Provably Efficient Reinforcement Learning: From Single-Agent MDPs to Markov Games

by

Shuang Qiu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2021

Doctoral Committee:

Professor Jieping Ye, Chair
Professor Satinder Singh Baveja
Professor Joyce Chai
Associate Professor Ambuj Tewari
Professor Ji Zhu

Shuang Qiu

qiush@umich.edu

ORCID iD: 0000-0002-9651-1061

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincerest gratitude to my advisor Prof. Jieping Ye for his support and guidance in my research. Prof. Ye has provided me a lot of great opportunities to collaborate with excellent researchers in the area of machine learning and he always encourages me to explore the fundamental problems in machine learning theory. His passion for work and research has significantly influenced me. I am so fortunate to be a member of Prof. Ye's lab, where I can always access great research resources and acquire cutting-edge research topics. This dissertation would not have been possible without Prof. Ye's help and support.

I am extremely grateful to all my fantastic collaborators during my Ph.D. study: Dr. Ming Lin, Dr. Ji Liu, Dr. Jie Wang, Dr. Zhaoran Wang, Dr. Xiaohan Wei, Zhuoran Yang, and all other collaborators. I would like to thank Dr. Ming Lin, Dr. Ji Liu, and Dr. Jie Wang for leading me into the area of optimization theory, training me on the mathematical foundations of machine learning, and spending a lot of time on detailed discussions with me. I would also like to thank Dr. Zhaoran Wang, Dr. Xiaohan Wei, and Zhuoran Yang for their guidance in the areas of statistical learning and reinforcement learning. They are always willing to discuss technical details with me and also share their research experience. They have inspired me greatly with their excellent research ideas and insightful thoughts on these topics.

I would like to express my gratitude to all my friends and my previous and current labmates for their valuable academic discussions and for the wonderful time we spent together. I would also like to thank all my dissertation committee members for reviewing my dissertation and providing valuable feedback and suggestions on research topics during my defense.

Last but not least, I would like to give my deepest thanks to my parents and all other family members. Their unconditional and endless support and love always give me courage and strength to overcome various difficulties and get through hard times.

TABLE OF CONTENTS

ACKNOWLEDGMENTS ii

ABSTRACT vi

CHAPTER

1 Introduction 1

1.1 Main Problems and Contributions 2
1.2 Thesis Organization 5

2 Background 6

2.1 Single-Agent Markov Decision Process 6
2.2 Two-Player Zero-Sum Markov Game 7

3 Constrained MDP with Adversarial Loss 10

3.1 Introduction 10
3.2 Problem Setup 13
3.3 Proposed Algorithm 16
3.3.1 Computing Optimistic Policies 17
3.3.2 Efficient Solver for Subproblem 19
3.4 Main Results 20
3.5 Theoretical Analysis 22
3.5.1 Proof of Regret Bound 22
3.5.2 Proof of Constraint Violation Bound 25
3.6 Conclusion 26
3.7 Proofs of Lemmas for Regret Bound 26
3.7.1 Proof of Lemma 3.11 26
3.7.2 Proof of Lemma 3.12 32
3.7.3 Proof of Lemma 3.13 38
3.7.4 Proof of Lemma 3.15 45
3.8 Proofs of Lemmas for Constraint Violation Bound 47
3.8.1 Proof of Lemma 3.16 47
3.8.2 Proof of Lemma 3.17 48

4 Policy Optimization for Zero-Sum Markov Games with Structured Transitions . . . 52

4.1 Introduction 52

4.2 Preliminaries . . . 55
4.2.1 Problem Setup . . . 55
4.2.2 Structured Transition Models . . . 57
4.3 Markov Game with Factored Independent Transition . . . 58
4.3.1 Main Results . . . 61
4.4 Markov Game with Single-Controller Transition . . . 62
4.4.1 Main Results . . . 65
4.5 Theoretical Analysis . . . 66
4.5.1 Proofs of Theorems 4.2 and 4.3 . . . 66
4.5.2 Proofs of Theorems 4.5 and 4.6 . . . 67
4.6 Conclusion . . . 68
4.7 Proofs for Markov Game with Factored Independent Transition . . . 69
4.7.1 Other Supporting Lemmas . . . 85
4.8 Proofs for Markov Game with Single-Controller Transition . . . 87
4.8.1 Other Supporting Lemmas . . . 108

5 Reward-Free RL with Kernel and Neural Function Approximations . . . 111

5.1 Introduction . . . 111
5.2 Preliminaries . . . 113
5.2.1 Problem Setup . . . 113
5.2.2 Reproducing Kernel Hilbert Space . . . 114
5.2.3 Overparameterized Neural Network . . . 115
5.3 Single-Agent MDP Setting . . . 116
5.3.1 Kernel Function Approximation . . . 116
5.3.2 Neural Function Approximation . . . 118
5.3.3 Main Results for Single-Agent MDP . . . 119
5.4 Markov Game Setting . . . 122
5.4.1 Kernel Function Approximation . . . 122
5.4.2 Neural Function Approximation . . . 123
5.4.3 Main Results for Markov Game . . . 124
5.5 Theoretical Analysis . . . 125
5.5.1 Proof Sketches of Theorem 5.3 and Theorem 5.5 . . . 125
5.5.2 Proof Sketches of Theorem 5.6 and Theorems 5.7 . . . 127
5.6 Conclusion . . . 128
5.7 Discussion of Function Space Complexity . . . 128
5.8 Proofs for Single-Agent MDP with Kernel Function Approximation . . . 132
5.8.1 Lemmas . . . 132
5.8.2 Proof of Theorem 5.3 . . . 146
5.9 Proofs for Single-Agent MDP with Neural Function Approximation . . . 148
5.9.1 Lemmas . . . 148
5.9.2 Proof of Theorem 5.5 . . . 164
5.10 Proofs for Markov Game with Kernel Function Approximation . . . 166
5.10.1 Lemmas . . . 166
5.10.2 Proof of Theorem 5.6 . . . 172
5.11 Proofs for Markov Game with Neural Function Approximation . . . 174

5.11.1 Lemmas 174
5.11.2 Proof of Theorem 5.7 178
5.12 Other Supporting Lemmas 180

6 Conclusion 182

6.1 Future Direction 183

BIBLIOGRAPHY 185

ABSTRACT

Reinforcement Learning (RL) has achieved tremendous empirical successes in real-world decision-making problems. Along with its great empirical achievements, recently, the question of *how to design efficient RL algorithms with provable theoretical guarantees* has attracted increasing attention. To answer the above question, this thesis proposes and analyzes novel RL algorithms for single-agent Markov Decision Processes (MDPs) and Markov games, where the Markov game is the multi-agent extension of single-agent MDPs. This thesis covers two paradigms in RL: *reward-based online RL* and *reward-free RL*. The reward-based online RL is a standard online learning framework for RL. In this framework, the agents keep interacting with the *unknown* environment and receiving rewards. Meanwhile, the agents continuously update the policies with the current information obtained from the environment to achieve their goals of learning. The reward-free RL is a different framework where the agents first aim to thoroughly explore the *unknown* environment *without* accessing any pre-specified rewards and then, given an *arbitrary* extrinsic reward function, the agents compute the target policy via a planning algorithm with data collected in the exploration phase.

Concretely, this thesis focuses on providing a theoretical analysis of three fundamental and challenging problems in RL: **(1)** *online learning for constrained MDPs*, **(2)** *policy optimization for Markov games*, **(3)** *reward-free RL with nonlinear function approximation*. The first two problems are studied in the scope of the reward-based online RL and the third one is an important problem under the reward-free RL setting. In these three directions, the main contributions of this thesis are summarized as follows. The first contribution is that this thesis proposes a provably efficient upper confidence primal-dual algorithm for the single-agent MDP online learning problem with *time-varying constraints*, where the transition model is *unknown* and the reward function is *adversarial*. This thesis further proves the upper bounds of the regret and the constraint violation for learning the constrained MDPs. As the second contribution, this thesis proposes new optimistic policy optimization algorithms for two-player zero-sum Markov games with *structured but unknown* transitions and theoretically analyzes both players' regret bounds, which generalizes the recent studies on policy optimization for single-agent MDPs in a stationary environment. The third contribution is that this thesis tackles the reward-free RL problem for both single-agent MDPs and two-player zero-sum Markov games under the context of function approximation, leveraging powerful nonlinear approximators: *kernel* and *neural* function approximators. Specifically, this thesis proposes to explore the unknown environment via an optimistic variant of the value-iteration algorithm incorporating kernel and neural function approximations and designs effective planning algorithms, which are theoretically justified to be able to generate the target policies when given an arbitrary extrinsic reward function.

CHAPTER 1

Introduction

In view of the tremendous successes of RL algorithms on real-world decision-making problems, the theoretical understanding of RL algorithms has been gaining increasing attention from researchers. Thus, how to design sample- (and computationally) efficient RL algorithms with provable theoretical guarantees becomes a core question in the recent studies of RL theory. In particular, a higher sample efficiency indicates fewer interactions with the environments for sampling data to achieve the desired learning accuracy. This thesis focuses on proposing and analyzing provably efficient algorithms for both single-agent and multi-agent RL problems. The analysis of single-agent RL algorithms is based on the single-agent MDP model, where an agent can interact with the environment following a certain policy and the environment returns the next state and the reward to the agent following a transition model and a reward function. Furthermore, this thesis goes beyond the single-agent scenario and further investigates the multi-agent RL setting. Specifically, this thesis studies the two-player zero-sum Markov game model, where the transition model and the reward function have a dependence on both players' actions and their state. Under such a setting, one player aims to learn a policy to maximize the expected cumulative rewards while the other player, in contrast, intends to minimize them. Therefore, the two-player zero-sum Markov game is a non-trivial extension of single-agent MDPs to a multi-agent scenario in a competitive and non-stationary environment, where the non-stationarity results from the potentially adversarial actions or policies of the two players.

For both single-agent MDPs and Markov games, this thesis studies two paradigms of RL: *reward-based online RL* and *reward-free RL*. In the reward-based online RL, the agents keep interacting with the environment to collect the data including the state-action trajectories and the rewards, and meanwhile, the agents continuously update their policies with the online data. To efficiently learn the target policies, it is necessary to design an online algorithm that can effectively exploit the collected information from the environment and at the same time, encourage the exploration of the states and actions of high uncertainty. In the framework of reward-based online RL, this thesis makes attempt to solve two fundamental problems: online learning for constrained MDPs and policy optimization for Markov games.

This thesis further investigates the reward-free RL [Jin et al., 2020a], a novel RL paradigm motivated by the following scenario. In real-world RL applications, the reward function is often designed by the learner based on the domain knowledge. The learner might have a set of reward functions to choose from or use an adaptive algorithm for reward design [Laud, 2004, Grzes, 2017]. In such a scenario, it is often desirable to collect an offline dataset that has a wider coverage of the trajectories associated with a set of reward functions and target policies. With such a benign offline dataset, for an arbitrary reward function, the agents have sufficient information to estimate the corresponding target policies. Thus, the reward-free RL is composed of two learning phases: *exploration phase* and *planning phase*. In the exploration phase, the agents aim to thoroughly explore the environment without accessing any pre-specified reward function in a principled manner. In the planning phase, when given an arbitrary extrinsic reward function, the planning algorithm generates the target policies by effectively making use of the collected offline data. Moreover, this thesis tackles the reward-free RL problem under the context of function approximation, leveraging powerful nonlinear function approximators. Then, within the framework of reward-free RL, this thesis studies the problem of the reward-free RL with nonlinear function approximations for single-agent MDPs and two-player zero-sum Markov games.

The following section elaborates the aforementioned three main problems and the associated contributions of this thesis.

1.1 Main Problems and Contributions

Online Learning for Constrained MDPs. Online learning for MDPs has been broadly studied in previous works which pay more attention to the unconstrained MDP model. This thesis considers online learning for single-agent MDPs with multiple constraints. Constrained MDPs play an important role in control and planning, which aim at maximizing a reward or minimizing a penalty metric over the set of all available policies subject to constraints. The constraints can enforce the fairness or safety of the policies so that over time the behaviors of the learned policy are under control. Previous works (e.g., Wei et al. [2018], Zheng and Ratliff [2020]) solve this problem under the restrictive assumption that the transition model of the MDP is known a priori. This thesis considers a more realistic and challenging setting that the transition model is unknown to the agent, the loss function can vary arbitrarily across the episodes, and the constraints are stochastically time-varying.

In this thesis, a new upper confidence primal-dual algorithm is proposed for learning constrained MDPs. With the trajectories of states and actions that the agent collects by interacting with the environment, the proposed algorithm estimates the unknown transition model with maintaining a confidence set inspired by the idea of Upper Confidence Bound (UCB), which is also shown effective to achieve tight regret bounds for learning unconstrained MDPs. Then, it incorporates the confidence set into the online primal-dual type method for learning the policies. The proposed algorithm is proved to achieve $\widetilde{\mathcal{O}}(\sqrt{K})$[1] upper bounds for the regret and the constraint violation simultaneously, which demonstrates the power of *"optimism in the face of uncertainty"* [Auer et al., 2002, Bubeck and Cesa-Bianchi, 2012]. Here K is the number of episodes. Moreover, the regret bound nearly matches the lower bound of the regret for learning MDPs. The analysis incorporates a new high-probability drift analysis of Lagrange multiplier processes into the regret and constraint violation proofs for the proposed upper confidence algorithm. The study of online learning for constrained MDPs in this thesis is based on joint work with Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang [Qiu et al., 2020].

Policy Optimization for Zero-Sum Markov Games. While single-agent Policy Optimization (PO) in a fixed environment has attracted a lot of research attention recently in the reinforcement learning community, much less is known theoretically when there are multiple agents playing in a potentially competitive environment. To study such a problem, this thesis considers learning the two-player zero-sum Markov games. In most of the recent works, the proposed algorithms for learning zero-sum Markov games are typically value-based methods (e.g, Bai and Jin [2020]) that can achieve tight $\widetilde{\mathcal{O}}(\sqrt{K})$ regrets and they assume there is a central agent available for solving certain subproblems each step, which introduces extra computational costs. Here, K denotes the number of episodes. As opposed to the value-based methods, the PO algorithms aim to directly update the policies of agents via only executing a mirror descent or ascent step separately on each agent with high computational efficiency. Although there has been great progress on understanding single-agent PO algorithms, directly extending single-agent PO methods to the multi-agent setting encounters the main challenge of non-stationary environments.

This thesis takes steps forward by proposing and analyzing new provable optimistic PO algorithms for two-player zero-sum Markov games with structured but unknown transitions. In particular, two classes of transition structures are considered here: *factored independent transition* and *single-controller transition*. The proposed algorithms feature a combination of UCB-type optimism and policy optimization updating rules adapted to the structured transitions in a multi-agent non-stationary environment. To handle the non-stationarity resulting from the opponent's varying state, both players under the factored independent transition setting and Player 2 under the single-controller setting demand to estimate the opponent's state reaching probability. For both transition structures, this thesis provides $\widetilde{\mathcal{O}}(\sqrt{K})$ regret bounds after K episodes. The regret of each player is measured against a potentially adversarial opponent who can choose a single best policy in hindsight if observing the full policy sequence. The $\widetilde{\mathcal{O}}(\sqrt{K})$ regret bounds in this thesis also match the

[1] In this thesis, we use $\widetilde{\mathcal{O}}$ to hide the logarithmic dependence.

regrets of the value-based methods when translating their results in terms of the regret definition in this thesis. If both players adopt the proposed algorithms, the overall optimality gap is upper bounded by $\widetilde{\mathcal{O}}(\sqrt{K})$. Moreover, this thesis proposes novel value difference decomposition by taking the transition structures and the state reaching probability estimation error into consideration. The study of the second problem, i.e., policy optimization for zero-sum Markov games, is based on the joint work with Xiaohan Wei, Jieping Ye, Zhaoran Wang, and Zhuoran Yang [Qiu et al., 2021a].

Reward-Free RL with Kernel and Neural Function Approximations. The framework of the reward-free RL consists of an exploration phase and a planning phase, which needs to efficiently explore the underlying environment without accessing any pre-specified reward function and to effectively generate the target policies with the collected dataset when given an arbitrary reward function. On the other hand, when the state and action spaces are large, it is a common practice to combine the RL algorithms with the idea of function approximation, especially the powerful nonlinear function approximators such as neural networks. Thus, this thesis considers the reward-free RL with two classic nonlinear function approximators, i.e., kernel and neural function approximators. This further motivates us to design provably exploration and planning algorithms that can incorporate kernel and neural function approximations into the framework of the reward-free RL.

Recently, many works focus on designing provably sample-efficient reward-free RL algorithms, including the tabular case [Jin et al., 2020a, Kaufmann et al., 2020, Ménard et al., 2020, Zhang et al., 2020] and the linear function approximation case [Zanette et al., 2020b, Wang et al., 2020a] for the single-agent MDP, which can achieve $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexities for obtaining an ε-suboptimal policy. However, reward-free RL combined with nonlinear function approximators remains not fully explored. On the other hand, reward-free RL algorithms for the multi-player Markov games [Bai and Jin, 2020, Liu et al., 2020] in the tabular case have been studied recently, which is shown to achieve an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity for obtaining an ε-approximate Nash equilibrium. Though, there is still a lack of works theoretically studying multi-agent scenarios with function approximation.

This thesis first proposes sample- and computationally efficient reward-free RL algorithms with kernel and neural function approximations for single-agent MDPs. The proposed exploration algorithm is an optimistic variant of the least-square value iteration algorithm incorporating kernel and neural function approximators inspired by the idea of UCB. Further with the planning phase, which is a single-episode optimistic value iteration algorithm, the proposed method achieves an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity to generate an ε-suboptimal policy for an *arbitrary* extrinsic reward function. Moreover, this thesis extends the proposed method from the single-agent setting to the two-player zero-sum Markov game setting, which can achieve an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity to generate an ε-approximate Nash equilibrium. Particularly, in the planning phase for the Markov game setting, the proposed algorithm only involves finding the Nash equilibrium of matrix games formed by Q-function that can be solved *efficiently*, which is of independent interest. The sample complexities of our methods match the $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ results in existing works for tabular or linear function approximation settings. To the best of our knowledge, we establish the first provably efficient reward-free RL algorithms with kernel and neural function approximators for both single-agent and multi-agent settings. The study of this problem, i.e., reward-free RL with kernel and neural function approximations, is based on the joint work with Jieping Ye, Zhaoran Wang, and Zhuoran Yang [Qiu et al., 2021b].

1.2 Thesis Organization

The rest of this thesis is organized as follows. Chapter 2 provides the fundamental backgrounds of the single-agent MDP and the two-player zero-sum Markov game, which are two basic models that the three main problems in this thesis are based on. Chapter 3 presents the first contribution of this thesis, including the proposed learning algorithm and the theoretical analysis of the regret and the constraint violation with detailed proofs for the online constrained MDP learning problem. The second problem and the related contribution are presented in Chapter 4, which proposes the policy optimization algorithms for the zero-sum Markov games with structured transitions and provides detailed regret analysis for the proposed algorithms. Chapter 5 investigates the third problem, i.e., reward-free RL with kernel and neural function approximations for MDP and Markov games, and presents the proposed algorithms along with proving their sample complexities. The last chapter, Chapter 6, concludes this thesis by summarizing the main problems and contributions and then suggests several potential future research directions.

CHAPTER 2

Background

In this thesis, our works study single-agent MDPs and two-player zero-sum Markov games in an *episodic* (finite-horizon) setting. In this chapter, we introduce the episodic MDPs and episodic two-player zero-sum Markov Games in the following two sections.

Notation. Throughout this thesis, we let $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ be ℓ_1, ℓ_2, and ℓ_∞-norm for a vector. Let $\|\cdot\|_2$ be the spectral norm for a matrix. We define $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top A\mathbf{x}}$ for a vector $\mathbf{x}$ and a matrix A if $\mathbf{x}^\top A\mathbf{x} \geq 0$. We let $\|f\|_\infty = \sup_{x\in X} |f(x)|$ for any function f defined on the set X. We define $[n] := \{1, 2, \ldots, n\}$. For any vectors $\mathbf{x}, \mathbf{y}$, the inner product of $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\langle \mathbf{x}, \mathbf{y}\rangle$, which is also written as $\mathbf{x}^\top\mathbf{y}$. Given $a > 0$, we define the operation $\min\{x, a\}^+ := \min\{\max\{x, 0\}, a\}$ for any x.

2.1 Single-Agent Markov Decision Process

An episodic single-agent MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the action space of the agent, H is the length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is the transition model with $\mathbb{P}_h(s'|s, a)$ denoting the transition probability at the h-th step from the state $s \in \mathcal{S}$ to the state $s' \in \mathcal{S}$ when the agent takes action $a \in \mathcal{A}$, and $r = \{r_h\}_{h=1}^H$ with $r_h : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}$ denoting the reward function at the h-step. Following a certain policy, an agent interacts with the environment. The policy of an agent is a collection of probability distributions $\pi = \{\pi_h\}_{h=1}^H$ where $\pi_h(a|s)$ is the probability of taking action $a \in \mathcal{A}$ at the state $s \in \mathcal{S}$ at the h-th step.

Value Function. For a specific policy $\{\pi_h\}_{h=1}^H$ and reward function $\{r_h\}_{h=1}^H$, under the transition model $\{\mathbb{P}_h\}_{h=1}^H$, we define the associated value function $V_h^\pi(s, r) : \mathcal{S} \mapsto \mathbb{R}$ at the h-th step as follows

$$V_h^\pi(s, r) := \mathbb{E}\left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \,\middle|\, s_h = s, \pi, \mathbb{P}\right], \quad \forall s \in \mathcal{S}.$$

The corresponding action-value function (Q-function) $Q_h^\pi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is then defined as follows

$$Q_h^\pi(s,a,r) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \,\middle|\, s_h = s, a_h = a, \pi, \mathbb{P}\right], \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Then, we have the Bellman equation as follows

$$V_h^\pi(s,r) = \langle Q_h^\pi(s,\cdot,r), \pi_h(\cdot|s)\rangle_{\mathcal{A}}, \quad \forall s \in \mathcal{S}, \tag{2.1}$$

$$Q_h^\pi(s,a,r) = r_h(s,a) + \langle \mathbb{P}_h(\cdot|s,a), V_{h+1}^\pi(\cdot,r)\rangle_{\mathcal{S}}, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \tag{2.2}$$

where we let $\langle\cdot,\cdot\rangle_{\mathcal{S}}$, $\langle\cdot,\cdot\rangle_{\mathcal{A}}$ denote the inner product over the spaces $\mathcal{S}$, $\mathcal{A}$. The above Bellman equation holds for all $h \in [H]$ with setting $V_{H+1}^\pi(s) = 0, \forall s \in \mathcal{S}$.

Optimal Policy and ε-Suboptimal Policy. Without loss of generality, in this thesis, we assume the agent starts from a fixed state s_1 at $h = 1$. We also make the same assumption for the Markov game setting in the next section. In an MDP learning problem, an agent aims to learn a policy to maximize the value function $V_1^\pi(s_1, r)$. Then, we denote π_r^* as the *optimal policy* w.r.t. r such that π_r^* maximizes $V_1^\pi(s_1, r)$, i.e.,

$$\pi_r^* := \operatorname*{argmax}_\pi V_1^\pi(s_1, r).$$

Then, we define $Q_h^*(s,a,r) := Q_h^{\pi_r^*}(s,a,r)$ as well as $V_h^*(s,r) := V_h^{\pi_r^*}(s,r)$. We say $\widetilde{\pi}$ is an ε-*suboptimal policy* if it satisfies

$$V_1^*(s_1, r) - V_1^{\widetilde{\pi}}(s_1, r) \leq \varepsilon.$$

To simplify the notations, for the rest of this thesis, we rewrite $\langle \mathbb{P}_h(\cdot|s,a), V_{h+1}(\cdot,r)\rangle_{\mathcal{S}} = \mathbb{P}_h V_{h+1}(s,a,r)$ for any transition probability $\mathbb{P}_h$ and value function $V(\cdot,r)$. In addition, $V_h^\pi(s,r)$ and $Q_h^\pi(s,a,r)$ will be simplified as $V_h^\pi(s)$ and $Q_h^\pi(s,a)$ when their dependence on the reward function are clear from the context in a chapter.

2.2 Two-Player Zero-Sum Markov Game

The two-player zero-sum Markov game is an extension of the single-agent MDP to a multi-agent competitive scenario. Specifically, an episodic two-player zero-sum Markov game is characterized by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \mathbb{P}, r)$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ and $\mathcal{B}$ are the action spaces for the two players, H is the length of each episode, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is the transition model with $\mathbb{P}_h(s'|s,a,b)$ denoting the transition probability at the h-th step from the state s to the state s'

when Player 1 takes action $a \in \mathcal{A}$ and Player 2 takes action $b \in \mathcal{B}$, and $r = \{r_h\}_{h=1}^H$ with $r_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}$ denoting the reward function at the h-step. The policy of Player 1 is a collection of probability distributions $\pi = \{\pi_h\}_{h=1}^H$ with $\pi(\cdot|s)$ being the probability of taking action $a \in \mathcal{A}$ at the state $s \in \mathcal{S}$ at the h-th step. Analogously, the policy of Player 2 is a collection of probability distributions $\nu = \{\nu_h\}_{h=1}^H$ with $\nu(b|s)$ being the probability of taking action $b \in \mathcal{B}$ at the state $s \in \mathcal{S}$ at the h-th step. As we can see from the above definitions, the reward function and the transition model for the two-player zero-sum Markov Game depends on both players' actions (a, b) and their state s, which introduce challenges of *non-stationarity*.

Value Function. For a specific policy π and ν and reward function $\{r_h\}_{h=1}^H$, under the transition model $\{\mathbb{P}_h\}_{h=1}^H$, we define the value function $V_h^{\pi,\nu}(s, r) : \mathcal{S} \mapsto \mathbb{R}$ at the h-th step as follows

$$V_h^{\pi,\nu}(s, r) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'}) \,\middle|\, s_h = s, \pi, \nu, \mathbb{P}\right], \quad \forall s \in \mathcal{S}.$$

We further define the Q-function $Q_h^{\pi,\nu} : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto \mathbb{R}$ for all $\forall(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ as follows

$$Q_h^{\pi,\nu}(s, a, b, r) := \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}, b_{h'}) \,|\, (s_h, a_h, b_h) = (s, a, b), \pi, \nu, \mathbb{P}\right].$$

Thus, we have the Bellman equation for all $h \in [H]$ as follows

$$V_h^{\pi,\nu}(s, r) = \mathbb{E}_{a\sim\pi_h(\cdot|s), b\sim\nu_h(\cdot|s)}[Q_h^{\pi,\nu}(s, a, b, r)], \quad \forall s \in \mathcal{S}, \tag{2.3}$$

$$Q_h^{\pi,\nu}(s, a, b, r) = r_h(s, a, b) + \mathbb{P}_h V_{h+1}^{\pi,\nu}(s, a, b, r), \quad \forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}, \tag{2.4}$$

where, for simplicity, we also let $\mathbb{P}_h V_{h+1}^{\pi,\nu}(s, a, b, r) = \langle \mathbb{P}_h(\cdot|s, a, b), V_{h+1}^{\pi,\nu}(\cdot, r)\rangle_{\mathcal{S}}$.

Nash Equilibrium and ε-Approximate Nash Equilibrium. In the zero-sum Markov Game learning problem, Player 1 aims to learn a policy to maximize the value function and Player 2 tries to learn a policy to minimize the value function. Thus, each individual player faces a *competitive* environment which is affected by the opponent's potentially adversarial actions and policies. Under this setting, we define the *Nash equilibrium* (NE) $(\pi^\dagger, \nu^\dagger)$ as a solution to the following minimax problem

$$\max_{\pi} \min_{\nu} V_1^{\pi,\nu}(s_1),$$

such that we have

$$V_1^{\pi^\dagger,\nu^\dagger}(s_1, r) = \max_{\pi} \min_{\nu} V_1^{\pi,\nu}(s_1) = \min_{\nu} \max_{\pi} V_1^{\pi,\nu}(s_1).$$

We let $V_h^\dagger(s,r) := V_h^{\pi^\dagger,\nu^\dagger}(s,r)$ and $Q_h^\dagger(s,a,b,r) := Q_h^{\pi^\dagger,\nu^\dagger}(s,a,b,r)$ denote the value function and Q-function under the NE $(\pi^\dagger,\nu^\dagger)$ at h-th step. We further define the *best response* $\mathrm{br}(\pi)$ for Player 1 with the policy π and the best response $\mathrm{br}(\nu)$ for Player 2 with the policy ν as

$$\mathrm{br}(\pi) := \operatorname*{argmin}_{\nu} V_1^{\pi,\nu}(s_1,r), \quad \mathrm{br}(\nu) := \operatorname*{argmax}_{\pi} V_1^{\pi,\nu}(s_1,r).$$

Then, one can see that $\widetilde{\nu} = \mathrm{br}(\widetilde{\pi})$ and $\widetilde{\pi} = \mathrm{br}(\widetilde{\nu})$ for the NE $(\widetilde{\pi},\widetilde{\nu})$. Moreover, We say $(\widetilde{\pi},\widetilde{\nu})$ is an ε*-approximate NE* if it satisfies

$$V_1^{\mathrm{br}(\widetilde{\nu}),\widetilde{\nu}}(s_1,r) - V_1^{\widetilde{\pi},\mathrm{br}(\widetilde{\pi})}(s_1,r) \leq \varepsilon,$$

where the weak duality $V_1^{\mathrm{br}(\widetilde{\nu}),\widetilde{\nu}}(s_1,r) \geq V_1^\dagger(s_1,r) \geq V_1^{\widetilde{\pi},\mathrm{br}(\widetilde{\pi})}(s_1,r)$ always holds. Similarly, $V_h^\pi(s,r)$ and $Q_h^\pi(s,a,b,r)$ will be simplified as $V_h^\pi(s)$ and $Q_h^\pi(s,a,b)$ when their dependence on the reward function are clear from the context in a chapter.

Remark 2.1. For MDPs, if a policy π_h is deterministic, i.e., for a state $s \in \mathcal{S}$, there always exists an action $a \in \mathcal{A}$ such that $\pi_h(a|s) = 1$, then for ease of analysis, we slightly abuse the notion by letting $\pi_h : \mathcal{S} \mapsto \mathcal{A}$ such that $\pi_h(s)$ is the deterministic action which an agent will take at the state s at the step h. A similar notation is also defined for the Markov game setting. For the two player Markov game, if the policy π_h for Player 1 and the policy ν_h for Player 2 are deterministic, then we slightly abuse their notions by letting $\pi_h : \mathcal{S} \mapsto \mathcal{A}$ and $\nu_h : \mathcal{S} \mapsto \mathcal{B}$ such that $\pi_h(s)$ and $\nu_h(s)$ are the deterministic action which the players will take at the state s at the step h.

CHAPTER 3

Constrained MDP with Adversarial Loss

3.1 Introduction

Constrained Markov Decision Processes (CMDPs) play an important role in control and planning. It aims at maximizing a reward or minimizing a penalty metric over the set of all available policies subject to constraints on other relevant metrics. The constraints aim at enforcing the fairness or safety of the policies so that over time the behaviors of the chosen policy are under control. For example, in an edge cloud serving network [Urgaonkar et al., 2015, Wang et al., 2015], one would like to minimize the average cost of serving the moving targets subject to a constraint on the average serving delay. In an autonomous vehicle control problem [Le et al., 2019], one might be interested in minimizing the driving time subject to certain fuel efficiency or driving safety constraints.

Classical treatment of CMDPs dates back to Fox [1966], Altman [1999] reformulating the problem into a linear program via stationary state-action occupancy measures. However, to formulate such a linear program, one requires the full knowledge of the transition model, reward, and constraint functions, and also assumes them to be fixed. Leveraging the episodic structure of a class of MDPs, Neely [2012] develops online renewal optimization which potentially allows the loss and constraint functions to be stochastically varying and unknown, while still relying on the transition model to solve the subproblem within the episode. More recently, policy-search type algorithms have received much attention, attaining state-of-art performance in various tasks. While most of the algorithms focus on unconstrained RL problems, there are efforts to develop policy-based methods in CMDPs where constraints are known with limited theoretical guarantees. The work Chow et al. [2017] develops a primal-dual type algorithm which is shown to converge to some constraint satisfying policy. The work Achiam et al. [2017] develops a trust-region type algorithm, which requires solving an optimization problem with both trust-region and safety constraints during each update. Generalizing ideas from the fitted-Q iteration, Le et al. [2019] develops a batch offline primal-dual algorithm which guarantees only the time average primal-dual gap converges.

The goal of this chapter is to efficiently solve constrained episodic MDPs with more generality where not only transition models are unknown, but also the loss and constraint functions can change online. In particular, the losses can be arbitrarily time-varying and adversarial. Let K be the number of episodes and T the number of steps[1]. When assuming the transition model is known, Even-Dar et al. [2009] achieves $\widetilde{\mathcal{O}}(\varrho^2\sqrt{T})$ regret with ϱ being the mixing time of MDPs, and the work Yu et al. [2009] achieves $\widetilde{\mathcal{O}}(T^{2/3})$ regret. These two papers consider a continuous setting (non-episodic setting) that is different to the episodic setting that we consider in this chapter. The work Zimin and Neu [2013] further studies the episodic MDP and achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ regret. For the constrained case with known transitions, the work Wei et al. [2018] achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ regret and constraint violations, and the work Zheng and Ratliff [2020] attains $\widetilde{\mathcal{O}}(T^{3/4})$ for the non-episodic setting.

There are several concurrent works also focusing on CMDPs with unknown transition models. The work Efroni et al. [2020a] studies episodic tabular MDPs with unknown but fixed reward and constraint functions. Leveraging Upper Confidence Bound (UCB) on the reward, constraints, and transitions, they obtain an $\mathcal{O}(\sqrt{K})$ regret and constraint violation via linear program as well as primal-dual optimization. In another work, Ding et al. [2021] studies the constrained episodic MDPs with a linear structure and adversarial losses via a primal-dual-type policy optimization algorithm, achieving $\widetilde{\mathcal{O}}(\sqrt{K})$ regret and constraint violation. While their scenario is more general than ours, their results' dependence on the sizes of state and action spaces and the length of the episode is worse when applied to the tabular case. Both of these two works rely on Slater condition which is also more restrictive than that of the method in this chapter.

On the other hand, for unconstrained online MDPs, the idea of UCB is shown to be effective and helps to achieve tight regret bounds without knowing the transition model, e.g., Jaksch et al. [2010], Azar et al. [2017], Rosenberg and Mansour [2019a,b], Jin et al. [2019]. The main idea here is to sequentially refine a confidence set of the transition model and choose a model in the interval which performs the best in optimizing the current value.

The main contribution of this chapter is to show that incorporating the confidence set of the transition model into primal-dual type approaches can achieve $\widetilde{\mathcal{O}}(H|\mathbf{S}|\sqrt{|\mathbf{A}|K})$ regret and constraint violation simultaneously in online CMDPs when the transition model is unknown, the loss function is adversarial, and the constraints are stochastic. Here $|\mathbf{S}|$ is the state space size with $\mathbf{S}$ denoting the state space as defined later. We also let $|\mathbf{A}|$ be the size of the action spaces and H be the length of an episode. This result nearly matches the lower bound $\Omega(\sqrt{H|\mathbf{S}||\mathbf{A}|K})$ for the regret [Jaksch et al., 2010] up to an $\mathcal{O}(\sqrt{H|\mathbf{S}|})$ factor. Under the hood is a new Lagrange multiplier

[1] In the non-episodic setting, T denotes the total number of steps, which is different from the aforementioned K for the episodic setting. However, we can analogously compute the total number of steps as $T = KH$ in the episodic setting, where H is the episode length. Thus, T and K are comparable since in the episodic setting, only an extra constant factor H is involved in T.

condition together with a new drift analysis on the Lagrange multipliers leading to low constraint violation. Our setup is challenging compared to classical constrained optimization in particular due to **(1)** the unknown loss and constraint functions from the online setup; **(2)** the time-varying decision sets resulting from moving confidence set estimation. The decision sets can potentially be much larger than or even inconsistent with the true decision set knowing the model, resulting in a potentially large constraint violation. The main idea is to utilize a Lagrange multiplier condition as well as a confidence set of the model to construct a probabilistic bound on an online dual multiplier. We then explicitly take into account the laziness nature of the confidence set estimation in our algorithm to argue that the bound on the dual multiplier gives the $\widetilde{\mathcal{O}}(\sqrt{K})$ bound on constraint violation.

Related Work. In this chapter, we are interested in a class of online MDP problems where the loss functions are arbitrarily changing, or adversarial. With a known transition model, adversarial losses, and full-information feedbacks (as opposed to bandit feedbacks), Even-Dar et al. [2009] achieves $\widetilde{\mathcal{O}}(\varrho^2\sqrt{T})$ regret with ϱ being the mixing time of MDPs, and the work Yu et al. [2009] achieves $\widetilde{\mathcal{O}}(T^{2/3})$ regret, which consider a different non-episodic setting. The work Zimin and Neu [2013] further studies the episodic MDP and achieves an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret.

In contrast to the aforementioned works, a more challenging setting is that the transition model is not known a priori. Under such a setting, there are several works studying the online episodic MDP problems with adversarial losses and full-information feedbacks. Neu et al. [2012] obtains $\widetilde{\mathcal{O}}(H|\mathbf{S}||\mathbf{A}|\sqrt{K})$ regret by proposing a Follow the Perturbed Optimistic Policy (FPOP) algorithm. The recent work Rosenberg and Mansour [2019a] improves the regret to $\widetilde{\mathcal{O}}(H|\mathbf{S}|\sqrt{|\mathbf{A}|K})$ by proposing an online upper confidence mirror descent algorithm. This regret bound nearly matches the lower bound $\Omega(\sqrt{H|\mathbf{S}||\mathbf{A}|K})$ [Jaksch et al., 2010] up to $\mathcal{O}(\sqrt{H|\mathbf{S}|})$ and some logarithm factors. This chapter is along this line of research, and further considers the setup that there exist stochastic constraints observed at each episode during the learning process.

Besides, a number of papers also investigate online episodic MDPs with bandit feedbacks. Assuming the transition model is known and the losses are adversarial, Neu et al. [2010] achieves $\widetilde{\mathcal{O}}(\sqrt{K}/\beta)$ regret, where $\beta > 0$ is the probability with which all states are reachable under all policies. Under the same setting, Neu et al. [2010] achieves $\widetilde{\mathcal{O}}(K^{2/3})$ regret without the dependence on β, and Zimin and Neu [2013] obtains $\widetilde{\mathcal{O}}(\sqrt{K})$ regret. Furthermore, with assuming the transition model is not known and the losses are adversarial, Rosenberg and Mansour [2019b] obtains $\widetilde{\mathcal{O}}(K^{3/4})$ regret and also $\widetilde{\mathcal{O}}(\sqrt{K}/\beta)$. Jin et al. [2019] further achieves $\widetilde{\mathcal{O}}(\sqrt{K})$ regret without β under the same setting.

On the other hand, instead of adversarial losses, extensive works have studied the setting where the feedbacks of the losses are stochastic and have fixed expectations, e.g., Jaksch et al. [2010], Azar et al. [2017], Ouyang et al. [2017], Jin et al. [2018], Fruit et al. [2018], Wei et al. [2020], Zhang and Ji [2019], Dong et al. [2019]. With assuming that the transition model is known, Zheng and Ratliff [2020] studies online CMDPs under the non-episodic setting and attains an $\widetilde{\mathcal{O}}(T^{3/4})$ regret in T steps which is suboptimal in terms of T. The concurrent work Efroni et al. [2020a] studies episodic MDPs with unknown transitions and stochastic bandits feedbacks of the losses and the constraints, and obtains an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret and constraint violation.

In addition to the aforementioned papers, there is also a line of policy-search type works, focusing on solving online MDP problems via directly optimizing policies. Along this direction, there have been great numbers of works studying unconstrained MDPs without knowing transition models, e.g., Williams [1992], Baxter and Bartlett [2000], Konda and Tsitsiklis [2000], Kakade [2002], Schulman et al. [2015], Lillicrap et al. [2015], Schulman et al. [2017], Sutton and Barto [2018], Fazel et al. [2018], Abbasi-Yadkori et al. [2019a,b], Bhandari and Russo [2019], Cai et al. [2019], Wang et al. [2019a], Liu et al. [2019], Agarwal et al. [2019], Efroni et al. [2020b]. Efforts have also been made in several works [Chow et al., 2017, Achiam et al., 2017, Le et al., 2019] to investigate CMDP problems via policy-based methods, but with known transition models. In another concurrent work, assuming the transition model is unknown, Ding et al. [2021] studies CMDPs with linear function approximation and proposes a primal-dual policy optimization algorithm.

3.2 Problem Setup

To study the episodic MDP learning problem, we consider a Stochastic Shortest Path (SSP) model [Neu et al., 2010, 2012]. This chapter is build upon the SSP model because the proposed method in this chapter has a close connection with a line of the recent research [Rosenberg and Mansour, 2019a,b, Jin et al., 2019] studying the MDP problem based on SSP. In this chapter, we adopt the definition of SSP presented in Rosenberg and Mansour [2019a]. The connection between the SSP and the epsodic MDP defined in Section 2.1 of Chapter 2 is then discussed in Remark 3.2. In the loop-free SSP, we have a finite state space $\mathbf{S}$ and a finite action space $\mathbf{A}$ at each state over a finite horizon of K episodes. Each episode starts with a fixed initial state s_0 and ends with a terminal state s_H. The transition probability is $P : \mathbf{S}\times\mathbf{S}\times\mathbf{A} \mapsto [0,1]$, where $P(s'|s,a)$ gives the probability of transition from s to s' under an action a. This underlying transition model P is assumed to be *unknown*. The state space is divided into layers with a loop-free structure, i.e., $\mathbf{S} := \mathbf{S}_0 \cup \mathbf{S}_1 \cup \cdots \cup \mathbf{S}_H$ with a singleton initial layer $\mathbf{S}_0 = \{s_0\}$ and terminal layer $X_H = \{s_H\}$. Furthermore, we have $\mathbf{S}_h \cap \mathbf{S}_{h'} = \varnothing$ for $h \neq h'$, and transitions are only allowed between consecutive layers, which is $P(s'|s,a) > 0$ only if $s' \in \mathbf{S}_{h+1}$, $s \in \mathbf{S}_h$, and $a \in \mathbf{A}$, $\forall h \in \{0,1,\ldots,H-1\}$. Such an assumption enforces that each path from the initial state to the terminal state takes a fixed length H. This is not an excessively restrictive assumption as any loop-free MDP with bounded varying path lengths can be transformed into one with a fixed path length (see György et al. [2007] for details).

The loss function for each episode is $f^k : \mathbf{S} \times \mathbf{A} \times \mathbf{S} \mapsto \mathbb{R}$, where $f^k(s, a, s')$ denotes the loss received at episode k for any $s \in \mathbf{S}_h,\ s' \in \mathbf{S}_{h+1}$, and $a \in \mathbf{A}$, $\forall h \in \{0, 1, 2, \ldots, H-1\}$. We assume f_t can be *arbitrarily varying* with potentially no fixed probability distribution. There are I stochastic constraint (or budget consumption) functions: $g_i^k : \mathbf{S} \times \mathbf{A} \times \mathbf{S} \mapsto \mathbb{R}$, $\forall i \in [I]$, where $g_i^k(s, a, s')$ denotes the price to pay at episode k for any (s, a, s'). Each stochastic function g_i^k at episode k is sampled according to a random variable $\xi_i^k \sim \mathcal{D}_i$, namely $g_i^k(s, a, s') = g_i(s, a, s'; \xi_i^k)$. Then, we define $g_i(s, a, s') := \mathbb{E}[g_i^k(s, a, s')] = \mathbb{E}[g_i(s, a, s'; \xi_i^k)]$ where the expectation is taken over the randomness of $\xi_i^k \sim \mathcal{D}_i$. For abbreviation, we denote $g_i = \mathbb{E}[g_i^k]$. In addition, the functions f^k and g_i^k, $\forall i \in [I]$, are mutually independent and independent of the Markov transition. Both the loss functions and the budget consumption functions are revealed at the end of each episode.

Remark 3.1. It might be tempting to consider the more general scenario that both losses and constraints are arbitrarily time-varying. For such a setting, however, there exist counterexamples [Mannor et al., 2009] in the arguably simpler constrained online learning scenario that no algorithm can achieve sublinear regret and constraint violation simultaneously. Therefore, we seek to put extra assumptions on the problem so that obtaining sublinear regret and constraint violation is feasible, one of which is to assert constraints to be stochastic instead of arbitrarily varying.

A policy π is the conditional probability $\pi(a|s)$ of choosing an action $a \in \mathbf{A}$ at any given state $s \in \mathbf{S}$. At the k-episode, for any policy π, letting $(s_h, a_h, s_{h+1}) \in \mathbf{S}_h \times \mathbf{A} \times \mathbf{S}_{h+1}$ denote a random tuple following the transition model P and the policy π, the corresponding expected loss and the budget costs are written as

$$\mathbb{E}\left[\sum_{h=0}^{H-1} f^k(s_h, a_h, s_{h+1}) \,\middle|\, \pi, P\right], \text{ and } \quad \mathbb{E}\left[\sum_{h=0}^{H-1} g_i^k(s_h, a_h, s_{h+1}) \,\middle|\, \pi, P\right], \quad i \in [I], \tag{3.1}$$

where the expectations are taken w.r.t. the randomness of the tuples (s_h, a_h, s_{h+1}).

Remark 3.2. The above definition for the SSP problem has a close connection with the episodic MDP $(\mathcal{S}, \mathbf{A}, H, \mathbb{P}, r)$ which is defined in Section 2.1 of Chapter 2. To show how such a relation exists, we can let $\mathbf{A} = \mathcal{A}$ and $\mathbf{S}_h = \mathcal{S}$ for any $h \in \{0, \ldots, H-1\}$ with a fixed starting state and relaxing the restriction of $\mathbf{S}_h \cap \mathbf{S}_{h'} = \varnothing$ for $h \neq h'$. Note that the notations of P, π, f^k, g_i^k do not need to depend on the layer (or step) h because of this restriction. We further let $\mathbb{P}_{h+1}(s'|s, a) = P(s'|s, a)$ where $s' \in \mathbf{S}_{h+1}, s \in \mathbf{S}_h$ for all $h \in \{0, \ldots, H-1\}$ and $s', s \in \mathcal{S}$. Based on such the above conversion, shifting the index h by 1 and removing the dependence on s_{h+1} in the loss function f^k, we can view the above expected loss in (3.1) as a value function $V_1^\pi(s_1, f^k)$ as defined in Section 2.1. A similar argument can also be applied to the budget costs, which corresponds to the value function $V_1^\pi(s_1, g_i^k)$ for all $i \in [I]$. Our analysis can be applied to the setting that the loss function f^k and the budget cost functions g_i^k for all $i \in [I]$ do not depend on the next state

s_{h+1} with no barrier. In addition, this chapter studies the problem of minimizing the overall losses, which can be interpreted as maximizing the overall negative losses w.r.t. the function $-f^k$. With the above analysis, we can convert the SSP problem to the episodic MDP defined in Chapter 2.

In this chapter, we adopt the occupancy measure $\theta(s, a, s')$ for our analysis. In general, the occupancy measure $\theta(s, a, s')$ is a joint probability of the tuple $(s, a, s') \in \mathbf{S} \times \mathbf{A} \times \mathbf{S}$ under some certain policy and transition model. Particularly, with the true transition P, we define the set as

$$\Delta = \{\theta \ : \ \theta \text{ satisfies the conditions (a), (b), and (c)}\},$$

where the conditions (a) (b) (c) [Altman, 1999] are

(a) $\sum_{s \in \mathbf{S}_h} \sum_{a \in \mathbf{A}} \sum_{s' \in \mathbf{S}_{h+1}} \theta(s, a, s') = 1, \ \forall h \in \{0, \ldots, H-1\}$, and $\theta(s, a, s') \geq 0$.

(b) $\sum_{s \in \mathbf{S}_h} \sum_{a \in \mathbf{A}} \theta(s, a, s') = \sum_{a \in \mathbf{A}} \sum_{s'' \in \mathbf{S}_{h+2}} \theta(s', a, s''), \ \forall s' \in \mathbf{S}_{h+1}, \forall h \in \{0, \ldots, H-2\}$.

(c) $\frac{\theta(s,a,s')}{\sum_{s'' \in \mathbf{S}_{h+1}} \theta(s,a,s'')} = P(s'|s, a), \ \forall (s, a, s') \in \mathbf{S}_h \times \mathbf{A} \times \mathbf{S}_{h+1}, \forall h \in \{0, \ldots, H-1\}$.

We can further recover a policy π from an occupancy measure θ via

$$\pi(a|s) = \frac{\sum_{s' \in \mathbf{S}_{h+1}} \theta(s, a, s')}{\sum_{s' \in \mathbf{S}_{h+1}, a \in \mathbf{A}} \theta(s, a, s')}, \ \forall (s, a) \in \mathbf{S}_h \times \mathbf{A}, \ \forall h \in \{0, \ldots, H-1\}.$$

In addition, we define $\overline{\theta}^k(s, a, s')$ to be the occupancy measure at episode k w.r.t. the true transition P, resulting from a policy π^k at episode k. Given the definition of occupancy measure, we can rewrite the expected loss and the budget cost as $\mathbb{E}[\sum_{h=0}^{H-1} f^k(s_h, a_h, s_{h+1})|\pi^k, P] = \langle f^k, \overline{\theta}^k \rangle$ where $\langle f^k, \overline{\theta}^k \rangle = \sum_{s,a,s'} f^k(s, a, s')\overline{\theta}^k(s, a, s')$ and $\mathbb{E}[\sum_{h=0}^{H-1} g_i^k(s_h, a_h, s_{h+1})|\pi^k, P] = \langle g_i^k, \overline{\theta}^k \rangle$ with $\langle g_i^k, \overline{\theta}^k \rangle = \sum_{s,a,s'} f^k(s, a, s')\overline{\theta}^k(s, a, s')$. We aim to solve the following constrained optimization, and let $\overline{\theta}^*$ be one solution which is further viewed as a reference point to define the regret:

$$\underset{\theta \in \Delta}{\text{minimize}} \sum_{k=1}^{K} \langle f^k, \theta \rangle, \ \text{subject to} \ \langle g_i, \theta \rangle \leq c_i, \ \forall i \in [I], \tag{3.2}$$

where $\sum_{k=1}^{K} \langle f^k, \theta \rangle = \sum_{k=1}^{K} \mathbb{E}[\sum_{h=0}^{H-1} f^k(s_h, a_h, s_{h+1})|\pi, P]$ is the overall loss in K episodes and constraints are enforced on the budget cost $\langle g_i, \theta \rangle = \mathbb{E}[\sum_{h=0}^{H-1} g_i(s_h, a_h, s_{h+1})|\pi, P]$ based on the expected budget consumption functions $g_i, \forall i \in [I]$. To measure the regret and the constraint violation respectively for solving (3.2) in an online setting, we define the following two metrics:

$$\text{Regret}(K) := \sum_{k=1}^{K} \langle f^k, \overline{\theta}^k - \overline{\theta}^* \rangle, \ \text{and} \ \text{Violation}(K) := \left\| \left[\sum_{k=1}^{K} \left(\mathbf{g}(\overline{\theta}^k) - \mathbf{c} \right) \right]_+ \right\|_2, \tag{3.3}$$

where the notation $[\mathbf{v}]_+$ denotes the entry-wise application of $\max\{\cdot, 0\}$ for any vector $\mathbf{v}$. For abbreviation, we let $\mathbf{g}^k(\theta) := [\langle g_1^k, \theta\rangle, \cdots, \langle g_I^k, \theta\rangle]^\top$, and $\mathbf{c} := [c_1, \cdots, c_I]^\top$.

The goal is to attain a sublinear regret bound and constraint violation on this problem w.r.t. *any fixed stationary policy π, which does not change over episodes.* In another word, we compare to the best policy π^* in hindsight whose corresponding occupancy measure $\overline{\theta}^* \in \Delta$ solves problem (3.2). We make the following assumption on the existence of a solution to (3.2).

Assumption 3.3. *There exists at least one fixed policy π such that the corresponding occupancy measure $\theta \in \Delta$ is feasible, i.e., $\langle g_i, \theta\rangle \leq c_i, \forall i \in [I]$.*

WLOG, we assume boundedness on function values for simplicity of notation.

