tau = \mathcal{K}$ and

$$x_k = \mathbb{E}[f(\theta_k) - f^\star],\ y_k = \mathbb{E}\left[\|z_k\|^2\right],\ a_k = \lambda\alpha_k,\ b_k = \frac{L^2\alpha_k}{2},\ c_k = \frac{25L^2B^3}{2}\tau_k^2\alpha_k\alpha_{k-\tau_k},$$
$$d_k = \frac{\lambda\beta_k}{2},\ e_k = 0,\ f_k = 24C_2(\|\theta_0\|^2 + B^2\alpha^2 + 1)\tau_k^2\beta_{k-\tau_k}\beta_k\log^2(k+1) + \frac{2L^2B^2\alpha_k^2}{\lambda\beta_k}.$$

In this case, one can verify that we can choose $A = \frac{L^2}{\lambda}$ if the step size sequences satisfy $\frac{\alpha_k}{\beta_k} \leqslant \frac{1}{4}$. As a result of Lemma A.4 case 1), we have for all $k \geqslant \mathcal{K}$

$$\mathbb{E}\left[f(\theta_k) - f^\star\right] \leqslant \left(\mathbb{E}\left[f(\theta_\mathcal{K}) - f^\star\right] + \frac{2L^2\alpha_\mathcal{K}}{\lambda\beta_\mathcal{K}}\mathbb{E}[\|z_\mathcal{K}\|^2]\right)\prod_{t=\mathcal{K}}^{k-1}(1 - \frac{\lambda\alpha_t}{2}) + \sum_{\ell=\mathcal{K}}^{k-1}\prod_{t=\ell+1}^{k-1}(1 - \frac{\lambda\alpha_t}{2})$$
$$\times \left(\frac{25L^2B^3}{2}\tau_k^2\alpha_\ell\alpha_{\ell-\tau_k} + \frac{48C_2L^2}{\lambda}(\|\theta_0\|^2 + B^2\alpha^2 + 1)\tau_k^2\beta_{k-\tau_k}\alpha_k\log^2(k+1) + \frac{4L^4B^2\alpha_k^3}{\lambda^2\beta_k^2}\right).$$

Plugging in the step sizes to the second term, we have

$$\begin{aligned}
\mathbb{E}\left[f(\theta_k) - f^\star\right] &\leqslant \left(\mathbb{E}\left[f(\theta_\mathcal{K}) - f^\star\right] + \frac{2L^2\alpha_{\tau_k}}{\lambda\beta_{\tau_k}}\mathbb{E}[\|z_\mathcal{K}\|^2]\right)\prod_{t=\mathcal{K}}^{k-1}(1 - \frac{\lambda\alpha_t}{2}) \\
&\quad + \tau_k^2\sum_{\ell=\mathcal{K}}^{k-1}\Big(\frac{25L^2B^3\alpha_0^2}{2c_\tau(\ell+1)^2} + \frac{4L^4B^2\alpha_0^3}{\lambda^2\beta_0^2(\ell+1)^{5/3}} \\
&\qquad + \frac{48C_2L^2}{\lambda}(\|\theta_0\|^2 + B^2\alpha_0^2 + 1)\frac{\alpha_0\beta_0}{c_\tau(\ell+1)^{5/3}}\log^2(k+1)\Big)\prod_{t=\ell+1}^{k-1}(1 - \frac{\lambda\alpha_t}{2}) \\
&\leqslant \left(\mathbb{E}\left[f(\theta_\mathcal{K}) - f^\star\right] + \frac{2L^2\alpha_\mathcal{K}}{\lambda\beta_\mathcal{K}}\mathbb{E}[\|z_\mathcal{K}\|^2]\right)\prod_{t=\mathcal{K}}^{k-1}(1 - \frac{\lambda\alpha_t}{2}) \\
&\quad + \tau_k^2\log^2(k+1)\sum_{\ell=\mathcal{K}}^{k-1}\frac{C_3}{(\ell+1)^{5/3}}\prod_{t=\ell+1}^{k-1}(1 - \frac{\lambda\alpha_t}{2}),
\end{aligned} \tag{A.17}$$

where we use the fact that $\frac{1}{\log^2(k+1)} \leqslant 12$ for all $k > 0$ and the definition of C_3.

Since $1 + c \leqslant \exp(c)$ for any scalar c, we have

$$\begin{aligned}
\prod_{t=\mathcal{K}}^{k-1}(1 - \frac{\lambda\alpha_t}{2}) &\leqslant \prod_{t=\mathcal{K}}^{k-1}\exp(-\frac{\lambda\alpha_t}{2}) = \exp(-\sum_{t=\mathcal{K}}^{k-1}\frac{\lambda\alpha_t}{2}) \leqslant \exp(-\frac{\lambda\alpha_0}{2}\sum_{t=\mathcal{K}}^{k-1}\frac{1}{t+1}) \\
&\leqslant \exp(-\frac{\lambda\alpha_0}{2}\log(\frac{k+1}{\mathcal{K}+1})) \leqslant (\frac{\mathcal{K}+1}{k+1})^{\frac{\lambda\alpha_0}{2}} \leqslant \frac{\mathcal{K}+1}{k+1},
\end{aligned} \tag{A.18}$$

where the last inequality results from $\alpha_0 \geqslant \frac{2}{\lambda}$, and the third inequality follows from $\sum_{t=k_1}^{k_2} \frac{1}{t+1} \geqslant \log(\frac{k_2+2}{k_1+1})$. Similarly, we have

$$\prod_{t=\ell+1}^{k-1} (1 - \frac{\lambda\alpha_t}{2}) \leqslant \frac{2\ell+1}{k+1} \leqslant \frac{2(\ell+1)}{k+1}. \tag{A.19}$$

Using Equation A.18 and Equation A.19 in Equation A.17,

$$\begin{aligned}
\mathbb{E}\left[f(\theta_k) - f^\star\right] &\leqslant \left(\mathbb{E}\left[f(\theta_{\mathcal{K}}) - f^\star\right] + \frac{2L^2\alpha_{\mathcal{K}}}{\lambda\beta_{\mathcal{K}}}\mathbb{E}[\|z_{\mathcal{K}}\|^2]\right)\frac{\mathcal{K}}{k+1} + \frac{\log^2(k+1)\tau_k^2}{k+1}\sum_{\ell=\mathcal{K}}^{k-1}\frac{2C_3}{(\ell+1)^{2/3}} \\
&\leqslant \frac{\mathcal{K}+1}{k+1}\left(\mathbb{E}\left[f(\theta_{\mathcal{K}}) - f^\star\right] + \frac{2L^2\alpha_0}{\lambda\beta_0}\mathbb{E}[\|z_{\mathcal{K}}\|^2]\right) + \frac{2C_3\log^2(k+1)\tau_k^2}{3(k+1)^{2/3}},
\end{aligned}$$

where the second inequality is a result of the relation $\sum_{t=0}^{t'} \frac{1}{(t+1)^{2/3}} \leqslant \frac{(t'+1)^{1/3}}{3}$ for any $t' \geqslant 0$. The claimed result follows from this and Equation 2.20.

□

A.2 Proof of Additional Lemmas

A.2.1 Proof of Lemma A.1

Our Markov process is a time-varying one (they depend on the iterates θ). Therefore, one cannot directly utilize Assumption 2.5 to analyze the bias of G in Algorithm 2.1 since the mixing time is defined for a fixed Markov chain (see Definition 2.1). To handle this difficulty, we introduce an auxiliary Markov chain $\{\widetilde{X}_k\}$ generated under the decision variable $\theta_{k-\tau_k}$ starting from $X_{k-\tau_k}$ as follows

$$X_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k}} \widetilde{X}_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k}} \cdots \widetilde{X}_{k-1} \xrightarrow{\theta_{k-\tau_k}} \widetilde{X}_k. \tag{A.20}$$

For clarity, we recall original the time-varying Markov processes $\{X_k\}$ generated by Algorithm Algorithm 2.1

$$X_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k+1}} X_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k+2}} \cdots \xrightarrow{\theta_{k-1}} X_{k-1} \xrightarrow{\theta_k} X_k.$$

Using the shorthand notation $y_k = \nabla f(\theta_k)$, we define the following quantities

$$\begin{aligned}
T_1 &= \mathbb{E}[\langle y_k - y_{k-\tau_k}, \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_k, \omega^\star(\theta_k), \hat{X})] - H(\theta_k, \omega^\star(\theta_k), X_k)\rangle] \\
T_2 &= \mathbb{E}[\langle y_{k-\tau_k}, H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), X_k) - H(\theta_k, \omega^\star(\theta_k), X_k)\rangle] \\
T_3 &= \mathbb{E}[\langle y_{k-\tau_k}, H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \widetilde{X}_k) - H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), X_k)\rangle] \\
T_4 &= \mathbb{E}[\langle y_{k-\tau_k}, \mathbb{E}_{\bar{X}\sim\mu_{\theta_{k-\tau_k}}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \bar{X})] - H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \widetilde{X}_k)\rangle] \\
T_5 &= \mathbb{E}[\langle y_{k-\tau_k}, \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \hat{X})] - \mathbb{E}_{\bar{X}\sim\mu_{\theta_{k-\tau_k}}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \bar{X})]\rangle] \\
T_6 &= \mathbb{E}[\langle y_{k-\tau_k}, \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_k, \omega^\star(\theta_k), \hat{X})] - \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \hat{X}]\rangle].
\end{aligned}$$

It is easy to see that

$$-\mathbb{E}\left[\langle \nabla f(\theta_k), \Delta H_k\rangle\right] = T_1 + T_2 + T_3 + T_4 + T_5 + T_6. \tag{A.21}$$

We analyze the terms of Equation A.21 individually. First, we treat T_1 using the boundedness of H and the Lipschitz continuity of ∇f

$$\begin{aligned}
T_1 &\leqslant \mathbb{E}\Big[\, \|y_k - y_{k-\tau_k}\| \left\| H(\theta_k, \omega^\star(\theta_k), X_k) - \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_k, \omega^\star(\theta_k), \hat{X})] \right\| \,\Big] \\
&\leqslant L\mathbb{E}\Big[\, \|\theta_k - \theta_{k-\tau_k}\| \,\Big]] \cdot 2B \leqslant 2B^2 L \tau_k \alpha_{k-\tau_k},
\end{aligned} \tag{A.22}$$

where the last inequality follows from

$$\|\theta_k - \theta_{k-\tau_k}\| \leqslant \sum_{t=k-\tau_k}^{k} \|\alpha_t H(\theta_t, \omega_t, X_t)\| \leqslant B\tau_k\alpha_{k-\tau_k}.$$

Similarly, for T_2 we have

$$\begin{aligned} T_2 &\leqslant \mathbb{E}\big[\|y_{k-\tau_k}\|\|H(\theta_k, \omega^\star(\theta_k), X_k) - H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), X_k)\|\big] \\ &\leqslant B\mathbb{E}\big[\|H(\theta_k, \omega^\star(\theta_k), X_k) - H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), X_k)\|\big] \\ &\leqslant BL\mathbb{E}\big[\|\theta_k - \theta_{k-\tau_k}\| + \|\omega^\star(\theta_k) - \omega^\star(\theta_{k-\tau_k})\|\big] \\ &\leqslant BL(L+1)\mathbb{E}\big[\|\theta_k - \theta_{k-\tau_k}\|\big] \leqslant B^2L(L+1)\tau_k\alpha_{k-\tau_k}. \end{aligned} \tag{A.23}$$

To analyze T_3, we utilize the law of total expectation: given $\mathcal{F} \subseteq \mathcal{F}'$ and a random variable X we have $\mathbb{E}[X \mid \mathcal{F}] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{F}'] \mid \mathcal{F}]$.

Let $\mathcal{F}_k$ be $\mathcal{F}_k = \{X_0, \ldots, X_k, \theta_0, \ldots, \theta_k, \omega_0, ..., \omega_k\}$, and for convenience we denote

$$p_k(x) = P(X_k = x \mid \mathcal{F}_{k-1}) \quad \text{and} \quad \tilde{p}_k(x) = P(\widetilde{X}_k = x \mid \mathcal{F}_{k-1}).$$

Then, we have

$$\begin{aligned} &\mathbb{E}\Big[\Big\langle y_{k-\tau_k}, H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \widetilde{X}_k) - H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), X_k)\Big\rangle \mid \mathcal{F}_{k-\tau_k}\Big] \\ &\leqslant \|y_{k-\tau_k}\| \, \|\mathbb{E}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), X_k) - H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \widetilde{X}_k) \mid \mathcal{F}_{k-\tau_k}]\| \\ &= \|y_{k-\tau_k}\| \, \big\|\mathbb{E}\big[\mathbb{E}\big[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), X_k) - H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \widetilde{X}_k) \,|\, \mathcal{F}_{k-1}\big] \mid \mathcal{F}_{k-\tau_k}\big]\big\| \\ &\leqslant B\mathbb{E}[\int_{\mathcal{X}} H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), x)(p_k(x) - \widetilde{p}_k(x))dx \mid \mathcal{F}_{k-\tau_k}] \\ &\leqslant 2B^2\mathbb{E}[d_{TV}(p_k(\cdot), \widetilde{p}_k(\cdot)) \mid \mathcal{F}_{k-\tau_k}] \\ &\leqslant 2B^2\mathbb{E}\big[d_{TV}(p_{k-1}(\cdot), \widetilde{p}_{k-1}(\cdot)) + L\|\theta_k - \theta_{k-\tau_k}\| \mid \mathcal{F}_{k-\tau_k}\big], \end{aligned}$$

where the second inequality uses the definition of the TV distance in Equation 2.19, and the last inequality is a result of Assumption 2.6. Recursively applying this inequality and taking the expectation, we get

$$T_3 \leqslant 2B^2L \sum_{t=k-\tau_k+1}^{k-1} \mathbb{E}[\|\theta_t - \theta_{k-\tau_k}\|] \leqslant 2B^3L\tau_k^2\alpha_{k-\tau_k}. \tag{A.24}$$

Similarly, to bound T_4, we again use the definition of TV distance

$$\begin{aligned}
&\mathbb{E}[\langle y_{k-\tau_k}, \mathbb{E}_{\bar{X}\sim\mu_{\theta_{k-\tau_k}}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \bar{X})] - H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \widetilde{X}_k)\rangle \mid \mathcal{F}_{k-\tau_k}] \\
&\leqslant \|y_{k-\tau_k}\| \left\|\mathbb{E}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \widetilde{X}_k) - \mathbb{E}_{\bar{X}\sim\mu_{\theta_{k-\tau_k}}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \bar{X})] \,|\, \mathcal{F}_{k-\tau_k}]\right\| \\
&\leqslant B \cdot 2B\mathbb{E}[d_{TV}(\widetilde{p}_k(\cdot), \mu_{\theta_{k-\tau_k}}) \mid \mathcal{F}_{k-\tau_k}].
\end{aligned}$$

Taking the expectation and using the definition of the mixing time 2.1,

$$T_4 \leqslant 2B^2\mathbb{E}[d_{TV}(P(\widetilde{X}_k = \cdot), \mu_{\theta_{k-\tau_k}})] \leqslant 2B^2\alpha_k. \tag{A.25}$$

We next consider T_5

$$\begin{aligned}
&\mathbb{E}[\langle y_{k-\tau_k}, \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \hat{X})] - \mathbb{E}_{\bar{X}\sim\mu_{\theta_{k-\tau_k}}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \bar{X})]\rangle \,|\, \mathcal{F}_{k-\tau_k}] \\
&\leqslant B\|\mathbb{E}[\mathbb{E}_{\bar{X}\sim\mu_{\theta_{k-\tau_k}}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \bar{X})] - \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \hat{X})] \,|\, \mathcal{F}_{k-\tau_k}]\| \\
&\leqslant 2B^2\mathbb{E}[d_{TV}(\mu_{\theta_{k-\tau_k}}, \mu_{\theta_k}) \mid \mathcal{F}_{k-\tau_k}],
\end{aligned}$$

where the last inequality again comes from the definition of the TV distance in Equation 2.19. By Equation 2.22 in Assumption 2.6, we have

$$T_5 \leqslant 2B^2\mathbb{E}[d_{TV}(\mu_{\theta_{k-\tau_k}}, \mu_{\theta_k})] \leqslant 2B^2L\mathbb{E}[\|\theta_k - \theta_{k-\tau_k}\|] \leqslant 2B^3L\tau_k\alpha_k. \tag{A.26}$$

Finally, we bound T_6 using the boundedness of ∇f and the Lipschitz continuity of H

$$\begin{aligned}
T_6 &\leqslant \mathbb{E}\big[\|y_{k-\tau_k}\|\|\mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}\big[H(\theta_{k-\tau_k}, \omega^\star(\theta_{k-\tau_k}), \hat{X})\big] - \mathbb{E}_{\hat{X}\sim\mu_{\theta_k}}\big[H(\theta_k, \omega^\star(\theta_k), \hat{X})\big]\|\big] \\
&\leqslant BL\mathbb{E}[\|\theta_{k-\tau_k} - \theta_k\| + \|\omega^\star(\theta_{k-\tau_k}) - \omega^\star(\theta_k)\|] \leqslant 2L^2B\mathbb{E}[\|\theta_{k-\tau_k} - \theta_k\|] \leqslant 2L^2B^2\tau_k\alpha_{k-\tau_k}.
\end{aligned}$$

The claimed result follows from plugging the bounds on T_1-T_6 into Equation A.21.

□

A.2.2 Proof of Lemma A.3

As a result of Equation 2.24, for any $k \geqslant 0$

$$\|\omega_{k+1}\| - \|\omega_k\| \leqslant \|\omega_{k+1} - \omega_k\| = \|\beta_k G(\theta_{k+1}, \omega_k, X_k)\| \leqslant D\beta_k \left(\|\theta_{k+1}\| + \|\omega_k\| + 1\right). \quad \text{(A.27)}$$

Define $h_k = \|\omega_k\| + \|\theta_{k+1}\|$. We have for all $k \geqslant 1$

$$\begin{aligned} h_k &= \|\omega_{k-1} + \beta_{k-1} G(\theta_k, \omega_{k-1}, X_{k-1})\| + \|\theta_k + \alpha_k H(\theta_k, \omega_k, X_k)\| \\ &\leqslant \|\omega_{k-1}\| + D\beta_{k-1}(\|\theta_k\| + \|\omega_{k-1}\| + 1) + \|\theta_k\| + B\alpha_k \\ &\leqslant (1 + D\beta_{k-1}) h_{k-1} + (B + D)\beta_{k-1} \end{aligned}$$

where the second inequality follows from Equation 2.24 and Assumption 2.1, and the last inequality is due to $\alpha_k \leqslant \beta_k$ and the fact that $\{\beta_k\}$ is a decaying sequence.

Since $1 + x \leqslant e^x$ for all $x \geqslant 0$, we have for all $k \geqslant \tau_k$ and $k - \tau_k \leqslant t \leqslant k$

$$\begin{aligned} h_t &\leqslant (1 + D\beta_{t-1}) h_{t-1} + (B + D)\beta_{t-1} \\ &\leqslant (1 + D\beta_{k-\tau_k})^{\tau_k + t - k} h_{k-\tau_k} + (B + D)\beta_{k-\tau_k} \sum_{t'=k-\tau_k}^{t-1} (1 + D\beta_{k-\tau_k})^{t-t'-1} \\ &\leqslant (1 + D\beta_{k-\tau_k})^{\tau_k} h_{k-\tau_k} + (B + D)\beta_{k-\tau_k}\tau_k (1 + D\beta_{k-\tau_k})^{\tau_k} \\ &\leqslant e^{D\beta_{k-\tau_k}\tau_k} h_{k-\tau_k} + (B + D)\beta_{k-\tau_k}\tau_k e^{D\beta_{k-\tau_k}\tau_k} \leqslant 2h_{k-\tau_k} + \frac{1}{3}, \end{aligned}$$

where the last inequality follows from the step size $2(B + D)\beta_{k-\tau_k}\tau_k \leqslant \frac{1}{3} \leqslant \log(2)$. Combining this inequality with Equation A.27, we have for all $k \geqslant \tau_k$

$$\begin{aligned} &\|\omega_k - \omega_{k-\tau_k}\| \leqslant \sum_{t=k-\tau_k}^{k-1} \|\omega_{t+1} - \omega_t\| \leqslant D \sum_{t=k-\tau_k}^{k-1} \beta_t \left(h_t + 1\right) \leqslant D\beta_{k-\tau_k} \sum_{t=k-\tau_k}^{k-1} \left(2h_{k-\tau_k} + \frac{4}{3}\right) \\ &\leqslant 2D\beta_{k-\tau_k}\tau_k (\|\omega_{k-\tau_k}\| + \|\theta_{k-\tau_k+1}\| + \frac{2}{3}) \leqslant 2D\beta_{k-\tau_k}\tau_k (\|\omega_{k-\tau_k}\| + \|\theta_k\| + B\beta_{k-\tau_k}(\tau_k - 1) + \frac{2}{3}) \\ &\leqslant 2D\beta_{k-\tau_k}\tau_k \left(\|\omega_k - \omega_{k-\tau_k}\| + \|\omega_k\| + \|\theta_k\| + 1\right). \end{aligned}$$

Re-arranging terms and again using the step size condition $2D\beta_{k-\tau_k}\tau_k \leqslant \frac{1}{3}$, we get

$$\|\omega_k - \omega_{k-\tau_k}\| \leqslant 3D\beta_{k-\tau_k}\tau_k \left(\|\omega_k\| + \|\theta_k\| + 1\right).$$

□

APPENDIX B

SUPPLEMENTARY MATERIAL FOR RESULTS IN CHAPTER 3

B.1 Computation Details of Examples in Section 3.3

First, we look at the example in Section 3.3 which illustrates that deterministic optimal policy may not exist in multi-task RL. As we discussed, it is easy to see that the optimal policy in state S_2 and S_4 is to always take action L in order to reach the positive reward or to stay away from the negative reward, and all that is left to be figured out is the policy at state S_3.

There are 2 possible deterministic policies in state S_3, to always take action L or to always take action R. First, consider one policy $\pi_{d,l}$, which is to always take L.

We have $V_1^{\pi_{d,l}}(S_3) = \gamma$ as the agent reaches S_1 in 2 steps under $\pi_{d,l}$ and claims the $+1$ reward. However, this policy produces a zero value in environment 2, $V_2^{\pi_d}(S_3) = 0$, since an agent will move back and forth between S_3 and S_4 forever. Therefore, this deterministic policy achieves

$$V_1^{\pi_{d,l}}(S_3) + V_2^{\pi_{d,l}}(S_3) = \gamma + 0 = \gamma.$$

By symmetry, the value of the policy $\pi_{d,r}$, which is to always take action R in state S_3, is

$$V_1^{\pi_{d,r}}(S_3) + V_2^{\pi_{d,r}}(S_3) = 0 + \gamma = \gamma.$$

Now, let's consider a stochastic policy π_s, which we will show performs better than the two deterministic policies. This policy π_s takes the same deterministic actions as $\pi_{d,l}$ and

$\pi_{d,r}$ in state S_2, S_4, and is defined as follows for state S_3.

$$\pi_s(a|S_3) = \begin{cases} p, & a = \text{left} \\ 1-p, & a = \text{right} \end{cases}$$

We compute cumulative rewards under π_s.

$$\begin{aligned} V_1^{\pi_s}(S_3) &= p\gamma + p(1-p)\gamma^3 + p(1-p)^2\gamma^5 + ... \\ &= p\gamma \sum_{k=0}^{\infty} \left((1-p)\gamma^2\right)^k \\ &= \frac{p\gamma}{1-(1-p)\gamma^2}. \end{aligned}$$

Similarly,

$$\begin{aligned} V_2^{\pi_s}(S_3) &= (1-p)\gamma + (1-p)p\gamma^3 + (1-p)p^2\gamma^5 + ... \\ &= (1-p)\gamma \sum_{k=0}^{\infty} \left(p\gamma^2\right)^k \\ &= \frac{(1-p)\gamma}{1-p\gamma^2}. \end{aligned}$$

Then,

$$V_1^{\pi_s}(S_3) + V_2^{\pi_s}(S_3) = \frac{p\gamma}{1-(1-p)\gamma^2} + \frac{(1-p)\gamma}{1-p\gamma^2}.$$

Taking the derivative with respect to p and setting it to 0, we get

$$\frac{1}{(1-(1-p)\gamma^2)^2} = \frac{1}{(1-p\gamma^2)^2}, \tag{B.1}$$

which leads to $p = 0.5$.

The value of policy π_s at state S_3 is

$$\begin{aligned} V_1^{\pi_s}(S_3) + V_2^{\pi_s}(S_3) &= \frac{p\gamma}{1-(1-p)\gamma^2} + \frac{(1-p)\gamma}{1-p\gamma^2} \\ &= \frac{2\gamma}{2-\gamma^2}. \end{aligned}$$

Then, we explain how the three stationary points are computed in the second example in Section 3.3. Note that the gradient of the value function can be expressed as

$$\frac{\partial}{\partial \theta_{s,a}} V_i^{\pi_\theta}(\rho_i) = \frac{1}{1-\gamma_i} d_{i,\rho_i}^{\pi_\theta}(s) \pi_\theta(a|s) A_i^{\pi_\theta}(s,a). \tag{B.2}$$

We define $D_i^{\pi_\theta}$ to be the $|\mathcal{S}_i| \times |\mathcal{S}_i|$ matrix where the entry (i,j) is $d_i^{\pi_\theta}(s_i|s_j)$. It can be easily seen that

$$d_{i,\rho_i}^{\pi_\theta}(s) = D_i^{\pi_\theta} \rho_i. \tag{B.3}$$

Given $P_i^{\pi_\theta}$ the transition probability matrix of task i under policy π_θ (whose entry (j,k) denotes $P_i(j \mid k)$), the matrix $D_i^{\pi_\theta}$ can be computed as

$$D_i^{\pi} = (1 - \gamma P_i^{\pi})^{-1}. \tag{B.4}$$

Given the small scale and the known dynamics of the problem, we can also compute the value function and the Q function of the policy π_θ in the two tasks by solving the Bellman equation, from which we get $A_i^{\pi_\theta}(s,a)$. Specifically, under a policy π, the value functions

associated with the first and second tasks are

$$V_1^\pi = (I - \gamma(P_1^\pi)^\top)^{-1} \begin{bmatrix} 0 \\ 1-p \\ 0 \\ -p \\ 0 \end{bmatrix}, \quad \text{and} \quad V_2^\pi = (I - \gamma(P_2^\pi)^\top)^{-1} \begin{bmatrix} 0 \\ -p \\ 0 \\ 1-p \\ 0 \end{bmatrix}. \tag{B.5}$$

In addition, we can compute the Q functions

$$\begin{aligned}
Q_1^\pi(\cdot, L) &= [0, \quad (1-p) + \gamma p V_1^\pi(S_3), \quad \gamma V_1^\pi(S_2), \quad \gamma(1-p)V_1^\pi(S_3) - p, \quad 0]^\top, \\
Q_1^\pi(\cdot, R) &= [0, \quad (1-p) + \gamma p V_1^\pi(S_3), \quad \gamma V_1^\pi(S_4), \quad \gamma(1-p)V_1^\pi(S_3) - p, \quad 0]^\top, \\
Q_2^\pi(\cdot, L) &= [0, \quad \gamma(1-p)V_2^\pi(S_3) - p, \quad \gamma V_2^\pi(S_2), \quad \gamma p V_2^\pi(S_3) + (1-p), \quad 0]^\top, \\
Q_2^\pi(\cdot, R) &= [0, \quad \gamma(1-p)V_2^\pi(S_3) - p, \quad \gamma V_2^\pi(S_4), \quad \gamma p V_2^\pi(S_3) + (1-p), \quad 0]^\top,
\end{aligned} \tag{B.6}$$

from which the advantage function can be easily computed by taking the difference between the Q functions and the value functions. We also know $\pi_\theta(s, a)$ of the policy for which we would like to evaluate the gradient. Therefore, we can compute all the quantities in the gradient expression Equation B.2. Now we go through all three parameterizations and calculate the gradient and the cumulative return.

We first consider the policy π_1 under the parameterization $\theta_{S_3,L} = 1, \theta_{S_3,R} = \infty$, which implies $\pi_1(L \mid S_3) = 0$ and $\pi_1(R \mid S_3) = 1$. First, we can easily see that the transition

probability matrices are

$$P_1^{\pi_1} = \begin{bmatrix} 1 & 1-p & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & p & 0 & 1-p & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & p & 1 \end{bmatrix}, \quad \text{and} \quad P_2^{\pi_1} = \begin{bmatrix} 1 & p & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1-p & 1 \end{bmatrix}.$$

Computing $D_i^{\pi_1}$ according to Equation B.4 using Gaussian elimination, we can derive

$$D_1^{\pi_1} = \begin{bmatrix} 1 & \gamma(1-p) & 0 & 0 & 0 \\ 0 & 1-\gamma & 0 & 0 & 0 \\ 0 & \frac{\gamma p(1-\gamma)}{(\gamma^2 p-\gamma^2+1)} & \frac{1-\gamma}{\gamma^2 p-\gamma^2+1} & \frac{\gamma(1-\gamma)(1-p)}{\gamma^2 p-\gamma^2+1} & 0 \\ 0 & \frac{\gamma^2 p(1-\gamma)}{(\gamma^2 p-\gamma^2+1)} & \frac{\gamma(1-\gamma)}{\gamma^2 p-\gamma^2+1} & \frac{1-\gamma}{\gamma^2 p-\gamma^2+1} & 0 \\ 0 & \frac{\gamma^3 p^2}{(\gamma^2 p-\gamma^2+1)} & \frac{\gamma^2 p}{\gamma^2 p-\gamma^2+1} & \frac{\gamma p}{\gamma^2 p-\gamma^2+1} & 1 \end{bmatrix},$$

$$D_2^{\pi_1} = \begin{bmatrix} 1 & \gamma p & 0 & 0 & 0 \\ 0 & 1-\gamma & 0 & 0 & 0 \\ 0 & \frac{\gamma(1-\gamma)(1-p)}{1-\gamma^2 p} & \frac{1-\gamma}{1-\gamma^2 p} & \frac{\gamma(1-\gamma)p}{1-\gamma^2 p} & 0 \\ 0 & \frac{\gamma^2(1-\gamma)(1-p)}{1-\gamma^2 p} & \frac{\gamma(1-\gamma)}{1-\gamma^2 p} & \frac{1-\gamma}{1-\gamma^2 p} & 0 \\ 0 & \frac{\gamma^3(1-p)^2}{1-\gamma^2 p} & \frac{\gamma^2(1-p)}{1-\gamma^2 p} & \frac{\gamma(1-p)}{1-\gamma^2 p} & 1 \end{bmatrix}.$$

As explained in Equation B.5 and Equation B.6, we can compute the advantage functions

$$A_1^{\pi_1}(\cdot, L) = \left[0, 0, \quad \frac{\gamma(-\gamma^2 p^2 + (1-p)(\gamma^2 p - \gamma^2 + 1) + p)}{\gamma^2 p - \gamma^2 + 1}, 0, \quad 0\right]^\top,$$

$$A_1^{\pi_1}(\cdot, R) = \left[0, \quad 0, \quad 0, \quad 0, \quad 0\right]^\top,$$

$$A_2^{\pi_1}(\cdot, L) = \left[0, \quad 0, \quad \frac{\gamma(\gamma^2(1-p)^2 + p(\gamma^2 p - 1) - (1-p))}{1-\gamma^2 p}, \quad 0, \quad 0\right]^\top,$$

$$A_2^{\pi_1}(\cdot, R) = [0, \quad 0, \quad 0, \quad 0, \quad 0]^\top.$$

Recall Equation B.2, which implies

$$\begin{aligned}\frac{\partial}{\partial \theta_{S_3,L}}(V_1^{\pi_1}(\rho_1) + V_2^{\pi_1}(\rho_2)) &= \frac{1}{1-\gamma} d_{1,\rho_1}^{\pi_1}(S_3)\pi_1(L|S_3)A_1^{\pi_1}(S_3, L) \\ &\quad + \frac{1}{1-\gamma} d_{2,\rho_2}^{\pi_1}(S_3)\pi_1(L|S_3)A_2^{\pi_1}(S_3, L) \\ &= 0,\end{aligned}$$

since $\pi_1(L \mid S_3) = 0$. Similarly, we have

$$\begin{aligned}\frac{\partial}{\partial \theta_{S_3,R}}(V_1^{\pi_1}(\rho_1) + V_2^{\pi_1}(\rho_2)) &= \frac{1}{1-\gamma} d_{1,\rho_1}^{\pi_1}(S_3)\pi_1(R|S_3)A_1^{\pi_1}(S_3, R) \\ &\quad + \frac{1}{1-\gamma} d_{2,\rho_2}^{\pi_1}(S_3)\pi_1(R|S_3)A_2^{\pi_1}(S_3, R) \\ &= 0,\end{aligned}$$

since $A_1^{\pi_1}(S_3, R) = A_2^{\pi_1}(S_3, R) = 0$. The cumulative return under this policy is

$$V_1^{\pi_1}(\rho_1) + V_2^{\pi_1}(\rho_2) = V_1^{\pi_1}(S_3) + V_2^{\pi_1}(S_3) = \frac{\gamma(-2\gamma^2 p + \gamma^2 + 2p - 1)}{\gamma^4 p^2 - \gamma^4 p + \gamma^2 - 1}.$$

By symmetry, the second policy π_2 under parameterization $\theta_{S_3,L} = \infty, \theta_{S_3,R} = 1$ is also a stationary point and has a cumulative return

$$V_1^{\pi_2}(\rho_1) + V_2^{\pi_2}(\rho_2) = \frac{\gamma(-2\gamma^2 p + \gamma^2 + 2p - 1)}{\gamma^4 p^2 - \gamma^4 p + \gamma^2 - 1}.$$

Finally, we look at the policy π_3 under parameterization $\theta_{S_3,L} = 1, \theta_{S_3,R} = 1$, which implies $\pi_3(L \mid S_3) = \pi_3(R \mid S_3) = 0.5$. We can see that the transition probability matrices

are

$$P_1^{\pi_3} = \begin{bmatrix} 1 & 1-p & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & p & 0 & 1-p & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & p & 1 \end{bmatrix}, \quad \text{and} \quad P_2^{\pi_3} = \begin{bmatrix} 1 & p & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 1-p & 1 \end{bmatrix}.$$

Computing $D_i^{\pi_3}$ according to Equation B.4 using Gaussian elimination, we can derive

$$D_1^{\pi_3} = \begin{bmatrix} 1 & \frac{\gamma(-\gamma^2p^2+2\gamma^2p-\gamma^2-2p+2)}{2-\gamma^2} & \frac{\gamma^2(1-p)}{2-\gamma^2} & \frac{\gamma^3(1-p)^2}{2-\gamma^2} & 0 \\ 0 & \frac{(1-\gamma)(\gamma^2p-\gamma^2+2)}{2-\gamma^2} & \frac{\gamma(1-\gamma)}{2-\gamma^2} & \frac{\gamma^2(1-\gamma)(1-p)}{2-\gamma^2} & 0 \\ 0 & \frac{2\gamma(1-\gamma)(1-p)}{2-\gamma^2} & \frac{2(1-\gamma)}{2-\gamma^2} & \frac{2\gamma(1-\gamma)(1-p)}{2-\gamma^2} & 0 \\ 0 & \frac{\gamma^2(1-\gamma)p}{2-\gamma^2} & \frac{\gamma(1-\gamma)}{2-\gamma^2} & \frac{(1-\gamma)(2-\gamma^2p)}{2-\gamma^2} & 0 \\ 0 & \frac{\gamma^3p^2}{2-\gamma^2} & \frac{\gamma^2p}{2-\gamma^2} & \frac{\gamma p(2-\gamma^2p)}{2-\gamma^2} & 1 \end{bmatrix},$$

$$D_2^{\pi_3} = \begin{bmatrix} 1 & \frac{\gamma p(2-\gamma^2p)}{2-\gamma^2} & \frac{\gamma^2p}{2-\gamma^2} & \frac{\gamma^2(1-\gamma)p}{2-\gamma^2} & 0 \\ 0 & \frac{(1-\gamma)(2-\gamma^2p)}{2-\gamma^2} & \frac{\gamma(1-\gamma)}{2-\gamma^2} & \frac{\gamma^2(1-\gamma)p}{2-\gamma^2} & 0 \\ 0 & \frac{2\gamma(1-\gamma)(1-p)}{2-\gamma^2} & \frac{2(1-\gamma)}{2-\gamma^2} & \frac{2\gamma(1-\gamma)p}{2-\gamma^2} & 0 \\ 0 & \frac{\gamma^2(1-\gamma)(1-p)}{2-\gamma^2} & \frac{\gamma(1-\gamma)}{2-\gamma^2} & \frac{(1-\gamma)(2-\gamma^2+\gamma^2p)}{2-\gamma^2} & 0 \\ 0 & \frac{\gamma^3(1-p)^2}{2-\gamma^2} & \frac{\gamma^2(1-p)}{2-\gamma^2} & \frac{\gamma(2-\gamma^2p^2+2\gamma^2p-\gamma^2-2p)}{2-\gamma^2} & 1 \end{bmatrix}.$$

The advantage functions are

$$A_1^{\pi_3}(\cdot, L) = \left[0, 0, \quad \frac{\gamma(-2\gamma^2p^2 + 2\gamma^2p - \gamma^2 + 1)}{2-\gamma^2}, 0, \quad 0\right]^\top,$$
$$A_1^{\pi_3}(\cdot, R) = \left[0, \quad 0, \quad \frac{\gamma(2\gamma^2p^2 - 2\gamma^2p + \gamma^2 - 1)}{2-\gamma^2}, \quad 0, \quad 0\right]^\top,$$
$$A_2^{\pi_3}(\cdot, L) = \left[0, \quad 0, \quad \frac{\gamma(2\gamma^2p^2 - 2\gamma^2p + \gamma^2 - 1)}{2-\gamma^2}, \quad 0, \quad 0\right]^\top,$$

$$A_2^{\pi_3}(\cdot, R) = \left[0, \quad 0, \quad \frac{\gamma(-2\gamma^2 p^2 + 2\gamma^2 p - \gamma^2 + 1)}{2 - \gamma^2}, \quad 0, \quad 0\right]^\top,$$

From Equation B.2, we have

$$\begin{aligned}
\frac{\partial}{\partial \theta_{S_3,L}}(V_1^{\pi_3}(\rho_1) + V_2^{\pi_3}(\rho_2)) &= \frac{1}{1-\gamma}\pi_3(L|S_3)\left(d_{1,\rho_1}^{\pi_3}(S_3)A_1^{\pi_3}(S_3, L) + d_{2,\rho_2}^{\pi_3}(S_3)A_2^{\pi_3}(S_3, L)\right) \\
&= \frac{0.5}{1-\gamma} \cdot \frac{2(1-\gamma)}{2-\gamma^2} \cdot \frac{\gamma(-2\gamma^2 p^2 + 2\gamma^2 p - \gamma^2 + 1)}{2-\gamma^2} \\
&\quad + \frac{0.5}{1-\gamma} \cdot \frac{2(1-\gamma)}{2-\gamma^2} \cdot \frac{\gamma(2\gamma^2 p^2 - 2\gamma^2 p + \gamma^2 - 1)}{2-\gamma^2} \\
&= 0.
\end{aligned}$$

Similarly,

$$\frac{\partial}{\partial \theta_{S_3,R}}(V_1^{\pi_3}(\rho_1) + V_2^{\pi_3}(\rho_2)) = 0.$$

The cumulative return under this policy is

$$V_1^{\pi_3}(\rho_1) + V_2^{\pi_3}(\rho_2) = V_1^{\pi_1}(S_3) + V_2^{\pi_1}(S_3) = \frac{\gamma(2-4p)}{2-\gamma^2}.$$

For computational simplicity, we choose $\gamma = \sqrt{0.5}$. Then,

$$\begin{aligned}
&V_1^{\pi_1}(\rho_1) + V_2^{\pi_1}(\rho_2) = \frac{\gamma(-2\gamma^2 p + \gamma^2 + 2p - 1)}{\gamma^4 p^2 - \gamma^4 p + \gamma^2 - 1} = \frac{2p-1}{8\sqrt{2}(p-2)(p+1)}, \\
\text{and} \quad &V_1^{\pi_3}(\rho_1) + V_2^{\pi_3}(\rho_2) = V_1^{\pi_1}(S_3) + V_2^{\pi_1}(S_3) = \frac{\gamma(2-4p)}{2-\gamma^2} = \frac{4-8p}{3}.
\end{aligned}$$

If $p > 0.5$,

$$\begin{aligned}
V_1^{\pi_1}(\rho_1) + V_2^{\pi_1}(\rho_2) &= V_1^{\pi_2}(\rho_1) + V_2^{\pi_2}(\rho_2) \\
&= \frac{2p-1}{8\sqrt{2}(p-2)(p+1)} > \frac{4-8p}{3} = V_1^{\pi_3}(\rho_1) + V_2^{\pi_3}(\rho_2).
\end{aligned}$$

B.2 Lipschitz, Gradient Lipschitz, and Hessian Lipschitz Constants

In this section, we show that the value function and the relative entropy regularizer are Lipschitz and have Lipschitz continuous gradients and Hessians. We present the result in two lemmas as well as their proofs.