Assumption 3.4. *We assume the following quantities are bounded. For any $k \geq 1$, **(1)** $\sup_{s,a,s'} |f^k(s,a,s')| \leq 1$, **(2)** $\sum_{i=1}^I \sup_{s,a,s'} |g_i^k(s,a,s')| \leq 1$, **(3)** $\sum_{i=1}^I |c_i| \leq H$.*

When the transition model P is known and Slater's condition holds (i.e., the existence of a policy which satisfies all stochastic inequality constraints with a constant ε-slackness), this stochastically constrained online linear program can be solved via similar methods as Wei et al. [2018], Yu et al. [2017] with a regret bound that depends polynomially on the cardinalities of state and action spaces, which is highly suboptimal especially when the state or action space is large. The main challenge we will address in this chapter is to *solve this problem without knowing the model P, or losses and constraints before making decisions, while tightening the dependency on both state and action spaces in the resulting performance bound.*

3.3 Proposed Algorithm

In this section, we introduce our proposed algorithm, namely, Upper Confidence Primal-Dual (UCPD) algorithm, as presented in Algorithm 1. It adopts a primal-dual mirror descent type algorithm solving constrained problems but with an important difference: we maintain a confidence set via past sample trajectories, which contains the true MDP model P with high probability, and choose the policy to minimize the proximal Lagrangian using the most optimistic model from the confidence set. Such an idea, known as "optimism in the face of uncertainty", is reminiscent of UCB algorithms [Auer et al., 2002, Bubeck and Cesa-Bianchi, 2012] for stochastic multi-armed bandit (MAB) and is used by Jaksch et al. [2010] to obtain a near-optimal regret for reinforcement learning problems.

In the algorithm, we introduce epochs, which are back-to-back time intervals that span several episodes. We use $\ell \in \{1, 2, \cdots\}$ to index the epochs and use $\ell(k)$ to denote a mapping from the episode index k to the epoch index, indicating which epoch the k-th episode lives. Next, let

$N_\ell(s,a)$ and $M_\ell(s,a,s')$ be two global counters which indicate the number of times the tuples (s,a) and (s,a,s') appear before the ℓ-th epoch. Let $n_\ell(s,a)$, $m_\ell(s,a,s')$ be two local counters which indicate the number of times the tuples (s,a) and (s,a,s') appear in the ℓ-th epoch. We start a new epoch whenever there exists (s,a) such that $n_{\ell(k)}(s,a) \geq N_{\ell(k)}(s,a)$. Otherwise, set $\ell(k+1) = \ell(k)$. Such an update rule follows from Jaksch et al. [2010]. Then, we define the empirical transition model $\widehat{P}_\ell$ at any epoch $\ell > 0$ as

$$\widehat{P}_\ell(s'|s,a) := \frac{M_\ell(s,a,s')}{\max\{1, N_\ell(s,a)\}}, \ \forall s, s' \in \mathbf{S}, \ a \in \mathbf{A}.$$

As shown in Remark 3.18, introducing the notion of epoch is necessary to achieve an $\widetilde{\mathcal{O}}(\sqrt{K})$ constraint violation.

The next lemma shows that with high probability, the true transition model P is contained in a confidence interval around the empirical one, which is adapted from Lemma 1 of Neu et al. [2012].

Lemma 3.5 (Lemma 1 of Neu et al. [2012])**.** *For any $\zeta \in (0,1)$, we have that with probability at least $1-\zeta$, for all epoch $\ell \leq \ell(K+1)$ and any state and action pair $(s,a) \in \mathbf{S} \times \mathbf{A}$, $\left\|P(\cdot|s,a) - \widehat{P}_\ell(\cdot|s,a)\right\|_1 \leq \varepsilon_\ell^\zeta(s,a)$, with the error $\varepsilon_\ell^\zeta(s,a)$ being*[2]

$$\varepsilon_\ell^\zeta(s,a) := \sqrt{\frac{2|\mathbf{S}_{h(s)+1}| \log[(K+1)|\mathbf{S}||\mathbf{A}|/\zeta]}{\max\{1, N_\ell(s,a)\}}}, \tag{3.4}$$

where $h(s)$ is a map from state s to the layer that s belongs to.

3.3.1 Computing Optimistic Policies

Next, we show how to compute the policy at each episode. Formally, we introduce a new occupancy measure at episode k, namely $\theta^k(s,a,s'), \ s,s' \in \mathbf{S}, \ a \in \mathbf{A}$. It should be emphasized that this is different from $\overline{\theta}^k(s,a,s')$ defined in the previous section as $\theta^k(s,a,s')$ is chosen by the decision maker at episode k to construct the policy. In particular, $\theta^k(s,a,s')$ does not have to satisfy the local balance equation (c). Once getting $\theta^k(s,a,s')$ (which will be detailed below), we construct the policy by

$$\pi^k(a|s) = \frac{\sum_{s'} \theta^k(s,a,s')}{\sum_{s',a} \theta^k(s,a,s')}, \ \forall a \in \mathbf{A}, \ s \in \mathbf{S}. \tag{3.5}$$

Next, we demonstrate the proposed method computing $\theta^k(s,a,s')$. First, we introduce an online dual multiplier $Q_i(k)$ for each constraint in (3.2), which is 0 when $k=1$ and is updated as follows

[2] We use log to denote the natural logarithm.

Algorithm 1 Upper-Confidence Primal-Dual (UCPD) Mirror Descent

```
Input: Let $V, \alpha > 0$, $\lambda \in [0,1)$ be some trade-off parameters. Fix $\zeta \in (0,1)$.
Initialize: $Q_i(1) = 0,\ \forall i = 1,\dots,I$. $\theta^1(s,a,s') = 1/(|\mathbf{S}_h||\mathbf{S}_{h+1}||\mathbf{A}|),\ \forall (s,a,s') \in \mathbf{S}_h \times \mathbf{A} \times \mathbf{S}_{h+1}$. $\ell(1) = 1$. $n_1(s,a) = 0,\ N_1(s,a) = 0,\ \forall (s,a) \in \mathbf{S} \times \mathbf{A}$. $m_1(s,a,s') = 0,\ M_1(s,a,s') = 0,\ \forall (s,a,s') \in \mathbf{S} \times \mathbf{A} \times \mathbf{S}$.
for $k = 1,2,3,\dots$ do
    Compute $\theta^k$ via (3.7) and the corresponding policy $\pi^k$ via (3.5).
    Sample a path $(s_0^k, a_0^k, \cdots, s_{H-1}^k, a_{H-1}^k, s_H^k)$ following the policy $\pi^k$.
    Update each dual multiplier $Q_i(k)$ via (3.6) and update the local counters:
        $$n_{\ell(k)}(s_h^k, a_h^k) = n_{\ell(k)}(s_h^k, a_h^k) + 1,\ m_{\ell(k)}(s_h^k, a_h^k, s_{h+1}^k) = m_{\ell(k)}(s_h^k, a_h^k, s_{h+1}^k) + 1.$$
    Observe the loss function $f^k$ and constraint functions $\{g_i^k\}_{i=1}^I$.
    if $\exists (s,a) \in \mathbf{S} \times \mathbf{A},\ n_{\ell(k)}(s,a) \geq N_{\ell(k)}(s,a)$, then
        Start a new epoch:
        Set $\ell(k+1) = \ell(k) + 1$, and update the global counters for all $s, s' \in \mathbf{S},\ a \in \mathbf{A}$ by
            $$N_{\ell(k+1)}(s,a) = N_{\ell(k)}(s,a) + n_{\ell(k)}(s,a),$$
            $$M_{\ell(k+1)}(s,a,s') = M_{\ell(k)}(s,a,s') + m_{\ell(k)}(s,a,s').$$
        Construct the empirical transition $\widehat{P}_{\ell(k+1)}(s'|s,a) := \frac{M_{\ell(k+1)}(s,a,s')}{\max\{1, N_{\ell(k+1)}(s,a)\}}, \forall (s,a,s')$.
        Initialize $n_{\ell(k+1)}(s,a) = 0,\ m_{\ell(k+1)}(s,a,s') = 0,\ \forall (s,a,s') \in \mathbf{S} \times \mathbf{A} \times \mathbf{S}$.
    else
        Set $\ell(k+1) = \ell(k)$.
    end if
end for
```

for $k \geq 2$,

$$Q_i(k) = \max\{Q_i(k-1) + \left\langle g_i^{k-1}, \theta^k \right\rangle - c_i,\ 0\}. \tag{3.6}$$

At each episode, we compute the occupancy measure $\theta^k(s,a,s')$ by solving an optimistic regularized linear program with tuning parameters $\lambda,\ V,\ \alpha > 0$. Specifically, we update θ^k for all $k \geq 2$ by solving the following minimization problem

$$\theta^k = \operatorname*{argmin}_{\theta \in \Delta(\ell(k), \zeta)} \left\langle V f^{k-1} + \sum_{i=1}^{I} Q_i(k-1) g_i^{k-1}, \theta \right\rangle + \alpha D(\theta, \widetilde{\theta}^{k-1}), \quad \forall k \geq 2. \tag{3.7}$$

For $k = 1$, we let $\theta^1(s,a,s') = 1/(|\mathbf{S}_h||\mathbf{S}_{h+1}||\mathbf{A}|), \forall (s,a,s') \in \mathbf{S}_h \times \mathbf{A} \times \mathbf{S}_{h+1}$. The above updating rule (3.7) introduces extra notations $\Delta(\ell(k), \zeta)$, $\widetilde{\theta}^{k-1}$, and $D(\cdot,\cdot)$, which will be elaborated below. Specifically, we denote by $D(\cdot,\cdot)$ the unnormalized Kullback-Leibler (KL) divergence for two different occupancy measures θ and θ', which is defined as

$$D(\theta, \theta') := \sum_{s,a,s'} [\theta(s,a,s') \log \frac{\theta(s,a,s')}{\theta'(s,a,s')} - \theta(s,a,s') + \theta'(s,a,s')]. \tag{3.8}$$

In addition, for $\forall h = \{0, \ldots, H-1\}$ and $\forall s \in \mathbf{S}_h, a \in \mathbf{A}, s' \in \mathbf{S}_{h+1}$, we compute $\widetilde{\theta}^{k-1}$ via $\widetilde{\theta}^{k-1}(s,a,s') = (1-\lambda)\theta^{k-1}(s,a,s') + \lambda/(|\mathbf{S}_h||\mathbf{S}_{h+1}||\mathbf{A}|)$, where $0 \leq \lambda \leq 1$. This equation introduces a probability mixing, pushing the update away from the boundary and encouraging explorations.

Furthermore, since for any epoch $\ell > 0$, we can compute the empirical transition model $\widehat{P}_\ell$ with the confidence set as defined in (3.4), we let every $\theta \in \Delta(\ell, \zeta)$ satisfy that

$$\left\| \frac{\theta(s,a,\cdot)}{\sum_{s'} \theta(s,a,s')} - \widehat{P}_\ell(\cdot|s,a) \right\|_1 \leq \varepsilon_\ell^\zeta(s,a),\ \forall s \in \mathbf{S}, a \in \mathbf{A}, \tag{3.9}$$

such that we can define the feasible set $\Delta(\ell, \zeta)$ for the optimization problem (3.7) as follows

$$\Delta(\ell, \zeta) := \{\theta : \theta \text{ satisfies conditions (a), (b), and (3.9) }\}. \tag{3.10}$$

By this definition, we know that $\theta^k \in \Delta(\ell(k), \zeta)$ at the epoch $\ell(k)$. On the other hand, according to Lemma 3.5, we have that with probability at least $1-\zeta$, for all epoch ℓ, $\Delta \subseteq \Delta(\ell, \zeta)$ holds. By Rosenberg and Mansour [2019a], the problem (3.7) is essentially a linear programming with a special structure that can be solved efficiently. We present the efficient solver for the problem (3.7) in the following subsection.

3.3.2 Efficient Solver for Subproblem

In this subsection, we provide the details on how to efficiently solve the subproblem (3.7). We can rewrite (3.7) into the following equivalent form

$$\theta^k = \operatorname*{argmin}_{\theta \in \Delta(\ell(k), \zeta)} \ \alpha^{-1}\langle \varphi^{k-1}, \theta \rangle + D(\theta, \widetilde{\theta}^{k-1}),$$

where we let $\varphi^{k-1} := V f^{k-1} + \sum_{i=1}^I Q_i(k-1) g_i^{k-1}$. According to Rosenberg and Mansour [2019a], solving the above problem is based on the following two steps

$$\underline{\theta}^k = \operatorname*{argmin}_{\theta} \ \alpha^{-1}\langle \varphi^{k-1}, \theta \rangle + D(\theta, \widetilde{\theta}^{k-1}), \tag{3.11}$$

$$\theta^k = \operatorname*{argmin}_{\theta \in \Delta(\ell(k), \zeta)} \ D(\theta, \underline{\theta}^k). \tag{3.12}$$

Note that the first step, i.e., (3.11), is an unconstrained problem, which has a closed-form solution

$$\underline{\theta}^k(s,a,s') = \widetilde{\theta}^{k-1}(s,a,s')e^{-\varphi^{k-1}/\alpha},\ \forall(s,a,s') \in \mathbf{S}_h \times \mathbf{A} \times \mathbf{S}_{h+1},\ \forall h = 0,\ldots,H-1. \quad (3.13)$$

The second step, i.e., (3.12), can be viewed as a projection of $\underline{\theta}^k(s,a,s')$ onto the feasible set $\Delta(\ell(k),\zeta)$. With the definition of the feasible set as in (3.10), further by Theorem 4.2 of Rosenberg and Mansour [2019a] and Lemma 7 of Jin et al. [2019], and plugging in $\underline{\theta}^k$ computed as (3.13), we have the following equation

$$\theta^k(s,a,s') = \frac{\widetilde{\theta}^{k-1}(s,a,s')}{Z_k^{h(s)}(\mu^k,\beta^k)} e^{B^k_{\mu^k,\beta^k}(s,a,s')}, \quad (3.14)$$

where $h(s)$ is a mapping for state s to its associated layer index, and $W_h^k(\mu,\beta)$ and $B^k_{\mu,\beta}$ are defined as follows

$$\begin{aligned} B^k_{\mu,\beta}(s,a,s') &= \mu^-(s,a,s') - \mu^+(s,a,s') + (\mu^+(s,a,s') + \mu^-(s,a,s'))\varepsilon^\zeta_{\ell(k)}(s,a) + \beta(s') \\ &\quad - \beta(s) - \varphi^{k-1}(s,a,s')/\alpha - \sum_{s''\in\mathbf{S}_{h(s)+1}} \widehat{P}_{\ell(k)}(s''|s,a)(\mu^-(s,a,s'') - \mu^+(s,a,s'')), \\ W_h^k(\mu,\beta) &= \sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s'\in\mathbf{S}_{h+1}} \widetilde{\theta}^{k-1}(s,a,s')e^{B^k_{\mu,\beta}(s,a,s')}, \end{aligned}$$

where $\beta : \mathbf{S} \to \mathbb{R}$ and $\mu = (\mu^+,\mu^-)$ with $\mu^+,\mu^- : \mathbf{S}\times\mathbf{A}\times\mathbf{S} \to \mathbb{R}_{\geq 0}$. Specifically, μ^k and β^k in (3.14) are obtained by solving a convex optimization with only non-negativity constraints, which is

$$\mu^k, \beta^k = \underset{\mu,\beta\geq 0}{\text{argmin}} \sum_{h=0}^{H-1} \log W_h^k(\mu,\beta). \quad (3.15)$$

Therefore, after solving (3.15), we can eventually compute θ^k by (3.14). Since (3.15) is associated with a convex optimization with only non-negativity constraints, it can be solved much efficiently.

3.4 Main Results

Before presenting our theoretical results, we first make assumption on the existence of Lagrange multipliers. We define a partial average function starting from any episode k as $f^{(k,\tau)} := \frac{1}{\tau}\sum_{j=0}^{\tau-1} f^{k+j}$. Then, we consider the following static optimization problem (recalling $g_i := \mathbb{E}[g_i^k]$)

$$\underset{\theta\in\Delta}{\text{minimize}}\ \langle f^{(k,\tau)},\theta\rangle \text{ s.t. } \langle g_i,\theta\rangle \leq c_i,\ \forall i \in [I]. \quad (3.16)$$

Denote the solution to this program as $\theta^*_{k,\tau}$. Define the Lagrangian dual function of (3.16) as

$$q^{(k,\tau)}(\eta) := \min_{\theta\in\Delta} \langle f^{(k,\tau)}, \theta\rangle + \sum_{i=1}^{I} \eta_i(\langle g_i, \theta\rangle - c_i),$$

where $\eta = [\eta_1, \ldots, \eta_I]^\top \in \mathbb{R}^I$ is a dual variable. We are ready to state our assumption.

Assumption 3.6. *For any episode k and any period τ, the set of primal optimal solution to* (3.16) *is non-empty. Furthermore, the set of Lagrange multipliers, which is $\mathcal{V}^*_{k,\tau} :=$ argmax$_{\eta\in\mathbb{R}^I_+} q^{(k,\tau)}(\eta)$, is non-empty and bounded. Any vector in $\mathcal{V}^*_{k,\tau}$ is called a Lagrange multiplier associated with* (3.16)*. Furthermore, let $B > 0$ be a constant such that for any $k \in \{1, \ldots, K\}$ and $\tau = \sqrt{K}$, the dual optimal set $\mathcal{V}^*_{k,\tau}$ defined above satisfies* $\max_{\eta\in\mathcal{V}^*_{k,\tau}} \|\eta\|_2 \leq B$.

We have the following simple sufficient condition which is a direct corollary of Lemma 1 in Nedić and Ozdaglar [2009]:

Lemma 3.7. *Suppose that the problem* (3.16) *is feasible. Then, the set of Lagrange multipliers $\mathcal{V}^*_{k,\tau}$ defined in Assumption 3.6 is nonempty and bounded if the Slater condition holds, i.e., $\exists\theta \in \Delta,\ \varepsilon > 0$ such that $\langle g_i, \theta\rangle \leq c_i - \varepsilon,\ \forall i \in [I]$.*

In fact, it can be shown that some certain constraint qualification condition more general than Slater condition can imply the boundedness of Lagrange multipliers (see, for example, Lemma 18 of Wei et al. [2019]). According to Wei et al. [2019], Assumption 3.6 is weaker than the Slater condition commonly adopted in previous constrained online learning works. The motivation for such a Lagrange multiplier condition is that it is a sufficient condition of a key structural property on the dual function $q^{(k,\tau)}(\eta)$, namely, the error bound condition. Formally, we have the following definition.

Definition 3.8 (Error Bound Condition (EBC))**.** Let $F(\mathbf{x})$ be a concave function over $\mathbf{x} \in \mathcal{C}$, where the set $\mathcal{C}$ is closed and convex. Suppose $\Lambda^* := \mathrm{argmax}_{\mathbf{x}\in\mathcal{C}} F(\mathbf{x})$ is non-empty. The function $F(\mathbf{x})$ satisfies the EBC if there exists constants $\vartheta,\ \sigma > 0$ such that for any $\mathbf{x} \in \mathcal{C}$ satisfying[3] $\mathrm{dist}(\mathbf{x}, \Lambda^*) \geq \vartheta$,

$$F(\mathbf{x}^*) - F(\mathbf{x}) \geq \sigma \cdot \mathrm{dist}(\mathbf{x}, \Lambda^*) \text{ with } \mathbf{x}^* \in \Lambda^*.$$

Note that in Definition 3.8, Λ^* is a closed convex set, which follows from the fact that $F(\mathbf{x})$ is a concave function and thus all superlevel sets are closed and convex. The following lemma, whose proof can be found in Lemma 5 of Wei et al. [2019], shows the relation between the Lagrange multiplier condition and the dual function.

[3] We let $\mathrm{dist}(\mathbf{x}, \Lambda^*) := \min_{\mathbf{x}'\in\Lambda^*} \frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2$ as the Euclidean distance between a point $\mathbf{x}$ and the set Λ^*.

Lemma 3.9. *Fix $K \geq 1$. Under Assumption 3.6, for any $k \in \{0, \ldots, K-1\}$ and $\tau = \sqrt{K}$, the dual function $q^{(k,\tau)}(\eta)$ satisfies EBC with $\sigma > 0$ and $\vartheta > 0$, i.e., for any $\eta \in \mathbb{R}^I$ satisfying $\operatorname{dist}(\eta, \mathcal{V}^*_{k,\tau}) \geq \vartheta$, we have*

$$q^{(k,\tau)}(\eta^*_{k,\tau}) - q^{(k,\tau)}(\eta) \geq \sigma \cdot \operatorname{dist}(\eta, \mathcal{V}^*_{k,\tau}), \ \forall\, \eta^*_{k,\tau} \in \mathcal{V}^*_{k,\tau}.$$

We define $\operatorname{dist}(\eta, \mathcal{V}^*_{k,\tau}) := \min_{\eta' \in \mathcal{V}^*_{k,\tau}} \frac{1}{2}\|\eta - \eta'\|_2^2$ as Euclidean distance between a point η and the set $\mathcal{V}^*_{k,\tau}$.

Based on the above assumptions and lemmas, we present results of the regret and constraint violation.

Theorem 3.10. *Consider any fixed horizon $K \geq |\mathbf{S}||\mathbf{A}|$ with $|\mathbf{S}|, |\mathbf{A}| > 1$. Suppose Assumption 3.3, 3.4, 3.6 hold and there exist absolute constants $\overline{\sigma}$ and $\overline{\vartheta}$ such that $\sigma \geq \overline{\sigma}$ and $\vartheta \leq \overline{\vartheta}$ for all σ, ϑ in Lemma 3.9 over $k = \{0, 1, \ldots, K-1\}$ and $\tau = \sqrt{K}$. If setting $\alpha = KH$, $V = H\sqrt{K}$, $\lambda = 1/K$ and $\zeta \in (0, 1/(4 + 8H/\overline{\sigma})]$ in Algorithm 1, with probability at least $1 - 4\zeta$, we have*

$$\operatorname{Regret}(K) \leq \widetilde{\mathcal{O}}\Big(H|\mathbf{S}|\sqrt{K|\mathbf{A}|}\Big), \qquad \operatorname{Violation}(K) \leq \widetilde{\mathcal{O}}\Big(H|\mathbf{S}|\sqrt{K|\mathbf{A}|}\Big),$$

where $\widetilde{\mathcal{O}}(\cdot)$ hides the logarithmic factors $\log^{3/2}(K/\zeta)$ and $\log(K|\mathbf{S}||\mathbf{A}|/\zeta)$.

For unconstrained episodic MDPs with the unknown transition and adversarial losses, the recent work Rosenberg and Mansour [2019a] achieves a tight regret bound of $\widetilde{\mathcal{O}}(H|\mathbf{S}|\sqrt{|\mathbf{A}|K})$, almost matches the lower bound $\Omega(\sqrt{H|\mathbf{S}||\mathbf{A}|K})$ [Jaksch et al., 2010] up to an $\mathcal{O}(\sqrt{H|\mathbf{S}|})$ factor. Comparing to aforementioned works, for CMDPs, our proposed algorithm can maintain the $\widetilde{\mathcal{O}}(H|\mathbf{S}|\sqrt{|\mathbf{A}|K})$ regret bound and also achieve a constraint violation bound of $\widetilde{\mathcal{O}}(H|\mathbf{S}|\sqrt{|\mathbf{A}|K})$ under the setting of the unknown transition model, the adversarial losses, and stochastic constraints.

3.5 Theoretical Analysis

3.5.1 Proof of Regret Bound

Lemma 3.11. *The updating rules in Algorithm 1 ensure that with probability at least $1 - 2\zeta$,*

$$\sum_{k=1}^{K} \left\|\theta^k - \overline{\theta}^k\right\|_1 \leq (\sqrt{2}+1)H|\mathbf{S}|\sqrt{2K|\mathbf{A}|\log\frac{2K|\mathbf{S}||\mathbf{A}|}{\zeta}} + 2H^2\sqrt{2K\log\frac{H}{\zeta}}.$$

Lemma 3.12. *The updating rules in Algorithm 1 ensure that with probability at least* $1-\zeta$*,*

$$\sum_{k=1}^{K} \left\langle f^k, \theta^k - \overline{\theta}^* \right\rangle \leq \frac{4H^2K + (\lambda T + 1)\alpha H \log |\mathbf{S}|^2|\mathbf{A}|}{V} + 2\lambda K H + \frac{KH}{2\alpha} + \frac{1}{V}\sum_{k=1}^{K} \langle \mathbf{Q}(k), \mathbf{g}^k(\overline{\theta}^*) - \mathbf{c} \rangle.$$

Here we let $\mathbf{Q}(k) := [Q_1(k),\ Q_2(k),\ \cdots,\ Q_I(k)]^\top$. Next, we present Lemma 3.13, which is one of the key lemmas in our proof. Then, this lemma indicates that $\|\mathbf{Q}(k)\|_2$ is bounded by $\mathcal{O}(\sqrt{K})$ with high probability when setting the parameters τ, V, α, λ as in Theorem 3.10. Thus, introducing stochastic constraints retains the $\mathcal{O}(\sqrt{K})$ regret. Moreover, this lemma will lead to constraint violation in the level of $\mathcal{O}(\sqrt{K})$. Lemma 3.13 is proved by making use of Assumption 3.6 and Lemma 3.9.

Lemma 3.13. *Letting* $\tau = \sqrt{K}$ *and* ζ *satisfy* $\overline{\sigma}/4 \geq \zeta(\overline{\sigma}/2 + 2H)$*, the updating rules in Algorithm 1 ensure that with probability at least* $1 - K\delta$*, the following inequality holds for all* $k \in [K+1]$*,*

$$\|\mathbf{Q}(k)\|_2 \leq \omega := \psi + \tau \frac{512H^2}{\overline{\sigma}} \log\left(1 + \frac{128H^2}{\overline{\sigma}^2} e^{\overline{\sigma}/(32H)}\right) + \tau \frac{64H^2}{\overline{\sigma}} \log \frac{1}{\delta} + 2\tau H,$$

where we define $\psi := (2\tau H + C_{V,\alpha,\lambda})/\overline{\sigma} + 2\alpha H \log(|\mathbf{S}|^2|\mathbf{A}|/\lambda)/(\overline{\sigma}\tau) + \tau\overline{\sigma}/2$ *and* $C_{V,\alpha,\lambda} := 2(\overline{\sigma}B + \overline{\sigma}\,\overline{\vartheta})V + (6 + 4\overline{\vartheta})VH + VH/\alpha + 4H\lambda V + 2\alpha\lambda H \log |\mathbf{S}|^2|\mathbf{A}| + 8H^2$.

The upper bound of $||Q(k)||_2$ is a convex function w.r.t. τ, which thus indicates that there exists a tight upper bound of $||Q(k)||_2$ if τ is chosen by finding the minimizer of this upper bound. In this chapter, we directly set $\tau = \sqrt{T}$, which suffices to give an $\widetilde{\mathcal{O}}(\sqrt{T})$ upper bound.

Remark 3.14. We discuss the upper bound of the term $\log\big(1 + \frac{128H^2}{\overline{\sigma}^2} e^{\overline{\sigma}/(32H)}\big)$ in the following way: **(1)** if $\frac{128H^2}{\overline{\sigma}^2} e^{\overline{\sigma}/(32H)} \geq 1$, then this term is bounded by $\log\big(\frac{256H^2}{\overline{\sigma}^2} e^{\overline{\sigma}/(32H)}\big) = \frac{\overline{\sigma}}{32H} + \log \frac{256H^2}{\overline{\sigma}^2}$; **(2)** if $\frac{128H^2}{\overline{\sigma}^2} e^{\overline{\sigma}/(32H)} < 1$, then the term is bounded by $\log 2$. Thus, combining the two cases, we have $\log\big(1 + \frac{128H^2}{\overline{\sigma}^2} e^{\overline{\sigma}/(32H)}\big) \leq \log 2 + \frac{\overline{\sigma}}{32H} + \log \frac{256H^2}{\overline{\sigma}^2}$. This discussion shows that the $\log$ term in the result of Lemma 3.13 will not introduce extra dependence on H except a $\log H$ term.

With the bound of $\|\mathbf{Q}(k)\|_2$ in Lemma 3.13, we further obtain the following lemma.

Lemma 3.15. *By Algorithm 1, if* $\overline{\sigma}/4 \geq \zeta(\overline{\sigma}/2 + 2H)$*, then with probability at least* $1 - 2K\delta$*,*

$$\sum_{k=1}^{K} \langle \mathbf{Q}(k), \mathbf{g}^k(\overline{\theta}^*) - \mathbf{c} \rangle \leq 2H\omega\sqrt{K \log \frac{1}{K\delta}},$$

with ω *defined as the same as in Lemma 3.13.*

Proof of Regret Bound in Theorem 3.10. Recall that θ^k is the probability vector chosen by the decision maker, and $\overline{\theta}^k$ is the true occupancy measure at episode k while $\overline{\theta}^*$ is the solution to the problem (3.2). The main idea is to decompose the regret as follows

$$\begin{aligned}\sum_{k=1}^{K} \big\langle f^k, \overline{\theta}^k - \overline{\theta}^* \big\rangle &= \sum_{k=1}^{K} \big(\big\langle f^k, \overline{\theta}^k - \theta^k \big\rangle + \big\langle f^k, \theta^k - \overline{\theta}^* \big\rangle \big) \\ &\leq \underbrace{\sum_{k=1}^{K} \big\| \overline{\theta}^k - \theta^k \big\|_1}_{\text{Term(I)}} + \underbrace{\sum_{k=1}^{K} \big\langle f^k, \theta^k - \overline{\theta}^* \big\rangle}_{\text{Term(II)}},\end{aligned} \tag{3.17}$$

where we use Assumption 3.4 such that $\big\langle f^k, \overline{\theta}^k - \theta^k \big\rangle \leq \|f^k\|_\infty \big\| \overline{\theta}^k - \theta^k \big\|_1 \leq \big\| \overline{\theta}^k - \theta^k \big\|_1$. Thus, it suffices to bound the Term(I) and Term(II).

We first show the bound for Term(I). According to Lemma 3.11, by the fact that $H \leq |\mathbf{S}|$ and $|\mathbf{S}|, |\mathbf{A}| \geq 1$, we have that with probability at least $1 - 2\zeta$, the following holds

$$\text{Term(I)} \leq \mathcal{O}\big(H|\mathbf{S}|\sqrt{K|\mathbf{A}|} \log^{\frac{1}{2}}(K|\mathbf{S}||\mathbf{A}|/\zeta)\big). \tag{3.18}$$

For Term(II), setting $V = H\sqrt{K}$, $\alpha = KH$, $\tau = \sqrt{K}$, and $\lambda = 1/K$, by Lemma 3.12, we obtain

$$\text{Term(II)} \leq 8H\sqrt{K|\mathbf{S}||\mathbf{A}|} + \frac{1}{H\sqrt{K}} \sum_{k=1}^{K} \langle \mathbf{Q}(k), \mathbf{g}^k(\overline{\theta}^*) - \mathbf{c} \rangle,$$

where we use the inequality that $\log|\mathbf{S}||\mathbf{A}| \leq \sqrt{|\mathbf{S}||\mathbf{A}|}$ with the inequality $\sqrt{x} \geq \log x$. Thus, we further need to bound the last term of the above inequality. By Lemma 3.15 and Remark 3.14, with probability at least $1 - 2K\delta$ for all $k \in \{1, \ldots, K\}$, we have

$$\frac{1}{H\sqrt{K}} \sum_{k=1}^{K} \langle \mathbf{Q}(k), \mathbf{g}^k(\overline{\theta}^*) - \mathbf{c} \rangle \leq \mathcal{O}\big(H|\mathbf{S}|\sqrt{K|\mathbf{A}|} \log^{\frac{3}{2}}(K/\delta)\big),$$

by the facts that $H \leq |\mathbf{S}|$, $|\mathbf{S}| > 1$, $|\mathbf{A}| > 1$, and the assumption $K \geq |\mathbf{S}||\mathbf{A}|$, as well as the computation of ψ as $\psi = \mathcal{O}\big(H^2\sqrt{K} + H \log|\mathbf{S}||\mathbf{A}| + H^2\sqrt{K}\log(K|\mathbf{S}||\mathbf{A}|)\big)$. Therefore, with probability at least $1 - 2K\delta$, the following holds

$$\text{Term(II)} \leq \mathcal{O}\big(H|\mathbf{S}|\sqrt{K|\mathbf{A}|} \log^{\frac{3}{2}}(K/\delta)\big). \tag{3.19}$$

Combining (3.18) and (3.19) with (3.17), and letting $\delta = \zeta/K$, by the union bound, we eventually obtain that with probability at least $1-4\zeta$, the regret bound $\mathrm{Regret}(K) \leq \widetilde{\mathcal{O}}\big(H|\mathbf{S}|\sqrt{K|\mathbf{A}|}\big)$ holds, where the notation $\widetilde{\mathcal{O}}(\cdot)$ hides the logarithmic factors. We further let $\zeta \leq 1/(4 + 8H/\overline{\sigma}) < 1/4$

(such that $\overline{\sigma}/4 \geq \zeta(\overline{\sigma}/2 + 2H)$ is guaranteed). This completes the proof. □

3.5.2 Proof of Constraint Violation Bound

Lemma 3.16. *The updating rules in Algorithm 1 ensure*

$$\left\| \left[\sum_{k=1}^{K} (\mathbf{g}^k(\theta^k) - \mathbf{c}) \right]_+ \right\|_2 \leq \|\mathbf{Q}(K+1)\|_2 + \sum_{k=1}^{K} \left\|\theta^{k+1} - \theta^k\right\|_1.$$

Lemma 3.17. *The updating rules in Algorithm 1 ensure*

$$\sum_{k=1}^{K} \left\|\theta^{k+1} - \theta^k\right\|_1 \leq 3H\sqrt{K|\mathbf{S}||\mathbf{A}|} \log \frac{8K}{|\mathbf{S}||\mathbf{A}|} + \frac{2H}{(1-\lambda)^2\alpha} \sum_{k=1}^{K} \|\mathbf{Q}(k)\|_2 + \frac{2KHV}{(1-\lambda)^2\alpha} + \frac{2\lambda KH}{1-\lambda} + \frac{\sqrt{8\lambda \log |\mathbf{S}|^2|\mathbf{A}|}}{1-\lambda} KH.$$

Remark 3.18. The proof of Lemma 3.17 uses the fact that the confidence set of P changes only $\sqrt{K|\mathbf{S}||\mathbf{A}|}\log_2(8K/(|\mathbf{S}||\mathbf{A}|))$ times as shown in Lemma 3.26, thanks to the doubling of the epoch length in Algorithm 1. Within each epoch where the confidence set is unchanged, we further show $\|\theta^{k+1} - \theta^k\|_1$ is sufficiently small. Therefore, we can show that the cumulative update difference $\sum_{k=1}^{K} \left\|\theta^{k+1} - \theta^k\right\|_1$ grows in the order of $\sqrt{K}$, which further leads to an $\mathcal{O}(\sqrt{K})$ constraint violation according to Lemma 3.16.

Proof of Constraint Violation Bound in Theorem 3.10. We decompose the constraint violation as

$$\begin{aligned} \left\| \left[\sum_{k=1}^{K} \left(\mathbf{g}^k(\overline{\theta}^k) - \mathbf{c} \right) \right]_+ \right\|_2 &\leq \sum_{k=1}^{K} \left\|\mathbf{g}^k(\theta^k) - \mathbf{g}^k(\overline{\theta}^k)\right\|_2 + \left\| \left[\sum_{k=1}^{K} \left(\mathbf{g}^k(\theta^k) - \mathbf{c} \right) \right]_+ \right\|_2 \\ &\leq \underbrace{\sum_{k=1}^{K} \left\|\overline{\theta}^k - \theta^k\right\|_1}_{\text{Term(III)}} + \underbrace{\left\| \left[\sum_{k=1}^{K} \left(\mathbf{g}^k(\theta^k) - \mathbf{c} \right) \right]_+ \right\|_2}_{\text{Term(IV)}}, \end{aligned} \tag{3.20}$$

where the second inequality is due to Assumption 3.4 that $\|\mathbf{g}^k(\theta^k) - \mathbf{g}^k(\overline{\theta}^k)\|_2 = (\sum_{i=1}^{I} |\langle g_i^k, \theta^k - \overline{\theta}^k \rangle|^2)^{\frac{1}{2}} \leq \sum_{i=1}^{I} \|g_i^k\|_\infty \|\theta^k - \overline{\theta}^k\|_1 \leq \|\theta^k - \overline{\theta}^k\|_1$. Thus, it suffices to bound Terms (III) and (IV).

For Term(III), we already have its bound as (3.18). Then, we focus on proving the upper bound of Term(IV). Set $V = H\sqrt{K}$, $\alpha = KH$, $\tau = \sqrt{K}$, and $\lambda = 1/K$ as in the proof of the regret bound. By Lemma 3.16, we know that to bound Term(IV) requires bounding the terms $\|\mathbf{Q}(K+1)\|_2$ and $\sum_{k=1}^{K} \|\theta^{k+1} - \theta^k\|_1$. By Lemma 3.13, combining it with Remark 3.14 and $\psi = \mathcal{O}\big(H^2\sqrt{K} + H \log |\mathbf{S}||\mathbf{A}| + H^2 \log(K|\mathbf{S}||\mathbf{A}|)/\sqrt{K}\big)$ as shown in the proof of the regret bound, letting $\overline{\sigma}/4 \geq \zeta(\overline{\sigma}/2 + 2H)$, with probability $1 - K\delta$, for all $k \in [K+1]$, the following inequality holds

$$\|\mathbf{Q}(k)\|_2 \leq \mathcal{O}\big(H^2\sqrt{K}\log(H/\delta)\big), \tag{3.21}$$

where we use $\log x \leq \sqrt{x}$. This gives the upper bound of $\|\mathbf{Q}(K+1)\|_2$ which is $\|\mathbf{Q}(K+1)\|_2 \leq \mathcal{O}\big(H^2\sqrt{K}\log(H/\delta)\big)$.

Furthermore, by Lemma 3.17, we know that the the key to bound $\sum_{k=1}^{K}\|\theta^{k+1} - \theta^k\|_1$ is also the drift bound for $\mathbf{Q}(k)$. Therefore, by (3.21) and the settings of the parameters α, λ, V, we have

$$\sum_{k=1}^{K}\|\theta^{k+1} - \theta^k\|_1 \leq \mathcal{O}\big(H|\mathbf{S}|\sqrt{|\mathbf{A}|K}\log(K|\mathbf{S}||\mathbf{A}|/\delta)\big), \tag{3.22}$$

by the facts that $H \leq |\mathbf{S}|$, $|\mathbf{S}| > 1$, $|\mathbf{A}| > 1$ and the condition $|\mathbf{S}||\mathbf{A}| \leq K$. Thus combining (3.21) and (3.22) with Lemma 3.16, and letting $\delta = \zeta/K$, then with probability at least $1 - \zeta$, we have

$$\text{Term(IV)} \leq \mathcal{O}\big(H|\mathbf{S}|\sqrt{|\mathbf{A}|K}\log(K|\mathbf{S}||\mathbf{A}|/\delta)\big).$$

Combining results for Term(III) and Term(IV) with (3.20), by the union bound, with probability at least $1 - 4\zeta$, the constraint violation $\mathrm{Violation}(K) \leq \widetilde{\mathcal{O}}\big(H|\mathbf{S}|\sqrt{K|\mathbf{A}|}\big)$ holds. This finishes the proof. □

3.6 Conclusion

In this chapter, we propose a new upper confidence primal-dual algorithm to solve online constrained episodic MDPs with adversarial losses and stochastically changing constraints. In particular, our algorithm does not require the true transition models of MDPs and achieves $\widetilde{\mathcal{O}}(H|\mathbf{S}|\sqrt{|\mathbf{A}|K})$ regret and constraint violation.

3.7 Proofs of Lemmas for Regret Bound

3.7.1 Proof of Lemma 3.11

We first provide Lemmas 3.19 and 3.20 below. Then, we give the proof of Lemma 3.11 based on the two lemmas.

Lemma 3.19 (Lemma 19 in Jaksch et al. [2010]). *For any sequence of numbers $x_1, \dots, x_n$ with*

$0 \leq x_k \leq X_{k-1} := \max\left\{1, \sum_{i=1}^{k-1} x_i\right\}$, *the following inequality holds*

$$\sum_{k=1}^{n} \frac{x_k}{\sqrt{X_{k-1}}} \leq (\sqrt{2}+1)\sqrt{X_n}.$$

Lemma 3.20. *Let $\widehat{d}_k(s)$ and $d_k(s)$ be the state stationary distributions associated with θ^k and $\overline{\theta}^k$ respectively, and $\widehat{P}_{\ell(k)}(s'|a,s)$ and $P(s'|a,s)$ be the corresponding transition distributions. Denote $\pi^k(a|s)$ as the policy at episode k. There are $\theta^k(s,a,s') = \widehat{d}_k(s)\pi^k(a|s)\widehat{P}_{\ell(k)}(s'|a,s)$ and $\overline{\theta}^k(s,a,s) = d_k(s)\pi^k(a|s)P(s'|a,s)$. On the other hand, there are also $\widehat{d}_k(s') = \sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\theta^k(s,a,s'), \forall s' \in \mathbf{S}_{h+1}$, and $d_k(s') = \sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\theta^k(s,a,s'), \forall s' \in \mathbf{S}_{h+1}$. Then, we have the following inequality*

$$\|\theta^k - \overline{\theta}^k\|_1 \leq \sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}} \mu_k(s,a)\|\widehat{P}_{\ell(k)}(\cdot|s,a) - P(\cdot|s,a)\|_1,$$

where we let $\mu_k(s,a) = d_k(s)\pi^k(a|s)$.

Proof. By the definitions of $\widehat{d}_k$, d_k, $\widehat{P}_{\ell(k)}$, P, and π^k shown in Lemma 3.20, we have

$$\begin{aligned}
\|\theta^k - \overline{\theta}^k\|_1 &= \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}} \|\theta^k(a,s,\cdot) - \overline{\theta}^k(a,s,\cdot)\|_1 \\
&= \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}} \pi^k(a|s)\|\widehat{P}_{\ell(k)}(\cdot|a,s)\widehat{d}_k(s) - P(\cdot|a,s)d_k(s)\|_1 \\
&= \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}} \pi^k(a|s)\|\widehat{P}_{\ell(k)}(\cdot|a,s)\widehat{d}_k(s) - \widehat{P}_{\ell(k)}(\cdot|a,s)d_k(s) \\
&\quad + \widehat{P}_{\ell(k)}(\cdot|a,s)d_k(s) - P(\cdot|a,s)d_k(s)\|_1.
\end{aligned}$$

Thus, by triangle inequality for $\|\cdot\|_1$, we can bound the term $\|\theta^k - \overline{\theta}^k\|_1$ in the following way

$$\begin{aligned}
\|\theta^k - \overline{\theta}^k\|_1 &\leq \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}} \pi^k(a|s)[\|\widehat{P}_{\ell(k)}(\cdot|a,s)\widehat{d}_k(s) - \widehat{P}_{\ell(k)}(\cdot|a,s)d_k(s)\|_1 \\
&\quad + \|\widehat{P}_{\ell(k)}(\cdot|a,s)d_k(s) - P(\cdot|a,s)d_k(s)\|_1] \\
&\leq \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}} \pi^k(a|s)d_k(s)\|\widehat{P}_{\ell(k)}(\cdot|a,s) - P(\cdot|a,s)\|_1 \\
&\quad + \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}} \pi^k(a|s)\|\widehat{P}_{\ell(k)}(\cdot|a,s)\|_1 \cdot |\widehat{d}_k(s) - d_k(s)|.
\end{aligned} \tag{3.23}$$

Then we need to bound the last two terms of (3.23) respectively. For the first term on the right-hand side of (3.23), we have

$$\begin{aligned}&\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\pi^k(a|s)d_k(s)\|\widehat{P}_{\ell(k)}(\cdot|a,s)-P(\cdot|a,s)\|_1\\&\qquad=\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\mu_k(s,a)\|\widehat{P}_{\ell(k)}(\cdot|a,s)-P(\cdot|a,s)\|_1,\end{aligned}\tag{3.24}$$

where $\mu_k(s,a)=\pi^k(a|s)d_k(s)$ denotes the joint distribution probability of (s,a).

Next, we bound the last term on the right-hand side of (3.23), which is

$$\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\pi^k(a|s)\|\widehat{P}_{\ell(k)}(\cdot|a,s)\|_1\cdot|\widehat{d}_k(s)-d_k(s)|=\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\pi^k(a|s)|\widehat{d}_k(s)-d_k(s)|,$$

since $\|\widehat{P}_{\ell(k)}(\cdot|a,s)\|_1=\sum_{s'\in\mathbf{S}_{h+1}}\widehat{P}_{\ell(k)}(s'|a,s)=1$. Furthermore, we can bound the last term above as

$$\begin{aligned}&\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\pi^k(a|s)|\widehat{d}_k(s)-d_k(s)|\\&\qquad=\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}|\widehat{d}_k(s)-d_k(s)|=\sum_{h=1}^{H-1}\sum_{s\in\mathbf{S}_h}|\widehat{d}_k(s)-d_k(s)|\\&\qquad=\sum_{h=1}^{H-1}\sum_{s\in\mathbf{S}_h}\Big|\sum_{s''\in\mathbf{S}_{h-1}}\sum_{a\in\mathbf{A}}\theta^k(s'',a,s)-\sum_{s''\in\mathbf{S}_{h-1}}\sum_{a\in\mathbf{A}}\overline{\theta}^k(s'',a,s)\Big|,\end{aligned}$$

where the first equality is due to $\sum_{a\in\mathbf{A}}\pi^k(a|s)=1$, the second equality is due to $\widehat{d}_k(s_0)=d_k(s_0)=1$, and the third equality is by the relations $\widehat{d}_k(s)=\sum_{s''\in\mathbf{S}_{h-1}}\sum_{a\in\mathbf{A}}\theta^k(s'',a,s)$ and $d_k(s)=\sum_{s''\in\mathbf{S}_{h-1}}\sum_{a\in\mathbf{A}}\overline{\theta}^k(s'',a',s)$, $\forall s\in\mathbf{S}_h$. Further bounding the last term of the above equation gives

$$\begin{aligned}&\sum_{h=1}^{H-1}\sum_{s\in\mathbf{S}_h}\Big|\sum_{s''\in\mathbf{S}_{h-1}}\sum_{a\in\mathbf{A}}\theta^k(s'',a,s)-\sum_{s''\in\mathbf{S}_{h-1}}\sum_{a\in\mathbf{A}}\overline{\theta}^k(s'',a,s)\Big|\\&\qquad\le\sum_{h=1}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{s''\in\mathbf{S}_{h-1}}\sum_{a\in\mathbf{A}}\Big|\theta^k(s'',a,s)-\overline{\theta}^k(s'',a,s)\Big|\\&\qquad=\sum_{h=1}^{H-1}\sum_{s''\in\mathbf{S}_{h-1}}\sum_{a\in\mathbf{A}}\big\|\theta^k(s'',a,\cdot)-\overline{\theta}^k(s'',a,\cdot)\big\|_1=\sum_{h=0}^{H-2}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\big\|\theta^k(s,a,\cdot)-\overline{\theta}^k(s,a,\cdot)\big\|_1,\end{aligned}$$

which eventually implies that the last term on the right-hand side of (3.23) can be bounded as

$$\begin{aligned}&\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\pi^k(a|s)\|\widehat{P}_{\ell(k)}(\cdot|a,s)\|_1\cdot|\widehat{d}_k(s)-d_k(s)|\\&\qquad\le\sum_{h=0}^{H-2}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\big\|\theta^k(s,a,\cdot)-\overline{\theta}^k(s,a,\cdot)\big\|_1.\end{aligned}\tag{3.25}$$

Therefore, plugging the bounds (3.24) and (3.25) in (3.23), we have

$$\begin{aligned}\|\theta^k-\bar{\theta}^k\|_1&=\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\big\|\theta^k(a,s,\cdot)-\overline{\theta}^k(a,s,\cdot)\big\|_1\\&\le\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\mu_k(s,a)\big\|\widehat{P}_{\ell(k)}(\cdot|a,s)-P(\cdot|a,s)\big\|_1\\&\quad+\sum_{h=0}^{H-2}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\big\|\theta^k(s,a,\cdot)-\overline{\theta}^k(s,a,\cdot)\big\|_1.\end{aligned}$$

Recursively applying the above inequality, we obtain

$$\|\theta^k-\overline{\theta}^k\|_1\le\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\mu_k(s,a)\Big\|\widehat{P}_{\ell(k)}(\cdot|s,a)-P(\cdot|s,a)\Big\|_1,$$

which completes the proof. □

Now, we are in position to give the proof of Lemma 3.11.