Lemma B.1. *Under the tabular softmax policy, $V_i^{\pi_\theta}(\mu)$ is Lipschitz, has a Lipschitz gradient and a Lipschtz Hessian for all i and μ, i.e.*

$$
\begin{aligned}
&||V_i^{\pi_{\theta'}}(\mu) - V_i^{\pi_{\theta''}}(\mu)|| \leqslant \frac{2}{(1-\gamma_i)^2}||\theta' - \theta''||,\\
&||\nabla_{\theta'} V_i^{\pi_{\theta'}}(\mu) - \nabla_{\theta''} V_i^{\pi_{\theta''}}(\mu)|| \leqslant \frac{8}{(1-\gamma_i)^3}||\theta' - \theta''||, \text{ and}\\
&||\nabla^2_{\theta'} V_i^{\pi_{\theta'}}(\mu) - \nabla^2_{\theta''} V_i^{\pi_{\theta''}}(\mu)|| \leqslant \frac{48}{(1-\gamma_i)^4}||\theta' - \theta''||.
\end{aligned}
$$

Lemma B.2. *The cross entropy regularizer is Lipschitz, has a Lipschitz gradient and a Lipschtz Hessian, i.e.*

$$
\begin{aligned}
&||\lambda RE(\pi'_\theta) - \lambda RE(\pi''_\theta)|| \leqslant \lambda(\frac{1}{\sqrt{|\mathcal{A}|}} + 1)||\theta' - \theta''||,\\
&||\nabla_{\theta'} \lambda RE(\pi'_\theta) - \nabla_{\theta''} \lambda RE(\pi''_\theta)|| \leqslant \frac{2\lambda}{|\mathcal{S}|}||\theta' - \theta''||, \text{ and}\\
&||\nabla^2_{\theta'} \lambda RE(\pi'_\theta) - \nabla^2_{\theta''} \lambda RE(\pi''_\theta)|| \leqslant \frac{6\lambda}{|\mathcal{S}|}||\theta' - \theta''||.
\end{aligned} \tag{B.7}
$$

B.2.1 Proof of Lemma B.1

The proof of Lemma B.1 employs an intermediate result, which we state below.

Lemma B.3. *Let $\pi_\alpha \triangleq \pi_{\boldsymbol{\theta}+\alpha \boldsymbol{u}}$, where $\boldsymbol{u}$ is a unit vector and $\tilde{V}_i(\alpha) \triangleq V_i^{\pi_\alpha}(s_i)$. If*

$$
\sum_{a\in\mathcal{A}} \left| \frac{d\pi_\alpha(a|s_0)}{d\alpha}\bigg|_{\alpha=0} \right| \leqslant C', \ \sum_{a\in\mathcal{A}} \left| \frac{d^2\pi_\alpha(a|s_0)}{d\alpha^2}\bigg|_{\alpha=0} \right| \leqslant C'', \ \sum_{a\in\mathcal{A}} \left| \frac{d^3\pi_\alpha(a|s_0)}{d\alpha^3}\bigg|_{\alpha=0} \right| \leqslant C''',
$$

then we have

$$\max_{||\boldsymbol{u}||_2=1} \left| \frac{d\tilde{V}_i(\alpha)}{d\alpha}\bigg|_{\alpha=0} \right| \leqslant \frac{C'}{(1-\gamma_i)^2},$$
$$\max_{||\boldsymbol{u}||_2=1} \left| \frac{d^2\tilde{V}_i(\alpha)}{d\alpha^2}\bigg|_{\alpha=0} \right| \leqslant \frac{C''}{(1-\gamma_i)^2} + \frac{2\gamma_i C'^2}{(1-\gamma_i)^3},$$
$$\max_{||\boldsymbol{u}||_2=1} \left| \frac{d^3\tilde{V}_i(\alpha)}{d\alpha^3}\bigg|_{\alpha=0} \right| \leqslant \frac{C'''}{(1-\gamma_i)^2} + \frac{6\gamma_i C' C''}{(1-\gamma_i)^3} + \frac{6\gamma_i^2 C'^3}{(1-\gamma_i)^4}$$

To show a function is Lipschitz, we show the derivative of the Hessian with respect to θ is bounded. Under the softmax parameterization, we have

$$\nabla_{\theta_s} \pi_\theta(a|s) = \pi_\theta(a|s) \left(e_a - \pi(\cdot|s)\right), \tag{B.8}$$

$$\nabla^2_{\theta_s} \pi_\theta(a|s) = \pi_\theta(a|s) \left(e_a e_a^\top - e_a \pi(\cdot|s)^\top - \pi(\cdot|s) e_a^\top + 2\pi(\cdot|s)\pi(\cdot|s)^\top - \mathrm{diag}(\pi(\cdot|s))\right), \tag{B.9}$$

$$\begin{aligned}\frac{\partial}{\partial \theta_{s,a'}} \nabla^2_{\theta_s} \pi_\theta(a|s) = \pi_\theta(a|s)(\mathbf{1}(a=a') - \pi_\theta(a'|s)) \big(e_a e_a^\top - e_a \pi(\cdot|s)^\top - \pi(\cdot|s) e_a^\top \\ + 2\pi(\cdot|s)\pi(\cdot|s)^\top - \mathrm{diag}(\pi(\cdot|s))\big) \\ + \pi_\theta(a|s)(-e_a \pi_\theta(a'|s) e_{a'}^T + e_a \pi_\theta(a'|s)\pi_\theta(\cdot|s)^T - e_{a'} \pi_\theta(a'|s) e_a^T \\ + \pi_\theta(\cdot|s))\pi_\theta(a'|s) e_a^T + 4\pi_\theta(\cdot|s)\pi_\theta(a'|s) e_{a'}^T - 4\pi_\theta(\cdot|s)\pi_\theta \pi_\theta(\cdot|s)^T \\ + \mathrm{diag}(\pi_\theta(a'|s) e_a) - \mathrm{diag}(\pi_\theta(a'|s)\pi_\theta(\cdot|s)^T))\end{aligned} \tag{B.10}$$

where e_a is a vector with all 0 and 1 at action a. Then, for any s,

$$\begin{aligned}\sum_{a\in\mathcal{A}} \left| \frac{d\pi_\alpha(a|s)}{d\alpha}\bigg|_{\alpha=0} \right| &\leqslant \sum_{a\in\mathcal{A}} \left| \boldsymbol{u}^T \nabla_{\theta+\alpha\boldsymbol{u}} \pi_\alpha(a|s)\big|_{\alpha=0} \right| \\ &\leqslant \sum_{a\in\mathcal{A}} \pi_\theta(a|s) \left| \boldsymbol{u}_s^T e_a - \boldsymbol{u}_s^T \pi(\cdot|s) \right|\end{aligned}$$

$$\leqslant \max_{a\in\mathcal{A}} \left(\left| \boldsymbol{u}_s^T e_a \right| + \left| \boldsymbol{u}_s^T \pi(\cdot|s) \right| \right) \leqslant 2, \tag{B.11}$$

$$\begin{aligned}
\sum_{a\in\mathcal{A}} \left| \frac{d^2\pi_\alpha(a|s)}{d\alpha^2}\Big|_{\alpha=0} \right| &\leqslant \sum_{a\in\mathcal{A}} \left| \boldsymbol{u}^T \nabla^2_{\theta+\alpha\boldsymbol{u}} \pi_\alpha(a|s)\big|_{\alpha=0} \boldsymbol{u} \right| \\
&\leqslant \max_{a\in\mathcal{A}} \left(\left| \boldsymbol{u}_s^T e_a e_a^T \boldsymbol{u}_s \right| + \left| \boldsymbol{u}_s^T e_a \pi(\cdot|s)^T \boldsymbol{u}_s \right| + \left| \boldsymbol{u}_s^T \pi(\cdot|s) e_a^T \boldsymbol{u}_s \right| \right. \\
&\quad \left. + 2 \left| u_s^\top \pi(\cdot|s) \pi(\cdot|s)^\top u_s \right| + \left| u_s^\top \text{diag}(\pi(\cdot|s)) u_s \right| \right) \\
&\leqslant 6.
\end{aligned} \tag{B.12}$$

Similarly,

$$\begin{aligned}
\sum_{a\in\mathcal{A}} \left| \frac{d^3\pi_\alpha(a|s)}{d\alpha^3}\Big|_{\alpha=0} \right| &\leqslant \sum_{a\in\mathcal{A}} \sum_{a'\in\mathcal{A}} \left| \boldsymbol{u}_{a'} \boldsymbol{u}^T \nabla^3_{\theta+\alpha\boldsymbol{u}} \pi_\alpha(a|s)\big|_{\alpha=0} \boldsymbol{u} \right| \\
&\leqslant 26
\end{aligned} \tag{B.13}$$

Then we can use Lemma B.3 with $C' = 2, C'' = 6, C''' = 26$, and get

$$\begin{aligned}
\max_{||\boldsymbol{u}||_2=1} \left| \frac{d\tilde{V}_i(\alpha)}{d\alpha}\Big|_{\alpha=0} \right| &\leqslant \frac{2}{(1-\gamma_i)^2}, \\
\max_{||\boldsymbol{u}||_2=1} \left| \frac{d^2\tilde{V}_i(\alpha)}{d\alpha^2}\Big|_{\alpha=0} \right| &\leqslant \frac{6}{(1-\gamma_i)^2} + \frac{8\gamma_i}{(1-\gamma_i)^3} \leqslant \frac{8}{(1-\gamma_i)^3}, \\
\max_{||\boldsymbol{u}||_2=1} \left| \frac{d^3\tilde{V}_i(\alpha)}{d\alpha^3}\Big|_{\alpha=0} . \right| &\leqslant \frac{26}{(1-\gamma_i)^2} + \frac{72\gamma_i}{(1-\gamma_i)^3} + \frac{48\gamma_i^2}{(1-\gamma_i)^4} \leqslant \frac{48}{(1-\gamma_i)^4}
\end{aligned} \tag{B.14}$$

This is equivalent to

$$\begin{aligned}
&||V_i^{\pi_{\theta'}}(\mu) - V_i^{\pi_{\theta''}}(\mu)|| \leqslant \frac{2}{(1-\gamma_i)^2} ||\theta' - \theta''||, \\
&||\nabla V_i^{\pi_{\theta'}}(\mu) - \nabla V_i^{\pi_{\theta''}}(\mu)|| \leqslant \frac{8}{(1-\gamma_i)^3} ||\theta' - \theta''||, \text{ and} \\
&||\nabla^2 V_i^{\pi_{\theta'}}(\mu) - \nabla^2 V_i^{\pi_{\theta''}}(\mu)|| \leqslant \frac{48}{(1-\gamma_i)^4} ||\theta' - \theta''||.
\end{aligned} \tag{B.15}$$

B.2.2 Proof of Lemma B.2

Define

$$\zeta(\theta) \triangleq -\lambda \text{RE}(\pi_\theta) = \frac{\lambda}{|\mathcal{S}||\mathcal{A}|} \sum_{s,a} \log \pi_\theta(a|s). \tag{B.16}$$

We have

$$\begin{aligned}
\nabla_{\theta_s} \zeta(\theta) &= \frac{\lambda}{|\mathcal{S}|}(\frac{1}{|\mathcal{A}|}\mathbf{1} - \pi_\theta(\cdot|s)), \\
\nabla^2_{\theta_s} \zeta(\theta) &= \frac{\lambda}{|\mathcal{S}|}(-\text{diag}(\pi_\theta(\cdot|s)) + \pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T), \\
\frac{\partial}{\partial \theta_{s,a'}} \nabla^2_{\theta_s} \zeta(\theta) &= \frac{\lambda}{|\mathcal{S}|}(-\pi_\theta(a'|s) e_{a'} e_{a'}^T + \pi_\theta(a'|s)\text{diag}(\pi_\theta(\cdot|s)) \\
&\qquad + 2\pi_\theta(a'|s)\pi_\theta(\cdot|s)e_{a'}^T - 2\pi_\theta(a'|s)\pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T).
\end{aligned} \tag{B.17}$$

Now we can bound the norm of the gradient, the norm of the Hessian, and the norm of the third level gradient.

$$\begin{aligned}
||\nabla_\theta \zeta(\theta)|| &= \sum_s ||\nabla_{\theta_s} \zeta(\theta)|| \\
&\leqslant \frac{\lambda}{|\mathcal{S}|} \sum_s ||\frac{1}{|\mathcal{A}|}\mathbf{1} - \pi_\theta(\cdot|s)|| \\
&\leqslant \frac{\lambda}{|\mathcal{S}|} \sum_s \left(||\frac{1}{|\mathcal{A}|}\mathbf{1}|| + ||\pi_\theta(\cdot|s)|| \right) \\
&\leqslant \frac{\lambda}{|\mathcal{S}|} \sum_s \left(\frac{1}{\sqrt{|\mathcal{A}|}} + 1 \right) \\
&\leqslant \lambda(\frac{1}{\sqrt{|\mathcal{A}|}} + 1).
\end{aligned} \tag{B.18}$$

For any vector $u \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ with $||u||_2 = 1$,

$$\left| u^T \nabla^2_\theta \zeta(\theta) u \right| = \left| \sum_s u_s^T \nabla^2_{\theta_s} \zeta(\theta) u_s \right|$$

$$
\begin{aligned}
&\leqslant \frac{\lambda}{|\mathcal{S}|}\sum_s \left|u_s^T \text{diag}(\pi_\theta(\cdot|s))u_s - u_s^T\pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T u_s\right| \\
&\leqslant \frac{2\lambda}{|\mathcal{S}|}\sum_s ||u_s||_\infty^2 \\
&\leqslant \frac{2\lambda}{|\mathcal{S}|}||u||_2^2 \\
&\leqslant \frac{2\lambda}{|\mathcal{S}|}, \qquad \text{(B.19)}
\end{aligned}
$$

where the first equality follows since $\nabla_{\theta_{s'}}\nabla_{\theta_{s''}}\zeta(\theta) = 0,\ \forall s' \neq s''$. Using this method, we can further get

$$
\begin{aligned}
\left|\sum_{s',a'} u_{s',a'} u^T \nabla_\theta^2 \zeta(\theta) u\right| &= \left|\sum_s \sum_{a'} u_{s,a'} u_s^T \nabla_{\theta_s}^2 \zeta(\theta) u_s\right| \\
&\leqslant \frac{\lambda}{|\mathcal{S}|}\sum_s \Bigg| -\sum_{a'} u_{s,a'} u_s^T \pi_\theta(a'|s) e_{a'} e_{a'}^T u_s \\
&\qquad + \sum_{a'} u_{s,a'} u_s^T \pi_\theta(a'|s) \text{diag}(\pi_\theta(\cdot|s)) u_s \\
&\qquad + 2\sum_{a'} u_{s,a'} u_s^T \pi_\theta(a'|s)\pi_\theta(\cdot|s) e_{a'}^T u_s \\
&\qquad - 2\sum_{a'} u_{s,a'} u_s^T \pi_\theta(a'|s)\pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T u_s \Bigg| \\
&\leqslant \frac{6\lambda}{|\mathcal{S}|}\sum_s ||u_s||_\infty^3 \\
&\leqslant \frac{6\lambda}{|\mathcal{S}|},
\end{aligned}
$$

where the last inequality follows from $||u||_\infty \leqslant ||u||_2 = 1$. This implies that $\zeta(\theta)$ is $\lambda(\frac{1}{\sqrt{|\mathcal{A}|}} + 1)$-Lipschitz, $\frac{2\lambda}{|\mathcal{S}|}$-smooth, and has $\frac{6\lambda}{|\mathcal{S}|}$-Lipschitz Hessian.

□

B.3 Proof of Theorems

In this section, we provide complete analysis for the results stated in the main paper. We first introduce the following notations.

$$\boldsymbol{\theta} \triangleq \left[\theta_1^T, \theta_2^T, ..., \theta_N^T\right]^T \in \mathbb{R}^{N|\mathcal{S}||\mathcal{A}|}, \qquad \boldsymbol{V}(\boldsymbol{\theta}; \boldsymbol{\rho}) \triangleq \begin{pmatrix} V_1^{\pi_{\theta_1}}(\rho_1) \\ V_2^{\pi_{\theta_2}}(\rho_2) \\ \vdots \\ V_N^{\pi_{\theta_N}}(\rho_N) \end{pmatrix} \in \mathbb{R}^N, \tag{B.20}$$

$$\boldsymbol{\rho} = [\rho_1^T, \rho_2^T, ..., \rho_N^T]^T, \quad \boldsymbol{\mu} = [\mu_1^T, \mu_2^T, ..., \mu_N^T]^T, \quad \overline{\nabla \boldsymbol{V}}(\boldsymbol{\theta}; \boldsymbol{\rho}) = \frac{1}{N}\sum_{i=1}^N \nabla_{\theta_i} V_i^{\pi_{\theta_i}}(\rho_i).$$

B.3.1 Proof of Theorem 3.1

Define $D = 2N\lambda + \sum_{i=1}^N \frac{1}{(1-\gamma_i)^2}$. In the proof, we will need the following lemmas. The proof of Lemma B.4 is in Subsection B.5.2. Lemma B.5 is a standard result and its proof can be found in the existing literature such as [21].

Lemma B.4. *For all k and $\boldsymbol{\mu}$, $||\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k; \boldsymbol{\mu})|| \leqslant D$.*

Lemma B.5. *Let $\bar{\theta}^k = \frac{1}{N}\sum_{i=1}^N \theta_i^k$. If each agent starts with the same initialization, i.e. $\theta_1^0 = \theta_2^0 = ... = \theta_N^0$, then*

$$||\theta_i^k - \bar{\theta}^k|| \leqslant \frac{\alpha D}{1-\sigma_2}, \quad \forall i, k.$$

We made the assumption in Theorem 3.1 that the agents start with the same initialization. We denote $\theta^0 = \theta_i^0$, $\forall i$.

We define the Lyapunov function

$$\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}; \boldsymbol{\mu}) \triangleq -\mathbf{1}^T \boldsymbol{L}^\lambda(\boldsymbol{\theta}; \boldsymbol{\mu}) + \frac{1}{2\alpha}||\boldsymbol{\theta}||^2_{I-W}, \tag{B.21}$$

where $||\boldsymbol{\theta}||^2_{I-W} \triangleq \boldsymbol{\theta}^T((I-W)\otimes I)\boldsymbol{\theta}$.

Note that the sequence $\{\boldsymbol{\theta}^k\}$ generated by the distributed policy gradient algorithm is the same as the sequence generated by applying gradient descent on $\xi_{\alpha,\lambda}(\boldsymbol{\theta})$, if both algorithms use fixed step size α. This can be observed by re-writing the update equation Equation 3.6.

$$\begin{aligned}
\boldsymbol{\theta}^{k+1} &= (W\otimes I)\boldsymbol{\theta}^k + \alpha\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu}) \\
&= \boldsymbol{\theta}^k + \alpha\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu}) - ((I-W)\otimes I)\boldsymbol{\theta}^k \\
&= \boldsymbol{\theta}^k - \alpha(-\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu}) + \frac{1}{\alpha}((I-W)\otimes I)\boldsymbol{\theta}^k) \\
&= \boldsymbol{\theta}^k - \alpha\nabla \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) \qquad \text{(B.22)}
\end{aligned}$$

We have to establish the smoothness constant of $\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu})$. Combining Lemma B.1 and Lemma B.2, $L_i^\lambda(\theta_i)$ is β_i^λ-smooth with

$$\beta_i^\lambda = \frac{8}{(1-\gamma_i)^3} + \frac{2\lambda}{|\mathcal{S}|},$$

which implies $\sum_{i=1}^N L_i^\lambda(\theta_i)$ is β^λ-smooth, where

$$\beta^\lambda = \sum_{i=1}^N \left(\frac{8}{(1-\gamma_i)^3} + \frac{2\lambda}{|\mathcal{S}|}\right). \qquad \text{(B.23)}$$

In addition, we know $\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu})$ is $\beta^{\boldsymbol{\xi}_{\alpha,\lambda}}$-smooth, with

$$\beta^{\boldsymbol{\xi}_{\alpha,\lambda}} = \beta^\lambda + \frac{1}{\alpha}\sigma_{\max}(I-W) = \beta^\lambda + \alpha^{-1}(1-\sigma_N). \qquad \text{(B.24)}$$

By the $\beta^{\boldsymbol{\xi}_{\alpha,\lambda}}$-smoothness of $\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta})$, we have

$$\begin{aligned}
\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^{k+1};\boldsymbol{\mu}) &\leqslant \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) + \langle\nabla\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}), \boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k\rangle + \frac{\beta^{\xi_{\alpha,\lambda}}}{2}||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2 \\
&= \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) + \langle -\frac{\boldsymbol{\theta}_{k+1}-\boldsymbol{\theta}_k}{\alpha}, \boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k\rangle + \frac{\beta^{\boldsymbol{\xi}_{\alpha,\lambda}}}{2}||\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^k||^2
\end{aligned}$$

$$= \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) + (\frac{\beta^{\boldsymbol{\xi}_{\alpha,\lambda}}}{2} - \frac{1}{\alpha})||\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k||^2$$
$$= \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) - \frac{1}{2}(\alpha^{-1}(1+\sigma_N) - \beta^{\lambda})||\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k||^2$$

Since $\alpha \leqslant \frac{1+\sigma_N}{2\sum_{i=1}^{N}(\frac{8}{(1-\gamma_i)^3} + \frac{2\lambda}{|\mathcal{S}|})} = \frac{1+\sigma_N}{2\beta^{\lambda}}$, we know $\frac{1}{2}(\alpha^{-1}(1+\sigma_N) - \beta^{\lambda}) \geqslant 0$, $\forall k$. This implies $\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})$ is a non-increasing sequence. Let $\tilde{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \xi_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu})$. We have

$$\begin{aligned}\sum_{k=0}^{K-1} ||\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k||^2 &\leqslant \sum_{k=0}^{K-1} 2(\alpha^{-1}(1+\sigma_N) - \beta^{\lambda})^{-1}(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^{k+1};\boldsymbol{\mu})) \\ &= c_1(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^{K-1};\boldsymbol{\mu})) \\ &\leqslant c_1(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\tilde{\boldsymbol{\theta}};\boldsymbol{\mu})),\end{aligned}$$

where we define $c_1 = 2(\alpha^{-1}(1+\sigma_N) - \beta^{\lambda})^{-1}$.

This implies

$$\min_{k<K} ||\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k||^2 \leqslant \frac{c_1}{K}(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\tilde{\boldsymbol{\theta}};\boldsymbol{\mu})).$$

From Equation B.22, $||\alpha\nabla\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 = ||\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k||^2$. Thus,

$$\min_{k<K} ||\nabla\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2 = \frac{1}{\alpha^2}\min_{k<K} ||\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k||^2 \leqslant \frac{c_1}{K\alpha^2}(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0;\boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\tilde{\boldsymbol{\theta}};\boldsymbol{\mu})). \tag{B.25}$$

Taking derivative of Equation B.21,

$$\nabla\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}) = -\nabla\boldsymbol{L}^{\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}) + \frac{1}{\alpha}((I-W)\otimes I)\boldsymbol{\theta},$$

Observe that $\mathbf{1}^T(I-W) = \mathbf{0}$ due to the double stochasticity of W, which leads to

$$\overline{\nabla\boldsymbol{\xi}}_{\alpha,\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}) = -\overline{\nabla\boldsymbol{L}}^{\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}) + \frac{1}{N\alpha}(\mathbf{1}^T(I-W)\otimes I)\boldsymbol{\theta} = -\overline{\nabla\boldsymbol{L}}^{\lambda}(\boldsymbol{\theta};\boldsymbol{\mu}).$$

Now we can bound the gradient $\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k; \boldsymbol{\mu})$.

$$
\begin{aligned}
&\min_{k<K} ||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k; \boldsymbol{\mu})||^2 \\
&= \min_{k<K} ||\overline{\nabla \boldsymbol{\xi}}_{\alpha,\lambda}(\boldsymbol{\theta}^k; \boldsymbol{\mu})||^2 \\
&\leqslant \min_{k<K} ||\nabla \boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^k; \boldsymbol{\mu})||^2 \\
&\leqslant \frac{c_1}{K\alpha^2}(\boldsymbol{\xi}_{\alpha,\lambda}(\boldsymbol{\theta}^0; \boldsymbol{\mu}) - \boldsymbol{\xi}_{\alpha,\lambda}(\tilde{\boldsymbol{\theta}}; \boldsymbol{\mu})) \\
&= \frac{c_1}{K\alpha^2}(-\sum_{i=1}^{N} L_i^{\lambda}(\boldsymbol{\theta}^0; \mu_i) + \frac{1}{2\alpha}||\boldsymbol{\theta}^0||_{I-W}^2 + \sum_{i=1}^{N} L_i^{\lambda}(\tilde{\boldsymbol{\theta}}; \mu_i) - \frac{1}{2\alpha}||\tilde{\boldsymbol{\theta}}||_{I-W}^2) \\
&\leqslant \frac{c_1}{K\alpha^2} \sum_{i=1}^{N} (L_i^{\lambda}(\tilde{\boldsymbol{\theta}}; \mu_i) - L_i^{\lambda}(\boldsymbol{\theta}^0; \mu_i)) \\
&\leqslant \frac{c_1}{K\alpha^2} \sum_{i=1}^{N} (V_i^{\pi_{\tilde{\boldsymbol{\theta}}_i}}(\mu_i) - V_i^{\pi_{\theta_i^0}}(\mu_i) + \lambda \text{RE}(\pi_{\theta_i^0})) \\
&\leqslant \frac{c_1}{K\alpha^2} \sum_{i=1}^{N} (\frac{1}{1-\gamma_i} + \lambda \text{RE}(\pi_{\theta^0})).
\end{aligned} \tag{B.26}
$$

The third line comes from Equation B.25. The fifth line uses our assumption that all agents start with the same parameter initialization, making $||\boldsymbol{\theta}^0||_{I-W}^2 = 0$.The second last inequality is from the fact that relative entropy is non-negative. The last inequality comes from the bounded value function in Equation 3.3.

This implies

$$
\begin{aligned}
\min_{k<K} ||\overline{\nabla V}(\boldsymbol{\theta}^k; \boldsymbol{\mu})||^2 &= \min_{k<K} ||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k; \boldsymbol{\mu}) + \frac{\lambda}{N} \sum_{i=1}^{N} \nabla \text{RE}(\pi_{\theta_i^k})||^2 \\
&\leqslant 2 \min_{k<K} ||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k; \boldsymbol{\mu})||^2 + \frac{2}{N} \sum_{i=1}^{N} ||\nabla \lambda \text{RE}(\pi_{\theta_i^k})||^2.
\end{aligned}
$$

The second term depends on the smoothness of the regularizer, which we establish in Lemma B.2. The first term is bounded in Equation B.26. Therefore,

$$
\min_{k<K} ||\overline{\nabla V}(\boldsymbol{\theta}^k; \boldsymbol{\mu})||^2 \leqslant 2 \min_{k<K} ||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k; \boldsymbol{\mu})||^2 + \frac{2}{N} \sum_{i=1}^{N} ||\nabla \lambda \text{RE}(\pi_{\theta_i^k})||^2
$$

$$\leqslant \frac{2c_1}{K\alpha^2}\sum_{i=1}^{N}(\frac{1}{1-\gamma_i}+\lambda \text{RE}(\pi_{\theta^0}))+\frac{2}{N}\left(\frac{\lambda}{\sqrt{|\mathcal{A}|}}+\lambda\right)^2$$
$$\leqslant \frac{2c_1}{K\alpha^2}\sum_{i=1}^{N}(\frac{1}{1-\gamma_i}+\lambda \text{RE}(\pi_{\theta^0}))+\frac{8\lambda^2}{N} \tag{B.27}$$

Using the smoothness of V_i, which we show in Lemma B.1, we have

$$\begin{aligned}
&\min_{k<K}||\frac{1}{N}\sum_{j=1}^{N}\nabla V_j(\theta_i^k;\mu_j)||^2 \\
&=\min_{k<K}||\frac{1}{N}\sum_{j=1}^{N}\nabla V_j(\theta_j^k;\mu_j)-\left(\nabla V_j(\theta_j^k;\mu_j)-\nabla V_j(\theta_i^k;\mu_j)\right)||^2 \\
&\leqslant \min_{k<K}2||\frac{1}{N}\sum_{j=1}^{N}\nabla V_j(\theta_j^k;\mu_j)||^2+\frac{2}{N}\sum_{j=1}^{N}||\nabla V_j(\theta_j^k;\mu_j)-\nabla V_j(\theta_i^k;\mu_j)||^2 \\
&\leqslant 2\min_{k<K}||\overline{\nabla V}(\boldsymbol{\theta}^k;\boldsymbol{\mu})||^2+\frac{2}{N}\sum_{j=1}^{N}\frac{64}{(1-\gamma_j)^6}||\theta_i^k-\theta_j^k||^2.
\end{aligned} \tag{B.28}$$

From Lemma B.5, we have

$$\begin{aligned}
||\theta_i^k-\theta_j^k|| &= ||(\theta_i^k-\bar{\theta}^k)-(\bar{\theta}^k-\theta_j^k)|| \\
&\leqslant ||\theta_i^k-\bar{\theta}^k||+||\theta_j^k-\bar{\theta}^k|| \\
&\leqslant \frac{2\alpha D}{1-\sigma_2}.
\end{aligned}$$

Plugging this inequality and Equation B.27 into Equation B.28, we get

$$\begin{aligned}
&\min_{k<K}||\frac{1}{N}\sum_{j=1}^{N}\nabla V_j(\theta_i^k;\mu_j)||^2 \\
&\quad\leqslant \frac{4c_1}{K\alpha^2}\sum_{j=1}^{N}(\frac{1}{1-\gamma_j}+\lambda \text{RE}(\pi_{\theta^0}))+\frac{16\lambda^2}{N}+\frac{2}{N}\sum_{j=1}^{N}\frac{64}{(1-\gamma_j)^6}\frac{4\alpha^2D^2}{(1-\sigma_2)^2} \\
&\quad\leqslant \frac{16}{K\alpha}\sum_{j=1}^{N}\left(\frac{1}{1-\gamma_j}+\lambda \text{RE}(\pi_{\theta^0})\right)+\frac{16\lambda^2}{N}+\sum_{j=1}^{N}\frac{512D^2\alpha^2}{N(1-\sigma_2)(1-\gamma_j)^6}.
\end{aligned}$$

The proof is completed by recognizing $\rho_i=\mu_i,\ \forall i$.

□

B.3.2 Proof of Theorem 3.2

When condition in Equation 3.8 is observed, we can establish the global optimality condition under the tabular policy.

Proposition B.1. *Let* $\theta^* = \max_\theta V(\theta; \boldsymbol{\rho})$. *For policy parameter* θ, *if* $|| \sum_{i=1}^N \nabla L_i^\lambda(\theta; \mu_i)|| \leqslant \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|}$, *we have*

$$V(\theta^*; \boldsymbol{\rho}) - V(\theta; \boldsymbol{\rho}) \leqslant 2\lambda N \max_{s\in\mathcal{S}, i: s\in\mathcal{S}_i} \{\frac{d_{\rho_i}^{\pi_{\theta^\star}}(s)}{(1-\gamma_i)\mu_i(s)}\}$$

if the environment and the initial state distributions $\boldsymbol{\rho}$ *and* $\boldsymbol{\mu}$ *jointly satisfies the discounted visitation match assumption.*

The proof of this proposition is in Subsection B.4.1. Using the proposition, we can guarantee that θ_i^k is an ϵ-optimal solution in the objective function by setting $\epsilon = 2N\lambda \max_{j,s}\{\frac{d_{\rho_j}^{\pi_{\theta^\star}}(s)}{(1-\gamma_j)\mu_j(s)}\}$ and ensuring $|| \sum_{j=1}^N \nabla L_j^\lambda(\theta_i^k; \mu_j)|| \leqslant \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|}$. Denoting

$$c_2 = \frac{1}{\max_{j,s}\{\frac{d_{\rho_j}^{\pi_{\theta^\star}}(s)}{(1-\gamma_j)\mu_j(s)}\}}$$

and solving for λ in terms of ϵ, we get

$$\lambda = \frac{\epsilon}{2N \max_{j,s}\{\frac{d_{\rho_j}^{\pi_{\theta^\star}}(s)}{(1-\gamma_j)\mu_j(s)}\}} = \frac{\epsilon c_2}{2N}.$$

Now we bound the norm of the gradient.

$$\begin{aligned}
\min_{k<K} || \sum_{j=1}^N \nabla L_j^\lambda(\theta_i^k; \mu_j)|| &= \min_{k<K} || \sum_{j=1}^N \nabla L_j^\lambda(\theta_j^k; \mu_j) + \sum_{j=1}^N \left(\nabla L_j^\lambda(\theta_i^k; \mu_j) - \nabla L_j^\lambda(\theta_j^k)\right) || \\
&\leqslant \min_{k<K} ||N\overline{\nabla \boldsymbol{L}}^\lambda(\boldsymbol{\theta}^k; \boldsymbol{\mu})|| + \sum_{j=1}^N ||\nabla L_j^\lambda(\theta_i^k; \mu_j) - \nabla L_j^\lambda(\theta_j^k; \mu_j)||
\end{aligned}$$

$$\leqslant N \min_{k<K} ||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k; \boldsymbol{\mu})|| + \sum_{j=1}^{N} \beta_i^{\lambda} ||\theta_i^k - \theta_j^k||, \tag{B.29}$$

where the last inequality uses the smoothness property of L_i^{λ}. Combining Lemma B.1 and Lemma B.2, $\beta_i^{\lambda} = \frac{8}{(1-\gamma_i)^3} + \frac{2\lambda}{|\mathcal{S}|}$. We have a bound on the first term in Equation B.26, and now we bound the second term using Lemma B.5.

$$\begin{aligned}
||\theta_i^k - \theta_j^k|| &= ||(\theta_i^k - \bar{\theta}^k) - (\bar{\theta}^k - \theta_j^k)|| \\
&\leqslant ||\theta_i^k - \bar{\theta}^k|| + ||\theta_j^k - \bar{\theta}^k|| \\
&\leqslant \frac{2\alpha D}{1-\sigma_2}
\end{aligned}$$

Plug this into Equation B.29,

$$\begin{aligned}
\min_{k<K} ||\sum_{j=1}^{N} \nabla L_j^{\lambda}(\theta_i^k; \mu_j)|| &\leqslant N \min_{k<K} ||\overline{\nabla \boldsymbol{L}}^{\lambda}(\boldsymbol{\theta}^k; \boldsymbol{\mu})|| + \sum_{j=1}^{N} \beta_i^{\lambda} ||\theta_i^k - \theta_j^k|| \\
&\leqslant N \sqrt{\frac{c_1}{K\alpha^2} \sum_{j=1}^{N} (\frac{1}{1-\gamma_j} + \lambda \text{RE}(\pi_{\theta^0}))} + \sum_{j=1}^{N} \beta_i^{\lambda} \frac{2\alpha D}{1-\sigma_2} \\
&\leqslant N \sqrt{\frac{c_1}{K\alpha^2} \sum_{j=1}^{N} (\frac{1}{1-\gamma_j} + \lambda \text{RE}(\pi_{\theta^0}))} + \frac{2\alpha\beta^{\lambda} D}{1-\sigma_2}
\end{aligned}$$

To ensure $\min_{k<K} ||\sum_{j=1}^{N} \nabla L_j^{\lambda}(\theta_i^k; \mu_j)|| \leqslant \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|}$, we make

$$N \sqrt{\frac{c_1}{K\alpha^2} \sum_{j=1}^{N} (\frac{1}{1-\gamma_j} + \lambda \text{RE}(\pi_{\theta^0}))} + \frac{2\alpha\beta^{\lambda} D}{1-\sigma_2} \leqslant \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|}$$

Solving for K, we get

$$K \geqslant \frac{c_1 N^2 \left(\sum_{j=1}^{N} (\frac{1}{1-\gamma_j} + \lambda \text{RE}(\pi_{\theta^0})) \right)}{\alpha^2 \left(\frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} - \frac{2\alpha\beta^{\lambda} D}{1-\sigma_2} \right)^2}$$

$$= \frac{c_1 N^2 \left(\sum_{j=1}^{N} (\frac{1}{1-\gamma_j} + \lambda \text{RE}(\pi_{\theta^0})) \right)}{\alpha^2 \left(\frac{\epsilon c_2}{4|\mathcal{S}||\mathcal{A}|} - \frac{2\alpha D}{1-\sigma_2} \sum_{j=1}^{N} \left(\frac{8}{(1-\gamma_j)^3} + \frac{\epsilon c_2}{N|\mathcal{S}|} \right) \right)^2},$$

where we used the fact that $\frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} - \frac{2\alpha\beta^\lambda D}{1-\sigma_2} > 0$, if $\alpha < \frac{\lambda N(1-\sigma_2)}{4\beta^\lambda D|\mathcal{S}||\mathcal{A}|}$.