Proof of Lemma 3.11. The proof for Lemma 3.11 adopts similar ideas in Neu et al. [2012], Rosenberg and Mansour [2019a]. By Lemma 3.20, one can show that

$$\begin{aligned}\|\theta^k-\overline{\theta}^k\|_1&\le\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\mu_k(s,a)\big\|\widehat{P}_{\ell(k)}(\cdot|s,a)-P(\cdot|s,a)\big\|_1\\&=\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\Big[\big(\mu_k(s,a)-\mathbb{1}\{s_j^k=s,\ a_j^k=a\}\big)\big\|\widehat{P}_{\ell(k)}(\cdot|s,a)-P(\cdot|s,a)\big\|_1\\&\quad+\mathbb{1}\{s_j^k=s,\ a_j^k=a\}\big\|\widehat{P}_{\ell(k)}(\cdot|s,a)-P(\cdot|s,a)\big\|_1\Big],\end{aligned}$$

where we denote $\mathbb{1}\{s_j^k=s,\ a_j^k=a\}$ the indicator random variable that equals 1 with probability $\mu_k(s,a),\forall s\in X_j,a\in\mathbf{A}$ and 0 otherwise. Denote $\xi^k(s,a)=\|\widehat{P}_{\ell(k)}(\cdot|s,a)-P(\cdot|s,a)\|_1$ for abbreviation. We can see that $\xi^k(s,a)\le\|\widehat{P}_{\ell(k)}(\cdot|s,a)\|_1+\|P(\cdot|s,a)\|_1=2$. Taking summation over K time slots on both sides of the above inequality, we obtain

$$\begin{aligned}\sum_{k=1}^{K}\|\theta^k-\overline{\theta}^k\|_1 \leq & \sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}(\mu_k(s,a)-\mathbb{1}\{s_j^k=s,\ a_j^k=a\})\xi^k(s,a)\\ &+\sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\mathbb{1}\{s_j^k=s,\ a_j^k=a\}\xi^k(s,a).\end{aligned} \tag{3.26}$$

Next, we bound the first term on the right-hand side of (3.26). Let $\mathcal{F}^{k-1}$ be the system history up to $(k-1)$-th episode. Then, we have

$$\mathbb{E}\Big\{\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}(\mu_k(s,a)-\mathbb{1}\{s_j^k=s,\ a_j^k=a\})\xi^k(s,a)\ \Big|\ \mathcal{F}^{k-1}\Big\}=0,$$

since ξ^k is only associated with system randomness history up to $k-1$ episodes. Thus, the term $\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}(\mu_k(s,a)-\mathbb{1}\{s_j^k=s,\ a_j^k=a\})\xi^k(s,a)$ is a martingale difference sequence with respect to $\mathcal{F}^{k-1}$. Furthermore, by $\xi^k(s,a)\leq 2$ and $\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\mathbb{1}\{s_j^k=s,\ a_j^k=a\})=1$, there will be

$$\begin{aligned}&\left|\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}(\mu_k(s,a)-\mathbb{1}\{s_j^k=s,\ a_j^k=a\})\xi^k(s,a)\right|\\ &\qquad\leq\left|\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\mathbb{1}\{s_j^k=s,\ a_j^k=a\}\right|\xi^k(s,a)+\left|\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\mu_k(s,a)\right|\xi^k(s,a)\leq 4.\end{aligned}$$

Thus, by Hoeffding-Azuma inequality, we obtain that with probability at least $1-\zeta/H$,

$$\sum_{k=1}^{K}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}(\mu_k(s,a)-\mathbb{1}\{s_j^k=s,\ a_j^k=a\})\xi^k(s,a)\leq 4\sqrt{2K\log\frac{H}{\zeta}}.$$

According to the union bound, we further have that with probability at least $1-\zeta$, the above inequality holds for all $j=0,...,H-1$. This implies that with probability at least $1-\zeta$, the following inequality holds

$$\sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}(\mu_k(s,a)-\mathbb{1}\{s_j^k=s,\ a_j^k=a\})\xi^k(s,a)\leq 2H^2\sqrt{2K\log\frac{H}{\zeta}}. \tag{3.27}$$

Furthermore, we adopt the same argument as the first part of the proof of Lemma 5 in Neu et al. [2012] to show the upper bound of $\sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\mathbb{1}\{s_j^k=s,\ a_j^k=a\}\xi^k(s,a)$ in (3.26). Recall that $\ell(k)$ denotes the epoch that the k-th episode belongs to. By the definition of the state-action pair counter $N_\ell(s,a)$ and $n_\ell(s,a)$, we have $N_q(s,a) = \sum_{\ell=0}^{q-1} n_\ell(s,a)$. According to Lemma 3.19, we have

$$\sum_{q=1}^{\ell(k)} \frac{n_q(s,a)}{\max\{1, \sqrt{N_q(s,a)}\}} \leq (\sqrt{2}+1)\sqrt{\sum_{q=1}^{\ell(k)} n_q(s,a)}. \tag{3.28}$$

Since we can rewrite

$$\sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}} \mathbb{1}\{s_j^k = s,\ a_j^k = a\}\xi^k(s,a) = \sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{j=0}^{h} \|\widehat{P}_{\ell(k)}(\cdot|s_j^k, a_j^k) - P(\cdot|s_j^k, a_j^k)\|_1,$$

then by Lemma 3.5 and $K+1 \leq 2K$, the following holds with probability at least $1-\zeta$,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}} \mathbb{1}\{s_j^k = s,\ a_j^k = a\}\xi^k(s,a) \\
&\quad \leq \sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{k=1}^{K} \sqrt{\frac{2|\mathbf{S}_{j+1}|\log(2K|\mathbf{S}||\mathbf{A}|/\zeta)}{\max\{1, N_{\ell(k)}(s_j^k, a_j^k)\}}} \\
&\quad \leq \sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{q=1}^{\ell(K)}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}} n_q(s,a)\sqrt{\frac{2|\mathbf{S}_{j+1}|\log(2K|\mathbf{S}||\mathbf{A}|/\zeta)}{\max\{1, N_q(s,a)\}}} \\
&\quad \leq \sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}} (\sqrt{2}+1)\sqrt{2\sum_{q=1}^{\ell(K)} n_q(s,a)|\mathbf{S}_{j+1}|\log\frac{2K|\mathbf{S}||\mathbf{A}|}{\zeta}},
\end{aligned}$$

where the first inequality is due to Lemma 3.5, the second inequality is by the definitions of the local counter $n_\ell(s,a)$ and the global counter $N_\ell(s,a)$, and the last inequailty is by (3.28). Thus, further bounding the last term of the above inequality yields

$$\begin{aligned}
&\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}} (\sqrt{2}+1)\sqrt{2\left[\sum_{q=1}^{\ell(K)} n_q(s,a)\right]|\mathbf{S}_{j+1}|\log\frac{2K|\mathbf{S}||\mathbf{A}|}{\zeta}} \\
&\quad \leq \sum_{h=0}^{H-1}\sum_{j=0}^{h} (\sqrt{2}+1)\sqrt{2\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\left[\sum_{q=1}^{\ell(K)} n_q(s,a)\right]|\mathbf{S}_j||\mathbf{S}_{j+1}||\mathbf{A}|\log\frac{2K|\mathbf{S}||\mathbf{A}|}{\zeta}} \\
&\quad \leq \sum_{h=0}^{H-1}\sum_{j=0}^{h} (\sqrt{2}+1)\sqrt{2K|\mathbf{S}_j||\mathbf{S}_{j+1}||\mathbf{A}|\log\frac{2K|\mathbf{S}||\mathbf{A}|}{\zeta}} \\
&\quad \leq (\sqrt{2}+1)H|\mathbf{S}|\sqrt{2K|\mathbf{A}|\log\frac{2K|\mathbf{S}||\mathbf{A}|}{\zeta}},
\end{aligned}$$

where the first inequality is due to Jensen's inequality, the second inequality is by $\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\sum_{q=1}^{\ell(K)} n_q(s,a) \leq K$, and the last inequality is by $\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sqrt{|\mathbf{S}_j||\mathbf{S}_{j+1}|} \leq \sum_{h=0}^{H-1}\sum_{j=0}^{h}(|\mathbf{S}_j|+|\mathbf{S}_{j+1}|)/2 \leq H|\mathbf{S}|$. The above results imply that with probability at least $1-\zeta$, the following inequality holds

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=0}^{H-1}\sum_{j=0}^{h}\sum_{s\in\mathbf{S}_j}\sum_{a\in\mathbf{A}}\mathbb{1}\{s_j^k=s,\ a_j^k=a\}\xi^k(s,a)\\&\qquad\leq(\sqrt{2}+1)H|\mathbf{S}|\sqrt{2K|\mathbf{A}|\log\frac{2K|\mathbf{S}||\mathbf{A}|}{\zeta}}.\end{aligned}\tag{3.29}$$

By the union bound, combining (3.26), (3.27) and (3.29), we obtain with probability at least $1-2\zeta$,

$$\sum_{k=1}^{K}\|\theta^k-\overline{\theta}^k\|_1\leq(\sqrt{2}+1)H|\mathbf{S}|\sqrt{2K|\mathbf{A}|\log\frac{2K|\mathbf{S}||\mathbf{A}|}{\zeta}}+2H^2\sqrt{2K\log\frac{H}{\zeta}}.$$

This completes the proof. □

3.7.2 Proof of Lemma 3.12

We provide Lemmas 3.21, 3.22, and 3.23 first. Then, we give the proof of Lemma 3.12 based on these lemmas.

Lemma 3.21 (Lemma 14 in Wei et al. [2019]). *Let Λ and Λ^o denote a compact convex set and the relative interior of the set Λ respectively. Assuming $\mathbf{y}\in\Lambda^o$, and letting $\mathcal{C}\subseteq\Lambda$, then the following inequality holds*

$$F(\mathbf{x}^{\mathrm{opt}})+\alpha D(\mathbf{x}^{\mathrm{opt}},\mathbf{y})\leq F(\mathbf{z})+\alpha D(\mathbf{z},\mathbf{y})-\alpha D(\mathbf{z},\mathbf{x}^{\mathrm{opt}}),\ \forall\mathbf{z}\in\mathcal{C},$$

where $\mathbf{x}^{\mathrm{opt}}\in\arg\min_{\mathbf{x}\in\mathcal{C}}F(\mathbf{x})+\alpha D(\mathbf{x},\mathbf{y})$, $F(\cdot)$ is a convex function, and $D(\cdot,\cdot)$ is the Bregman divergence.

Lemma 3.21 is an extension of Lemma 14 in Wei et al. [2019], whose proof follows the one in Wei et al. [2019]. We slightly abuse the notation of D in the above lemma and it becomes the unnormalized KL divergence when we apply this lemma in our problem, which is a special case of the Bregman divergence.

Lemma 3.22. *For any θ and θ' satisfying $\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s'\in\mathbf{S}_{h+1}}\theta(s,a,s')=1,$ and $\theta(s,a,s')\geq 0,\forall h\in\{0,\dots,H-1\}$, we let $\theta_h:=[\theta(s,a,s')]_{s\in\mathbf{S}_h,a\in\mathbf{A},s'\in\mathbf{S}_{h+1}}$ denote the vector formed by the elements $\theta(s,a,s')$ for all $(s,a,s')\in\mathbf{S}_h\times\mathbf{A}\times\mathbf{S}_{h+1}$. We also let $\theta'_h:=[\theta'(s,a,s')]_{s\in\mathbf{S}_h,a\in\mathbf{A},s'\in\mathbf{S}_{h+1}}$*

similarly denote a vector formed by $\theta'(s,a,s')$. *Then, we have*

$$D(\theta,\theta') \geq \frac{1}{2}\sum_{h=0}^{H-1}\|\theta_h-\theta'_h\|_1^2 \geq \frac{1}{2H}\|\theta-\theta'\|_1^2,$$

where $D(\cdot,\cdot)$ *is defined as in* (3.8).

Proof. We prove the lemma by the following inequality

$$\begin{aligned}
D(\theta,\theta') &= \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}}\theta(s,a,s')\frac{\theta(s,a,s')}{\theta'(s,a,s')} - \theta(s,a,s') + \theta'(s,a,s') \\
&= \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}}\theta(s,a,s')\frac{\theta(s,a,s')}{\theta'(s,a,s')} \\
&\geq \frac{1}{2}\sum_{h=0}^{H-1}\|\theta_h-\theta'_h\|_1^2 \geq \frac{1}{2H}\bigg(\sum_{h=0}^{H-1}\|\theta_h-\theta'_h\|_1\bigg)^2 \geq \frac{1}{2H}\|\theta-\theta'\|_1^2,
\end{aligned}$$

where the inequality is due to the Pinsker's inequality since θ_h and θ'_h are two probability distributions such that $\|\theta_h\|_1 = 1$ and $\|\theta'_h\|_1 = 1$. This completes the proof. □

Lemma 3.23. *For any* θ *and* θ' *satisfying* $\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s'\in\mathbf{S}_{h+1}}\theta(s,a,s') = 1,$ *and* $\theta(s,a,s') \geq 0, \forall h \in \{0,\ldots,H-1\}$ *, letting* $\widetilde{\theta}'(s,a,s') = (1-\lambda)\theta'(s,a,s') + \frac{\lambda}{|\mathbf{A}||\mathbf{S}_h||\mathbf{S}_{h+1}|}, \forall(s,a,s') \in \mathbf{S}_h \times \mathbf{A} \times \mathbf{S}_{h+1}, \forall h = 1,\ldots,H-1$ *with* $0 < \lambda \leq 1$*, then we have*

$$D(\theta,\widetilde{\theta}') - D(\theta,\theta') \leq \lambda H \log|\mathbf{S}|^2|\mathbf{A}|, \quad D(\theta,\widetilde{\theta}') \leq H\log\frac{|\mathbf{S}|^2|\mathbf{A}|}{\lambda},$$

where $D(\cdot,\cdot)$ *is defined as in* (3.8).

Proof. We start our proof as follows

$$\begin{aligned}
D(\theta,\widetilde{\theta}') - D(\theta,\theta') &= \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}}\theta(s,a,s')\Big(\log\frac{\theta(s,a,s')}{\widetilde{\theta}'(s,a,s')} - \log\frac{\theta(s,a,s')}{\theta'(s,a,s')}\Big) \\
&\quad + \widetilde{\theta}'(s,a,s') - \theta'(s,a,s') \\
&= \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}}\theta(s,a,s')\big(\log\theta'(s,a,s') - \log\widetilde{\theta}'(s,a,s')\big) \\
&= \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}}\theta(s,a,s')\Big(\log\theta'(s,a,s') \\
&\quad - \log[(1-\lambda)\theta'(s,a,s') + \lambda/(|\mathbf{S}_h||\mathbf{S}_{h+1}||\mathbf{A}|)]\Big),
\end{aligned}$$

where the last equality is by substituting $\widetilde{\theta}'(s,a,s') = (1-\lambda)\theta'(s,a,s') + \frac{\lambda}{|\mathbf{A}||\mathbf{S}_h||\mathbf{S}_{h+1}|}, \forall (s,a,s') \in \mathbf{S}_h \times \mathbf{A} \times \mathbf{S}_{h+1}, \forall h = 1,\ldots,H-1$. Thus, by bounding the last term above, we further have

$$
\begin{aligned}
D(\theta,\widetilde{\theta}') - D(\theta,\theta') \le & \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}} \theta(s,a,s')\bigg(\log\theta'(s,a,s') \\
& - (1-\lambda)\log\theta'(s,a,s') - \lambda\log\frac{1}{|\mathbf{S}_h||\mathbf{S}_{h+1}||\mathbf{A}|}\bigg) \\
= & \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}} \lambda\theta(s,a,s')\big(\log\theta'(s,a,s') + \log(|\mathbf{S}_h||\mathbf{S}_{h+1}||\mathbf{A}|)\big) \\
\le & \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}} \lambda\theta(s,a,s')\log(|\mathbf{S}_h||\mathbf{S}_{h+1}||\mathbf{A}|) \le \lambda H \log|\mathbf{S}|^2|\mathbf{A}|,
\end{aligned}
$$

where the first inequality is by Jensen's inequality and the second inequality is due to $\log\theta'(s,a,s') \le 0$ since $0 < \theta'(s,a,s') < 1$, and the last inequality is due to Hölder's inequality that $\langle \mathbf{x},\mathbf{y}\rangle \le \|\mathbf{x}\|_1\|\mathbf{y}\|_\infty$ and $|\mathbf{S}_h||\mathbf{S}_{h+1}| \le |\mathbf{S}|^2$.

Moreover, we have

$$
\begin{aligned}
D(\theta,\widetilde{\theta}') = & \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}} \theta(s,a,s')\log\frac{\theta(s,a,s')}{\widetilde{\theta}'(s,a,s')} - \theta(s,a,s') + \theta'(s,a,s') \\
= & \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}} \theta(s,a,s')\big(\log\theta(s,a,s') - \log\widetilde{\theta}'(s,a,s')\big) \\
\le & -\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}} \theta(s,a,s')\log\widetilde{\theta}'(s,a,s') \\
= & -\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}} \theta(s,a,s')\big(\log[(1-\lambda)\theta'(s,a,s') + \lambda/(|\mathbf{S}_h||\mathbf{S}_{h+1}||\mathbf{A}|)]\big) \\
\le & -\sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s\in\mathbf{S}_{h+1}} \theta(s,a,s')\cdot\log\frac{\lambda}{|\mathbf{S}_h||\mathbf{S}_{h+1}||\mathbf{A}|} \le H\log\frac{|\mathbf{S}|^2|\mathbf{A}|}{\lambda},
\end{aligned}
$$

where the first inequality is due to $\log\theta(s,a,s') \le 0$, the second inequality is due to the monotonicity of logarithm function, and the third inequality is by as well as $|\mathbf{S}_h||\mathbf{S}_{h+1}| \le |\mathbf{S}|^2$. This completes the proof. □

Now we are ready to provide the proof of Lemma 3.12.

Proof of Lemma 3.12. First of all, by Lemma 3.5, we know that

$$\|P(\cdot|s,a) - \widehat{P}_\ell(\cdot|s,a)\|_1 \le \varepsilon_\ell^\zeta(s,a),$$

with probability at least $1-\zeta$, for all epochs ℓ and any state and action pair $(s,a) \in \mathbf{S} \times \mathbf{A}$. Thus, we have that for any epoch $\ell \le \ell(K+1)$,

$$\Delta \subseteq \Delta(\ell, \zeta)$$

holds with probability at least $1-\zeta$.

This can be easily proved in the following way: if any $\overline{\theta} \in \Delta$, then for all $h \in \{0, \dots, H-1\}$, $s \in \mathbf{S}_h$, and $a \in \mathbf{A}$, we have

$$\frac{\overline{\theta}(s,a,\cdot)}{\sum_{s' \in \mathbf{S}_{h+1}} \overline{\theta}(s,a,s')} = P(\cdot|s,a).$$

Then, we obtain with probability at least $1-\zeta$,

$$\left\| \frac{\overline{\theta}(s,a,\cdot)}{\sum_{s' \in \mathbf{S}_{h+1}} \overline{\theta}(s,a,s')} - \widehat{P}_\ell(\cdot|s,a) \right\|_1 = \left\| P(\cdot|s,a) - \widehat{P}_\ell(\cdot|s,a) \right\|_1 \le \varepsilon_\ell^\zeta(s,a).$$

where the last inequality is by Lemma 3.5. Therefore, we know that $\overline{\theta} \in \Delta(\ell,\zeta)$, which proves the above claim.

Thus, we define the following event

$$\text{Event } \mathcal{D}_K : \Delta \subseteq \cap_{\ell=1}^{\ell(K+1)} \Delta(\ell, \zeta), \tag{3.30}$$

by which we have

$$\Pr(\mathcal{D}_K) \ge 1 - \zeta.$$

For any $\overline{\theta}^*$ which is a solution to problem (3.2), we have $\overline{\theta}^* \in \Delta$. If event $\mathcal{D}_K$ happens, then $\overline{\theta}^* \in \cap_{\ell=1}^{\ell(K+1)} \Delta(\ell,\zeta)$. Now we have that the updating rule of θ follows $\theta^k = \arg\min_{\theta \in \Delta(\ell(k),\zeta)} \langle V f^{k-1} + \sum_{i=1}^I Q_i(k-1) g_i^{k-1}, \theta \rangle + \alpha D(\theta, \widetilde{\theta}^{k-1})$ as shown in (3.7), and also $\overline{\theta}^* \in \cap_{\ell=1}^{\ell(K+1)} \Delta(\ell,\zeta)$ holds with probability at least $1-\zeta$. According to Lemma 3.21, letting $\mathbf{x}^{\text{opt}} = \theta^k$, $\mathbf{z} = \overline{\theta}^*$, $\mathbf{y} = \widetilde{\theta}^{k-1}$ and $F(\theta) = \langle V f^{k-1} + \sum_{i=1}^I Q_i(k-1) g_i^{k-1}, \theta \rangle$, we have that with probability at least $1-\zeta$, the following inequality holds for episodes $2 \leq k \leq K+1$

$$\begin{aligned}&\left\langle Vf^{k-1} + \sum_{i=1}^{I} Q_i(k-1) g_i^{k-1}, \theta^k \right\rangle + \alpha D(\theta^k, \widetilde{\theta}^{k-1}) \\ &\qquad \leq \left\langle Vf^{k-1} + \sum_{i=1}^{I} Q_i(k-1) g_i^{k-1}, \overline{\theta}^* \right\rangle + \alpha D(\overline{\theta}^*, \widetilde{\theta}^{k-1}) - \alpha D(\overline{\theta}^*, \theta^k).\end{aligned} \tag{3.31}$$

Thus, once given the event $\mathcal{D}_K$ happens, the inequality (3.31) will hold.

On the other hand, according to the updating rule of $\mathbf{Q}(\cdot)$ in (3.6), which is $Q_i(k) = \max\{Q_i(k-1) + \langle g_i^{k-1}, \theta^k \rangle - c_i,\ 0\}$, we know that

$$Q_i(k)^2 = \left(\max\{Q_i(k-1) + \left\langle g_i^{k-1}, \theta^k \right\rangle - c_i,\ 0\}\right)^2 \leq \left(Q_i(k-1) + \left\langle g_i^{k-1}, \theta^k \right\rangle - c_i\right)^2,$$

which further leads to

$$Q_i(k)^2 - Q_i(k-1)^2 \leq 2Q_i(k-1)\left(\left\langle g_i^{k-1}, \theta^k \right\rangle - c_i\right) + \left(\left\langle g_i^{k-1}, \theta^k \right\rangle - c_i\right)^2.$$

Taking summation on both sides of the above inequality from $i=1$ to I, we have

$$\begin{aligned}&\frac{1}{2}\left(\|\mathbf{Q}(k)\|_2^2 - \|\mathbf{Q}(k-1)\|_2^2\right) \\ &\qquad \leq \sum_{i=1}^{I} \left\langle Q_i(k-1) g_i^{k-1}, \theta^k \right\rangle - \sum_{i=1}^{I} Q_i(k-1) c_i + \frac{1}{2}\sum_{i=1}^{I}\left(\left\langle g_i^{k-1}, \theta^k \right\rangle - c_i\right)^2 \\ &\qquad \leq \sum_{i=1}^{I} \left\langle Q_i(k-1) g_i^{k-1}, \theta^k \right\rangle - \sum_{i=1}^{I} Q_i(k-1) c_i + 2H^2,\end{aligned} \tag{3.32}$$

where we let $\|\mathbf{Q}(k)\|_2^2 = \sum_{i=1}^{I} Q_i^2(k)$ and $\|\mathbf{Q}(k-1)\|_2^2 = \sum_{i=1}^{I} Q_i^2(k-1)$, and the last inequality is due to

$$\begin{aligned}&\sum_{i=1}^{I}(\langle g_i^{k-1}, \theta^k \rangle - c_i)^2 \\ &\qquad \leq 2\sum_{i=1}^{I}[(\langle g_i^{k-1}, \theta^k \rangle)^2 + c_i^2] \leq 2\sum_{i=1}^{I}[\|g_i^{k-1}\|_\infty^2 \|\theta^k\|_1^2 + c_i^2] \\ &\qquad \leq 2\sum_{i=1}^{I}[H^2\|g_i^{k-1}\|_\infty^2 + c_i^2] \leq 2[H^2(\sum_{i=1}^{I}\|g_i^{k-1}\|_\infty)^2 + (\sum_{i=1}^{I}|c_i|)^2] \leq 4H^2\end{aligned}$$

by Assumption 3.4 and the facts that $\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s'\in\mathbf{S}_{h+1}} \theta^k(s,a,s') = 1$ and $\theta^k(s,a,s') \geq 0$.

Thus, summing up (3.31) and (3.32), and then subtracting $\langle V f^{k-1}, \theta^{k-1}\rangle$ from both sides, we have

$$
\begin{aligned}
&V\big\langle f^{k-1}, \theta^k - \theta^{k-1}\big\rangle + \frac{1}{2}\big(\|\mathbf{Q}(k)\|_2^2 - \|\mathbf{Q}(k-1)\|_2^2\big) + \alpha D(\theta^k, \widetilde{\theta}^{k-1})\\
&\quad \le V\big\langle f^{k-1}, \overline{\theta}^* - \theta^{k-1}\big\rangle + \sum_{i=1}^{I} Q_i(k-1)(\langle g_i^{k-1}, \overline{\theta}^*\rangle - c_i) + \alpha D(\overline{\theta}^*, \widetilde{\theta}^{k-1}) - \alpha D(\overline{\theta}^*, \theta^k) + 4H^2.
\end{aligned}
$$

We further need to show the lower bound of the term $V\big\langle f^{k-1}, \theta^k - \theta^{k-1}\big\rangle + \alpha D(\theta^k, \widetilde{\theta}^{k-1})$ on the left-hand side of the above inequality. Specifically, we have

$$
\begin{aligned}
&V\big\langle f^{k-1}, \theta^k - \theta^{k-1}\big\rangle + \alpha D(\theta^k, \widetilde{\theta}^{k-1})\\
&\quad = V\big\langle f^{k-1}, \theta^k - \widetilde{\theta}^{k-1}\big\rangle + V\big\langle f^{k-1}, \widetilde{\theta}^{k-1} - \theta^{k-1}\big\rangle + \alpha D(\theta^k, \widetilde{\theta}^{k-1})\\
&\quad \ge -V\|f^{k-1}\|_\infty \cdot \|\theta^k - \widetilde{\theta}^{k-1}\|_1 - V\|f^{k-1}\|_\infty \cdot \|\widetilde{\theta}^{k-1} - \theta^{k-1}\|_1 + \frac{\alpha}{2}\sum_{h=0}^{H-1} \|\theta_h^k - \widetilde{\theta}_h^{k-1}\|_1^2\\
&\quad \ge -V\sum_{h=0}^{H-1} \|\theta_h^k - \widetilde{\theta}_h^{k-1}\|_1 - 2H\lambda V + \frac{\alpha}{2}\sum_{h=0}^{H-1} \|\theta_h^k - \widetilde{\theta}_h^{k-1}\|_1^2\\
&\quad \ge -\frac{HV}{2\alpha} - 2H\lambda V,
\end{aligned}
$$

where the first inequality uses Hölder's inequality and Lemma 3.22 that $D(\theta, \theta') = \sum_{h=0}^{H-1} D(\theta_h, \theta'_h) \ge \frac{1}{2}\sum_{h=0}^{H-1} \|\theta_h - \theta'_h\|_1^2$ with $\theta_h := [\theta(s, a, s')]_{s\in\mathbf{S}_h, a\in\mathbf{A}, s'\in\mathbf{S}_{h+1}}$, the second inequality is due to $\widetilde{\theta}_h^{k-1} = (1-\lambda)\theta_h^{k-1} + \lambda\frac{\mathbf{1}}{|\mathbf{A}||\mathbf{S}_h||\mathbf{S}_{h+1}|}$, the second inequality is due to $\|\widetilde{\theta}^{k-1} - \theta^{k-1}\|_1 = \sum_{h=0}^{H-1} \|\widetilde{\theta}_h^{k-1} - \theta_h^{k-1}\|_1 = \lambda\sum_{h=0}^{H-1} \big\|\theta_h^{k-1} - \frac{\mathbf{1}}{|\mathbf{A}||\mathbf{S}_h||\mathbf{S}_{h+1}|}\big\|_1 \le \lambda\sum_{h=0}^{H-1} \big(\big\|\theta_h^{k-1}\big\|_1 + \big\|\frac{\mathbf{1}}{|\mathbf{A}||\mathbf{S}_h||\mathbf{S}_{h+1}|}\big\|_1\big) \le 2\lambda H$, and the third inequality is by finding the minimal value of a quadratic function $-Vx + \frac{\alpha}{2}x^2$.

Therefore, one can show that with probability at least $1-\zeta$, the following inequality holds for all $k \le K+1$,

$$
\begin{aligned}
&\frac{1}{2}\big(\|\mathbf{Q}(k)\|_2^2 - \|\mathbf{Q}(k-1)\|_2^2\big) - \frac{HV}{2\alpha} - 2H\lambda V \qquad (3.33)\\
&\quad \le V\big\langle f^{k-1}, \overline{\theta}^* - \theta^{k-1}\big\rangle + \sum_{i=1}^{I} Q_i(k-1)(\langle g_i^{k-1}, \overline{\theta}^*\rangle - c_i) + \alpha D(\overline{\theta}^*, \widetilde{\theta}^{k-1}) - \alpha D(\overline{\theta}^*, \theta^k) + 4H^2.
\end{aligned}
$$

Note that according to Lemma 3.23, we have

$$
\begin{aligned}
D(\overline{\theta}^*, \widetilde{\theta}^{k-1}) - D(\overline{\theta}^*, \theta^k) &= D(\overline{\theta}^*, \widetilde{\theta}^{k-1}) - D(\overline{\theta}^*, \theta^{k-1}) + D(\overline{\theta}^*, \theta^{k-1}) - D(\overline{\theta}^*, \theta^k)\\
&\le \lambda H \log|\mathbf{S}|^2|\mathbf{A}| + D(\overline{\theta}^*, \theta^{k-1}) - D(\overline{\theta}^*, \theta^k).
\end{aligned}
$$

Therefore, plugging the above inequality into (3.33) and rearranging the terms, we further get

$$V \left\langle f^{k-1}, \theta^{k-1} - \overline{\theta}^* \right\rangle \leq \frac{1}{2}\left(\|\mathbf{Q}(k-1)\|_2^2 - \|\mathbf{Q}(k)\|_2^2\right) + \sum_{i=1}^{I} Q_i(k-1)(\langle g_i^{k-1}, \overline{\theta}^* \rangle - c_i) + 4H^2 + \alpha\lambda H \log |\mathbf{S}|^2|\mathbf{A}| + \alpha D(\overline{\theta}^*, \theta^{k-1}) - \alpha D(\overline{\theta}^*, \theta^k) + \frac{HV}{2\alpha} + 2H\lambda V.$$

Thus, taking summation on both sides of the above inequality from 2 to $K+1$, by $\mathbf{Q}(1) = \mathbf{0}$, we obtain that with probability at least $1-\zeta$,

$$\sum_{k=2}^{K+1} \left\langle f^{k-1}, \theta^{k-1} - \overline{\theta}^* \right\rangle \leq \frac{1}{V} \sum_{k=2}^{K+1} \sum_{i=1}^{I} Q_i(k-1)(\langle g_i^{k-1}, \overline{\theta}^* \rangle - c_i) + \frac{K\alpha\lambda H \log |\mathbf{S}|^2|\mathbf{A}|}{V} + \frac{\alpha D(\overline{\theta}^*, \theta^1) + 4H^2 K}{V} + \frac{KH}{2\alpha} + 2H\lambda K. \tag{3.34}$$

It is not difficult to compute that $D(\overline{\theta}^*, \theta^1) \leq H \log |\mathbf{S}|^2|\mathbf{A}|$ according to the initialization of θ^1 by the uniform distribution. Rearranging the terms and shifting the index, we rewrite (3.34) as

$$\sum_{k=1}^{K} \left\langle f^{k}, \theta^{k} - \overline{\theta}^* \right\rangle \leq \frac{1}{V} \sum_{k=1}^{K} \sum_{i=1}^{I} Q_i(k)(\langle g_i^{k}, \overline{\theta}^* \rangle - c_i) + \frac{4H^2 K + (\lambda K + 1)\alpha H \log |\mathbf{S}|^2|\mathbf{A}|}{V} + \frac{KH}{2\alpha} + 2H\lambda K.$$

This completes the proof. □

3.7.3 Proof of Lemma 3.13

We first provide Lemmas 3.24 below. Then, we give the proof of Lemma 3.13 based on this lemma.

Lemma 3.24 (Lemma 5 of Yu et al. [2017]). *Let $\{Z(k), k \geq 0\}$ be a discrete time stochastic process adapted to a filtration $\{\mathcal{U}^k, k \geq 0\}$ with $Z(0) = 0$ and $\mathcal{U}^0 = \{\varnothing, \Omega\}$. Suppose there exists an integer $\tau > 0$, real constants $\theta > 0$, $\rho_{\max} > 0$ and $0 < \kappa \leq \rho_{\max}$ such that*

$$|Z(k+1) - Z(k)| \leq \rho_{\max},$$
$$\mathbb{E}[Z(k+\tau) - Z(k) \,|\, \mathcal{U}^k] \leq \begin{cases} \tau\rho_{\max}, & \textit{if } Z(k) < \psi \\ -\tau\kappa, & \textit{if } Z(k) \geq \psi \end{cases}$$

hold for all $k \in \{1, 2, ...\}$. Then for any constant $0 < \delta < 1$, with probability at least $1-\delta$, we *have*

$$Z(k) \leq \psi + \tau\frac{4\rho_{\max}^2}{\kappa}\log\left(1+\frac{8\rho_{\max}^2}{\kappa^2}e^{\kappa/(4\rho_{\max})}\right) + \tau\frac{4\rho_{\max}^2}{\kappa}\log\frac{1}{\delta},\ \forall k \in \{1,2,...\}.$$

Now, we are in position to give the proof of Lemma 3.13.

Proof of Lemma 3.13. The proof of this Lemma is based on applying the lemma 3.24 to our problem. Thus, this proof mainly focuses on showing that the variable $\|\mathbf{Q}(k)\|_2$ satisfies the condition of Lemma 3.24.

According to the updating rule of $Q_i(k)$, which is $Q_i(k+1) = \max\{Q_i(k) + \langle g_i^k, \theta^{k+1}\rangle - c_i, 0\}$, we have

$$\begin{aligned}|\|\mathbf{Q}(k+1)\|_2 - \|\mathbf{Q}(k)\|_2| \leq& \|\mathbf{Q}(k+1) - \mathbf{Q}(k)\|_2 \\ =& \sqrt{\sum_{i=1}^{I}|Q_i(k+1) - Q_i(k)|^2} \leq \sqrt{\sum_{i=1}^{I}|\langle g_i^k, \theta^{k+1}\rangle - c_i|^2},\end{aligned}$$

where the first inequality is due to triangle inequality, and the second inequality is by the fact that $|\max\{a+b,0\} - a| \leq |b|$ if $a \geq 0$. Then, by Assumption 3.4, we further have

$$\sqrt{\sum_{i=1}^{I}|\langle g_i^k, \theta^{k+1}\rangle - c_i|^2} \leq \sum_{i=1}^{I}|\langle g_i^k, \theta^{k+1}\rangle - c_i| \leq \sum_{i=1}^{I}(\|g_i^k\|_\infty\|\theta^{k+1}\|_1 + |c_i|) \leq 2H,$$

which therefore implies

$$|\|\mathbf{Q}(k+1)\|_2 - \|\mathbf{Q}(k)\|_2| \leq 2H. \tag{3.35}$$

Thus, with the above inequality, we have

$$\begin{aligned}\|\mathbf{Q}(k+\tau)\|_2 - \|\mathbf{Q}(k)\|_2 \leq& |\|\mathbf{Q}(k+\tau)\|_2 - \|\mathbf{Q}(k)\|_2| \\ \leq& \sum_{\tau=1}^{\tau}|\|\mathbf{Q}(k+\tau)\|_2 - \|\mathbf{Q}(k+\tau-1)\|_2| \leq 2\tau H,\end{aligned} \tag{3.36}$$

such that

$$\mathbb{E}[\|\mathbf{Q}(k+\tau)\|_2 - \|\mathbf{Q}(k)\|_2|\mathcal{F}^{k-1}] \leq 2\tau H, \tag{3.37}$$

where $\mathcal{F}^{k-1}$ represents the system randomness up to the $(k-1)$-th episode and $\mathbf{Q}(k)$ depends on $\mathcal{F}^{k-1}$ according to its updating rule.

Next, we need to show that there exist ψ and κ such that $\mathbb{E}[\|\mathbf{Q}(k+\tau)\|_2 - \|\mathbf{Q}(k)\|_2|\mathcal{F}^{k-1}] \leq -\tau\kappa$ if $\|\mathbf{Q}(k)\|_2 \geq \psi$. Recall the definition of the event $\mathcal{D}_K$ in (3.30). Therefore, we have that with probability at least $1-\zeta$, the event $\mathcal{D}_K$ happens, such that for all $2 \leq k' \leq K+1$ and any $\theta \in \cap_{\ell=1}^{\ell(K+1)} \Delta(\ell, \zeta)$, the following holds

$$V\big\langle f^{k'-1}, \theta^{k'-1} - \overline{\theta}^*\big\rangle \leq \frac{1}{2}\left(\|\mathbf{Q}(k'-1)\|_2^2 - \|\mathbf{Q}(k')\|_2^2\right) + \sum_{i=1}^{I} Q_i(k'-1)(\langle g_i^{k'-1}, \theta\rangle - c_i) \\ + \alpha\lambda H \log|\mathbf{S}|^2|\mathbf{A}| + \alpha D(\theta, \widetilde{\theta}^{k'-1}) - \alpha D(\theta, \theta^{k'}) + 4H^2 + \frac{HV}{2\alpha} + 2H\lambda V,$$

which adopts similar proof techniques to (3.33). Then, by rearranging the terms, the above inequality further leads to the following inequality

$$\|\mathbf{Q}(k')\|_2^2 - \|\mathbf{Q}(k'-1)\|_2^2 \leq -2V\big\langle f^{k'-1}, \theta^{k'-1} - \theta\big\rangle + 2\sum_{i=1}^{I} Q_i(k'-1)(\langle g_i^{k'-1}, \theta\rangle - c_i) \\ + 2\alpha\lambda H \log|\mathbf{S}|^2|\mathbf{A}| + 2\alpha D(\theta, \widetilde{\theta}^{k'-1}) - 2\alpha D(\theta, \theta^{k'}) + 8H^2 + \frac{HV}{\alpha} + 4H\lambda V.$$

Taking summation from $k+1$ to $\tau+k$ on both sides of the above inequality, the following inequality holds with probability $1-\zeta$ for any $\tau > 0$ and k satisfying $1 \leq k \leq K+1-\tau$,

$$\begin{aligned} &\|\mathbf{Q}(\tau+k)\|_2^2 - \|\mathbf{Q}(k)\|_2^2 \\ &\leq -2V\sum_{k'=k+1}^{\tau+k}\big\langle f^{k'-1}, \theta^{k'-1} - \theta\big\rangle + 2\sum_{k'=k+1}^{\tau+k}\sum_{i=1}^{I} Q_i(k'-1)(\langle g_i^{k'-1}, \theta\rangle - c_i) + 2\alpha D(\theta, \widetilde{\theta}^{k}) \\ &\quad - 2\alpha D(\theta, \widetilde{\theta}^{\tau+k}) + \sum_{k'=k+1}^{\tau+k} 2\alpha[D(\theta, \widetilde{\theta}^{k'-1}) - D(\theta, \theta^{k'-1})] + 8\tau H^2 + \frac{\tau HV}{\alpha} + 4\tau H\lambda V. \end{aligned} \tag{3.38}$$

Particularly, in (3.38), the term $-2\alpha D(\theta, \theta^{k'-1}) \leq 0$ due to the non-negativity of unnormalized KL divergence. By Lemma 3.23, we have

$$\sum_{\tau=t+1}^{\tau+k} 2\alpha[D(\theta, \widetilde{\theta}^{k'-1}) - D(\theta, \theta^{k'-1})] \leq 2\alpha\tau H \log|\mathbf{S}|^2|\mathbf{A}|.$$

For the term $2\alpha D(\theta, \widetilde{\theta}^k)$, by Lemma 3.23, we can bound it as

$$2\alpha D(\theta, \widetilde{\theta}^k) \leq 2\alpha H \log(|\mathbf{S}|^2|\mathbf{A}|/\lambda).$$

Moreover, we can decompose the term $2V\sum_{k'=k+1}^{\tau+k}\langle f^{k'-1}, \theta - \theta^{k'-1}\rangle + 2\sum_{k'=k+1}^{\tau+k}\sum_{i=1}^{I} Q_i(k' -$

$1)(\langle g_i^{k'-1}, \overline{\theta}^* \rangle - c_i)$ in (3.38) as

$$
\begin{aligned}
&2V \sum_{k'=k+1}^{\tau+k} \left\langle f^{k'-1}, \theta - \theta^{k'-1} \right\rangle + 2 \sum_{k'=k+1}^{\tau+k} \sum_{i=1}^{I} Q_i(k'-1)(\langle g_i^{k'-1}, \theta \rangle - c_i) \\
&\quad = 2V \sum_{k'=k+1}^{\tau+k} \left\langle f^{k'-1}, \theta - \theta^{k'-1} \right\rangle + 2 \sum_{i=1}^{I} Q_i(k) \sum_{k'=k+1}^{\tau+k} (\langle g_i^{k'-1}, \theta \rangle - c_i) \\
&\qquad + 2 \sum_{k'=k+2}^{\tau+k} \sum_{i=1}^{I} [Q_i(k'-1) - Q_i(k)](\langle g_i^{k'-1}, \theta \rangle - c_i) \\
&\quad \leq 2V \sum_{k'=k+1}^{\tau+k} \left\langle f^{k'-1}, \theta \right\rangle + 2 \sum_{i=1}^{I} Q_i(k) \sum_{k'=k+1}^{\tau+k} (\langle g_i^{k'-1}, \theta \rangle - c_i) + 2H\tau^2 + 2VH\tau,
\end{aligned}
$$

where the last inequality is due to

$$
-2V \sum_{k'=k+1}^{\tau+k} \left\langle f^{k'-1}, \theta^{k'-1} \right\rangle \leq 2V \sum_{k'=k+1}^{\tau+k} \sum_{h=0}^{H-1} \sum_{s \in \mathbf{S}_h} \sum_{a \in \mathbf{A}} \sum_{s' \in \mathbf{S}_{h+1}} f^{k'-1}(s,a,s') \theta^{k'-1}(s,a,s') \leq 2VH\tau
$$

as well as

$$
\begin{aligned}
&2 \sum_{k'=k+2}^{\tau+k} \sum_{i=1}^{I} [Q_i(k'-1) - Q_i(k)](\langle g_i^{k'-1}, \theta \rangle - c_i) \\
&\quad \leq 2 \sum_{k'=k+2}^{\tau+k} \sum_{i=1}^{I} \sum_{r=k}^{k'-2} |\langle g_i^r, \theta^{r+1} \rangle - c_i| \cdot |\langle g_i^{k'-1}, \theta \rangle - c_i| \\
&\quad \leq \sum_{k'=k+2}^{\tau+k} \sum_{r=k}^{k'-2} \sqrt{\sum_{i=1}^{I} |\langle g_i^r, \theta^{r+1} \rangle - c_i|^2} + \sum_{k'=k+2}^{\tau+k} \sum_{r=k}^{k'-2} \sqrt{\sum_{i=1}^{I} |\langle g_i^{k'-1}, \theta \rangle - c_i|^2} \leq 2H\tau^2
\end{aligned}
$$

by $Q_i(k+1) = \max\{Q_i(k) + \langle g_i^k, \theta^{k+1} \rangle - c_i, 0\}$ and $|\max\{a+b, 0\} - a| \leq |b|$ if $a \geq 0$ for the first inequality and Assumption 3.4 for the last inequality. Taking conditional expectation on both sides of (3.38) and combining the above bounds for terms in (3.38), we have for any $\theta \in \cap_{\ell=1}^{\ell(K+1)} \Delta(\ell, \zeta)$,

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{Q}(\tau+k)\|_2^2 - \|\mathbf{Q}(k)\|_2^2 | \mathcal{F}^{k-1}, \mathcal{D}_K] \\
&\quad \leq 2\tau^2 H + 2\alpha H \log(|\mathbf{S}|^2 |\mathbf{A}|/\lambda) \\
&\qquad + 2V\tau \mathbb{E}\left[\frac{1}{\tau} \sum_{k'=k+1}^{\tau+k} \langle f^{k'-1}, \theta \rangle + \frac{1}{\tau} \sum_{i=1}^{I} \frac{Q_i(k)}{V} \sum_{k'=k+1}^{\tau+k} (\langle g_i^{k'-1}, \theta \rangle - c_i) \middle| \mathcal{F}^{k-1}, \mathcal{D}_K \right] \qquad (3.39) \\
&\qquad + 2\alpha\lambda\tau H \log|\mathbf{S}|^2|\mathbf{A}| + 8\tau H^2 + \frac{\tau H V}{\alpha} + 4\tau H \lambda V + 2VH\tau.
\end{aligned}
$$

Thus, it remains to bound the term $\mathbb{E}[\frac{1}{\tau}\sum_{k'=k+1}^{\tau+k}\langle f^{k'-1},\theta\rangle + \frac{1}{\tau}\sum_{i=1}^{I}\frac{Q_i(k)}{V}\sum_{k'=k+1}^{\tau+k}(\langle g_i^{k'-1},\theta\rangle - c_i)|\mathcal{F}^{k-1},\mathcal{D}_K]$ so as to give an upper bound of the right-hand side of (3.39). Given the event $\mathcal{D}_K$ happens such that $\Delta \subseteq \cap_{\ell=1}^{\ell(K+1)}\Delta(\ell,\zeta) \neq \varnothing$, and since θ is any vector in the set $\cap_{\ell=1}^{\ell(K+1)}\Delta(\ell,\zeta)$, we can give an upper bound of (3.39) by bounding a term $q^{(k,\tau)}\Big(\frac{\mathbf{Q}(k)}{V}\Big)$, which is due to

$$\begin{aligned}
&\min_{\theta\in\cap_{\ell=1}^{\ell(K+1)}\Delta(\ell,\zeta)} \mathbb{E}\Big[\frac{1}{\tau}\sum_{k'=k+1}^{\tau+k}\big\langle f^{k'-1},\theta\big\rangle + \frac{1}{\tau}\sum_{i=1}^{I}\frac{Q_i(k)}{V}\sum_{k'=k+1}^{\tau+k}(\langle g_i^{k'-1},\theta\rangle - c_i)\Big|\mathcal{F}^{k-1},\mathcal{D}_K\Big]\\
&\quad= \min_{\theta\in\cap_{\ell=1}^{\ell(K+1)}\Delta(\ell,\zeta)}\big\langle f^{(k,\tau)},\theta\big\rangle + \sum_{i=1}^{I}\frac{Q_i(k)}{V}(\langle g_i,\theta\rangle - c_i)\\
&\quad\le \min_{\theta\in\Delta}\big\langle f^{(k,\tau)},\theta\big\rangle + \sum_{i=1}^{I}\frac{Q_i(k)}{V}(\langle g_i,\theta\rangle - c_i) = q^{(k,\tau)}\Big(\frac{\mathbf{Q}(k)}{V}\Big),
\end{aligned}$$

where the inequality is due to $\Delta \subseteq \cap_{\ell=1}^{\ell(K+1)}\Delta(\ell,\zeta)$ given $\mathcal{D}_K$ happens and the last equality is obtained according to the definition of the dual function q in Section 3.4. Next, we bound $q^{(k,\tau)}\big(\frac{\mathbf{Q}(k)}{V}\big)$.

According to Assumption 3.6, we assume that one dual solution is $\eta_{k,\tau}^* \in \mathcal{V}_{k,\tau}^*$. We let $\overline{\vartheta}$ be the maximum of all ϑ and $\overline{\sigma}$ be the minimum of all σ. Thus, when $\mathrm{dist}(\frac{\mathbf{Q}(k)}{V},\mathcal{V}_{k,\tau}^*) \ge \overline{\vartheta}$, we have

$$\begin{aligned}
q^{(k,\tau)}\Big(\frac{\mathbf{Q}(k)}{V}\Big) =& q^{(k,\tau)}\Big(\frac{\mathbf{Q}(k)}{V}\Big) - q^{(k,\tau)}(\eta_{k,\tau}^*) + q^{(k,\tau)}(\eta_{k,\tau}^*)\\
\le& -\overline{\sigma}\Big\|\eta_{k,\tau}^* - \frac{\mathbf{Q}(k)}{V}\Big\|_2 + \big\langle f^{(k,\tau)},\theta_{k,\tau}^*\big\rangle\\
\le& -\overline{\sigma}\Big\|\frac{\mathbf{Q}(k)}{V}\Big\|_2 + \overline{\sigma}\|\eta_{k,\tau}^*\|_2 + \sum_{h=0}^{H-1}\sum_{s\in\mathbf{S}_h}\sum_{a\in\mathbf{A}}\sum_{s'\in\mathbf{S}_{h+1}} f^{(k,\tau)}(s,a,s')\theta_{k,\tau}^*(s,a,s')\\
\le& -\overline{\sigma}\Big\|\frac{\mathbf{Q}(k)}{V}\Big\|_2 + \overline{\sigma}B + H,
\end{aligned}$$

where the first inequality is due to the error bound condition in Lemma 3.9 and the weak duality relation $q^{(k,\tau)}(\eta_{k,\tau}^*) \le \big\langle f^{(k,\tau)},\theta_{k,\tau}^*\big\rangle$ for the Lagrangian duality (see, e.g., Bertsekas [2009]) with $\theta_{k,\tau}^*$ being a primal solution, the second inequality is by triangle inequality, and the third inequality is by Assumption 3.4 and Assumption 3.6. On the other hand, when $\mathrm{dist}(\frac{\mathbf{Q}(k)}{V},\mathcal{V}_{k,\tau}^*) \le \overline{\vartheta}$, we have

$$\begin{aligned}
q^{(k,\tau)}\Big(\frac{\mathbf{Q}(k)}{V}\Big) =& \min_{\theta\in\Delta}\big\langle f^{(k,\tau)},\theta\big\rangle + \sum_{i=1}^{I}\frac{Q_i(k)}{V}(\langle g_i,\theta\rangle - c_i)\\
=& \min_{\theta\in\Delta}\big\langle f^{(k,\tau)},\theta\big\rangle + \sum_{i=1}^{I}[\eta_{k,\tau}^*]_i(\langle g_i,\theta\rangle - c_i) + \sum_{i=1}^{I}\Big(\frac{Q_i(k)}{V} - [\eta_{k,\tau}^*]_i\Big)(\langle g_i,\theta\rangle - c_i)\\
\le& q^{(k,\tau)}(\eta_{k,\tau}^*) + \Big\|\frac{\mathbf{Q}(k)}{V} - \eta_{k,\tau}^*\Big\|_2\|\mathbf{g}(\theta) - \mathbf{c}\|_2 \le H + 2\overline{\vartheta}H,
\end{aligned}$$

where the first inequality is by the definition of $q^{(k,\tau)}(\eta^*_{k,\tau})$ and Cauchy-Schwarz inequality, and the second inequality is due to weak duality relation and Assumption 3.4 such that

$$q^{(k,\tau)}(\eta^*_{k,\tau}) \le \langle f^{(k,\tau)}, \theta^*_{k,\tau}\rangle \le \big\|f^{(k,\tau)}\big\|_\infty \|\theta^*_{k,\tau}\|_1 \le H,$$
$$\Big\|\frac{\mathbf{Q}(k)}{V} - \eta^*_{k,\tau}\Big\|_2 \|\mathbf{g}(\theta)-\mathbf{c}\|_2 \le \overline{\vartheta}\sqrt{\sum_{i=1}^{I} \Big|\langle g_i,\theta\rangle - c_i\Big|^2} \le \overline{\vartheta}\sum_{i=1}^{I}(\|g_i\|_\infty\|\theta\|_1 + |c_i|) \le 2\overline{\vartheta}H.$$

Now we can combine the two cases as follows

$$q^{(k,\tau)}\Big(\frac{\mathbf{Q}(k)}{V}\Big) \le -\overline{\sigma}\Big\|\frac{\mathbf{Q}(k)}{V}\Big\|_2 + \overline{\sigma}B + 2H + 2\overline{\vartheta}H + \overline{\sigma}\overline{\vartheta}. \tag{3.40}$$

The bound in (3.40) is due to

(1) When $\mathrm{dist}\big(\frac{\mathbf{Q}(k)}{V}, \mathcal{V}^*_{k,\tau}\big) \ge \overline{\vartheta}$, we have

$$q^{(k,\tau)}\left(\frac{\mathbf{Q}(k)}{V}\right) \le -\overline{\sigma}\left\|\frac{\mathbf{Q}(k)}{V}\right\|_2 + \overline{\sigma}B + H \le -\overline{\sigma}\left\|\frac{\mathbf{Q}(k)}{V}\right\|_2 + \overline{\sigma}B + 2H + 2\overline{\vartheta}H + \overline{\sigma}\overline{\vartheta}.$$

(2) When $\mathrm{dist}\big(\frac{\mathbf{Q}(k)}{V}, \mathcal{V}^*_{k,\tau}\big) < \overline{\vartheta}$, we have

$$q^{(k,\tau)}\left(\frac{\mathbf{Q}(k)}{V}\right) \le H + 2\overline{\vartheta}H \le -\overline{\sigma}\Big\|\frac{\mathbf{Q}(k)}{V}\Big\|_2 + \overline{\sigma}B + 2H + 2\overline{\vartheta}H + \overline{\sigma}\overline{\vartheta},$$

since $-\overline{\sigma}\left\|\frac{\mathbf{Q}(k)}{V}\right\|_2 + \overline{\sigma}\overline{\vartheta} + \overline{\sigma}B \ge -\overline{\sigma}\cdot\mathrm{dist}\big(\frac{\mathbf{Q}(k)}{V}, \mathcal{V}^*_{k,\tau}\big) + \overline{\sigma}\overline{\vartheta} + \overline{\sigma}B - \overline{\sigma}B = \overline{\sigma}\big[-\mathrm{dist}\big(\frac{\mathbf{Q}(k)}{V}, \mathcal{V}^*_{k,\tau}\big) + \overline{\vartheta}\big] \ge 0$.

Therefore, plugging (3.40) into (3.39), we can obtain that given the event $\mathcal{D}_K$ happens, the following holds

$$\begin{aligned}&\mathbb{E}[\|\mathbf{Q}(\tau+k)\|_2^2 - \|\mathbf{Q}(k)\|_2^2|\mathcal{F}^{k-1}, \mathcal{D}_K]\\ &\quad\le 2\tau^2H + \tau C_{V,\alpha,\lambda} + 2\alpha H\log(|\mathbf{S}|^2|\mathbf{A}|/\lambda) - 2\tau\overline{\sigma}\|\mathbf{Q}(k)\|_2,\end{aligned} \tag{3.41}$$

where we define

$$C_{V,\alpha,\lambda} := 2(\overline{\sigma}B + \overline{\sigma}\,\overline{\vartheta})V + (6+4\overline{\vartheta})VH + \frac{VL}{\alpha} + 4H\lambda V + 2\alpha\lambda H\log|\mathbf{S}|^2|\mathbf{A}| + 8H^2$$

We can see that if $\|\mathbf{Q}(k)\|_2 \ge (2\tau H + C_{V,\alpha,\lambda})/\overline{\sigma} + 2\alpha\lambda H\log(|\mathbf{S}|^2|\mathbf{A}|/\lambda)/(\overline{\sigma}\tau) + \tau\overline{\sigma}/2$, then according to (3.41), there is

$$\begin{aligned}\mathbb{E}[\|\mathbf{Q}(\tau+k)\|_2^2|\mathcal{F}^{k-1},\mathcal{D}_K] &\leq \|\mathbf{Q}(k)\|_2^2 - \tau\overline{\sigma}\|\mathbf{Q}(k)\|_2 - \frac{\overline{\sigma}^2\tau^2}{2} \\ &\leq \|\mathbf{Q}(k)\|_2^2 - \tau\overline{\sigma}\|\mathbf{Q}(k)\|_2 + \frac{\overline{\sigma}^2\tau^2}{4} \\ &\leq \Big(\|\mathbf{Q}(k)\|_2 - \frac{\tau\overline{\sigma}}{2}\Big)^2.\end{aligned}$$

Due to $\|\mathbf{Q}(k)\|_2 \geq \frac{\tau\overline{\sigma}}{2}$ and by Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}[\|\mathbf{Q}(\tau+k)\|_2|\mathcal{F}^{k-1},\mathcal{D}_K] &\leq \sqrt{\mathbb{E}[\|\mathbf{Q}(\tau+k)\|_2^2|\mathcal{F}^{k-1},\mathcal{D}_K]} \\ &\leq \|\mathbf{Q}(k)\|_2 - \frac{\tau\overline{\sigma}}{2}.\end{aligned} \tag{3.42}$$

Then we can compute the expectation $\mathbb{E}[\|\mathbf{Q}(\tau+k)\|_2^2 - \|\mathbf{Q}(k)\|_2^2|\mathcal{F}^{k-1}]$ according to the law of total expectation. With (3.36) and (3.42), we can obtain that

$$\begin{aligned}&\mathbb{E}[\|\mathbf{Q}(\tau+k)\|_2 - \|\mathbf{Q}(k)\|_2|\mathcal{F}^{k-1}] \\ &\quad= P(\mathcal{D}_K)\mathbb{E}[\|\mathbf{Q}(\tau+k)\|_2 - \|\mathbf{Q}(k)\|_2|\mathcal{F}^{k-1},\mathcal{D}_K] \\ &\qquad+ P(\overline{\mathcal{D}}_K)\mathbb{E}[\|\mathbf{Q}(\tau+k)\|_2 - \|\mathbf{Q}(k)\|_2|\mathcal{F}^{k-1},\overline{\mathcal{D}}_K] \\ &\quad\leq -\frac{\tau\overline{\sigma}}{2}(1-\zeta) + 2\zeta\tau H \\ &\quad= -\tau\Big[\frac{\overline{\sigma}}{2} - \zeta\Big(\frac{\overline{\sigma}}{2} + 2H\Big)\Big] \leq -\frac{\overline{\sigma}}{4}\tau,\end{aligned}$$

where we let $\overline{\sigma}/4 \geq \zeta(\overline{\sigma}/2 + 2H)$.

Summarizing the above results, we have that if $\overline{\sigma}/4 \geq \zeta(\overline{\sigma}/2 + 2H)$, then

$$\begin{aligned}&|\|\mathbf{Q}(k+1)\|_2 - \|\mathbf{Q}(k)\|_2| \leq 2H, \\ &\mathbb{E}[\|\mathbf{Q}(k+\tau)\|_2 - \|\mathbf{Q}(k)\|_2|\mathcal{F}^{k-1}] \leq \begin{cases} 2\tau H, & \text{if } \|\mathbf{Q}(k)\|_2 < \psi \\ -\overline{\sigma}\tau/4, & \text{if } \|\mathbf{Q}(k)\|_2 \geq \psi \end{cases},\end{aligned}$$

where we let

$$\begin{aligned}&\psi = \frac{2\tau H + C_{V,\alpha,\lambda}}{\overline{\sigma}} + \frac{2\alpha H\log(|\mathbf{S}|^2|\mathbf{A}|/\lambda)}{\overline{\sigma}\tau} + \frac{\tau\overline{\sigma}}{2}, \\ &C_{V,\alpha,\lambda} = 2(\overline{\sigma}B + \overline{\sigma}\,\overline{\vartheta})V + (6+4\overline{\vartheta})VH + \frac{VL}{\alpha} + 4H\lambda V + 2\alpha\lambda H\log|\mathbf{S}|^2|\mathbf{A}| + 8H^2.\end{aligned}$$

By Lemma 3.24, for a certain $k \in [K+1-\tau]$ the following inequality holds with probability at least $1-\delta$,

$$\|\mathbf{Q}(k)\|_2 \leq \psi + \tau\frac{512H^2}{\overline{\sigma}}\log\left(1+\frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}\right)+\tau\frac{64H^2}{\overline{\sigma}}\log\frac{1}{\delta}. \tag{3.43}$$

Further by the union bound, we have that with probability at least $1-(K+1-\tau)\delta \geq 1-K\delta$, for any $k\in[K+1-\tau]$, the above inequality (3.43) holds. Note that (3.43) only holds when $k\in[K+1-\tau]$. For $K+2-\tau\leq k\leq K+1$, when (3.43) holds for $k\in[K+1-\tau]$, combining (3.43) and (3.35), we have

$$\|\mathbf{Q}(k)\|_2 \leq \psi + \tau\frac{512H^2}{\overline{\sigma}}\log\left(1+\frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}\right)+\tau\frac{64H^2}{\overline{\sigma}}\log\frac{1}{\delta}+2\tau H. \tag{3.44}$$

Thus, with probability at least $1-K\delta$, for any k satisfying $1\leq k\leq K+1$, the inequality (3.44) holds. We can discuss the upper bound of the term $\log\big(1+\frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}\big)$ in the following way: **(1)** if $\frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}\geq 1$, then this term is bounded by $\log\big(\frac{256H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}\big)=\frac{\overline{\sigma}}{32H}+\log\frac{256H^2}{\overline{\sigma}^2}$; **(2)** if $\frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}<1$, then the term is bounded by $\log 2$. Thus, we have

$$\log\left(1+\frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}\right)\leq \log 2+\frac{\overline{\sigma}}{32H}+\log\frac{256H^2}{\overline{\sigma}^2}.$$

This discussion shows that the $\log$ term in (3.44) will not introduce extra dependency on H except a $\log H$ term. This completes our proof. □

3.7.4 Proof of Lemma 3.15

We first provide Lemmas 3.25 below. Then, we give the proof of Lemma 3.15 based on this lemma.

Lemma 3.25 (Lemma 9 of Yu et al. [2017]). *Let $\{Z(k),k\geq 0\}$ be a supermartingale adapted to a filtration $\{\mathcal{U}^k,k\geq 0\}$ with $Z(0)=0$ and $\mathcal{U}^0=\{\varnothing,\Omega\}$, i.e., $\mathbb{E}[Z(k+1)\,|\,\mathcal{U}^k]\leq Z(k)$, $\forall k\geq 0$. Suppose there exists a constant $\varsigma>0$ such that $\{|Z(k+1)-Z(k)|>\varsigma\}\subset\{Y(k)>0\}$, where $Y(k)$ is process with $Y(k)$ adpated to $\mathcal{F}^k$ for all $k\geq 0$. Then, for all $z>0$, we have*

$$\Pr(Z(k)\geq z)\leq e^{-z^2/(2k\varsigma^2)}+\sum_{\tau=0}^{k-1}\Pr(Y(\tau)>0),\forall k\geq 1.$$

We are in position to give the proof of Lemma 3.15.

Proof of Lemma 3.15. Now we compute the upper bound of the term $\sum_{k=1}^K\sum_{i=1}^I Q_i(k)(\langle g_i^k,\overline{\theta}^*\rangle - c_i)$. Note that $Z(k):=\sum_{\tau=1}^k\sum_{i=1}^I Q_i(\tau)(\langle g_i^\tau,\overline{\theta}^*\rangle-c_i)$ is supermartingale which can be verified by

$$
\begin{aligned}
\mathbb{E}[Z(k)|\mathcal{F}^{k-1}] =& \mathbb{E}\Big[\sum_{\tau=1}^{k}\sum_{i=1}^{I} Q_i(\tau)(\langle g_i^\tau, \overline{\theta}^*\rangle - c_i)\Big|\mathcal{F}^{k-1}\Big]\\
=& \sum_{i=1}^{I}\mathbb{E}[Q_i(k)|\mathcal{F}^{k-1}](\langle\mathbb{E}[g_i^k|\mathcal{F}^{k-1}], \overline{\theta}^*\rangle - c_i)\\
&+\sum_{\tau=1}^{k-1}\sum_{i=1}^{I} Q_i(\tau)(\langle g_i^\tau, \overline{\theta}^*\rangle - c_i)\\
\le& \sum_{\tau=1}^{k-1}\sum_{i=1}^{I} Q_i(\tau)(\langle g_i^\tau, \overline{\theta}^*\rangle - c_i) = Z(k-1),
\end{aligned}
$$

where $Q_i(k)$ and g_i^k are independent variables with $Q_i(k) \ge 0$ and $\langle\mathbb{E}[g_i^k|\mathcal{F}^{k-1}], \overline{\theta}^*\rangle = \langle g_i, \overline{\theta}^*\rangle \le c_i$. On the other hand, we know that the random process has bounded drifts as

$$
\begin{aligned}
|Z(k+1) - Z(k)| =& \sum_{i=1}^{I} Q_i(k+1)(\langle g_i^{k+1}, \overline{\theta}^*\rangle - c_i)\\
\le& \|\mathbf{Q}(k+1)\|_2\sqrt{\sum_{i=1}^{I}\big|\langle g_i^{k+1}, \overline{\theta}^*\rangle - c_i\big|^2}\\
\le& \|\mathbf{Q}(k+1)\|_2\sum_{i=1}^{I}(\|g_i^{k+1}\|_\infty\|\overline{\theta}^*\|_1 + |c_i|) \le 2H\|\mathbf{Q}(k+1)\|_2,
\end{aligned}
$$

where the first inequality is by Cauchy-Schwarz inequality, and the last inequality is by Assumption 3.4. This also implies that for an arbitrary ς, we have $\{|Z(k+1) - Z(k)| > \varsigma\} \subset \{Y(k) := \|\mathbf{Q}(k+1)\|_2 - \varsigma/(2H) > 0\}$ since $|Z(k+1) - Z(k)| > \varsigma$ implies $2H\|\mathbf{Q}(k+1)\|_2 > \varsigma$ according to the above inequality. Thus, by Lemma 3.25, we have

$$
\begin{aligned}
&\Pr\bigg(\sum_{k=1}^{K}\sum_{i=1}^{I} Q_i(k)(\langle g_i^k, \overline{\theta}^*\rangle - c_i) \ge z\bigg)\\
&\le e^{-\frac{z^2}{2K\varsigma^2}} + \sum_{k=0}^{K-1}\Pr\bigg(\|\mathbf{Q}(k+1)\|_2 > \frac{\varsigma}{2H}\bigg) = e^{-\frac{z^2}{2K\varsigma^2}} + \sum_{k=1}^{K}\Pr\bigg(\|\mathbf{Q}(k)\|_2 > \frac{\varsigma}{2H}\bigg),
\end{aligned}
\tag{3.45}
$$

where we can see that bounding $\|\mathbf{Q}(k)\|_2$ is the key to obtaining the bound of $\sum_{k=1}^{K}\sum_{i=1}^{I} Q_i(k)(\langle g_i^k, \overline{\theta}^*\rangle - c_i)$.

Next, we will show the upper bound of the term $\|\mathbf{Q}(k)\|_2$. According to the proof of Lemma

3.13 in , if $\overline{\sigma}/4 \geq \zeta(\overline{\sigma}/2 + 2H)$, setting

$$\psi = \frac{2\tau H + C_{V,\alpha,\lambda}}{\overline{\sigma}} + \frac{2\alpha H \log(|\mathbf{S}|^2|\mathbf{A}|/\lambda)}{\overline{\sigma}\tau} + \frac{\tau\overline{\sigma}}{2},$$
$$C_{V,\alpha,\lambda} := 2V\left(\overline{\sigma}B + 3H + 2\overline{\vartheta}H + \overline{\sigma}\overline{\vartheta} + \frac{H}{2\alpha} + 2H\lambda + \frac{\alpha\lambda H \log|\mathbf{S}|^2|\mathbf{A}| + 4H^2}{V}\right),$$

we have that with probability at least $1-\delta$, for a certain $k \in [K+1-\tau]$,

$$\|\mathbf{Q}(k)\|_2 \leq \psi + \tau\frac{512H^2}{\overline{\sigma}}\log[1 + \frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}] + \tau\frac{64H^2}{\overline{\sigma}}\log\frac{1}{\delta} + 2\tau H.$$

Thus, combining (3.35) and the above inequality at $k = K+1-\tau$, with probability at least $1-\delta$, for a certain k satisfying $K+2-\tau \leq k \leq K+1$, the above inequality also holds. The above inequality is equivalent to

$$\Pr\left(\|\mathbf{Q}(k)\|_2 > \psi + \tau\frac{512H^2}{\overline{\sigma}}\log[1 + \frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}] + \tau\frac{64H^2}{\overline{\sigma}}\log\frac{1}{\delta} + 2\tau H\right) \leq \delta.$$

Setting $\varsigma = 2H\psi + \tau\frac{1024H^3}{\overline{\sigma}}\log\left[1 + \frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}\right] + \tau\frac{128H^3}{\overline{\sigma}}\log\frac{1}{\delta} + 4\tau H^2$ and $z = \sqrt{2K\varsigma^2\log\frac{1}{K\delta}}$ in (3.45), then the following probability hold with probability at least $1-2K\delta$ with

$$\sum_{k=1}^{K}\sum_{i=1}^{I} Q_i(k)(\langle g_i^k, \overline{\theta}^*\rangle - c_i)$$
$$\leq \left(2H\psi + \tau\frac{1024H^3}{\overline{\sigma}}\log\left[1 + \frac{128H^2}{\overline{\sigma}^2}e^{\overline{\sigma}/(32H)}\right] + \tau\frac{128H^3}{\overline{\sigma}}\log\frac{1}{\delta} + 4\tau H^2\right)\sqrt{K\log\frac{1}{K\delta}},$$

which completes the proof. □

3.8 Proofs of Lemmas for Constraint Violation Bound

3.8.1 Proof of Lemma 3.16

Proof of Lemma 3.16. We start our proof with the updating rule of $\mathbf{Q}(\cdot)$ as follows

$$\begin{aligned} Q_i(k) =& \max\{Q_i(k-1) + \langle g_i^{k-1}, \theta^k\rangle - c_i, 0\} \\ \geq& Q_i(k-1) + \langle g_i^{k-1}, \theta^k\rangle - c_i \\ \geq& Q_i(k-1) + \langle g_i^{k-1}, \theta^{k-1}\rangle - c_i + \langle g_i^{k-1}, \theta^k - \theta^{k-1}\rangle. \end{aligned}$$

Rearranging the terms in the above inequality futher leads to

$$\langle g_i^{k-1}, \theta^{k-1}\rangle - c_i \leq Q_i(k) - Q_i(k-1) - \langle g_i^{k-1}, \theta^k - \theta^{k-1}\rangle.$$

Thus, taking summation on both sides of the above inequality from 2 to $K+1$ leads to

$$\begin{aligned}\sum_{k=1}^{K}(\langle g_i^k, \theta^k\rangle - c_i) &\leq Q_i(K+1) - \sum_{k=1}^{K}\langle g_i^k, \theta^{k+1} - \theta^k\rangle \\ &\leq Q_i(K+1) + \sum_{k=1}^{K}\|g_i^k\|_\infty\|\theta^{k+1} - \theta^k\|_1,\end{aligned}$$

where the second inequality is due to Hölder's inequality. Note that $Q_i(K+1)$ is no less than 0 according to its updating rule $Q_i(k) = \max\{Q_i(k-1) + \langle g_i^{k-1}, \theta^k\rangle - c_i, 0\} \geq 0$. Thus, we have

$$\left[\sum_{k=1}^{K}(\langle g_i^k, \theta^k\rangle - c_i)\right]_+ \leq Q_i(K+1) + \sum_{k=1}^{K}\|g_i^k\|_\infty\|\theta^{k+1} - \theta^k\|_1,$$

where $[\,\cdot\,]_+$ is an entry-wise application of the operation $\max\{\cdot, 0\}$ for any vector.

Defining $\mathbf{g}^k(\theta^k) := [\langle g_1^k, \theta^k\rangle, \cdots, \langle g_I^k, \theta^k\rangle]^\top$ and $\mathbf{c} := [c_1, \cdots, c_I]^\top$, we would obtain

$$\begin{aligned}\left\|\left[\sum_{k=1}^{K}(\mathbf{g}^k(\theta^k) - \mathbf{c})\right]_+\right\|_2 &\leq \|\mathbf{Q}(K+1)\|_2 + \sum_{k=1}^{K}\sqrt{\sum_{i=1}^{I}\|g_i^k\|_\infty^2}\|\theta^{k+1} - \theta^k\|_1 \\ &\leq \|\mathbf{Q}(K+1)\|_2 + \sum_{k=1}^{K}\sum_{i=1}^{I}\|g_i^k\|_\infty\|\theta^{k+1} - \theta^k\|_1 \\ &\leq \|\mathbf{Q}(K+1)\|_2 + \sum_{k=1}^{K}\|\theta^{k+1} - \theta^k\|_1,\end{aligned}$$

where the third inequality is due to Assumption 3.4. This completes the proof. □

3.8.2 Proof of Lemma 3.17

Lemma 3.26 (Proposition 18 of Jaksch et al. [2010])**.** *The number of epochs in K episodes with $K \geq |\mathbf{S}||\mathbf{A}|$ is upper bounded by*

$$\ell(K) \leq |\mathbf{S}||\mathbf{A}|\log_2\left(\frac{8K}{|\mathbf{S}||\mathbf{A}|}\right) \leq \sqrt{K|\mathbf{S}||\mathbf{A}|}\log_2\left(\frac{8K}{|\mathbf{S}||\mathbf{A}|}\right),$$

where $\ell(\cdot)$ is a mapping from a certain episode to the epoch where it lives.

We are ready to give the proof of Lemma 3.17.

Proof of Lemma 3.17. We need to discuss the upper bound of the term $\|\theta^{k+1}-\theta^k\|_1$ for $k \in [K]$ in two different cases:

(1) $\ell(k+1)=\ell(k)$, i.e., episodes $k+1$ and k are in the same epoch;

(2) $\ell(k+1)>\ell(k)$, i.e., episodes $k+1$ and k are in two different epochs.

We first consider case **(1)**. According to Lemma 3.21 and the updating rule (3.7), letting $\mathbf{x}^{\mathrm{opt}}=\theta^k$, $\mathbf{y}=\widetilde{\theta}^{k-1}$, $\mathbf{z}=\theta^{k-1}$ and $F(\theta)=\big\langle Vf^{k-1}+\sum_{i=1}^I Q_i(k-1)g_i^{k-1},\theta\big\rangle$ with $k\geq 2$ and $\ell(k)=\ell(k-1)$, we have

$$\begin{aligned}&\Big\langle Vf^{k-1}+\sum_{i=1}^I Q_i(k-1)g_i^{k-1},\theta^k\Big\rangle+\alpha D(\theta^k,\widetilde{\theta}^{k-1})\\&\qquad\leq\Big\langle Vf^{k-1}+\sum_{i=1}^I Q_i(k-1)g_i^{k-1},\theta^{k-1}\Big\rangle+\alpha D(\theta^{k-1},\widetilde{\theta}^{k-1})-\alpha D(\theta^{k-1},\theta^k).\end{aligned}$$

Rearranging the terms and dropping the last term (due to $D(\theta^{k-1},\theta^k)\geq 0$) yield

$$\begin{aligned}\alpha D(\theta^k,\widetilde{\theta}^{k-1})\leq&\Big\langle Vf^{k-1}+\sum_{i=1}^I Q_i(k-1)g_i^{k-1},\theta^{k-1}-\theta^k\Big\rangle+\alpha D(\theta^{k-1},\widetilde{\theta}^{k-1})\\\leq&\Big(V\|f^{k-1}\|_\infty+\sum_{i=1}^I Q_i(k-1)\|g_i^{k-1}\|_\infty\Big)\|\theta^{k-1}-\theta^k\|_1+\alpha D(\theta^{k-1},\widetilde{\theta}^{k-1})\\\leq&\left(V+\|\mathbf{Q}(k-1)\|_2\sqrt{\sum_{i=1}^I\|g_i^{k-1}\|_\infty^2}\right)\|\theta^{k-1}-\theta^k\|_1+\alpha D(\theta^{k-1},\widetilde{\theta}^{k-1})\\\leq&(V+\|\mathbf{Q}(k-1)\|_2)\|\theta^{k-1}-\theta^k\|_1+\alpha\lambda H\log|\mathbf{S}|^2|\mathbf{A}|,\end{aligned}$$

where the second inequality is by Hölder's inequality and triangle inequality, the third inequality is by Cauchy–Schwarz inequality and Assumption 3.4, and the last inequality is due to Assumption 3.4 and the first inequality in Lemma 3.23 with setting $\theta=\theta'=\theta^{k-1}$ and $\widetilde{\theta}'=\widetilde{\theta}^{k-1}$. Note that by Lemma 3.22, there is

$$D(\theta^k,\widetilde{\theta}^{k-1})\geq\frac{1}{2H}\|\theta^k-\widetilde{\theta}^{k-1}\|_1^2.$$

Thus, combining the previous two inequalities, we obtain

$$\|\theta^k-\widetilde{\theta}^{k-1}\|_1^2\leq\frac{2HV+2H\|\mathbf{Q}(k-1)\|_2}{\alpha}\|\theta^{k-1}-\theta^k\|_1+2\lambda H^2\log|\mathbf{S}|^2|\mathbf{A}|,$$

which further leads to

$$\|\theta^k - \widetilde{\theta}^{k-1}\|_1 \leq \sqrt{\frac{2HV + 2H\|\mathbf{Q}(k-1)\|_2}{\alpha}\|\theta^{k-1} - \theta^k\|_1} + \sqrt{2\lambda H^2 \log |\mathbf{S}|^2|\mathbf{A}|}.$$

Since there is

$$\begin{aligned}\|\theta^k - \widetilde{\theta}^{k-1}\|_1 &= \sum_{h=0}^{H-1}\left\|\theta_h^k - (1-\lambda)\theta_h^{k-1} - \lambda\frac{\mathbf{1}}{|\mathbf{S}|^2|\mathbf{A}|}\right\|_1 \\ &\geq (1-\lambda)\|\theta^k - \theta^{k-1}\|_1 - \lambda H,\end{aligned}$$

where $\theta_h := [\theta(s,a,s')]_{s\in\mathbf{S}_h, a\in\mathbf{A}, s'\in\mathbf{S}_{h+1}}$, combining it with the last inequality, we further have

$$\begin{aligned}\|\theta^k - \theta^{k-1}\|_1 &\leq \sqrt{\frac{2HV + 2H\|\mathbf{Q}(k-1)\|_2}{\alpha(1-\lambda)^2}\|\theta^{k-1} - \theta^k\|_1} + \frac{\sqrt{2\lambda H^2 \log |\mathbf{S}|^2|\mathbf{A}|}}{(1-\lambda)} + \frac{\lambda H}{1-\lambda} \\ &\leq \frac{2HV + 2H\|\mathbf{Q}(k-1)\|_2}{2(1-\lambda)^2\alpha} + \frac{1}{2}\|\theta^{k-1} - \theta^k\|_1 + \frac{\sqrt{2\lambda H^2 \log |\mathbf{S}|^2|\mathbf{A}|}}{1-\lambda} + \frac{\lambda H}{1-\lambda},\end{aligned}$$

where the last inequality is due to $\sqrt{ab} \leq |a|/2 + |b|/2$. Rearranging the terms in the above inequality gives for $k \geq 2$ with $\ell(k) = \ell(k-1)$,

$$\|\theta^k - \theta^{k-1}\|_1 \leq \frac{2HV + 2H\|\mathbf{Q}(k-1)\|_2}{(1-\lambda)^2\alpha} + \frac{\sqrt{8\lambda H^2 \log |\mathbf{S}|^2|\mathbf{A}|}}{1-\lambda} + \frac{2\lambda H}{1-\lambda}.$$

Shifting the index in the above inequality, we further have for $k \in [K]$ with $\ell(k+1) = \ell(k)$,

$$\|\theta^{k+1} - \theta^k\|_1 \leq \frac{2HV + 2H\|\mathbf{Q}(k)\|_2}{(1-\lambda)^2\alpha} + \frac{\sqrt{8\lambda H^2 \log |\mathbf{S}|^2|\mathbf{A}|}}{1-\lambda} + \frac{2\lambda H}{1-\lambda}. \tag{3.46}$$

Next, we consider case **(2)** where $\ell(k+1) > \ell(k)$ with $k \in [K]$. It is difficult to know whether the two solutions θ^{k+1} and θ^k are in the same feasible set since $\Delta(\ell(k+1),\zeta) \neq \Delta(\ell(k),\zeta)$. Thus, the above result does not hold. Then, we give a bound for the term $\|\theta^{k+1} - \theta^k\|_1$ as follows

$$\begin{aligned}\|\theta^{k+1} - \theta^k\|_1 &\leq \|\theta^{k+1}\|_1 + \|\theta^k\|_1 \\ &= \sum_{h=0}^{H-1}\sum_{s\in\mathcal{S}_h}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}_{h+1}}[\theta^{k+1}(s,a,s') + \theta^k(s,a,s')] = 2H,\end{aligned} \tag{3.47}$$

However, we can observe that $\ell(k+1) > \ell(k)$ only happens when episode $k+1$ is a starting episode for a new epoch. The number of starting episodes for new epochs in $K+1$ episodes is bounded by $\ell(K)$, namely the total number of epochs in K episodes. According to Lemma

3.26, the total number of epochs $\ell(K)$ is bounded by $\ell(K) \leq \sqrt{K|\mathbf{S}||\mathbf{A}|}\log_2[8K/(|\mathbf{S}||\mathbf{A}|)] \leq 1.5\sqrt{K|\mathbf{S}||\mathbf{A}|}\log_2[8K/(|\mathbf{S}||\mathbf{A}|)]$, which only grows in the order of $\sqrt{K}\log K$.

Thus, we can decompose the term $\sum_{k=1}^K \|\theta^{k+1}-\theta^k\|_1$ in the following way

$$\begin{aligned}\sum_{k=1}^K \|\theta^{k+1}-\theta^k\|_1 = &\sum_{\substack{k:\ k\leq K,\\ \ell(k+1)>\ell(k)}} \|\theta^{k+1}-\theta^k\|_1 + \sum_{\substack{k:\ k\leq K,\\ \ell(k+1)=\ell(k)}} \|\theta^{k+1}-\theta^k\|_1\\ \leq &2H\ell(K) + \sum_{\substack{k:\ k\leq K,\\ \ell(k+1)=\ell(k)}} \|\theta^{k+1}-\theta^k\|_1,\end{aligned}$$

where the inequality is due to (3.47) and the fact that $\sum_{\substack{k:\ k\leq K,\\ \ell(k+1)>\ell(k)}} 1 \leq \ell(K)$. By (3.46), we can further bound the last term in the above inequality as

$$\sum_{\substack{k:\ k\leq K,\\ \ell(k+1)=\ell(k)}} \|\theta^{k+1}-\theta^k\|_1 \leq \frac{2KHV + 2H\sum_{k=1}^K \|\mathbf{Q}(k)\|_2}{(1-\lambda)^2\alpha} + \frac{\sqrt{8\lambda\log|\mathbf{S}|^2|\mathbf{A}|}}{1-\lambda}KH + \frac{2\lambda}{1-\lambda}KH,$$

where we relax the summation on the right-hand side to $\sum_{k=1}^K$. Thus, we eventually obtain

$$\begin{aligned}\sum_{k=1}^K \|\theta^{k+1}-\theta^k\|_1 &\leq 2H\ell(K) + \sum_{\substack{k:\ k\leq K,\\ \ell(k+1)=\ell(k)}} \|\theta^{k+1}-\theta^k\|_1\\ &\leq 3H\sqrt{K|\mathbf{S}||\mathbf{A}|}\log\frac{8K}{|\mathbf{S}||\mathbf{A}|} + \frac{2H}{(1-\lambda)^2\alpha}\sum_{k=1}^K \|\mathbf{Q}(k)\|_2 + \frac{2KHV}{(1-\lambda)^2\alpha}\\ &\quad + \frac{2\lambda KH}{1-\lambda} + \frac{\sqrt{8\lambda\log|\mathbf{S}|^2|\mathbf{A}|}}{1-\lambda}KH,\end{aligned}$$

where we use the result in Lemma 3.26 to bound the number of epoch, i.e., $\ell(K)$. This completes the proof. □

CHAPTER 4

Policy Optimization for Zero-Sum Markov Games with Structured Transitions

4.1 Introduction

Widely applied in multi-agent reinforcement learning [Sutton and Barto, 2018, Bu et al., 2008], Policy Optimization (PO) has achieved tremendous empirical success [Foerster et al., 2016, Leibo et al., 2017, Silver et al., 2016, 2017, Berner et al., 2019, Vinyals et al., 2019], due to its high efficiency and easiness to combine with different optimization techniques. Despite these empirical successes, theoretical understanding of multi-agent policy optimization, especially the zero-sum Markov game [Littman, 1994] via policy optimization, lags rather behind. Most recent works studying zero-sum Markov games (e.g. Xie et al. [2020], Bai and Jin [2020]) focus on value-based methods achieving $\widetilde{\mathcal{O}}(\sqrt{K})$ regrets and they assume there is a central controller available solving for coarse correlated equilibrium or Nash equilibrium at each step, which brings extra computational cost. Here we let K denotes the total number of episodes.[1] On the other hand, although there has been great progress on understanding single-agent PO algorithms [Sutton et al., 2000, Kakade, 2002, Schulman et al., 2015, Papini et al., 2018, Cai et al., 2019, Bhandari and Russo, 2019, Liu et al., 2019], directly extending the single-agent PO to the multi-agent setting encounters the main challenge of non-stationary environments caused by agents changing their own policies simultaneously [Bu et al., 2008, Zhang et al., 2019a]. In this chapter, we aim to answer the following challenging question:

Can policy optimization probably solve two-player zero-sum Markov games
to achieve $\mathcal{O}(\sqrt{K})$ regrets?

As an initial attempt to tackle the problem, in this chapter, we focus on two *non-trivial* classes of zero-sum Markov games with structured transitions: *factored independent transition* and *single-*

[1] The dependence on K is equivalent to the dependence of the total number of steps T in the same order, as we have $T := KH$ with H denoting the episode's length.

controller transition. For the game with the factored independent transition, the transition model is factored into two independent parts, and each player makes transition following their own transition model. The single-controller zero-sum game assumes that the transition model is entirely controlled by the actions of Player 1. In both settings, the rewards received are decided jointly by the actions of both players. These two problems capture the non-stationarity of the multi-agent reinforcement learning in the following aspects: **(1)** the rewards depend on both players' potentially adversarial actions and policies in both settings; **(2)** the rewards further depend on both players' states in the factored independent transition setting; **(3)** Player 2 in the single-controller transition setting faces non-stationary states determined by Player 1's policies. In addition to the non-stationarity, practically, the true transition model of the environment could be unknown to players and only bandit feedback is accessible to players. Thus, the non-stationarity, as well as the unknown transition model and reward function, poses great challenges to the design and theoretical analysis of the multi-agent PO algorithms.

In this chapter, we propose two novel optimistic policy optimization algorithms for the games with factored independent transition and single-controller zero-sum games respectively. Our algorithms are motivated by the close connection between the multi-agent PO and Fictitious Play (FP) framework. Specifically, FP [Robinson, 1951] is a classical framework for solving games based on simultaneous policy updates, which includes two major steps: inferring the opponent and taking the best response policy against the policy of the opponent. As an extension of FP to Markov games, our proposed PO algorithms possess two phases of learning, i.e., policy evaluation and policy improvement. The policy evaluation phase involves exchanging the policies of the previous episode[2], which is motivated by the step of inferring the opponent in FP. By making use of the policies from the previous episode, the algorithms further compute the value function and the Q-function with the estimated reward function and transition model. By the principle of "optimism in the face of uncertainty" [Auer et al., 2002, Bubeck and Cesa-Bianchi, 2012], their estimations incorporate UCB bonus terms to handle the non-stationarity of the environment as well as the uncertainty arising from only observing finite historical data. Furthermore, the policy improvement phase corresponds to taking the (regularized) best response policy via a mirror descent/ascent step (where the regularization comes from KL divergence), which can be viewed as a soft-greedy step based on the historical information about the opponent and the environment. This step resembles the smoothed FP [Fudenberg and Levine, 1995, Perolat et al., 2018, Zhang et al., 2019a] for normal form games (or matrix games). During this phase, both players in the factored independent transition setting and Player 2 in the single-controller setting demand to estimate the opponent's state reaching probability to handle the non-stationarity.

For each player, we measure the performance of its algorithm by the regret of the learned policy

[2] For ease of theoretical analysis, we assume there exists an oracle exchanging the players' policies.

sequence comparing against the best policy in hindsight after K episodes. In the two settings, our proposed algorithms can achieve an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret for both players, matching the regret of value-based algorithms. Furthermore, with both players running the proposed PO algorithms, they have $\widetilde{\mathcal{O}}(\sqrt{K})$ optimality gap. This chapter also partially solves one open question in Bai and Jin [2020] that how to solve a zero-sum Markov game of multiple steps $(H \geq 2)$ with an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret via mirror descent/ascent type (policy optimization) algorithms.

Related Work. There have been a large number of classical works studying the games with the independent transition model, e.g., Altman et al. [2005, 2008], Flesch et al. [2008], Singh and Hemachandra [2014]. In addition, the single-controller games are also broadly investigated in many existing works, .e.g., Parthasarathy and Raghavan [1981], Filar and Raghavan [1984], Rosenberg et al. [2004], Guan et al. [2016]. Most of the aforementioned works do not focus on the non-asymptotic regret analysis. Guan et al. [2016] studies the regret of the single-controller zero-sum game but with an assumption that the transition model is known to players. In contrast, this chapter provides a regret analysis for both transition models under a more realistic setting that the transition model is unknown. Games with the two structured transition models are closely associated with the applications in communications. The game with the factored independent transition [Altman et al., 2005] finds applications in wireless communications. An application example of the single-controller game is the attack-defense modeling in communications [Eldosouky et al., 2016].

Recently, many works are focusing on the non-asymptotic analysis of Markov games [Heinrich and Silver, 2016, Guan et al., 2016, Wei et al., 2017, Perolat et al., 2018, Zhang et al., 2019b, Xie et al., 2020, Bai and Jin, 2020]. Some of them aim to propose sample-efficient algorithms with theoretical regret guarantees for zero-sum games. Wei et al. [2017] proposes an algorithm extending single-agent UCRL2 algorithm [Jaksch et al., 2010], which requires solving a constrained optimization problem each round. Zhang et al. [2019b] also studies PO algorithms but does not provide regret analysis, which also assumes an extra linear quadratic structure and a known transition model. In addition, recent works on Markov games [Xie et al., 2020, Bai and Jin, 2020, Liu et al., 2020, Bai et al., 2020] propose value-based algorithms under the assumption that there exists a central controller that specifies the policies of agents by finding the coarse correlated equilibrium or Nash equilibrium for a set of matrix games in each episode. Bai and Jin [2020] also makes an attempt to investigate PO algorithms in zero-sum games. However, their work shows restrictive results where each player only plays one step in each episode. Right prior to our work, Daskalakis et al. [2021] also studies the policy optimization algorithm for a two-player zero-sum Markov game under an assumption of bounded distribution mismatch coefficient in a non-episodic setting. To achieve a certain error ε for the convergence measure defined in their work, their proposed algorithm requires an $\mathcal{O}(\varepsilon^{-12.5})$ sample complexity. A concurrent work [Tian et al., 2020] studies zero-sum games under a different online agnostic setting with PO methods and achieves an $\widetilde{\mathcal{O}}(K^{3/4})$ regret. Motivated by classical fictitious play works [Robinson, 1951, Fudenberg and Levine, 1995, Heinrich et al., 2015, Perolat et al., 2020], for the episodic Markov game, we focus on the setting where there is no central controller which determines the policies of the two players and we propose a policy optimization algorithm where each player updates its own policy based solely on the historical information at hand. Moreover, our result matches the $\mathcal{O}(\sqrt{K})$ regret upper bounds in Xie et al. [2020], Bai and Jin [2020] that are obtained by value-based methods.

Furthermore, we note that the game for each individual player can be viewed as a special case of MDPs with adversarial rewards and bandit feedbacks due to the adversarial actions of opponents. For such a class of MDP models in general, Jin et al. [2019] proposes an algorithm based on mirror descent involving occupancy measures and attains an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret. However, each update step of the algorithm requires solving another optimization problem which is more computationally demanding than our PO method. Besides, it is also unclear whether the algorithm in Jin et al. [2019] can be extended to zero-sum games. Moreover, for the same MDP model, Efroni et al. [2020b] proposes an optimistic policy optimization algorithm that achieves an $\widetilde{\mathcal{O}}(K^{2/3})$ regret. Thus, directly applying this result would yield an $\widetilde{\mathcal{O}}(K^{2/3})$ regret. In fact, regarding the problem as an MDP with adversarial rewards neglects the fact that such "adversarial reward functions" are determined by the actions and policies of the opponent. Thus, since each player knows the past actions taken and policies executed by the opponent under the FP framework, both players can construct accurate estimators of the environment after a sufficiently large number of episodes. As we will show in Sections 4.3 and 4.4, the proposed PO methods explicitly utilize the information of the opponent in the policy evaluation step, which is critical for the methods to obtain an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret.

4.2 Preliminaries

In this section, we formally introduce notations and setups. Then, we describe the two transition structures in detail.

4.2.1 Problem Setup

We consider a tabular episodic two-player zero-sum Markov game $(\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \mathbb{P}, r)$ defined as in Section 2.2 of Chapter 2 with *finite* action spaces $\mathcal{A}, \mathcal{B}$ and state space $\mathcal{S}$. The policy for Player 1 is denoted by π and the policy for Player 2 is denoted by ν. The value function $V^{\pi,\nu}(s)$ and the Q-function $Q^{\pi,\nu}(s, a, b)$ are defined the same as in Chapter 2 with omitting the reward r in the notation. Then, we further define the Bellman equation, NE, and ε-approximate NE the same as in

Chapter 2.

For ease of theoretical analysis, we normalize the value of the reward function $r = \{r_h\}_{h=1}^H$ in the range $[0, 1]$, i.e., $r_h(s, a, b) \in [0, 1]$ for all $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. We study a practical and challenging setting that the true transition model $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ and reward function $r = \{r_h\}_{h=1}^H$ are *unknown* to both players. And only the bandit feedbacks of the reward function are accessible to the players. At episode k, we let $\pi^k = \{\pi_h^k\}_{h=1}^H$ and $\nu^k = \{\nu_h^k\}_{h=1}^H$ be the policies for Players 1 and 2, and the two players move *simultaneously* with their own policies. By the end of the k-th episode, each player observes only the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$ and the bandit feedbacks along the trajectory. The bandit setting is more challenging than the full-information setting, where only the reward values $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$ on the trajectory are observed rather than the exact value function $r_h(s, a, b)$ for all $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. Moreover, the rewards $r_h^k(\cdot, \cdot, \cdot) \in [0, 1]$ is time-varying with an expectation $r_h = \mathbb{E}[r_h^k]$ which can be adversarially affected by the opponent's action or policy, indicating the non-stationarity of the environment.

Throughout this chapter, we let $\langle \cdot, \cdot \rangle_{\mathcal{S}}$, $\langle \cdot, \cdot \rangle_{\mathcal{A}}$, and $\langle \cdot, \cdot \rangle_{\mathcal{B}}$ denote the inner product over $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{B}$ respectively. At the h-th step in an episode, for any policy pair (π, ν), we denote $\pi_h(\cdot|s)$ and $\nu_h(\cdot|s)$ as column vectors over the space $\mathcal{A}$ and the space $\mathcal{B}$ respectively. We also denote $Q_h^{\pi,\nu}(s, \cdot, \cdot)$ as a matrix over the space $\mathcal{A} \times \mathcal{B}$ for any Q-function Q. Then, the expectation $\mathbb{E}_{a\sim\pi_h(\cdot|s), b\sim\nu_h(\cdot|s)}[Q_h^{\pi,\nu}(s, a, b)]$ can be equivalently rewritten as $[\pi_h(\cdot|s)]^\top Q_h^{\pi,\nu}(s, \cdot, \cdot)\nu_h(\cdot|s)$.

Basic Learning Framework. At the beginning of the k-th episode, each player observes the opponent's policy during the $(k-1)$-th episode. For simplicity of theoretical analysis, we assume there exists an oracle which can exchange players' policies in the last episode. Then, they take regularized best response policies via a mirror descent/ascent step for the current episode and make simultaneous moves.

Regret and Optimality Gap. The goal for Player 1 is to learn a sequence of policies, $\{\pi^k\}_{k>0}$, to have a small regret as possible in K episodes, which is defined as

$$\mathrm{Regret}_1(K) := \sum_{k=1}^K \Big[V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1)\Big]. \tag{4.1}$$

Here $\{\nu^k\}_{k=1}^K$ is any possible and potentially adversarial policy sequence of Player 2. The policy π^* is *the best policy in hindsight*, which is defined as $\pi^* := \mathrm{argmax}_\pi \sum_{k=1}^K V_1^{\pi,\nu^k}(s_1)$ for any specific $\{\nu^k\}_{k=1}^K$. Similarly, Player 2 aims to learn a sequence of policies, $\{\nu^k\}_{k>0}$, to have a small regret defined as

$$\mathrm{Regret}_2(K) := \sum_{k=1}^K \Big[V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1)\Big], \tag{4.2}$$

where $\{\pi^k\}_{k=1}^K$ is any possible policy sequence of Player 1. The policy ν^* is also *the best policy in hindsight* which is defined as $\nu^* := \mathrm{argmin}_\nu \sum_{k=1}^K V_1^{\pi^k,\nu}(s_1)$ for any specific $\{\pi^k\}_{k=1}^K$. Note that π^* and ν^* depend on opponents' policy sequence and is non-deterministic, and we drop such a dependency in the notation for simplicity. We further define the *optimality gap* $\mathrm{Gap}(K)$ as follows

$$\mathrm{Gap}(K) := \mathrm{Regret}_1(K) + \mathrm{Regret}_2(K). \tag{4.3}$$

Our definition of the optimality gap is consistent with a certain form of the regret to measure the learning performance of zero-sum games defined in Bai and Jin [2020, Definition 8]. Specifically, when the two players executes their algorithms to have small regrets, i.e., $\mathrm{Regret}_1(K)$ and $\mathrm{Regret}_2(K)$ are small, then their optimality gap $\mathrm{Gap}(K)$ is small as well.

On the other hand, letting the uniform mixture policies $\widehat{\pi} \sim \mathrm{Unif}(\pi^1, \ldots, \pi^K)$ and $\widehat{\nu} \sim \mathrm{Unif}(\nu^1, \ldots, \nu^K)$ be random policies sampled uniformly from the learned policies, then $(\widehat{\pi}, \widehat{\nu})$ can be viewed as an ε-approximate NE if $\mathrm{Regret}(K)/K \leq \varepsilon$. This build a connection between the approximate NE and the optimality gap.

4.2.2 Structured Transition Models

Factored Independent Transition. Consider a two-player Markov game where the state space is factored as $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2$ such that a state can be represented as $s = (s^1, s^2)$ with $s^1 \in \mathcal{S}_1$ and $s^2 \in \mathcal{S}_2$. Then, $\mathcal{S}_1$ and $\mathcal{S}_2$ are the state spaces for Player 1 and Player 2 respectively. Under this setting, the transition model is factored into two independent components, i.e.,

$$\mathbb{P}_h(s' \,|\, s, a, b) = \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\mathbb{P}_h^2(s^{2\prime} \,|\, s^2, b), \tag{4.4}$$

where we also have $s' = (s^{1\prime}, s^{2\prime})$, and $\mathbb{P}_h(s^{1\prime} \,|\, s^1, a)$ is the transition model for Player 1 and $\mathbb{P}_h(s^{2\prime} \,|\, s^2, b)$ for Player 2. Additionally, we consider the case where the policy of Player 1 only depends on its own state s^1 such that we have $\pi(a|s) = \pi(a|s^1)$ and meanwhile Player 2 similarly has the policy of the form $\nu(b|s) = \nu(b|s^2)$. Though the transitions, policies, and state spaces of two players are independent of each other, the reward function still depends on both players' actions and states, i.e., $r_h(s, a, b) = r_h(s^1, s^2, a, b)$.

Single-Controller Transition. In this setting, we consider that the transition model is controlled by the action of one player, e.g., Player 1, which is thus characterized by

$$\mathbb{P}_h(s' \,|\, s, a, b) = \mathbb{P}_h(s' \,|\, s, a). \tag{4.5}$$

In addition, the policies remain to be $\pi(a|s)$ and $\nu(b|s)$. The reward $r_h(s, a, b)$ is determined by both players' actions and the state s decided by the transition model controlled by Player 1.

Remark 4.1 (Misspecification). When the above models are not ideally satisfied, one can potentially consider scenarios that the transition model satisfies, for example, $\max_{s'} |\mathbb{P}_h(s' \mid s, a, b) - \mathbb{P}_h^1(s^{1\prime} \mid s^1, a)\mathbb{P}_h^2(s^{2\prime} \mid s^2, b)| \leq \varrho$ or $\max_{s'} |\mathbb{P}_h(s' \mid s, a, b) - \mathbb{P}_h(s' \mid s, a)| \leq \varrho$, $\forall (s, a, b, h)$, with a misspecification error ϱ. One can still follow the techniques in this chapter to analyze such misspecified scenarios and obtain regrets with an extra bias term depending on the misspecification error ϱ. When ϱ is small, it implies that the MG has approximately factored independent transition or single-controller transition structures, and then the bias term depending on ϱ should be small.

4.3 Markov Game with Factored Independent Transition

In this section, we propose and analyze optimistic policy optimization algorithms for both players under the setting of the factored independent transition.

Algorithm for Player 1. The algorithm for Player 1 is illustrated in Algorithm 2. Assume that the game starts from a fixed state $s_1 = (s_1^1, s_1^2)$ each round. We also assume that the true transition model $\mathbb{P}$ is not known to Player 1, and Player 1 can only access the bandit feedback of the rewards along this trajectory instead of the full information. Thus, Player 1 needs to empirically estimate the reward function and the transition model for all (s, a, b, s') and $h \in [H]$ via

$$\begin{aligned}
\widehat{r}_h^k(s, a, b) &= \frac{\sum_{\tau=1}^k \mathbb{1}\{(s, a, b) = (s_h^\tau, a_h^\tau, b_h^\tau)\} r_h^k(s, a, b)}{\max\{N_h^k(s, a, b), 1\}}, \\
\widehat{\mathbb{P}}_h^{1,k}(s^{1\prime}|s^1, a) &= \frac{\sum_{\tau=1}^k \mathbb{1}\{(s^1, a, s^{1\prime}) = (s_h^{1,\tau}, a_h^\tau, s_{h+1}^{1,\tau})\}}{\max\{N_h^k(s^1, a), 1\}}, \\
\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime}|s^2, b) &= \frac{\sum_{\tau=1}^k \mathbb{1}\{(s^2, b, s^{2\prime}) = (s_h^{2,\tau}, b_h^\tau, s_{h+1}^{2,\tau})\}}{\max\{N_h^k(s^2, b), 1\}}, \\
\widehat{\mathbb{P}}_h^k(s'|s, a, b) &= \widehat{\mathbb{P}}_h^{1,k}(s^{1\prime}|s^1, a)\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime}|s^2, b),
\end{aligned} \tag{4.6}$$

where we denote $\mathbb{1}\{\cdot\}$ as an indicator function, and $N_h^k(s, a, b)$ counts the empirical number of observation for a certain tuple (s, a, b) at step h until the k-th iteration as well as $N_h^k(s^1, a)$ for (s^1, a) and $N_h^k(s^2, b)$ for (s^2, b). For simplicity of presentation, in this chapter, we let $s = (s^1, s^2)$ and we use s^1, s^2 separately when necessary.

Based on the estimations of the transition model and reward function, we further estimate the Q-function and value-function as shown in Lines 7 and 8 in Algorithm 2. In terms of the principle of "optimism in the face of uncertainty", bonus terms are introduced to construct an estimated

Q-function as shown in Line 7 of Algorithm 2. We set the bonus term as

$$\beta_h^k(s,a,b) = \beta_h^{r,k}(s,a,b) + \beta_h^{\mathbb{P},k}(s,a,b), \tag{4.7}$$

where we define

$$\beta_h^{r,k}(s,a,b) := \sqrt{\frac{4\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}},$$
$$\beta_h^{\mathbb{P},k}(s,a,b) := \sqrt{\frac{2H^2|\mathcal{S}_1|\log(2|\mathcal{S}_1||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s^1,a),1\}}} + \sqrt{\frac{2H^2|\mathcal{S}_2|\log(2|\mathcal{S}_2||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s^2,b),1\}}},$$

with $\delta \in (0,1)$. Here, we decompose $\beta_h^k(s,a,b)$ into two terms where $\beta_h^{r,k}(s,a,b)$ is the bonus term for the reward and $\beta_h^{\mathbb{P},k}(s,a)$ for the transition estimation. As shown in Lemmas 4.10 and 4.11, the bonus terms $\beta_h^{r,k}(s,a,b)$ and $\beta_h^{\mathbb{P},k}(s,a,b)$ are obtained by using Hoeffding's inequality. Note that the two terms in the definition of $\beta_h^{\mathbb{P},k}$ stem from the uncertainties of estimating both transitions $\mathbb{P}_h^1(s^{1\prime} \mid s^1, a)$ and $\mathbb{P}_h^2(s^{2\prime} \mid s^2, b)$.

Next, we introduce the notion of the state reaching probability $q^{\nu^k,\mathbb{P}^2}(s^2)$ for any state $s^2 \in \mathcal{S}_2$ under the policy ν^k and the true transition $\mathbb{P}^2$, which is defined as

$$q_h^{\nu^k,\mathbb{P}^2}(s^2) := \Pr(s_h^2 = s^2 \mid \nu^k, \mathbb{P}^2, s_1^2), \forall h \in [H].$$

To handle non-stationarity of the opponent, as in Line 10, Player 1 needs to estimate the state reaching probability of Player 2 by the empirical reaching probability under the empirical transition model $\widehat{\mathbb{P}}^{2,k}$ for Player 2, i.e.,

$$d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2) = \Pr(s_h^2 = s^2 \mid \nu^k, \widehat{\mathbb{P}}^{2,k}, s_1^2), \quad \forall h \in [H].$$

The empirical reaching probability can be simply computed dynamically from $h = 1$ to H by $d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2) = \sum_{s^{2\prime}\in\mathcal{S}_2}\sum_{b'\in\mathcal{B}} d_{h-1}^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^{2\prime})\nu_{h-1}^k(b'|s^{2\prime})\widehat{\mathbb{P}}_{h-1}^{2,k}(s^2|s^{2\prime},b')$. Based on the estimated state reaching probability, the policy improvement step is associated with solving the following optimization problem (denoting by D_{KL} the KL divergence)

$$\underset{\pi}{\operatorname{maximize}} \sum_{h=1}^{H} [\overline{G}_h^{k-1}(\pi_h) - \eta^{-1} D_{\mathrm{KL}}(\pi_h(\cdot|s^1), \pi_h^{k-1}(\cdot|s^1))], \tag{4.8}$$

where $\overline{G}_h^{k-1}(\pi_h) := \langle \pi_h(\cdot|s^1) - \pi_h^{k-1}(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k-1}(s^1,s^2,\cdot) d_h^{\nu^{k-1},\widehat{\mathbb{P}}^{2,k-1}}(s^2)\rangle_{\mathcal{A}}$ with letting $F_h^{1,k-1}(s^1,s^2,a) = \langle \overline{Q}_h^{k-1}(s^1,s^2,a,\cdot), \nu_h^{k-1}(\cdot|s^2)\rangle_{\mathcal{B}}$. One can see that (4.8) is a mirror ascent step

Algorithm 2 Optimistic Policy Optimization for Player 1

```
Initialize: For all $h \in [H]$, $(s^1, s^2, a, b) \in \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{A} \times \mathcal{B}$: $\pi_h^0(\cdot|s^1) = \mathbf{1}/|\mathcal{A}|$, $\widehat{\mathbb{P}}_h^{1,0}(\cdot|s^1, a) = \mathbf{1}/|\mathcal{S}_1|$, $\widehat{\mathbb{P}}_h^{2,0}(\cdot|s^2, b) = \mathbf{1}/|\mathcal{S}_2|$, $\widehat{r}_h^0(\cdot,\cdot,\cdot) = \beta_h^0(\cdot,\cdot,\cdot) = \mathbf{0}$.
for episode $k = 1, \ldots, K$ do
    Observe Player 2's policy $\{\nu_h^{k-1}\}_{h=1}^H$.
    Start from state $s_1 = (s_1^1, s_1^2)$, set $\overline{V}_{H+1}^{k-1}(\cdot) = \mathbf{0}$.
    for step $h = H, H-1, \ldots, 1$ do
        Estimate the transition and reward function by $\widehat{\mathbb{P}}_h^{k-1}(\cdot|\cdot,\cdot)$ and $\widehat{r}_h^{k-1}(\cdot,\cdot,\cdot)$ as (4.6).
        Update Q-function $\forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$:
            $$\overline{Q}_h^{k-1}(s,a,b) = \min\{(\widehat{r}_h^{k-1} + \widehat{\mathbb{P}}_h^{k-1}\overline{V}_{h+1}^{k-1} + \beta_h^{k-1})(s,a,b), H-h+1\}^+.$$
        Update value-function $\forall s \in \mathcal{S}$:
            $$\overline{V}_h^{k-1}(s) = \big[\pi_h^{k-1}(\cdot|s)\big]^\top \overline{Q}_h^{k-1}(s,\cdot,\cdot)\nu_h^{k-1}(\cdot|s).$$
    end for
    Estimate the state reaching probability of Player 2 by $d_h^{\nu^{k-1},\widehat{\mathbb{P}}^{2,k-1}}(s^2)$, $\forall s^2 \in \mathcal{S}_2, h \in [H]$.
    Update policy $\pi_h^k(a|s^1)$ by solving (4.8), $\forall (s^1, a) \in \mathcal{S}_1 \times \mathcal{A}, h \in [H]$.
    Take actions following $a_h^k \sim \pi_h^k(\cdot|s_h^{1,k})$, $\forall h \in [H]$.
    Observe the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$, and rewards $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$.
end for
```

and admits a closed-form solution for all $(h, s^1, a) \in [H] \times \mathcal{S}_1 \times \mathcal{A}$ as follows

$$\pi_h^k(a|s^1) = (\overline{Y}_h^{k-1})^{-1}\pi_h^{k-1}(a \mid s^1) \cdot \exp\Big\{\eta \sum_{s^2 \in \mathcal{S}_2} F_h^{1,k-1}(s^1, s^2, a) d_h^{\nu^{k-1},\widehat{\mathbb{P}}^{2,k-1}}(s^2)\Big\},$$

where $\overline{Y}_h^{k-1}$ is a probability normalization term.