□

B.3.3 Proof of Theorem 3.3

We denote

$$\begin{aligned}
&\overline{\theta}^{k+1} = \frac{1}{N}\sum_{i=1}^{N} \theta_i^{k+1}, \quad \overline{\pi}^{k+1} = \frac{\exp\left(\overline{\theta}^{k+1}(s,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\overline{\theta}^{k+1}(s,a')\right)}, \quad \boldsymbol{\pi}^k = [\pi_1^k, \ldots, \pi_N^k], \\
&Q_g^{\boldsymbol{\pi}^k} = [Q_1^{\pi_1^k}, \ldots, Q_N^{\pi_N^k}], \quad V_g^{\boldsymbol{\pi}^k} = [V_1^{\pi_1^k}, \ldots, V_N^{\pi_N^k}], \\
&Q_{L,k}^{\boldsymbol{\pi}^k} = \sum_{i=1}^{N} (\frac{1}{N} + \lambda_i^k - \nu_i^k) Q_i^{\pi_i^k}, \quad V_{L,k}^{\boldsymbol{\pi}^k} = \sum_{i=1}^{N} (\frac{1}{N} + \lambda_i^k - \nu_i^k) V_i^{\pi_i^k}.
\end{aligned} \tag{B.30}$$

Our analysis relies on the lemmas below. The first lemma establishes the Lipschitz continuity of the value function and Q function. The second lemma bounds the consensus error. The last three lemmas establish some technical immediate convergence results.

Lemma B.6 (Lemma 8 of [156])**.** *For any policy π_1, π_2 and $i = 1, \ldots, N$*

$$\|Q_i^{\pi_1} - Q_i^{\pi_2}\| \leqslant \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}\|\pi_1 - \pi_2\|, \quad \|V_i^{\pi_1} - V_i^{\pi_2}\| \leqslant \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}\|\pi_1 - \pi_2\|,$$
$$|Q_i^{\pi_1}(s,a) - Q_i^{\pi_2}(s,a)| \leqslant \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2}\|\pi_1 - \pi_2\|, \quad |V_i^{\pi_1}(s) - V_i^{\pi_2}(s)| \leqslant \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2}\|\pi_1 - \pi_2\|.$$

Lemma B.7. *The policy iterates $\{\pi_i^k\}$ generated by Algorithm Algorithm 3.2 satisfy*

$$\|\overline{\pi}^k - \pi_i^k\| \leqslant \mathcal{O}\left(\frac{\sqrt{N}\alpha}{1-\sigma_2(W)}\right), \quad \textit{for all } k = 0, \ldots, K-1 \textit{ and } i = 1, \ldots, N.$$

Lemma B.8. *The iterates of Algorithm 3.2 satisfy for all $k = 0, \ldots, K-1$*

$$V_{L,k}^{\bar{\pi}^{k+1}}(\zeta) - V_{L,k}^{\bar{\pi}^k}(\zeta)$$
$$\geqslant \frac{N}{\alpha}\mathbb{E}_{s\sim\zeta}\left[\log Z_k(s) - \frac{\alpha}{N}V_{L,k}^{\bar{\pi}^k}(s)\right] - \frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^3}\sum_{i=1}^{N}\|\bar{\pi}^k - \pi_i^k\|.$$

Lemma B.9. *The iterates of Algorithm 3.2 satisfy for all $k = 0, \ldots, K-1$*

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_{L,k}^{\pi^\star}(\rho) - V_{L,k}^{\bar{\pi}^k}(\rho)\right) \leqslant \frac{N\log|\mathcal{A}|}{(1-\gamma)K\alpha} + \frac{2NB_\lambda}{(1-\gamma)^2K} + \frac{4N\eta}{(1-\gamma)^3K}$$
$$+ \frac{3\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^4K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\bar{\pi}^k - \pi_i^k\|.$$

Lemma B.10 (Theorem 6 of [112])**.** *Suppose that Assumption 3.3 holds. Let the constant C obey $C \geqslant 2\|\lambda^\star\|_\infty$ and $C \geqslant 2\|\nu^\star\|_\infty$. Then, given a policy π, if there exists a constant $\delta > 0$ such that*

$$V_0^{\pi^\star}(\rho) - V_0^\pi(\rho) + C\sum_{i=1}^{N}\left([\ell_i - V_i^\pi(\rho)]_+ + [V_i^\pi(\rho) - u_i]_+\right) \leqslant \delta,$$

then we have

$$\sum_{i=1}^{N}\left([\ell_i - V_i^\pi(\rho)]_+ + [V_i^\pi(\rho) - u_i]_+\right) \leqslant \frac{2\delta}{C}.$$

We have from Equation 3.17

$$\bar{\pi}^{k+1}(a \mid s) = \bar{\pi}^k(a \mid s)\frac{\exp(\frac{\alpha}{N}\sum_{i=1}^{N}(\frac{1}{N} + \lambda_i^k - \nu_i^k)Q_i^{\pi_i^k}(s,a))}{Z^k(s)}, \tag{B.31}$$

where $Z^k(s) = \sum_{a'\in\mathcal{A}}\bar{\pi}^k(a' \mid s)\exp(\frac{\alpha}{N}\sum_{i=1}^{N}(\frac{1}{N} + \lambda_i^k - \nu_i^k)Q_i^{\pi_i^k}(s,a'))$.

Objective function convergence. From the dual update Equation 3.18, we have

$$
\begin{aligned}
0 \leqslant \|\lambda^K\|^2 &= \sum_{k=0}^{K-1} \left(\|\lambda^{k+1}\|^2 - \|\lambda^k\|^2\right) \\
&= \sum_{k=0}^{K-1} \left(\left\| \Pi_{[0,B_\lambda]} \left(\lambda^k - \eta \left(\sum_{s,a} \rho(s) \operatorname{diag}(\boldsymbol{\pi}^k(a \mid s)) Q_g^{\boldsymbol{\pi}^k}(s,a) - \ell \right) \right) \right\|^2 - \|\lambda^k\|^2 \right) \\
&\leqslant \sum_{k=0}^{K-1} \left(\left\| \lambda^k - \eta \left(\sum_{s,a} \rho(s) \operatorname{diag}(\boldsymbol{\pi}^k(a \mid s)) Q_g^{\boldsymbol{\pi}^k}(s,a) - \ell \right) \right\|^2 - \|\lambda^k\|^2 \right) \\
&= -2\eta \sum_{k=0}^{K-1} (\lambda^k)^\top \left(\sum_{s,a} \rho(s) \operatorname{diag}(\boldsymbol{\pi}^k(a \mid s)) Q_g^{\boldsymbol{\pi}^k}(s,a) - \ell \right) \\
&\quad + \eta^2 \sum_{k=0}^{K-1} \left\| \sum_{s,a} \rho(s) \operatorname{diag}(\boldsymbol{\pi}^k(a \mid s)) Q_g^{\boldsymbol{\pi}^k}(s,a) - \ell \right\|^2 \\
&= -2\eta \sum_{k=0}^{K-1} (\lambda^k)^\top \left(V_g^{\boldsymbol{\pi}^k}(\rho) - \ell \right) \\
&\quad + \eta^2 \sum_{k=0}^{K-1} \left\| \sum_{s,a} \rho(s) \operatorname{diag}(\boldsymbol{\pi}^k(a \mid s)) Q_g^{\boldsymbol{\pi}^k}(s,a) - \ell \right\|^2 .
\end{aligned} \tag{B.32}
$$

Since the value function and constant ℓ_i are within $[0, \frac{1}{1-\gamma}]$, the second term of Equation B.32 obeys

$$
\begin{aligned}
&\sum_{k=0}^{K-1} \left\| \sum_{s,a} \rho(s) \operatorname{diag}(\boldsymbol{\pi}^k(a \mid s)) Q_g^{\boldsymbol{\pi}^k}(s,a) - \ell \right\|^2 \\
&= \sum_{k=0}^{K-1} \sum_{i=1}^{N} \left(\sum_{s,a} \rho(s) \pi_i^k(a \mid s) Q_i^{\pi_i^k}(s,a) - \ell_i \right)^2 \\
&\leqslant 2 \sum_{k=0}^{K-1} \sum_{i=1}^{N} \left(\left(\sum_{s,a} \rho(s) \pi_i^k(a \mid s) Q_i^{\pi_i^k}(s,a) \right)^2 + (\ell_i)^2 \right) \\
&\leqslant \frac{4KN}{(1-\gamma)^2}.
\end{aligned} \tag{B.33}
$$

Equation B.32 and Equation B.33 imply

$$
\begin{aligned}
0 &\leqslant -2\eta \sum_{k=0}^{K-1} (\lambda^k)^\top \left(V_g^{\boldsymbol{\pi}^k}(\rho) - \ell \right) + \frac{4KN\eta^2}{(1-\gamma)^2} \\
&\leqslant 2\eta \sum_{k=0}^{K-1} (\lambda^k)^\top \left(V_g^{\pi^\star}(\rho) - V_g^{\bar{\pi}^k}(\rho) \right) + 2\eta \sum_{k=0}^{K-1} (\lambda^k)^\top \left(V_g^{\bar{\pi}^k}(\rho) - V_g^{\boldsymbol{\pi}^k}(\rho) \right) + \frac{4KN\eta^2}{(1-\gamma)^2} \\
&\leqslant 2\eta \sum_{k=0}^{K-1} (\lambda^k)^\top \left(V_g^{\pi^\star}(\rho) - V_g^{\bar{\pi}^k}(\rho) \right) + \frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}B_\lambda \eta}{(1-\gamma)^2} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\| + \frac{4KN\eta^2}{(1-\gamma)^2},
\end{aligned}
$$

where the second inequality follows from the fact that the optimal policy satisfies the constraints, i.e. $V_i^{\pi^\star}(\rho) \geqslant \ell_i$ for all $i = 1, \cdots, N$, and the third inequality is applies Lemma B.6.

Re-arranging this inequality and dividing by $2K\eta$ lead to

$$
\frac{1}{K} \sum_{k=0}^{K-1} (\lambda^k)^\top \left(V_g^{\pi^\star}(\rho) - V_g^{\bar{\pi}^k}(\rho) \right) \geqslant -\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}B_\lambda}{(1-\gamma)^2 K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\| - \frac{2N\eta}{(1-\gamma)^2}. \tag{B.34}
$$

A similar analysis on ν^k implies

$$
-\frac{1}{K} \sum_{k=0}^{K-1} (\nu^k)^\top \left(V_g^{\pi^\star}(\rho) - V_g^{\bar{\pi}^k}(\rho) \right) \geqslant -\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}B_\lambda}{(1-\gamma)^2 K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\| - \frac{2N\eta}{(1-\gamma)^2}. \tag{B.35}
$$

Combining Equation B.34, Equation B.35, and Lemma B.9, we have

$$
\begin{aligned}
&\frac{1}{K} \sum_{k=0}^{K-1} \left(V_0^{\pi^\star}(\rho) - V_0^{\bar{\pi}^k}(\rho) \right) \\
&\leqslant \frac{N \log |\mathcal{A}|}{(1-\gamma)K\alpha} + \frac{2NB_\lambda}{(1-\gamma)^2 K} + \frac{4N\eta}{(1-\gamma)^3 K} + \frac{3\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^4 K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\| \\
&\quad + \frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}B_\lambda}{(1-\gamma)^2 K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\| + \frac{4N\eta}{(1-\gamma)^2}
\end{aligned}
$$

$$\leqslant \frac{N\log|\mathcal{A}|}{(1-\gamma)K\alpha} + \frac{2NB_\lambda}{(1-\gamma)^2K} + \frac{8N\eta}{(1-\gamma)^3} + \frac{5\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda+1/N)}{(1-\gamma)^4K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\overline{\pi}^k - \pi_i^k\|.$$

By the bound on consensus error in Lemma B.7 and the Lipschitz continuity of the value function in Lemma B.6, this implies for any agent $j = 1, \cdots, N$

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(V_0^{\pi^\star}(\rho) - V_0^{\pi_j^k}(\rho)\right) \leqslant \mathcal{O}\left(\frac{N}{K\alpha} + N\eta + \frac{N^{3/2}\alpha}{1-\sigma_2(W)}\right).$$

Constraint violation convergence. For any $\lambda \in [0, B_\lambda]^N$, since the projection operator $\Pi_{[0,B_\lambda]}$ is non-expansive, we have

$$\begin{aligned}
\|\lambda^{k+1} - \lambda\|^2 &= \|\Pi_{[0,B_\lambda]}(\lambda^k - \eta(V_g^{\boldsymbol{\pi}^k}(\rho) - \ell)) - \lambda\|^2 \\
&\leqslant \|\lambda^k - \eta(V_g^k(\rho) - \ell) - \lambda\|^2 \\
&= \|\lambda^k - \lambda\|^2 - 2\eta(\lambda^k - \lambda)^\top(V_g^{\boldsymbol{\pi}^k}(\rho) - \ell) + \eta^2\sum_{i=1}^{N}\|\sum_{s,a}\rho(s)\pi_i^k(a \mid s)Q_i^{\pi_i^k}(\rho) - \ell_i\|^2 \\
&\leqslant \|\lambda^k - \lambda\|^2 - 2\eta(\lambda^k - \lambda)^\top(V_g^{\boldsymbol{\pi}^k}(\rho) - \ell) + \frac{4N\eta^2}{(1-\gamma)^2},
\end{aligned}$$

where the last inequality bounds the quadratic term using an approach similar to Equation B.33.

Re-arranging the terms and summing up from $k = 0$ to $k = K-1$, we get

$$\begin{aligned}
\frac{1}{K}\sum_{k=0}^{K-1}(\lambda^k - \lambda)^\top(V_g^{\boldsymbol{\pi}^k}(\rho) - b) &\leqslant \frac{1}{K}\left(\|\lambda^0 - \lambda\|^2 - \|\lambda^K - \lambda\|^2\right) + \frac{2N\eta}{(1-\gamma)^2} \\
&\leqslant \frac{1}{2K\eta}\|\lambda^0 - \lambda\|^2 + \frac{2N\eta}{(1-\gamma)^2},
\end{aligned}$$

which implies

$$\frac{1}{K}\sum_{k=0}^{K-1}(\lambda^k - \lambda)^\top(V_g^{\bar{\pi}^k}(\rho) - \ell)$$

$$= \frac{1}{K}\sum_{k=0}^{K-1}(\lambda^k-\lambda)^\top(V_g^{\boldsymbol{\pi}^k}(\rho)-\ell)+\frac{1}{K}\sum_{k=0}^{K-1}(\lambda^k-\lambda)^\top(V_g^{\bar{\pi}^k}(\rho)-V_g^{\boldsymbol{\pi}_k}(\rho))$$
$$\leqslant \frac{1}{2K\eta}\|\lambda^0-\lambda\|^2+\frac{2N\eta}{(1-\gamma)^2}+\frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}B_\lambda}{(1-\gamma)^2K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\bar{\pi}^k-\pi_i^k\|. \tag{B.36}$$

Similarly, we can show for any $\nu \in [0, B_\lambda]^N$

$$\frac{1}{K}\sum_{k=0}^{K-1}(\lambda^k-\lambda)^\top(u-V_g^{\bar{\pi}^k}(\rho))$$
$$\leqslant \frac{1}{2K\eta}\|\nu^0-\nu\|^2+\frac{2N\eta}{(1-\gamma)^2}+\frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}B_\lambda}{(1-\gamma)^2K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\bar{\pi}^k-\pi_i^k\|. \tag{B.37}$$

Since λ^k, ν^k are non-negative, we have from Equation B.36, Equation B.37, and Lemma B.9

$$\frac{1}{K}\sum_{k=0}^{K-1}\Big(V_0^{\pi^\star}(\rho)-V_0^{\bar{\pi}^k}(\rho)+\lambda^\top(\ell-V_g^{\bar{\pi}^k}(\rho))+\nu^\top(V_g^{\bar{\pi}^k}(\rho)-u)\Big)$$
$$\leqslant \frac{1}{K}\sum_{k=0}^{K-1}\Big(V_0^{\pi^\star}(\rho)-V_0^{\bar{\pi}^k}(\rho)+(\lambda^k)^\top\left(V_g^{\pi^\star}(\rho)-\ell\right)+\lambda^\top(\ell-V_g^{\bar{\pi}^k}(\rho))$$
$$+(\nu^k)^\top\left(u-V_g^{\pi^\star}(\rho)\right)+\nu^\top(V_g^{\bar{\pi}^k}(\rho)-u)\Big)$$
$$= \frac{1}{K}\sum_{k=0}^{K-1}\Big(V_0^{\pi^\star}(\rho)-V_0^{\bar{\pi}^k}(\rho)+(\lambda^k-\nu^k)^\top\left(V_g^{\pi^\star}(\rho)-V_g^{\bar{\pi}^k}(\rho)\right)$$
$$+(\lambda^k-\lambda)^\top(V_g^{\bar{\pi}^k}(\rho)-\ell)+(\nu^k-\nu)^\top(u-V_g^{\bar{\pi}^k}(\rho))\Big)$$
$$\leqslant \frac{N\log|\mathcal{A}|}{(1-\gamma)K\alpha}+\frac{3\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda+1/N)}{(1-\gamma)^4K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\bar{\pi}^k-\pi_i^k\|+\frac{2NB_\lambda}{(1-\gamma)^2K}+\frac{4N\eta}{(1-\gamma)^3K}$$
$$+\frac{1}{2K\eta}\|\lambda^0-\lambda\|^2+\frac{2N\eta}{(1-\gamma)^2}+\frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}B_\lambda}{(1-\gamma)^2K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\bar{\pi}^k-\pi_i^k\|$$
$$+\frac{1}{2K\eta}\|\nu^0-\nu\|^2+\frac{2N\eta}{(1-\gamma)^2}+\frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}B_\lambda}{(1-\gamma)^2K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\bar{\pi}^k-\pi_i^k\| \tag{B.38}$$
$$\leqslant \frac{N\log|\mathcal{A}|}{(1-\gamma)K\alpha}+\frac{2NB_\lambda}{(1-\gamma)^2K}+\frac{8N\eta}{(1-\gamma)^3}+\frac{\|\lambda^0-\lambda\|^2+\|\nu^0-\nu\|^2}{2K\eta}$$
$$+\frac{7\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda+1/N)}{(1-\gamma)^4K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\bar{\pi}^k-\pi_i^k\|. \tag{B.39}$$

Now, choosing λ and ν such that

$$\lambda_i = \begin{cases} B_\lambda, & \text{if } \ell_i - V_i^{\pi_k}(\rho) \geqslant 0 \\ 0, & \text{else} \end{cases} \qquad \nu_i = \begin{cases} B_\lambda, & \text{if } V_i^{\pi_k}(\rho) - u_i \geqslant 0 \\ 0, & \text{else} \end{cases}$$

Then, Equation B.39 leads to

$$\begin{aligned} &\frac{1}{K}\sum_{k=0}^{K-1}\left(V_0^{\pi^\star}(\rho) - V_0^{\bar{\pi}^k}(\rho)\right) + \frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{N} B_\lambda\left(\left[\ell_i - V_i^{\bar{\pi}^k}(\rho)\right]_+ + \left[V_i^{\bar{\pi}^k}(\rho) - u_i\right]_+\right) \\ &\leqslant \frac{N\log|\mathcal{A}|}{(1-\gamma)K\alpha} + \frac{2NB_\lambda}{(1-\gamma)^2K} + \frac{8N\eta}{(1-\gamma)^3} + \frac{NB_\lambda^2}{K\eta} \\ &\quad + \frac{7\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^4K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\bar{\pi}^k - \pi_i^k\|. \end{aligned} \tag{B.40}$$

Note that there always exists a policy $\widetilde{\pi}^K$ such that $d_\rho^{\widetilde{\pi}^K} = \frac{1}{K}\sum_{k=0}^{K-1} d_\rho^{\bar{\pi}^k}$, which implies

$$V_i^{\widetilde{\pi}^K} = \frac{1}{K}\sum_{k=0}^{K-1} V_i^{\bar{\pi}^k} \quad \forall i = 0, 1, \cdots, N.$$

As a result, Equation B.40 becomes

$$\begin{aligned} &\left(V_0^{\pi^\star}(\rho) - V_0^{\widetilde{\pi}^K}(\rho)\right) + B_\lambda\sum_{i=1}^{N}\left(\left[\ell_i - V_i^{\widetilde{\pi}^K}(\rho)\right]_+ + \left[V_i^{\widetilde{\pi}^K}(\rho) - u_i\right]_+\right) \\ &\leqslant \frac{N\log|\mathcal{A}|}{(1-\gamma)K\alpha} + \frac{2NB_\lambda}{(1-\gamma)^2K} + \frac{8N\eta}{(1-\gamma)^3} + \frac{NB_\lambda^2}{K\eta} \\ &\quad + \frac{7\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^4K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\|\bar{\pi}^k - \pi_i^k\|. \end{aligned} \tag{B.41}$$

Recall that Lemma 3.1 states that $2\|\lambda^\star\|_\infty \leqslant B_\lambda$ and $2\|\nu^\star\|_\infty \leqslant B_\lambda$. Applying Lemma B.10 with $C = B_\lambda$ and δ being the terms on the left hand side of Equation B.41, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\sum_{i=1}^{N}\left(\left[\ell_i - V_i^{\bar{\pi}^k}(\rho)\right]_+ + \left[V_i^{\bar{\pi}_k}(\rho) - u_i\right]_+\right)$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \Big(\big[\ell_i - V_i^{\tilde{\pi}^K}(\rho)\big]_+ + \big[V_i^{\tilde{\pi}^K}(\rho) - u_i\big]_+ \Big) \\
&\leqslant \frac{2}{B_\lambda} \Big(\frac{N \log |\mathcal{A}|}{(1-\gamma) K \alpha} + \frac{2 N B_\lambda}{(1-\gamma)^2 K} + \frac{8 N \eta}{(1-\gamma)^3} + \frac{N B_\lambda^2}{K \eta} \\
&\quad + \frac{7\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^4 K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\overline{\pi}^k - \pi_i^k\| \Big).
\end{aligned}
$$

By the bound on consensus error in Lemma B.7 and the Lipschitz continuity of the value function in Lemma B.6, this implies for any agent $j = 1, \cdots, N$

$$
\begin{aligned}
&\max\{ \frac{1}{K} \sum_{k=0}^{K-1} \Big(V_0^{\pi^\star}(\rho) - V_0^{\pi_j^k}(\rho) \Big), \frac{1}{K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \Big(\big[\ell_i - V_i^{\pi_j^k}(\rho)\big]_+ + \big[V_i^{\pi_j^k}(\rho) - u_i\big]_+ \Big) \} \\
&\leqslant \mathcal{O}\left(\frac{N}{K\alpha} + N\eta + \frac{N}{K\eta} + \frac{N^{3/2}\alpha}{1 - \sigma_2(W)} \right).
\end{aligned}
$$

Choosing the step sizes as $\alpha = \mathcal{O}(\frac{\sqrt{1-\sigma_2(W)}}{N^{1/4}\sqrt{K}})$ and $\eta = \mathcal{O}(1/\sqrt{K})$ leads to the claimed result.

□

B.4 Proof of Propositions

B.4.1 Proof of Proposition B.1

From Assumption 3.2, we define

$$
\frac{d_{i,\rho_i}^{\pi_\theta *}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)} = \frac{d_{j,\rho_j}^{\pi_\theta *}(s)}{d_{j,\mu_j}^{\pi_\theta}(s)} \triangleq \tilde{d}(s), \qquad \forall s : s \in \mathcal{S}_i \cap \mathcal{S}_j, \ \forall i, j.
$$

Our analysis uses the following performance difference lemma introduced in [157].

Lemma B.11. *For any policy π and $\tilde{\pi}$ operating in environment i under the initial state distribution ρ_i,*

$$
V_i^{\pi}(\rho_i) - V_i^{\tilde{\pi}}(\rho_i) = \frac{1}{1-\gamma_i} \mathbb{E}_{s \sim d_{i,\rho_i}^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)} \Big[A^{\pi'}(s,a) \Big].
$$

By Lemma B.11,

$$
\begin{aligned}
V(\theta^*;\boldsymbol{\rho}) - V(\theta;\boldsymbol{\rho}) &= \sum_{i=1}^{N} \frac{1}{1-\gamma_i} \sum_{s\in\mathcal{S}_i} \sum_{a\in\mathcal{A}} d_{i,\rho_i}^{\pi_{\theta^\star}}(s)\pi_{\theta^\star}(a\,|\,s) A_i^{\pi_\theta}(s,a) \\
&= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \pi_{\theta^\star}(a\,|\,s) \sum_{i:s\in\mathcal{S}_i} \frac{1}{1-\gamma_i} d_{i,\rho_i}^{\pi_{\theta^\star}}(s) A_i^{\pi_\theta}(s,a) \\
&\leqslant \sum_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \sum_{i:s\in\mathcal{S}_i} \frac{1}{1-\gamma_i} d_{i,\rho_i}^{\pi_{\theta^\star}}(s) A_i^{\pi_\theta}(s,a) \\
&= \sum_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \sum_{i:s\in\mathcal{S}_i} \frac{d_{i,\rho_i}^{\pi_{\theta^\star}}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)} \frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i} A_i^{\pi_\theta}(s,a) \\
&= \sum_{s\in\mathcal{S}} \tilde{d}(s) \max_{a\in\mathcal{A}} \sum_{i:s\in\mathcal{S}_i} \frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i} A_i^{\pi_\theta}(s,a) \\
&\leqslant \max_{s\in\mathcal{S}, i:s\in\mathcal{S}_i} \{\frac{d_{i,\rho_i}^{\pi_\theta *}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)}\} \sum_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \sum_{i:s\in\mathcal{S}_i} \frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i} A_i^{\pi_\theta}(s,a) \\
&\leqslant \max_{s\in\mathcal{S}, i:s\in\mathcal{S}_i} \{\frac{d_{i,\rho_i}^{\pi_\theta *}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)}\} |\mathcal{S}| \frac{2\lambda N}{|\mathcal{S}|} \\
&= 2\lambda N \max_{s\in\mathcal{S}, i:s\in\mathcal{S}_i} \{\frac{d_{i,\rho_i}^{\pi_\theta *}(s)}{d_{i,\mu_i}^{\pi_\theta}(s)}\} \\
&= 2\lambda N \max_{s\in\mathcal{S}, i:s\in\mathcal{S}_i} \{\frac{d_{i,\rho_i}^{\pi_\theta *}(s)}{(1-\gamma_i)\mu_i(s)}\}
\end{aligned}
$$

The sixth line follows since $\max_{a\in\mathcal{A}} \sum_{i:s\in\mathcal{S}_i} \frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i} A_i^{\pi_\theta}(s,a) \geqslant 0,\ \forall s$. The last inequality uses the fact that $d_{i,\mu_i}^{\pi}(s) \geqslant (1-\gamma_i)\mu_i(s)$, element-wise, $\forall \pi$, which simply follows from the definition of $d_{i,\mu_i}^{\pi}(s)$. The seventh line uses

$$
\max_{a\in\mathcal{A}} \sum_{i:s\in\mathcal{S}_i} \frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i} A_i^{\pi_\theta}(s,a) \leqslant \frac{2\lambda N}{|\mathcal{S}|}, \tag{B.42}
$$

which we now prove. To show this, we only have to prove this is true for those (s,a) where $\sum_{i:s\in\mathcal{S}_i} \frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i} A_i^{\pi_\theta}(s,a) \geqslant 0$. The gradient of θ under the softmax parameterization in environment i is

$$
\frac{\partial L_i^\lambda(\theta;\mu_i)}{\partial \theta_{s,a}} = \frac{1}{1-\gamma_i} d_{i,\mu_i}^{\pi_\theta}(s)\pi_\theta(a\,|\,s) A_i^{\pi_\theta}(s,a) + \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right). \tag{B.43}
$$

From our assumption $||\sum_{i=1}^{N} \nabla L_i^{\lambda}(\theta;\mu_i)|| \leqslant \frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|}$, we know that for all (s,a) such that $\sum_{i:s\in\mathcal{S}_i} \frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i} A_i^{\pi_\theta}(s,a) \geqslant 0$,

$$\begin{aligned}
\frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} &\geqslant \sum_{i=1}^{N} \frac{\partial L_i^{\lambda}(\theta;\mu_i)}{\partial \theta_{s,a}} \\
&= \sum_{i:s\in\mathcal{S}_i} \frac{1}{1-\gamma_i} d_{i,\mu_i}^{\pi_\theta}(s)\pi_\theta(a|s)A_i^{\pi_\theta}(s,a) + \sum_{i=1}^{N} \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right) \\
&\geqslant 0 + \sum_{i=1}^{N} \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right) \\
&\geqslant \frac{\lambda N}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|} - \pi_\theta(a|s)\right).
\end{aligned}$$

Rearranging the terms,

$$\pi_\theta(a \,|\, s) \geqslant \frac{1}{|\mathcal{A}|} - \frac{|\mathcal{S}|}{\lambda N}\frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} \geqslant \frac{1}{2|\mathcal{A}|}. \tag{B.44}$$

Re-writing Equation B.43 and summing over environments,

$$\begin{aligned}
\sum_{i=1}^{N} \frac{d_{i,\mu_i}^{\pi_\theta}(s)}{1-\gamma_i} A_i^{\pi_\theta}(s,a) &= \sum_{i:s\in\mathcal{S}_i} \frac{1}{\pi_\theta(a|s)} \frac{\partial L_i^{\lambda}(\theta;\mu_i)}{\partial \theta_{s,a}} - \sum_{i=1}^{N} \frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{\pi_\theta(a \,|\, s)|\mathcal{A}|} - 1\right) \\
&\leqslant \frac{1}{\pi_\theta(a|s)} \sum_{i:s\in\mathcal{S}_i} \frac{\partial L_i^{\lambda}(\theta;\mu_i)}{\partial \theta_{s,a}} + \sum_{i=1}^{N} \frac{\lambda}{|\mathcal{S}|} \\
&\leqslant 2|\mathcal{A}|\frac{\lambda N}{2|\mathcal{S}||\mathcal{A}|} + \frac{\lambda N}{|\mathcal{S}|} \\
&\leqslant \frac{2\lambda N}{|\mathcal{S}|},
\end{aligned}$$

where the second last line uses inequality Equation B.44.

B.5 Proof of Additional Lemmas

B.5.1 Proof of Lemma B.3

The proof uses a similar technique to Lemma E.2 of [12], which proves the second derivative is bounded. Here we also show the first and the third derivative is bounded. We use $\tilde{P}_i(\alpha)$ to denote the state-action transition matrix in environment i.

$$[\tilde{P}_i(\alpha)]_{(s,a)\to(s',a')} = \pi_\alpha\left(a'|s'\right) P_i\left(s'|s,a\right) \tag{B.45}$$

Differentiating with respect to α, we get

$$\left[\left.\frac{d\tilde{P}_i(\alpha)}{d\alpha}\right|_{\alpha=0}\right]_{(s,a)\to(s',a')} = \left.\frac{d\pi_\alpha\left(a'|s'\right)}{d\alpha}\right|_{\alpha=0} P_i\left(s'|s,a\right), \tag{B.46}$$

which implies that for any $\boldsymbol{x}$,

$$\left[\left.\frac{d\tilde{P}_i(\alpha)}{d\alpha}\right|_{\alpha=0} x\right]_{s,a} = \sum_{a',s'} \left.\frac{d\pi_\alpha\left(a'|s'\right)}{d\alpha}\right|_{\alpha=0} P_i\left(s'|s,a\right) x_{a',s'} \tag{B.47}$$

We can bound the ℓ_∞ norm of this as

$$\begin{aligned}
\max_{||\boldsymbol{u}||_2=1} \left\|\frac{d\tilde{P}_i(\alpha)}{d\alpha}\boldsymbol{x}\right\|_\infty &= \max_{s,a} \max_{||\boldsymbol{u}||_2=1} \left|\left[\left.\frac{d\tilde{P}_i(\alpha)}{d\alpha}\right|_{\alpha=0} \boldsymbol{x}\right]_{s,a}\right| \\
&= \max_{s,a} \max_{||\boldsymbol{u}||_2=1} \left|\sum_{a',s'} \left.\frac{d\pi_\alpha\left(a'|s'\right)}{d\alpha}\right|_{\alpha=0} P_i\left(s'|s,a\right) \boldsymbol{x}_{a',s'}\right| \\
&\leqslant \max_{s,a} \sum_{a',s'} \left|\left.\frac{d\pi_\alpha\left(a'|s'\right)}{d\alpha}\right|_{\alpha=0}\right| P_i\left(s'|s,a\right) |\boldsymbol{x}_{a',s'}| \\
&\leqslant \max_{s,a} \sum_{s'} P_i\left(s'|s,a\right) ||\boldsymbol{x}||_\infty \sum_{a'} \left|\left.\frac{d\pi_\alpha\left(a'|s'\right)}{d\alpha}\right|_{\alpha=0}\right| \\
&\leqslant C'||\boldsymbol{x}||_\infty
\end{aligned} \tag{B.48}$$

Using the same approach, we can bound

$$\max_{||\boldsymbol{u}||_2=1} \left\| \frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2}\boldsymbol{x} \right\|_\infty \leqslant C'''||\boldsymbol{x}||_\infty, \text{ and } \max_{||\boldsymbol{u}||_2=1} \left\| \frac{d^3\tilde{P}_i(\alpha)}{d\alpha^3}\boldsymbol{x} \right\|_\infty \leqslant C''''||\boldsymbol{x}||_\infty. \tag{B.49}$$

With $M(\alpha) := (\boldsymbol{I} - \gamma_i\tilde{P}_i(\alpha))^{-1}$, we re-writing the Bellman equation in the matrix form,

$$Q^\alpha(s_0, a_0) = e^T_{(s_0,a_0)}(\boldsymbol{I} - \gamma_i\tilde{P}_i(\alpha))^{-1}r = e^T_{(s_0,a_0)}M(\alpha)r. \tag{B.50}$$

Taking the first, second, and third derivative of $Q^\alpha(s_0, a_0)$ with respect to α,

$$\frac{dQ^\alpha(s_0, a)}{d\alpha} = \gamma_i e^T_{(s_0,a)}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)r, \tag{B.51}$$

$$\begin{aligned}\frac{d^2Q^\alpha(s_0, a_0)}{(d\alpha)^2} = 2\gamma_i^2 e^T_{(s_0,a_0)}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)r \\ +\gamma_i e^T_{(s_0,a_0)}M(\alpha)\frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2}M(\alpha)r,\end{aligned} \tag{B.52}$$

$$\begin{aligned}\frac{d^3Q^\alpha(s_0, a_0)}{(d\alpha)^3} = 6\gamma_i^3 e^T_{(s_0,a_0)}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)r \\ +3\gamma_i^2 e^T_{(s_0,a_0)}M(\alpha)\frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)r \\ +3\gamma_i^2 e^T_{(s_0,a_0)}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)\frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2}M(\alpha)r \\ +\gamma_i e^T_{(s_0,a_0)}M(\alpha)\frac{d^3\tilde{P}_i(\alpha)}{d\alpha^3}M(\alpha)r\end{aligned} \tag{B.53}$$

Using $M(\alpha)\mathbf{1} = (\mathrm{I} - \gamma_i\tilde{P}_i(\alpha))^{-1}\mathbf{1} = \sum_{n=0}^\infty \gamma_i^n\tilde{P}_i(\alpha)^n\mathbf{1} = \frac{1}{1-\gamma}\mathbf{1}$, Equation B.48, and Equation B.49, we have

$$\max_{||\boldsymbol{u}||_2=1} \left| \frac{dQ^\alpha(s_0, a)}{d\alpha}\bigg|_{\alpha=0} \right| \leqslant \left\| \gamma_i M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)r \right\|_\infty$$

$$\leqslant \frac{\gamma_i C'}{(1-\gamma_i)^2},$$

$$\begin{aligned}\max_{||\boldsymbol{u}||_2=1}\left|\frac{d^2Q^{\alpha}\left(s_0,a_0\right)}{d\alpha^2}\right|_{\alpha=0}| &\leqslant 2\gamma_i^2\left\|M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)r\right\|_{\infty}\\ &\quad+\gamma_i\left\|M(\alpha)\frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2}M(\alpha)r\right\|_{\infty}\\ &\leqslant \frac{2\gamma_i^2C'^2}{(1-\gamma_i)^3}+\frac{\gamma_iC''}{(1-\gamma_i)^2}\end{aligned}$$

$$\begin{aligned}\max_{||\boldsymbol{u}||_2=1}\left|\frac{d^3Q^{\alpha}\left(s_0,a_0\right)}{d\alpha^3}\right|_{\alpha=0}| &\leqslant 6\gamma_i^3\left\|M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)r\right\|_{\infty}\\ &\quad+3\gamma_i^2\left\|M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)\frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2}M(\alpha)r\right\|_{\infty}\\ &\quad+3\gamma_i^2\left\|M(\alpha)\frac{d^2\tilde{P}_i(\alpha)}{d\alpha^2}M(\alpha)\frac{d\tilde{P}_i(\alpha)}{d\alpha}M(\alpha)r\right\|_{\infty}\\ &\quad+\gamma_i\left\|M(\alpha)\frac{d^3\tilde{P}_i(\alpha)}{d\alpha^3}M(\alpha)r\right\|_{\infty}\\ &\leqslant \frac{6\gamma_i^3C'^3}{(1-\gamma_i)^4}+\frac{3\gamma_i^2C'C''}{(1-\gamma_i)^3}+\frac{3\gamma_i^2C'C''}{(1-\gamma_i)^3}+\frac{\gamma_iC'''}{(1-\gamma_i)^2}\\ &= \frac{6\gamma_i^3C'^3}{(1-\gamma_i)^4}+\frac{6\gamma_i^2C'C''}{(1-\gamma_i)^3}+\frac{\gamma_iC'''}{(1-\gamma_i)^2}\end{aligned}$$