Algorithm for Player 2. For the setting of MG with factored independent transition, the algorithm for Player 2 is trying to minimize the expected cumulative reward w.r.t. $r_h(\cdot,\cdot,\cdot)$. In another word, Player 2 is maximizing the expected cumulative reward w.r.t. $-r_h(\cdot,\cdot,\cdot)$. From this perspective, one can view the algorithm for Player 2 as a *'symmetric'* version of Algorithm 2. The algorithm for Player 2 is summarized in Algorithm 3. Specifically, in this algorithm, Player 2 also estimates the transition model and the reward function the same as (4.10). Since Player 2 is minimizing the expected cumulative reward, the bonus terms as (4.7) are subtracted in the Q-function estimation step by the optimism principle. The algorithm further estimates the state reaching probability of Player 1, $q_h^{\pi^k,\mathbb{P}^1}(s^1)$, by the empirical one $d_h^{\pi^k,\widehat{\mathbb{P}}^{1,k}}(s^1)$, which can be dynamically computed. For the policy improvement step, Algorithm 3 performs a mirror descent step based on the empirical

Algorithm 3 Optimistic Policy Optimization for Player 2

```
**Initialize:** For all $h \in [H]$, $(s^1, s^2, a, b) \in \mathcal{S}_1 \times \mathcal{S}_2 \times \mathcal{A} \times \mathcal{B}$: $\pi_h^0(\cdot|s^1) = \mathbf{1}/|\mathcal{A}|$, $\widehat{\mathbb{P}}_h^{1,0}(\cdot|s^1, a) = \mathbf{1}/|\mathcal{S}_1|$, $\widehat{\mathbb{P}}_h^{2,0}(\cdot|s^2, b) = \mathbf{1}/|\mathcal{S}_2|$, $\widehat{r}_h^0(\cdot,\cdot,\cdot) = \beta_h^0(\cdot,\cdot,\cdot) = \mathbf{0}$.
**for** episode $k = 1, \ldots, K$ **do**
    Observe Player 1's policy $\{\pi_h^{k-1}\}_{h=1}^H$.
    Start from state $s_1 = (s_1^1, s_1^2)$, set $\overline{V}_{H+1}^{k-1}(\cdot) = \mathbf{0}$.
    **for** step $h = H, H-1, \ldots, 1$ **do**
        Estimate the transition and reward function by $\widehat{\mathbb{P}}_h^{k-1}(\cdot|\cdot,\cdot)$ and $\widehat{r}_h^{k-1}(\cdot,\cdot,\cdot)$ as (4.6).
        Update Q-function $\forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$:
            $\underline{Q}_h^{k-1}(s, a, b) = \min\{(\widehat{r}_h^{k-1} + \widehat{\mathbb{P}}_h^{k-1}\underline{V}_{h+1}^{k-1} - \beta_h^{k-1})(s, a, b), H - h + 1\}^+.$
        Update value-function $\forall s \in \mathcal{S}$:
            $\underline{V}_h^{k-1}(s) = \big[\pi_h^{k-1}(\cdot|s)\big]^\top \underline{Q}_h^{k-1}(s, \cdot, \cdot)\nu_h^{k-1}(\cdot|s).$
    **end for**
    Estimate the state reaching probability of Player 1 by $d_h^{\pi^{k-1}, \widehat{\mathbb{P}}^{1,k-1}}(s^1)$, $\forall s^1 \in \mathcal{S}_1, h \in [H]$.
    Update policy $\nu_h^k(b|s^2)$ by solving (4.9), $\forall (s^2, b) \in \mathcal{S}_2 \times \mathcal{B}, h \in [H])$.
    Take actions following $b_h^k \sim \nu_h^k(\cdot|s_h^{2,k})$, $\forall h \in [H]$.
    Observe the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$, and rewards $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$.
**end for**
```

reaching probability, which is associated with solving the following optimization problem

$$\underset{\pi}{\text{minimize}} \sum_{h=1}^{H} [\underline{G}_h^{k-1}(\nu_h) + \gamma^{-1} D_{\mathrm{KL}}(\nu_h(\cdot|s^2), \nu_h^{k-1}(\cdot|s^2))], \tag{4.9}$$

where $\underline{G}_h^{k-1}(\pi_h) := \langle \nu_h(\cdot|s^2) - \nu_h^{k-1}(\cdot|s^2), \sum_{s^1 \in \mathcal{S}_1} F_h^{2,k-1}(s^1, s^2, \cdot) d_h^{\pi^{k-1}, \widehat{\mathbb{P}}^{1,k-1}}(s^1) \rangle_{\mathcal{B}}$ with letting $F_h^{2,k-1}(s^1, s^2, b) = \langle \underline{Q}_h^{k-1}(s^1, s^2, \cdot, b), \pi_h^{k-1}(\cdot|s^1) \rangle_{\mathcal{A}}$. Here (4.9) is a standard mirror descent step and admits a closed-form solution for all $(h, s^2, b) \in [H] \times \mathcal{S}_2 \times \mathcal{B}$ as follows

$$\nu_h^k(b|s^2) = (\underline{Y}_h^{k-1})^{-1} \nu_h^{k-1}(b \,|\, s^2) \cdot \exp\Big\{ -\gamma \sum_{s^1 \in \mathcal{S}_1} F_h^{2,k-1}(s^1, s^2, b) d_h^{\pi^{k-1}, \widehat{\mathbb{P}}^{1,k-1}}(s^1) \Big\},$$

where $\underline{Y}_h^{k-1}$ is a probability normalization term.

4.3.1 Main Results

In this subsection, we show our main results of the upper bounds of the regrets for each player under the setting of the factored independent transition model.

Theorem 4.2. *By setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, with probability at least $1-4\delta$, Algorithm 2 ensures the sublinear regret bound for Player 1*[3] *i.e.,* $\mathrm{Regret}_1(K) \leq \widetilde{\mathcal{O}}\big(C\sqrt{T}\big)$, *where $T = KH$ is the number of steps, and the constant factor is $C = \sqrt{(|\mathcal{S}_1|^2|\mathcal{A}| + |\mathcal{S}_2|^2|\mathcal{B}|)H^3} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H}$.*

Theorem 4.2 shows that Player 1 can obtain an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret by Algorithm 2, when the opponent, Player 2, takes actions following potentially adversarial policies.

Theorem 4.3. *By setting $\gamma = \sqrt{\log|\mathcal{B}|/(KH^2)}$, with probability at least $1-4\delta$, Algorithm 3 ensures the sublinear regret bound for Player 2, i.e.,* $\mathrm{Regret}_2(K) \leq \widetilde{\mathcal{O}}\big(C\sqrt{T}\big)$, *where $T = KH$ is the number of steps, and the constant factor is $C = \sqrt{(|\mathcal{S}_1|^2|\mathcal{A}| + |\mathcal{S}_2|^2|\mathcal{B}|)H^3} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H}$.*

Theorem 4.3 shows that $\mathrm{Regret}_2(K)$ admits the same $\widetilde{\mathcal{O}}(\sqrt{K})$ regret as Theorem 4.2 given any arbitrary and adversarial policies of the opponent Player 1, due to the symmetric nature of the two algorithms.

From the perspective of each individual player, the game can be viewed as a special case of an MDP with adversarial bandit feedback due to the potentially adversarial actions or policies of the opponent. For MDPs with adversarial bandit feedback, Jin et al. [2019] attains an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret via an occupancy measure based method, which requires solving a constrained optimization problem in each update step that is more computationally demanding than PO. Efroni et al. [2020b] proposes a PO method for the same MDP model, achieving an $\widetilde{\mathcal{O}}(K^{2/3})$ regret. Thus, directly applying this result would yield an $\widetilde{\mathcal{O}}(K^{2/3})$ regret. However, for the problem of zero-sum games, regarding the problem faced by one player as an MDP with adversarial rewards neglects the fact that such "adversarial reward functions" are determined by the actions and policies of the opponent. Thus, under the FP framework, by utilizing the past actions and policies of the opponent, Algorithm 2 and 3 obtain an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret.

In particular, if Player 1 runs Algorithm 2 and Player 2 runs Algorithm 3 *simultaneously*, then we have the following corollary of Theorems 4.2 and 4.3.

Corollary 4.4. *By setting η and γ as in Theorem 4.2 and Theorem 4.3, letting $T = KH$, with probability at least $1-8\delta$, Algorithm 2 and Algorithm 3 ensures the following optimality gap* $\mathrm{Gap}(K) \leq \widetilde{\mathcal{O}}\big(\sqrt{T}\big)$.

4.4 Markov Game with Single-Controller Transition

In this section, we propose and analyze optimistic policy optimization algorithms for the single-controller game.

[3] Hereafter, we use $\widetilde{\mathcal{O}}$ to hide the logarithmic factors on $|\mathcal{S}|$, $|\mathcal{A}|$, $|\mathcal{B}|$, H, K, and $1/\delta$.

Algorithm for Player 1. The algorithm for Player 1 is illustrated in Algorithm 4. Since transition model is unknown and only bandit feedback of the rewards is available, Player 1 needs to empirically estimate the reward function and the transition model for all (s, a, b, s') and $h \in [H]$ via

$$\begin{aligned} \widehat{r}_h^k(s,a,b) &= \frac{\sum_{\tau=1}^k \mathbb{1}\{(s,a,b) = (s_h^\tau, a_h^\tau, b_h^\tau)\} r_h^k(s,a,b)}{\max\{N_h^k(s,a,b), 1\}}, \\ \widehat{\mathbb{P}}_h^k(s'|s,a) &= \frac{\sum_{\tau=1}^k \mathbb{1}\{(s,a,s') = (s_h^\tau, a_h^\tau, s_{h+1}^\tau)\}}{\max\{N_h^k(s,a), 1\}}. \end{aligned} \tag{4.10}$$

Based on the estimations, Algorithm 4 further estimates the Q-function and value-function for policy evaluation. In terms of the optimism principle, the bonus term is added to construct an estimated Q-function as shown in Line 7 of Algorithm 4. The bonus terms are computed as

$$\beta_h^k(s,a,b) = \beta_h^{r,k}(s,a,b) + \beta_h^{\mathbb{P},k}(s,a), \tag{4.11}$$

where the two bonus terms above are expressed as

$$\beta_h^{r,k}(s,a,b) := \sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}, \qquad \beta_h^{\mathbb{P},k}(s,a) := \sqrt{\frac{2H^2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}},$$

for $\delta \in (0, 1)$. Here we also decompose $\beta_h^k(s,a,b)$ into two terms with $\beta_h^{r,k}(s,a,b)$ denoting the bonus term for the reward and $\beta_h^{\mathbb{P},k}(s,a)$ for the transition estimation. Note that the transition bonus are only associated with (s, a) due to the single-controller structure. The bonus terms are derived in Lemmas 4.23 and 4.24.

Different from Algorithm 2, in this algorithm for Player 1, there is no need to estimate the state reaching probability of the opponent as the transition only depends on Player 1. The policy improvement step is then associated with solving the following optimization problem

$$\underset{\pi}{\text{maximize}} \sum_{h=1}^H [\overline{L}_h^{k-1}(\pi_h) - \eta^{-1} D_{\mathrm{KL}}\big(\pi_h(\cdot|s), \pi_h^{k-1}(\cdot|s)\big)], \tag{4.12}$$

where we define the function $\overline{L}_h^{k-1}(\pi_h) := \big[\pi_h(\cdot|s) - \pi_h^{k-1}(\cdot|s)\big]^\top \overline{Q}_h^{k-1}(s,\cdot,\cdot)\nu_h^{k-1}(\cdot|s)$. This is a mirror ascent step and admits the closed-form solution for all $(h, s, a) \in [H] \times \mathcal{S} \times \mathcal{A}$ as follows

$$\pi_h^k(a|s) = (\overline{Z}_h^{k-1})^{-1} \pi_h^{k-1}(a\,|\,s) \exp\{\eta \langle \overline{Q}_h^{k-1}(s,a,\cdot), \nu_h^{k-1}(\cdot\,|\,s)\rangle_{\mathcal{B}}\},$$

where $\overline{Z}_h^{k-1}$ is a probability normalization term.

Algorithm 4 Optimistic Policy Optimization for Player 1

```
Initialize: $\pi_h^0(\cdot|s) = \mathbf{1}/|\mathcal{A}|$ for all $s \in \mathcal{S}$ and $h \in [H]$. $\widehat{\mathbb{P}}_h^0(\cdot|s,a) = \mathbf{1}/|\mathcal{S}|$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$. $\widehat{r}_h^0(\cdot,\cdot,\cdot) = \beta_h^0(\cdot,\cdot,\cdot) = \mathbf{0}$ for all $h \in [H]$.
for episode $k = 1, \dots, K$ do
    Observe Player 2's policy $\{\nu_h^{k-1}\}_{h=1}^H$.
    Start from $s_1^k = s_1$, and set $\overline{V}_{H+1}^{k-1}(\cdot) = \mathbf{0}$.
    for step $h = H, H-1, \dots, 1$ do
        Estimate the transition and reward function by $\widehat{\mathbb{P}}_h^{k-1}(\cdot|\cdot,\cdot)$ and $\widehat{r}_h^{k-1}(\cdot,\cdot,\cdot)$ as (4.10).
        Update Q-function $\forall(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$:
            $$\overline{Q}_h^{k-1}(s,a,b) = \min\{\widehat{r}_h^{k-1}(s,a,b) + \widehat{\mathbb{P}}_h^{k-1}\overline{V}_{h+1}^{k-1}(s,a) + \beta_h^{k-1}(s,a,b), H-h+1\}^+.$$
        Update value-function $\forall s \in \mathcal{S}$:
            $$\overline{V}_h^{k-1}(s) = \big[\pi_h^{k-1}(\cdot|s)\big]^\top \overline{Q}_h^{k-1}(s,\cdot,\cdot)\nu_h^{k-1}(\cdot|s).$$
    end for
    Update policy $\pi_h^k(a|s)$ by solving (4.12), $\forall(s,a) \in \mathcal{S}\times\mathcal{A}, h \in [H]$.
    Take actions following $a_h^k \sim \pi_h^k(\cdot|s_h^k)$, $\forall h \in [H]$.
    Observe the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$, and rewards $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$.
end for
```

Algorithm for Player 2. The algorithm for Player 2 is illustrated in Algorithm 5. Player 2 also estimates the transition model and the reward function the same as (4.10). However, due to the *asymmetric* nature of the single-controller transition model, Player 2 has a different way to learning the policy. The main differences to Algorithm 4 are summarized in the following three aspects: First, according to our theoretical analysis shown in Lemma 4.21, no transition model estimation is involved. Instead, only a reward function estimation is considered in Line 7 of Algorithm 5. Second, in the policy improvement step, Player 2 needs to approximate the state reaching probability $q_h^{\pi^k,\mathbb{P}}(s) := \Pr(s_h = s \,|\, \pi^k, \mathbb{P}, s_1)$ under π^k and true transition $\mathbb{P}$ by the empirical reaching probability $d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s) = \Pr(s_h = s \,|\, \pi^k, \widehat{\mathbb{P}}^k, s_1)$ with the empirical transition model $\widehat{\mathbb{P}}^k$, which can be computed dynamically from $h = 1$ to H. Third, we subtract a reward bonus term $\beta_h^{r,k-1}$ in Line 7 instead of adding the bonus. Similar to our discussion in Section 4.3, it is still a UCB estimation if viewing Player 2 is maximizing the cumulative reward w.r.t. a negative reward function $-r$.

Particularly, the policy improvement step of Algorithm 5 is associated with solving the following minimization problem

$$\underset{\nu}{\text{minimize}} \sum_{h=1}^{H} \{\underline{L}_h^{k-1}(\nu_h) + \gamma^{-1} D_{\mathrm{KL}}\big(\nu_h(\cdot|s), \nu_h^{k-1}(\cdot|s)\big)\}, \tag{4.13}$$

Algorithm 5 Optimistic Policy Optimization for Player 2

```
Initialize: $\nu_h^0(\cdot|s) = \mathbf{1}/|\mathcal{B}|$ for all $s \in \mathcal{S}$ and $h \in [H]$. $\widehat{\mathbb{P}}_h^0(\cdot|s,a) = \mathbf{1}/|\mathcal{S}|$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $h \in [H]$. $\widehat{r}_h^0(\cdot,\cdot,\cdot) = \beta_h^{r,0}(\cdot,\cdot,\cdot) = \mathbf{0}$ for all $h \in [H]$.
for episode $k = 1, \ldots, K$ do
    Observe Player 1's policy $\{\pi_h^{k-1}\}_{h=1}^H$.
    Start from the initial state $s_1^k = s_1$.
    for step $h = 1, 2, \ldots, H$ do
        Estimate the transition and reward function by $\widehat{\mathbb{P}}_h^{k-1}$ and $\widehat{r}_h^{k-1}$ as (4.10).
        Update $\widetilde{r}_h^{k-1}$, $\forall (s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$:
            $$\widetilde{r}_h^{k-1}(s,a,b) = \max\big\{\widehat{r}_h^{k-1}(s,a,b) - \beta_h^{r,k-1}(s,a,b), 0\big\}.$$
        Estimate the state reaching probability by $d_h^{\pi^{k-1},\widehat{\mathbb{P}}^{k-1}}(s)$, $\forall s \in \mathcal{S}, h \in [H]$.
    end for
    Update policy $\nu_h^k(b|s)$ by solving (4.13), $\forall (s,b) \in \mathcal{S}\times\mathcal{B}, h \in [H]$.
    Take actions following $b_h^k \sim \nu_h^k(\cdot|s_h^k), \forall h \in [H]$.
    Observe the trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$, and rewards $\{r_h^k(s_h^k, a_h^k, b_h^k)\}_{h=1}^H$.
end for
```

where we define $\underline{L}_h^{k-1}(\nu_h) := d_h^{\pi^{k-1},\widehat{\mathbb{P}}^{k-1}}(s)[\pi_h^{k-1}(\cdot|s)]^\top \cdot \widetilde{r}_h^{k-1}(s,\cdot,\cdot)[\nu_h(\cdot|s) - \nu_h^{k-1}(\cdot|s)]$. This is a mirror descent step with the closed-form solution for all $(h,s,b) \in [H]\times\mathcal{S}\times\mathcal{B}$ as

$$\nu_h^k(b|s) = (\underline{Z}_h^{k-1})^{-1} \cdot \nu_h^{k-1}(b\,|\,s)\exp\{-\gamma d_h^{\pi^{k-1},\widehat{\mathbb{P}}^{k-1}}(s)\langle \widetilde{r}_h^{k-1}(s,\cdot,b), \pi_h^{k-1}(\cdot\,|\,s)\rangle_{\mathcal{A}}\},$$

with the denominator $\underline{Z}_h^{k-1}$ being a normalization term.

4.4.1 Main Results

Next, we present the main results of the regrets for the single-controller transition model.

Theorem 4.5. *By setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, with probability at least $1-3\delta$, Algorithm 4 ensures the following regret bound for Player 1 $\mathrm{Regret}_1(K) \le \widetilde{\mathcal{O}}\big(C\sqrt{T}\big)$, where $T = KH$ is the total number of steps, and the constant factor is $C = \sqrt{|\mathcal{S}|^2|\mathcal{A}|H^3} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H}$.*

Theorem 4.5 shows that $\mathrm{Regret}_1(K)$ is in the level of $\widetilde{\mathcal{O}}(\sqrt{K})$, for arbitrary policies of Player 2. Similar to the discussion after Theorem 4.3, from the perspective of Player 1, the game can also be viewed as a special case of an MDP with adversarial bandit feedback. Under the FP framework, by utilizing the past policies of Player 2, Algorithm 4 can achieve an $\widetilde{\mathcal{O}}(\sqrt{K})$ regret, comparing to $\widetilde{\mathcal{O}}(K^{2/3})$ regret by the PO method [Efroni et al., 2020b] and $\widetilde{\mathcal{O}}(\sqrt{K})$ regret by a computationally demanding non-PO method [Jin et al., 2019] for MDP with adversarial rewards.

Theorem 4.6. *By setting $\gamma = \sqrt{|\mathcal{S}| \log |\mathcal{B}|/K}$, with probability at least $1-2\delta$, Algorithm 5 ensures the sublinear regret bound for Player 2, i.e.,* $\mathrm{Regret}_2(K) \leq \widetilde{\mathcal{O}}(C\sqrt{T})$*, where $T = KH$ is the total number of steps, and the constant factor is $C = \sqrt{|\mathcal{S}|^2|\mathcal{A}|H^3} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H}$.*

Interestingly, Theorem 4.6 also shows that $\mathrm{Regret}_2(K)$ has the same bound (including the constant factor C) as $\mathrm{Regret}_1(K)$ given any opponent's policy, though the transition model bonus is not involved in Algorithm 5 and the learning process for two players are essentially different. In fact, although the bonus term for estimating the transition is not involved in this algorithm, approximating the state reaching probability of Player 1 implicitly reflects the gap between the empirical transition $\widehat{\mathbb{P}}^k$ and the true transition $\mathbb{P}$, which can explain the same upper bound in Theorems 4.5 and 4.6.

Moreover, if Player 1 runs Algorithm 4 and Player 2 runs Algorithm 5 *simultaneously*, we have the following corollary of the above two theorems.

Corollary 4.7. *By setting η and γ as in Theorem 4.5 and Theorem 4.6, letting $T = KH$, with probability at least $1 - 5\delta$, Algorithm 4 and Algorithm 5 ensures the optimality gap* $\mathrm{Gap}(K) \leq \widetilde{\mathcal{O}}(\sqrt{T})$.

4.5 Theoretical Analysis

4.5.1 Proofs of Theorems 4.2 and 4.3

Proof. To bound $\mathrm{Regret}_1(K)$, we need to analyze the value function difference for the instantaneous regret at the k-th episode, i.e., $V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1)$. By Lemma 4.8, we decompose the difference between $V_1^{\pi^*,\nu^k}(s_1)$ and $V_1^{\pi^k,\nu^k}(s_1)$ into four terms

$$
\begin{aligned}
&V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1) \\
&\quad \leq \underbrace{\overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1)}_{\mathrm{Err}_k(\mathrm{I.1})} + \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\{[\pi_h^*(\cdot|s_h)]^\top \overline{\iota}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h) \,|\, s_1\}}_{\mathrm{Err}_k(\mathrm{I.2})} \\
&\qquad + \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}^1}\{\langle \pi_h^*(\cdot|s_h^1) - \pi_h^k(\cdot|s_h^1), M_h^k(s_h^1,\cdot)\rangle_{\mathcal{A}} \,|\, s_1\}}_{\mathrm{Err}_k(\mathrm{I.3})} + \underbrace{2H\sum_{h=1}^{H}\sum_{s_h^2\in\mathcal{S}_2} |q_h^{\nu^k,\mathbb{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)|}_{\mathrm{Err}_k(\mathrm{I.4})},
\end{aligned}
$$

where $M_h^k(s_h^1,\cdot) := \sum_{s_h^2\in\mathcal{S}_2} F_h^{1,k}(s_h^1,s_h^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)$. Here we define the model prediction error of Q-function as $\overline{\iota}_h^k(s,a,b) = r_h(s,a,b) + \mathbb{P}_h\overline{V}_{h+1}^k(s,a,b) - \overline{Q}_h^k(s,a,b)$. Let s_h^1, s_h^2, a_h, b_h be random variables for states and actions.

Specifically, $\mathrm{Err}_k(\mathrm{I.1})$ is the difference between the estimated value function and the true value function, $\mathrm{Err}_k(\mathrm{I.2})$ is associated with the model prediction error $\bar{\iota}_h^k(s,a,b)$ of Q-function, $\mathrm{Err}_k(\mathrm{I.3})$ is the error from the policy mirror ascent step, and $\mathrm{Err}_k(\mathrm{I.4})$ is the error related to the reaching probability estimation. According to Lemmas 4.9, 4.13, 4.15, we have that $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{I.1}) \leq \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}_1|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}_2|^2|\mathcal{B}|H^4K} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H^2K})$, the third error term is bounded as $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{I.3}) \leq \mathcal{O}(\sqrt{H^4K\log|\mathcal{A}|})$, and the last error term is bounded as $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{I.4}) \leq \widetilde{\mathcal{O}}(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K})$. Moreover, as shown in Lemma 4.12, since the estimated Q-function is a UCB estimate, then we have that the model prediction error $\bar{\iota}_h^k(s,a,b) \leq 0$ with high probability, which leads to $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{I.2}) \leq 0$. This shows the significance of the principle of "optimism in the face of uncertainty". By the union bound, all the above inequalities hold with probability at least $1-4\delta$. Therefore, letting $T = KH$, by the relation that $\mathrm{Regret}_1(K) = \sum_{k=1}^K [V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1)] \leq \sum_{k=1}^K [\mathrm{Err}_k(\mathrm{I.1}) + \mathrm{Err}_k(\mathrm{I.2}) + \mathrm{Err}_k(\mathrm{I.3}) + \mathrm{Err}_k(\mathrm{I.4})]$, we can obtain the result in Theorem 4.2.

Due to the symmetry of Algorithm 2 and Algorithm 3 as we discussed in Section 4.3, the proof for Theorem 4.3 exactly follows the proof of Theorem 4.3. This completes the proof. □

4.5.2 Proofs of Theorems 4.5 and 4.6

Proof. We first show the proof of Theorem 4.5. By lemma 4.20, we have

$$V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1) \leq \underbrace{\overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1)}_{\mathrm{Err}_k(\mathrm{II.1})} + \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\left[\bar{\varsigma}_h^k(s_h,a_h,b_h) \,\middle|\, s_1\right]}_{\mathrm{Err}_k(\mathrm{II.2})}$$
$$+ \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}}[\langle \pi_h^*(\cdot|s_h) - \pi_h^k(\cdot|s_h), U_h^k(s_h,\cdot)\rangle_{\mathcal{A}} \,|\, s_1]}_{\mathrm{Err}_k(\mathrm{II.3})},$$

where s_h, a_h, b_h are random variables for states and actions, $U_h^k(s,a) := \langle \overline{Q}_h^k(s,a,\cdot), \nu_h^k(\cdot\,|\,s)\rangle_{\mathcal{B}}$, and we define the model prediction error of Q-function as $\bar{\varsigma}_h^k(s,a,b) = r_h(s,a,b) + \mathbb{P}_h\overline{V}_{h+1}^k(s,a) - \overline{Q}_h^k(s,a,b)$.

Particularly, $\mathrm{Err}_k(\mathrm{II.1})$ is the difference between the estimated value function and the true value function, $\mathrm{Err}_k(\mathrm{II.2})$ is associated with the model prediction error $\bar{\varsigma}_h^k(s,a,b)$ for Q-function, and $\mathrm{Err}_k(\mathrm{II.3})$ characterizes the error from the policy mirror ascent step. As shown in Lemma 4.26, $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{II.1}) \leq \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K})$ with probability at least $1-\delta$. In addition, we have $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{II.2}) \leq 0$ with probability at least $1-2\delta$ as shown in Lemma 4.25, which is due to the UCB estimation of the Q-function. Furthermore, Lemma 4.22 shows the cumulative error for the mirror ascent step is $\sum_{k=1}^K \mathrm{Err}_k(\mathrm{II.3}) \leq \mathcal{O}(\sqrt{H^4K\log|\mathcal{A}|})$ with setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$. Therefore, letting $T = KH$, further by the relation that $\mathrm{Regret}_1(K) \leq$

$\sum_{k=1}^K[\mathrm{Err}_k(\text{II.1})+\mathrm{Err}_k(\text{II.2})+\mathrm{Err}_k(\text{II.3})]$, we can obtain the result in Theorem 4.5 with probability at least $1-3\delta$ by the union bound.

Next, we show the proof of Theorem 4.6. By Lemma 4.21, we can decompose the difference between $V_1^{\pi^k,\nu^k}(s_1)$ and $V_1^{\pi^k,\nu^*}(s_1)$ into four terms

$$
\begin{aligned}
&V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1) \\
&\quad \le \underbrace{2\sum_{h=1}^H \mathbb{E}_{\pi^k,\mathbb{P},\nu^k}[\beta_h^{r,k}(s_h,a_h,b_h)\,|\,s_1]}_{\mathrm{Err}_k(\text{III.1})} + \underbrace{\sum_{h=1}^H\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top \underline{\varsigma}_h^k(s,\cdot,\cdot)\nu_h^*(\cdot|s)}_{\mathrm{Err}_k(\text{III.2})} \\
&\qquad + \underbrace{\sum_{h=1}^H\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s)-\nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}}}_{\mathrm{Err}_k(\text{III.3})} + \underbrace{2\sum_{h=1}^H\sum_{s\in\mathcal{S}} |q_h^{\pi^k,\mathbb{P}}(s) - d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)|}_{\mathrm{Err}_k(\text{III.4})},
\end{aligned}
$$

with $W_h^k(s,b) = \langle \widetilde{r}_h^k(s,\cdot,b), \pi_h^k(\cdot\,|\,s)\rangle_{\mathcal{A}}$ and $\underline{\iota}_h^k(s,a,b) = \widetilde{r}_h^k(s,a,b) - r_h(s,a,b)$. The above inequality holds for all $k\in[K]$ with probability at least $1-\delta$. Due to the single-controller structure, distinct from the value function decomposition above for Theorem 4.5, here we have that $\mathrm{Err}_k(\text{III.1})$ is the expectation of reward bonus term, $\mathrm{Err}_k(\text{III.2})$ is associated with the reward prediction error $\underline{\varsigma}_h^k$, $\mathrm{Err}_k(\text{III.3})$ is the error from the policy mirror descent step, and $\mathrm{Err}_k(\text{III.4})$ is the difference between the true state reaching probability and the empirical one. Technically, in the proof of this decomposition, we can show $V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1) = \sum_{h=1}^H\sum_{s\in\mathcal{S}} q_h^{\pi^k,\mathbb{P}}(s)[\pi_h^k(\cdot|s)]^\top r_h(s,\cdot,\cdot)(\nu_h^k-\nu_h^*)(\cdot|s)$, where the value function difference is only related to the reward function $r_h(s,\cdot,\cdot)$ instead of the Q-function. This is the reason why only the reward bonus and reward-based mirror descent appear in Algorithm 5.

As shown in Lemmas 4.27, 4.30, and 4.31, we can obtain upper bounds that $\sum_{k=1}^K \mathrm{Err}_k(\text{III.1}) \le \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K})$, $\sum_{k=1}^K \mathrm{Err}_k(\text{III.3}) \le \mathcal{O}(\sqrt{H^2|\mathcal{S}|K\log|\mathcal{B}|})$, and also $\sum_{k=1}^K \mathrm{Err}_k(\text{III.4}) \le \widetilde{\mathcal{O}}(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K})$ by taking summation from $k=1$ to K for the three error terms $\mathrm{Err}_k(\text{III.1})$, $\mathrm{Err}_k(\text{III.2})$, $\mathrm{Err}_k(\text{III.3})$. For $\mathrm{Err}_k(\text{III.2})$, by Lemma 4.28, with probability at least $1-\delta$, we have that $\sum_{k=1}^K \mathrm{Err}_k(\text{III.1}) \le 0$, which is due to the UCB estimation of the reward function, i.e., $\widetilde{r}$. The above inequalities hold with probability at least $1-2\delta$ by the union bound. Therefore, letting $T=KH$, further by $\mathrm{Regret}_1(K) \le \sum_{k=1}^K[\mathrm{Err}_k(\text{III.1}) + \mathrm{Err}_k(\text{III.2}) + \mathrm{Err}_k(\text{III.3}) + \mathrm{Err}_k(\text{III.4})]$, we can obtain the result in Theorem 4.6. This completes the proof. □

4.6 Conclusion

In this chapter, we propose and analyze new optimistic policy optimization algorithms for two-player zero-sum Markov games with structured but unknown transitions. We consider two classes of transition structures: factored independent transition and single-controller transition. For both scenarios, we prove $\widetilde{\mathcal{O}}(\sqrt{T})$ regret bounds for each player after T steps in a two-agent competitive game scenario. When both players adopt the proposed algorithms, their overall optimality gap is $\widetilde{\mathcal{O}}(\sqrt{T})$.

4.7 Proofs for Markov Game with Factored Independent Transition

Lemma 4.8. *At the k-th episode of Algorithm 2, the difference between value functions $V_1^{\pi^*,\nu^k}(s_1)$ and $V_1^{\pi^k,\nu^k}(s_1)$ is bounded as*

$$
\begin{aligned}
&V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1) \\
&\quad = \overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1) + \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\big\{[\pi_h^*(\cdot|s_h)]^\top \bar{\iota}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h) \,\big|\, s_1\big\} \\
&\qquad + \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}^1}\Big\{\Big\langle \pi_h^*(\cdot|s_h^1) - \pi_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^{1,k}(s_h^1,s_h^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}} \,\Big|\, s_1^1, s_1^2\Big\} \\
&\qquad + 2H \sum_{h=1}^{H} \sum_{s_h^2\in\mathcal{S}_2} \Big| q_h^{\nu^k,\mathbb{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big|,
\end{aligned}
$$

where s_h, a_h, b_h are random variables for state and actions, $F_h^{1,k}(s^1,s^2,a) := \langle \overline{Q}_h^k(s^1,s^2,a,\cdot), \nu_h^k(\cdot|s^2)\rangle_{\mathcal{B}}$, and we define the model prediction error of Q-function as

$$
\bar{\iota}_h^k(s,a,b) = r_h(s,a,b) + \mathbb{P}_h \overline{V}_{h+1}^k(s,a,b) - \overline{Q}_h^k(s,a,b). \tag{4.14}
$$

Proof. The proof of this lemma starts with decomposing the value function difference as

$$
V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1) = V_1^{\pi^*,\nu^k}(s_1) - \overline{V}_1^k(s_1) + \overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1). \tag{4.15}
$$

Here the term $\overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1)$ is the bias between the estimated value function $\overline{V}_1^k(s_1)$ generated by Algorithm 2 and the value function $V_1^{\pi^k,\nu^k}(s_1)$ under the true transition model $\mathbb{P}$ at the k-th episode. We first analyze the term $V_1^{\pi^*,\nu^k}(s_1) - \overline{V}_1^k(s_1)$. For any h and s, we consider to decompose the term $V_h^{\pi^*,\nu^k}(s) - \overline{V}_h^k(s)$, which gives

$$\begin{aligned}
&V_h^{\pi^*,\nu^k}(s) - \overline{V}_h^k(s) \\
&\qquad = [\pi_h^*(\cdot|s)]^\top Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\pi_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) \\
&\qquad = [\pi_h^*(\cdot|s)]^\top Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - [\pi_h^*(\cdot|s)]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) \\
&\qquad\quad + [\pi_h^*(\cdot|s)]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\pi_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) \\
&\qquad = [\pi_h^*(\cdot|s)]^\top \big[Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot) - \overline{Q}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s) \\
&\qquad\quad + \big[\pi_h^*(\cdot|s) - \pi_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s),
\end{aligned} \tag{4.16}$$

where the first inequality is by the definition of $V_h^{\pi^*,\nu^k}$ in (2.3) and the definition of $\overline{V}_h^k$ in Line 8 of Algorithm 2. In addition, by the definition of $Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot)$ in (2.4) and the definition of the model prediction error $\bar{\iota}_h^k$ for Player 1 in (4.14), we have

$$\begin{aligned}
&[\pi_h^*(\cdot|s)]^\top \big[Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot) - \overline{Q}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s) \\
&\qquad = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^*(a|s)\bigg[\sum_{s'\in\mathcal{S}} \mathbb{P}_h(s'|s,a,b)\big[V_{h+1}^{\pi^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big] + \bar{\iota}_h^k(s,a,b)\bigg]\nu_h^k(b|s) \\
&\qquad = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^*(a|s)\bigg[\sum_{s'\in\mathcal{S}} \mathbb{P}_h(s'|s,a,b)\big[V_{h+1}^{\pi^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big]\bigg]\nu_h^k(b|s) \\
&\qquad\quad + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^*(a|s)\bar{\iota}_h^k(s,a,b)\nu_h^k(b|s).
\end{aligned}$$

Combining this equality with (4.16) gives

$$\begin{aligned}
V_h^{\pi^*,\nu^k}(s) - \overline{V}_h^k(s) &= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^*(a|s)\bigg[\sum_{s'\in\mathcal{S}} \mathbb{P}_h(s'|s,a,b)\big[V_{h+1}^{\pi^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big]\bigg]\nu_h^k(b|s) \\
&\quad + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^*(a|s)\bar{\iota}_h^k(s,a,b)\nu_h^k(b|s) \\
&\quad + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \big[\pi_h^*(a|s) - \pi_h^k(a|s)\big]\overline{Q}_h^k(s,a,b)\nu_h^k(b|s).
\end{aligned} \tag{4.17}$$

The inequality (4.17) indicates a recursion of the value function difference $V_h^{\pi^*,\nu^k}(s) - \overline{V}_h^k(s)$. As we have defined $V_{H+1}^{\pi^*,\nu^k}(s) = 0$ and $\overline{V}_{H+1}^k(s) = 0$, by recursively applying (4.17) from $h = 1$ to

H, we obtain

$$\begin{aligned} V_1^{\pi^*,\nu^k}(s_1) - \overline{V}_1^k(s_1) = &\sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\big\{[\pi_h^*(\cdot|s_h)]^\top \overline{l}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h) \,\big|\, s_1\big\} \\ &+ \underbrace{\sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\big\{\big[\pi_h^*(\cdot|s_h) - \pi_h^k(\cdot|s_h)\big]^\top \overline{Q}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h) \,\big|\, s_1\big\}}_{\text{Term(I)}}, \end{aligned} \tag{4.18}$$

where s_h are a random variables denoting the state at the h-th step following a distribution determined jointly by $\pi^*, \mathbb{P}, \nu^k$. Note that we have the factored independent transition model structure $\mathbb{P}_h(s'|s,a,b) = \mathbb{P}_h^1(s^{1\prime}|s^1,a)\mathbb{P}_h^2(s^{2\prime}|s^2,b)$ with $s = (s^1, s^2)$ and $s' = (s^{1\prime}, s^{2\prime})$, and $\pi_h(a|s) = \pi_h(a|s^1)$ as well as $\nu_h(b|s) = \nu_h(b|s^2)$. Here we also have the state reaching probability $q^{\nu^k,\mathbb{P}^2}(s^2) = \big\{q_h^{\nu^k,\mathbb{P}^2}(s^2)\big\}_{h=1}^H$ under ν^k and true transition $\mathbb{P}^2$ for Player 2, and define the empirical reaching probability $d^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2) = \{d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)\}_{h=1}^H$ under the empirical transition model $\widehat{\mathbb{P}}^{2,k}$ for Player 2, where we let $\widehat{\mathbb{P}}_h^k(s'|s,a,b) = \widehat{\mathbb{P}}_h^{1,k}(s^{1\prime}|s^1,a)\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime}|s^2,b)$. Then, for Term(I), we have

$$\begin{aligned} \text{Term(I)} &= \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\big\{\big[\pi_h^*(\cdot|s_h) - \pi_h^k(\cdot|s_h)\big]^\top \overline{Q}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h) \,\big|\, s_1\big\} \\ &= \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}^1,\mathbb{P}^2,\nu^k}\big\{\big[\pi_h^*(\cdot|s_h^1) - \pi_h^k(\cdot|s_h^1)\big]^\top \overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)\nu_h^k(\cdot|s_h^2) \,\big|\, s_1^1, s_1^2\big\} \\ &= \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}^1}\big\{\sum_{s_h^2\in\mathcal{S}_2} \underbrace{\big[\pi_h^*(\cdot|s_h^1) - \pi_h^k(\cdot|s_h^1)\big]^\top \overline{Q}_h^k(s_h^1,s_h^2,\cdot,\cdot)\nu_h^k(\cdot|s_h^2)}_{=:\overline{E}_h^k(s_h^1,s_h^2)} q_h^{\nu^k,\mathbb{P}^2}(s_h^2) \,\big|\, s_1^1, s_1^2\big\}. \end{aligned} \tag{4.19}$$

The last term of the above inequality (4.19) can be further bounded as

$$\begin{aligned} &\sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}^1}\big\{\sum_{s_h^2\in\mathcal{S}_2} \overline{E}_h^k(s_h^1,s_h^2) q_h^{\nu^k,\mathbb{P}^2}(s_h^2) \,\big|\, s_1^1, s_1^2\big\} \\ &\quad = \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}^1}\big\{\sum_{s_h^2\in\mathcal{S}_2} \overline{E}_h^k(s_h^1,s_h^2)[d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2) + q_h^{\nu^k,\mathbb{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)] \,\big|\, s_1^1, s_1^2\big\} \\ &\quad \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}^1}\big\{\sum_{s_h^2\in\mathcal{S}_2} \overline{E}_h^k(s_h^1,s_h^2) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2) \,\big|\, s_1^1, s_1^2\big\} + 2H \sum_{h=1}^{H}\sum_{s_h^2\in\mathcal{S}_2} \Big| q_h^{\nu^k,\mathbb{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big|, \end{aligned}$$

where the factor H in the last term is due to $|\overline{Q}_h^k(s_h^1, s_h^2, \cdot, \cdot)| \leq H$. Combining the above inequality with (4.19), we have

$$\text{Term(I)} \leq \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}^1}\big\{\big[\pi_h^*(\cdot|s_h^1) - \pi_h^k(\cdot|s_h^1)\big]^\top \sum_{s_h^2\in\mathcal{S}_2} \overline{Q}_h^k(s_h^1, s_h^2, \cdot, \cdot)\nu_h^k(\cdot|s_h^2) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2) \,\big|\, s_1^1, s_1^2\big\}$$
$$+ 2H\sum_{h=1}^{H}\sum_{s_h^2\in\mathcal{S}_2} \Big| q_h^{\nu^k,\mathbb{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big|. \tag{4.20}$$

Further combining (4.20) with (4.15), we eventually have

$$\begin{aligned} &V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1)\\ &\quad\leq \overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1) + \sum_{h=1}^{H}\mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\big\{[\pi_h^*(\cdot|s_h)]^\top \overline{\iota}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\, s_1\big\}\\ &\qquad + \sum_{h=1}^{H}\mathbb{E}_{\pi^*,\mathbb{P}^1}\Big\{\Big\langle \pi_h^*(\cdot|s_h^1) - \pi_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^{1,k}(s_h^1, s_h^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}}\,\Big|\, s_1^1, s_1^2\Big\}\\ &\qquad + 2H\sum_{h=1}^{H}\sum_{s_h^2\in\mathcal{S}_2}\Big| q_h^{\nu^k,\mathbb{P}^2}(s_h^2) - d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big|, \end{aligned}$$

where $F_h^{1,k}(s_h^1, s_h^2, a) := \langle \overline{Q}_h^k(s_h^1, s_h^2, a, \cdot), \nu_h^k(\cdot|s_h^2)\rangle_{\mathcal{B}}$ for any $a\in\mathcal{A}$. This completes our proof. □

Lemma 4.9. *With setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, the mirror ascent steps of Algorithm 2 lead to*

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\mathbb{P}^1}\Big\{\Big\langle \pi_h^*(\cdot|s_h^1) - \pi_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^{1,k}(s_h^1, s_h^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}}\,\Big|\, s_1^1, s_1^2\Big\}\\ &\qquad\leq \mathcal{O}\left(\sqrt{H^4K\log|\mathcal{A}|}\right).\end{aligned}$$

Proof. As shown in (4.8), the mirror ascent step at the k-th episode is to solve the following maximization problem

$$\begin{aligned}\underset{\pi}{\text{maximize}}\sum_{h=1}^{H}\Big\langle \pi_h(\cdot|s^1) - \pi_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s^1, s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)\Big\rangle_{\mathcal{A}}\\ - \frac{1}{\eta}\sum_{h=1}^{H} D_{\mathrm{KL}}\big(\pi_h(\cdot|s^1), \pi_h^k(\cdot|s^1)\big),\end{aligned}$$

with $F_h^{1,k}(s^1, s^2, a) := \langle \overline{Q}_h^k(s^1, s^2, a, \cdot), \nu_h^k(\cdot|s^2)\rangle_{\mathcal{B}}$. We equivalently rewrite this maximization problem to a minimization problem as

$$\begin{aligned}\underset{\pi}{\text{minimize}} - \sum_{h=1}^{H} \Big\langle \pi_h(\cdot|s^1) - \pi_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s^1,s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2) \Big\rangle_{\mathcal{A}} \\ + \frac{1}{\eta}\sum_{h=1}^{H} D_{\mathrm{KL}}\big(\pi_h(\cdot|s^1), \pi_h^k(\cdot|s^1)\big).\end{aligned}$$

Note that the closed-form solution $\pi_h^{k+1}(\cdot|s^1), \forall s^1 \in \mathcal{S}_1$, to this minimization problem is guaranteed to stay in the relative interior of a probability simplex if initializing $\pi_h^0(\cdot|s^1) = \mathbf{1}/|\mathcal{A}|$. Thus, we apply Lemma 4.16 and obtain that for any $\pi = \{\pi_h\}_{h=1}^H$, the following inequality holds

$$\begin{aligned} -\eta\Big\langle \pi_h^{k+1}(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s^1,s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)\Big\rangle_{\mathcal{A}} + \eta\Big\langle \pi_h(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s^1,s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)\Big\rangle_{\mathcal{A}} \\ \leq D_{\mathrm{KL}}\big(\pi_h(\cdot|s^1), \pi_h^k(\cdot|s^1)\big) - D_{\mathrm{KL}}\big(\pi_h(\cdot|s^1), \pi_h^{k+1}(\cdot|s^1)\big) - D_{\mathrm{KL}}\big(\pi_h^{k+1}(\cdot|s^1), \pi_h^k(\cdot|s^1)\big).\end{aligned}$$

Then, by rearranging the terms and letting $\pi_h = \pi_h^*$, we have

$$\begin{aligned}\eta\Big\langle \pi_h^*(\cdot|s^1) - \pi_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s^1,s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}} \\ \leq D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s^1), \pi_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s), \pi_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi_h^{k+1}(\cdot|s), \pi_h^k(\cdot|s)\big) \\ + \eta\Big\langle \pi_h^{k+1}(\cdot|s^1) - \pi_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s^1,s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}}.\end{aligned} \tag{4.21}$$

Due to Pinsker's inequality, we have

$$-D_{\mathrm{KL}}\big(\pi_h^{k+1}(\cdot|s^1), \pi_h^k(\cdot|s^1)\big) \leq -\frac{1}{2}\big\|\pi_h^{k+1}(\cdot|s^1) - \pi_h^k(\cdot|s^1)\big\|_1^2.$$

Further by Cauchy-Schwarz inequality, we have

$$\eta\Big\langle \pi_h^{k+1}(\cdot|s^1) - \pi_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s^1,s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)\Big\rangle_{\mathcal{A}} \leq \eta H \big\|\pi_h^{k+1}(\cdot|s^1) - \pi_h^k(\cdot|s^1)\big\|_1.$$

since we have

$$\begin{aligned}\Big\| \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s^1,s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)\Big\|_\infty &= \max_{a\in\mathcal{A}} \sum_{s^2\in\mathcal{S}_2} \langle \overline{Q}_h^k(s^1,s^2,a,\cdot), \nu_h^k(\cdot|s^2)\rangle_{\mathcal{B}} \cdot d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2) \\ &\leq \sum_{s^2\in\mathcal{S}_2} H\cdot d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2) = H.\end{aligned}$$

Thus, we further obtain

$$
\begin{aligned}
&- D_{\mathrm{KL}}\big(\pi_h^{k+1}(\cdot|s^1), \pi_h^k(\cdot|s^1)\big) + \eta\big\langle \pi_h^{k+1}(\cdot|s^1) - \pi_h^k(\cdot|s^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^{1,k}(s^1,s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)\big\rangle_{\mathcal{A}} \\
&\qquad \leq -\frac{1}{2}\big\|\pi_h^{k+1}(\cdot|s^1) - \pi_h^k(\cdot|s^1)\big\|_1^2 + \eta H \big\|\pi_h^{k+1}(\cdot|s^1) - \pi_h^k(\cdot|s^1)\big\|_1 \leq \frac{1}{2}\eta^2 H^2,
\end{aligned}
\tag{4.22}
$$

where the last inequality is by viewing $\big\|\pi_h^{k+1}(\cdot|s^1) - \pi_h^k(\cdot|s^1)\big\|_1$ as a variable x and finding the maximal value of $-1/2 \cdot x^2 + \eta H x$ to obtain the upper bound $1/2 \cdot \eta^2 H^2$.

Thus, combing (4.22) with (4.21), the policy improvement step in Algorithm 2 implies

$$
\begin{aligned}
&\eta\Big\langle \pi_h^*(\cdot|s^1) - \pi_h^k(\cdot|s^1), \sum_{s^2\in\mathcal{S}_2} F_h^{1,k}(s^1,s^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)\Big\rangle_{\mathcal{A}} \\
&\qquad \leq D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s^1), \pi_h^k(\cdot|s^1)\big) - D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s^1), \pi_h^{k+1}(\cdot|s^1)\big) + \frac{1}{2}\eta^2 H^2,
\end{aligned}
$$

which further leads to

$$
\begin{aligned}
&\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}^1}\Big\{\Big\langle \pi_h^*(\cdot|s_h^1) - \pi_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^{1,k}(s_h^1,s_h^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}} \,\Big|\, s_1^1, s_1^2\Big\} \\
&\qquad \leq \frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}^1}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s_h^1), \pi_h^k(\cdot|s_h^1)\big) - D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s_h^1), \pi_h^{k+1}(\cdot|s_h^1)\big)\big] + \frac{1}{2}\eta H^3.
\end{aligned}
$$

Taking summation from $k=1$ to K of both sides, we obtain

$$
\begin{aligned}
&\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}^1}\Big\{\Big\langle \pi_h^*(\cdot|s_h^1) - \pi_h^k(\cdot|s_h^1), \sum_{s_h^2\in\mathcal{S}_2} F_h^{1,k}(s_h^1,s_h^2,\cdot) d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}} \,\Big|\, s_1^1, s_1^2\Big\} \\
&\qquad \leq \frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}^1}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s_h^1), \pi_h^1(\cdot|s_h^1)\big) - D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s_h^1), \pi_h^{K+1}(\cdot|s_h^1)\big)\big] + \frac{1}{2}\eta K H^3 \\
&\qquad \leq \frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}^1}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s_h^1), \pi_h^1(\cdot|s_h^1)\big)\big] + \frac{1}{2}\eta K H^3,
\end{aligned}
$$

where the last inequality is by non-negativity of KL divergence. With the initialization in Algorithm 2, it is guaranteed that $\pi_h^1(\cdot|s^1) = \mathbf{1}/|\mathcal{A}|$, which thus leads to $D_{\mathrm{KL}}\left(\pi_h^*(\cdot|s^1), \pi_h^1(\cdot|s^1)\right) \leq \log|\mathcal{A}|$ for any s^1. Then, with setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, we bound the last term as

$$
\frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}^1}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s_h^1), \pi_h^1(\cdot|s_h^1)\big)\big] + \frac{1}{2}\eta K H^3 \leq \mathcal{O}\left(\sqrt{H^4 K \log|\mathcal{A}|}\right),
$$

which gives

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\mathbb{P}^1}\Big\{\Big\langle \pi_h^*(\cdot|s_h^1)-\pi_h^k(\cdot|s_h^1),\sum_{s_h^2\in\mathcal{S}_2}F_h^{1,k}(s_h^1,s_h^2,\cdot)d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s_h^2)\Big\rangle_{\mathcal{A}}\,\Big|\,s_1^1,s_1^2\Big\}&\\ \le \mathcal{O}\left(\sqrt{H^4K\log|\mathcal{A}|}\right).&\end{aligned}$$

This completes the proof. □

Lemma 4.10. *For any $k\in[K]$, $h\in[H]$ and all $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, with probability at least $1-\delta$, we have*

$$\big|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)\big|\le\sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}.$$

Proof. The proof for this theorem is a direct application of Hoeffding's inequality. For $k\ge 1$, the definition of $\widehat{r}_h^k$ in (4.10) indicates that $\widehat{r}_h^k(s,a,b)$ is the average of $N_h^k(s,a,b)$ samples of the observed rewards at (s,a,b) if $N_h^k(s,a,b)>0$. Then, for fixed $k\in[K],h\in[H]$ and state-action tuple $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, when $N_h^k(s,a,b)>0$, according to Hoeffding's inequality, with probability at least $1-\delta'$ where $\delta'\in(0,1]$, we have

$$\big|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)\big|\le\sqrt{\frac{\log(2/\delta')}{2N_h^k(s,a,b)}},$$

where we also use the facts that the observed rewards $r_h^k\in[0,1]$ for all k and h, and $\mathbb{E}\big[\widehat{r}_h^k\big]=r_h$ for all k and h. For the case where $N_h^k(s,a,b)=0$, by (4.10), we know $\widehat{r}_h^k(s,a,b)=0$ such that $|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)|=|r_h(s,a,b)|\le 1$. On the other hand, we have $\sqrt{2\log(2/\delta')}\ge 1>|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)|$. Thus, combining the above results, with probability at least $1-\delta'$, for fixed $k\in[K],h\in[H]$ and state-action tuple $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, we have

$$\big|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)\big|\le\sqrt{\frac{2\log(2/\delta')}{\max\{N_h^k(s,a,b),1\}}}.$$

Moreover, by the union bound, letting $\delta=|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK\delta'/2$, assuming $K>1$, with probability at least $1-\delta$, for any $k\in[K],h\in[H]$ and any state-action tuple $(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$, we have

$$\big|\widehat{r}_h^k(s,a,b)-r_h(s,a,b)\big|\le\sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}.$$

This completes the proof. □

In (4.7), we factor the state as $s = (s^1, s^2)$ such that we have $|\mathcal{S}| = |\mathcal{S}_1||\mathcal{S}_2|$. Thus, we set $\beta_h^{r,k}(s,a,b) = \sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}} = \sqrt{\frac{4\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s^1,s^2,a,b),1\}}}$, which equals the bound in Lemma 4.10. The counter $N_h^k(s,a,b)$ is equivalent to $N_h^k(s^1, s^2, a, b)$.

Lemma 4.11. *For any $k \in [K]$, $h \in [H]$ and all $(s,a) \in \mathcal{S} \times \mathcal{A}$, with probability at least $1-\delta$, we have*

$$\left\|\widehat{\mathbb{P}}_h^k(\cdot \,|\, s,a,b) - \mathbb{P}_h(\cdot \,|\, s,a,b)\right\|_1 \leq \sqrt{\frac{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}},$$

where we have a factored state space $s = (s^1, s^2)$, $s' = (s^{1\prime}, s^{2\prime})$, and an independent state transition $\mathbb{P}_h(s' \,|\, s,a,b) = \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\mathbb{P}_h^1(s^{2\prime} \,|\, s^2, b)$ and $\widehat{\mathbb{P}}_h^k(\cdot \,|\, s,a,b) = \widehat{\mathbb{P}}_h^{1,k}(s^{1\prime} \,|\, s^1, a)\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b)$.

Proof. Since the state space and the transition model are factored, we need to decompose the term as follows

$$\begin{aligned}
&\left\|\widehat{\mathbb{P}}_h^k(\cdot \,|\, s,a,b) - \mathbb{P}_h(\cdot \,|\, s,a,b)\right\|_1 \\
&\quad = \sum_{s^{1\prime}, s^{2\prime}} \left|\widehat{\mathbb{P}}_h^{1,k}(s^{1\prime} \,|\, s^1, a)\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b) - \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\mathbb{P}_h^2(s^{2\prime} \,|\, s^2, b)\right| \\
&\quad = \sum_{s^{1\prime}, s^{2\prime}} \bigg| \left[\widehat{\mathbb{P}}_h^{1,k}(s^{1\prime} \,|\, s^1, a) - \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\right]\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b) \\
&\qquad + \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\left[\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b) - \mathbb{P}_h^2(s^{2\prime} \,|\, s^2, b)\right] \bigg|.
\end{aligned}$$

We can further bound the last term in the above equality as follows

$$\begin{aligned}
&\sum_{s^{1\prime}, s^{2\prime}} \bigg| \left[\widehat{\mathbb{P}}_h^{1,k}(s^{1\prime} \,|\, s^1, a) - \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\right]\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b) \\
&\qquad + \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\left[\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b) - \mathbb{P}_h^2(s^{2\prime} \,|\, s^2, b)\right] \bigg| \\
&\quad \leq \sum_{s^{1\prime}, s^{2\prime}} \bigg\{ \left|\widehat{\mathbb{P}}_h^{1,k}(s^{1\prime} \,|\, s^1, a) - \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\right|\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b) \\
&\qquad + \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\left|\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b) - \mathbb{P}_h^2(s^{2\prime} \,|\, s^2, b)\right| \bigg\} \\
&\quad \leq \sum_{s^{1\prime}} \left|\widehat{\mathbb{P}}_h^{1,k}(s^{1\prime} \,|\, s^1, a) - \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\right| + \sum_{s^{2\prime}} \left|\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b) - \mathbb{P}_h^2(s^{2\prime} \,|\, s^2, b)\right| \\
&\quad = \left\|\widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a)\right\|_1 + \left\|\widehat{\mathbb{P}}_h^{2,k}(\cdot \,|\, s^2, b) - \mathbb{P}_h^2(\cdot \,|\, s^2, b)\right\|_1,
\end{aligned}$$

where the last inequality is due to $\sum_{s^{2\prime}} \widehat{\mathbb{P}}_h^{2,k}(s^{2\prime} \,|\, s^2, b) = 1$ and $\sum_{s^{1\prime}} \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a) = 1$. Thus, we need to bound the two terms $\|\widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a)\|_1$ and $\|\widehat{\mathbb{P}}_h^{2,k}(\cdot \,|\, s^2, b) - \mathbb{P}_h^2(\cdot \,|\, s^2, b)\|_1$ separately.

For $k \geq 1$, we have $\|\widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a)\|_1 = \max_{\|\mathbf{z}\|_\infty \leq 1} \langle \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(s^{1\prime} \,|\, s^1, a), \mathbf{z} \rangle_{\mathcal{S}_1}$ by the duality. We construct an ϵ-cover for the set $\{\mathbf{z} \in \mathbb{R}^{|\mathcal{S}_1|} : \|\mathbf{z}\|_\infty \leq 1\}$ with the distance induced by $\|\cdot\|_\infty$, denoted as $\mathcal{C}_\infty(\epsilon)$, such that for any $\mathbf{z} \in \mathbb{R}^{|\mathcal{S}_1|}$, there always exists $\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)$ satisfying $\|\mathbf{z} - \mathbf{z}'\|_\infty \leq \epsilon$. The covering number is $\mathcal{N}_\infty(\epsilon) = |\mathcal{C}_\infty(\epsilon)| = 1/\epsilon^{|\mathcal{S}_1|}$. Thus, we know that for any $(s^1, a) \in \mathcal{S}_1 \times \mathcal{A}$ and any $\mathbf{z}$ with $\|\mathbf{z}\|_\infty \leq 1$, there exists $\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)$ such that $\|\mathbf{z}' - \mathbf{z}\|_\infty \leq \epsilon$ and

$$\begin{aligned}
&\left\langle \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z} \right\rangle_{\mathcal{S}_1} \\
&\quad = \left\langle \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z}' \right\rangle_{\mathcal{S}_1} + \left\langle \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z} - \mathbf{z}' \right\rangle_{\mathcal{S}_1} \\
&\quad \leq \left\langle \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z}' \right\rangle_{\mathcal{S}_1} + \epsilon \left\| \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a) \right\|_1,
\end{aligned}$$

such that we further have

$$\begin{aligned}
&\left\| \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a) \right\|_1 \\
&\quad = \max_{\|\mathbf{z}\|_\infty \leq 1} \left\langle \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a)), \mathbf{z} \right\rangle_{\mathcal{S}_1} \\
&\quad \leq \max_{\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)} \left\langle \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z}' \right\rangle_{\mathcal{S}_1} + \epsilon \left\| \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a) \right\|_1.
\end{aligned} \tag{4.23}$$

By Hoeffding's inequality and the union bound over all $\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)$, when $N_h^k(s^1, a) > 0$, with probability at least $1 - \delta'$ where $\delta' \in (0, 1]$,

$$\max_{\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)} \left\langle \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a), \mathbf{z}' \right\rangle_{\mathcal{S}_1} \leq \sqrt{\frac{|\mathcal{S}_1| \log(1/\epsilon) + \log(1/\delta')}{2N_h^k(s^1, a)}}. \tag{4.24}$$

Letting $\epsilon = 1/2$, by (4.23) and (4.24), with probability at least $1 - \delta'$, we have

$$\left\| \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a) \right\|_1 \leq 1 \sqrt{\frac{|\mathcal{S}| \log 2 + \log(1/\delta')}{2N_h^k(s^1, a)}}.$$

When $N_h^k(s^1, a) = 0$, we have $\left\| \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a) \right\|_1 = \|\mathbb{P}_h^1(\cdot \,|\, s^1, a)\|_1 = 1$ such that $2\sqrt{\frac{|\mathcal{S}| \log 2 + \log(1/\delta')}{2}} > 1 = \left\| \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a) \right\|_1$ always holds. Thus, with probability at least $1 - \delta'$,

$$\left\| \widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a) \right\|_1 \leq 2 \sqrt{\frac{|\mathcal{S}_1| \log 2 + \log(1/\delta')}{2 \max\{N_h^k(s^1, a), 1\}}} \leq \sqrt{\frac{2|\mathcal{S}_1| \log(2/\delta')}{\max\{N_h^k(s^1, a), 1\}}}.$$

Then, by the union bound, assuming $K > 1$, letting $\delta'' = |\mathcal{S}_1||\mathcal{A}|HK\delta'/2$, with probability at least

$1-\delta''$, for any $(s^1, a) \in \mathcal{S}_1 \times \mathcal{A}$ and any $h \in [H]$ and $k \in [K]$, we have

$$\left\|\widehat{\mathbb{P}}_h^{1,k}(\cdot \,|\, s^1, a) - \mathbb{P}_h^1(\cdot \,|\, s^1, a)\right\|_1 \leq \sqrt{\frac{2|\mathcal{S}_1| \log(|\mathcal{S}_1||\mathcal{A}|HK/\delta'')}{\max\{N_h^k(s^1, a), 1\}}}.$$

Similarly, we can also obtain that with probability at least $1-\delta''$, for any $(s^2, a) \in \mathcal{S}_2 \times \mathcal{B}$ and any $h \in [H]$ and $k \in [K]$, we have

$$\left\|\widehat{\mathbb{P}}_h^{2,k}(\cdot \,|\, s^2, b) - \mathbb{P}_h^2(\cdot \,|\, s^2, b)\right\|_1 \leq \sqrt{\frac{2|\mathcal{S}_2| \log(|\mathcal{S}_2||\mathcal{B}|HK/\delta'')}{\max\{N_h^k(s^2, b), 1\}}}.$$

Further by the union bound, we have with probability at least $1-\delta$ where $\delta = 2\delta''$,

$$\left\|\widehat{\mathbb{P}}_h^{k}(\cdot \,|\, s, a, b) - \mathbb{P}_h(\cdot \,|\, s, a, b)\right\|_1 \leq \sqrt{\frac{2|\mathcal{S}_1| \log(2|\mathcal{S}_1||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s^1, a), 1\}}} + \sqrt{\frac{2|\mathcal{S}_2| \log(2|\mathcal{S}_2||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s^2, b), 1\}}}.$$

This completes the proof. □

In (4.7), we set $\beta_h^{\mathbb{P},k}(s,a,b) = \sqrt{\frac{2H^2|\mathcal{S}_1| \log(2|\mathcal{S}_1||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s^1,a),1\}}} + \sqrt{\frac{2H^2|\mathcal{S}_2| \log(2|\mathcal{S}_2||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s^2,b),1\}}}$, which equals the product of the upper bound in Lemma 4.11 and the factor H.

Lemma 4.12. *With probability at least* $1-2\delta$*, Algorithm 2 ensures that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\left[\bar{\iota}_h^k(s_h, a_h, b_h) \,\middle|\, s_1\right] \leq 0.$$

Proof. We prove the upper bound of the model prediction error term. As defined in (4.14), we have the instantaneous prediction error at the h-step of the k-th episode as

$$\bar{\iota}_h^k(s,a,b) = r_h(s,a,b) + \left\langle \mathbb{P}_h(\cdot \,|\, s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \overline{Q}_h^k(s,a,b), \tag{4.25}$$

where the equality is by the definition of the prediction error in (4.14). By plugging in the definition of $\overline{Q}_h^k$ in Line 7 of Algorithm 2, for any (s,a,b), we bound the following term as

$$\begin{aligned} &r_h(s,a,b) + \left\langle \mathbb{P}_h(\cdot \,|\, s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \overline{Q}_h^k(s,a,b) \\ &\quad \leq r_h(s,a,b) + \left\langle \mathbb{P}_h(\cdot \,|\, s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} \\ &\qquad - \min\left\{\widehat{r}_h^k(s,a,b) + \left\langle \widehat{\mathbb{P}}_h^k(\cdot|s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \beta_h^k, H-h+1\right\} \\ &\quad \leq \max\left\{r_h(s,a,b) - \widehat{r}_h^k(s,a,b) + \left\langle \mathbb{P}_h(\cdot \,|\, s,a,b) - \widehat{\mathbb{P}}_h^k(\cdot|s,a,b), \overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}} - \beta_h^k, 0\right\}, \end{aligned} \tag{4.26}$$

where the inequality holds because

$$\begin{aligned}
& r_h(s,a,b) + \big\langle \mathbb{P}_h(\cdot \mid s,a,b), \overline{V}^k_{h+1}(\cdot) \big\rangle_{\mathcal{S}} \\
& \qquad \le r_h(s,a,b) + \big\|\mathbb{P}_h(\cdot \mid s,a,b)\big\|_1 \|\overline{V}^k_{h+1}(\cdot)\|_\infty \le 1 + \max_{s' \in \mathcal{S}} \big|\overline{V}^k_{h+1}(s')\big| \le 1 + H - h,
\end{aligned}$$

since $\big\|\mathbb{P}_h(\cdot \mid s,a,b)\big\|_1 = 1$ and also the truncation step as shown in Line 7 of Algorithm 2 for $\overline{Q}^k_{h+1}$ such that for any $s' \in \mathcal{S}$

$$\begin{aligned}
\big|\overline{V}^k_{h+1}(s')\big| &= \Big| \big[\pi^k_{h+1}(\cdot|s')\big]^\top \overline{Q}^k_{h+1}(s',\cdot,\cdot)\nu^k_{h+1}(\cdot|s') \Big| \\
&\le \big\|\pi^k_{h+1}(\cdot|s')\big\|_1 \big\|\overline{Q}^k_{h+1}(s',\cdot,\cdot)\nu^k_{h+1}(\cdot|s')\big\|_\infty \\
&\le \max_{a,b} \big|\overline{Q}^k_{h+1}(s',a,b)\big| \le H.
\end{aligned} \tag{4.27}$$

Combining (4.25) and (4.26) gives

$$\begin{aligned}
\overline{\iota}^k_h(s,a,b) \le \max\Big\{ & r_h(s,a,b) - \widehat{r}^k_h(s,a,b) \\
& + \big\langle \mathbb{P}_h(\cdot \mid s,a,b) - \widehat{\mathbb{P}}^k_h(\cdot|s,a,b), \overline{V}^k_{h+1}(\cdot) \big\rangle_{\mathcal{S}} - \beta^k_h, 0 \Big\}.
\end{aligned} \tag{4.28}$$

Note that as shown in (4.7), we have

$$\beta^k_h(s,a,b) = \beta^{r,k}_h(s,a,b) + \beta^{\mathbb{P},k}_h(s,a,b).$$

Then, with probability at least $1-\delta$, we have

$$\begin{aligned}
& r_h(s,a,b) - \widehat{r}^k_h(s,a,b) - \beta^{r,k}_h(s,a,b) \\
& \qquad \le \big| r_h(s,a,b) - \widehat{r}^k_h(s,a,b) \big| - \beta^{r,k}_h(s,a,b) \\
& \qquad \le \beta^{r,k}_h(s,a,b) - \beta^{r,k}_h(s,a,b) = 0,
\end{aligned}$$

where the last inequality is by Lemma 4.10 and the setting of the bonus for the reward. Moreover, with probability at least $1-\delta$, we have

$$\begin{aligned}
& \big\langle \mathbb{P}_h(\cdot \mid s,a,b) - \widehat{\mathbb{P}}^k_h(\cdot|s,a,b), \overline{V}^k_{h+1}(\cdot) \big\rangle_{\mathcal{S}} - \beta^{\mathbb{P},k}_h(s,a,b) \\
& \qquad \le \big\|\mathbb{P}_h(\cdot \mid s,a,b) - \widehat{\mathbb{P}}^k_h(\cdot|s,a,b)\big\|_1 \big\|\overline{V}^k_{h+1}(\cdot)\big\|_\infty - \beta^{\mathbb{P},k}_h(s,a,b) \\
& \qquad \le H \big\|\mathbb{P}_h(\cdot \mid s,a,b) - \widehat{\mathbb{P}}^k_h(\cdot|s,a)\big\|_1 - \beta^{\mathbb{P},k}_h(s,a,b) \\
& \qquad \le \beta^{\mathbb{P},k}_h(s,a,b) - \beta^{\mathbb{P},k}_h(s,a,b) = 0,
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the second inequality is due to

$\max_{s'\in\mathcal{S}} \big\|\overline{V}^k_{h+1}(s')\big\|_\infty \le H$ as shown in (4.27), and the last inequality is by the setting of $\beta_h^{\mathbb{P},k}$ in (4.7) and also Lemma 4.11. Thus, with probability at least $1-2\delta$, the following inequality holds

$$r_h(s,a,b) - \widehat{r}^k_h(s,a,b) + \big\langle \mathbb{P}_h(\cdot\,|\,s,a,b) - \widehat{\mathbb{P}}^k_h(\cdot|s,a,b), \overline{V}^k_{h+1}(\cdot)\big\rangle_{\mathcal{S}} - \beta^k_h(s,a,b) \le 0.$$

Combining the above inequality with (4.28), we have that with probability at least $1-2\delta$, for any $h\in[H]$ and $k\in[K]$, the following inequality holds

$$\overline{\iota}^k_h(s,a,b) \le 0,\ \forall (s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B},$$

which leads to

$$\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\big[\overline{\iota}^k_h(s_h,a_h,b_h)\,\big|\,s_1\big] \le 0.$$

This completes the proof. □

Lemma 4.13. *With probability at least* $1-\delta$*, Algorithm 2 ensures that*

$$\sum_{k=1}^K \overline{V}^k_1(s_1) - \sum_{k=1}^K V_1^{\pi^k,\nu^k}(s_1) \le \widetilde{\mathcal{O}}\big(\sqrt{|\mathcal{S}_1|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}_2|^2|\mathcal{B}|H^4K} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H^2K}\big).$$

Proof. We assume that a trajectory $\{(s^k_h,a^k_h,b^k_h,s^k_{h+1})\}_{h=1}^H$ for all $k\in[K]$ is generated following the policies π^k, ν^k, and the true transition model $\mathbb{P}$. Thus, we expand the bias term at the h-th step of the k-th episode, which is

$$\begin{aligned}
&\overline{V}^k_h(s^k_h) - V_h^{\pi^k,\nu^k}(s^k_h)\\
&\qquad= \big[\pi^k_h(\cdot|s^k_h)\big]^\top\big[\overline{Q}^k_h(s^k_h,\cdot,\cdot) - Q_h^{\pi^k,\nu^k}(s^k_h,\cdot,\cdot)\big]\nu^k_h(\cdot|s^k_h)\\
&\qquad= \zeta^k_h + \overline{Q}^k_h(s^k_h,a^k_h,b^k_h) - Q_h^{\pi^k,\nu^k}(s^k_h,a^k_h,b^k_h) \qquad\qquad (4.29)\\
&\qquad= \zeta^k_h + \big\langle \mathbb{P}_h(\cdot\,|\,s^k_h,a^k_h,b^k_h), \overline{V}^k_{h+1}(\cdot) - V_{h+1}^{\pi^k,\nu^k}(\cdot)\big\rangle_{\mathcal{S}} - \overline{\iota}^k_h(s^k_h,a^k_h,b^k_h)\\
&\qquad= \zeta^k_h + \xi^k_h + \overline{V}^k_{h+1}(s^k_{h+1}) - V_{h+1}^{\pi^k,\nu^k}(s^k_{h+1}) - \overline{\iota}^k_h(s^k_h,a^k_h,b^k_h),
\end{aligned}$$

where the first equality is by Line 8 of Algorithm 4 and (2.3), the third equality is by plugging in (2.4) and (4.14). Specifically, in the above equality, we introduce two martingale difference sequence, namely, $\{\zeta^k_h\}_{h\ge0,k\ge0}$ and $\{\xi^k_h\}_{h\ge0,k\ge0}$, which are defined as

$$\begin{aligned}
\zeta^k_h &:= \big[\pi^k_h(\cdot|s^k_h)\big]^\top\big[\overline{Q}^k_h(s^k_h,\cdot,\cdot) - Q_h^{\pi^k,\nu^k}(s^k_h,\cdot,\cdot)\big]\nu^k_h(\cdot|s^k_h) - \big[\overline{Q}^k_h(s^k_h,a^k_h,b^k_h) - Q_h^{\pi^k,\nu^k}(s^k_h,a^k_h,b^k_h)\big],\\
\xi^k_h &:= \big\langle \mathbb{P}_h(\cdot\,|\,s^k_h,a^k_h,b^k_h), \overline{V}^k_{h+1}(\cdot) - V_{h+1}^{\pi^k,\nu^k}(\cdot)\big\rangle_{\mathcal{S}} - \big[\overline{V}^k_{h+1}(s^k_{h+1}) - V_{h+1}^{\pi^k,\nu^k}(s^k_{h+1})\big],
\end{aligned}$$

such that

$$\mathbb{E}_{a_h^k\sim\pi_h^k(\cdot|s_h^k),b_h^k\sim\nu_h^k(\cdot|s_h^k)}\big[\zeta_h^k\,\big|\,\mathcal{F}_h^k\big]=0,$$
$$\mathbb{E}_{s_{h+1}^k\sim\mathbb{P}_h(\cdot\,|\,s_h^k,a_h^k,b_h^k)}\big[\xi_h^k\,\big|\,\widetilde{\mathcal{F}}_h^k\big]=0,$$

with $\mathcal{F}_h^k$ being the filtration of all randomness up to $(h-1)$-th step of the k-th episode plus s_h^k, and $\widetilde{\mathcal{F}}_h^k$ being the filtration of all randomness up to $(h-1)$-th step of the k-th episode plus s_h^k, a_h^k, b_h^k.

The equality (4.29) forms a recursion for $\overline{V}_h^k(s_h^k)-V_h^{\pi^k,\nu^k}(s_h^k)$. We also have $\overline{V}_{H+1}^k(\cdot)=\mathbf{0}$ and $V_{H+1}^{\pi^k,\nu^k}(\cdot)=\mathbf{0}$. Thus, recursively apply (4.29) from $h=1$ to H leads to the following equality

$$\overline{V}_1^k(s_1)-V_1^{\pi^k,\nu^k}(s_1)=\sum_{h=1}^H\zeta_h^k+\sum_{h=1}^H\xi_h^k-\sum_{h=1}^H\overline{\iota}_h^k(s_h^k,a_h^k,b_h^k). \tag{4.30}$$

Moreover, by (4.14) and Line 7 of Algorithm 2, we have

$$\begin{aligned}-\overline{\iota}_h^k(s_h^k,a_h^k,b_h^k)&=-r_h(s_h^k,a_h^k,b_h^k)-\big\langle\mathbb{P}_h(\cdot\,|\,s_h,a_h,b_h),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\\&\quad+\min\big\{\widehat{r}_h^k(s_h^k,a_h^k,b_h^k)+\big\langle\widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}+\beta_h^k(s_h^k,a_h^k,b_h^k),H\big\}.\end{aligned}$$

Then, we can further bound $-\overline{\iota}_h^k(s_h^k,a_h^k,b_h^k)$ as follows

$$\begin{aligned}-\overline{\iota}_h^k(s_h^k,a_h^k,b_h^k)&\le-r_h(s_h^k,a_h^k,b_h^k)-\big\langle\mathbb{P}_h(\cdot\,|\,s_h^k,a_h^k,b_h^k),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}+\widehat{r}_h^k(s_h^k,a_h^k,b_h^k)\\&\quad+\big\langle\widehat{\mathbb{P}}_h^k(\cdot|s_h^k,a_h^k,b_h^k),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}+\beta_h^k(s_h^k,a_h^k,b_h^k)\\&\le\big|\widehat{r}_h^k(s_h^k,a_h^k,b_h^k)-r_h(s_h^k,a_h^k,b_h^k)\big|\\&\quad+\Big|\big\langle\mathbb{P}_h(\cdot\,|\,s_h^k,a_h^k,b_h^k)-\widehat{\mathbb{P}}_h^k(\cdot\,|\,s_h^k,a_h^k,b_h^k),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\Big|+\beta_h^k(s_h^k,a_h^k,b_h^k),\end{aligned}$$

where the first inequality is due to $\min\{x,y\}\le x$. Additionally, we have

$$\begin{aligned}&\Big|\big\langle\mathbb{P}_h(\cdot\,|\,s_h^k,a_h^k,b_h^k)-\widehat{\mathbb{P}}_h^k(\cdot\,|\,s_h^k,a_h^k,b_h^k),\overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\Big|\\&\quad\le\big\|\overline{V}_{h+1}^k(\cdot)\big\|_\infty\big\|\mathbb{P}_h(\cdot\,|\,s_h^k,a_h^k,b_h^k)-\widehat{\mathbb{P}}_h^k(\cdot\,|\,s_h^k,a_h^k,b_h^k)\big\|_1\\&\quad\le H\big\|\mathbb{P}_h(\cdot\,|\,s_h^k,a_h^k,b_h^k)-\widehat{\mathbb{P}}_h^k(\cdot\,|\,s_h^k,a_h^k,b_h^k)\big\|_1,\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality and the second inequality is by (4.54).

Thus, putting the above together, we obtain

$$\begin{aligned}-\overline{\iota}_h^k(s_h^k, a_h^k, b_h^k) &\leq \big|\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) - r_h(s_h^k, a_h^k, b_h^k)\big| \\ &\quad + H\big\|\mathbb{P}_h(\cdot \mid s_h^k, a_h^k, b_h^k) - \mathbb{P}_h(\cdot \mid s_h^k, a_h^k, b_h^k)\big\|_1 + \beta_h^k(s_h^k, a_h^k, b_h^k) \\ &\leq 2\beta_h^{r,k}(s_h^k, a_h^k, b_h^k) + 2\beta_h^{\mathbb{P},k}(s_h^k, a_h^k, a_h^k),\end{aligned}$$

where the second inequality is by Lemma 4.10, Lemma 4.11, and the decomposition of the bonus term β_h^k as (4.7). Due to Lemma 4.10 and Lemma 4.11, by union bound, for any $h \in [H], k \in [K]$ and $(s_h, a_h, b_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, the above inequality holds with probability with probability at least $1 - 2\delta$. Therefore, by (4.30), with probability at least $1 - 2\delta$, we have

$$\begin{aligned}&\sum_{k=1}^{K} \big[\overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1)\big] \\ &\quad \leq \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k + \sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k + 2\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{r,k}(s_h^k, a_h^k, b_h^k) + 2\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{\mathbb{P},k}(s_h^k, a_h^k, b_h^k).\end{aligned} \tag{4.31}$$

By Azuma-Hoeffding inequality, with probability at least $1 - \delta$, the following inequalities hold

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k \leq \mathcal{O}\left(\sqrt{H^3 K \log \frac{1}{\delta}}\right), \quad \sum_{k=1}^{K}\sum_{h=1}^{H} \xi_h^k \leq \mathcal{O}\left(\sqrt{H^3 K \log \frac{1}{\delta}}\right),$$

where we use the facts that $|\overline{Q}_h^k(s_h^k, a_h^k, b_h^k) - Q_h^{\pi^k,\nu^k}(s_h^k, a_h^k, b_h^k)| \leq 2H$ and $|\overline{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k,\nu^k}(s_{h+1}^k)| \leq 2H$. Next, we need to bound $\sum_{k=1}^K \sum_{h=1}^H \beta_h^{r,k}(s_h^k, a_h^k, b_h^k)$ and $\sum_{k=1}^K \sum_{h=1}^H \beta_h^{\mathbb{P},k}(s_h^k, a_h^k, b_h^k)$ in (4.31). We show that

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{r,k}(s_h^k, a_h^k, b_h^k) &= C\sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\frac{\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s_h^{1,k}, s_h^{2,k}, a_h^k, b_h^k), 1\}}} \\ &= C\sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\frac{\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{N_h^k(s_h^{1,k}, s_h^{2,k}, a_h^k, b_h^k)}} \\ &\leq C\sum_{h=1}^{H} \sum_{\substack{(s^1,s^2,a,b)\in\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{A}\times\mathcal{B} \\ N_h^K(s^1,s^2,a,b)>0}} \sum_{n=1}^{N_h^K(s^1,s^2,a,b)} \sqrt{\frac{\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{n}},\end{aligned}$$

where the second equality is because $(s_h^{1,k}, s_h^{2,k}, a_h^k, b_h^k)$ is visited such that $N_h^k(s_h^{1,k}, s_h^{2,k}, a_h^k, b_h^k) \geq$

1. In addition, we have

$$\begin{aligned}
&\sum_{h=1}^{H} \sum_{\substack{(s^1,s^2,a,b)\in\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{A}\times\mathcal{B}\\ N_h^K(s^1,s^2,a,b)>0}} \sum_{n=1}^{N_h^K(s^1,s^2,a,b)} \sqrt{\frac{\log(|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK/\delta)}{n}} \\
&\qquad \le \sum_{h=1}^{H} \sum_{(s^1,s^2,a,b)\in\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{A}\times\mathcal{B}} \mathcal{O}\left(\sqrt{N_h^K(s^1,s^2,a,b)\log\frac{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|HK}{\delta}}\right) \\
&\qquad \le \mathcal{O}\left(H\sqrt{K|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|\log\frac{|\mathcal{S}_1||\mathcal{S}_2|\mathcal{A}||\mathcal{B}|HK}{\delta}}\right),
\end{aligned}$$

where the last inequality is based on the consideration that $\sum_{(s^1,s^2,a,b)\in\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{A}\times\mathcal{B}} N_h^K(s^1,s^2,a,b) = K$ such that $\sum_{(s^1,s^2,a,b)\in\mathcal{S}_1\times\mathcal{S}_2\times\mathcal{A}\times\mathcal{B}} \sqrt{N_h^K(s^1,s^2,a,b)} \le \mathcal{O}\left(\sqrt{K|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|}\right)$ when K is sufficiently large. Putting the above together, we obtain

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{r,k}(s_h^k, a_h^k, b_h^k) \le \mathcal{O}\left(H\sqrt{K|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|\log\frac{|\mathcal{S}_1||\mathcal{S}_2|\mathcal{A}||\mathcal{B}|HK}{\delta}}\right).$$

Similarly, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \beta_h^{\mathbb{P},k}(s_h^k, a_h^k, b_h^k) \\
&\quad = \sum_{k=1}^{K}\sum_{h=1}^{H} \left(\sqrt{\frac{2H^2|\mathcal{S}_1|\log(2|\mathcal{S}_1||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s_h^{1,k}, a_h^k), 1\}}} + \sqrt{\frac{2H^2|\mathcal{S}_2|\log(2|\mathcal{S}_2||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s_h^{2,k}, b_h^k), 1\}}}\right) \\
&\quad \le \mathcal{O}\left(H\sqrt{K|\mathcal{S}_1|^2|\mathcal{A}|H^2\log\frac{2|\mathcal{S}_1||\mathcal{A}|HK}{\delta}} + H\sqrt{K|\mathcal{S}_2|^2|\mathcal{B}|H^2\log\frac{2|\mathcal{S}_2||\mathcal{B}|HK}{\delta}}\right).
\end{aligned}$$

Thus, by (4.31), with probability at least $1-\delta$, we have

$$\sum_{k=1}^{K}\overline{V}_1^k(s_1) - \sum_{k=1}^{K} V_1^{\pi^k,\nu^k}(s_1) \le \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}_1|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}_2|^2|\mathcal{B}|H^4K} + \sqrt{|\mathcal{S}_1||\mathcal{S}_2||\mathcal{A}||\mathcal{B}|H^2K}),$$

where $\widetilde{\mathcal{O}}$ hides logarithmic terms. This completes the proof. □

Before presenting the next lemma, we first show the following definition of confidence set for the proof of the next lemma.

Definition 4.14 (Confidence Set for Player 2)**.** Define the following confidence set for transition models for Player 2

$$\Upsilon^{2,k} := \Big\{\widetilde{\mathbb{P}} : \Big|\widetilde{\mathbb{P}}_h(s^{2\prime}|s^2, b) - \widehat{\mathbb{P}}_h^{2,k}(s^{2\prime}|s^2, b)\Big| \leq \epsilon_h^{2,k},\ \|\widetilde{\mathbb{P}}_h(\cdot|s^2, b)\|_1 = 1,$$
$$\text{and } \widetilde{\mathbb{P}}_h(s^{2\prime}|s^2, b) \geq 0,\ \forall (s^2, b, s^{2\prime}) \in \mathcal{S}_2 \times \mathcal{B} \times \mathcal{S}_2, \forall k \in [K]\Big\}$$

where we define

$$\epsilon_h^{2,k} := 2\sqrt{\frac{\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime}|s^2, b)\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2, b) - 1, 1\}}} + \frac{14\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{3\max\{N_h^k(s^2, b) - 1, 1\}}$$

with $N_h^k(s^2, b) := \sum_{\tau=1}^k \mathbb{1}\{(s^2, b) = (s_h^{2,\tau}, b_h^\tau)\}$, and $\widehat{\mathbb{P}}^{2,k}$ being the empirical transition model for Player 2.

Lemma 4.15. *With probability at least $1 - \delta$, the difference between $q_h^{\nu^k, \mathbb{P}^2}$ and $d_h^{\nu^k, \widehat{\mathbb{P}}^{2,k}}$ is bounded as*

$$\sum_{k=1}^K \sum_{h=1}^H \sum_{s^2 \in \mathcal{S}_2} \Big| q_h^{\nu^k, \mathbb{P}^2}(s^2) - d_h^{\nu^k, \widehat{\mathbb{P}}^{2,k}}(s^2) \Big| \leq \widetilde{\mathcal{O}}\left(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K}\right).$$

Proof. By the definition of state distribution for Player 2, we have

$$\begin{aligned}
&\sum_{k=1}^K \sum_{h=1}^H \sum_{s^2 \in \mathcal{S}_2} \Big| q_h^{\nu^k, \mathbb{P}^2}(s^2) - d_h^{\nu^k, \widehat{\mathbb{P}}^{2,k}}(s^2) \Big| \\
&\qquad = \sum_{k=1}^K \sum_{h=1}^H \sum_{s^2 \in \mathcal{S}_2} \left| \sum_{b \in \mathcal{B}} w_h^{2,k}(s^2, b) - \sum_{b \in \mathcal{B}} \widehat{w}_h^{2,k}(s^2, b) \right| \\
&\qquad \leq \sum_{k=1}^K \sum_{h=1}^H \sum_{s^2 \in \mathcal{S}_2} \sum_{b \in \mathcal{B}} \big| w_h^{2,k}(s, a) - \widehat{w}_h^{2,k}(s^2, b) \big|.
\end{aligned}$$

where $\widehat{w}_h^{2,k}(s^2, b)$ is the occupancy measure under the empirical transition model $\widehat{\mathbb{P}}^{2,k}$ and the policy ν^k. Then, since $\widehat{\mathbb{P}}^{2,k} \in \Upsilon^{2,k}$ always holds for any k, by Lemma 4.19, we can bound the last term of the bound inequality such that with probability at least $1 - 6\delta'$,

$$\sum_{k=1}^K \sum_{h=1}^H \sum_{s^2 \in \mathcal{S}_2} \Big| q_h^{\nu^k, \mathbb{P}^2}(s^2) - d_h^{\nu^k, \widehat{\mathbb{P}}^{2,k}}(s^2) \Big| \leq \mathcal{E}_1 + \mathcal{E}_2.$$

Then, we compute $\mathcal{E}_1$ by Lemma 4.18. With probability at least $1-2\delta'$, we have

$$\begin{aligned}
\mathcal{E}_1 &= \mathcal{O}\left[\sum_{h=2}^{H}\sum_{h'=1}^{h-1}\sum_{k=1}^{K}\sum_{s^2\in\mathcal{S}_2}\sum_{b\in\mathcal{B}} w_h^k(s^2,b)\left(\sqrt{\frac{|\mathcal{S}_2|\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2,b),1\}}}+\frac{\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2,b),1\}}\right)\right]\\
&= \mathcal{O}\left[\sum_{h=2}^{H}\sum_{h'=1}^{h-1}\sqrt{|\mathcal{S}_2|}\left(\sqrt{|\mathcal{S}_2||\mathcal{B}|K}+|\mathcal{S}_2||\mathcal{B}|\log K+\log\frac{H}{\delta'}\right)\log\frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}\right]\\
&= \mathcal{O}\left[\left(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K}+H^2|\mathcal{S}_2|^{3/2}|\mathcal{B}|\log K+H^2\sqrt{|\mathcal{S}_2|}\log\frac{H}{\delta'}\right)\log\frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}\right]\\
&= \widetilde{\mathcal{O}}\left(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K}\right),
\end{aligned}$$

where we ignore $\log K$ when K is sufficiently large such that $\sqrt{K}$ dominates, and $\widetilde{\mathcal{O}}$ hides logarithm dependence on $|\mathcal{S}_2|$, $|\mathcal{B}|$, H, K, and $1/\delta'$. In addition, $\mathcal{E}_2$ depends on $\mathrm{ploy}(H,|\mathcal{S}_2|,|\mathcal{B}|)$ except the factor $\log\frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}$ as shown in Lemma 4.19. Thus, $\mathcal{E}_2$ can be ignored comparing to $\mathcal{E}_1$ if K is sufficiently large. Therefore, we obtain that with probability at least $1-8\delta'$, the following inequality holds

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s^2\in\mathcal{S}_2}\left|q_h^{\nu^k,\mathbb{P}^2}(s^2)-d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)\right|\le\widetilde{\mathcal{O}}\left(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K}\right).$$

We further let $\delta=8\delta'$ such that $\log\frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}=\log\frac{8|\mathcal{S}_2||\mathcal{B}|HK}{\delta}$ which does not change the order as above. Then, with probability at least $1-\delta$, we have $\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s^2\in\mathcal{S}_2}|q_h^{\nu^k,\mathbb{P}^2}(s^2)-d_h^{\nu^k,\widehat{\mathbb{P}}^{2,k}}(s^2)|\le\widetilde{\mathcal{O}}(H^2|\mathcal{S}_2|\sqrt{|\mathcal{B}|K})$. This completes the proof. □

4.7.1 Other Supporting Lemmas

Lemma 4.16. *Let $f:\Lambda\mapsto\mathbb{R}$ be a convex function, where Λ is the probability simplex defined as $\Lambda:=\{\mathbf{x}\in\mathbb{R}^d:\|\mathbf{x}\|_1=1$ and $\mathbf{x}_i\ge 0,\forall i\in[d]\}$. For any $\alpha\ge 0$, $\mathbf{z}\in\Lambda$, and $\mathbf{y}\in\Lambda^o$ where $\Lambda^o\subset\Lambda$ with only relative interior points of Λ, supposing $\mathbf{x}^{\mathrm{opt}}=\mathrm{argmin}_{\mathbf{x}\in\Lambda}f(\mathbf{x})+\alpha D_{\mathrm{KL}}(\mathbf{x},\mathbf{y})$, then the following inequality holds*

$$f(\mathbf{x}^{\mathrm{opt}})+\alpha D_{\mathrm{KL}}(\mathbf{x}^{\mathrm{opt}},\mathbf{y})\le f(\mathbf{z})+\alpha D_{\mathrm{KL}}(\mathbf{z},\mathbf{y})-\alpha D_{\mathrm{KL}}(\mathbf{z},\mathbf{x}^{\mathrm{opt}}).$$

This lemma is for mirror descent algorithms, whose proof can be obtained by slight modification from existing works [Tseng, 2008, Nemirovski et al., 2009, Wei et al., 2019].

The following lemmas are adapted from the recent papers [Efroni et al., 2020b, Jin et al., 2019], where we can find their detailed proofs.

Lemma 4.17. *With probability at least $1-4\delta'$, the true transition model $\mathbb{P}^2$ satisfies that for any $k \in [K]$,*

$$\mathbb{P} \in \Upsilon^{2,k}.$$

This lemma indicates that the estimated transition model $\widehat{\mathbb{P}}_h^{2,k}(s^{2\prime}|s^2, b)$ for Player 2 by (4.10) is closed to the true transition model $\mathbb{P}_h^2(s^{2\prime}|s^2, b)$ with high probability. The upper bound is by empirical Bernstein's inequality and the union bound.

The next lemma is adapted from Lemma 10 in Jin et al. [2019].

Lemma 4.18. *We let $w_h^{2,k}(s^2, b)$ denote the occupancy measure at the h-th step of the k-th episode under the true transition model $\mathbb{P}^2$ and the current policy ν^k. Then, with probability at least $1-2\delta'$ we have for all $h \in [H]$, the following results hold*

$$\sum_{k=1}^{K} \sum_{s^2 \in \mathcal{S}_2} \sum_{b \in \mathcal{B}} \frac{w_h^k(s^2, b)}{\max\{N_h^k(s^2, b), 1\}} = \mathcal{O}\left(|\mathcal{S}_2||\mathcal{B}| \log K + \log \frac{H}{\delta'}\right),$$

and

$$\sum_{k=1}^{K} \sum_{s^2 \in \mathcal{S}_2} \sum_{b \in \mathcal{B}} \frac{w_h^k(s^2, b)}{\sqrt{\max\{N_h^k(s^2, b), 1\}}} = \mathcal{O}\left(\sqrt{|\mathcal{S}_2||\mathcal{B}|K} + |\mathcal{S}_2||\mathcal{B}| \log K + \log \frac{H}{\delta'}\right).$$

By Lemma 4.17 and Lemma 4.18, we have the following lemma to show the difference of two occupancy measures, which is modified from parts of the proof of Lemma 4 in Jin et al. [2019].

Lemma 4.19. *For Player 2, we let $w_h^{2,k}(s^2, b)$ be the occupancy measure at the h-th step of the k-th episode under the true transition model $\mathbb{P}^2$ and the current policy ν^k, and $\widetilde{w}_h^{2,k}(s^2, b)$ be the occupancy measure at the h-th step of the k-th episode under any transition model $\widetilde{\mathbb{P}}^{2,k} \in \Upsilon^{2,k}$ and the current policy ν^k for any k. Then, with probability at least $1-6\delta'$ we have for all $h \in [H]$, the following inequality holds*

$$\sum_{k=1}^{K} \sum_{h=1}^{K} \sum_{s \in \mathcal{S}_2} \sum_{b \in \mathcal{B}} \left|\widetilde{w}_h^{2,k}(s^2, b) - w_h^{2,k}(s^2, b)\right| \leq \mathcal{E}_1 + \mathcal{E}_2,$$

where $\mathcal{E}_1$ and $\mathcal{E}_2$ are in the level of

$$\mathcal{E}_1 = \mathcal{O}\left[\sum_{h=2}^{H} \sum_{h'=1}^{h-1} \sum_{k=1}^{K} \sum_{s^2 \in \mathcal{S}_2} \sum_{b \in \mathcal{B}} w_h^k(s^2, b) \left(\sqrt{\frac{|\mathcal{S}_2| \log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2, b), 1\}}} + \frac{\log(|\mathcal{S}_2||\mathcal{B}|HK/\delta')}{\max\{N_h^k(s^2, b), 1\}}\right)\right]$$

and

$$\mathcal{E}_2 = \mathcal{O}\left(\mathrm{poly}(H, |\mathcal{S}_2|, |\mathcal{B}|) \cdot \log \frac{|\mathcal{S}_2||\mathcal{B}|HK}{\delta'}\right),$$

where $\mathrm{poly}(H, |\mathcal{S}_2|, |\mathcal{B}|)$ *denotes the polynomial dependency on* $H, |\mathcal{S}_2|, |\mathcal{B}|$.

4.8 Proofs for Markov Game with Single-Controller Transition

Lemma 4.20. *At the k-th episode of Algorithm 4, the difference between value functions $V_1^{\pi^*,\nu^k}(s_1)$ and $V_1^{\pi^k,\nu^k}(s_1)$ is*

$$\begin{aligned} &V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1) \\ &\quad = \overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1) + \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}}\Big[\big\langle \pi_h^*(\cdot|s_h) - \pi_h^k(\cdot|s_h), U_h^k(s_h,\cdot)\big\rangle_{\mathcal{A}} \,\Big|\, s_1\Big] \\ &\qquad + \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\big[\overline{\varsigma}_h^k(s_h,a_h,b_h) \,\big|\, s_1\big]. \end{aligned}$$

where s_h, a_h, b_h are random variables for state and actions, $U_h^k(s,a) := \langle \overline{Q}_h^k(s,a,\cdot), \nu_h^k(\cdot \mid s)\rangle_{\mathcal{B}}$, and we define the model prediction error of Q-function as

$$\overline{\varsigma}_h^k(s,a,b) = r_h(s,a,b) + \mathbb{P}_h\overline{V}_{h+1}^k(s,a) - \overline{Q}_h^k(s,a,b). \tag{4.32}$$

Proof. We start the proof by decomposing the value function difference as

$$V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1) = V_1^{\pi^*,\nu^k}(s_1) - \overline{V}_1^k(s_1) + \overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1). \tag{4.33}$$

Note that the term $\overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1)$ is the bias between the estimated value function $\overline{V}_1^k(s_1)$ generated by Algorithm 4 and the value function $V_1^{\pi^k,\nu^k}(s_1)$ under the true transition model $\mathbb{P}$ at the k-th episode.

We focus on analyzing the other term $V_1^{\pi^*,\nu^k}(s_1) - \overline{V}_1^k(s_1)$ in this proof. For any h and s, we have the following decomposition

$$
\begin{aligned}
& V_h^{\pi^*,\nu^k}(s) - \overline{V}_h^k(s) \\
&\quad = [\pi_h^*(\cdot|s)]^\top Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\pi_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) \\
&\quad = [\pi_h^*(\cdot|s)]^\top Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - [\pi_h^*(\cdot|s)]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) \\
&\qquad + [\pi_h^*(\cdot|s)]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\pi_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) \\
&\quad = [\pi_h^*(\cdot|s)]^\top \big[Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot) - \overline{Q}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s) \\
&\qquad + \big[\pi_h^*(\cdot|s) - \pi_h^k(\cdot|s)\big]^\top \overline{Q}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s),
\end{aligned}
\tag{4.34}
$$

where the first inequality is by the definition of $V_h^{\pi^*,\nu^k}$ in (2.3) and the definition of $\overline{V}_h^k$ in Line 8 of Algorithm 4. Moreover, by the definition of $Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot)$ in (2.4) and the model prediction error $\overline{\varsigma}_h^k$ for Player 1 in (4.32), we have

$$
\begin{aligned}
& [\pi_h^*(\cdot|s)]^\top \big[Q_h^{\pi^*,\nu^k}(s,\cdot,\cdot) - \overline{Q}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s) \\
&\quad = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^*(a|s)\bigg[\sum_{s'\in\mathcal{S}} \mathbb{P}_h(s'|s,a)\big[V_{h+1}^{\pi^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big] + \overline{\varsigma}_h^k(s,a,b)\bigg]\nu_h^k(b|s) \\
&\quad = \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \pi_h^*(a|s)\mathbb{P}_h(s'|s,a)\big[V_{h+1}^{\pi^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big] + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^*(a|s)\overline{\varsigma}_h^k(s,a,b)\nu_h^k(b|s).
\end{aligned}
$$

where the last equality holds due to $\sum_{b\in\mathcal{B}}\nu_h^k(b\,|\,s) = 1$. Combining this equality with (4.34) gives

$$
\begin{aligned}
V_h^{\pi^*,\nu^k}(s) - \overline{V}_h^k(s) &= \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \pi_h^*(a|s)\mathbb{P}_h(s'|s,a)\big[V_{h+1}^{\pi^*,\nu^k}(s') - \overline{V}_{h+1}^k(s')\big] \\
&\quad + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^*(a|s)\overline{\varsigma}_h^k(s,a,b)\nu_h^k(b|s) \\
&\quad + \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \big[\pi_h^*(a|s) - \pi_h^k(a|s)\big]\overline{Q}_h^k(s,a,b)\nu_h^k(b|s).
\end{aligned}
\tag{4.35}
$$

Note that (4.35) indicates a recursion of the value function difference $V_h^{\pi^*,\nu^k}(s) - \overline{V}_h^k(s)$. Since we define $V_{H+1}^{\pi^*,\nu^k}(s) = 0$ and $\overline{V}_{H+1}^k(s) = 0$, by recursively applying (4.35) from $h = 1$ to H, we obtain

$$
\begin{aligned}
V_1^{\pi^*,\nu^k}(s_1) - \overline{V}_1^k(s_1) &= \sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}}\big\{[\pi_h^*(\cdot|s_h)]^\top \overline{\varsigma}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\,s_1\big\} \\
&\quad + \sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}}\big\{\big[\pi_h^*(\cdot|s_h) - \pi_h^k(\cdot|s_h)\big]^\top \overline{Q}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\,s_1\big\},
\end{aligned}
\tag{4.36}
$$

where s_h are a random variables denoting the state at the h-th step following a distribution determined jointly by $\pi^*, \mathbb{P}$. Further combining (4.36) with (4.33), we eventually have

$$
\begin{aligned}
&V_1^{\pi^*,\nu^k}(s_1) - V_1^{\pi^k,\nu^k}(s_1)\\
&\qquad = \overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1) + \sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}}\big\{[\pi_h^*(\cdot|s_h)]^\top \overline{\varsigma}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\, s_1\big\}\\
&\qquad\quad + \sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}}\big\{\big[\pi_h^*(\cdot|s_h) - \pi_h^k(\cdot|s_h)\big]^\top \overline{Q}_h^k(s_h,\cdot,\cdot)\nu_h^k(\cdot|s_h)\,\big|\, s_1\big\},
\end{aligned}
$$

which is equivalent to the result in this lemma. This completes our proof. □

Lemma 4.21. *At the k-th episode of Algorithm 5, with probability at least $1-\delta$, the difference between the value functions $V_1^{\pi^k,\nu^k}(s_1)$ and $V_1^{\pi^k,\nu^*}(s_1)$ for all $k \in [K]$ is decomposed as*

$$
\begin{aligned}
&V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1)\\
&\quad \leq 2\sum_{h=1}^H \mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\big[\beta_h^{r,k}(s_h,a_h,b_h)\,\big|\, s_1\big] + \sum_{h=1}^H\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top \underline{\varsigma}_h^k(s,\cdot,\cdot)\nu_h^*(\cdot|s)\\
&\qquad + \sum_{h=1}^H\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}} + 2\sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big|q_h^{\pi^k,\mathbb{P}}(s) - d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\Big|,
\end{aligned}
$$

where s_h, a_h, b_h are random variables for state and actions, $W_h^k(s,b) = \langle \widetilde{r}_h^k(s,\cdot,b), \pi_h^k(\cdot\,|\,s)\rangle_{\mathcal{A}}$, and we define the error term as

$$
\underline{\varsigma}_h^k(s,a,b) = \widetilde{r}_h^k(s,a,b) - r_h(s,a,b). \tag{4.37}
$$

Proof. We start our proof by decomposing the value difference term for any h and s as follows

$$
\begin{aligned}
&V_h^{\pi^k,\nu^k}(s) - V_h^{\pi^k,\nu^*}(s)\\
&\qquad = \big[\pi_h^k(\cdot|s)\big]^\top Q_h^{\pi^k,\nu^k}(s,\cdot,\cdot)\nu_h^k(\cdot|s) - \big[\pi_h^k(\cdot|s)\big]^\top Q_h^{\pi^k,\nu^*}(s,\cdot,\cdot)\nu_h^*(\cdot|s)\\
&\qquad = \big[\pi_h^k(\cdot|s)\big]^\top Q_h^{\pi^k,\nu^k}(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big]\\
&\qquad\quad + \big[\pi_h^k(\cdot|s)\big]^\top \big[Q_h^{\pi^k,\nu^k}(s,\cdot,\cdot) - Q_h^{\pi^k,\nu^*}(s,\cdot,\cdot)\big]\nu_h^*(\cdot|s),
\end{aligned}
\tag{4.38}
$$

where the first equality is by the Bellman equation for $V_h^{\pi,\nu}(s)$ in (2.3) and the second equality is obtained by subtracting and adding the term $\big[\pi_h^k(\cdot|s)\big]^\top Q_h^{\pi^k,\nu^k}(s,\cdot,\cdot)\nu_h^*(\cdot|s)$ in the first equality.