By the definition of $\tilde{V}_i(\alpha)$,

$$\tilde{V}_i(\alpha)=\sum_a \pi_{\alpha}\left(a|s_0\right)Q^{\alpha}\left(s_0,a\right).$$

Taking the first derivative of $\tilde{V}_i(\alpha)$ with respect to α,

$$\frac{d\tilde{V}_i(\alpha)}{d\alpha}=\sum_a \frac{d\pi_{\alpha}\left(a|s_0\right)}{d\alpha}Q_i^{\alpha}\left(s_0,a\right)+\sum_a \pi_{\alpha}\left(a|s_0\right)\frac{dQ_i^{\alpha}\left(s_0,a\right)}{d\alpha}.$$

Taking the second derivative of $\tilde{V}_i(\alpha)$ with respect to α,

$$\frac{d^2\tilde{V}_i(\alpha)}{d\alpha^2} = \sum_a \frac{d^2\pi_\alpha(a|s_0)}{d\alpha^2} Q_i^\alpha(s_0,a) + 2\sum_a \frac{d\pi_\alpha(a|s_0)}{d\alpha}\frac{dQ_i^\alpha(s_0,a)}{d\alpha} + \sum_a \pi_\alpha(a|s_0)\frac{d^2Q_i^\alpha(s_0,a)}{d\alpha^2}.$$

Taking the third derivative of $\tilde{V}_i(\alpha)$ with respect to α,

$$\frac{d^3\tilde{V}_i(\alpha)}{d\alpha^3} = \sum_a \frac{d^3\pi_\alpha(a|s_0)}{d\alpha^3} Q^\alpha(s_0,a) + 3\sum_a \frac{d^2\pi_\alpha(a|s_0)}{d\alpha^2}\frac{dQ^\alpha(s_0,a)}{d\alpha} + 3\sum_a \frac{d\pi_\alpha(a|s_0)}{d\alpha}\frac{d^2Q^\alpha(s_0,a)}{d\alpha^2} + \sum_a \pi_\alpha(a|s_0)\frac{d^3Q^\alpha(s_0,a)}{d\alpha^3}.$$

Finally, we have

$$\max_{||\boldsymbol{u}||_2=1}\left|\frac{d\tilde{V}_i(\alpha)}{d\alpha}\bigg|_{\alpha=0}\right| \leqslant \frac{C'}{1-\gamma_i} + \frac{\gamma_i C'}{(1-\gamma_i)^2} = \frac{C'}{(1-\gamma_i)^2}$$

,

$$\max_{||\boldsymbol{u}||_2=1}\left|\frac{d^2\tilde{V}_i(\alpha)}{d\alpha^2}\bigg|_{\alpha=0}\right| \leqslant \frac{C''}{1-\gamma_i} + \frac{2C'^2}{(1-\gamma_i)^2} + \left(\frac{2\gamma_i C'^2}{(1-\gamma_i)^3} + \frac{\gamma_i C''}{(1-\gamma_i)^2}\right) = \frac{C''}{(1-\gamma_i)^2} + \frac{2\gamma_i C'^2}{(1-\gamma_i)^3}$$

, and

$$\begin{aligned}\max_{||\boldsymbol{u}||_2=1}\left|\frac{d^3\tilde{V}_i(\alpha)}{d\alpha^3}\bigg|_{\alpha=0}\right| &\leqslant \frac{C'''}{1-\gamma_i} + \frac{3\gamma_i C'C''}{(1-\gamma_i)^2} + 3C'\left(\frac{2\gamma_i^2 C'^2}{(1-\gamma_i)^3} + \frac{\gamma_i C''}{(1-\gamma_i)^2}\right) \\ &\quad + \frac{6\gamma_i^3 C'^3}{(1-\gamma_i)^4} + \frac{6\gamma_i^2 C'C''}{(1-\gamma_i)^3} + \frac{\gamma_i C'''}{(1-\gamma_i)^2} \\ &= \frac{C'''}{1-\gamma_i} + \frac{\gamma_i(6C'C''+C''')}{(1-\gamma_i)^2} + \frac{6\gamma_i^2(C'^3+C'C'')}{(1-\gamma_i)^3} + \frac{6\gamma_i^3 C'^3}{(1-\gamma_i)^4} \\ &= \frac{C'''}{(1-\gamma_i)^2} + \frac{6\gamma_i C'C''}{(1-\gamma_i)^3} + \frac{6\gamma_i^2 C'^3}{(1-\gamma_i)^4}\end{aligned}$$

□

B.5.2 Proof of Lemma B.4

By Equation B.43,

$$\begin{aligned}||\nabla L_i^\lambda(\theta_i^k;\mu_i)|| &\leqslant \sum_{s,a}\left|\frac{\partial L_i^\lambda(\theta_i^k;\mu_i)}{\partial \theta_{i\,s,a}^k}\right| \\ &\leqslant \sum_{s,a}\left|\frac{1}{1-\gamma_i}d_{\mu_j}^{\pi_\theta}(s)\pi_\theta(a\,|\,s)A_i^{\pi_\theta}(s,a)+\frac{\lambda}{|\mathcal{S}|}\left(\frac{1}{|\mathcal{A}|}-\pi_\theta(a|s)\right)\right| \\ &\leqslant \sum_{s,a}\frac{d_{\mu_j}^{\pi_\theta}(s)\pi_\theta(a\,|\,s)}{1-\gamma_i}\frac{1}{1-\gamma_i}+\sum_{s,a}\frac{\lambda}{|\mathcal{S}||\mathcal{A}|}+\sum_{s,a}\frac{\lambda}{|\mathcal{S}|}\pi_\theta(a\,|\,s) \\ &\leqslant \frac{1}{(1-\gamma_i)^2}+2\lambda,\end{aligned}$$

where the second last inequality follows from Equation 3.3. By the triangular inequality,

$$||\nabla \boldsymbol{L}^\lambda(\boldsymbol{\theta}^k;\boldsymbol{\mu})|| \leqslant \sum_{i=1}^N ||\nabla L_i^\lambda(\theta_i^k;\mu_i)|| \leqslant 2N\lambda+\sum_{i=1}^N\frac{1}{(1-\gamma_i)^2}.$$

□

B.5.3 Proof of Lemma B.7

We denote $g_i^k=(\frac{1}{N}+\lambda_i^k-\nu_i^k)Q_i^{\pi_i^k}\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $g^k=[(g_1^k)^\top,\ldots,(g_N^k)^\top]^\top\in\mathbb{R}^{N|\mathcal{S}||\mathcal{A}|}$. It is easy to see

$$\|g_i^k\|\leqslant\left|\frac{1}{N}+\lambda_i^k-\nu_i^k\right|\|Q_i^{\pi_i^k}\|\leqslant\frac{(B_\lambda+\frac{1}{N})\sqrt{|\mathcal{S}||\mathcal{A}|}}{1-\gamma},$$

which implies $\|g^k\|\leqslant\frac{(B_\lambda+\frac{1}{N})\sqrt{N|\mathcal{S}||\mathcal{A}|}}{1-\gamma}$ for all k. Then, using an argument similar to the one in [21][Lemma 1], we can get

$$\|\bar{\theta}^k-\theta_i^k\|\leqslant\frac{(B_\lambda+\frac{1}{N})\sqrt{N|\mathcal{S}||\mathcal{A}|}\alpha}{(1-\gamma)(1-\sigma_2(W))}. \tag{B.54}$$

The softmax function is Lipschitz with constant 1, i.e.

$$\|\pi_\theta - \pi_{\theta'}\| \leqslant \|\theta - \theta'\|, \quad \forall \theta, \theta',$$

Recall the definition of $\bar{\pi}^k$ in Equation B.30. The Lipschitz continuity and Equation B.54 imply the claimed result.

□

B.5.4 Proof of Lemma B.8

The performance difference lemma states that for any policies π_1, π_2, initial distribution ζ, and $i = 0, \cdots, N$

$$V_i^{\pi_1}(\zeta) - V_i^{\pi_2}(\zeta) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\zeta^{\pi^\star}, a \sim \pi^\star(\cdot|s)} \left[A_0^{\pi_k}(s,a) \right]. \tag{B.55}$$

By this lemma,

$$\begin{aligned}
&V_0^{\bar{\pi}^{k+1}}(\zeta) - V_0^{\bar{\pi}^k}(\zeta) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\zeta^{\bar{\pi}^{k+1}}, a \sim \bar{\pi}^{k+1}(\cdot|s)} \left[A_0^{\bar{\pi}^k}(s,a) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\zeta^{\bar{\pi}^{k+1}}, a \sim \bar{\pi}^{k+1}(\cdot|s)} \left[Q_0^{\bar{\pi}^k}(s,a) \right] - \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\zeta^{\bar{\pi}^{k+1}}} \left[V_0^{\bar{\pi}^k}(s) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\zeta^{\bar{\pi}^{k+1}}, a \sim \bar{\pi}^{k+1}(\cdot|s)} \left[Q_{L,k}^{\boldsymbol{\pi}^k}(s,a) \right] \\
&\quad + \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\zeta^{\bar{\pi}^{k+1}}, a \sim \bar{\pi}^{k+1}(\cdot|s)} \left[Q_{L,k}^{\bar{\pi}^k}(s,a) - Q_{L,k}^{\boldsymbol{\pi}^k}(s,a) \right] \\
&\quad - \frac{(\lambda^k - \nu^k)^\top}{1-\gamma} \mathbb{E}_{s \sim d_\zeta^{\bar{\pi}^{k+1}}, a \sim \bar{\pi}^{k+1}(\cdot|s)} \left[Q_g^{\bar{\pi}^k}(s,a) \right] - \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\zeta^{\bar{\pi}^{k+1}}} \left[V_0^{\bar{\pi}^k}(s) \right].
\end{aligned}$$

Note that the update rule in Equation B.31 implies

$$Q_{L,k}^{\boldsymbol{\pi}^k}(s,a) = \frac{N}{\alpha} \log \left(\frac{\bar{\pi}^{k+1}(a \mid s)}{\bar{\pi}^k(a \mid s)} Z_k(s) \right).$$

Combining the two equalities above, we have

$$
\begin{aligned}
&V_0^{\bar{\pi}^{k+1}}(\zeta) - V_0^{\bar{\pi}^k}(\zeta) \\
&= \frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}, a\sim\bar{\pi}^{k+1}(\cdot|s)} \left[\log \left(\frac{\bar{\pi}^{k+1}(a \mid s)}{\bar{\pi}^k(a \mid s)} Z_k(s) \right) \right] \\
&\quad + \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}, a\sim\bar{\pi}^{k+1}(\cdot|s)} \left[Q_{L,k}^{\bar{\pi}^k}(s,a) - Q_{L,k}^{\boldsymbol{\pi}^k}(s,a) \right] \\
&\quad - \frac{(\lambda^k - \nu^k)^\top}{1-\gamma} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}, a\sim\bar{\pi}^{k+1}(\cdot|s)} \left[Q_g^{\bar{\pi}^k}(s,a) \right] - \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}} \left[V_0^{\bar{\pi}^k}(s) \right] \\
&\geqslant \frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}} \left[D_{KL}(\bar{\pi}^{k+1}(\cdot \mid s) || \bar{\pi}^k(\cdot \mid s)) \right] + \frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}} \left[\log Z_k(s) \right] \\
&\quad - \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^3} \sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\| - \frac{(\lambda^k - \nu^k)^\top}{1-\gamma} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}, a\sim\bar{\pi}^{k+1}(\cdot|s)} \left[A_g^{\bar{\pi}^k}(s,a) \right] \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}} \left[V_{L,k}^{\bar{\pi}^k}(s) \right] \\
&\geqslant \frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}} \left[\log Z_k(s) \right] - \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^3} \sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\| \\
&\quad - (\lambda^k - \nu^k)^\top \left(V_g^{\bar{\pi}^{k+1}}(\zeta) - V_g^{\bar{\pi}^k}(\zeta) \right) - \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}} \left[V_{L,k}^{\bar{\pi}^k}(s) \right],
\end{aligned}
$$

where the last inequality applies the performance difference lemma. Rearranging this inequality leads to

$$
\begin{aligned}
V_{L,k}^{\bar{\pi}^{k+1}}(\zeta) - V_{L,k}^{\bar{\pi}^k}(\zeta) \geqslant \frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}} \left[\log Z_k(s) \right] - \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^3} \sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\| \\
- \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\zeta^{\bar{\pi}^{k+1}}} \left[V_{L,k}^{\bar{\pi}^k}(s) \right]. \qquad \text{(B.56)}
\end{aligned}
$$

From the definition of Z^k and Jensen's inequality,

$$
\begin{aligned}
\log Z^k(s) &= \log \left(\sum_{a'\in\mathcal{A}} \bar{\pi}^k(a' \mid s) \exp\left(\frac{\alpha}{N} \sum_{i=1}^N \left(\frac{1}{N} + \lambda_i^k - \nu_i^k\right) Q_i^{\pi_i^k}(s,a')\right) \right) \\
&\geqslant \sum_{a'\in\mathcal{A}} \bar{\pi}^k(a' \mid s) \log \left(\exp\left(\frac{\alpha}{N} \sum_{i=1}^N \left(\frac{1}{N} + \lambda_i^k - \nu_i^k\right) Q_i^{\pi_i^k}(s,a')\right) \right) \\
&= \frac{\alpha}{N} \sum_{a'\in\mathcal{A}} \bar{\pi}^k(a' \mid s) \sum_{i=1}^N \left(\frac{1}{N} + \lambda_i^k - \nu_i^k\right) Q_i^{\pi_i^k}(s,a')
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\alpha}{N} \sum_{a' \in \mathcal{A}} \bar{\pi}^k(a' \mid s) \sum_{i=1}^{N} (\frac{1}{N} + \lambda_i^k - \nu_i^k) Q_i^{\bar{\pi}^k}(s, a') \\
&\quad + \frac{\alpha}{N} \sum_{a' \in \mathcal{A}} \bar{\pi}^k(a' \mid s) \sum_{i=1}^{N} (\frac{1}{N} + \lambda_i^k - \nu_i^k)(Q_i^{\pi_i^k}(s, a') - Q_i^{\bar{\pi}^k}(s, a')) \\
&\geqslant \frac{\alpha}{N} V_{L,k}^{\bar{\pi}^k}(s) - \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)\alpha}{N(1-\gamma)^2} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\|.
\end{aligned}
$$

This bound on $\log Z_k(s)$ implies

$$
\begin{aligned}
&\frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s \sim d_\zeta^{\bar{\pi}^{k+1}}} \left[\log Z_k(s)\right] - \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\zeta^{\bar{\pi}^{k+1}}} \left[V_{L,k}^{\bar{\pi}^k}(s)\right] \\
&= \frac{N}{\alpha(1-\gamma)} \sum_s d_\zeta^{\bar{\pi}^{k+1}}(s) \left(\log Z_k(s) - \frac{\alpha}{N} V_{L,k}^{\bar{\pi}^k}(s)\right) \\
&= \frac{N}{\alpha(1-\gamma)} \sum_s d_\zeta^{\bar{\pi}^{k+1}}(s) \big(\log Z_k(s) - \frac{\alpha}{N} V_{L,k}^{\bar{\pi}^k}(s) + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)\alpha}{N(1-\gamma)^2} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\|\big) \\
&\quad - \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^3} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\| \\
&\geqslant \frac{N}{\alpha} \sum_s \zeta(s) \left(\log Z_k(s) - \frac{\alpha}{N} V_{L,k}^{\bar{\pi}^k}(s)\right) - \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^3} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\|,
\end{aligned}
$$

where the inequality follows from the fact that $d_\zeta^\pi \geqslant (1-\gamma)\zeta$ elementwise for any policy π. Plugging this bound into Equation B.56, we have

$$
\begin{aligned}
&V_{L,k}^{\bar{\pi}^{k+1}}(\zeta) - V_{L,k}^{\bar{\pi}^k}(\zeta) \\
&\geqslant \frac{N}{\alpha} \mathbb{E}_{s \sim \zeta} \left[\log Z_k(s) - \frac{\alpha}{N} V_{L,k}^{\bar{\pi}^k}(s)\right] - \frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^3} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\|.
\end{aligned}
$$

□

B.5.5 Proof of Lemma B.9

By the performance difference lemma in Equation B.55,

$$
V_0^{\pi^\star}(\rho) - V_0^{\bar{\pi}^k}(\rho)
$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i=1}^{N}(V_i^{\pi^\star}(\rho) - V_i^{\bar{\pi}^k}(\rho)) \\
&= \frac{1}{N}\sum_{i=1}^{N}(V_i^{\pi^\star}(\rho) - V_i^{\pi_i^k}(\rho)) + \frac{1}{N}\sum_{i=1}^{N}(V_i^{\pi_i^k}(\rho) - V_i^{\bar{\pi}^k}(\rho)) \\
&\leqslant \frac{1}{N(1-\gamma)}\sum_{i=1}^{N}\mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)}\left[A_i^{\pi_i^k}(s,a)\right] + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{N(1-\gamma)^3}\sum_{i=1}^{N}\|\bar{\pi}^k - \pi_i^k\| \\
&= \frac{1}{N(1-\gamma)}\sum_{i=1}^{N}\mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)}\left[Q_i^{\pi_i^k}(s,a)\right] - \frac{1}{N(1-\gamma)}\sum_{i=1}^{N}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[V_i^{\pi_i^k}(s)\right] \\
&\quad + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{N(1-\gamma)^3}\sum_{i=1}^{N}\|\bar{\pi}^k - \pi_i^k\|.
\end{aligned}
$$

Plugging in the update rule of the policy,

$$
\begin{aligned}
&V_0^{\pi^\star}(\rho) - V_0^{\bar{\pi}^k}(\rho) \\
&\leqslant \frac{1}{N(1-\gamma)}\sum_{i=1}^{N}\mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)}\left[Q_i^{\pi_i^k}(s,a)\right] - \frac{1}{N(1-\gamma)}\sum_{i=1}^{N}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[V_i^{\pi_i^k}(s)\right] \\
&\quad + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{N(1-\gamma)^3}\sum_{i=1}^{N}\|\bar{\pi}^k - \pi_i^k\| \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)}\left[\sum_{i=1}^{N}(\frac{1}{N} + \lambda_i^k - \nu_i^k)Q_i^{\pi_i^k}(s,a)\right] \\
&\quad - \frac{(\lambda^k - \nu^k)^\top}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)}\left[Q_g^{\boldsymbol{\pi}^k}(s,a)\right] \\
&\quad - \frac{1}{N(1-\gamma)}\sum_{i=1}^{N}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[V_i^{\pi_i^k}(s)\right] + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{N(1-\gamma)^3}\sum_{i=1}^{N}\|\bar{\pi}^k - \pi_i^k\| \\
&= \frac{N}{\alpha(1-\gamma)}\mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)}\left[\log\left(\frac{\bar{\pi}^{k+1}(a\mid s)}{\bar{\pi}^k(a\mid s)}Z_k(s)\right)\right] \\
&\quad - \frac{(\lambda^k - \nu_k)^\top}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)}\left[A_g^{\boldsymbol{\pi}^k}(s,a)\right] - \frac{(\lambda^k - \nu_k)^\top}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[V_g^{\boldsymbol{\pi}^k}(s)\right] \\
&\quad - \frac{1}{N(1-\gamma)}\sum_{i=1}^{N}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[V_i^{\pi_i^k}(s)\right] + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{N(1-\gamma)^3}\sum_{i=1}^{N}\|\bar{\pi}^k - \pi_i^k\| \\
&\leqslant \frac{N}{\alpha(1-\gamma)}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[D_{\mathrm{KL}}(\pi^\star(\cdot\mid s)||\bar{\pi}^k(\cdot\mid s)) - D_{\mathrm{KL}}(\pi^\star(\cdot\mid s)||\bar{\pi}^{k+1}(\cdot\mid s))\right] \\
&\quad + \frac{N}{\alpha(1-\gamma)}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\log Z^k(s)\right]
\end{aligned}
$$

$$- \frac{(\lambda^k - \nu^k)^\top}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)} \left[A_g^{\boldsymbol{\pi}^k}(s,a) \right] - \frac{(\lambda^k - \nu^k)^\top}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi^\star}} \left[V_g^{\boldsymbol{\pi}^k}(s) \right]$$
$$- \frac{1}{N(1-\gamma)} \sum_{i=1}^N \mathbb{E}_{s\sim d_\rho^{\pi^\star}} \left[V_i^{\pi_i^k}(s) \right] + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{N(1-\gamma)^3} \sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\|.$$

Re-grouping the terms,

$$\begin{aligned}
&V_0^{\pi^\star}(\rho) - V_0^{\bar{\pi}^k}(\rho) \\
&\leqslant \frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s\sim d_\rho^{\pi^\star}} \left[D_{\mathrm{KL}}(\pi^\star(\cdot \mid s) || \bar{\pi}^k(\cdot \mid s)) - D_{\mathrm{KL}}(\pi^\star(\cdot \mid s) || \bar{\pi}^{k+1}(\cdot \mid s)) \right] \\
&\quad + \frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s\sim d_\rho^{\pi^\star}} \left[\log Z^k(s) \right] - \frac{(\lambda^k - \nu_k)^\top}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)} \left[A_g^{\bar{\pi}^k}(s,a) \right] \\
&\quad + \frac{(\lambda^k - \nu^k)^\top}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi^\star}, a\sim\pi^\star(\cdot|s)} \left[A_g^{\bar{\pi}^k}(s,a) - A_g^{\boldsymbol{\pi}^k}(s,a) \right] - \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi^\star}} \left[V_{L,k}^{\boldsymbol{\pi}^k}(s) \right] \\
&\quad + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{N(1-\gamma)^3} \sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\| \\
&\leqslant \frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s\sim d_\rho^{\pi^\star}} \left[D_{\mathrm{KL}}(\pi^\star(\cdot \mid s) || \bar{\pi}^k(\cdot \mid s)) - D_{\mathrm{KL}}(\pi^\star(\cdot \mid s) || \bar{\pi}^{k+1}(\cdot \mid s)) \right] \\
&\quad + \frac{N}{\alpha(1-\gamma)} \mathbb{E}_{s\sim d_\rho^{\pi^\star}} \left[\log Z^k(s) \right] - \frac{(\lambda^k - \nu_k)^\top}{1-\gamma} \left(V_g^{\pi^\star}(s) - V_g^{\bar{\pi}^k}(s) \right) \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_\rho^{\pi^\star}} \left[V_{L,k}^{\boldsymbol{\pi}^k}(s) \right] + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^3} \sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\|, \qquad \text{(B.57)}
\end{aligned}$$

where the second inequality follows from the performance difference lemma and the Lipschitz continuity of the advantage.

Applying Lemma B.8 with $\zeta = d_\rho^{\pi^\star}$,

$$\begin{aligned}
\frac{N}{\alpha} \mathbb{E}_{s\sim d_\rho^{\pi^\star}} \left[\log Z_k(s) - \frac{\alpha}{N} V_{L,k}^{\bar{\pi}^k}(s) \right] &\leqslant V_{L,k}^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) - V_{L,k}^{\bar{\pi}^k}(d_\rho^{\pi^\star}) \\
&\quad + \frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^3} \sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\|. \qquad \text{(B.58)}
\end{aligned}$$

Combining Equation B.57 and Equation B.58,

$$V_0^{\pi^\star}(\rho) - V_0^{\bar{\pi}^k}(\rho)$$

$$
\begin{aligned}
&\leqslant \frac{N}{\alpha(1-\gamma)}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[D_{\mathrm{KL}}(\pi^\star(\cdot\mid s)||\bar{\pi}^k(\cdot\mid s)) - D_{\mathrm{KL}}(\pi^\star(\cdot\mid s)||\bar{\pi}^{k+1}(\cdot\mid s))\right] \\
&\quad + \frac{1}{1-\gamma}\left(V_{L,k}^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) - V_{L,k}^{\bar{\pi}^k}(d_\rho^{\pi^\star})\right) + \frac{2\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^4}\sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\| \\
&\quad - \frac{(\lambda^k-\nu^k)^\top}{1-\gamma}\left(V_g^{\pi^\star}(s) - V_g^{\bar{\pi}^k}(s)\right) + \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{N(1-\gamma)^3}\sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\|,
\end{aligned}
$$

which implies

$$
\begin{aligned}
&V_{L,k}^{\pi^\star}(\rho) - V_{L,k}^{\bar{\pi}^k}(\rho) \\
&\leqslant \frac{N}{\alpha(1-\gamma)}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[D_{\mathrm{KL}}(\pi^\star(\cdot\mid s)||\bar{\pi}^k(\cdot\mid s)) - D_{\mathrm{KL}}(\pi^\star(\cdot\mid s)||\bar{\pi}^{k+1}(\cdot\mid s))\right] \\
&\quad + \frac{1}{1-\gamma}\left(V_{L,k}^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) - V_{L,k}^{\bar{\pi}^k}(d_\rho^{\pi^\star})\right) + \frac{3\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^4}\sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\|.
\end{aligned}
$$

Taking the average from $k=0$ to $k=K-1$, we have

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\left(V_{L,k}^{\pi^\star}(\rho) - V_{L,k}^{\bar{\pi}^k}(\rho)\right) \\
&\leqslant \frac{N}{(1-\gamma)K\alpha}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[D_{\mathrm{KL}}(\pi^\star(\cdot\mid s)||\pi_0(\cdot\mid s))\right] + \frac{1}{(1-\gamma)K}\sum_{k=0}^{K-1}\left(V_{L,k}^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) - V_{L,k}^{\bar{\pi}^k}(d_\rho^{\pi^\star})\right) \\
&\quad + \frac{3\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^4 K}\sum_{k=0}^{K-1}\sum_{i=1}^N \|\bar{\pi}^k - \pi_i^k\|. \qquad \text{(B.59)}
\end{aligned}
$$

The second term on the right hand side can be decomposed as follows

$$
\begin{aligned}
&\frac{1}{(1-\gamma)K}\sum_{k=0}^{K-1}\left(V_{L,k}^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) - V_{L,k}^{\bar{\pi}^k}(d_\rho^{\pi^\star})\right) \\
&\leqslant \frac{1}{(1-\gamma)K}\sum_{k=0}^{K-1}\left(V_0^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) - V_0^{\bar{\pi}^k}(d_\rho^{\pi^\star})\right) \\
&\quad + \frac{1}{(1-\gamma)K}\sum_{k=0}^{K-1}(\lambda^k-\nu^k)^\top\left(V_g^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) - V_g^{\bar{\pi}^k}(d_\rho^{\pi^\star})\right) \\
&= \frac{V_0^{\bar{\pi}^K}(d_\rho^{\pi^\star})}{(1-\gamma)K} + \frac{1}{(1-\gamma)K}\sum_{k=0}^{K-1}\left((\lambda^{k+1}-\nu^{k+1})^\top V_g^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) - (\lambda^k-\nu^k)^\top V_g^{\bar{\pi}^k}(d_\rho^{\pi^\star})\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{(1-\gamma)K} \sum_{k=0}^{K-1} \left(\lambda^k - \nu^k - \lambda^{k+1} + \nu^{k+1}\right)^\top V_g^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) \\
&= \frac{V_0^{\bar{\pi}^K}(d_\rho^{\pi^\star})}{(1-\gamma)K} + \frac{1}{(1-\gamma)K} \sum_{i=1}^{N} (\lambda_i^K - \nu_i^K) V_i^{\bar{\pi}^K}(d_\rho^{\pi^\star}) \\
&\quad + \frac{1}{(1-\gamma)K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} (\lambda_i^k - \nu_i^k - \lambda_i^{k+1} + \nu_i^{k+1}) V_i^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}). \qquad \text{(B.60)}
\end{aligned}
$$

We know that the value functions are bounded between $[0, \frac{1}{1-\gamma}]$. The projection in the update of the dual variable in Equation 3.18 guarantees $\lambda_i^k \in [0, B_\lambda]$. It is also straightforward to see that for all i and k

$$
|\lambda_{i,k} - \lambda_{i,k+1}| \leqslant \frac{\eta}{1-\gamma} + B\eta, \; |\nu_{i,k} - \nu_{i,k+1}| \leqslant \frac{\eta}{1-\gamma} + B\eta.
$$

Using these bounds in Equation B.60, we get

$$
\begin{aligned}
&\frac{1}{(1-\gamma)K} \sum_{k=0}^{K-1} \left(V_{L,k}^{\bar{\pi}^{k+1}}(d_\rho^{\pi^\star}) - V_{L,k}^{\bar{\pi}^k}(d_\rho^{\pi^\star}) \right) \\
&\leqslant \frac{1}{(1-\gamma)^2 K} + \frac{NB_\lambda}{(1-\gamma)^2 K} + \frac{2N\eta}{(1-\gamma)^3 K} + \frac{2NB\eta}{(1-\gamma)^2 K} \\
&\leqslant \frac{2NB_\lambda}{(1-\gamma)^2 K} + \frac{4N\eta}{(1-\gamma)^3 K}. \qquad \text{(B.61)}
\end{aligned}
$$

Finally, combining Equation B.59 and Equation B.61 yields

$$
\begin{aligned}
&\frac{1}{K} \sum_{k=0}^{K-1} \left(V_{L,k}^{\pi^\star}(\rho) - V_{L,k}^{\bar{\pi}^k}(\rho) \right) \\
&\leqslant \frac{N}{(1-\gamma)K\alpha} \mathbb{E}_{s \sim d_\rho^{\pi^\star}} \left[D_{\text{KL}}(\pi^\star(\cdot \mid s) || \pi_0(\cdot \mid s)) \right] \\
&\quad + \frac{3\sqrt{|\mathcal{S}||\mathcal{A}|}(B_\lambda + 1/N)}{(1-\gamma)^4 K} \sum_{k=0}^{K-1} \sum_{i=1}^{N} \|\bar{\pi}^k - \pi_i^k\| + \frac{2NB_\lambda}{(1-\gamma)^2 K} + \frac{4N\eta}{(1-\gamma)^3 K},
\end{aligned}
$$

which leads to the claimed result by recognizing the fact that for $D_{\text{KL}}(p_1||p_2) \leqslant \log |\mathcal{A}|$ for $p_1, p_2 \in \Delta_{\mathcal{A}}$ if p_2 is a uniform distribution.

□

APPENDIX C

SUPPLEMENTARY MATERIAL FOR RESULTS IN CHAPTER 4

C.1 Proof of Theorems and Corollaries

We frequently use the following inequalities which hold for all $\tau \geqslant 0$, $\pi \in \Delta_{\mathcal{S}}^{\mathcal{A}}$, and $\phi \in \Delta_{\mathcal{S}}^{\mathcal{B}}$,

$$J_\tau(\pi, \phi_\tau(\pi)) \leqslant J_\tau(\pi, \phi), \quad J_\tau(\pi_\tau(\phi), \phi) \geqslant J_\tau(\pi, \phi).$$

We use $H(\cdot)$ to denote the entropy of a distribution. For example,

$$H(\pi(\cdot \mid s)) = -\sum_a \pi(a \mid s) \log \pi(a \mid s), \quad H(\phi(\cdot \mid s)) = -\sum_b \phi(b \mid s) \log \phi(b \mid s). \tag{C.1}$$

Due to the uniqueness of $\phi_\tau(\cdot)$, Danskin's Theorem guarantees that $g_\tau(\pi_\theta)$ defined in Equation 4.2 is differentiable with respect to θ [158]

$$\nabla_\theta g_\tau(\pi_\theta) = \nabla_\theta J_\tau(\pi_\theta, \phi), \quad \phi = \phi_\tau(\pi_\theta), \quad \forall \theta \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}. \tag{C.2}$$

We also introduce a few lemmas that will be applied regularly in the rest of the paper.

Lemma C.1. *Let $L_V = \frac{8}{(1-\gamma)^3}$. The value function J is L_V-Lipschitz continuous and has L_V-Lipschitz gradients, i.e. we have for all $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ and $\psi_1, \psi_2 \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{B}|}$*

$$\begin{aligned}
&\|\nabla_\theta J(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta J(\pi_{\theta_2}, \phi_{\psi_2})\| \leqslant L_V(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|), \\
&\|\nabla_\psi J(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\psi J(\pi_{\theta_2}, \phi_{\psi_2})\| \leqslant L_V(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|), \\
&\|J(\pi_{\theta_1}, \phi_{\psi_1}) - J(\pi_{\theta_2}, \phi_{\psi_2})\| \leqslant L_V(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|).
\end{aligned}$$

Lemma C.2. *Let $L_{\mathcal{H}} = \frac{4+8\log|\mathcal{A}|}{(1-\gamma)^3}$. The regularization functions $\mathcal{H}_\pi$ and $\mathcal{H}_\phi$ are $L_{\mathcal{H}}$-Lipschitz continuous and has $L_{\mathcal{H}}$-Lipschitz gradients.*

Lemmas C.1 and C.2 imply that $\forall \tau \geqslant 0$, $\nabla_\theta J_\tau$ is Lipschitz continuous, i.e. for any $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, $\psi_1, \psi_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{B}|}$

$$\begin{aligned}\|\nabla_\theta J_\tau(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta J_\tau(\pi_{\theta_2}, \phi_{\psi_2})\| \leqslant{}& \|\nabla_\theta J(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta J(\pi_{\theta_2}, \phi_{\psi_2})\| \\ &+ \tau\|\nabla_\theta \mathcal{H}_\pi(s, \pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta \mathcal{H}_\pi(s, \pi_{\theta_2}, \phi_{\psi_2})\| \\ &+ \tau\|\nabla_\theta \mathcal{H}_\phi(s, \pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta \mathcal{H}_\phi(s, \pi_{\theta_2}, \phi_{\psi_2})\| \\ \leqslant{}& (L_V + 2\tau L_{\mathcal{H}})(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|). \end{aligned} \tag{C.3}$$

Lemma C.3. *For any $0 \leqslant a \leqslant 1$ and integer $k > 0$, we have*

$$\frac{1}{(k+h)^a} - \frac{1}{(k+1+h)^a} \leqslant \frac{8}{3(k+h)^{a+1}}.$$

C.1.1 Proof of Theorem 4.1

The definition of the constant L and Equation C.3 imply for any $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, $\psi_1, \psi_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{B}|}$

$$\|\nabla_\theta J_\tau(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta J_\tau(\pi_{\theta_2}, \phi_{\psi_2})\| \leqslant L(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|). \tag{C.4}$$

We will use an induction argument to prove the convergence of $3\delta_k^\pi + \delta_k^\phi$. The base case is $3\delta_0^\pi + \delta_0^\phi \leqslant 3\delta_0^\pi + \delta_0^\phi$, which obviously holds. Now, suppose

$$3\delta_k^\pi + \delta_k^\phi \leqslant (1 - \frac{\alpha(1-\gamma)\tau\rho_{\min}^2 c^2}{32|\mathcal{S}|})^k (3\delta_0^\pi + \delta_0^\phi) \tag{C.5}$$

holds. We aim to show

$$3\delta_{k+1}^{\pi} + \delta_{k+1}^{\phi} \leqslant (1 - \frac{\alpha(1-\gamma)\tau\rho_{\min}^2 c^2}{32|\mathcal{S}|})^{k+1}(3\delta_0^{\pi} + \delta_0^{\phi}).$$

We introduce the following technical lemmas.

Lemma C.4. *Suppose Equation C.5 holds. Then, we have*

$$-\left(\min_{s,a} \pi_{\theta_k}(a \mid s)\right)^2 \leqslant -\frac{3c^2}{8}, \tag{C.6}$$

$$-\left(\min_{s,b} \phi_{\psi_k}(b \mid s)\right)^2 \leqslant -\frac{3c^2}{8}. \tag{C.7}$$

Lemma C.5. *Suppose Equation C.5 holds. Under Assumption 4.1 and the step size $\alpha_k \leqslant (L + \frac{2\sqrt{|\mathcal{S}|}L^2}{\sqrt{(1-\gamma)\rho_{\min}}\tau c})^{-1}$, we have*

$$\begin{aligned}
&g_\tau(\theta_k) - g_\tau(\theta_{k+1}) \\
&= J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - J_\tau(\pi_{\theta_{k+1}}, \phi_\tau(\pi_{\theta_{k+1}})) \\
&= \frac{\alpha_k}{2}\left(\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2\right).
\end{aligned}$$

By the lemma above, we have

$$\begin{aligned}
&\delta_{k+1}^{\pi} - \delta_k^{\pi} \\
&= J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - J_\tau(\pi_{\theta_{k+1}}, \phi_\tau(\pi_{\theta_{k+1}})) \\
&\leqslant \frac{\alpha_k}{2}\left(\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2\right).
\end{aligned} \tag{C.8}$$

Similarly, we consider the decay of δ_k^{ϕ}.

$$\begin{aligned}
\delta_{k+1}^{\phi} - \delta_k^{\phi} &= J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - g_\tau(\pi_{\theta_{k+1}}) - J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) + g_\tau(\pi_{\theta_k}) \\
&= \left(J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k})\right)
\end{aligned}$$

$$+ \left(J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k}) - J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\right) + \left(g_\tau(\pi_{\theta_k}) - g_\tau(\pi_{\theta_{k+1}})\right). \tag{C.9}$$

Using the L-smoothness of the value function derived in Equation C.4

$$\begin{aligned}
&J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k}) \\
&\leqslant \langle \nabla_\psi J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k}), \psi_{k+1} - \psi_k \rangle + \frac{L}{2}\|\psi_{k+1} - \psi_k\|^2 \\
&= -\beta_k \|\nabla_\psi J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k})\|^2 + \frac{L\beta_k^2}{2}\|\nabla_\psi J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k})\|^2 \\
&\leqslant -\frac{\beta_k}{2}\|\nabla_\psi J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k})\|^2 \\
&\leqslant -\frac{(1-\gamma)\beta_k \tau \rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,b} \phi_{\psi_k}(b \mid s)\right)^2 \left(J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) - J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\right) \\
&= -\frac{(1-\gamma)\beta_k \tau \rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,b} \phi_{\psi_k}(b \mid s)\right)^2 \delta_k^\phi,
\end{aligned}$$

where the second inequality uses $\beta_k \leqslant \frac{1}{L}$ and the third inequality follows from Lemma 4.4 and the fact that $d_\rho^{\pi,\phi}(s) \leqslant 1$ for all $s \in \mathcal{S}$ and policies π, ϕ.

Using Equation C.7 of Lemma C.4 to further simplify this inequality,

$$J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k}) \leqslant -\frac{3(1-\gamma)\beta_k \tau \rho_{\min}^2 c^2}{8|\mathcal{S}|}\delta_k^\phi. \tag{C.10}$$

For the second term of Equation C.9, we have from the L-smoothness of the value function derived in Equation C.4

$$\begin{aligned}
J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k}) - J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) &\leqslant \langle \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}), \theta_{k+1} - \theta_k \rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= \alpha_k \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 + \frac{L\alpha_k^2}{2}\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \\
&\leqslant \frac{3\alpha_k}{2}\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2,
\end{aligned} \tag{C.11}$$

where in the last inequality we use $\alpha_k L \leqslant 1$.

Similarly to Equation C.8, the last term of Equation C.9 is bounded as

$$
\begin{aligned}
g_\tau(\pi_{\theta_k}) - g_\tau(\pi_{\theta_{k+1}}) &= g_\tau(\pi_{\theta_k}) - g_\tau(\pi_{\theta_{k+1}}) + g_\tau(\pi_{\theta_{k+1}}) - g_\tau(\pi_{\theta_{k+1}}) \\
&\leqslant \frac{\alpha_k}{2}\left(\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2\right)
\end{aligned} \tag{C.12}
$$

Using Equation C.10-Equation C.12 in Equation C.9, we have

$$
\begin{aligned}
\delta^\phi_{k+1} &= \left(J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k})\right) \\
&\quad + \left(J_\tau(\pi_{\theta_{k+1}}, \phi_{\psi_k}) - J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\right) + \left(g_\tau(\pi_{\theta_k}) - g_\tau(\pi_{\theta_{k+1}})\right) \\
&\leqslant (1 - \frac{3(1-\gamma)\beta_k \tau \rho^2_{\min} c^2}{8|\mathcal{S}|})\delta^\phi_k + \frac{3\alpha_k}{2}\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \\
&\quad + \frac{\alpha_k}{2}\left(\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2\right).
\end{aligned} \tag{C.13}
$$

Combining Equation C.8 and Equation C.13,

$$
\begin{aligned}
3\delta^\pi_{k+1} + \delta^\phi_{k+1} &\leqslant 3\delta^\pi_k + \frac{3\alpha_k}{2}\left(\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2\right) \\
&\quad + (1 - \frac{3(1-\gamma)\beta_k \tau \rho^2_{\min} c^2}{8|\mathcal{S}|})\delta^\phi_k + 2\delta^\phi_k)\delta^\phi_k + \frac{3\alpha_k}{2}\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \\
&\quad + \frac{\alpha_k}{2}\left(\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2\right) \\
&\leqslant 3\delta^\pi_k + (1 - \frac{3(1-\gamma)\beta_k \tau \rho^2_{\min} c^2}{8|\mathcal{S}|})\delta^\phi_k + \frac{3\alpha_k}{2}\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \\
&\quad + 2\alpha_k\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - 2\alpha_k\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2.
\end{aligned}
$$

Simplifying this inequality with

$$
\begin{aligned}
\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 &= \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - (\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}))\|^2 \\
&\leqslant \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2 + \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \\
&\quad + 2\langle \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})), \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\rangle
\end{aligned}
$$

$$\leqslant \frac{5}{4}\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2 + 5\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2,$$

we have

$$3\delta_{k+1}^\pi + \delta_{k+1}^\phi \leqslant 3\delta_k^\pi + (1 - \frac{3(1-\gamma)\beta_k\tau\rho_{\min}^2 c^2}{8|\mathcal{S}|})\delta_k^\phi - \frac{\alpha_k}{8}\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2 + \frac{19\alpha_k}{2}\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2. \tag{C.14}$$

Using Lemma 4.4 to bound $-\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2$,

$$\begin{aligned}
&- \|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2 \\
&\leqslant -\frac{2(1-\gamma)\tau\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 (J_\tau(\pi_\tau(\phi_\tau(\pi_{\theta_k})), \phi_\tau(\pi_{\theta_k})) - J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))) \\
&\leqslant -\frac{2(1-\gamma)\tau\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 (J_\tau(\pi_\tau^\star, \phi_\tau^\star) - J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))), \qquad \text{(C.15)}
\end{aligned}$$

where the second inequality follows from

$$J_\tau(\pi_\tau(\phi_\tau(\pi_{\theta_k})), \phi_\tau(\pi_{\theta_k})) = \max_\pi J_\tau(\pi, \phi_\tau(\pi_{\theta_k})) \geqslant \max_\pi \min_\phi J_\tau(\pi, \phi) = J_\tau(\pi_\tau^\star, \phi_\tau^\star).$$

From Lemma C.4 Equation C.6, $-\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 \leqslant -\frac{3c^2}{8}$, which further simplifies Equation C.15

$$\begin{aligned}
-\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\|^2 &\leqslant -\frac{2(1-\gamma)\tau\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 (J_\tau(\pi_\tau^\star, \phi_\tau^\star) - J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))) \\
&= -\frac{2(1-\gamma)\tau\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 \delta_k^\pi \leqslant -\frac{3(1-\gamma)\tau\rho_{\min}^2 c^2}{4|\mathcal{S}|}\delta_k^\pi.
\end{aligned}$$

For $\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2$, we have from the L-smoothness of the value function derived in Equation C.4

$$\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \leqslant L^2\|\phi_\tau(\pi_{\theta_k}) - \phi_{\psi_k}\|^2$$

$$\leqslant \frac{2\log(2)L^2}{\tau\rho_{\min}}\left(J_\tau(\pi_{\theta_k},\phi_{\psi_k}) - J_\tau(\pi_{\theta_k},\phi_\tau(\pi_{\theta_k}))\right)$$
$$= \frac{2\log(2)L^2}{\tau\rho_{\min}}\delta_k^\phi$$

Using the bound on $-\|\nabla_\theta J_\tau(\pi_{\theta_k},\phi_\tau(\pi_{\theta_k}))\|^2$ and $\|\nabla_\theta J_\tau(\pi_{\theta_k},\phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k},\phi_{\psi_k})\|^2$ in Equation C.14,

$$\begin{aligned}
3\delta_{k+1}^\pi + \delta_{k+1}^\phi &\leqslant 3\delta_k^\pi + (1 - \frac{3(1-\gamma)\beta_k\tau\rho_{\min}^2c^2}{8|\mathcal{S}|})\delta_k^\phi - \frac{\alpha_k}{8}\|\nabla_\theta J_\tau(\pi_{\theta_k},\phi_\tau(\pi_{\theta_k}))\|^2 \\
&\quad + \frac{19\alpha_k}{2}\|\nabla_\theta J_\tau(\pi_{\theta_k},\phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k},\phi_{\psi_k})\|^2 \\
&\leqslant 3\delta_k^\pi + (1 - \frac{3(1-\gamma)\beta_k\tau\rho_{\min}^2c^2}{8|\mathcal{S}|})\delta_k^\phi - \frac{3\alpha_k(1-\gamma)\tau\rho_{\min}^2c^2}{32|\mathcal{S}|}\delta_k^\pi + \frac{19\log(2)L^2\alpha_k}{\tau\rho_{\min}}\delta_k^\phi \\
&= 3(1 - \frac{\alpha_k(1-\gamma)\tau\rho_{\min}^2c^2}{32|\mathcal{S}|})\delta_k^\pi + (1 - \frac{3(1-\gamma)\beta_k\tau\rho_{\min}^2c^2}{8|\mathcal{S}|} + \frac{19\log(2)L^2\alpha_k}{\tau\rho_{\min}})\delta_k^\phi.
\end{aligned}$$

With the step sizes $\alpha_k = \alpha$, $\beta_k = \beta$ such that $\frac{\alpha}{\beta} \leqslant \min\{\frac{(1-\gamma)\tau^2\rho_{\min}^3c^2}{152|\mathcal{S}|\log(2)L^2}, 8\}$, we can simplify the inequality above

$$\begin{aligned}
3\delta_{k+1}^\pi + \delta_{k+1}^\phi &\leqslant 3(1 - \frac{\alpha_k(1-\gamma)\tau\rho_{\min}^2c^2}{32|\mathcal{S}|})\delta_k^\pi + (1 - \frac{3(1-\gamma)\beta_k\tau\rho_{\min}^2c^2}{8|\mathcal{S}|} + \frac{19\log(2)L^2\alpha_k}{\tau\rho_{\min}})\delta_k^\phi \\
&\leqslant 3(1 - \frac{\alpha(1-\gamma)\tau\rho_{\min}^2c^2}{32|\mathcal{S}|})\delta_k^\pi + (1 - \frac{(1-\gamma)\beta\tau\rho_{\min}^2c^2}{4|\mathcal{S}|})\delta_k^\phi \\
&\leqslant (1 - \frac{\alpha(1-\gamma)\tau\rho_{\min}^2c^2}{32|\mathcal{S}|})(3\delta_k^\pi + \delta_k^\phi) \\
&\leqslant (1 - \frac{\alpha(1-\gamma)\tau\rho_{\min}^2c^2}{32|\mathcal{S}|})^{k+1}(3\delta_0^\pi + \delta_0^\phi).
\end{aligned}$$

□

C.1.2 Proof of Corollary 4.1

As a result of Lemma 4.3, it is easy to verify

$$(3\delta_{t+1,0}^\pi + \delta_{t+1,0}^\phi) - (3\delta_{t,K_t}^\pi + \delta_{t,K_t}^\phi)$$

$$
\begin{aligned}
&= (3J_{\tau_{t+1}}(\pi^\star_{\tau_{t+1}}, \phi^\star_{\tau_{t+1}}) - 3J_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}, \phi_{\tau_{t+1}}(\pi_{\theta_{t+1,0}})) \\
&\quad + J_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}, \phi_{\psi_{t+1,0}}) - J_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}, \phi_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}))) \\
&\quad - (3J_{\tau_t}(\pi^\star_{\tau_t}, \phi^\star_{\tau_t}) - 3J_{\tau_t}(\pi_{\theta_{t,K_t}}, \phi_{\tau_t}(\pi_{\theta_{t,K_t}})) \\
&\quad + J_{\tau_t}(\pi_{\theta_{t,K_t}}, \phi_{\psi_{t,K_t}}) - J_{\tau_t}(\pi_{\theta_{t,K_t}}, \phi_{\tau_t}(\pi_{\theta_{t,K_t}})) \\
&= (3J_{\tau_{t+1}}(\pi^\star_{\tau_{t+1}}, \phi^\star_{\tau_{t+1}}) - 3J_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}, \phi_{\tau_{t+1}}(\pi_{\theta_{t+1,0}})) \\
&\quad + J_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}, \phi_{\psi_{t+1,0}}) - J_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}, \phi_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}))) \\
&\quad - (3J_{\tau_t}(\pi^\star_{\tau_t}, \phi^\star_{\tau_t}) - 3J_{\tau_t}(\pi_{\theta_{t+1,0}}, \phi_{\tau_t}(\pi_{\theta_{t+1,0}})) \\
&\quad + J_{\tau_t}(\pi_{\theta_{t+1,0}}, \phi_{\psi_{t+1,0}}) - J_{\tau_t}(\pi_{\theta_{t+1,0}}, \phi_{\tau_t}(\pi_{\theta_{t+1,0}}))) \\
&= 3(J_{\tau_{t+1}}(\pi^\star_{\tau_{t+1}}, \phi^\star_{\tau_{t+1}}) - J_{\tau_t}(\pi^\star_{\tau_t}, \phi^\star_{\tau_t})) \\
&\quad - 4(J_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}, \phi_{\tau_{t+1}}(\pi_{\theta_{t+1,0}})) - J_{\tau_t}(\pi_{\theta_{t+1,0}}, \phi_{\tau_t}(\pi_{\theta_{t+1,0}}))) \\
&\quad + (J_{\tau_{t+1}}(\pi_{\theta_{t+1,0}}, \phi_{\psi_{t+1,0}}) - J_{\tau_t}(\pi_{\theta_{t+1,0}}, \phi_{\psi_{t+1,0}})) \\
&\leqslant L_\delta(\tau_t - \tau_{t+1}). \qquad \text{(C.16)}
\end{aligned}
$$

We can choose τ_0 large enough that

$$3\delta^\pi_{0,0} + \delta^\phi_{0,0} \leqslant C_1\tau_0$$

holds. For any $t \geqslant 0$, if we run the inner loop for K_t iterations such that

$$3\delta^\pi_{t,K_t} + \delta^\phi_{t,K_t} \leqslant \frac{1}{2}(3\delta^\pi_{t,0} + \delta^\phi_{t,0}) \leqslant \frac{C_1\tau_t}{2},$$

then we have

$$
\begin{aligned}
3\delta^\pi_{t+1,0} + \delta^\phi_{t+1,0} &\leqslant 3\delta^\pi_{t,K_t} + \delta^\phi_{t,K_t} + L_\delta(\tau_t - \tau_{t+1}) \leqslant \frac{C_1\tau_t}{2} + L_\delta(\tau_t - \tau_{t+1}) \\
&= \frac{(C_1 + L_\delta)C_1}{C_1 + 2L_\delta}\tau_{t+1} + \frac{C_1 L_\delta}{C_1 + 2L_\delta}\tau_{t+1} = C_1\tau_{t+1},
\end{aligned}
$$

where the first equality plugs in $\tau_t = \frac{2C_1+2L_\delta}{C_1+2L_\delta}\tau_{t+1}$. This means that the initial condition

Equation 4.12 is observed at the beginning of the every outer loop iteration.

Applying the inequality recursively,

$$3\delta^{\pi}_{T,0} + \delta^{\phi}_{T,0} \leqslant C_1 \tau_T.$$

With an argument similar to the one in Equation C.16, we can show

$$\begin{aligned}(3(J(\pi^\star, \phi^\star) - J(\pi_{\theta_{T,0}}, \phi_0(\pi_{\theta_{T,0}}))) \\ + (J(\pi_{\theta_{T,0}}, \phi_{\psi_{T,0}}) - J(\pi_{\theta_{T,0}}, \phi_0(\pi_{\theta_{T,0}})))) - (3\delta^{\pi}_{T,0} + \delta^{\phi}_{T,0}) \leqslant L_\delta \tau_T.\end{aligned}$$

In order to achieve Equation 4.13, it suffices to guarantee $3\delta^{\pi}_{T,0} + \delta^{\phi}_{T,0} + L_\delta \tau_T \leqslant \epsilon$, or $(C_1 + L_\delta)\tau_T \leqslant \epsilon$. This implies that we need $\tau_T = \mathcal{O}(\epsilon)$, or equivalently, $T = \mathcal{O}(\log(\epsilon^{-1}))$ since $\tau_T = \left(\frac{C_1 + 2L_\delta}{2C_1 + 2L_\delta}\right)^T \tau_0$.

Ultimately we are interested in bounding $\sum_{t=0}^{T} K_t$. Note that K_t needs to be at most

$$K_t \leqslant \lceil \frac{\log(\frac{1}{2})}{\log(1 - \frac{\alpha_t (1-\gamma) \tau_t \rho^2_{\min} c^2}{32|\mathcal{S}|})} \rceil.$$

To apply Theorem 4.1, we need to select the step sizes that satisfy the required condition. Since $\{\tau_t\}$ is a decaying sequence, the smoothness constant $L = 3L_{\mathcal{H}} \max\{\tau_0, 1\}$ is valid across all outer loop iterations t.

We use $L_t = 3L_{\mathcal{H}} \max\{\tau_t, 1\}$ to denote the smoothness constant of the regularized value function in outer loop iteration t and use T_1 to denote the index of the outer loop iteration such that $\tau_{T_1} \geqslant 1$ and $\tau_{T_1+1} < 1$. Note that T_1 is an absolute constant that only depends on the structure of the Markov game. From iterations $t = 0$ to $t = T_1$, the smoothness constant is proportional to regularization weight $L_t = 3L_{\mathcal{H}} \max\{\tau_t, 1\} = 3L_{\mathcal{H}} \tau_t$. We need to choose α_t, β_t such that

$$\beta_t \leqslant \frac{1}{L_t} = \frac{1}{3L_{\mathcal{H}} \tau_t}, \quad \frac{\alpha_t}{\beta_t} \leqslant \min\{\frac{(1-\gamma)\rho^3_{\min} c^2 \tau_t^2}{152 \log(2) |\mathcal{S}| L_t^2}, 8\} = \min\{\frac{(1-\gamma)\rho^3_{\min} c^2}{1368 \log(2) |\mathcal{S}| L_{\mathcal{H}}^2}, 8\},$$

$$\alpha_t \leqslant \min\{(L_t + \frac{2\sqrt{|\mathcal{S}|}L_t^2}{\sqrt{(1-\gamma)\rho_{\min}}\tau_t c})^{-1}, \frac{16|\mathcal{S}|}{(1-\gamma)\rho_{\min}^2 c^2 \tau_t}\}$$
$$= \min\{(3L_{\mathcal{H}}\tau_t + \frac{18\sqrt{|\mathcal{S}|}L_{\mathcal{H}}^2\tau_t}{\sqrt{(1-\gamma)\rho_{\min}}c})^{-1}, \frac{16|\mathcal{S}|}{(1-\gamma)\rho_{\min}^2 c^2 \tau_t}\}.$$

Then it is obvious that we can choose $\alpha_t = \mathcal{O}(\tau_t^{-1})$, implying $\alpha_t\tau_t = \mathcal{O}(1)$. Therefore, for all $0 \leqslant t \leqslant T_1$,

$$K_t \leqslant \lceil \frac{\log(\frac{1}{2})}{\log(1 - \frac{\alpha_t(1-\gamma)\tau_t\rho_{\min}^2 c^2}{32|\mathcal{S}|})} \rceil = \mathcal{O}(1). \tag{C.17}$$

From iterations $t = T_1$ until $t = T$, the smoothness constant is $L_t = 3L_{\mathcal{H}} \max\{\tau_t, 1\} = 3L_{\mathcal{H}}$. Note that there is an upper and lower bound on β_t. In order for the upper bound to be no smaller than the lower bound, we need

$$\frac{152\log(2)|\mathcal{S}|L^2\alpha_t}{(1-\gamma)\rho_{\min}^3 c^2\tau_t^2} \leqslant \frac{1}{L}.$$

This means that we should choose $\alpha_t = \mathcal{O}(\tau_t^2)$, implying $\alpha_t\tau_t = \mathcal{O}(\tau_t^3)$. Plugging it in Equation C.17,

$$K_t = \lceil \frac{\log(\frac{1}{2})}{\log(1 - \frac{\alpha_t(1-\gamma)\tau_t\rho_{\min}^2 c^2}{32|\mathcal{S}|})} \rceil = \mathcal{O}(\frac{1}{\log(1-\tau_t^3)}) \leqslant \mathcal{O}(\tau_t^{-3}),$$

where the last inequality follows from the fact that $1 + x \leqslant \exp(x)$ for any scalar x.

Since $\tau_t = \tau_T(\frac{2C_1+2L_\delta}{C_1+2L_\delta})^{T-t}$,

$$\sum_{t=0}^{T} K_t = \sum_{t=0}^{T_1} K_t + \sum_{t=T_1}^{T} K_t \leqslant \sum_{t=0}^{T} \mathcal{O}(\tau_t^{-3}) = \mathcal{O}(1) + \sum_{t=T_1}^{T} \mathcal{O}(\tau_T^{-3}(\frac{2C_1+2L_\delta}{C_1+2L_\delta})^{-3(T-t)})$$
$$\leqslant \mathcal{O}(\tau_T^{-3}\sum_{t=0}^{T}(\frac{C_1+2L_\delta}{2C_1+2L_\delta})^{3(T-t)}) = \mathcal{O}(\tau_T^{-3}\sum_{t=0}^{T}(\frac{C_1+2L_\delta}{2C_1+2L_\delta})^{3t})$$
$$\leqslant \mathcal{O}(\tau_T^{-3}\frac{1}{1-(\frac{C_1+2L_\delta}{2C_1+2L_\delta})^3}) = \mathcal{O}(\tau_T^{-3}).$$

Since $\tau_T = \mathcal{O}(\epsilon)$,

$$\sum_{t=0}^{T} K_t \leqslant \mathcal{O}(\tau_T^{-3}) = \mathcal{O}(\epsilon^{-3}).$$

□

C.1.3 Proof of Theorem 4.2

Define $L_0 = L_{\mathcal{H}}(2\tau_0 + 1)$. The exact conditions on the initial step sizes, regularization weight, and h are

$$\delta_0^{\pi} + \delta_0^{\phi} \leqslant \frac{C_1 \tau_0}{h^{\frac{1}{3}}}, \tag{C.18}$$

$$\alpha_0 = \frac{65536\log(2)(\log|\mathcal{A}| + \log|\mathcal{B}|) + 96(1-\gamma)\rho_{\min}c^2}{3(1-\gamma)^2\rho_{\min}^3 c^4 \tau_0}, \tag{C.19}$$

$$\frac{\alpha_0}{h^{\frac{2}{3}}} \leqslant (2L_{\mathcal{H}} + 4L_{\mathcal{H}}^2 C_2)\frac{\tau_0}{h^{\frac{1}{3}}} + (L_{\mathcal{H}} + 4L_{\mathcal{H}}^2 C_2) + \frac{L_{\mathcal{H}}^2 C_2 h^{\frac{1}{3}}}{\tau_0}, \tag{C.20}$$

$$\beta_0 \leqslant \frac{1}{L_0}, \quad \frac{\alpha_0}{\beta_0} \leqslant \min\{\frac{(1-\gamma)\tau_0^2\rho_{\min}^3 c^2}{152\log(2)|\mathcal{S}|L_0^2}, 1\}. \tag{C.21}$$

In Remark C.1 at the end of this section, we show that there always exist α_0, β_0, τ_0, and h that observe the conditions.

Equation C.3 implies that for any $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, $\psi_1, \psi_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{B}|}$, and $k \geqslant 0$,

$$\begin{aligned}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta J_{\tau_k}(\pi_{\theta_2}, \phi_{\psi_2})\| &\leqslant (L_V + 2\tau_k L_{\mathcal{H}})(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|) \\ &\leqslant L_0(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|),\end{aligned} \tag{C.22}$$

where the last inequality follows from $\tau_k \leqslant \tau_0$.

Convergence of $3\delta_k^{\pi} + \delta_k^{\phi}$:

We will first use an induction argument to prove

$$3\delta_k^{\pi} + \delta_k^{\phi} \leqslant \frac{\rho_{\min}\tau_0 c^2}{64\log(2)(k+h)^{1/3}}, \quad \forall k \geqslant 0.$$

The base case is $3\delta_0^\pi + \delta_0^\phi \leqslant \frac{\rho_{\min} c^2 \tau_0}{64 \log(2) h^{\frac{1}{3}}}$, which holds by the initial condition. Now, suppose

$$3\delta_k^\pi + \delta_k^\phi \leqslant \frac{\rho_{\min} \tau_0 c^2}{64 \log(2)(k+h)^{1/3}} \tag{C.23}$$

holds. We aim to show

$$3\delta_{k+1}^\pi + \delta_{k+1}^\phi \leqslant \frac{\rho_{\min} \tau_0 c^2}{64 \log(2)(k+1+h)^{1/3}}.$$

We introduce the following technical lemmas.

Lemma C.6. *Suppose Equation C.23 holds. Then, we have*

$$-\left(\min_{s,a} \pi_{\theta_k}(a \mid s)\right)^2 \leqslant -\frac{3c^2}{8}, \tag{C.24}$$

$$-\left(\min_{s,b} \phi_{\psi_k}(b \mid s)\right)^2 \leqslant -\frac{3c^2}{8}. \tag{C.25}$$

Lemma C.7. *Suppose Equation C.23 holds. Under Assumption 4.1 and Assumption 4.2 and the step sizes of Theorem 4.2, we have*

$$\begin{aligned}
&g_{\tau_k}(\theta_k) - g_{\tau_k}(\theta_{k+1}) \\
&= J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_k}(\pi_{\theta_{k+1}})) \\
&\leqslant \frac{\alpha_k}{2}\left(\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\|^2\right).
\end{aligned}$$

We perform the following decomposition

$$\begin{aligned}
&\delta_{k+1}^\pi - \delta_k^\pi \\
&= J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_{k+1}}(\pi_{\theta_{k+1}}, \phi_{\tau_{k+1}}(\pi_{\theta_{k+1}})) + J_{\tau_{k+1}}(\pi^\star_{\tau_{k+1}}, \phi^\star_{\tau_{k+1}}) - J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) \\
&= J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_k}(\pi_{\theta_{k+1}})) \\
&\quad + J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_k}(\pi_{\theta_{k+1}})) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_{k+1}}(\pi_{\theta_{k+1}}))
\end{aligned}$$

$$
\begin{aligned}
&+ J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_{k+1}}(\pi_{\theta_{k+1}})) - J_{\tau_{k+1}}(\pi_{\theta_{k+1}}, \phi_{\tau_{k+1}}(\pi_{\theta_{k+1}})) \\
&+ J_{\tau_{k+1}}(\pi^\star_{\tau_{k+1}}, \phi^\star_{\tau_{k+1}}) - J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) \\
\leqslant\ & J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_k}(\pi_{\theta_{k+1}})) + \frac{\tau_k - \tau_{k+1}}{1-\gamma} \log|\mathcal{A}| + (\tau_k - \tau_{k+1}) \log|\mathcal{B}| \\
\leqslant\ & \frac{\alpha_k}{2} \left(\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\|^2 \right) \\
&+ \frac{\tau_k - \tau_{k+1}}{1-\gamma} (\log|\mathcal{A}| + \log|\mathcal{B}|)
\end{aligned} \tag{C.26}
$$

where the first inequality comes from $J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_k}(\pi_{\theta_{k+1}})) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_{k+1}}(\pi_{\theta_{k+1}})) \leqslant 0$ by the definition of $\phi_\tau(\cdot)$ and the bound on $J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_{k+1}}(\pi_{\theta_{k+1}})) - J_{\tau_{k+1}}(\pi_{\theta_{k+1}}, \phi_{\tau_{k+1}}(\pi_{\theta_{k+1}}))$ and $J_{\tau_{k+1}}(\pi^\star_{\tau_{k+1}}, \phi^\star_{\tau_{k+1}}) - J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k})$ from Lemma 4.3 Equation 4.8 and Equation 4.6. The second inequality uses Lemma C.7.

Similarly, we consider the decay of δ^ϕ_k.

$$
\begin{aligned}
\delta^\phi_{k+1} - \delta^\phi_k &= J_{\tau_{k+1}}(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - g_{\tau_{k+1}}(\pi_{\theta_{k+1}}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) + g_{\tau_k}(\pi_{\theta_k}) \\
&= \left(J_{\tau_{k+1}}(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}})\right) + \left(J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_k})\right) \\
&\quad + \left(J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\right) + \left(g_{\tau_k}(\pi_{\theta_k}) - g_{\tau_{k+1}}(\pi_{\theta_{k+1}})\right).
\end{aligned} \tag{C.27}
$$

By Lemma 4.3 Equation 4.8,

$$
J_{\tau_{k+1}}(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) \leqslant \frac{\tau_k - \tau_{k+1}}{1-\gamma} \log|\mathcal{B}|. \tag{C.28}
$$

Using the L_0-smoothness of the value function derived in Equation C.22

$$
\begin{aligned}
&J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_{k+1}}) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_k}) \\
&\leqslant \left\langle \nabla_\psi J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_k}), \psi_{k+1} - \psi_k \right\rangle + \frac{L_0}{2} \|\psi_{k+1} - \psi_k\|^2 \\
&= -\beta_k \|\nabla_\psi J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_k})\|^2 + \frac{L_0 \beta_k^2}{2} \|\nabla_\psi J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_k})\|^2 \\
&\leqslant -\frac{\beta_k}{2} \|\nabla_\psi J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\psi_k})\|^2
\end{aligned}
$$

$$\leqslant -\frac{(1-\gamma)\beta_k\tau_k\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,b}\phi_{\psi_k}(b\mid s)\right)^2\left(J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})-J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\right)$$
$$= -\frac{(1-\gamma)\beta_k\tau_k\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,b}\phi_{\psi_k}(b\mid s)\right)^2\delta_k^{\phi},$$

where the second inequality uses $\beta_k \leqslant \frac{1}{L_0}$ and the third inequality follows from Lemma 4.4.

Using Equation C.25 of Lemma C.6 to further simplify this inequality,

$$J_{\tau_k}(\pi_{\theta_{k+1}},\phi_{\psi_{k+1}})-J_{\tau_k}(\pi_{\theta_{k+1}},\phi_{\psi_k})\leqslant -\frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2c^2}{8|\mathcal{S}|}\delta_k^{\phi}. \tag{C.29}$$

For the third term of Equation C.27, we have from the L_0-smoothness of the value function derived in Equation C.22

$$\begin{aligned}J_{\tau_k}(\pi_{\theta_{k+1}},\phi_{\psi_k})-J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k}) &\leqslant \langle\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k}),\theta_{k+1}-\theta_k\rangle+\frac{L_0}{2}\|\theta_{k+1}-\theta_k\|^2\\ &= \alpha_k\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2+\frac{L_0\alpha_k^2}{2}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2\\ &\leqslant \frac{3\alpha_k}{2}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2,\end{aligned} \tag{C.30}$$

where in the last inequality we use $\alpha_k L_0 \leqslant 1$.

Using Lemma C.7 and Lemma 4.3 Equation 4.7, we bound the last term of Equation C.27

$$\begin{aligned}&g_{\tau_k}(\pi_{\theta_k})-g_{\tau_{k+1}}(\pi_{\theta_{k+1}})\\ &= g_{\tau_k}(\pi_{\theta_k})-g_{\tau_k}(\pi_{\theta_{k+1}})+g_{\tau_k}(\pi_{\theta_{k+1}})-g_{\tau_{k+1}}(\pi_{\theta_{k+1}})\\ &\leqslant \frac{\alpha_k}{2}\left(\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))-\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2-\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\|^2\right)\\ &\quad+(\tau_k-\tau_{k+1})\log|\mathcal{A}|\end{aligned} \tag{C.31}$$

Using Equation C.28-Equation C.31 in Equation C.27, we have

$$\begin{aligned}\delta_{k+1}^{\phi} = \delta_k^{\phi}&+\left(J_{\tau_{k+1}}(\pi_{\theta_{k+1}},\phi_{\psi_{k+1}})-J_{\tau_k}(\pi_{\theta_{k+1}},\phi_{\psi_{k+1}})\right)+\left(J_{\tau_k}(\pi_{\theta_{k+1}},\phi_{\psi_{k+1}})-J_{\tau_k}(\pi_{\theta_{k+1}},\phi_{\psi_k})\right)\\ &+\left(J_{\tau_k}(\pi_{\theta_{k+1}},\phi_{\psi_k})-J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\right)+\left(g_{\tau_k}(\pi_{\theta_k})-g_{\tau_{k+1}}(\pi_{\theta_{k+1}})\right)\end{aligned}$$

$$
\begin{aligned}
&\leqslant \delta_k^\phi + \frac{\tau_k - \tau_{k+1}}{1-\gamma}\log|\mathcal{B}| - \frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2 c^2}{8|\mathcal{S}|}\delta_k^\phi + \frac{3\alpha_k}{2}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 \\
&\quad + \frac{\alpha_k}{2}\left(\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 - \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\|^2\right) \\
&\quad + (\tau_k - \tau_{k+1})\log|\mathcal{A}| \\
&\leqslant (1 - \frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2 c^2}{8|\mathcal{S}|})\delta_k^\phi + \frac{3\alpha_k}{2}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 \\
&\quad + \frac{\alpha_k}{2}\left(\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 - \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\|^2\right) \\
&\quad + \frac{\tau_k - \tau_{k+1}}{1-\gamma}(\log|\mathcal{A}| + \log|\mathcal{B}|). \qquad \text{(C.32)}
\end{aligned}
$$

Combining Equation C.26 and Equation C.32,

$$
\begin{aligned}
&3\delta_{k+1}^\pi + \delta_{k+1}^\phi \\
&\leqslant 3\delta_k^\pi + \frac{3\alpha_k}{2}\left(\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 - \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\|^2\right) \\
&\quad + \frac{3(\tau_k - \tau_{k+1})}{1-\gamma}(\log|\mathcal{A}| + \log|\mathcal{B}|) + (1 - \frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2 c^2}{8|\mathcal{S}|})\delta_k^\phi \\
&\quad + \frac{\alpha_k}{2}\left(\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 - \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\|^2\right) \\
&\quad + \frac{3\alpha_k}{2}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 + \frac{\tau_k - \tau_{k+1}}{1-\gamma}(\log|\mathcal{A}| + \log|\mathcal{B}|) \\
&\leqslant 3\delta_k^\pi + (1 - \frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2 c^2}{8|\mathcal{S}|})\delta_k^\phi + \frac{3\alpha_k}{2}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 \\
&\quad + 2\alpha_k\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 - 2\alpha_k\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\|^2 \\
&\quad + \frac{4(\tau_k - \tau_{k+1})}{1-\gamma}(\log|\mathcal{A}| + \log|\mathcal{B}|).
\end{aligned}
$$

Simplifying this inequality with

$$
\begin{aligned}
\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 &= \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - (\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k}))\|^2 \\
&\leqslant \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\|^2 + \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 \\
&\quad + 2\langle \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})), \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\rangle \\
&\leqslant \frac{5}{4}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\|^2 + 5\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2
\end{aligned}
$$

we have

$$3\delta_{k+1}^{\pi} + \delta_{k+1}^{\phi} \leqslant 3\delta_k^{\pi} + (1 - \frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2 c^2}{8|\mathcal{S}|})\delta_k^{\phi} - \frac{\alpha_k}{8}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\|^2 + \frac{19\alpha_k}{2}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2 + \frac{4(\tau_k - \tau_{k+1})}{1-\gamma}(\log|\mathcal{A}| + \log|\mathcal{B}|) \tag{C.33}$$

Using Lemma 4.4 to bound $-\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\|^2$,

$$\begin{aligned}
&-\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\|^2 \\
&\leqslant -\frac{2(1-\gamma)\tau_k\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 (J_{\tau_k}(\pi_{\tau_k}(\phi_{\tau_k}(\pi_{\theta_k})), \phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))) \\
&\leqslant -\frac{2(1-\gamma)\tau_k\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 \left(J_{\tau_k}(\pi_{\tau_k}^\star, \phi_{\tau_k}^\star) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\right),
\end{aligned} \tag{C.34}$$

where the second inequality follows from

$$J_{\tau_k}(\pi_{\tau_k}(\phi_{\tau_k}(\pi_{\theta_k})), \phi_{\tau_k}(\pi_{\theta_k})) = \max_\pi J_{\tau_k}(\pi, \phi_{\tau_k}(\pi_{\theta_k})) \geqslant \max_\pi \min_\phi J_{\tau_k}(\pi, \phi) = J_{\tau_k}(\pi_{\tau_k}^\star, \phi_{\tau_k}^\star).$$

From Equation C.24 of Lemma C.6 , $-\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 \leqslant -\frac{3c^2}{8}$, which further simplifies Equation C.34

$$\begin{aligned}
&-\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\|^2 \\
&\leqslant -\frac{2(1-\gamma)\tau_k\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 \left(J_{\tau_k}(\pi_{\tau_k}^\star, \phi_{\tau_k}^\star) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\right) \\
&= -\frac{2(1-\gamma)\tau_k\rho_{\min}^2}{|\mathcal{S}|}\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 \delta_k^{\pi} \leqslant -\frac{3(1-\gamma)\tau_k\rho_{\min}^2 c^2}{4|\mathcal{S}|}\delta_k^{\pi}.
\end{aligned} \tag{C.35}$$

For $\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2$, we have from the L_0-smoothness of the value function derived in Equation C.22

$$\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \leqslant L_0^2\|\phi_{\tau_k}(\pi_{\theta_k}) - \phi_{\psi_k}\|^2$$

$$\leqslant \frac{2\log(2)L_0^2}{\tau_k\rho_{\min}}\left(J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})-J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\right)$$
$$= \frac{2\log(2)L_0^2}{\tau_k\rho_{\min}}\delta_k^\phi, \tag{C.36}$$

where the second inequality follows from Lemma Equation 4.4 of 4.1 .