Moreover, by the Bellman equation for $Q_h^{\pi,\nu}$ in (2.4), we can expand the last term in (4.38) as

$$\begin{aligned} &\big[\pi_h^k(\cdot|s)\big]^\top \big[Q_h^{\pi^k,\nu^k}(s,\cdot,\cdot) - Q_h^{\pi^k,\nu^*}(s,\cdot,\cdot)\big]\nu_h^*(\cdot|s) \\ &\quad = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^k(a|s) \sum_{s'\in\mathcal{S}} \mathbb{P}_h(s'|s,a)\big[V_{h+1}^{\pi^k,\nu^k}(s') - V_{h+1}^{\pi^k,\nu^*}(s')\big]\nu_h^*(b|s) \\ &\quad = \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \pi_h^k(a|s)\mathbb{P}_h(s'|s,a)\big[V_{h+1}^{\pi^k,\nu^k}(s') - V_{h+1}^{\pi^k,\nu^*}(s')\big]. \end{aligned} \tag{4.39}$$

where the last equality holds due to $\sum_{b\in\mathcal{B}} \nu_h^*(b \mid s) = 1$. Combining (4.39) with (4.38) gives

$$\begin{aligned} V_h^{\pi^k,\nu^k}(s) - V_h^{\pi^k,\nu^*}(s) &= \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} \pi_h^k(a|s) Q_h^{\pi^k,\nu^k}(s,a,b)\big[\nu_h^k(b|s) - \nu_h^*(b|s)\big] \\ &\quad + \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \pi_h^k(a|s)\mathbb{P}_h(s'|s,a)\big[V_{h+1}^{\pi^k,\nu^k}(s') - V_{h+1}^{\pi^k,\nu^*}(s')\big]. \end{aligned} \tag{4.40}$$

Note that (4.40) indicates a recursion of the value function difference $V_h^{\pi^k,\nu^k}(s) - V_h^{\pi^k,\nu^*}(s)$. Since we define $V_{H+1}^{\pi,\nu}(s) = 0$ for any π and ν, by recursively applying (4.40) from $h = 1$ to H, we obtain

$$\begin{aligned} &V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1) \\ &\quad = \sum_{h=1}^{H} \mathbb{E}_{\pi^k,\mathbb{P}}\big\{\big[\pi_h^k(\cdot|s_h)\big]^\top Q_h^{\pi^k,\nu^k}(s_h,\cdot,\cdot)\big[\nu_h^k(\cdot|s_h) - \nu_h^*(\cdot|s_h)\big] \,\big|\, s_1\big\}, \end{aligned} \tag{4.41}$$

where s_h are a random variables following a distribution determined jointly by $\pi^k, \mathbb{P}$. Note that since we have defined the distribution of s_h under π^k and $\mathbb{P}$ as

$$q_h^{\pi^k,\mathbb{P}}(s) = \Pr\big(s_h = s \,\big|\, \pi^k, \mathbb{P}, s_1\big),$$

we can rewrite (4.41) as

$$\begin{aligned} &V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1) \\ &\quad = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} q_h^{\pi^k,\mathbb{P}}(s)\pi_h^k(a|s) Q_h^{\pi^k,\nu^k}(s,a,b)\big[\nu_h^k(b|s) - \nu_h^*(b|s)\big]. \end{aligned} \tag{4.42}$$

By plugging the Bellman equation for Q-function as (2.4) into (4.42), we further expand (4.42) as

$$
\begin{aligned}
&V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1)\\
&\quad = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} q_h^{\pi^k,\mathbb{P}}(s)\pi_h^k(a|s)\big[r_h(s,a,b) + \big\langle \mathbb{P}_h(\cdot|s,a), V_{h+1}^{\pi^k,\nu^k}(\cdot)\big\rangle\big]\left[\nu_h^k(b|s) - \nu_h^*(b|s)\right]\\
&\quad = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} q_h^{\pi^k,\mathbb{P}}(s)\pi_h^k(a|s)\left[r_h(s,a,b)\right]\left[\nu_h^k(b|s) - \nu_h^*(b|s)\right]\\
&\quad = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} q_h^{\pi^k,\mathbb{P}}(s)[\pi_h^k(\cdot|s)]^\top r_h(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big],
\end{aligned}
$$

where the second equality by

$$
\begin{aligned}
&\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} q_h^{\pi^k,\mathbb{P}}(s)\pi_h^k(a|s)\big\langle \mathbb{P}_h(\cdot|s,a), V_{h+1}^{\pi^k,\nu^k}(\cdot)\big\rangle_{\mathcal{S}}\left[\nu_h^k(b|s) - \nu_h^*(b|s)\right]\\
&\qquad = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} q_h^{\pi^k,\mathbb{P}}(s)\pi_h^k(a|s)\big\langle \mathbb{P}_h(\cdot|s,a), V_{h+1}^{\pi^k,\nu^k}(\cdot)\big\rangle_{\mathcal{S}}\sum_{b\in\mathcal{B}}\left[\nu_h^k(b|s) - \nu_h^*(b|s)\right]\\
&\qquad = 0.
\end{aligned}
$$

In particular, the last equality above is due to

$$
\sum_{b\in\mathcal{B}}\big[\nu_h^k(b|s) - \nu_h^*(b|s)\big] = 1 - 1 = 0.
$$

Thus, we have

$$
V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1) = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} q_h^{\pi^k,\mathbb{P}}(s)\big[\pi_h^k(\cdot|s)\big]^\top r_h(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big]. \tag{4.43}
$$

Recall that we also define the estimate of the state reaching probability $q_h^{\pi^k,\mathbb{P}}(s)$ as

$$
d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s) = \Pr\big(s_h = s \,\big|\, \pi^k, \widehat{\mathbb{P}}^k, s_1\big).
$$

Now we define the following term associated with $\widehat{\mathbb{P}}^k$, $\widehat{r}^k$, π^k, ν^k, and the initial state s_1 as

$$
\underline{V}_1^k := \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s),
$$

with $\widetilde{r}$ defined in Line 7 of Algorithm 5, which is

$$\widetilde{r}_h^k(s,a,b) = \max\big\{\widehat{r}_h^k(s,a,b) - \beta_h^{r,k}(s,a,b),\ 0\big\}.$$

Thus, by (4.43), we have the following decomposition

$$\begin{aligned}
&V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1)\\
&= V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1) - \underline{V}_1^k + \underline{V}_1^k\\
&= \underbrace{\sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big\{q_h^{\pi^k,\mathbb{P}}(s)\big[\pi_h^k(\cdot|s)\big]^\top r_h(s,\cdot,\cdot)\nu_h^k(\cdot|s) - d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\Big\}}_{\text{Term(I)}}\\
&\quad + \underbrace{\sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big\{d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s) - q_h^{\pi^k,\mathbb{P}}(s)\big[\pi_h^k(\cdot|s)\big]^\top r_h(s,\cdot,\cdot)\nu_h^*(\cdot|s)\Big\}}_{\text{Term(II)}}.
\end{aligned} \tag{4.44}$$

We first bound Term(I) as

$$\begin{aligned}
\text{Term(I)} &= \sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big\{q_h^{\pi^k,\mathbb{P}}(s)\big[\pi_h^k(\cdot|s)\big]^\top r_h(s,\cdot,\cdot)\nu_h^k(\cdot|s) - d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\Big\}\\
&= \sum_{h=1}^H\sum_{s\in\mathcal{S}} q_h^{\pi^k,\mathbb{P}}(s)\big[\pi_h^k(\cdot|s)\big]^\top\big[r_h(s,\cdot,\cdot) - \widetilde{r}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s)\\
&\quad + \sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big[q_h^{\pi^k,\mathbb{P}}(s) - d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\Big]\big[\pi_h^k(\cdot|s)\big]^\top \widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\\
&\le 2\sum_{h=1}^H \mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\big[\beta_h^{r,k}(s,a,b)\big] + \sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big|q_h^{\pi^k,\mathbb{P}}(s) - d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\Big|,
\end{aligned} \tag{4.45}$$

where the inequality is due to $|\widehat{r}_h^k(s,a,b) - r_h(s,a,b)| \le \beta_h^{r,k}(s,a,b)$ with probability at least $1-\delta$ by Lemma 4.23 such that we have

$$\begin{aligned}
r_h(s,a,b) - \widetilde{r}_h^k(s,a,b) &= r_h(s,a,b) - \max\big\{\widehat{r}_h^k(s,a,b) - \beta_h^{r,k}(s,a,b), 0\big\}\\
&= \min\big\{r_h(s,a,b) - \widehat{r}_h^k(s,a,b) + \beta_h^{r,k}(s,a,b), r_h(s,a,b)\big\}\\
&\le r_h(s,a,b) - \widehat{r}_h^k(s,a,b) + \beta_h^{r,k}(s,a,b) \le 2\beta_h^{r,k}(s,a,b)
\end{aligned}$$

and then

$$\sum_{s\in\mathcal{S}} q_h^{\pi^k,\mathbb{P}}(s)\big[\pi_h^k(\cdot|s)\big]^\top\big[r_h(s,\cdot,\cdot)-\widetilde{r}_h^k(s,\cdot,\cdot)\big]\nu_h^k(\cdot|s)\le 2\mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\big[\beta_h^{r,k}(s,a,b)\big].$$

In addition, the inequality in (4.45) is also due to

$$\begin{aligned}\Big|\big[\pi_h^k(\cdot|s)\big]^\top\widetilde{r}_h^k(s,\cdot,\cdot)\nu_h^k(\cdot|s)\Big| &\le \Big|\sum_a\sum_b \pi_h^k(a|s)\widetilde{r}_h^k(s,a,b)\nu_h^k(b|s)\Big|\\ &\le \sum_a\sum_b \pi_h^k(a|s)\cdot\big|\widetilde{r}_h^k(s,a,b)\big|\cdot\nu_h^k(b|s)\le 1,\end{aligned}$$

because of $0\le\widetilde{r}_h^k(s,a,b)=\max\big\{\widehat{r}_h^k(s,a,b)-\beta_h^{r,k}(s,a,b),0\big\}\le\widehat{r}_h^k(s,a,b)\le 1$. Therefore, with probability at least $1-\delta$, we have

$$\text{Term(I)}\le 2\sum_{h=1}^H\mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\big[\beta_h^{r,k}(s_h,a_h,b_h)\big]+\sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big|q_h^{\pi^k,\mathbb{P}}(s)-d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\Big|. \tag{4.46}$$

Next, we bound Term(II) in the following way

$$\begin{aligned}\text{Term(II)} &= \sum_{h=1}^H\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top\widetilde{r}_h^k(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s)-\nu_h^*(\cdot|s)\big]\\ &\quad+\sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big[d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)-q_h^{\pi^k,\mathbb{P}}(s)\Big]\big[\pi_h^k(\cdot|s)\big]^\top r_h(s,\cdot,\cdot)\nu_h^*(\cdot|s)\\ &\quad+\sum_{h=1}^H\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top\varsigma_h^k(s,\cdot,\cdot)\nu_h^*(\cdot|s),\end{aligned}$$

where $\varsigma_h^k(s,a,b)$ is defined in (4.37). Here the first term in the above equality is associated with the mirror descent step in Algorithm 5. The second term can be similarly bounded by $\sum_{h=1}^H\sum_{s\in\mathcal{S}}|q_h^{\pi^k,\mathbb{P}}(s)-d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)|$. Thus, we have

$$\begin{aligned}\text{Term(II)} &\le \sum_{h=1}^H\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top\widetilde{r}_h^k(s,\cdot,\cdot)\big[\nu_h^k(\cdot|s)-\nu_h^*(\cdot|s)\big] \qquad (4.47)\\ &\quad+\sum_{h=1}^H\sum_{s\in\mathcal{S}}\Big|q_h^{\pi^k,\mathbb{P}}(s)-d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\Big|+\sum_{h=1}^H\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big[\pi_h^k(\cdot|s)\big]^\top\varsigma_h^k(s,\cdot,\cdot)\nu_h^*(\cdot|s).\end{aligned}$$

Combining (4.46), (4.47) with (4.44), we obtain that with probability at least $1-\delta$, the following inequality holds

$$
\begin{aligned}
&V_1^{\pi^k,\nu^k}(s_1) - V_1^{\pi^k,\nu^*}(s_1) \\
&\quad \le 2\sum_{h=1}^{H} \mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\left[\beta_h^{r,k}(s_h,a_h,b_h) \,\middle|\, s_1\right] + \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\left[\pi_h^k(\cdot|s)\right]^\top \underline{\varsigma}_h^k(s,\cdot,\cdot)\nu_h^*(\cdot|s) \\
&\qquad + \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}} + 2\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} \left| q_h^{\pi^k,\mathbb{P}}(s) - d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\right|,
\end{aligned}
$$

where $W_h^k(s,b) = \langle \widetilde{r}_h^k(s,\cdot,b), \pi_h^k(\cdot\mid s)\rangle_{\mathcal{A}}$. This completes our proof. □

Lemma 4.22. *With setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, the mirror ascent steps of Algorithm 4 lead to*

$$
\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P}}\Big[\big\langle \pi_h^*(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}\Big] \le \mathcal{O}\left(\sqrt{H^4 K \log|\mathcal{A}|}\right),
$$

where $U_h^k(s,a) = \langle \overline{Q}_h^k(s,a,\cdot), \nu_h^k(\cdot|s)\rangle_{\mathcal{B}}$, $\forall (s,a)\in\mathcal{S}\times\mathcal{A}$.

Proof. As shown in (4.12), the mirror ascent step at the k-th episode is to solve the following maximization problem

$$
\underset{\pi}{\text{maximize}} \sum_{h=1}^{H} \big\langle \pi_h(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} - \frac{1}{\eta}\sum_{h=1}^{H} D_{\mathrm{KL}}\big(\pi_h(\cdot|s), \pi_h^k(\cdot|s)\big),
$$

with $U_h^k(s,a) = \langle \overline{Q}_h^k(s,a,\cdot), \nu_h^k(\cdot|s)\rangle_{\mathcal{B}}$. We can further equivalently rewrite this maximization problem as a minimization problem as

$$
\underset{\pi}{\text{minimize}} - \sum_{h=1}^{H} \big\langle \pi_h(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} + \frac{1}{\eta}\sum_{h=1}^{H} D_{\mathrm{KL}}\big(\pi_h(\cdot|s), \pi_h^k(\cdot|s)\big).
$$

Note that the closed-form solution $\pi_h^{k+1}(\cdot|s), \forall s\in\mathcal{S}$, to this minimization problem is guaranteed to stay in the relative interior of a probability simplex when initialize $\pi_h^0(\cdot|s) = \mathbf{1}/|\mathcal{A}|$. Thus, we can apply Lemma 4.16 and obtain that for any $\pi = \{\pi_h\}_{h=1}^H$, the following inequality holds

$$
\begin{aligned}
&-\eta\big\langle \pi_h^{k+1}(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} + \eta\big\langle \pi_h(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \\
&\quad \le D_{\mathrm{KL}}\big(\pi_h(\cdot|s), \pi_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi_h(\cdot|s), \pi_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi_h^{k+1}(\cdot|s), \pi_h^k(\cdot|s)\big).
\end{aligned}
$$

Then, by rearranging the terms, we have

$$\begin{aligned}
&\eta\big\langle \pi_h^*(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \\
&\qquad \le D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s), \pi_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s), \pi_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi_h^{k+1}(\cdot|s), \pi_h^k(\cdot|s)\big) \\
&\qquad\quad + \eta\big\langle \pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}.
\end{aligned} \tag{4.48}$$

Due to Pinsker's inequality, we have

$$- D_{\mathrm{KL}}\big(\pi_h^{k+1}(\cdot|s), \pi_h^k(\cdot|s)\big) \le -\frac{1}{2}\big\|\pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s)\big\|_1^2.$$

Moreover, by Cauchy-Schwarz inequality, we have

$$\eta\big\langle \pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \le \eta H\big\|\pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s)\big\|_1.$$

Thus, we have

$$\begin{aligned}
&- D_{\mathrm{KL}}\big(\pi_h^{k+1}(\cdot|s), \pi_h^k(\cdot|s)\big) + \eta\big\langle \pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \\
&\qquad \le -\frac{1}{2}\big\|\pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s)\big\|_1^2 + \eta H\big\|\pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s)\big\|_1 \le \frac{1}{2}\eta^2 H^2,
\end{aligned} \tag{4.49}$$

where the last inequality is by viewing $\big\|\pi_h^{k+1}(\cdot|s) - \pi_h^k(\cdot|s)\big\|_1$ as a variable x and finding the maximal value of $-1/2 \cdot x^2 + \eta H x$ to obtain the upper bound $1/2 \cdot \eta^2 H^2$.

Thus, combing (4.49) with (4.48), the policy improvement step in Algorithm 4 implies

$$\begin{aligned}
&\eta\big\langle \pi_h^*(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}} \\
&\qquad \le D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s), \pi_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s), \pi_h^{k+1}(\cdot|s)\big) + \frac{1}{2}\eta^2 H^2,
\end{aligned}$$

which further leads to

$$\begin{aligned}
&\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}}\Big[\big\langle \pi_h^*(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}\Big] \\
&\qquad \le \frac{1}{\eta}\sum_{h=1}^H \mathbb{E}_{\pi^*,\mathbb{P}}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s), \pi_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s), \pi_h^{k+1}(\cdot|s)\big)\big] + \frac{1}{2}\eta H^3.
\end{aligned}$$

Moreover, we take summation from $k = 1$ to K of both sides and then obtain

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\mathbb{P}}\Big[\big\langle \pi_h^*(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}\Big] \\
&\quad\le \frac{1}{\eta}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\mathbb{P}}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s),\pi_h^1(\cdot|s)\big) - D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s),\pi_h^{K+1}(\cdot|s)\big)\big] + \frac{1}{2}\eta K H^3 \\
&\quad\le \frac{1}{\eta}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\mathbb{P}}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s),\pi_h^1(\cdot|s)\big)\big] + \frac{1}{2}\eta K H^3,
\end{aligned}$$

where the last inequality is non-negativity of KL divergence. By the initialization in Algorithm 4, it is guaranteed that $\pi_h^1(\cdot|s) = \mathbf{1}/|\mathcal{A}|$, which thus leads to $D_{\mathrm{KL}}\left(\pi_h^*(\cdot|s),\pi_h^1(\cdot|s)\right) \le \log|\mathcal{A}|$. Then, with setting $\eta = \sqrt{\log|\mathcal{A}|/(KH^2)}$, we bound the last term as

$$\frac{1}{\eta}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\mathbb{P}}\big[D_{\mathrm{KL}}\big(\pi_h^*(\cdot|s),\pi_h^1(\cdot|s)\big)\big] + \frac{1}{2}\eta K H^3 \le \mathcal{O}\left(\sqrt{H^4 K\log|\mathcal{A}|}\right),$$

which gives

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\mathbb{P}}\Big[\big\langle \pi_h^*(\cdot|s) - \pi_h^k(\cdot|s), U_h^k(s,\cdot)\big\rangle_{\mathcal{A}}\Big] \le \mathcal{O}\left(\sqrt{H^4 K\log|\mathcal{A}|}\right),$$

This completes the proof. □

Lemma 4.23. *For any $k \in [K]$, $h \in [H]$ and all $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$, with probability at least $1-\delta$, we have*

$$\big|\widehat{r}_h^k(s,a,b) - r_h(s,a,b)\big| \le \sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}.$$

This lemma is the same as Lemma 4.10. We rewrite it here for the completeness of the proofs in this section. In (4.11), we set $\beta_h^{r,k}(s,a,b) = \sqrt{\frac{4\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s,a,b),1\}}}$, which equals the bound in Lemma 4.23.

Lemma 4.24. *For any $k \in [K]$, $h \in [H]$ and all $(s,a) \in \mathcal{S}\times\mathcal{A}$, with probability at least $1-\delta$, we have*

$$\left\|\widehat{\mathbb{P}}_h^k(\cdot\,|\,s,a) - \mathbb{P}_h(\cdot\,|\,s,a)\right\|_1 \le \sqrt{\frac{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}}.$$

Proof. For $k \ge 1$, we have $\|\widehat{\mathbb{P}}_h^k(\cdot\,|\,s,a) - \mathbb{P}_h(\cdot\,|\,s,a)\|_1 = \max_{\|\mathbf{z}\|_\infty\le 1}\langle\widehat{\mathbb{P}}_h^k(\cdot\,|\,s,a) - \mathbb{P}_h(\cdot\,|\,s,a),\mathbf{z}\rangle_{\mathcal{S}}$

by the duality. We construct an ϵ-cover for the set $\{\mathbf{z} \in \mathbb{R}^{|\mathcal{S}|} : \|\mathbf{z}\|_\infty \leq 1\}$ with the distance induced by $\|\cdot\|_\infty$, denoted as $\mathcal{C}_\infty(\epsilon)$, such that for any $\mathbf{z} \in \mathbb{R}^{|\mathcal{S}|}$, there always exists $\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)$ satisfying $\|\mathbf{z} - \mathbf{z}'\|_\infty \leq \epsilon$. The covering number is $\mathcal{N}_\infty(\epsilon) = |\mathcal{C}_\infty(\epsilon)| = 1/\epsilon^{|\mathcal{S}|}$. Thus, we have for any $(s, a) \in \mathcal{S} \times \mathcal{A}$ and any $\mathbf{z}$ with $\|\mathbf{z}\|_\infty \leq 1$, there exists $\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)$ such that $\|\mathbf{z}' - \mathbf{z}\|_\infty \leq \epsilon$ and

$$
\begin{aligned}
&\big\langle \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a), \mathbf{z} \big\rangle_{\mathcal{S}} \\
&\quad = \big\langle \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a), \mathbf{z}' \big\rangle_{\mathcal{S}} + \big\langle \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a), \mathbf{z} - \mathbf{z}' \big\rangle_{\mathcal{S}} \\
&\quad \leq \big\langle \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a), \mathbf{z}' \big\rangle_{\mathcal{S}} + \epsilon \left\| \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a) \right\|_1,
\end{aligned}
$$

such that we further have

$$
\begin{aligned}
&\left\| \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a) \right\|_1 \\
&\quad = \max_{\|\mathbf{z}\|_\infty \leq 1} \big\langle \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a), \mathbf{z} \big\rangle_{\mathcal{S}} \\
&\quad \leq \max_{\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)} \big\langle \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a), \mathbf{z}' \big\rangle_{\mathcal{S}} + \epsilon \left\| \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a) \right\|_1.
\end{aligned}
\tag{4.50}
$$

By Hoeffding's inequality and the union bound over all $\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)$, when $N_h^k(s, a) > 0$, with probability at least $1 - \delta'$ where $\delta' \in (0, 1]$,

$$
\max_{\mathbf{z}' \in \mathcal{C}_\infty(\epsilon)} \big\langle \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a), \mathbf{z}' \big\rangle_{\mathcal{S}} \leq \sqrt{\frac{|\mathcal{S}| \log(1/\epsilon) + \log(1/\delta')}{2N_h^k(s, a)}}. \tag{4.51}
$$

Letting $\epsilon = 1/2$, by (4.50) and (4.51), with probability at least $1 - \delta'$, we have

$$
\left\| \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a) \right\|_1 \leq 1 \sqrt{\frac{|\mathcal{S}| \log 2 + \log(1/\delta')}{2N_h^k(s, a)}}.
$$

When $N_h^k(s, a) = 0$, we have $\big\| \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a) \big\|_1 = \|\mathbb{P}_h(\cdot \,|\, s, a)\|_1 = 1$ such that $2\sqrt{\frac{|\mathcal{S}| \log 2 + \log(1/\delta')}{2}} > 1 = \big\| \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a) \big\|_1$ always holds. Thus, with probability at least $1 - \delta'$,

$$
\left\| \widehat{\mathbb{P}}_h^k(\cdot \,|\, s, a) - \mathbb{P}_h(\cdot \,|\, s, a) \right\|_1 \leq 2 \sqrt{\frac{|\mathcal{S}| \log 2 + \log(1/\delta')}{2 \max\{N_h^k(s, a), 1\}}} \leq \sqrt{\frac{2|\mathcal{S}| \log(2/\delta')}{\max\{N_h^k(s, a), 1\}}}.
$$

Then, by the union bound, assuming $K > 1$, letting $\delta = |\mathcal{S}||\mathcal{A}|HK\delta'/2$, with probability at least

$1-\delta$, for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ and any $h\in[H]$ and $k\in[K]$, we have

$$\left\|\widehat{\mathbb{P}}_h^k(\cdot\,|\,s,a)-\mathbb{P}_h(\cdot\,|\,s,a)\right\|_1\leq\sqrt{\frac{2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}},$$

This completes the proof. □

In (4.11), we set $\beta_h^{\mathbb{P},k}(a,b)=\sqrt{\frac{2H^2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s,a),1\}}}$, which equals the product of the upper bound in Lemma 4.24 and the factor H.

Lemma 4.25. *With probability at least* $1-2\delta$*, Algorithm 4 ensures that*

$$\sum_{k=1}^K\sum_{h=1}^H\mathbb{E}_{\pi^*,\mathbb{P},\nu^k}\left[\overline{\varsigma}_h^k(s_h,a_h,b_h)\,\middle|\,s_1\right]\leq 0.$$

Proof. We prove the upper bound of the model prediction error term. We can decompose the instantaneous prediction error at the h-step of the k-th episode as

$$\overline{\varsigma}_h^k(s,a,b)=r_h(s,a,b)+\left\langle\mathbb{P}_h(\cdot\,|\,s,a),\overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}}-\overline{Q}_h^k(s,a,b),\tag{4.52}$$

where the equality is by the definition of the prediction error in (4.32). By plugging in the definition of $\overline{Q}_h^k$ in Line 7 of Algorithm 4, for any (s,a,b), we bound the following term as

$$\begin{aligned}&r_h(s,a,b)+\left\langle\mathbb{P}_h(\cdot\,|\,s,a),\overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}}-\overline{Q}_h^k(s,a,b)\\&\quad\leq r_h(s,a,b)+\left\langle\mathbb{P}_h(\cdot\,|\,s,a),\overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}}\\&\qquad-\min\left\{\widehat{r}_h^k(s,a,b)+\left\langle\widehat{\mathbb{P}}_h^k(\cdot|s,a),\overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}}-\beta_h^k,H-h+1\right\}\\&\quad\leq\max\left\{r_h(s,a,b)-\widehat{r}_h^k(s,a,b)+\left\langle\mathbb{P}_h(\cdot\,|\,s,a)-\widehat{\mathbb{P}}_h^k(\cdot|s,a),\overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}}-\beta_h^k,0\right\},\end{aligned}\tag{4.53}$$

where the inequality holds because

$$\begin{aligned}&r_h(s,a,b)+\left\langle\mathbb{P}_h(\cdot\,|\,s_h,a_h),\overline{V}_{h+1}^k(\cdot)\right\rangle_{\mathcal{S}}\\&\quad\leq r_h(s,a,b)+\left\|\mathbb{P}_h(\cdot\,|\,s_h,a_h)\right\|_1\|\overline{V}_{h+1}^k(\cdot)\|_\infty\leq 1+\max_{s'\in\mathcal{S}}\left|\overline{V}_{h+1}^k(s')\right|\leq 1+H-h,\end{aligned}$$

since $\left\|\mathbb{P}_h(\cdot\,|\,s_h,a_h)\right\|_1=1$ and also the truncation step as shown in Line 7 of Algorithm 4 for $\overline{Q}_{h+1}^k$

such that for any $s' \in \mathcal{S}$

$$\begin{aligned}\big|\overline{V}_{h+1}^k(s')\big| &= \Big|\big[\pi_{h+1}^k(\cdot|s')\big]^\top \overline{Q}_{h+1}^k(s',\cdot,\cdot)\nu_{h+1}^k(\cdot|s')\Big| \\ &\le \big\|\pi_{h+1}^k(\cdot|s')\big\|_1 \big\|\overline{Q}_{h+1}^k(s',\cdot,\cdot)\nu_{h+1}^k(\cdot|s')\big\|_\infty \\ &\le \max_{a,b} \big|\overline{Q}_{h+1}^k(s',a,b)\big| \le H-h.\end{aligned} \tag{4.54}$$

Combining (4.52) and (4.53) gives

$$\begin{aligned}\overline{\varsigma}_h^k(s,a,b) \le \max\Big\{ & r_h(s,a,b) - \widehat{r}_h^k(s,a,b) \\ & + \big\langle \mathbb{P}_h(\cdot \mid s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} - \beta_h^k, 0\Big\}.\end{aligned} \tag{4.55}$$

Note that as shown in (4.11), we have

$$\beta_h^k(s,a,b) = \beta_h^{r,k}(s,a,b) + \beta_h^{\mathbb{P},k}(s,a).$$

Then, with probability at least $1-\delta$, we have

$$\begin{aligned}& r_h(s,a,b) - \widehat{r}_h^k(s,a,b) - \beta_h^{r,k}(s,a,b) \\ &\quad \le \big|r_h(s,a,b) - \widehat{r}_h^k(s,a,b)\big| - \beta_h^{r,k}(s,a,b) \\ &\quad \le \beta_h^{r,k}(s,a,b) - \beta_h^{r,k}(s,a,b) = 0,\end{aligned}$$

where the last inequality is by Lemma 4.23 and the setting of the bonus for the reward. Moreover, with probability at least $1-\delta$, we have

$$\begin{aligned}& \big\langle \mathbb{P}_h(\cdot \mid s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} - \beta_h^{\mathbb{P},k}(s,a) \\ &\quad \le \big\|\mathbb{P}_h(\cdot \mid s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a)\big\|_1 \big\|\overline{V}_{h+1}^k(\cdot)\big\|_\infty - \beta_h^{\mathbb{P},k}(s,a) \\ &\quad \le H\big\|\mathbb{P}_h(\cdot \mid s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a)\big\|_1 - \beta_h^{\mathbb{P},k}(s,a) \\ &\quad \le \beta_h^{\mathbb{P},k}(s,a) - \beta_h^{\mathbb{P},k}(s,a) = 0,\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the second inequality is due to $\max_{s'\in\mathcal{S}} \big\|\overline{V}_{h+1}^k(s')\big\|_\infty \le H$ as shown in (4.54), and the last inequality is by the setting of $\beta_h^{\mathbb{P},k}$ and also Lemma 4.24. Thus, with probability at least $1-2\delta$, the following inequality holds

$$r_h(s,a,b) - \widehat{r}_h^k(s,a,b) + \big\langle \mathbb{P}_h(\cdot \mid s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} - \beta_h^k(s,a,b) \le 0.$$

Combining the above inequality with (4.55), we have that with probability at least $1-2\delta$, for any

$h \in [H]$ and $k \in [K]$, the following inequality holds

$$\overline{\varsigma}_h^k(s,a,b) \leq 0, \ \forall (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B},$$

which leads to

$$\sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\mathbb{P},\nu^k} \left[\overline{\varsigma}_h^k(s_h,a_h,b_h) \,\middle|\, s_1 \right] \leq 0.$$

This completes the proof. □

Lemma 4.26. *With probability at least* $1-\delta$*, Algorithm 4 ensures that*

$$\sum_{k=1}^{K} \overline{V}_1^k(s_1) - \sum_{k=1}^{K} V_1^{\pi^k,\nu^k}(s_1) \leq \widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{S}|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K} \right).$$

Proof. We assume that a trajectory $\{(s_h^k, a_h^k, b_h^k, s_{h+1}^k)\}_{h=1}^H$ for all $k \in [K]$ is generated following the policies π^k, ν^k, and the true transition model $\mathbb{P}$. Thus, we expand the bias term at the h-th step of the k-th episode, which is

$$\begin{aligned} \overline{V}_h^k(s_h^k) - V_h^{\pi^k,\nu^k}(s_h^k) &= \left[\pi_h^k(\cdot|s_h^k)\right]^\top \left[\overline{Q}_h^k(s_h^k,\cdot,\cdot) - Q_h^{\pi^k,\nu^k}(s_h^k,\cdot,\cdot)\right]\nu_h^k(\cdot|s_h^k) \\ &= \zeta_h^k + \overline{Q}_h^k(s_h^k,a_h^k,b_h^k) - Q_h^{\pi^k,\nu^k}(s_h^k,a_h^k,b_h^k) \\ &= \zeta_h^k + \left\langle \mathbb{P}_h(\cdot \,|\, s_h^k,a_h^k), \overline{V}_{h+1}^k(\cdot) - V_{h+1}^{\pi^k,\nu^k}(\cdot) \right\rangle_{\mathcal{S}} - \overline{\varsigma}_h^k(s_h^k,a_h^k,b_h^k) \\ &= \zeta_h^k + \xi_h^k + \overline{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k,\nu^k}(s_{h+1}^k) - \overline{\varsigma}_h^k(s_h^k,a_h^k,b_h^k), \end{aligned} \tag{4.56}$$

where the first equality is by Line 8 of Algorithm 4 and (2.3), the third equality is by plugging in (2.4) and (4.32). Specifically, in the above equality, we introduce two martingale difference sequence, namely, $\{\zeta_h^k\}_{h\geq 0,k\geq 0}$ and $\{\xi_h^k\}_{h\geq 0,k\geq 0}$, which are defined as

$$\begin{aligned} \zeta_h^k &:= \left[\pi_h^k(\cdot|s_h^k)\right]^\top \left[\overline{Q}_h^k(s_h^k,\cdot,\cdot) - Q_h^{\pi^k,\nu^k}(s_h^k,\cdot,\cdot)\right]\nu_h^k(\cdot|s_h^k) - \left[\overline{Q}_h^k(s_h^k,a_h^k,b_h^k) - Q_h^{\pi^k,\nu^k}(s_h^k,a_h^k,b_h^k)\right], \\ \xi_h^k &:= \left\langle \mathbb{P}_h(\cdot \,|\, s_h^k,a_h^k), \overline{V}_{h+1}^k(\cdot) - V_{h+1}^{\pi^k,\nu^k}(\cdot) \right\rangle_{\mathcal{S}} - \left[\overline{V}_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k,\nu^k}(s_{h+1}^k)\right], \end{aligned}$$

such that

$$\mathbb{E}_{a_h^k \sim \pi_h^k(\cdot|s_h^k), b_h^k \sim \nu_h^k(\cdot|s_h^k)}\left[\zeta_h^k \,\middle|\, \mathcal{F}_h^k\right] = 0, \qquad \mathbb{E}_{s_{h+1}^k \sim \mathbb{P}_h(\cdot \,|\, s_h^k,a_h^k)}\left[\xi_h^k \,\middle|\, \widetilde{\mathcal{F}}_h^k\right] = 0,$$

with $\mathcal{F}_h^k$ being the filtration of all randomness up to $(h-1)$-th step of the k-th episode plus s_h^k, and $\widetilde{\mathcal{F}}_h^k$ being the filtration of all randomness up to $(h-1)$-th step of the k-th episode plus s_h^k, a_h^k, b_h^k.

We can observe that the equality (4.56) construct a recursion for $\overline{V}_h^k(s_h^k) - V_h^{\pi^k,\nu^k}(s_h^k)$. Moreover, we also have $\overline{V}_{H+1}^{k}(\cdot) = \mathbf{0}$ and $V_{H+1}^{\pi^k,\nu^k}(\cdot) = \mathbf{0}$. Thus, recursively apply (4.56) from $h = 1$ to H leads to the following equality

$$\overline{V}_1^k(s_1) - V_1^{\pi^k,\nu^k}(s_1) = \sum_{h=1}^{H} \zeta_h^k + \sum_{h=1}^{H} \xi_h^k - \sum_{h=1}^{H} \overline{\varsigma}_h^k(s_h^k, a_h^k, b_h^k). \tag{4.57}$$

Moreover, by (4.32) and Line 7 of Algorithm 4, we have

$$\begin{aligned}-\overline{\varsigma}_h^k(s_h^k, a_h^k, b_h^k) &= -r_h(s_h^k, a_h^k, b_h^k) - \big\langle \mathbb{P}_h(\cdot \,|\, s_h, a_h), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} \\ &\quad + \min\big\{\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) + \big\langle \widehat{\mathbb{P}}_h^k(\cdot|s_h, a_h), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \beta_h^k(s_h^k, a_h^k, b_h^k), H - h + 1\big\}.\end{aligned}$$

Then, we can further bound $-\overline{\varsigma}_h^k(s_h^k, a_h^k, b_h^k)$ as follows

$$\begin{aligned}-\overline{\varsigma}_h^k(s_h^k, a_h^k, b_h^k) &\leq -r_h(s_h^k, a_h^k, b_h^k) - \big\langle \mathbb{P}_h(\cdot \,|\, s_h^k, a_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \widehat{r}_h^k(s_h^k, a_h^k, b_h^k) \\ &\quad + \big\langle \widehat{\mathbb{P}}_h^k(\cdot|s_h^k, a_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}} + \beta_h^k(s_h^k, a_h^k, b_h^k) \\ &\leq \big|\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) - r_h(s_h^k, a_h^k, b_h^k)\big| \\ &\quad + \Big|\big\langle \mathbb{P}_h(\cdot \,|\, s_h^k, a_h^k) - \widehat{\mathbb{P}}_h^k(\cdot \,|\, s_h^k, a_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\Big| + \beta_h^k(s_h^k, a_h^k, b_h^k),\end{aligned}$$

where the first inequality is due to $\min\{x, y\} \leq x$. Additionally, we have

$$\begin{aligned}\Big|\big\langle \mathbb{P}_h(\cdot \,|\, s_h^k, a_h^k) - \widehat{\mathbb{P}}_h^k(\cdot \,|\, s_h^k, a_h^k), \overline{V}_{h+1}^k(\cdot)\big\rangle_{\mathcal{S}}\Big| &\leq \big\|\overline{V}_{h+1}^k(\cdot)\big\|_\infty \big\|\mathbb{P}_h(\cdot \,|\, s_h^k, a_h^k) - \widehat{\mathbb{P}}_h^k(\cdot \,|\, s_h^k, a_h^k)\big\|_1 \\ &\leq H\big\|\mathbb{P}_h(\cdot \,|\, s_h^k, a_h^k) - \widehat{\mathbb{P}}_h^k(\cdot \,|\, s_h^k, a_h^k)\big\|_1,\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality and the second inequality is by (4.54). Thus, putting the above together, we obtain

$$\begin{aligned}-\overline{\varsigma}_h^k(s_h^k, a_h^k, b_h^k) &\leq \big|\widehat{r}_h^k(s_h^k, a_h^k, b_h^k) - r_h(s_h^k, a_h^k, b_h^k)\big| + H\big\|\overline{V}_{h+1}^k(\cdot) - \overline{V}_{h+1}^k(\cdot)\big\|_1 + \beta_h^k(s_h^k, a_h^k, b_h^k) \\ &\leq 2\beta_h^{r,k}(s_h^k, a_h^k, b_h^k) + 2\beta_h^{\mathbb{P},k}(s_h^k, a_h^k),\end{aligned}$$

where the second inequality is by Lemma 4.23, Lemma 4.24, and the decomposition of the bonus term β_h^k as (4.11). Due to Lemma 4.23 and Lemma 4.24, by the union bound, for any $h \in [H], k \in [K]$ and $(s_h, a_h, b_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$, the above inequality holds with probability at least $1 - 2\delta$.

Therefore, by (4.57), with probability at least $1-2\delta$, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\left[\overline{V}_1^k(s_1)-V_1^{\pi^k,\nu^k}(s_1)\right] \\
&\qquad\le \sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^k+\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h^k+2\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^{r,k}(s_h^k,a_h^k,b_h^k)+2\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^{\mathbb{P},k}(s_h^k,a_h^k).
\end{aligned} \tag{4.58}$$

By Azuma-Hoeffding inequality, with probability at least $1-\delta$, the following inequalities hold

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\zeta_h^k \le \mathcal{O}\left(\sqrt{H^3K\log\frac{1}{\delta}}\right),\\
\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h^k \le \mathcal{O}\left(\sqrt{H^3K\log\frac{1}{\delta}}\right),
\end{aligned}$$

where we use the facts that $|\overline{Q}_h^k(s_h^k,a_h^k,b_h^k)-Q_h^{\pi^k,\nu^k}(s_h^k,a_h^k,b_h^k)|\le 2H$ and $|\overline{V}_{h+1}^k(s_{h+1}^k)-V_{h+1}^{\pi^k,\nu^k}(s_{h+1}^k)|\le 2H$. Next, we need to bound $\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^{r,k}(s_h^k,a_h^k,b_h^k)$ and $\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^{\mathbb{P},k}(s_h^k,a_h^k)$ in (4.58). We show that

$$\begin{aligned}
\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^{r,k}(s_h^k,a_h^k,b_h^k) &= C\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{\max\{N_h^k(s_h^k,a_h^k,b_h^k),1\}}}\\
&= C\sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{N_h^k(s_h^k,a_h^k,b_h^k)}}\\
&\le C\sum_{h=1}^{H}\sum_{\substack{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}\\ N_h^K(s,a,b)>0}}\sum_{n=1}^{N_h^K(s,a,b)}\sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{n}},
\end{aligned}$$

where the second equality is because (s_h^k,a_h^k,b_h^k) is visited such that $N_h^k(s_h^k,a_h^k,b_h^k)\ge 1$. In addition, we have

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{\substack{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}\\ N_h^K(s,a,b)>0}}\sum_{n=1}^{N_h^K(s,a,b)}\sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{n}}\\
&\qquad\le\sum_{h=1}^{H}\sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}}\mathcal{O}\left(\sqrt{N_h^K(s,a,b)\log\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK}{\delta}}\right)\\
&\qquad\le\mathcal{O}\left(H\sqrt{K|\mathcal{S}||\mathcal{A}||\mathcal{B}|\log\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK}{\delta}}\right),
\end{aligned}$$

where the last inequality is based on the consideration that $\sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} N_h^K(s,a,b) = K$ such that $\sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} \sqrt{N_h^K(s,a,b)} \le \mathcal{O}\left(\sqrt{K|\mathcal{S}||\mathcal{A}||\mathcal{B}|}\right)$ when K is sufficiently large. Putting the above together, we obtain

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^{r,k}(s_h^k,a_h^k,b_h^k) \le \mathcal{O}\left(H\sqrt{K|\mathcal{S}||\mathcal{A}||\mathcal{B}|\log\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK}{\delta}}\right).$$

Similarly, we have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H}\beta_h^{\mathbb{P},k}(s_h^k,a_h^k) &= \sum_{k=1}^{K}\sum_{h=1}^{H}\sqrt{\frac{H^2|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta)}{\max\{N_h^k(s_h^k,a_h^k),1\}}}\\ &\le \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathcal{O}\left(\sqrt{N_h^K(s,a)H^2|\mathcal{S}|\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta}}\right)\\ &\le \sum_{h=1}^{H}\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathcal{O}\left(\sqrt{\sum_{b\in B}N_h^K(s,a,b)H^2|\mathcal{S}|\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta}}\right)\\ &\le \mathcal{O}\left(H\sqrt{K|\mathcal{S}|^2|\mathcal{A}|H^2\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta}}\right),\end{aligned}$$

where the second inequality is due to $\sum_{b\in\mathcal{B}} N_h^K(s,a,b) = N_h^K(s,a)$, and the last inequality is based on the consideration that $\sum_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} N_h^K(s,a,b) = K$ such that $\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sqrt{\sum_{b\in\mathcal{B}} N_h^K(s,a,b)} \le \mathcal{O}(\sqrt{K|\mathcal{S}||\mathcal{A}|})$ when K is sufficiently large.

Thus, by (4.58), with probability at least $1-\delta$, we have

$$\sum_{k=1}^{K}\overline{V}_1^k(s_1) - \sum_{k=1}^{K}V_1^{\pi^k,\nu^k}(s_1) \le \widetilde{\mathcal{O}}(\sqrt{|\mathcal{S}|^2|\mathcal{A}|H^4K} + \sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K})$$

where $\widetilde{\mathcal{O}}$ hides logarithmic terms. This completes the proof. □

Lemma 4.27. *With setting $\gamma = \sqrt{|\mathcal{S}|\log|\mathcal{B}|/K}$, the mirror descent steps of Algorithm 5 lead to*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\langle W_h^k(s,\cdot),\nu_h^k(\cdot|s)-\nu_h^*(\cdot|s)\rangle \le \mathcal{O}\left(\sqrt{H^2|\mathcal{S}|K\log|\mathcal{B}|}\right),$$

where $W_h^k(s,b) = \langle\widetilde{r}_h^k(s,\cdot,b),\pi_h^k(\cdot\mid s)\rangle_{\mathcal{A}}$.

Proof. Similar to the proof of Lemma 4.22, and also by Lemma 4.16, for any $\nu=\{\nu_h\}_{h=1}^H$ and

$s \in \mathcal{S}$, the mirror descent step in Algorithm 5 leads to

$$\begin{aligned}&\gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^{k+1}(\cdot|s)\big\rangle_{\mathcal{B}} - \gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h(\cdot|s)\big\rangle_{\mathcal{B}}\\&\qquad \le D_{\mathrm{KL}}\big(\nu_h(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h(\cdot|s), \nu_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^{k+1}(\cdot|s), \nu_h^k(\cdot|s)\big),\end{aligned}$$

according to (4.13), where $W_h^k(s,b) = \big\langle \pi_h^k(\cdot|s), \widetilde{r}_h^k(s,\cdot,b)\big\rangle$. Then, by rearranging the terms, we have

$$\begin{aligned}&\gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}}\\&\qquad \le D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{k+1}(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^{k+1}(\cdot|s), \nu_h^k(\cdot|s)\big)\\&\qquad\quad - \gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\rangle_{\mathcal{B}}.\end{aligned} \tag{4.59}$$

Due to Pinsker's inequality, we have

$$-D_{\mathrm{KL}}\big(\nu_h^{k+1}(\cdot|s), \nu_h^k(\cdot|s)\big) \le -\frac{1}{2}\big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1^2. \tag{4.60}$$

Moreover, we have

$$\begin{aligned}&-\gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^{k+1}(\cdot|s)\big\rangle_{\mathcal{B}}\\&\qquad \le \gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\|W_h^k(s,\cdot)\big\|_\infty\big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1\\&\qquad \le \gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1,\end{aligned} \tag{4.61}$$

where the last inequality is by

$$\begin{aligned}\|W_h^k(s,\cdot)\|_\infty &= \max_{b\in\mathcal{B}} W_h^k(s,b) \le \max_{s\in\mathcal{S},b\in\mathcal{B}} W_h^k(s,b)\\&\le \max_{s\in\mathcal{S},b\in\mathcal{B}} \big\langle \widetilde{r}_h^{k-1}(s,\cdot,b), \pi_h^k(\cdot\,|\,s)\big\rangle\\&\le \max_{s\in\mathcal{S},b\in\mathcal{B}} \big\|\widetilde{r}_h^{k-1}(s,\cdot,b)\big\|_\infty\big\|\pi_h^k(\cdot\,|\,s)\big\|_1 \le 1.\end{aligned}$$

due to the definition of W_h^k and $0 \le \widetilde{r}_h^k(s,a,b) = \max\{\widehat{r}_h^k(s,a,b) - \beta_h^{r,k}, 0\} \le \widehat{r}_h^k(s,a,b) \le 1$. Combining (4.60) and (4.61) gives

$$\begin{aligned}&-D_{\mathrm{KL}}\big(\nu_h^{k+1}(\cdot|s), \nu_h^k(\cdot|s)\big) - \gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^{k+1}(\cdot|s)\big\rangle\\&\qquad \le -\frac{1}{2}\big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1^2 + \gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\big\|_1\\&\qquad \le \frac{1}{2}\big[d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big]^2\gamma^2 \le \frac{1}{2}d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\gamma^2,\end{aligned}$$

where the second inequality is obtained via solving $\max_x\{-1/2 \cdot x^2 + \gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s) \cdot x\}$ if letting $x = \|\nu_h^{k+1}(\cdot|s) - \nu_h^k(\cdot|s)\|_1$. Plugging the above inequality into (4.59) gives

$$\begin{aligned}&\gamma d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}}\\ &\qquad \leq D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{k+1}(\cdot|s)\big) + \frac{1}{2} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\gamma^2.\end{aligned}$$

Thus, the policy improvement step implies

$$\begin{aligned}&\sum_{h=1}^H \sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}}\\ &\qquad \leq \frac{1}{\gamma}\sum_{h=1}^H \sum_{s\in\mathcal{S}} \big[D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{k+1}(\cdot|s)\big)\big] + \frac{1}{\gamma}\sum_{h=1}^H \sum_{s\in\mathcal{S}} \frac{1}{2} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\gamma^2\\ &\qquad \leq \frac{1}{\gamma}\sum_{h=1}^H \sum_{s\in\mathcal{S}} \big[D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^k(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{k+1}(\cdot|s)\big)\big] + \frac{1}{2}H\gamma.\end{aligned}$$

Further taking summation from $k = 1$ to K on both sides of the above inequality gives

$$\begin{aligned}&\sum_{k=1}^K \sum_{h=1}^H \sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}}\\ &\qquad \leq \frac{1}{\gamma}\sum_{h=1}^H \sum_{s\in\mathcal{S}} \big[D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^1(\cdot|s)\big) - D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^{K+1}(\cdot|s)\big)\big] + \frac{1}{2}HK\gamma\\ &\qquad \leq \frac{1}{\gamma}\sum_{h=1}^H \sum_{s\in\mathcal{S}} D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^1(\cdot|s)\big) + \frac{1}{2}HK\gamma.\end{aligned}$$

Note that by the initialization in Algorithm 5, it is guaranteed that $\nu_h^1(\cdot|s) = \mathbf{1}/|\mathcal{B}|$, which thus leads to $D_{\mathrm{KL}}\left(\pi_h^*(\cdot|s), \pi_h^1(\cdot|s)\right) \leq \log|\mathcal{B}|$. By setting $\gamma = \sqrt{|\mathcal{S}|\log|\mathcal{B}|/K}$, we further bound the term as

$$\frac{1}{\gamma}\sum_{h=1}^H \sum_{s\in\mathcal{S}} D_{\mathrm{KL}}\big(\nu_h^*(\cdot|s), \nu_h^1(\cdot|s)\big) + \frac{1}{2}HK\gamma \leq \mathcal{O}\left(\sqrt{H^2|\mathcal{S}|K\log|\mathcal{B}|}\right),$$

which gives

$$\sum_{k=1}^K \sum_{h=1}^H \sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\big\langle W_h^k(s,\cdot), \nu_h^k(\cdot|s) - \nu_h^*(\cdot|s)\big\rangle_{\mathcal{B}} \leq \mathcal{O}\left(\sqrt{H^2|\mathcal{S}|K\log|\mathcal{B}|}\right).$$

This completes the proof. □

Lemma 4.28. *With probability at least* $1-\delta$*, Algorithm 5 ensures that*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)[\pi_h^k(\cdot|s)]^\top \varsigma_h^k(s,\cdot,\cdot)\nu_h^*(\cdot|s) \leq 0,$$

where $\varsigma_h^k(s,a,b) = \widetilde{r}_h^k(s,a,b) - r_h(s,a,b)$*.*

Proof. With probability at least $1-\delta$, for any $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}, h\in[H], k\in[K]$, we have

$$\begin{aligned}\varsigma_h^k(s,a,b) &= \widetilde{r}_h^k(s,a,b) - r_h(s,a,b)\\ &= \max\big\{\widehat{r}_h^{k-1}(s,a,b) - r_h(s,a,b) - \beta_h^{r,k-1}, -r_h(s,a,b)\big\}\\ &\leq \max\big\{0, -r_h(s,a,b)\big\} = 0,\end{aligned}$$

where $\widetilde{r}_h^k(s,a,b)$ is computed as in Algorithm 5 and the inequality is by $\widehat{r}_h^{k-1}(s,a,b) - r_h(s,a,b) - \beta_h^{r,k-1} \leq 0$ with probability at least $1-\delta$ by Lemma 4.23. The above result reflects the optimism of $\widetilde{r}_h^k$. Therefore, with probability at least $1-\delta$, we have

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)[\pi_h^k(\cdot|s)]^\top \varsigma_h^k(s,\cdot,\cdot)\nu_h^*(\cdot|s)\\ &\quad= \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\sum_{a,b}\pi_h^k(a|s)\big[\widetilde{r}_h^k(s,a,b) - r_h(s,a,b)\big]\nu_h^*(b|s)\\ &\quad\leq 0.\end{aligned}$$

This completes the proof. □

Before giving the next lemma, we first present the following definition for the proof of the next lemma.