Using Equation C.35 and Equation C.36 in Equation C.33,

$$\begin{aligned}
&3\delta_{k+1}^\pi + \delta_{k+1}^\phi \\
&\leqslant 3\delta_k^\pi + (1-\frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2c^2}{8|\mathcal{S}|})\delta_k^\phi - \frac{\alpha_k}{8}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k}))\|^2 \\
&\quad + \frac{19\alpha_k}{2}\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2 + \frac{4(\tau_k-\tau_{k+1})}{1-\gamma}(\log|\mathcal{A}|+\log|\mathcal{B}|) \\
&\leqslant 3\delta_k^\pi + (1-\frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2c^2}{8|\mathcal{S}|})\delta_k^\phi - \frac{3\alpha_k(1-\gamma)\tau_k\rho_{\min}^2c^2}{32|\mathcal{S}|}\delta_k^\pi \\
&\quad + \frac{19\log(2)L_0^2\alpha_k}{\tau_k\rho_{\min}}\delta_k^\phi + \frac{4(\tau_k-\tau_{k+1}))}{1-\gamma}(\log|\mathcal{A}|+\log|\mathcal{B}|) \\
&= 3(1-\frac{(1-\gamma)\alpha_k\tau_k\rho_{\min}^2c^2}{32|\mathcal{S}|})\delta_k^\pi + (1-\frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2c^2}{8|\mathcal{S}|} + \frac{19\log(2)L_0^2\alpha_k}{\tau_k\rho_{\min}})\delta_k^\phi \\
&\quad + \frac{4(\tau_k-\tau_{k+1})}{1-\gamma}(\log|\mathcal{A}|+\log|\mathcal{B}|).
\end{aligned} \tag{C.37}$$

With the step size rule $\frac{\alpha_0}{\beta_0} \leqslant \min\{\frac{(1-\gamma)\tau_0^2\rho_{\min}^3c^2}{152\log(2)L_0^2|\mathcal{S}|}, 1\}$, we can simplify Equation C.37,

$$\begin{aligned}
3\delta_{k+1}^\pi + \delta_{k+1}^\phi &\leqslant 3(1-\frac{(1-\gamma)\alpha_k\tau_k\rho_{\min}^2c^2}{32|\mathcal{S}|})\delta_k^\pi + (1-\frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2c^2}{8|\mathcal{S}|} + \frac{19\log(2)L_0^2\alpha_k}{\tau_k\rho_{\min}})\delta_k^\phi \\
&\quad + \frac{4(\tau_k-\tau_{k+1})}{1-\gamma}(\log|\mathcal{A}|+\log|\mathcal{B}|) \\
&\leqslant 3(1-\frac{(1-\gamma)\alpha_k\tau_k\rho_{\min}^2c^2}{32|\mathcal{S}|})\delta_k^\pi \\
&\quad + (1-\frac{3(1-\gamma)\beta_k\tau_k\rho_{\min}^2c^2}{8|\mathcal{S}|} + \frac{19\log(2)L_0^2}{\tau_k\rho_{\min}}\frac{(1-\gamma)\rho_{\min}^3c^2\tau_k^2\beta_k}{152\log(2)L_0^2|\mathcal{S}|})\delta_k^\phi \\
&\quad + \frac{4(\tau_k-\tau_{k+1})}{1-\gamma}(\log|\mathcal{A}|+\log|\mathcal{B}|) \\
&\leqslant 3(1-\frac{(1-\gamma)\alpha_k\tau_k\rho_{\min}^2c^2}{32|\mathcal{S}|})\delta_k^\pi + (1-\frac{(1-\gamma)\beta_k\tau_k\rho_{\min}^2c^2}{4|\mathcal{S}|})\delta_k^\phi \\
&\quad + \frac{4(\tau_k-\tau_{k+1})}{1-\gamma}(\log|\mathcal{A}|+\log|\mathcal{B}|)
\end{aligned}$$

$$\leqslant (1-\frac{(1-\gamma)\alpha_k\tau_k\rho_{\min}^2c^2}{32|\mathcal{S}|})(3\delta_k^\pi+\delta_k^\phi)+\frac{4(\tau_k-\tau_{k+1})}{1-\gamma}(\log|\mathcal{A}|+\log|\mathcal{B}|)$$
$$\leqslant (1-\frac{(1-\gamma)\rho_{\min}^2c^2\alpha_0\tau_0}{32|\mathcal{S}|(k+h)})\frac{C_1}{(k+h)^{1/3}}+\frac{32\tau_0}{3(1-\gamma)(k+h)^{4/3}}(\log|\mathcal{A}|+\log|\mathcal{B}|),$$

where the last inequality follows from Equation C.23 and Lemma C.3.

Letting $D_1=\frac{(1-\gamma)\rho_{\min}^2c^2}{32|\mathcal{S}|}$ and $D_2=\frac{32}{3(1-\gamma)}(\log|\mathcal{A}|+\log|\mathcal{B}|)$,

$$\begin{aligned}3\delta_{k+1}^\pi+\delta_{k+1}^\phi &\leqslant \left(1-\frac{D_1\alpha_0\tau_0}{k+h}\right)\frac{C_1\tau_0}{(k+h)^{1/3}}+\frac{D_2\tau_0}{(k+1)^{4/3}}\\ &=\left(k+h-D_1\alpha_0\tau_0+\frac{D_2}{C_1}\right)\frac{C_1\tau_0}{(k+h)^{4/3}}.\end{aligned}$$

By requiring

$$\tau_0=\frac{65536\log(2)(\log|\mathcal{A}|+\log|\mathcal{B}|)+96(1-\gamma)\rho_{\min}c^2}{3(1-\gamma)^2\rho_{\min}^3c^4\alpha_0}=\frac{1}{D_1\alpha_0}(1+\frac{D_2}{C_1}),$$

we have

$$\begin{aligned}3\delta_{k+1}^\pi+\delta_{k+1}^\phi &\leqslant \left(k+h-D_1\alpha_0\tau_0+\frac{D_2}{C_1}\right)\cdot\frac{C_1\tau_0}{(k+h)^{4/3}}\\ &=\left(k+h-(1+\frac{D_2}{C_1})+\frac{D_2}{C_1}\right)\cdot\frac{C_1\tau_0}{(k+h)^{4/3}}\\ &=\frac{C_1\tau_0(k-1+h)}{(k+h)^{4/3}},\end{aligned}$$

Since $(k-1+h)^3(k+1+h)\leqslant(k+h)^4$ for all $k\geqslant 0$ and $h\geqslant 1$, we have

$$\frac{k-1+h}{(k+h)^{4/3}}=\frac{(k-1+h)(k+1+h)^{1/3}}{(k+1)^{4/3}(k+1+h)^{1/3}}\leqslant\frac{(k+h)^{4/3}}{(k+h)^{4/3}(k+1+h)^{1/3}}=\frac{1}{(k+1+h)^{1/3}},$$

which leads to

$$3\delta_{k+1}^\pi+\delta_{k+1}^\phi\leqslant\frac{C_1\tau_0(k-1+h)}{(k+h)^{4/3}}\leqslant\frac{C_1\tau_0}{(k+1+h)^{1/3}}=\frac{\rho_{\min}\tau_0c^2}{64\log(2)(k+1+h)^{1/3}}.$$

This finishes our induction and implies that for all $k \geqslant 0$

$$J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) \leqslant \frac{C_1\tau_0}{3(k+h)^{1/3}},$$
$$J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) \leqslant \frac{C_1\tau_0}{(k+h)^{1/3}}.$$

Bounding the difference between value functions with and without the regularization:

Ultimately, we are interested in $J(\pi^\star, \phi^\star) - J(\pi_{\theta_k}, \phi_0(\pi_{\theta_k}))$ and $J(\pi_{\theta_k}, \phi_{\psi_k}) - J(\pi_{\theta_k}, \phi_0(\pi_{\theta_k}))$, which measure the performance of π_{θ_k} and ϕ_{ψ_k} in the original unregularized Markov game.

By Equation 4.6, Equation 4.7, and Equation 4.8,

$$J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) - J(\pi^\star, \phi^\star) \geqslant -\tau_k \log|\mathcal{B}|$$
$$J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J(\pi_{\theta_k}, \phi_0(\pi_{\theta_k})) \leqslant \tau_k \log|\mathcal{A}|$$
$$J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) - J(\pi_{\theta_k}, \phi_{\psi_k}) \geqslant -\frac{\tau_k}{1-\gamma}\log|\mathcal{B}|.$$

Therefore,

$$\begin{aligned}
J(\pi^\star, \phi^\star) - J(\pi_{\theta_k}, \phi_0(\pi_{\theta_k})) &= J(\pi^\star, \phi^\star) - J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) + J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) \\
&\quad + J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J(\pi_{\theta_k}, \phi_0(\pi_{\theta_k})) \\
&\leqslant \tau_k \log|\mathcal{B}| + \frac{C_1\tau_0}{3(k+h)^{1/3}} + \tau_k \log|\mathcal{A}| \\
&= \frac{C_1\tau_0 + 3(\log|\mathcal{A}| + \log|\mathcal{B}|)\tau_0}{3(k+h)^{1/3}},
\end{aligned}$$

and

$$\begin{aligned}
J(\pi_{\theta_k}, \phi_{\psi_k}) - J(\pi_{\theta_k}, \phi_0(\pi_{\theta_k})) &= J(\pi_{\theta_k}, \phi_{\psi_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) \\
&\quad + J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))
\end{aligned}$$

$$
\begin{aligned}
&\quad + J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J(\pi_{\theta_k}, \phi_0(\pi_{\theta_k})) \\
&\leqslant \frac{\tau_k}{1-\gamma} \log|\mathcal{B}| + \frac{C_1 \tau_0}{(k+h)^{1/3}} + \tau_k \log|\mathcal{A}| \\
&\leqslant \frac{(1-\gamma) C_1 \tau_0 + (\log|\mathcal{A}| + \log|\mathcal{B}|)\tau_0}{(1-\gamma)(k+h)^{1/3}}.
\end{aligned}
$$

Remark C.1. *To select α_0, β_0, τ_0, and h, we first make $\tau_0 = \lambda h^{1/3}$ for some $\lambda > 0$ large enough. This choice guarantees the validity of Equation C.18 (we just need $\delta_0^\pi + \delta_0^\phi \leqslant C_1 \lambda$). Viewing Equation C.19, it means*

$$
\alpha_0 = \frac{65536 \log(2)(\log|\mathcal{A}| + \log|\mathcal{B}|) + 96(1-\gamma)\rho_{\min} c^2}{3(1-\gamma)^2 \rho_{\min}^3 c^4 \lambda h^{\frac{1}{3}}}.
$$

Now that λ is fixed, to ensure Equation C.20, we choose h large enough to observe

$$
\begin{aligned}
\frac{65536 \log(2)(\log|\mathcal{A}| + \log|\mathcal{B}|) + 96(1-\gamma)\rho_{\min} c^2}{3(1-\gamma)^2 \rho_{\min}^3 c^4 \lambda h} &= \frac{\alpha_0}{h^{\frac{2}{3}}} \\
&\leqslant (2L_{\mathcal{H}} + 4L_{\mathcal{H}}^2 C_2)\lambda + (L_{\mathcal{H}} + 4L_{\mathcal{H}}^2 C_2) + \frac{L_{\mathcal{H}}^2 C_2}{\lambda}.
\end{aligned}
$$

Once λ and h are chosen, α_0, τ_0, and h are determined. Finally, since $\frac{(1-\gamma)\tau_0^2 \rho_{\min}^3 c^2}{152 \log(2)|\mathcal{S}| L_0^2} \leqslant 1$, we just need to select $\beta_0 \in [\frac{152 \log(2)|\mathcal{S}| L_0^2 \alpha_0}{(1-\gamma)\tau_0^2 \rho_{\min}^3 c^2}, \frac{1}{L_0}]$. Recall that $L_0 = L_{\mathcal{H}}(2\tau_0 + 1)$, it can be easily seen that the lower bound $\frac{152 \log(2)|\mathcal{S}| L_0^2 \alpha_0}{(1-\gamma)\tau_0^2 \rho_{\min}^3 c^2} = \mathcal{O}(\frac{1}{\lambda^3 h^{1/3}})$, which is much smaller than the upper bound $\frac{1}{L_0} = \mathcal{O}(\frac{1}{\tau_0}) = \mathcal{O}(\frac{1}{\lambda h^{1/3}})$ since λ was large enough.

□

C.2 Proof of Lemmas

C.2.1 Proof of Lemma 4.1

For a given ϕ, let $\hat{\pi} \in \pi_\tau(\phi)$ (which is a possibly non-unique maximizer).

According to [13][Lemma 26],

$$J_\tau(\hat{\pi}, \phi) - J_\tau(\pi, \phi) = \frac{\tau}{1-\gamma} \sum_{s \in \mathcal{S}} d_\rho^{\pi,\phi}(s) D_{KL}(\pi(\cdot \mid s) || \hat{\pi}(\cdot \mid s)).$$

The Pinsker's inequality states that for any two probability distributions p_1 and p_2

$$D_{KL}(p_1||p_2) \geqslant \frac{1}{2\log(2)} \|p_1 - p_2\|_1^2.$$

Using this inequality,

$$\begin{aligned}
J_\tau(\hat{\pi}, \phi) - J_\tau(\pi, \phi) &= \frac{\tau}{1-\gamma} \sum_{s \in \mathcal{S}} d_\rho^{\pi,\phi}(s) D_{KL}(\pi(\cdot \mid s) || \hat{\pi}(\cdot \mid s)) \\
&\geqslant \frac{\tau}{2\log(2)(1-\gamma)} \sum_{s \in \mathcal{S}} d_\rho^{\pi,\phi}(s) \|\pi(\cdot \mid s) - \hat{\pi}(\cdot \mid s)\|_1^2 \\
&\geqslant \frac{\tau}{2\log(2)(1-\gamma)} \sum_{s \in \mathcal{S}} (1-\gamma)\rho(s) \|\pi(\cdot \mid s) - \hat{\pi}(\cdot \mid s)\|_1^2 \\
&\geqslant \frac{\tau \min_{s \in \mathcal{S}} \rho(s)}{2\log(2)} \sum_{s \in \mathcal{S}} \|\pi(\cdot \mid s) - \hat{\pi}(\cdot \mid s)\|_1^2 \\
&\geqslant \frac{\tau \min_{s \in \mathcal{S}} \rho(s)}{2\log(2)} \|\pi - \hat{\pi}\|^2,
\end{aligned}$$

where the second inequality follows from the fact that $d_\rho^{\pi,\hat{\phi}}(s) \geqslant (1-\gamma)\rho(s)$ entry-wise. This inequality means that $\hat{\pi} \in \pi_\tau(\phi)$ has to be unique, as no other policy can achieve the same value function.

The same argument can be used to show Equation 4.4.

□

C.2.2 Proof of Lemma 4.2

Let (π_1, ϕ_1), (π_2, ϕ_2) be optimal solution pairs to the maximin and minimax problem, respectively,

$$(\pi_1, \phi_1) \in \operatorname*{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \operatorname*{argmin}_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} J_\tau(\pi, \phi) \quad \text{and} \quad (\pi_2, \phi_2) \in \operatorname*{argmin}_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} \operatorname*{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} J_\tau(\pi, \phi). \tag{C.38}$$

Since the policy simplex is a compact set, (π_1, ϕ_1) and (π_2, ϕ_2) exist and are well-defined. The following minimax inequality always holds

$$J_\tau(\pi_1, \phi_1) = \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \min_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} J_\tau(\pi, \phi) \leqslant \min_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} J_\tau(\pi, \phi) = J_\tau(\pi_2, \phi_2). \tag{C.39}$$

We first want to show that $\pi_1 = \pi_\tau(\phi_1)$ and $\phi_1 = \phi_\tau(\pi_1)$. Since

$$J_\tau(\pi_1, \phi_1) = \max_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \min_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} J_\tau(\pi, \phi) = \min_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} J_\tau(\pi_1, \phi) = J_\tau(\pi_1, \phi_\tau(\pi_1)),$$

we have $\phi_1 \in \phi_\tau(\pi_1)$, and Lemma 4.1 further implies $\phi_1 = \phi_\tau(\pi_1)$ is unique. In addition, we know that π_1 is the optimizer of g_τ defined in Equation 4.2. Let θ_1 be an softmax parameter for π_1 (e.g. $\theta_1(s, a) = \log \pi(a \mid s)$ for all s, a). Since π_1 is an optimizer of g_τ in policy space, θ_1 must also be an (not necessarily unique) optimizer of $\tilde{g}_\tau(\theta) = \min_\phi J_\tau(\pi_\theta, \phi)$ in the parameter space. Therefore, we have $\forall \theta \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$

$$0 \geqslant \langle \nabla_\theta g_\tau(\pi_{\theta_1}), \theta - \theta_1 \rangle = \langle \nabla_\theta J_\tau(\pi_{\theta_1}, \phi_1), \theta - \theta_1 \rangle, \tag{C.40}$$

where the first equality follows from Danskin's Theorem in Equation C.2. Since θ is not constrained, Equation C.40 means that

$$\nabla_\theta J_\tau(\pi_{\theta_1}, \phi_1) = 0,$$

implying that θ_1 is a stationary point of

$$\max_{\theta} J_\tau(\pi_\theta, \phi_1).$$

By Lemma 4.4, every stationary point is also globally optimal. Therefore, we have $\pi_1 = \pi_{\theta_1} = \pi_\tau(\phi_1)$.

A consequence of $\pi_1 = \pi_\tau(\phi_1)$ and $\phi_1 = \phi_\tau(\pi_1)$ is that (π_1, ϕ_1) is the unique optimal solution pair to the maximin problem, i.e. there does not exist $(\widehat{\pi}_1, \widehat{\phi}_1) \neq (\pi_1, \phi_1)$ such that $(\widehat{\pi}_1, \widehat{\phi}_1) \in \operatorname{argmax}_{\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}} \operatorname{argmin}_{\phi \in \Delta_{\mathcal{B}}^{\mathcal{S}}} J_\tau(\pi, \phi)$. To see this, let us suppose that such a pair $(\widehat{\pi}_1, \widehat{\phi}_1)$ does exist. Then, the only possibility is $\widehat{\pi}_1 \neq \pi_1$ and $\widehat{\phi}_1 \neq \phi_1$ by Lemma 4.1. Since $\widehat{\pi}_1 \neq \pi_\tau(\phi_1)$ and $\phi_1 \neq \phi_\tau(\widehat{\pi}_1)$, we have

$$J_\tau(\widehat{\pi}_1, \phi_1) < J_\tau(\pi_1, \phi_1) = J_\tau(\widehat{\pi}_1, \widehat{\phi}_1) < J_\tau(\widehat{\pi}_1, \phi_1),$$

which creates a contradiction.

Similarly, it can be shown that

$$\pi_2 = \phi_\tau(\phi_2), \quad \text{and} \quad \phi_2 = \phi_\tau(\pi_2),$$

and that (π_2, ϕ_2) is the unique optimal solution pair to the minimax problem.

We now aim prove that $(\pi_1, \phi_1) = (\pi_2, \phi_2)$, i.e. the minimax and maximin problem have the same solution. Suppose $(\pi_1, \phi_1) \neq (\pi_2, \phi_2)$, which means that $\pi_1 \neq \pi_2$ and $\phi_1 \neq \phi_2$ have to hold due to Lemma 4.1. Since $\pi_2 \neq \pi_\tau(\phi_1)$ and $\phi_1 \neq \phi_\tau(\pi_2)$, we have from Equation C.39

$$J_\tau(\pi_2, \phi_1) < J_\tau(\pi_1, \phi_1) \leqslant J_\tau(\pi_2, \phi_2) < J_\tau(\pi_2, \phi_1).$$

This is again a contradiction. Therefore, $(\pi_1, \phi_1) = (\pi_2, \phi_2)$ has to be true. Then, Equa-

tion C.39 leads to

$$\max_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}}\min_{\phi\in\Delta_{\mathcal{B}}^{\mathcal{S}}} J_\tau(\pi,\phi) = \max_{\pi\in\Delta_{\mathcal{A}}^{\mathcal{S}}}\min_{\phi\in\Delta_{\mathcal{B}}^{\mathcal{S}}} J_\tau(\pi,\phi).$$

We also know that the Nash equilibrium has to be unique in this case, as the maximin and minimax problems both have a unique solution pair that agrees with each other.

□

C.2.3 Proof of Lemma 4.3

By the definition of the value function,

$$\begin{aligned}
&J_\tau(\pi,\phi) - J_{\tau'}(\pi,\phi)\\
&= \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k\Big(r\left(s_k,a_k,b_k\right) - \tau\log\pi(a_k\mid s_k) + \tau\log\phi(b_k\mid s_k)\Big)\mid s_0\sim\rho\right]\\
&\quad - \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k\Big(r\left(s_k,a_k,b_k\right) - \tau'\log\pi(a_k\mid s_k) + \tau'\log\phi(b_k\mid s_k)\Big)\mid s_0\sim\rho\right]\\
&= \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k\Big((\tau-\tau')\log\pi(a_k\mid s_k) + (\tau-\tau')\log\phi(b_k\mid s_k)\Big)\mid s_0\sim\rho\right]\\
&= \frac{\tau-\tau'}{1-\gamma}\mathbb{E}_{s'\sim d_\rho^{\pi,\phi},a\sim\pi(\cdot|s'),b\sim\phi(\cdot|s')}\left[-\log\pi(a\mid s') + \log\phi(b\mid s')\right]\\
&= \frac{\tau-\tau'}{1-\gamma}\mathbb{E}_{s'\sim d_\rho^{\pi,\phi}}[H(\pi(\cdot\mid s')) - H(\phi(\cdot\mid s'))],
\end{aligned}$$

where H denotes the entropy and is defined in Equation C.1.

We have the following upper and lower bound on the entropy

$$0\leqslant H(\pi(\cdot\mid s'))\leqslant\log|\mathcal{A}|,\quad 0\leqslant H(\phi(\cdot\mid s'))\leqslant\log|\mathcal{B}|.$$

Therefore, if $\tau\geqslant\tau'\geqslant 0$,

$$-\frac{\tau-\tau'}{1-\gamma}\log|\mathcal{B}|\leqslant J_\tau(\pi,\phi) - J_{\tau'}(\pi,\phi)\leqslant\frac{\tau-\tau'}{1-\gamma}\log|\mathcal{A}|.$$

For any $\tau \geqslant \tau' \geqslant 0$,

$$
\begin{aligned}
&J_\tau(\pi^\star_\tau, \phi^\star_\tau) - J_{\tau'}(\pi^\star_{\tau'}, \phi^\star_{\tau'}) \\
&= \max_\pi \min_\phi J_\tau(\pi, \phi) - \min_\phi J_{\tau'}(\pi^\star_{\tau'}, \phi) \\
&\geqslant \min_\phi J_\tau(\pi^\star_{\tau'}, \phi) - \min_\phi J_{\tau'}(\pi^\star_{\tau'}, \phi) \\
&= \min_\phi \Big(J_{\tau'}(\pi^\star_{\tau'}, \phi) + (\tau - \tau')\mathcal{H}_\pi(\rho, \pi^\star_{\tau'}, \phi) - (\tau - \tau')\mathcal{H}_\phi(\rho, \pi^\star_{\tau'}, \phi) \Big) - \min_\phi J_{\tau'}(\pi^\star_{\tau'}, \phi) \\
&\geqslant \min_\phi J_{\tau'}(\pi^\star_{\tau'}, \phi) + (\tau - \tau') \min_\phi \mathcal{H}_\pi(\rho, \pi^\star_{\tau'}, \phi) + (\tau - \tau') \min_\phi -\mathcal{H}_\phi(\rho, \pi^\star_{\tau'}, \phi) - \min_\phi J_{\tau'}(\pi^\star_{\tau'}, \phi) \\
&= (\tau - \tau') \left(\min_\phi \mathcal{H}_\pi(\rho, \pi^\star_{\tau'}, \phi) - \max_\phi \mathcal{H}_\phi(\rho, \pi^\star_{\tau'}, \phi) \right) \\
&\geqslant (\tau - \tau')(0 - \log |\mathcal{B}|) \\
&= -(\tau - \tau') \log |\mathcal{B}|,
\end{aligned}
$$

where the second inequality comes from the fact that $\min_x f_1(x) + f_2(x) \geqslant \min_x f_1(x) + \min_x f_2(x)$ for any functions f_1, f_2 of the same domain.

It can be shown by a similar argument

$$
J_\tau(\pi^\star_\tau, \phi^\star_\tau) - J_{\tau'}(\pi^\star_{\tau'}, \phi^\star_{\tau'}) \leqslant (\tau - \tau') \log |\mathcal{A}|.
$$

In addition, for any $\tau \geqslant \tau' \geqslant 0$ and any policy π,

$$
\begin{aligned}
&J_\tau(\pi, \phi_\tau(\pi)) - J_{\tau'}(\pi, \phi_0(\pi)) \\
&= \min_\phi J_\tau(\pi, \phi) - \min_\phi J_{\tau'}(\pi, \phi) \\
&= \min_\phi \left(J_{\tau'}(\pi, \phi) + (\tau - \tau')\mathcal{H}_\pi(\rho, \pi, \phi) - (\tau - \tau')\mathcal{H}_\phi(\rho, \pi, \phi) \right) - \min_\phi J_{\tau'}(\pi, \phi) \\
&\leqslant \left(\min_\phi J_{\tau'}(\pi, \phi) + (\tau - \tau') \max_\phi \mathcal{H}_\pi(\rho, \pi, \phi) + (\tau - \tau') \max_\phi (-\mathcal{H}_\phi(\rho, \pi, \phi)) \right) - \min_\phi J_{\tau'}(\pi, \phi) \\
&= (\tau - \tau') \left(\max_\phi \mathcal{H}_\pi(\rho, \pi, \phi) - \min_\phi \mathcal{H}_\phi(\rho, \pi, \phi) \right) \\
&\leqslant (\tau - \tau') \log |\mathcal{A}|.
\end{aligned}
$$

It can be shown by a similar argument

$$J_\tau(\pi, \phi_\tau(\pi)) - J_{\tau'}(\pi, \phi_0(\pi)) \geqslant -(\tau - \tau') \log |\mathcal{B}|.$$

□

C.2.4 Proof of Lemma 4.4

Adapting [13][Lemma 15], we have for any $\theta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\psi \in \mathbb{R}^{\mathcal{S}\times\mathcal{B}}$

$$\|\nabla_\theta J_\tau(\pi_\theta, \phi_\psi)\|^2 \geqslant \frac{2\tau\rho_{\min}}{|\mathcal{S}|}\left(\min_{s,a}\pi_\theta(a \mid s)\right)^2 \left\|\frac{d_\rho^{\pi_\tau(\phi_\psi),\phi_\psi}}{d_\rho^{\pi_\theta,\phi_\psi}}\right\|_\infty^{-1} \left(J_\tau(\pi_\tau(\phi_\psi), \phi_\psi) - J_\tau(\pi_\theta, \phi_\psi)\right),$$

$$\|\nabla_\psi J_\tau(\pi_\theta, \phi_\psi)\|^2 \geqslant \frac{2\tau\rho_{\min}}{|\mathcal{S}|}\left(\min_{s,b}\phi_\psi(b \mid s)\right)^2 \left\|\frac{d_\rho^{\pi_\theta,\phi_\tau(\pi_\theta)}}{d_\rho^{\pi_\theta,\phi_\psi}}\right\|_\infty^{-1} \left(J_\tau(\pi_\theta, \phi_\psi) - J_\tau(\pi_\theta, \phi_\tau(\pi_\theta))\right).$$

Then, the first inequality follows from $d_\rho^{\pi_\tau(\phi_\psi),\phi_\psi}(s) \leqslant 1$ and $d_\rho^{\pi_\theta,\phi_\psi}(s) \geqslant (1-\gamma)\rho(s) \geqslant (1-\gamma)\rho_{\min}$ for all $s \in \mathcal{S}$, and the second inequality from $d_\rho^{\pi_\theta,\phi_\tau(\pi_\theta)} \leqslant 1$ and $d_\rho^{\pi_\theta,\phi_\psi} \geqslant (1-\gamma)\rho_{\min}$ for all $s \in \mathcal{S}$.

□

C.2.5 Proof of Lemma C.1

Lemma 7 and 14 of [13] establish the smoothness condition of the value function and the regularization entropy with respect to one player's policy, i.e.

$$\|\nabla_\theta J(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta J(\pi_{\theta_2}, \phi_{\psi_1})\| \leqslant L_V\|\theta_1 - \theta_2\|,$$
$$\|\nabla_\psi J(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\psi J(\pi_{\theta_1}, \phi_{\psi_2})\| \leqslant L_V\|\psi_1 - \psi_2\|.$$

Therefore, we only need to show

$$\|\nabla_\theta J(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta J(\pi_{\theta_1}, \phi_{\psi_2})\| \leqslant L_V\|\psi_1 - \psi_2\|,$$

$$\|\nabla_\psi J(\pi_{\theta_1}, \phi_\psi) - \nabla_\psi J(\pi_{\theta_2}, \phi_\psi)\| \leqslant L_V \|\theta_1 - \theta_2\|.$$

Given a fixed θ and ψ, with arbitrary vectors u and v such that $\|u\|_2 = \|v\|_2 = 1$, we define the shorthand notation

$$\pi_{\alpha,u} = \pi_{\theta+\alpha u}, \quad \phi_{\beta,v} = \pi_{\psi+\beta v}.$$

According to [35][Lemma B.5],

$$\sum_a \left| \frac{d\pi_{\alpha,u}(a \mid s)}{d\alpha} \right| \leqslant 2, \quad \sum_b \left| \frac{d\phi_{\beta,v}(b \mid s)}{d\beta} \right| \leqslant 2,$$
$$\sum_{a,b} \left| \frac{d\pi_\alpha(a \mid s)}{d\alpha} \frac{d\phi_{\beta,v}(b \mid s)}{d\beta} \right| \leqslant \left(\sum_a \left| \frac{d\pi_\alpha(a \mid s)}{d\alpha} \right| \right) \left(\sum_b \left| \frac{d\phi_\beta(b \mid s)}{d\beta} \right| \right) \leqslant 4.$$

Let $P(\alpha, \beta, u, v) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}||\mathcal{B}| \times |\mathcal{S}||\mathcal{A}||\mathcal{B}|}$ denote the state-action transition matrix induced by the policy pair $(\pi_{\alpha,u}, \phi_{\beta,v})$

$$P(\alpha, \beta, u, v)_{(s,a,b)\to(s',a',b')} = \mathcal{P}(s' \mid s, a, b)\pi_{\alpha,u}(a' \mid s')\phi_{\beta,v}(b' \mid s').$$

Differentiating with respect to α and β,

$$\left[\frac{d^2 P(\alpha, \beta, u, v)}{d\alpha d\beta} \right]_{(s,a,b)\to(s',a',b')} = \frac{d\pi_{\alpha,u}(a' \mid s')}{d\alpha} \frac{d\phi_{\beta,v}(b' \mid s')}{d\beta} \mathcal{P}(s' \mid s, a, b),$$

which implies for any vector x

$$\left[\frac{d^2 P(\alpha, \beta, u, v)}{d\alpha d\beta} x \right]_{s,a,b} = \sum_{s',a',b'} \frac{d\pi_\alpha(a' \mid s')}{d\alpha} \frac{d\phi_{\beta,v}(b' \mid s')}{d\beta} \mathcal{P}(s' \mid s, a, b) x_{s',a',b'}.$$

The ℓ_∞ norm of this quantity can be upper bounded

$$\max_{\|u\|_2 = \|v\|_2 = 1} \left\| \frac{d^2 P(\alpha, \beta, u, v)}{d\alpha d\beta} x \right\|_\infty$$

$$
\begin{aligned}
&= \max_{s,a,b} \max_{\|u\|_2=\|v\|_2=1} \left| \left[\frac{d^2 P(\alpha,\beta,u,v)}{d\alpha d\beta} x \right]_{s,a,b} \right| \\
&= \max_{s,a,b} \max_{\|u\|_2=\|v\|_2=1} \left| \sum_{s',a',b'} \frac{d\pi_\alpha(a' \mid s')}{d\alpha} \frac{d\phi_{\beta,v}(b' \mid s')}{d\beta} \mathcal{P}(s' \mid s,a,b) x_{s',a',b'} \right| \\
&\leqslant \max_{s,a,b} \sum_{s'} \mathcal{P}(s' \mid s,a,b) \|x\|_\infty \max_{\|u\|_2=\|v\|_2=1} \sum_{a',b'} \left| \frac{d\pi_\alpha(a' \mid s')}{d\alpha} \frac{d\phi_{\beta,v}(b' \mid s')}{d\beta} \right| \\
&\leqslant 4\|x\|_\infty. \qquad \text{(C.41)}
\end{aligned}
$$

Using an identical argument, we can show that

$$
\max_{\|u\|_2=\|v\|_2=1} \left\| \frac{dP(\alpha,\beta,u,v)}{d\alpha} x \right\|_\infty \leqslant \sum_a \left| \frac{d\pi_{\alpha,u}(a \mid s)}{d\alpha} \right| \|x\|_\infty \leqslant 2\|x\|_\infty, \qquad \text{(C.42)}
$$

$$
\max_{\|u\|_2=\|v\|_2=1} \left\| \frac{dP(\alpha,\beta,u,v)}{d\beta} x \right\|_\infty \leqslant \sum_b \left| \frac{d\pi_{\beta,v}(b \mid s)}{d\beta} \right| \|x\|_\infty \leqslant 2\|x\|_\infty. \qquad \text{(C.43)}
$$

With $M(\alpha,\beta,u,v) = (I - \gamma P(\alpha,\beta,u,v))^{-1}$ and $r = [r(s_0,a_0,b_0), \cdots, r(s_{|\mathcal{S}|}, a_{|\mathcal{A}|}, b_{|\mathcal{B}|})]$,

$$
Q^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b) = e_{s,a,b}^\top M(\alpha,\beta,u,v) r.
$$

Taking the derivatives,

$$
\begin{aligned}
\frac{dQ^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)}{d\alpha} &= \gamma e_{s,a,b}^\top M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\alpha} M(\alpha,\beta,u,v) r, \\
\frac{dQ^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)}{d\beta} &= \gamma e_{s,a,b}^\top M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\beta} M(\alpha,\beta,u,v) r.
\end{aligned}
$$

Taking the second-order derivative,

$$
\begin{aligned}
&\frac{d^2 Q^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)}{d\alpha d\beta} \\
&= \gamma^2 e_{s,a,b}^\top M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\alpha} M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\beta} M(\alpha,\beta,u,v) r \\
&\quad + \gamma^2 e_{s,a,b}^\top M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\beta} M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\alpha} M(\alpha,\beta,u,v) r
\end{aligned}
$$

$$+ \gamma e_{s,a,b}^{\top} M(\alpha,\beta,u,v) \frac{d^2 P(\alpha,\beta,u,v)}{d\alpha d\beta} M(\alpha,\beta,u,v) r$$

Using $M(\alpha,\beta,u,v)\mathbf{1} = (I - \gamma P(\alpha,\beta,u,v))^{-1}\mathbf{1} = \frac{1}{1-\gamma}\mathbf{1}$ and inequalities Equation C.41 and Equation C.43, we have

$$\max_{\|u\|_2=\|v\|_2=1} \left| \frac{dQ^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)}{d\alpha} \right| \leqslant \|\gamma M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\alpha} M(\alpha,\beta,u,v) r\|_\infty \leqslant \frac{2\gamma}{(1-\gamma)^2},$$
$$\max_{\|u\|_2=\|v\|_2=1} \left| \frac{dQ^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)}{d\beta} \right| \leqslant \|\gamma M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\beta} M(\alpha,\beta,u,v) r\|_\infty \leqslant \frac{2\gamma}{(1-\gamma)^2},$$

and

$$\begin{aligned}
&\max_{\|u\|_2=\|v\|_2=1} \left| \frac{d^2 Q^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)}{d\alpha d\beta} \right| \\
&\leqslant \|\gamma^2 M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\alpha} M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\beta} M(\alpha,\beta,u,v) r\|_\infty \\
&\quad + \|\gamma^2 M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\beta} M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\alpha} M(\alpha,\beta,u,v) r\|_\infty \\
&\quad + \|\gamma M(\alpha,\beta,u,v) \frac{d^2 P(\alpha,\beta,u,v)}{d\alpha d\beta} M(\alpha,\beta,u,v) r\|_\infty \\
&\leqslant \frac{2\gamma^2}{(1-\gamma)^3} + \frac{4\gamma}{(1-\gamma)^2}.
\end{aligned}$$

Since $V^{\pi_{\alpha,u},\phi_{\beta,v}}(s) = \sum_{a,b} \pi_{\alpha,u}(a \mid s) \phi_{\beta,v}(b \mid s) Q^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)$,

$$\begin{aligned}
\frac{d^2 V^{\pi_{\alpha,u},\phi_{\beta,v}}(s)}{d\alpha d\beta} = &\sum_{a,b} \frac{d\pi_{\alpha,u}(a \mid s)}{d\alpha} \frac{d\phi_{\beta,v}(b \mid s)}{d\beta} Q^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b) \\
&+ \sum_{a,b} \pi_{\alpha,u}(a \mid s) \phi_{\beta,v}(b \mid s) \frac{d^2 Q^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)}{d\alpha d\beta} \\
&+ \sum_{a,b} \frac{d\pi_{\alpha,u}(a \mid s)}{d\alpha} \phi_{\beta,v}(b \mid s) \frac{dQ^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)}{d\beta} \\
&+ \sum_{a,b} \pi_{\alpha,u}(a \mid s) \frac{d\phi_{\beta,v}(b \mid s)}{d\beta} \frac{dQ^{\pi_{\alpha,u},\phi_{\beta,v}}(s,a,b)}{d\alpha}.
\end{aligned}$$

Therefore,

$$\max_{\|u\|_2=\|v\|_2=1} \left| \frac{dV^{\pi_{\alpha,u},\phi_{\beta,v}}(s)}{d\alpha d\beta} \right| \leqslant \frac{4}{1-\gamma} + \left(\frac{2\gamma^2}{(1-\gamma)^3} + \frac{4\gamma}{(1-\gamma)^2} \right) + 2\frac{4\gamma}{(1-\gamma)^2} \leqslant \frac{8}{(1-\gamma)^3},$$

which implies

$$\|\nabla_\theta J(\pi_\theta, \phi_{\psi_1}) - \nabla_\theta J(\pi_\theta, \phi_{\psi_2})\| \leqslant \frac{8}{(1-\gamma)^3}\|\psi_1 - \psi_2\|.$$

Similarly, it follows by the same argument that

$$\|\nabla_\psi J(\pi_{\theta_1}, \phi_\psi) - \nabla_\psi J(\pi_{\theta_2}, \phi_\psi)\| \leqslant \frac{8}{(1-\gamma)^3}\|\theta_1 - \theta_2\|.$$

[35][Lemma B.5] implies

$$\|J(\pi_{\theta_1}, \phi_{\psi_1}) - J(\pi_{\theta_2}, \phi_{\psi_2})\| \leqslant \frac{2}{(1-\gamma)^2}(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|), \tag{C.44}$$

and we simply use $\frac{2}{(1-\gamma)^2} \leqslant L_V$.

□

C.2.6 Proof of Lemma C.2

We will prove the first two inequalities on the Lipschitz gradient of $\mathcal{H}_\pi$. The next two inequalities are completely symmetric and can be derived using an identical argument.

[13][Lemma 14] implies

$$\|\nabla_\theta \mathcal{H}_\pi(s, \pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta \mathcal{H}_\pi(s, \pi_{\theta_2}, \phi_{\psi_1})\| \leqslant L_{\mathcal{H}}\|\theta_1 - \theta_2\|,$$

so we just need to show

$$\begin{aligned}
\|\nabla_\theta \mathcal{H}_\pi(s, \pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\theta \mathcal{H}_\pi(s, \pi_{\theta_1}, \phi_{\psi_2})\| &\leqslant L_{\mathcal{H}} \|\psi_1 - \psi_2\|, \\
\|\nabla_\psi \mathcal{H}_\pi(s, \pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\psi \mathcal{H}_\pi(s, \pi_{\theta_2}, \phi_{\psi_1})\| &\leqslant L_{\mathcal{H}} \|\theta_1 - \theta_2\|, \\
\|\nabla_\psi \mathcal{H}_\pi(s, \pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\psi \mathcal{H}_\pi(s, \pi_{\theta_1}, \phi_{\psi_2})\| &\leqslant L_{\mathcal{H}} \|\psi_1 - \psi_2\|.
\end{aligned} \tag{C.45}$$

Given a fixed θ and ψ, with arbitrary vectors u and v such that $\|u\|_2 = \|v\|_2 = 1$, we define the shorthand notation

$$\pi_{\alpha,u} = \pi_{\theta+\alpha u}, \quad \phi_{\beta,v} = \pi_{\psi+\beta v}.$$

Note that to show Equation C.45, it suffices to show for any u, v

$$\left| \frac{d^2 \mathcal{H}_\pi(s, \pi_{\alpha,u}, \phi_{\beta,v})}{d\alpha d\beta} \right| \leqslant L_{\mathcal{H}}, \quad \left| \frac{d^2 \mathcal{H}_\pi(s, \pi_{\alpha,u}, \phi_{\beta,v})}{d\beta^2} \right| \leqslant L_{\mathcal{H}}.$$

We define the state transition matrix $P \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ such that

$$P(\alpha, \beta, u, v)_{s \to s'} = \sum_{a,b} \mathcal{P}(s' \mid s, a, b) \pi_{\alpha,u}(a \mid s) \phi_{\beta,v}(b \mid s).$$

Let $M(\alpha, \beta, u, v) = (I - \gamma P(\alpha, \beta, u, v))^{-1}$. Then, we can re-write $\mathcal{H}_\pi(s, \pi, \phi)$ in the matrix form

$$\mathcal{H}_\pi(s, \pi, \phi) = e_s^\top M(\alpha, \beta, u, v) h_{\alpha,u},$$

where $h_{\alpha,u} = [h_{\alpha,u}(s_0), \cdots, h_{\alpha,u}(s_{|\mathcal{S}|})] \in \mathbb{R}^{|\mathcal{S}|}$ is a vector with

$$h_{\alpha,u}(s) = -\sum_a \pi_{\alpha,u}(a \mid s) \log \pi_{\alpha,u}(a \mid s).$$

According to [13][Lemma 14],

$$\left\|\frac{dh_{\alpha,u}}{d\alpha}\right\|_\infty \leqslant 2\log|\mathcal{A}|\|u\|_2 = 2\log|\mathcal{A}|.$$

Taking the derivatives of $\mathcal{H}_\pi(s,\pi,\phi)$,

$$\begin{aligned}&\frac{d\mathcal{H}_\pi(s,\pi_{\alpha,u},\phi_{\beta,v})}{d\alpha}\\&= \gamma e_s^\top M(\alpha,\beta,u,v)\frac{dP(\alpha,\beta,u,v)}{d\alpha}M(\alpha,\beta,u,v)h_{\alpha,u} + e_s^\top M(\alpha,\beta,u,v)\frac{dh_{\alpha,u}}{d\alpha},\end{aligned}$$

and taking second order derivative

$$\begin{aligned}&\frac{d^2\mathcal{H}_\pi(s,\pi_{\alpha,u},\phi_{\beta,v})}{d\alpha d\beta}\\&= \gamma^2 e_s^\top M(\alpha,\beta,u,v)\frac{dP(\alpha,\beta,u,v)}{d\alpha}M(\alpha,\beta,u,v)\frac{dP(\alpha,\beta,u,v)}{d\beta}M(\alpha,\beta,u,v)h_{\alpha,u}\\&\quad+\gamma^2 e_s^\top M(\alpha,\beta,u,v)\frac{dP(\alpha,\beta,u,v)}{d\beta}M(\alpha,\beta,u,v)\frac{dP(\alpha,\beta,u,v)}{d\alpha}M(\alpha,\beta,u,v)h_{\alpha,u}\\&\quad+\gamma e_s^\top M(\alpha,\beta,u,v)\frac{d^2P(\alpha,\beta,u,v)}{d\alpha d\beta}M(\alpha,\beta,u,v)h_{\alpha,u}\\&\quad+\gamma e_s^\top M(\alpha,\beta,u,v)\frac{dP(\alpha,\beta,u,v)}{d\beta}M(\alpha,\beta,u,v)\frac{dh_{\alpha,u}}{d\alpha}.\end{aligned}$$

Using a similar line of argument to [13][Eq. (192)-(195)] and analysis in Lemma C.1 of our work, we can show that for any vector x

$$\left\|\frac{dP(\alpha,\beta,u,v)}{d\alpha}x\right\|_\infty \leqslant 2\|x\|_\infty,\ \left\|\frac{dP(\alpha,\beta,u,v)}{d\beta}\right\|_\infty \leqslant 2\|x\|_\infty,\ \left\|\frac{d^2P(\alpha,\beta,u,v)}{d\alpha d\beta}\right\|_\infty \leqslant 4\|x\|_\infty.$$

From the fact that $\|M(\alpha,\beta,u,v)x\|_\infty \leqslant \frac{1}{1-\gamma}\|x\|_\infty$, we have for any vectors u, v

$$\begin{aligned}&\left|\frac{d^2\mathcal{H}_\pi(s,\pi_{\alpha,u},\phi_{\beta,v})}{d\alpha d\beta}\right|\\&\leqslant \gamma^2\left\|M(\alpha,\beta,u,v)\frac{dP(\alpha,\beta,u,v)}{d\alpha}M(\alpha,\beta,u,v)\frac{dP(\alpha,\beta,u,v)}{d\beta}M(\alpha,\beta,u,v)h_{\alpha,u}\right\|\end{aligned}$$

$$
\begin{aligned}
&+ \gamma^2 \left\| M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\beta} M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\alpha} M(\alpha,\beta,u,v) h_{\alpha,u} \right\| \\
&+ \gamma \left\| M(\alpha,\beta,u,v) \frac{d^2P(\alpha,\beta,u,v)}{d\alpha d\beta} M(\alpha,\beta,u,v) h_{\alpha,u} \right\| \\
&+ \gamma \left\| M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\beta} M(\alpha,\beta,u,v) \frac{dh_{\alpha,u}}{d\alpha} \right\| \\
\leqslant\ & \frac{4\gamma^2 \log|\mathcal{A}|}{(1-\gamma)^3} + \frac{4\gamma^2 \log|\mathcal{A}|}{(1-\gamma)^3} + \frac{4\gamma \log|\mathcal{A}|}{(1-\gamma)^2} + \frac{2\gamma}{(1-\gamma)^2} \cdot 2\log|\mathcal{A}| \\
\leqslant\ & \frac{8\log|\mathcal{A}|}{(1-\gamma)^3}.
\end{aligned}
$$

Now it remains to be shown

$$
\left| \frac{d^2 \mathcal{H}_\pi(s, \pi_{\alpha,u}, \phi_{\beta,v})}{d\beta^2} \right| \leqslant L_{\mathcal{H}}.
$$

From the eye of the second player, $\mathcal{H}_\pi(s, \pi_\theta, \phi_\psi)$ is simply the value function of a regular MDP with itself as the only agent (the first player's policy combines with $\mathcal{P}$) with the reward function $r(s,b) = -\sum_{a\in\mathcal{A}} \pi_\theta(a \mid s) \log \pi_\theta(a \mid s) \in [0, \log|\mathcal{A}|]$. Therefore, by Lemma C.1 which is derived with reward bounded between 0 and 1, we know

$$
\left| \frac{d^2 \mathcal{H}_\pi(s, \pi_{\alpha,u}, \phi_{\beta,v})}{d\beta^2} \right| \leqslant \log|\mathcal{A}| L_V \leqslant L_{\mathcal{H}}.
$$

To show the Lipschitz continuity, we note that

$$
\begin{aligned}
&\left| \frac{d\mathcal{H}_\pi(s, \pi_{\alpha,u}, \phi_{\beta,v})}{d\alpha} \right| \\
=\ & \left| \gamma e_s^\top M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\alpha} M(\alpha,\beta,u,v) h_{\alpha,u} + e_s^\top M(\alpha,\beta,u,v) \frac{dh_{\alpha,u}}{d\alpha} \right| \\
\leqslant\ & \gamma \| M(\alpha,\beta,u,v) \frac{dP(\alpha,\beta,u,v)}{d\alpha} M(\alpha,\beta,u,v) h_{\alpha,u} \| + \| M(\alpha,\beta,u,v) \frac{dh_{\alpha,u}}{d\alpha} \| \\
\leqslant\ & \frac{4\gamma \log|\mathcal{A}|}{(1-\gamma)^2} + \frac{2\log|\mathcal{A}|}{1-\gamma} \leqslant L_{\mathcal{H}}.
\end{aligned}
$$

To show the Lipschitz continuity of $\mathcal{H}_\pi$ with respect to ψ, we use the same argument as above and note that from the eye of the second player, $\mathcal{H}_\pi(s, \pi_\theta, \phi_\psi)$ is simply the value

function of a regular MDP with itself as the only agent (the first player's policy combines with $\mathcal{P}$) with the reward function $r(s,b) = -\sum_{a\in\mathcal{A}} \pi_\theta(a \mid s) \log \pi_\theta(a \mid s) \in [0, \log|\mathcal{A}|]$. Adapting Equation C.44, we have

$$\left|\frac{d\mathcal{H}_\pi(s, \pi_{\alpha,u}, \phi_{\beta,v})}{d\beta}\right| \leqslant \frac{2}{(1-\gamma)^2} \cdot \log|\mathcal{A}| \leqslant L_{\mathcal{H}}.$$

□

C.2.7 Proof of Lemma C.3

We first show that for any $\tilde{k} > 0$, we have $\frac{1}{\tilde{k}^a} - \frac{1}{(\tilde{k}+1)^a} \leqslant \frac{8}{3(\tilde{k}+1)^{a+1}}$.

Since the integer $\tilde{k}$ is positive, it can be lower bound by $\frac{\tilde{k}+1}{2}$.

$$\begin{aligned}
&\frac{1}{\tilde{k}^a} - \frac{1}{(\tilde{k}+1)^a} \\
&= \frac{(\tilde{k}+1)^a - \tilde{k}^a}{\tilde{k}^a(\tilde{k}+1)^a} \leqslant \frac{2((\tilde{k}+1)^a - \tilde{k}^a)}{(\tilde{k}+1)^{2a}} = \frac{2((\tilde{k}+1)^a - \tilde{k}^a)\left((\tilde{k}+1)^{1-a} + \tilde{k}^{1-a}\right)}{(\tilde{k}+1)^{2a}\left((\tilde{k}+1)^{1-a} + \tilde{k}^{1-a}\right)} \\
&\leqslant \frac{2((\tilde{k}+1)^a - \tilde{k}^a)\left((\tilde{k}+1)^{1-a} + \tilde{k}^{1-a}\right)}{(\tilde{k}+1)^{2a}\left((\tilde{k}+1)^{1-a} + \frac{1}{2}(\tilde{k}+1)^{1-a}\right)} = \frac{4((\tilde{k}+1)^a - \tilde{k}^a)\left((\tilde{k}+1)^{1-a} + \tilde{k}^{1-a}\right)}{3(\tilde{k}+1)^{a+1}} \\
&= \frac{4\left((\tilde{k}+1) - \tilde{k}^a(\tilde{k}+1)^{1-a} + \tilde{k}^{1-a}(\tilde{k}+1)^a - \tilde{k}\right)}{3(\tilde{k}+1)^{a+1}} \\
&= \frac{4\left(1 - \tilde{k}^a(\tilde{k}+1)^{1-a} + \tilde{k}^{1-a}(\tilde{k}+1)^a\right)}{3(\tilde{k}+1)^{a+1}} \leqslant \frac{8}{3(\tilde{k}+1)^{a+1}},
\end{aligned}$$

where the last inequality follows from

$$\tilde{k}^{1-a}(\tilde{k}+1)^a - \tilde{k}^a(\tilde{k}+1)^{1-a} \leqslant (\tilde{k}+1)^{1-a}(\tilde{k}+1)^a - \tilde{k}^a\tilde{k}^{1-a} = \tilde{k}+1-\tilde{k} = 1.$$

Choosing $\tilde{k} = k + h$ yields

$$\frac{1}{(k+h)^a} - \frac{1}{(k+1+h)^a} \leqslant \frac{8}{3(k+1+h)^{a+1}} \leqslant \frac{8}{3(k+h)^{a+1}}.$$

□

C.2.8 Proof of Lemma C.4

The property of the min and max function implies that

$$\max_{s,a}(\pi_\tau^\star(a \mid s) - \pi_{\theta_k}(a \mid s)) + \min_{s,a} \pi_{\theta_k}(a \mid s) \geqslant \min_{s,a} \pi_\tau^\star(a \mid s).$$

Since the three terms are all non-negative, the inequality holds after taking the square

$$\begin{aligned}(\min_{s,a} \pi_\tau^\star(a \mid s))^2 &\leqslant (\max_{s,a}(\pi_\tau^\star(a \mid s) - \pi_{\theta_k}(a \mid s)) + \min_{s,a} \pi_{\theta_k}(a \mid s))^2 \\ &\leqslant \frac{4}{3}(\min_{s,a} \pi_{\theta_k}(a \mid s))^2 + 4(\max_{s,a}(\pi_\tau^\star(a \mid s) - \pi_{\theta_k}(a \mid s)))^2.\end{aligned}$$

Re-arranging the terms,

$$\begin{aligned}-\left(\min_{s,a} \pi_{\theta_k}(a \mid s)\right)^2 &\leqslant -\frac{3}{4}\left(\min_{s,a} \pi_\tau^\star(a \mid s)\right)^2 + 3\left(\max_{s,a} \pi_\tau^\star(a \mid s) - \pi_{\phi_k}(a \mid s)\right)^2 \\ &\leqslant -\frac{3}{4}\left(\min_{s,a} \pi_\tau^\star(a \mid s)\right)^2 + 3\|\pi_\tau^\star - \pi_{\phi_k}\|^2\end{aligned}$$

From Lemma 4.1,

$$\begin{aligned}-\left(\min_{s,a} \pi_{\theta_k}(a \mid s)\right)^2 &\leqslant -\frac{3}{4}\left(\min_{s,a} \pi_\tau^\star(a \mid s)\right)^2 + 3\|\pi_\tau^\star - \pi_{\phi_k}\|^2 \\ &\leqslant -\frac{3}{4}\left(\min_{s,a} \pi_\tau^\star(a \mid s)\right)^2 + \frac{6\log(2)}{\tau\rho_{\min}}(J_\tau(\pi_\tau^\star, \phi_\tau^\star) - J_\tau(\pi_{\theta_k}, \phi_\tau^\star)) \\ &\leqslant -\frac{3}{4}\left(\min_{s,a} \pi_\tau^\star(a \mid s)\right)^2 + \frac{6\log(2)}{\tau\rho_{\min}}(J_\tau(\pi_\tau^\star, \phi_\tau^\star) - J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})))\end{aligned}$$

$$= -\frac{3}{4}\left(\min_{s,a}\pi_\tau^\star(a \mid s)\right)^2 + \frac{6\log(2)}{\tau\rho_{\min}}\delta_k^\pi \tag{C.46}$$

Since $3\delta_k^\pi + \delta_k^\phi \leqslant (1 - \frac{\alpha(1-\gamma)\tau\rho_{\min}^2 c^2}{32|\mathcal{S}|})^k(3\delta_0^\pi + \delta_0^\phi) \leqslant 3\delta_0^\pi + \delta_0^\phi \leqslant \frac{\rho_{\min}c^2}{64\log(2)}$, we have $\delta_k^\pi \leqslant \frac{\rho_{\min}c^2}{64\log(2)}$. Then, Equation C.46 implies

$$-\left(\min_{s,a}\pi_{\theta_k}(a \mid s)\right)^2 \leqslant -\frac{3}{4}\left(\min_{s,a}\pi_\tau^\star(a \mid s)\right)^2 + \frac{6\log(2)}{\tau\rho_{\min}}\delta_k^\pi \leqslant -\frac{3c^2}{4} + \frac{3c^2}{32} \leqslant -\frac{3c^2}{8}.$$

Similarly, the property of the min and max function implies that

$$\max_{s,b}(\phi_\tau^\star(b \mid s) - \phi_{\psi_k}(b \mid s)) + \min_{s,b}\phi_{\psi_k}(b \mid s) \geqslant \min_{s,b}\phi_\tau^\star(b \mid s).$$

Again, all three terms are non-negative, which means that the inequality is preserved after taking the square

$$\begin{aligned}(\min_{s,b}\phi_\tau^\star(b \mid s))^2 &\leqslant (\min_{s,b}\phi_{\psi_k}(b \mid s) + \max_{s,b}(\phi_\tau^\star(b \mid s) - \phi_{\psi_k}(b \mid s)))^2 \\ &\leqslant \frac{4}{3}(\min_{s,b}\phi_{\psi_k}(b \mid s))^2 + 4(\max_{s,b}(\phi_\tau^\star(b \mid s) - \phi_{\psi_k}(b \mid s)))^2,\end{aligned}$$

which leads to

$$\begin{aligned}-(\min_{s,b}\phi_{\psi_k}(b \mid s))^2 &\leqslant -\frac{3}{4}(\min_{s,b}\phi_\tau^\star(b \mid s))^2 + 3(\max_{s,b}(\phi_\tau^\star(b \mid s) - \phi_{\psi_k}(b \mid s)))^2 \\ &\leqslant -\frac{3}{4}(\min_{s,b}\phi_\tau^\star(b \mid s))^2 + 3\|\phi_\tau^\star - \phi_{\psi_k}\|^2 \\ &\leqslant -\frac{3}{4}(\min_{s,b}\phi_\tau^\star(b \mid s))^2 + 6\|\phi_\tau(\pi_{\theta_k}) - \phi_{\psi_k}\|^2 + 6\|\phi_\tau^\star - \phi_\tau(\pi_{\theta_k})\|^2.\end{aligned} \tag{C.47}$$

From Lemma 4.1,

$$\|\phi_\tau(\pi_{\theta_k}) - \phi_{\psi_k}\|^2 \leqslant \frac{2\log(2)}{\tau\rho_{\min}}\left(J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) - J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k}))\right) = \frac{2\log(2)}{\tau\rho_{\min}}\delta_k^\phi, \tag{C.48}$$

and

$$\begin{aligned}\|\phi_\tau^\star - \phi_\tau(\pi_{\theta_k})\|^2 &\leqslant \frac{2\log(2)}{\tau\rho_{\min}}\left(J_\tau(\pi_{\theta_k},\phi_\tau^\star) - J_\tau(\pi_{\theta_k},\phi_\tau(\pi_{\theta_k}))\right)\\ &\leqslant \frac{2\log(2)}{\tau\rho_{\min}}\left(J_\tau(\pi_\tau^\star,\phi_\tau^\star) - J_\tau(\pi_{\theta_k},\phi_\tau(\pi_{\theta_k}))\right)\\ &= \frac{2\log(2)}{\tau\rho_{\min}}\delta_k^\pi, \end{aligned} \tag{C.49}$$

Using Equation C.48 and Equation C.49 in Equation C.47,

$$\begin{aligned}-(\min_{s,b}\phi_{\psi_k}(b\mid s))^2 &\leqslant -\frac{3}{4}(\min_{s,b}\phi_\tau^\star(b\mid s))^2 + 6\|\phi_\tau(\pi_{\theta_k}) - \phi_{\psi_k}\|^2 + 6\|\phi_\tau^\star - \phi_\tau(\pi_{\theta_k})\|^2\\ &\leqslant -\frac{3}{4}(\min_{s,b}\phi_\tau^\star(b\mid s))^2 + \frac{12\log(2)}{\tau\rho_{\min}}\delta_k^\phi + \frac{12\log(2)}{\tau\rho_{\min}}\delta_k^\pi\\ &= -\frac{3}{4}(\min_{s,b}\phi_\tau^\star(b\mid s))^2 + \frac{12\log(2)}{\tau\rho_{\min}}(\delta_k^\pi + \delta_k^\phi).\end{aligned}$$

$3\delta_k^\pi + \delta_k^\phi \leqslant (1 - \frac{\alpha(1-\gamma)\tau\rho_{\min}^2 c^2}{32|\mathcal{S}|})^k(3\delta_0^\pi + \delta_0^\phi) \leqslant 3\delta_0^\pi + \delta_0^\phi \leqslant \frac{\rho_{\min}c^2}{64\log(2)}$ guarantees $\delta_k^\pi + \delta_k^\phi \leqslant \frac{\rho_{\min}c^2}{32\log(2)}$. Using this in the inequality above, we have

$$-(\min_{s,b}\phi_{\psi_k}(b\mid s))^2 \leqslant -\frac{3}{4}(\min_{s,b}\phi_\tau^\star(b\mid s))^2 + \frac{12\log(2)}{\tau\rho_{\min}}(\delta_k^\pi + \delta_k^\phi) \leqslant -\frac{3c^2}{4} + \frac{3c^2}{8} \leqslant -\frac{3c^2}{8}.$$

□

C.2.9 Proof of Lemma C.5

From Lemma 4.4, for any $\psi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{B}|}$

$$\begin{aligned}J_\tau(\pi_{\theta_2},\phi_\psi) - J_\tau(\pi_{\theta_2},\phi_\tau(\pi_{\theta_2})) &\leqslant \frac{|\mathcal{S}|}{2\tau\rho_{\min}\left(\min_{s,a}\phi_\psi(a\mid s)\right)^2}\left\|\frac{d_\rho^{\pi_{\theta_2},\phi_\tau(\pi_{\theta_2})}}{d_\rho^{\pi_{\theta_2},\phi_\psi}}\right\|_\infty \|\nabla_\psi J_\tau(\pi_{\theta_2},\phi_\psi)\|^2\\ &\leqslant \frac{|\mathcal{S}|}{2\tau(1-\gamma)\left(\min_{s,a}\phi_\psi(a\mid s)\right)^2}\|\nabla_\psi J_\tau(\pi_{\theta_2},\phi_\psi)\|^2,\end{aligned}$$

where the second inequality follows by an argument similar to Equation C.34. Letting ψ be the parameter that parameterizes $\phi_\tau(\pi_{\theta_1})$, we have

$$\begin{aligned}
&J_\tau(\pi_{\theta_2}, \phi_\tau(\pi_{\theta_1})) - J_\tau(\pi_{\theta_2}, \phi_\tau(\pi_{\theta_2})) \\
&\leqslant \frac{|\mathcal{S}|}{2\tau(1-\gamma)\left(\min_{s,a}\phi_\tau(\pi_{\theta_1})(a \mid s)\right)^2}\|\nabla_\psi J_\tau(\pi_{\theta_2}, \phi_\tau(\pi_{\theta_1}))\|^2 \\
&= \frac{|\mathcal{S}|}{2\tau(1-\gamma)\left(\min_{s,a}\phi_\tau(\pi_{\theta_1})(a \mid s)\right)^2}\|\nabla_\psi J_\tau(\pi_{\theta_2}, \psi^\star_{\rho,\tau}(\pi_{\theta_1})) - \nabla_\psi J_\tau(\pi_{\theta_1}, \psi^\star_{\rho,\tau}(\pi_{\theta_1}))\|^2 \\
&\leqslant \frac{L^2|\mathcal{S}|}{2\tau(1-\gamma)\left(\min_{s,a}\phi_\tau(\pi_{\theta_1})(a \mid s)\right)^2}\|\theta_1 - \theta_2\|^2,
\end{aligned}$$

where the last inequality follows from the fact that for any $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, $\psi_1, \psi_2 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{B}|}$

$$\begin{aligned}
\|\nabla_\psi J_\tau(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\psi J_\tau(\pi_{\theta_2}, \phi_{\psi_2})\| &\leqslant \|\nabla_\psi J(\pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\psi J(\pi_{\theta_2}, \phi_{\psi_2})\| \\
&\quad + \tau\|\nabla_\psi \mathcal{H}_\pi(s, \pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\psi \mathcal{H}_\pi(s, \pi_{\theta_2}, \phi_{\psi_2})\| \\
&\quad + \tau\|\nabla_\psi \mathcal{H}_\phi(s, \pi_{\theta_1}, \phi_{\psi_1}) - \nabla_\psi \mathcal{H}_\phi(s, \pi_{\theta_2}, \phi_{\psi_2})\| \\
&\leqslant L(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|), \qquad \text{(C.50)}
\end{aligned}$$

which is a result of Lemmas C.1 and C.2.

By Lemma 4.1, we also have

$$J_\tau(\pi_{\theta_2}, \phi_\tau(\pi_{\theta_1})) - J_\tau(\pi_{\theta_2}, \phi_\tau(\pi_{\theta_2})) \geqslant \frac{\tau\rho_{\min}}{2\log(2)}\|\phi_\tau(\pi_{\theta_1}) - \phi_\tau(\pi_{\theta_2})\|^2.$$

Combining the two inequalities and re-arranging the terms, we have

$$\|\phi_\tau(\pi_{\theta_1}) - \phi_\tau(\pi_{\theta_2})\| \leqslant \frac{\sqrt{|\mathcal{S}|\log(2)}L}{\sqrt{(1-\gamma)\rho_{\min}}\tau\left(\min_{s,a}\phi_\tau(\pi_{\theta_1})(a \mid s)\right)}\|\theta_1 - \theta_2\|. \qquad \text{(C.51)}$$

Therefore, by Equation C.3,

$$\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_{k+1}}, \phi_\tau(\pi_{\theta_{k+1}}))\|$$

$$\leqslant L\|\theta_k - \theta_{k+1}\| + L\|\phi_\tau(\pi_{\theta_k}) - \phi_\tau(\pi_{\theta_{k+1}})\|$$
$$\leqslant L\left(1 + \frac{\sqrt{|\mathcal{S}|\log(2)}L}{\sqrt{(1-\gamma)\rho_{\min}}\tau\left(\min_{s,a}\phi_\tau(\pi_{\theta_k})(a \mid s)\right)}\right)\|\theta_k - \theta_{k+1}\|$$

Due to the Danskin's Theorem Equation C.2, this implies that we can perform the expansion

$$\begin{aligned}
&J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - J_\tau(\pi_{\theta_{k+1}}, \phi_\tau(\pi_{\theta_{k+1}})) \\
&\leqslant -\langle \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})), \theta_{k+1} - \theta_k\rangle \\
&\quad + \frac{L}{2}\left(1 + \frac{\sqrt{|\mathcal{S}|\log(2)}L}{\sqrt{(1-\gamma)\rho_{\min}}\tau\left(\min_{s,a}\phi_\tau(\pi_{\theta_k})(a \mid s)\right)}\right)\|\theta_{k+1} - \theta_k\|^2 \\
&\leqslant -\alpha_k\langle \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})), \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\rangle \\
&\quad + \frac{L\alpha_k^2}{2}\left(1 + \frac{\sqrt{|\mathcal{S}|\log(2)}L}{\sqrt{(1-\gamma)\rho_{\min}}\tau\left(\min_{s,a}\phi_\tau(\pi_{\theta_k})(a \mid s)\right)}\right)\|\nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k})\|^2. \quad \text{(C.52)}
\end{aligned}$$

Note that by the property of the min function

$$\begin{aligned}
\min_{s,a}\phi_\tau(\pi_{\theta_k})(a \mid s) &\geqslant \min_{s,a}\phi_\tau^\star(a \mid s) - \max_{s,a}(\phi_\tau^\star(a \mid s) - \phi_\tau(\pi_{\theta_k})(a \mid s)) \\
&\geqslant \min_{s,a}\phi_\tau^\star(a \mid s) - \|\phi_\tau^\star - \phi_\tau(\pi_{\theta_k})\| \\
&\geqslant c - \sqrt{\frac{2\log(2)}{\tau\rho_{\min}}(\delta_k^\pi + \delta_k^\phi)}, \quad \text{(C.53)}
\end{aligned}$$

where the last inequality uses the same argument as in Equation C.58. Since Equation C.23 implies $\delta_k^\pi + \delta_k^\phi \leqslant \frac{\rho_{\min}c^2\tau}{64\log(2)(k+1)^{1/3}}$, we further have

$$\min_{s,a}\phi_\tau(\pi_{\theta_k})(a \mid s) \geqslant c - \sqrt{\frac{2\log(2)}{\tau\rho_{\min}}(\delta_k^\pi + \delta_k^\phi)} \geqslant c(1 - \sqrt{\frac{1}{32}}) \geqslant \frac{c\sqrt{\log(2)}}{2}.$$

Using this bound in Equation C.52,

$$J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - J_\tau(\pi_{\theta_{k+1}}, \phi_\tau(\pi_{\theta_{k+1}}))$$

$$
\begin{aligned}
&\leqslant -\alpha_k \langle \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})), \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) \rangle \\
&\quad + \frac{L\alpha_k^2}{2} \left(1 + \frac{\sqrt{|\mathcal{S}| \log(2)} L}{\sqrt{(1-\gamma)\rho_{\min}} \tau \left(\min_{s,a} \phi_\tau(\pi_{\theta_1})(a \mid s) \right)} \right) \| \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) \|^2 \\
&\leqslant -\alpha_k \langle \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})), \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) \rangle \\
&\quad + \frac{L\alpha_k^2}{2} \left(1 + \frac{2\sqrt{|\mathcal{S}|} L}{\sqrt{(1-\gamma)\rho_{\min}} \tau c} \right) \| \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) \|^2, \qquad \text{(C.54)}
\end{aligned}
$$

With the step size choice $\alpha_k \leqslant \left(L + \frac{2\sqrt{|\mathcal{S}|} L^2}{\sqrt{(1-\gamma)\rho_{\min}} \tau c} \right)^{-1}$, we get

$$
\begin{aligned}
&J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - J_\tau(\pi_{\theta_{k+1}}, \phi_\tau(\pi_{\theta_{k+1}})) \\
&\leqslant -\alpha_k \langle \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})), \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) \rangle \\
&\quad + \frac{L\alpha_k^2}{2} \left(1 + \frac{2\sqrt{|\mathcal{S}|} L}{\sqrt{(1-\gamma)\rho_{\min}} \tau c} \right) \| \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) \|^2 \\
&\leqslant -\alpha_k \langle \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})), \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) \rangle \\
&\quad + \frac{\alpha_k}{2} \| \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) \|^2 \\
&= \frac{\alpha_k}{2} \| \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) - \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_{\psi_k}) \|^2 - \| \nabla_\theta J_\tau(\pi_{\theta_k}, \phi_\tau(\pi_{\theta_k})) \|^2.
\end{aligned}
$$

□

C.2.10 Proof of Lemma C.6

The property of the min and max function implies that

$$
\max_{s,a} (\pi^\star_{\tau_k}(a \mid s) - \pi_{\theta_k}(a \mid s)) + \min_{s,a} \pi_{\theta_k}(a \mid s) \geqslant \min_{s,a} \pi^\star_{\tau_k}(a \mid s).
$$

Since the three terms are all non-negative, the inequality holds after taking the square

$$
(\min_{s,a} \pi^\star_{\tau_k}(a \mid s))^2 \leqslant (\max_{s,a} (\pi^\star_{\tau_k}(a \mid s) - \pi_{\theta_k}(a \mid s)) + \min_{s,a} \pi_{\theta_k}(a \mid s))^2
$$

$$\leqslant \frac{4}{3}(\min_{s,a} \pi_{\theta_k}(a \mid s))^2 + 4(\max_{s,a}(\pi^\star_{\tau_k}(a \mid s) - \pi_{\theta_k}(a \mid s)))^2.$$

Re-arranging the terms,

$$\begin{aligned}
-\left(\min_{s,a} \pi_{\theta_k}(a \mid s)\right)^2 &\leqslant -\frac{3}{4}\left(\min_{s,a} \pi^\star_{\tau_k}(a \mid s)\right)^2 + 3\left(\max_{s,a} \pi^\star_{\tau_k}(a \mid s) - \pi_{\phi_k}(a \mid s)\right)^2 \\
&\leqslant -\frac{3}{4}\left(\min_{s,a} \pi^\star_{\tau_k}(a \mid s)\right)^2 + 3\|\pi^\star_{\tau_k} - \pi_{\phi_k}\|^2
\end{aligned}$$

From Lemma 4.1,

$$\begin{aligned}
-\left(\min_{s,a} \pi_{\theta_k}(a \mid s)\right)^2 &\leqslant -\frac{3}{4}\left(\min_{s,a} \pi^\star_{\tau_k}(a \mid s)\right)^2 + 3\|\pi^\star_{\tau_k} - \pi_{\phi_k}\|^2 \\
&\leqslant -\frac{3}{4}\left(\min_{s,a} \pi^\star_{\tau_k}(a \mid s)\right)^2 + \frac{6\log(2)}{\tau_k \rho_{\min}}(J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi^\star_{\tau_k})) \\
&\leqslant -\frac{3}{4}\left(\min_{s,a} \pi^\star_{\tau_k}(a \mid s)\right)^2 + \frac{6\log(2)}{\tau_k \rho_{\min}}(J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))) \\
&= -\frac{3}{4}\left(\min_{s,a} \pi^\star_{\tau_k}(a \mid s)\right)^2 + \frac{6\log(2)}{\tau_k \rho_{\min}}\delta^\pi_k, \qquad \text{(C.55)}
\end{aligned}$$

Since $3\delta^\pi_k + \delta^\phi_k \leqslant \frac{\rho\tau_k c^2}{64\log(2)}$, we have $\delta^\pi_k \leqslant \frac{\rho\tau_k c^2}{64\log(2)}$, which along with Equation C.55 implies

$$-\left(\min_{s,a} \pi_{\theta_k}(a \mid s)\right)^2 \leqslant -\frac{3}{4}\left(\min_{s,a} \pi^\star_{\tau_k}(a \mid s)\right)^2 + \frac{6\log(2)}{\tau_k \rho_{\min}}\delta^\pi_k \leqslant -\frac{3c^2}{4} + \frac{3c^2}{32} \leqslant -\frac{3c^2}{8}.$$

Similarly, the property of the min and max function implies that

$$\max_{s,b}(\phi^\star_{\tau_k}(b \mid s) - \phi_{\psi_k}(b \mid s)) + \min_{s,b} \phi_{\psi_k}(b \mid s) \geqslant \min_{s,b} \phi^\star_{\tau_k}(b \mid s).$$

Again, all three terms are non-negative, which means that the inequality is preserved after taking the square

$$(\min_{s,b} \phi^\star_{\tau_k}(b \mid s))^2 \leqslant (\min_{s,b} \phi_{\psi_k}(b \mid s) + \max_{s,b}(\phi^\star_{\tau_k}(b \mid s) - \phi_{\psi_k}(b \mid s)))^2$$

$$\leqslant \frac{4}{3}(\min_{s,b} \phi_{\psi_k}(b \mid s))^2 + 4(\max_{s,b}(\phi^\star_{\tau_k}(b \mid s) - \phi_{\psi_k}(b \mid s)))^2,$$

which leads to

$$\begin{aligned}
-(\min_{s,b} \phi_{\psi_k}(b \mid s))^2 &\leqslant -\frac{3}{4}(\min_{s,b} \phi^\star_{\tau_k}(b \mid s))^2 + 3(\max_{s,b}(\phi^\star_{\tau_k}(b \mid s) - \phi_{\psi_k}(b \mid s)))^2 \\
&\leqslant -\frac{3}{4}(\min_{s,b} \phi^\star_{\tau_k}(b \mid s))^2 + 3\|\phi^\star_{\tau_k} - \phi_{\psi_k}\|^2 \\
&\leqslant -\frac{3}{4}(\min_{s,b} \phi^\star_{\tau_k}(b \mid s))^2 + 6\|\phi_{\tau_k}(\pi_{\theta_k}) - \phi_{\psi_k}\|^2 + 6\|\phi^\star_{\tau_k} - \phi_{\tau_k}(\pi_{\theta_k})\|^2.
\end{aligned} \tag{C.56}$$

From Lemma 4.1,

$$\|\phi_{\tau_k}(\pi_{\theta_k}) - \phi_{\psi_k}\|^2 \leqslant \frac{2\log(2)}{\tau_k \rho_{\min}} \left(J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\right) = \frac{2\log(2)}{\tau_k \rho_{\min}} \delta_k^\phi, \tag{C.57}$$

and

$$\begin{aligned}
&\|\phi^\star_{\tau_k} - \phi_{\tau_k}(\pi_{\theta_k})\|^2 \\
&\leqslant \frac{2\log(2)}{\tau_k \rho_{\min}} \left(J_{\tau_k}(\pi_{\theta_k}, \phi^\star_{\tau_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\right) \\
&\leqslant \frac{2\log(2)}{\tau_k \rho_{\min}} \Big(\left(J_{\tau_k}(\pi_{\theta_k}, \phi^\star_{\tau_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\right) + \underbrace{\left(J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\right)}_{\delta_k^\phi} \Big) \\
&= \frac{2\log(2)}{\tau_k \rho_{\min}} \Big(\left(J_{\tau_k}(\pi_{\theta_k}, \phi^\star_{\tau_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\right) + \left(J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\right) + \delta_k^\phi \Big) \\
&\leqslant \frac{2\log(2)}{\tau_k \rho_{\min}} \left(J_{\tau_k}(\pi_{\theta_k}, \phi^\star_{\tau_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) + \delta_k^\phi\right) \\
&\leqslant \frac{2\log(2)}{\tau_k \rho_{\min}} \left(J_{\tau_k}(\pi^\star_{\tau_k}, \phi^\star_{\tau_k}) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) + \delta_k^\phi\right) \\
&= \frac{2\log(2)}{\tau_k \rho_{\min}} \left(\delta_k^\pi + \delta_k^\phi\right),
\end{aligned} \tag{C.58}$$

where the third inequality follows from $J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) \leqslant 0$.

Using Equation C.57 and Equation C.58 in Equation C.56,

$$\begin{aligned}-(\min_{s,b}\phi_{\psi_k}(b\mid s))^2 &\leqslant -\frac{3}{4}(\min_{s,b}\phi^\star_{\tau_k}(b\mid s))^2 + 6\|\phi_{\tau_k}(\pi_{\theta_k}) - \phi_{\psi_k}\|^2 + 6\|\phi^\star_{\tau_k} - \phi_{\tau_k}(\pi_{\theta_k})\|^2 \\ &\leqslant -\frac{3}{4}(\min_{s,b}\phi^\star_{\tau_k}(b\mid s))^2 + \frac{12\log(2)}{\tau_k\rho_{\min}}\delta_k^\phi + \frac{12\log(2)}{\tau_k\rho_{\min}}(\delta_k^\pi + \delta_k^\phi) \\ &= -\frac{3}{4}(\min_{s,b}\phi^\star_{\tau_k}(b\mid s))^2 + \frac{12\log(2)}{\tau_k\rho_{\min}}(\delta_k^\pi + 2\delta_k^\phi).\end{aligned}$$

$3\delta_k^\pi + \delta_k^\phi \leqslant \frac{\rho\tau_k c^2}{64\log(2)}$ implies that $\delta_k^\pi + 2\delta_k^\phi \leqslant \frac{\rho\tau_k c^2}{32\log(2)}$. Using this in the inequality above,

$$-(\min_{s,b}\phi_{\psi_k}(b\mid s))^2 \leqslant -\frac{3}{4}(\min_{s,b}\phi^\star_{\tau_k}(b\mid s))^2 + \frac{12\log(2)}{\tau_k\rho_{\min}}(\delta_k^\pi + 2\delta_k^\phi) \leqslant -\frac{3c^2}{4} + \frac{12c^2}{32} \leqslant -\frac{3c^2}{8}.$$

□

C.2.11 Proof of Lemma C.7

From Lemma 4.4, for any $\psi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{B}|}$

$$\begin{aligned}&J_{\tau_k}(\pi_{\theta_2},\phi_\psi) - J_{\tau_k}(\pi_{\theta_2},\phi_{\tau_k}(\pi_{\theta_2})) \\ &\leqslant \frac{|\mathcal{S}|}{2\tau_k\rho_{\min}\left(\min_{s,a}\phi_\psi(a\mid s)\right)^2}\left\|\frac{d_\rho^{\pi_{\theta_2},\phi_{\tau_k}(\pi_{\theta_2})}}{d_\rho^{\pi_{\theta_2},\phi_\psi}}\right\|_\infty \|\nabla_\psi J_{\tau_k}(\pi_{\theta_2},\phi_\psi)\|^2 \\ &\leqslant \frac{|\mathcal{S}|}{2\tau_k(1-\gamma)\left(\min_{s,a}\phi_\psi(a\mid s)\right)^2}\|\nabla_\psi J_{\tau_k}(\pi_{\theta_2},\phi_\psi)\|^2,\end{aligned}$$

where the second inequality follows by an argument similar to Equation C.34. Letting ψ be the parameter that parameterizes $\phi_{\tau_k}(\pi_{\theta_1})$ and defining $L_k = L_{\mathcal{H}}(2\tau_k+1)$, we have

$$\begin{aligned}&J_{\tau_k}(\pi_{\theta_2},\phi_{\tau_k}(\pi_{\theta_1})) - J_{\tau_k}(\pi_{\theta_2},\phi_{\tau_k}(\pi_{\theta_2})) \\ &\leqslant \frac{|\mathcal{S}|}{2\tau_k(1-\gamma)\left(\min_{s,a}\phi_{\tau_k}(\pi_{\theta_1})(a\mid s)\right)^2}\|\nabla_\psi J_{\tau_k}(\pi_{\theta_2},\phi_{\tau_k}(\pi_{\theta_1}))\|^2 \\ &= \frac{|\mathcal{S}|}{2\tau_k(1-\gamma)\left(\min_{s,a}\phi_{\tau_k}(\pi_{\theta_1})(a\mid s)\right)^2}\|\nabla_\psi J_{\tau_k}(\pi_{\theta_2},\psi^\star_{\rho,\tau_k}(\pi_{\theta_1})) - \nabla_\psi J_{\tau_k}(\pi_{\theta_1},\psi^\star_{\rho,\tau_k}(\pi_{\theta_1}))\|^2\end{aligned}$$

$$\leqslant \frac{L_k^2|\mathcal{S}|}{2\tau_k(1-\gamma)\left(\min_{s,a}\phi_{\tau_k}(\pi_{\theta_1})(a \mid s)\right)^2}\|\theta_1-\theta_2\|^2,$$

where the last inequality uses the same argument as Equation C.50.

By Lemma 4.1, we also have

$$J_{\tau_k}(\pi_{\theta_2},\phi_{\tau_k}(\pi_{\theta_1})) - J_{\tau_k}(\pi_{\theta_2},\phi_{\tau_k}(\pi_{\theta_2})) \geqslant \frac{\tau_k\rho_{\min}}{2\log(2)}\|\phi_{\tau_k}(\pi_{\theta_1})-\phi_{\tau_k}(\pi_{\theta_2})\|^2.$$

Combining the two inequalities and re-arranging the terms, we have

$$\|\phi_{\tau_k}(\pi_{\theta_1})-\phi_{\tau_k}(\pi_{\theta_2})\| \leqslant \frac{\sqrt{|\mathcal{S}|\log(2)}L_k}{\sqrt{(1-\gamma)\rho_{\min}}\tau_k\left(\min_{s,a}\phi_{\tau_k}(\pi_{\theta_1})(a \mid s)\right)}\|\theta_1-\theta_2\|. \tag{C.59}$$

Therefore, by Equation C.3,

$$\begin{aligned}
&\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_{k+1}},\phi_{\tau_k}(\pi_{\theta_{k+1}}))\| \\
&\leqslant L_k\|\theta_k-\theta_{k+1}\| + L_k\|\phi_{\tau_k}(\pi_{\theta_k})-\phi_{\tau_k}(\pi_{\theta_{k+1}})\| \\
&\leqslant L_k\left(1+\frac{\sqrt{|\mathcal{S}|\log(2)}L_k}{\sqrt{(1-\gamma)\rho_{\min}}\tau_k\left(\min_{s,a}\phi_{\tau_k}(\pi_{\theta_k})(a \mid s)\right)}\right)\|\theta_k-\theta_{k+1}\|
\end{aligned}$$

Due to the Danskin's Theorem Equation C.2, this implies that we can perform the expansion

$$\begin{aligned}
&J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_k}(\pi_{\theta_{k+1}},\phi_{\tau_k}(\pi_{\theta_{k+1}})) \\
&\leqslant -\langle\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})),\theta_{k+1}-\theta_k\rangle \\
&\quad + \frac{L_k}{2}\left(1+\frac{\sqrt{|\mathcal{S}|\log(2)}L_k}{\sqrt{(1-\gamma)\rho_{\min}}\tau_k\left(\min_{s,a}\phi_{\tau_k}(\pi_{\theta_k})(a \mid s)\right)}\right)\|\theta_{k+1}-\theta_k\|^2 \\
&\leqslant -\alpha_k\langle\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\tau_k}(\pi_{\theta_k})),\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\rangle \\
&\quad + \frac{L_k\alpha_k^2}{2}\left(1+\frac{\sqrt{|\mathcal{S}|\log(2)}L_k}{\sqrt{(1-\gamma)\rho_{\min}}\tau_k\left(\min_{s,a}\phi_{\tau_k}(\pi_{\theta_k})(a \mid s)\right)}\right)\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k},\phi_{\psi_k})\|^2.
\end{aligned} \tag{C.60}$$

Note that by the property of the min function

$$\begin{aligned}\min_{s,a} \phi_{\tau_k}(\pi_{\theta_k})(a \mid s) &\geqslant \min_{s,a} \phi^\star_{\tau_k}(a \mid s) - \max_{s,a}(\phi^\star_{\tau_k}(a \mid s) - \phi_{\tau_k}(\pi_{\theta_k})(a \mid s)) \\ &\geqslant \min_{s,a} \phi^\star_{\tau_k}(a \mid s) - \|\phi^\star_{\tau_k} - \phi_{\tau_k}(\pi_{\theta_k})\| \\ &\geqslant c - \sqrt{\frac{2\log(2)}{\tau_k \rho_{\min}}(\delta_k^\pi + \delta_k^\phi)}, \end{aligned} \tag{C.61}$$

where the last inequality uses the same argument as in Equation C.58. Since Equation C.23 implies $\delta_k^\pi + \delta_k^\phi \leqslant \frac{\rho_{\min} c^2 \tau_0}{64\log(2)(k+1)^{1/3}}$, we further have

$$\min_{s,a} \phi_{\tau_k}(\pi_{\theta_k})(a \mid s) \geqslant c - \sqrt{\frac{2\log(2)}{\tau_k \rho_{\min}}(\delta_k^\pi + \delta_k^\phi)} \geqslant c(1 - \sqrt{\frac{1}{32}}) \geqslant \frac{c\sqrt{\log(2)}}{2}.$$

Using this bound in Equation C.60,

$$\begin{aligned}
&J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_k}^{\pi_{\theta_{k+1}}, \phi_{\tau_k}(\pi_{\theta_{k+1}})}(\rho) \\
&\leqslant -\alpha_k \langle \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})), \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) \rangle \\
&\quad + \frac{L_k \alpha_k^2}{2} \left(1 + \frac{\sqrt{|\mathcal{S}|\log(2)} L_k}{\sqrt{(1-\gamma)\rho_{\min}} \tau_k \left(\min_{s,a} \phi_{\tau_k}(\pi_{\theta_1})(a \mid s)\right)}\right) \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \\
&\leqslant -\alpha_k \langle \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})), \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) \rangle \\
&\quad + \frac{L_k \alpha_k^2}{2} \left(1 + \frac{2\sqrt{|\mathcal{S}|} L_k}{\sqrt{(1-\gamma)\rho_{\min}} \tau_k c}\right) \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \\
&\leqslant -\alpha_k \langle \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})), \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) \rangle + \frac{\alpha_k^2}{2} \left(L_k + \frac{C_2 L_k^2}{\tau_k}\right) \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2.
\end{aligned} \tag{C.62}$$

The condition on h, which is $\frac{\alpha_0}{h^{2/3}} \leqslant (2L_{\mathcal{H}} + 4L_{\mathcal{H}}^2 C_2)\frac{\tau_0}{h^{1/3}} + (L_{\mathcal{H}} + 4L_{\mathcal{H}}^2 C_2) + \frac{L_{\mathcal{H}}^2 C_2 h^{1/3}}{\tau_0}$, can be equivalently expressed as $\alpha_0 \left(L_0 + \frac{C_2 L_0^2}{\tau_0}\right) \leqslant 1$. Since α_k decays faster than τ_k, this

guarantees for all $k \geqslant 0$

$$\alpha_k \left(L_k + \frac{C_2 L_k^2}{\tau_k} \right) \leqslant 1.$$

Using this inequality in Equation C.62, we get

$$\begin{aligned}
&J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - J_{\tau_k}(\pi_{\theta_{k+1}}, \phi_{\tau_k}(\pi_{\theta_{k+1}})) \\
&\leqslant -\alpha_k \langle \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})), \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k}) \rangle + \frac{\alpha_k}{2} \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2 \\
&= \frac{\alpha_k}{2} \left(\|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k})) - \nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\psi_k})\|^2 - \|\nabla_\theta J_{\tau_k}(\pi_{\theta_k}, \phi_{\tau_k}(\pi_{\theta_k}))\|^2 \right).
\end{aligned}$$

□

C.3 Experiment Details

We first discuss the design of the completely mixed Markov game. The dimension of state space is 2, and so is the dimension of the action spaces of both players. Using s_1, s_2 to denote the two states, we can essentially describe $\mathcal{P}$ as a $2 \times 2 \times 2 \times 2$ tensor where $\mathcal{P}(s' \mid s, \cdot, \cdot)$ is a 2×2 matrix for any $s, s' \in \mathcal{S}$ with rows corresponding to the action of the first player and columns corresponding to the second player

$$\mathcal{P}(s_1 \mid s_1, \cdot, \cdot) = \begin{bmatrix} 0.2 & 0.5 \\ 0.5 & 0.1 \end{bmatrix}, \quad \mathcal{P}(s_2 \mid s_1, \cdot, \cdot) = \begin{bmatrix} 0.8 & 0.5 \\ 0.5 & 0.9 \end{bmatrix},$$
$$\mathcal{P}(s_1 \mid s_2, \cdot, \cdot) = \begin{bmatrix} 0.3 & 0.2 \\ 0.6 & 0.2 \end{bmatrix}, \quad \mathcal{P}(s_2 \mid s_2, \cdot, \cdot) = \begin{bmatrix} 0.7 & 0.8 \\ 0.4 & 0.8 \end{bmatrix}.$$

Similarly, the reward function can be described by a $2 \times 2 \times 2$ tensor where $r(s, \cdot, \cdot)$ is a 2×2 matrix for any $s \in \mathcal{S}$ with rows corresponding to the action of the first player and

columns corresponding to the second player

$$r(s_1, \cdot, \cdot) = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, \quad r(s_2, \cdot, \cdot) = \begin{bmatrix} 6 & 4 \\ 3 & 10 \end{bmatrix}.$$

Under the initial distribution $\rho = [0.5, 0.5]^\top$ and discount factor $\gamma = 0.9$, the (approximate) Nash equilibrium of this Markov game is

$$\pi^\star(\cdot \mid s_1) = [0.812, 0.188], \quad \pi^\star(\cdot \mid s_2) = [0.837, 0.163],$$
$$\phi^\star(\cdot \mid s_1) = [0.880, 0.120], \quad \phi^\star(\cdot \mid s_2) = [0.597, 0.403].$$

To design the Markov game that does not observe Assumption 4.2, we use the same transition probability matrices as in the completely mixed Markov game case. The reward function is

$$r(s_1, \cdot, \cdot) = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad r(s_2, \cdot, \cdot) = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}.$$

Under the initial distribution $\rho = [0.5, 0.5]^\top$ and discount factor $\gamma = 0.9$, it can be easily seen that the Nash equilibrium of this Markov game is unique and is

$$\pi^\star(\cdot \mid s_1) = [0, 1], \quad \pi^\star(\cdot \mid s_2) = [0, 1],$$
$$\phi^\star(\cdot \mid s_1) = [1, 0], \quad \phi^\star(\cdot \mid s_2) = [1, 0].$$

Since the Nash equilibrium consists of a pair of deterministic policies, Assumption 4.2 is not satisfied in this case.

REFERENCES

[1] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] D. Silver, A. Huang, C. J. Maddison, *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, p. 484, 2016.

[3] OpenAI, C. Berner, G. Brockman, B. Chan, *et al.*, "Dota 2 with large scale deep reinforcement learning," 2019. arXiv: 1912.06680.

[4] C. Yu, J. Liu, and S. Nemati, "Reinforcement learning in healthcare: A survey," *arXiv preprint arXiv:1908.08796*, 2019.

[5] A. Esteva *et al.*, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[6] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[7] T. Haarnoja *et al.*, *Soft actor-critic algorithms and applications*, available at: https://arxiv.org/abs/1812.05905, 2019.

[8] H. Kretzschmar, M. Spies, C. Sprunk, and W. Burgard, "Socially compliant mobile robot navigation via inverse reinforcement learning," *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1289–1307, 2016.

[9] Y. Zhu *et al.*, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *2017 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2017, pp. 3357–3364.

[10] A. Anwar and A. Raychowdhury, "Autonomous navigation via deep reinforcement learning for resource constraint edge nodes using transfer learning," *IEEE Access*, vol. 8, pp. 26 549–26 560, 2020.

[11] J. Bhandari and D. Russo, "Global optimality guarantees for policy gradient methods," *arXiv preprint arXiv:1906.01786*, 2019.

[12] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "Optimality and approximation with policy gradient methods in markov decision processes," ser. Proceedings of Machine Learning Research, vol. 125, 2020, pp. 64–66.

[13] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *International Conference on Machine Learning*, PMLR, 2020, pp. 6820–6829.

[14] L. Espeholt *et al.*, "IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures," ser. Proceedings of Machine Learning Research, vol. 80, 2018, pp. 1407–1416.

[15] M. Hessel, H. Soyer, L. Espeholt, W. Czarnecki, S. Schmitt, and H. van Hasselt, "Multi-task deep reinforcement learning with popart," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3796–3803.

[16] J. Qi, Q. Zhou, L. Lei, and K. Zheng, "Federated reinforcement learning: Techniques, applications, and open challenges," *arXiv preprint arXiv:2108.11887*, 2021.

[17] K. Ovchinnikov, A. Semakova, and A. Matveev, "Decentralized multi-agent tracking of unknown environmental level sets by a team of nonholonomic robots," in *2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, IEEE, 2014, pp. 352–359.

[18] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[19] K. I. Tsianos and M. G. Rabbat, "Distributed strongly convex optimization," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2012, pp. 593–600.

[20] D. Jakovetić, J. Xavier, and J. M. Moura, "Fast distributed gradient methods," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[21] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

[22] T. Lin, C. Jin, and M. Jordan, "On gradient descent ascent for nonconvex-concave minimax problems," in *International Conference on Machine Learning*, PMLR, 2020, pp. 6083–6093.

[23] T. Lin, C. Jin, and M. I. Jordan, "Near-optimal algorithms for minimax optimization," in *Conference on Learning Theory*, PMLR, 2020, pp. 2738–2779.

[24] Y. Wang and J. Li, "Improved algorithms for convex-concave minimax optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4800–4810, 2020.

[25] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn, "Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems," *SIAM Journal on Optimization*, vol. 31, no. 4, pp. 2508–2538, 2021.

[26] C. Jin, P. Netrapalli, and M. Jordan, "What is local optimality in nonconvex-nonconcave minimax optimization?" In *International Conference on Machine Learning*, PMLR, 2020, pp. 4880–4889.

[27] M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn, "Solving a class of non-convex min-max games using iterative first order methods," *arXiv preprint arXiv:1902.08297*, 2019.

[28] J. Yang, N. Kiyavash, and N. He, "Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems," *arXiv preprint arXiv:2002.09621*, 2020.

[29] C. Daskalakis, D. J. Foster, and N. Golowich, "Independent policy gradient methods for competitive reinforcement learning," *Advances in neural information processing systems*, vol. 33, pp. 5527–5540, 2020.

[30] H. H. Zhuo, W. Feng, Y. Lin, Q. Xu, and Q. Yang, "Federated deep reinforcement learning," *arXiv preprint arXiv:1901.08277*, 2019.

[31] C. Nadiger, A. Kumar, and S. Abdelhak, "Federated reinforcement learning for fast personalization," in *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, IEEE, 2019, pp. 123–127.

[32] L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta, "Robust adversarial reinforcement learning," in *International Conference on Machine Learning*, PMLR, 2017, pp. 2817–2826.

[33] S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, and S. Russell, "Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4213–4220.

[34] S. Zeng, T. T. Doan, and J. Romberg, "A two-time-scale stochastic optimization framework with applications in control and reinforcement learning," *arXiv preprint arXiv:2109.14756*, 2021.

[35] S. Zeng, M. A. Anwar, T. T. Doan, A. Raychowdhury, and J. Romberg, "A decentralized policy gradient approach to multi-task reinforcement learning," in *Uncertainty in Artificial Intelligence*, PMLR, 2021, pp. 1002–1012.

[36] S. Zeng, T. T. Doan, and J. Romberg, "Finite-time analysis of decentralized stochastic approximation with applications in multi-agent and multi-task learning," in *IEEE Conference on Decision and Control (CDC)*, IEEE, 2021, pp. 2641–2646.

[37] S. Zeng, T. T. Doan, and J. Romberg, "Finite-time convergence rates of decentralized stochastic approximation with applications in multi-agent and multi-task learning," *IEEE Transactions on Automatic Control*, 2022.

[38] S. Zeng, T. T. Doan, and J. Romberg, "Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained markov decision processes," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, IEEE, 2022, pp. 4028–4033.

[39] S. Zeng, T. T. Doan, and J. Romberg, "Regularized gradient descent ascent for two-player zero-sum markov games," in *Advances in Neural Information Processing Systems*, 2022.

[40] S. Zeng, A. Kody, Y. Kim, K. Kim, and D. K. Molzahn, "A reinforcement learning approach to parameter selection for distributed optimal power flow," *Electric Power Systems Research*, vol. 212, p. 108 546, 2022.

[41] V. S. Borkar and S. P. Meyn, "The ODE method for convergence of stochastic approximation and reinforcement learning," *SIAM Journal on Control and Optimization*, vol. 38, no. 2, pp. 447–469, 2000.

[42] V. R. Konda and J. N. Tsitsiklis, "Convergence rate of linear two-time-scale stochastic approximation," *The Annals of Applied Probability*, vol. 14, no. 2, pp. 796–819, 2004.

[43] G. Dalal, G. Thoppe, B. Szörényi, and S. Mannor, "Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning," in *COLT*, 2018.

[44] G. Dalal, B. Szorenyi, and G. Thoppe, "A tale of two-timescale reinforcement learning with the tightest finite-time bound," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 3701–3708, Apr. 2020.

[45] T. T. Doan and J. Romberg, "Linear two-time-scale stochastic approximation a finite-time analysis," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2019, pp. 399–406.

[46] H. Gupta, R. Srikant, and L. Ying, "Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning," in *Advances in Neural Information Processing Systems*, 2019.

[47] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and H.-T. Wai, "Finite time analysis of linear two-timescale stochastic approximation with Markovian noise," in *Proceedings of Thirty Third Conference on Learning Theory*, vol. 125, 2020, pp. 2144–2203.

[48] A. Mokkadem and M. Pelletier, "Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms," *The Annals of Applied Probability*, vol. 16, no. 3, pp. 1671–1702, 2006.

[49] T. T. Doan, "Finite-time convergence rates of nonlinear two-time-scale stochastic approximation under Markovian noise," *arXiv:2104.01627*, 2021.

[50] B. Colson, P. Marcotte, and G. Savard, "An overview of bilevel optimization," *Annals of operations research*, vol. 153, no. 1, pp. 235–256, 2007.

[51] M. Hong, H.-T. Wai, Z. Wang, and Z. Yang, "A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic," *arXiv:2007.05170*, 2020.

[52] T. Chen, Y. Sun, Q. Xiao, and W. Yin, "A single-timescale method for stochastic bilevel optimization," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 2466–2488.

[53] M. Wang, E. X. Fang, and H. Liu, "Stochastic compositional gradient descent: Algorithms for minimizing compositions of expected-value functions," *Mathematical Programming*, vol. 161, no. 1, pp. 419–449, 2017.

[54] T. Chen, Y. Sun, and W. Yin, "Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization," *IEEE Transactions on Signal Processing*, vol. 69, pp. 4937–4948, 2021.

[55] S. Qiu, Z. Yang, J. Ye, and Z. Wang, "On finite-time convergence of actor-critic algorithm," *IEEE Journal on Selected Areas in Information Theory*, 2021.

[56] H. Kumar, A. Koppel, and A. Ribeiro, "On the sample complexity of actor-critic method for reinforcement learning with function approximation," *Machine Learning*, pp. 1–35, 2023.

[57] T. Xu, Z. Wang, and Y. Liang, "Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms," *arXiv:2005.03557*, 2020.

[58] Y. Wu, W. Zhang, P. Xu, and Q. Gu, "A finite time analysis of two time-scale actor critic methods," *arXiv preprint arXiv:2005.01350*, 2020.

[59] S.-i. Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing*, vol. 5, no. 4-5, pp. 185–196, 1993.

[60] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, "Sgd: General analysis and improved rates," in *International Conference on Machine Learning*, PMLR, 2019, pp. 5200–5209.

[61] A. Khaled, K. Mishchenko, and P. Richtárik, "Tighter theory for local sgd on identical and heterogeneous data," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 4519–4529.

[62] A. Ruszczyński, "A linearization method for nonsmooth stochastic programming problems," *Mathematics of Operations Research*, vol. 12, no. 1, pp. 32–49, 1987.

[63] S. Boyd and A. Mutapcic, "Stochastic subgradient methods," *Lecture Notes for EE364b, Stanford University*, 2008.

[64] A. Ruszczyński, "Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization," *Optimization Letters*, vol. 14, no. 7, pp. 1615–1625, 2020.

[65] Z. Chen, S. Zhang, T. T. Doan, J.-P. Clarke, and S. Theja Maguluri, "Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning," *arXiv e-prints*, arXiv–1905, 2019.

[66] T. Sun, Y. Sun, and W. Yin, "On markov chain gradient descent," *Advances in neural information processing systems*, vol. 31, 2018.

[67] S. Zou, T. Xu, and Y. Liang, "Finite-sample analysis for sarsa with linear function approximation," *Advances in neural information processing systems*, vol. 32, 2019.

[68] D. Bertsekas, *Dynamic programming and optimal control: Volume I*. Athena scientific, 2012, vol. 1.

[69] B. Gravell, P. M. Esfahani, and T. Summers, "Learning optimal controllers for linear systems with multiplicative noise via policy gradient," *IEEE Transactions on Automatic Control*, vol. 66, no. 11, pp. 5283–5298, 2020.

[70] Z. Yang, Y. Chen, M. Hong, and Z. Wang, "On the global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost," *arXiv preprint arXiv:1907.06246*, 2019.

[71] V. R. Konda, "Actor-critic algorithms," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[72] A. Barakat, P. Bianchi, and J. Lehmann, "Analysis of a target-based actor-critic algorithm with linear function approximation," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 991–1040.

[73] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, "Fast global convergence of natural policy gradient methods with entropy regularization," *Operations Research*, 2021.

[74] R. S. Sutton, C. Szepesvári, and H. R. Maei, "A convergent o (n) algorithm for off-policy temporal-difference learning with linear function approximation," *Advances in neural information processing systems*, vol. 21, no. 21, pp. 1609–1616, 2008.

[75] R. S. Sutton *et al.*, "Fast gradient-descent methods for temporal-difference learning with linear function approximation," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 993–1000.

[76] T. Xu, S. Zou, and Y. Liang, "Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[77] P. Karmakar and S. Bhatnagar, "Two time-scale stochastic approximation with controlled markov noise and off-policy temporal-difference learning," *Mathematics of Operations Research*, vol. 43, no. 1, pp. 130–151, 2018.

[78] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. American Mathematical Society, 2006.

[79] S. Zeng, T. T. Doan, and J. Romberg, "Connected superlevel set in (deep) reinforcement learning and its application to minimax theorems," *arXiv preprint arXiv:2303.12981*, 2023.

[80] B. T. Polyak, "Introduction to optimization. translations series in mathematics and engineering.," *Optimization Software, Inc, New York*, 1987.

[81] S. Lojasiewicz, "A topological property of real analytic subsets," *Coll. du CNRS, Les équations aux dérivées partielles*, vol. 117, pp. 87–89, 1963.

[82] H. Karimi, J. Nutini, and M. Schmidt, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2016, pp. 795–811.

[83] C. Liu, L. Zhu, and M. Belkin, "Toward a theory of optimization for over-parameterized systems of non-linear equations: The lessons of deep learning," *arXiv:2003.00307*, 2020.

[84] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," *arXiv preprint arXiv:1801.05039*, 2018.

[85] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, *Gradient surgery for multi-task learning*, available at: https://arxiv.org/abs/2001.06782, 2020.

[86] A. A. Rusu *et al.*, "Policy distillation," *arXiv preprint arXiv:1511.06295*, 2015.

[87] R. Traoré *et al.*, "Discorl: Continual reinforcement learning via policy distillation," *arXiv preprint arXiv:1907.05855*, 2019.

[88] L. T. Liu, U. Dogan, and K. Hofmann, "Decoding multitask dqn in the world of minecraft," in *The 13th European Workshop on Reinforcement Learning (EWRL) 2016*, 2016.

[89] A. Gupta, C. Devin, Y. Liu, P. Abbeel, and S. Levine, "Learning invariant feature spaces to transfer skills with reinforcement learning," *arXiv preprint arXiv:1703.02949*, 2017.

[90] C. DEramo, D. Tateo, A. Bonarini, M. Restelli, and J. Peters, "Sharing knowledge in multi-task deep reinforcement learning," in *Eighth International Conference on Learning Representations (ICLR 2020)*, 2020.

[91] V. Mnih *et al.*, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*, 2016, pp. 1928–1937.

[92] A. Nair, P. Srinivasan, S. Blackwell, *et al.*, "Massively parallel methods for deep reinforcement learning," Jul. 2015.

[93] M. Assran, J. Romoff, N. Ballas, J. Pineau, and M. Rabbat, "Gossip-based actor-learner architectures for deep reinforcement learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 13 320–13 330.

[94] J. X. Wang *et al.*, "Learning to reinforcement learn," *arXiv preprint arXiv:1611.05763*, 2016.

[95] A. Nagabandi *et al.*, "Learning to adapt in dynamic, real-world environments through meta-reinforcement learning," *arXiv preprint arXiv:1803.11347*, 2018.

[96] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," ser. Proceedings of Machine Learning Research, vol. 80, 2018, pp. 5872–5881.

[97] K. Zhang, Z. Yang, and T. Başar, *Multi-agent reinforcement learning: A selective overview of theories and algorithms*, available at: https://arxiv.org/abs/1911.10635, 2019.

[98] T. Chu, S. Chinchali, and S. Katti, "Multi-agent reinforcement learning for networked system control," in *International Conference on Learning Representations (ICLR)*, 2020.

[99] G. Qu and N. L. A. Wierman, *Scalable reinforcement learning of localized policies for multi-agent networked systems*, available at: https://arxiv.org/abs/1912.02906, 2019.

[100] T. T. Doan, S. T. Maguluri, and J. Romberg, "Finite-time performance of distributed temporal-difference learning with linear function approximation," *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 1, pp. 298–320, 2021.

[101] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanović, *Fast multi-agent temporal-difference learning via homotopy stochastic primal-dual optimization*, available at: https://arxiv.org/abs/1908.02805, 2019.

[102] W. Li, B. Jin, X. Wang, J. Yan, and H. Zha, *F2a2: Flexible fully-decentralized approximate actor-critic for cooperative multi-agent reinforcement learning*, available at: https://arxiv.org/abs/2004.11145, 2020.

[103] H.-T. Wai, Z. Yang, Z. Wang, and M. Hong, "Multi-agent reinforcement learning via double averaging primal-dual optimization," in *Annual Conference on Neural Information Processing Systems*, 2018, pp. 9672–9683.

[104] S. Kar, J. M. F. Moura, and H. V. Poor, "Qd-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations," *IEEE Trans. Signal Processing*, vol. 61, pp. 1848–1862, 2013.

[105] D. Lee, N. He, P. Kamalaruban, and V. Cevher, *Optimization for reinforcement learning: From single agent to cooperative agents*, available at: https://arxiv.org/abs/1912.00498, 2019.

[106] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, 1st. USA: John Wiley & Sons, Inc., 1994.

[107] A. Olshevsky, "Linear time average consensus on fixed graphs," *IFAC-PapersOnLine*, vol. 48, no. 22, pp. 94–99, 2015.

[108] "Https://icsrl.ece.gatech.edu/pedra,"

[109] F. Sadeghi and S. Levine, "Cad2rl: Real single-image flight without a single real image," *arXiv preprint arXiv:1611.04201*, 2016.

[110] E. Altman, *Constrained Markov decision processes*. Chapman and Hall/CRC Press, 1999, vol. 7.

[111] S. Paternain, L. F. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," *arXiv preprint arXiv:1910.13393*, 2019.

[112] D. Ding, K. Zhang, T. Basar, and M. Jovanovic, "Natural policy gradient primal-dual method for constrained markov decision processes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8378–8390, 2020.

[113] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "Optimality and approximation with policy gradient methods in markov decision processes," in *Conference on Learning Theory*, PMLR, 2020, pp. 64–66.

[114] M. Lanctot *et al.*, "Openspiel: A framework for reinforcement learning in games," *arXiv preprint arXiv:1908.09453*, 2019.

[115] O. Vinyals *et al.*, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[116] M. Riedmiller and T. Gabel, "On experiences in a complex and competitive gaming domain: Reinforcement learning meets RoboCup," in *2007 IEEE Symposium on Computational Intelligence and Games*, IEEE, 2007, pp. 17–23.

[117] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," *arXiv preprint arXiv:1610.03295*, 2016.

[118] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng, "Training gans with optimism," *arXiv preprint arXiv:1711.00141*, 2017.

[119] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, "Bridging the gap between value and policy based reinforcement learning," *Advances in neural information processing systems*, vol. 30, 2017.

[120] G. Neu, A. Jonsson, and V. Gómez, "A unified view of entropy-regularized markov decision processes," *arXiv preprint arXiv:1705.07798*, 2017.

[121] G. Lan, "Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes," *Mathematical programming*, pp. 1–48, 2022.

[122] S. Cen, Y. Wei, and Y. Chi, "Fast policy extragradient methods for competitive games with entropy regularization," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[123] J. Perolat, B. Scherrer, B. Piot, and O. Pietquin, "Approximate dynamic programming for two-player zero-sum markov games," in *International Conference on Machine Learning*, PMLR, 2015, pp. 1321–1329.

[124] Y. Bai and C. Jin, "Provable self-play algorithms for competitive reinforcement learning," in *International Conference on Machine Learning*, PMLR, 2020, pp. 551–560.

[125] Q. Xie, Y. Chen, Z. Wang, and Z. Yang, "Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium," in *Conference on Learning Theory*, PMLR, 2020, pp. 3674–3682.

[126] M. O. Sayin, F. Parise, and A. Ozdaglar, "Fictitious play in zero-sum stochastic games," *SIAM Journal on Control and Optimization*, vol. 60, no. 4, pp. 2095–2114, 2022.

[127] T. Chavdarova, G. Gidel, F. Fleuret, and S. Lacoste-Julien, "Reducing noise in gan training with variance reduced extragradient," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[128] A. Mokhtari, A. Ozdaglar, and S. Pattathil, "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 1497–1507.

[129] C. J. Li *et al.*, "On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 9793–9826.

[130] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo, "Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games," in *Conference on Learning Theory*, PMLR, 2021, pp. 4259–4299.

[131] Y. Zhao, Y. Tian, J. D. Lee, and S. S. Du, "Provably efficient policy gradient methods for two-player zero-sum markov games," *arXiv preprint arXiv:2102.08903*, 2021.

[132] Z. Chen, S. Ma, and Y. Zhou, "Sample efficient stochastic policy extragradient algorithm for zero-sum markov game," in *International Conference on Learning Representations*, 2021.

[133] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[134] D. Ying, Y. Ding, and J. Lavaei, "A dual approach to constrained markov decision processes with entropy regularization," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2022, pp. 1887–1909.

[135] R. D. McKelvey and T. R. Palfrey, "Quantal response equilibria for normal form games," *Games and economic behavior*, vol. 10, no. 1, pp. 6–38, 1995.

[136] T. Raghavan, "Completely mixed games and M-matrices," *Linear Algebra and its Applications*, vol. 21, no. 1, pp. 35–45, 1978.

[137] I. Kaplansky, "A contribution to von neumann's theory of games. ii," *Linear algebra and its applications*, vol. 226, pp. 371–373, 1995.

[138] P. Das, T. Parthasarathy, and G. Ravindran, "On completely mixed stochastic games," *arXiv preprint arXiv:1703.04619*, 2017.

[139] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[140] S. Mhanna, A. C. Chapman, and G. Verbic, "Component-based dual decomposition methods for the OPF problem," *Sustainable Energy, Grids and Networks*, vol. 16, pp. 91–110, 2018.

[141] B. He, H. Yang, and S. Wang, "Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities," *Journal of Optimization Theory and Applications*, vol. 106, no. 2, pp. 337–356, 2000.

[142] Z. Xu, M. Figueiredo, and T. Goldstein, "Adaptive ADMM with spectral penalty parameter selection," in *20th International Conference on Artificial Intelligence and Statistics*, PMLR, 2017, pp. 718–727.

[143] S. Mhanna, G. Verbic, and A. C. Chapman, "Adaptive ADMM for distributed AC optimal power flow," *IEEE Transactions on Power Systems*, vol. 34, no. 3, pp. 2025–2035, 2019.

[144] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for decision-making and control in power systems: Tutorial, review, and vision," *arXiv:2102.01168*, 2021.

[145] T. Chen *et al.*, "Learning to optimize: A primer and a benchmark," *Journal of Machine Learning Research*, vol. 23, no. 189, pp. 1–59, 2022.

[146] M. Andrychowicz *et al.*, "Learning to learn by gradient descent by gradient descent," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 3981–3989.

[147] D. Biagioni, P. Graf, X. Zhang, A. S. Zamzam, K. Baker, and J. King, "Learning-accelerated ADMM for distributed DC optimal power flow," *IEEE Control Systems Letters*, vol. 6, pp. 1–6, 2022.

[148] P. Graf *et al.*, "Distributed reinforcement learning with ADMM-RL," in *American Control Conference (ACC)*, 2019, pp. 4159–4166.

[149] X. Xie, J. Wu, G. Liu, Z. Zhong, and Z. Lin, "Differentiable linearized ADMM," in *International Conference on Machine Learning (ICML)*, PMLR, 2019, pp. 6902–6911.

[150] J. Ichnowski *et al.*, "Accelerating quadratic optimization with reinforcement learning," *Advances in Neural Information Processing Systems (NIPS)*, 2021.

[151] F. Li and Y. Du, "From AlphaGo to power system AI: What engineers can learn from solving the most complex board game," *IEEE Power and Energy Magazine*, vol. 16, no. 2, pp. 76–84, 2018.

[152] L. Duchesne, E. Karangelos, and L. Wehenkel, "Recent developments in machine learning for energy systems reliability management," *Proceedings of the IEEE*, vol. 108, no. 9, pp. 1656–1676, 2020.

[153] J. Giesen and S. Laue, "Distributed convex optimization with many convex constraints," *arXiv preprint arXiv:1610.02967*, 2016.

[154] Y. Tang and S. Agrawal, "Discretizing continuous action space for on-policy optimization," in *Proceedings of the aaai conference on artificial intelligence*, vol. 34, 2020, pp. 5981–5988.

[155] R. Zimmerman, C. Murillo-Sánchez, and R. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, Feb. 2011.

[156] S. Khodadadian, T. T. Doan, J. Romberg, and S. T. Maguluri, "Finite sample analysis of two-time-scale natural actor-critic algorithm," *IEEE Transactions on Automatic Control*, 2022.

[157] S. Kakade and J. Langford, "Approximately optimal approximate reinforcement learning," in *ICML*, vol. 2, 2002, pp. 267–274.

[158] P. Bernhard and A. Rapaport, "On a theorem of danskin with an application to a theorem of von neumann-sion," *Nonlinear Analysis: Theory, Methods & Applications*, vol. 24, no. 8, pp. 1163–1181, 1995.