Definition 4.29 (Confidence Set)**.** Define the following confidence set for transition models

$$\begin{aligned}\Upsilon^k := \Big\{\widetilde{\mathbb{P}} : \Big|\widetilde{\mathbb{P}}_h(s'|s,a) - \widehat{\mathbb{P}}_h^k(s'|s,a)\Big| \leq \epsilon_h^k,\ \|\widetilde{\mathbb{P}}_h(\cdot|s,a)\|_1 = 1,\\ \text{and } \widetilde{\mathbb{P}}_h(s'|s,a) \geq 0,\ \forall (s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}, \forall k\in[K]\Big\}\end{aligned}$$

where we define

$$\epsilon_h^k := 2\sqrt{\frac{\widehat{\mathbb{P}}_h^k(s'|s,a)\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a)-1,1\}}} + \frac{14\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{3\max\{N_h^k(s,a)-1,1\}}$$

with $N_h^k(s,a) := \sum_{\tau=1}^{k}\mathbb{1}\{(s,a) = (s_h^\tau, a_h^\tau)\}$ and $\widehat{\mathbb{P}}^k$ being the empirical transition model.

Lemma 4.30. *With probability at least* $1-\delta$*, the difference between* $q^{\pi^k,\mathbb{P}}$ *and* d^k *are bounded as*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left|q_h^{\pi^k,\mathbb{P}}(s)-d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\right|\leq\widetilde{\mathcal{O}}\left(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K}\right).$$

Proof. By the definition of state distribution, we first have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left|q_h^{\pi^k,\mathbb{P}}(s)-d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\right|\\
&\quad=\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left|\sum_{a\in\mathcal{A}}w_h^k(s,a)-\sum_{a\in\mathcal{A}}\widehat{w}_h^k(s,a)\right|\\
&\quad\leq\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left|w_h^k(s,a)-\widehat{w}_h^k(s,a)\right|.
\end{aligned}$$

where $\widehat{w}_h^k(s,a)$ is the occupancy measure under the empirical transition model $\widehat{\mathbb{P}}^k$ and the policy π^k. Then, since $\widehat{\mathbb{P}}^k\in\Upsilon^k$ always holds for any k, by Lemma 4.34, we can bound the last term of the bound inequality such that with probability at least $1-6\delta'$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left|q_h^{\pi^k,\mathbb{P}}(s)-d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\right|\leq\mathcal{E}_1+\mathcal{E}_2.$$

Next, we compute the order of $\mathcal{E}_1$ by Lemma 4.33. With probability at least $1-2\delta'$, we have

$$\begin{aligned}
\mathcal{E}_1&=\mathcal{O}\left[\sum_{h=2}^{H}\sum_{h'=1}^{h-1}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}w_h^k(s,a)\left(\sqrt{\frac{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a),1\}}}+\frac{\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a),1\}}\right)\right]\\
&=\mathcal{O}\left[\sum_{h=2}^{H}\sum_{h'=1}^{h-1}\sqrt{|\mathcal{S}|}\left(\sqrt{|\mathcal{S}||\mathcal{A}|K}+|\mathcal{S}||\mathcal{A}|\log K+\log\frac{H}{\delta'}\right)\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'}\right]\\
&=\mathcal{O}\left[\left(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K}+H^2|\mathcal{S}|^{3/2}|\mathcal{A}|\log K+H^2\sqrt{|\mathcal{S}|}\log\frac{H}{\delta'}\right)\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'}\right]\\
&=\widetilde{\mathcal{O}}\left(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K}\right),
\end{aligned}$$

where we ignore $\log K$ terms when K is sufficiently large such that $\sqrt{K}$ dominates, and $\widetilde{\mathcal{O}}$ hides logarithm dependence on $|\mathcal{S}|$, $|\mathcal{A}|$, H, K, and $1/\delta'$. On the other hand, $\mathcal{E}_2$ also depends on $\operatorname{ploy}(H,|\mathcal{S}|,|\mathcal{A}|)$ except the factor $\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'}$ as shown in Lemma 4.34. Thus, $\mathcal{E}_2$ can be ignored comparing to $\mathcal{E}_1$ if K is sufficiently large. Therefore, we eventually obtain that with probability at least $1-8\delta'$, the following inequality holds

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\left|q_h^{\pi^k,\mathbb{P}}(s)-d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)\right|\leq\widetilde{\mathcal{O}}\left(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K}\right).$$

We let $\delta=8\delta'$ such that $\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'}=\log\frac{8|\mathcal{S}||\mathcal{A}|HK}{\delta}$ without changing the order as shown above. Then, with probability at least $1-\delta$, we have $\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{s\in\mathcal{S}}|q_h^{\pi^k,\mathbb{P}}(s)-d_h^{\pi^k,\widehat{\mathbb{P}}^k}(s)|\leq\widetilde{\mathcal{O}}(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|K})$. This completes the proof. □

Lemma 4.31. *With probability at least $1-\delta$, the following inequality holds*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\left[\beta_h^{r,k}(s_h,a_h,b_h)\,\middle|\,s_1\right]\leq\widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K}\right).$$

Proof. Since we have

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\left[\beta_h^{r,k}(s_h,a_h,b_h)\,\middle|\,s_1\right]\\&\quad=\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\left[C\sqrt{\frac{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK/\delta)}{N_h^k(s,a,b)}}\right]\\&\quad=C\sqrt{\log\frac{|\mathcal{S}||\mathcal{A}||\mathcal{B}|HK}{\delta}}\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\left[\sqrt{\frac{1}{N_h^k(s,a,b)}}\right],\end{aligned}$$

then we can apply Lemma 4.35 and obtain

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\left[\beta_h^{r,k}(s_h,a_h,b_h)\,\middle|\,s_1\right]\leq\widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K}\right),$$

with probability at least $1-\delta$. Here $\widetilde{\mathcal{O}}$ hides logarithm dependence on $|\mathcal{S}|,|\mathcal{A}|,|\mathcal{B}|,H,K$, and $1/\delta$. This completes the proof. □

4.8.1 Other Supporting Lemmas

Lemma 4.32. *With probability at least $1-4\delta'$, the true transition model $\mathbb{P}$ satisfies that for any $k\in[K]$,*

$$\mathbb{P}\in\Upsilon^k.$$

This lemma implies that the estimated transition model $\widehat{\mathbb{P}}_h^k(s'|s,a)$ by (4.10) is closed to the true transition model $\mathbb{P}_h(s'|s,a)$ with high probability. The upper bound for their difference is by empirical Bernstein's inequality and the union bound.

The next lemma is modified from Lemma 10 in Jin et al. [2019].

Lemma 4.33. *We let $w_h^k(s,a)$ denote the occupancy measure at the h-th step of the k-th episode under the true transition model $\mathbb{P}$ and the current policy π^k. Then, with probability at least $1-2\delta'$ we have for all $h \in [H]$, the following results hold*

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{w_h^k(s,a)}{\max\{N_h^k(s,a),1\}} = \mathcal{O}\left(|\mathcal{S}||\mathcal{A}|\log K + \log\frac{H}{\delta'}\right),$$

and

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{w_h^k(s,a)}{\sqrt{\max\{N_h^k(s,a),1\}}} = \mathcal{O}\left(\sqrt{|\mathcal{S}||\mathcal{A}|K} + |\mathcal{S}||\mathcal{A}|\log K + \log\frac{H}{\delta'}\right).$$

Furthermore, by Lemma 4.32 and Lemma 4.33, we give the following lemma to characterize the difference of two occupancy measures, which is modified from parts of the proof of Lemma 4 in Jin et al. [2019].

Lemma 4.34. *Let $w_h^k(s,a)$ be the occupancy measure at the h-th step of the k-th episode under the true transition model $\mathbb{P}$ and the current policy π^k, and $\widetilde{w}_h^k(s,a)$ be the occupancy measure at the h-th step of the k-th episode under any transition model $\widetilde{\mathbb{P}}^k \in \Upsilon^k$ and the current policy π^k for any k. Then, with probability at least $1-6\delta'$ we have $\forall h \in [H]$, the following inequality holds*

$$\sum_{k=1}^{K}\sum_{h=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\left|\widetilde{w}_h^k(s,a) - w_h^k(s,a)\right| \leq \mathcal{E}_1 + \mathcal{E}_2,$$

where $\mathcal{E}_1$ and $\mathcal{E}_2$ are in the level of

$$\mathcal{E}_1 = \mathcal{O}\left[\sum_{h=2}^{H}\sum_{h'=1}^{h-1}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} w_h^k(s,a)\left(\sqrt{\frac{|\mathcal{S}|\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a),1\}}} + \frac{\log(|\mathcal{S}||\mathcal{A}|HK/\delta')}{\max\{N_h^k(s,a),1\}}\right)\right]$$

and

$$\mathcal{E}_2 = \mathcal{O}\left(\mathrm{poly}(H,|\mathcal{S}|,|\mathcal{A}|)\cdot\log\frac{|\mathcal{S}||\mathcal{A}|HK}{\delta'}\right),$$

where $\mathrm{poly}(H,|\mathcal{S}|,|\mathcal{A}|)$ denotes the polynomial dependency on $H, |\mathcal{S}|, |\mathcal{A}|$.

Lemma 4.35. *With probability at least $1-\delta$, the following inequality holds*

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^k,\mathbb{P},\nu^k}\left[\sqrt{\frac{1}{\max\{N_h^k(s,a,b),1\}}}\,\right]\leq\widetilde{\mathcal{O}}\left(\sqrt{|\mathcal{S}||\mathcal{A}||\mathcal{B}|H^2K}+|\mathcal{S}||\mathcal{A}||\mathcal{B}|H\right),$$

where $\widetilde{\mathcal{O}}$ hides logarithmic terms.

Proof. The zero-sum Markov game with single-controller transition can interpreted as a regular MDP learning problem with policies $w_h^k(a,b\,|\,s)=\pi_h^k(a|s)\nu_h^k(b|s)$ and a transition model $\mathbb{P}_h(s'|s,a,b)=\mathbb{P}_h(s'|s,a)$ with a joint action (a,b) in the action space of size $|\mathcal{A}||\mathcal{B}|$. Thus, we apply Lemma 19 of Efroni et al. [2020b], which extends lemmas in Zanette and Brunskill [2019], Efroni et al. [2019] to MDP with non-stationary dynamics by adding a factor of H, to obtain our lemma. This completes the proof. □

CHAPTER 5

Reward-Free RL with Kernel and Neural Function Approximations

5.1 Introduction

While RL with function approximations has achieved great empirical success [Mnih et al., 2015, Silver et al., 2016, 2017, Vinyals et al., 2019], its application is mostly enabled by massive interactions with the unknown environment, especially when the state space is large and function approximators such as neural networks are employed. To achieve sample efficiency, any RL algorithm needs to accurately learn the transition model either explicitly or implicitly, which brings the need for efficient exploration.

Under the setting of offline RL, agents aim to learn the target policy only from an offline dataset collected a priori, without any interactions with the environment. Thus, the collected offline dataset should have sufficient coverage of the trajectory generated by the optimal policy. However, in real-world RL applications, the reward function is often designed by the learner based on domain knowledge. The learner might have a set of reward functions to choose from or use an adaptive algorithm for reward design [Laud, 2004, Grzes, 2017]. In such a scenario, it is often desirable to collect an offline dataset that covers all the possible trajectories associated with a set of reward functions and the target policies. With such a benign offline dataset, for any arbitrary reward function, the RL agents have sufficient information to estimate the corresponding policy.

To study such a problem in a principled manner, we focus on the framework of reward-free RL, which consists of an exploration phase and a planning phase. Specifically, in the exploration phase, the agents interact with the environment without accessing pre-specified rewards and collect empirical trajectories for the subsequent planning phase. During the planning phase, using the offline data collected in the exploration phase, the agents compute the target policy when given an extrinsic reward function, without further interactions with the environment.

Recently, many works focus on designing provably sample-efficient reward-free RL algorithms. For the single-agent tabular case, Jin et al. [2020a], Kaufmann et al. [2020], Ménard et al. [2020], Zhang et al. [2020] achieve $\widetilde{\mathcal{O}}(\mathrm{poly}(H, |\mathcal{S}|, |\mathcal{A}|)/\varepsilon^2)$ sample complexity for obtaining ε-suboptimal policy, where $|\mathcal{S}|, |\mathcal{A}|$ are the sizes of state and action space, respectively. In view of the large action and state spaces, the works Zanette et al. [2020b], Wang et al. [2020a] theoretically analyze reward-free RL by applying the linear function approximation for the single-agent Markov decision process (MDP), which achieve $\widetilde{\mathcal{O}}(\mathrm{poly}(H, \mathfrak{d})/\varepsilon^2)$ sample complexity with $\mathfrak{d}$ denoting the dimension of the feature space. However, RL algorithms combined with nonlinear function approximators such as the kernel and neural function approximators have shown great empirical successes in a variety of application problems (e.g., Duan et al. [2016], Silver et al. [2016, 2017], Wang et al. [2018], Vinyals et al. [2019]), thanks to their expressive power. On the other hand, although reward-free RL algorithms for the multi-player Markov games in the tabular case have been studied in Bai and Jin [2020], Liu et al. [2020], there is still a lack of works theoretically studying multi-agent scenarios with the function approximation. Thus, the following question remains open:

Can we design provably efficient reward-free RL algorithms with kernel and neural function approximations for both single-agent MDPs and Markov games?

The main challenges of answering the above question lie in how to appropriately integrate nonlinear approximators into the framework of reward-free RL and how to incentivize the exploration by designing exploration rewards and bonuses that fit such approximation. In this chapter, we provide an affirmative answer to the above question by tackling these challenges. Our contributions are summarized as follows:

Contributions. In this chapter, we first propose provable sample and computationally efficient reward-free RL algorithms with kernel and neural function approximations for the single-agent MDP setting. Our exploration algorithm is an optimistic variant of the least-squares value iteration algorithm, incorporating kernel and neural function approximators, which adopts the associated (scaled) bonus as the exploration reward. Further with the planning phase, our method achieves an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity to generate an ε-suboptimal policy for an *arbitrary* extrinsic reward function. Moreover, we extend the proposed method for the single-agent setting to the zero-sum Markov game setting such that the algorithm can achieve an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity to generate a policy pair which is an ε-approximate Nash equilibrium. Particularly, in the planning phase for Markov games, our algorithm only involves finding the Nash equilibrium of matrix games formed by Q-function that can be solved *efficiently*, which is of independent interest. The sample complexities of our methods match the $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ results in existing works for tabular or linear function approximation settings. To the best of our knowledge, we establish the first provably efficient reward-free RL algorithms with kernel and neural function approximators for both single-agent and multi-agent settings.

Related Work. There have been a lot of works focusing on designing provably efficient reward-free RL algorithms for both single-agent and multi-agent RL problems. For the single-agent scenario, Jin et al. [2020a] formalizes the reward-free RL for the tabular setting and provide theoretical analysis for the proposed algorithm with an $\widetilde{\mathcal{O}}(\mathrm{poly}(H, |\mathcal{S}|, |\mathcal{A}|)/\varepsilon^2)$ sample complexity for achieving ε-suboptimal policy. The sample complexity for the tabular setting is further improved in several recent works [Kaufmann et al., 2020, Ménard et al., 2020, Zhang et al., 2020]. Recently, Zanette et al. [2020b], Wang et al. [2020a] study the reward-free RL from the perspective of the linear function approximation, which inspire us to design reward-free RL algorithms with more powerful nonlinear function approximators. For the multi-agent setting, Bai and Jin [2020] studies the reward-free exploration for the zero-sum Markov game for the tabular case. Liu et al. [2020] further proposes provable reward-free RL algorithms for multi-player general-sum games.

Our work in this chapter is also closely related to a line of works that study RL algorithms with function approximations. There are many works [Yang and Wang, 2019, 2020, Cai et al., 2019, Zanette et al., 2020a, Jin et al., 2020b, Wang et al., 2019b, Ayoub et al., 2020, Zhou et al., 2020, Kakade et al., 2020] studying different RL problems with the (generalized) linear function approximation. Furthermore, Wang et al. [2020b] studies an optimistic LSVI algorithm for general function approximation. Our work is most closely related to the recent work Yang et al. [2020], which studies optimistic LSVI algorithms with kernel and neural function approximations. However, this chapter studies an online single-agent RL problem where the exploration is executed with reward feedbacks, which cannot be directly applied to the reward-free RL problem. Inspired by Yang et al. [2020], this chapter extends the idea of kernel and neural function approximations to the reward-free RL setting and Markov games.

5.2 Preliminaries

In this section, we introduce the basic notations and problem backgrounds for this chapter.

5.2.1 Problem Setup

In this chapter, we first consider a tabular episodic MDP characterized by $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ as defined in Section 2.1 of Chapter 2. The policy for the agent is denoted by π. For any reward function r, the value function $V^\pi(s, r)$ and the Q-function $Q^\pi(s, a, r)$ are defined the same as in Chapter 2. We can further define the Bellman equation, optimal policy π_r^* for a certain reward function r, and ε-suboptimal policy as in Chapter 2. The value function and Q-function associated with the optimal policy π_r^* is then denoted as Q^* and V^*. Thus, we have $V_h^*(s, r) = V_h^{\pi_r^*}(s, r)$ and $Q_h^*(s, a, r) = Q_h^{\pi_r^*}(s, a, r)$. Here we consider the practical and challenging setting that the true transition model $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is *unknown* to the agent. At the k-step of the exploration phase, we let $\pi^k = \{\pi_h^k\}_{h=1}^H$ be the exploration policy of the agent. And in the planning phase, given an arbitrary reward function $r = \{r_h\}_{h=1}^H$, we let $\pi = \{\pi_h\}_{h=1}^H$ be the output (the learned policy) the algorithm. For ease of theoretical analysis, we assume the function value of r is normalized in the range $[0, 1]$, i.e., $r_h(s, a) \in [0, 1]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Furthermore, we study the reward-free RL for the two-player zero-sum Markov game, which is defined by $(\mathcal{S}, \mathcal{A}, \mathcal{B}, H, \mathbb{P}, r)$ as in Section 2.2 of Chapter 2. The policies for Player 1 and Player 2 are denoted by π and ν respectively. The value function $V^{\pi,\nu}(s, r)$ and the Q-function $Q^{\pi,\nu}(s, a, b, r)$ are defined the same as in Chapter 2 for any reward function r. Therefore, we further define the Bellman equation, NE, best response, and ε-approximate NE as in Chapter 2. Similarly, we consider the practical and challenging setting that the true transition model $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ is *unknown* to both players. At the k-step of the exploration phase, we let $\pi^k = \{\pi_h^k\}_{h=1}^H$ and $\nu^k = \{\nu_h^k\}_{h=1}^H$ be the exploration policies for the players. In the planning phase, given an arbitrary reward function $r = \{r_h\}_{h=1}^H$, we let $\pi = \{\pi_h\}_{h=1}^H$ and $\nu = \{\nu_h\}_{h=1}^H$ be the learned policy pair. In addition, we make an assumption that the function value of r is normalized in the range $[0, 1]$, i.e., $r_h(s, a, b) \in [0, 1]$ for all $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. By slightly abusing the notation, we define π_r^* and ν_r^* as the solution to the maximization problem $\max_{\pi,\nu} V_1^{\pi,\nu}(s_1)$ such that we let $V_h^*(s, r) = V_h^{\pi_r^*,\nu_r^*}(s, r)$ and $Q_h^*(s, a, b, r) = Q_h^{\pi_r^*,\nu_r^*}(s, a, b, r)$. Moreover, for simplicity of notation, the notation rule in this chapter also follows Remark 2.1 in Chapter 2.

5.2.2 Reproducing Kernel Hilbert Space

We study the kernel function approximation based on the reproducing kernel Hilbert space (RKHS). With slight abuse of notion, we let $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$ for the single-agent MDP setting and $\mathcal{Z} = \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ for the zero-sum game setting, such that $z = (s, a) \in \mathcal{Z}$ or $z = (s, a, b) \in \mathcal{Z}$ for different cases. We assume that the space $\mathcal{Z}$ is the input space of the approximation function, where $\mathcal{Z}$ is a compact space on $\mathbb{R}^d$. This can also be achieved if there is a preprocessing method to embed (s, a) or (s, a, b) into the space $\mathbb{R}^d$. We let $\mathcal{H}$ be a RKHS defined on the space $\mathcal{Z}$ with the kernel function $\ker : \mathcal{Z} \times \mathcal{Z} \mapsto \mathbb{R}$. We further define the inner product on the RKHS $\mathcal{H}$ as $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \mapsto \mathbb{R}$ and the norm $\| \cdot \|_{\mathcal{H}} : \mathcal{H} \mapsto \mathbb{R}$. We have a feature map $\phi : \mathcal{Z} \mapsto \mathcal{H}$ on the RKHS $\mathcal{H}$ and define the function $f(z) := \langle f, \phi(z) \rangle_{\mathcal{H}}$ for $f \in \mathcal{H}$. Then the kernel is defined as

$$\ker(z, z') := \langle \phi(z), \phi(z') \rangle_{\mathcal{H}}, \quad \forall z, z' \in \mathcal{Z}.$$

We assume that $\sup_{z \in \mathcal{Z}} \ker(z, z) \leq 1$ such that $\|\phi(z)\|_{\mathcal{H}} \leq 1$ for any $z \in \mathcal{Z}$.

5.2.3 Overparameterized Neural Network

This chapter further considers a function approximator utilizing the overparameterized neural network. Overparameterized neural networks have drawn a lot of attention recently in both theory and practice [Neyshabur et al., 2018, Allen-Zhu et al., 2018, Arora et al., 2019, Gao et al., 2019, Bai and Lee, 2019]. Specifically, in this chapter, we have a two-layer neural network $f(\cdot; b, W) : \mathcal{Z} \mapsto \mathbb{R}$ with $2m$ neurons and weights $(\boldsymbol{v}, W)$, which can be represented as

$$f(z; \boldsymbol{v}, W) = \frac{1}{\sqrt{2m}} \sum_{i=1}^{2m} v_i \cdot \texttt{act}(W_i^\top z), \tag{5.1}$$

where $\texttt{act}$ is the activation function, and $\boldsymbol{v} = [v_1, \cdots, v_{2m}]^\top$ and $W = [W_1, W_2, \cdots, W_{2m}]$. Here, we assume that $z = (s, a)$ or $z = (s, a, b)$ with $z \in \mathcal{Z}$ satisfies $\|z\|_2 = 1$, i.e., z is normalized on a unit hypersphere in $\mathbb{R}^d$. Let $W^{(0)}$ be the initial value of W and $\boldsymbol{v}^{(0)}$ be the initialization of $\boldsymbol{v}$. The initialization step for the above model is performed as follows: we let $v_i \sim \text{Unif}(\{-1, 1\})$ and $W_i^{(0)} \sim N(0, I_d/d)$ for all $i \in [m]$, where I_d is an identity matrix in $\mathbb{R}^{d\times d}$, and $v_i^{(0)} = -v_{i-m}^{(0)}$, $W_i^{(0)} = W_{i-m}^{(0)}$ for all $i \in \{m+1, 2m\}$. Here we let $N(0, I_d/d)$ denote Gaussian distribution. In this chapter, we let $\boldsymbol{v}$ be fixed as $\boldsymbol{v}^{(0)}$ and we only learn W for the ease of theoretical analysis. Thus, we represent $f(z; , \boldsymbol{v}, W)$ by $f(z; W)$ to simplify the notation. This neural network model is widely studied in recent papers on the analysis of neural networks, e.g., Gao et al. [2019], Bai and Lee [2019]. When the model is overparameterized, i.e., m is sufficiently large, we can characterized the dynamics of the training such neural network by neural tangent kernel (NTK) [Jacot et al., 2018]. Here we define

$$\varphi(z; W) := [\nabla_{W_1} f(z; W)^\top, \cdots, \nabla_{W_{2m}} f(z; W)^\top]^\top, \tag{5.2}$$

where we let $\nabla_{W_i} f(z; W)$ be a column vector such that $\varphi(z; W) \in \mathbb{R}^{2md}$. Thus, conditioned on the randomness in the initialization of W by $W^{(0)}$, we further define the kernel

$$\ker_m(z, z') = \langle \varphi(z; W^{(0)}), \varphi(z'; W^{(0)}) \rangle, \forall z, z' \in \mathcal{Z}.$$

In addition, we consider the linearization of the model $f(z, W)$ at the initial value $W^{(0)}$, which is defined as $f_{\texttt{lin}}(z; W) := f(z; W^{(0)}) + \langle \varphi(z; W^{(0)}), W - W^{(0)} \rangle$. Moreover, $f_{\texttt{lin}}(z; W)$ can be rewritten as $f_{\texttt{lin}}(z; W) = \langle \varphi(z; W^{(0)}), W - W^{(0)} \rangle$ since $f(z; W^{(0)}) = 0$ by the initialization scheme. We can see that the linearized function $f_{\texttt{lin}}(z; W)$ is a function on RKHS with the kernel $\ker_m(z, z')$. When the model is overparameterized with $m \to \infty$, the kernel $\ker_m(z, z')$ converges to an NTK kernel, which is defined as $\ker_{\texttt{ntk}} = \mathbb{E}_{\boldsymbol{\omega} \sim N(0, I_d/d)}[\texttt{act}'(\boldsymbol{\omega}^\top z) \cdot \texttt{act}'(\boldsymbol{\omega}^\top z') \cdot z^\top z']$, where $\texttt{act}'$ is the derivative of the activation function $\texttt{act}$.

Algorithm 6 Exploration Phase for Single-Agent MDP

```
Initialize: $\delta > 0$ and $\varepsilon > 0$.
for episode $k = 1, \ldots, K$ do
    Let $V_{H+1}^k(\cdot) = \mathbf{0}$ and $Q_{H+1}^k(\cdot, \cdot) = \mathbf{0}$.
    for step $h = H, H-1, \ldots, 1$ do
        Construct bonus term $u_h^k(\cdot, \cdot)$.
        Compute exploration reward $r_h^k(\cdot, \cdot) = u_h^k(\cdot, \cdot)/H$.
        Compute approximation function $f_h^k(\cdot, \cdot)$.
        $Q_h^k(\cdot, \cdot) = \Pi_{[0,H]}[(f_h^k + r_h^k + u_h^k)(\cdot, \cdot)]$.
        $V_h^k(\cdot) = \max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$.
        $\pi_h^k(\cdot) = \mathrm{argmax}_{a \in \mathcal{A}} Q_h^k(\cdot, a)$.
    end for
    Take actions following $a_h^k \sim \pi_h^k(s_h^k), \ \forall h \in [H]$.
end for
Return: $\{(s_h^k, a_h^k)\}_{(h,k) \in [H] \times [K]}$.
```

5.3 Single-Agent MDP Setting

In this section, we introduce our method under the single-agent MDP setting with kernel and neural function approximations. Then, we present our theoretical results.

5.3.1 Kernel Function Approximation

Our proposed method is composed of the reward-free exploration phase and planning phase with the given extrinsic reward function. The exploration phase and planning phase are summarized in Algorithm 6 and Algorithm 7.

Specifically, the exploration algorithm is an optimistic variant of the value-iteration algorithm with the function approximation. In Algorithm 6, we use Q_h^k and V_h^k to denote the optimistic Q-function and value function for the exploration rewards. During the exploration phase, the agent does not access the true reward function and explore the environment for K episodes based on the policy $\{\pi_h^k\}_{(h,k) \in [H] \times [K]}$ determined by the value function V_h^k, and collects the trajectories $\{s_h^k, a_h^k\}_{(h,k) \in [H] \times [K]}$ for the subsequent planning phase. Thus, instead of approximating the Q-function directly, we seek to approximate $\mathbb{P}_h V_{h+1}^k$ by a clipped function $f_h^k(s, a)$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, where $f_h^k(\cdot, \cdot)$ is estimated by solving a regularized kernel regression problem as below. Based on this kernel approximation, we construct an associated Upper Confidence Bound (UCB) bonus term u_h^k to facilitate exploration, whose form is specified by the kernel function approximator. Moreover, although the true reward is not available to the agent, to guide the exploration, we construct the exploration reward by scaling the bonus u_h^k, guiding the agent to explore state-action pairs with high uncertainties characterized by u_h^k. Then, the Q-function Q_h^k is a com-

Algorithm 7 Planning Phase for Single-Agent MDP

```
Initialize: Reward function $\{r_h\}_{h\in[H]}$ and exploration data $\{(s_h^k, a_h^k)\}_{(h,k)\in[H]\times[K]}$.
for step $h = H, H-1, \ldots, 1$ do
    Compute bonus term $u_h(\cdot,\cdot)$.
    Compute approximation function $f_h(\cdot,\cdot)$.
    $Q_h(\cdot,\cdot) = \Pi_{[0,H]}[(f_h + r_h + u_h)(\cdot,\cdot)]$.
    $V_h(\cdot) = \max_{a\in\mathcal{A}} Q_h(\cdot, a)$.
    $\pi_h(\cdot) = \mathrm{argmax}_{a\in\mathcal{A}} Q_h(\cdot, a)$.
end for
Return: $\{\pi_h\}_{h\in[H]}$.
```

bination of $r_h^k(s,a)$, $f_h^k(s,a)$, and $u_h^k(s,a)$ as shown in Line 5 of Algorithm 6. In this chapter, we define a clipping operator as $\Pi_{[0,H]}[x] := \min\{x, H\}^+ = \min\{\max\{x, 0\}, H\}$. Note that the exploration phase in Algorithm 6 is not restricted to the kernel case and can be combined with other approximators, e.g., neural networks, as will be shown later.

At the k-th episode, given the visited trajectories $\{(s_h^\tau, a_h^\tau)\}_{\tau=1}^{k-1}$, we construct the approximator for each $h \in [H]$ by solving the following regularized kernel regression problem

$$\widehat{f}_h^k = \min_{f\in\mathcal{H}} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(z_h^\tau)]^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

where $f(z_h^\tau) = \langle f, \phi(z_h^\tau)\rangle_{\mathcal{H}}$ with $z_h^\tau = (s_h^\tau, a_h^\tau)$, and λ is a hyperparameter to be determined later. As we will discuss in Lemma 5.10, the closed form solution to the above problem is $\widehat{f}_h^k(z) = \langle \widehat{f}_h^k, \phi(z)\rangle_{\mathcal{H}} = \psi_h^k(z)^\top(\lambda\cdot I + \mathcal{K}_h^k)^{-1}\mathbf{y}_h^k$, where we define $\psi_h^k(z) := [\mathrm{ker}(z, z_h^1), \cdots, \mathrm{ker}(z, z_h^{k-1})]^\top$, $\mathbf{y}_h^k := [V_{h+1}^k(s_{h+1}^1), \cdots, V_{h+1}^k(s_{h+1}^{k-1})]^\top$, and also $\mathcal{K}_h^k := [\psi_h^k(z_h^1), \cdots, \psi_h^k(z_h^{k-1})]$ (recalling that $z = (s,a)$).

We let $f_h^k(z) = \Pi_{[0,H]}[\widehat{f}_h^k(z)]$ by clipping operation to guarantee $f_h^k(z) \in [0,H]$ such that in Algorithm 6, we let

$$f_h^k(z) = \Pi_{[0,H]}[\psi_h^k(z)^\top(\lambda\cdot I + \mathcal{K}_h^k)^{-1}\mathbf{y}_h^k], \tag{5.3}$$

In addition, the associated bonus term is defined as

$$u_h^k(z) := \min\{\beta \cdot w_h^k(z), H\} \tag{5.4}$$

where β is a hyperparameter to be determined and we set

$$w_h^k(z) = \lambda^{-\frac{1}{2}}[\mathrm{ker}(z,z) - \psi_h^k(z)^\top(\lambda I + \mathcal{K}_h^k)^{-1}\psi_h^k(z)]^{\frac{1}{2}}.$$

The planning phase can be viewed as a single-episode version of optimistic value iteration algorithm. Using all the collected trajectories $\{s_h^k, a_h^k\}_{(h,k)\in[H]\times[K]}$, we can similarly construct the approximation of $\mathbb{P}_h V_{h+1}$ by solving

$$\widehat{f}_h = \operatorname*{argmin}_{f\in\mathcal{H}} \sum_{\tau=1}^{K} [V_{h+1}(s_{h+1}^\tau) - f(z_h^\tau)]^2 + \lambda\|f\|_{\mathcal{H}}^2. \tag{5.5}$$

Thus, the kernel approximation function can be estimated as

$$f_h(z) = \Pi_{[0,H]}[\widehat{f}_h(z)] = \Pi_{[0,H]}[\psi_h(z)^\top(\lambda\cdot I + \mathcal{K}_h)^{-1}\mathbf{y}_h],$$

and the bonus term is

$$u_h(z) := \min\{\beta\cdot w_h(z), H\}$$

with setting

$$w_h(z) = \lambda^{-\frac{1}{2}}[\ker(z,z) - \psi_h(z)^\top(\lambda I + \mathcal{K}_h)^{-1}\psi_h(z)]^{\frac{1}{2}},$$

where we define $\psi_h(z) := [\ker(z, z_h^1), \cdots, \ker(z, z_h^K)]^\top$, $\mathbf{y}_h := [V_{h+1}(s_{h+1}^1), \cdots, V_{h+1}(s_{h+1}^K)]^\top$, and also $\mathcal{K}_h := [\psi_h(z_h^1), \cdots, \psi_h(z_h^K)]$. Given an arbitrary reward function r_h, with the kernel approximator f_h and the bonus u_h, one can compute the optimistic Q-function Q_h and the associated value function V_h. The learned policy π_h is obtained by value iteration based on the optimistic Q-function. Algorithm 7 is also a general planning scheme that can be generalized to other function approximator, for example, the neural function approximator.

Remark 5.1. Note that in the kernel function approximation setting, we directly define the kernel $\ker(z, z')$ for the algorithms instead of the feature map $\phi(z)$ which potential lies in an infinite dimensional space.

5.3.2 Neural Function Approximation

For the neural function approximation setting, the agent also runs Algorithm 6 for exploration and Algorithm 7 for planning. Different from the kernel function approximation, in the exploration phase, at the k-th episode, given the visitation history $\{s_h^\tau, a_h^\tau\}_{\tau=1}^{k-1}$, we construct the approximation for each $h \in [H]$ by solving the following regularized regression problem

$$W_h^k = \operatorname*{argmin}_{W\in\mathbb{R}^{2md}} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(z_h^\tau; W)]^2 + \lambda\|W - W^{(0)}\|_2^2, \tag{5.6}$$

where we assume that there exists an optimization oracle that can return the global optimizer of the above problem. The initialization of $W^{(0)}$ and $\boldsymbol{v}^{(0)}$ for the function $f(z; W)$ follows the scheme as we discussed in Section 5.2.3. As shown in many recent works [Du et al., 2019, 2018, Arora et al., 2019], when m is sufficiently large, with random initialization, some common optimizers, e.g., gradient descent, can find the global minimizer of the empirical loss efficiently with a linear convergence rate. Once we obtain W_h^k, the approximation function is constructed as $f_h^k(z) = \Pi_{[0,H]}[f(z; W_h^k)]$. The related exploration bonus u_h^k is of the form $u_h^k(z) := \min\{\beta \cdot w_h^k(z), H\}$ where

$$w_h^k(z) = [\varphi(z; W_h^k)^\top (\Lambda_h^k)^{-1} \varphi(z; W_h^k)]^{\frac{1}{2}}. \tag{5.7}$$

Here we define the invertible matrix $\Lambda_h^k := \lambda I_{2md} + \sum_{\tau=1}^{k-1} \varphi(z_h^\tau; W_h^k)\varphi(z_h^\tau; W_h^k)^\top$with $\varphi(z_h^\tau; W)$ as (5.2).

In the planning phase, given the collection of trajectories in K episodes of exploration phase, we construct the neural approximation of $\mathbb{P}_h V_{h+1}(z)$ as solving a least square problem, i.e., W_h is the global optimizer of

$$\min_{W \in \mathbb{R}^{2md}} \sum_{\tau=1}^{K} [V_{h+1}(s_{h+1}^\tau) - f(z_h^\tau; W)]^2 + \lambda \|W - W^{(0)}\|_2^2,$$

such that $f_h(z) = \Pi_{[0,H]}[f(z; W_h)]$. Analogously, the bonus term for the planning phase is of the form $u_h(z) := \min\{\beta \cdot w_h(z), H\}$ where

$$w_h(z) = [\varphi(z; W_h)^\top (\Lambda_h)^{-1} \varphi(z; W_h)]^{\frac{1}{2}},$$

where we define the invertible matrix $\Lambda_h := \lambda I_{2md} + \sum_{\tau=1}^{K} \varphi(z_h^\tau; W_h)\varphi(z_h^\tau; W_h)^\top$.

5.3.3 Main Results for Single-Agent MDP

Kernel Function Approximation. In this subsection, we first present the result for the kernel function approximation setting. We make the following assumptions.

Assumption 5.2. *For any value function $V : \mathcal{S} \mapsto \mathbb{R}$, we assume that $\mathbb{P}_h V(z)$ is in a form of $\langle \phi(z), \mathbf{w}_h \rangle_{\mathcal{H}}$ for some $\mathbf{w}_h \in \mathcal{H}$. In addition, we assume there exists a fixed constant R_Q such that $\|\mathbf{w}_h\|_{\mathcal{H}} \leq R_Q H$.*

One example for this assumption is that the transition model is in a form of $\mathbb{P}_h(s'|z) = \langle \phi(z), \mathbf{w}'_h(s') \rangle_{\mathcal{H}}$ such that $\mathbb{P}_h V(z) = \int_{\mathcal{S}} V_{h+1}(s') \langle \phi(z), \mathbf{w}'_h(s') \rangle_{\mathcal{H}} \mathrm{d}s'$ where we can write $\mathbf{w}_h =$

$\int_{\mathcal{S}} V_{h+1}(s')\mathbf{w}'_h(s')\mathrm{d}s'$. This example can be viewed as a generalization of the linear transition model [Jin et al., 2020b] to the RKHS.

In this chapter, we use maximal information gain [Srinivas et al., 2009] to measure the function space complexity, i.e.,

$$\Gamma(\mathfrak{C}, \sigma; \mathrm{ker}) = \sup_{\mathcal{D}\subseteq\mathcal{Z}} 1/2 \cdot \log\det(I + \mathcal{K}_{\mathcal{D}}/\sigma),$$

where the supremum is taken over all possible sample sets $\mathcal{D} \subseteq \mathcal{Z}$ with $|\mathcal{D}| \leq \mathfrak{C}$, and $\mathcal{K}_{\mathcal{D}}$ is the Gram matrix induced by $\mathcal{D}$ based on some kernel ker of RKHS. The value of $\Gamma(\mathfrak{C}, \sigma; \mathrm{ker})$ reflects how fast the the eigenvalues of $\mathcal{H}$ decay to zero and can be viewed as a proxy of the dimension of $\mathcal{H}$ when $\mathcal{H}$ is infinite-dimensional. To characterize the complexity, we define a Q-function class $\overline{\mathcal{Q}}$ of the form

$$\overline{\mathcal{Q}}(c, R, B) = \{Q : Q \text{ satisfies the form of } Q^\sharp\}. \tag{5.8}$$

where we define $Q^\sharp$ in the following form $Q^\sharp(z) = \min\{c(z) + \Pi_{[0,H]}[\langle \mathbf{w}, \phi(z)\rangle_{\mathcal{H}}] + g(z), H\}^+$ with some $\mathbf{w}$ satisfying $\|\mathbf{w}\|_{\mathcal{H}} \leq R$, $\|\phi(z)\|_{\mathcal{H}} \leq 1$, and also $g(z) = B \cdot \min\{\|\phi(z)\|_{\Lambda_{\mathcal{D}}^{-1}}, H/\beta\}^+$. Here $\Lambda_{\mathcal{D}}$ is an adjoint operator with the form $\Lambda_{\mathcal{D}} = \lambda I_{\mathcal{H}} + \sum_{z'\in\mathcal{D}} \phi(z')\phi(z')^\top$ with $I_{\mathcal{H}}$ denoting identity mapping on $\mathcal{H}$ and $\mathcal{D} \subseteq \mathcal{Z}$ with $|\mathcal{D}| \leq K$. Here we define the ς-covering number of the class $\overline{\mathcal{Q}}$ w.r.t. the ℓ_∞-norm as $\overline{\mathcal{N}}_\infty(\varsigma; R, B)$ with an upper bound $\mathcal{N}_\infty(\varsigma; R, B)$. As formally discussed in Section 5.7, we compute the covering number upper bound $\mathcal{N}_\infty(\varsigma; R, B)$. As we can see in Algorithms 6 and 7, we have $Q_h^k \in \overline{\mathcal{Q}}(\mathbf{0}, R, (1 + 1/H)\beta)$ and $Q_h \in \overline{\mathcal{Q}}(r_h, R', \beta)$ for some R and R'. Based on the above assumptions and definitions, we have the following result.

Theorem 5.3. *Suppose that β satisfies the condition that $16H^2\big[R_Q^2 + \log\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 2\Gamma(K, \lambda; \mathrm{ker}) + 6\log(2KH) + 5\big] \leq \beta^2$. Under the kernel function approximation setting with a kernel ker, letting $\lambda = 1 + 1/K$, $R_K = 2H\sqrt{\Gamma(K, \lambda; \mathrm{ker})}$, and $\varsigma^* = H/K$, with probability at least $1 - (2K^2H^2)^{-1}$, the policy generated via Algorithm 7 satisfies $V_1^*(s_1, r) - V_1^\pi(s_1, r) \leq \mathcal{O}(\beta\sqrt{H^4[\Gamma(K, \lambda; \mathrm{ker}) + \log(KH)]}/\sqrt{K})$, after exploration for K episodes with Algorithm 6.*

The covering number $\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta)$ and the information gain $\Gamma(K, \lambda; \mathrm{ker})$ reflect the function class complexity. To understand the result in Theorem 5.3, we consider kernels ker with two different types of eigenvalue decay conditions: (i) γ-finite spectrum and (ii) γ-exponential spectral decay.

For the case of γ-finite spectrum with $\gamma \in \mathbb{Z}_+$, we have $\beta = \mathcal{O}(\gamma H\sqrt{\log(\gamma KH)})$, $\log\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}(\gamma^2\log(\gamma KH))$, and $\Gamma(K, \lambda; \mathrm{ker}) = \mathcal{O}(\gamma\log K)$, which further implies that to achieve $V_1^*(s_1, r) - V_1^\pi(s_1, r) \leq \varepsilon$, it requires $\widetilde{\mathcal{O}}(H^6\gamma^3/\varepsilon^2)$ rounds of exploration, where $\widetilde{\mathcal{O}}$ hides the logarithmic dependence on γ and $1/\varepsilon$.

Therefore, when the problem reduces to the setting of linear function approximation, the above result becomes $\widetilde{\mathcal{O}}(H^6\mathfrak{d}^3/\varepsilon^2)$ by letting $\gamma = \mathfrak{d}$, where $\mathfrak{d}$ is the feature dimension. This is consistent with the result in Wang et al. [2020a], which studies the linear approximation setting for reward-free RL. Furthermore, the sample complexity becomes $\widetilde{\mathcal{O}}(H^6|\mathcal{S}|^3|\mathcal{A}|^3/\varepsilon^2)$ by setting $\gamma = |\mathcal{S}||\mathcal{A}|$, when the problem reduces to the tabular setting.

For the case of γ-exponential spectral decay with $\gamma > 0$, we have $\log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}((\log K)^{1+2/\gamma} + (\log\log H)^{1+2/\gamma})$, $\beta = \mathcal{O}(H\sqrt{\log(KH)}(\log K)^{1/\gamma})$, and also $\Gamma(K, \lambda; \ker) = \mathcal{O}((\log K)^{1+1/\gamma})$. Therefore, to obtain an ε-suboptimal policy, it requires $\mathcal{O}(H^6 C_\gamma \cdot \log^{4+6/\gamma}(\varepsilon^{-1})/\varepsilon^2) = \widetilde{\mathcal{O}}(H^6 C_\gamma/\varepsilon^2)$ rounds of exploration, where C_γ is some constant depending on $1/\gamma$. Please see Section 5.7 for detailed definitions and discussions.

Neural Function Approximation. Next, we present the result for the neural function approximation setting.

Assumption 5.4. *For any value function V, we assume that $\mathbb{P}_h V(z)$ can be represented as $\mathbb{P}_h V(z) = \int_{\mathbb{R}^d} \texttt{act}'(\boldsymbol{\omega}^\top z) \cdot z^\top \boldsymbol{\alpha}_h(\boldsymbol{\omega}) \mathrm{d}p_0(\boldsymbol{\omega})$ for some $\boldsymbol{\alpha}_h(\boldsymbol{\omega})$ with $\boldsymbol{\alpha} : \mathbb{R}^d \mapsto \mathbb{R}^d$ and $\sup_{\boldsymbol{\omega}} \|\boldsymbol{\alpha}(\boldsymbol{\omega})\| \leq R_Q H/\sqrt{d}$. Here p_0 is the density of Gaussian distribution $N(0, I_d/d)$.*

As discussed in Gao et al. [2019], Yang et al. [2020], the function class characterized by $f(z) = \int_{\mathbb{R}^d} \texttt{act}'(\boldsymbol{\omega}^\top z) \cdot z^\top \boldsymbol{\alpha}_h(\boldsymbol{\omega}) \mathrm{d}p_0(\boldsymbol{\omega})$ is an expressive subset of RKHS. One example is that the transition model can be written as $\mathbb{P}_h(s'|z) = \int_{\mathbb{R}^d} \texttt{act}'(\boldsymbol{\omega}^\top z) \cdot z^\top \boldsymbol{\alpha}'_h(\boldsymbol{\omega}; s') \mathrm{d}p_0(\boldsymbol{\omega})$ such that we have $\boldsymbol{\alpha}_h(\boldsymbol{\omega}) = \int_{\mathcal{S}} \boldsymbol{\alpha}'_h(\boldsymbol{\omega}; s') V_{h+1}(s') \mathrm{d}s'$. This example also generalizes the linear transition model [Jin et al., 2020b] to the overparameterized neural network setting. Similar to (5.8), we also define a Q-function class based on a normalized version of $\varphi(z, W^{(0)})$, which further can be analyzed using the same notations $\overline{\mathcal{Q}}$ and $\mathcal{N}_\infty$ (See Lemma 5.19 for details).

Theorem 5.5. *Suppose that β satisfies the condition that $8H^2[R_Q^2(1+\sqrt{\lambda/d})^2 + 4\Gamma(K, \lambda; \ker_m) + 10 + 4\log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 12\log(2KH)] \leq \beta^2$ with $m = \Omega(K^{19}H^{14}\log^3 m)$. Under the overparameterized neural function approximation setting, letting $\lambda = C(1 + 1/K)$ for some constant $C \geq 1$, $R_K = H\sqrt{K}$, and $\varsigma^* = H/K$, with probability at least $1 - (2K^2H^2)^{-1} - 4m^{-2}$, the policy generated via Algorithm 7 satisfies $V_1^*(s_1, r) - V_1^\pi(s_1, r) \leq \mathcal{O}(\beta\sqrt{H^4[\Gamma(K, \lambda; \ker_m) + \log(KH)]}/\sqrt{K} + H^2\beta\iota)$ with $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$, after exploration for K episodes with Algorithm 6.*

In Theorem 5.5, there is an error term $H^2\beta\iota$ that depends on $m^{-1/12}$. In the regime of overparameterization, when m is sufficiently large, this term can be extremely small and $\iota \to 0$, $\ker_m \to \ker_{\texttt{ntk}}$ if $m \to \infty$. Here $\Gamma(K, \lambda; \ker_m)$ and $\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta)$ characterize the intrinsic complexity of the function class. In particular, when m is large, the overparamterized neural function setting can be viewed as a special case of RKHS with a misspecification error. If the eigenvalues of the kernel $\ker_m$ satisfy finite spectrum or exponential spectral decay, we know that β, $\Gamma(K, \lambda; \ker_m)$, and $\log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta)$ are of the same orders to the ones in the discussion after Theorem 5.3. Moreover, if m is sufficiently large such that $H^2\beta\iota \leq \varepsilon$, we obtain an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity to achieve an $\mathcal{O}(\varepsilon)$-suboptimal policy.

Overall, the above results show that with the kernel function approximation and overparameterized neural function approximation, Algorithms 8 and 9 guarantee $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity for achieving ε-suboptimal policy, which matches existing $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ results for the single-agent MDP for the tabular case or with linear function approximation in terms of ε.

5.4 Markov Game Setting

In this section, we introduce the algorithms under the Markov game setting with kernel and neural function approximations. We further present their theoretical results on the sample complexity.

5.4.1 Kernel Function Approximation

The exploration phase and planning phase for the zero-sum game are summarized in Algorithm 8 and Algorithm 9.

Specifically, in the exploration phase, the exploration policy for both players is obtained by taking maximum on Q-function over both action spaces. Thus, Algorithm 8 in essence is an extension of Algorithm 6 and performs the same exploration steps, if we view the pair (a, b) as an action $\boldsymbol{a} = (a, b)$ on the action space $\mathcal{A} \times \mathcal{B}$ and regard the exploration policy pair $(\pi_h^k(s), \nu_h^k(s))$ as a product policy $(\pi_h^k \otimes \nu_h^k)(s)$. Thus, the approximator $f_h^k(z)$ and the bonus term $u_h^k(z)$ share the same forms as (5.3) and (5.4) if we slightly abuse the notation by letting $z = (s, a, b)$.

In the planning phase, the algorithm generates the policies for two players in a separate manner. While maintaining two optimistic Q-functions, their policies are generated by finding NE of two games with payoff matrices $\overline{Q}$ and $\underline{Q}$ respectively, namely $(\pi_h(\cdot|s), \overline{D}_0(\cdot|s))$ is the solution to $\max_{\pi'} \min_{\nu'} \mathbb{E}_{a\sim\pi', b\sim\nu'}[\overline{Q}_h(s, a, b)]$ and $(\underline{D}_0(\cdot|s), \nu_h(\cdot|s))$ is the solution to $\max_{\pi'} \min_{\nu'} \mathbb{E}_{a\sim\pi', b\sim\nu'}[\underline{Q}_h(s, a, b)]$, which can be solved efficiently in computation by many existing algorithms (e.g., Koller et al. [1994]).

Moreover, we construct the approximation functions for Player 1 and Player 2 similarly via (5.5) by letting $z = (s, a, b)$ and placing the value function with $\overline{V}$ and $\underline{V}$ separately such that we have

$$\overline{f}_h(z) = \Pi_{[0,H]}[\psi_h(z)^\top(\lambda \cdot I + \mathcal{K}_h)^{-1}\overline{\mathbf{y}}_h],$$
$$\underline{f}_h(z) = \Pi_{[0,H]}[\psi_h(z)^\top(\lambda \cdot I + \mathcal{K}_h)^{-1}\underline{\mathbf{y}}_h],$$

Algorithm 8 Exploration Phase for Zero-Sum Markov Game

```
Initialize: $\delta > 0$ and $\varepsilon > 0$.
for episode $k = 1, \dots, K$ do
    Let $V_{H+1}^k(\cdot) = \mathbf{0}$ and $Q_{H+1}^k(\cdot,\cdot,\cdot) = \mathbf{0}$.
    for step $h = H, H-1, \dots, 1$ do
        Construct bonus term $u_h^k(\cdot,\cdot,\cdot)$.
        Exploration reward $r_h^k(\cdot,\cdot,\cdot) = u_h^k(\cdot,\cdot,\cdot)/H$.
        Compute approximation function $f_h^k(\cdot,\cdot,\cdot)$.
        $Q_h^k(\cdot,\cdot,\cdot) = \Pi_{[0,H]}[(f_h^k + r_h^k + u_h^k)(\cdot,\cdot,\cdot)]$.
        $V_h^k(\cdot) = \max_{a\in\mathcal{A},b\in\mathcal{B}} Q_h^k(\cdot,a,b)$.
        $(\pi_h^k(\cdot), \nu_h^k(\cdot)) = \mathrm{argmax}_{a\in\mathcal{A},b\in\mathcal{B}} Q_h^k(\cdot,a,b)$.
    end for
    Take actions following $a_h^k \sim \pi_h^k(s_h^k)$ and also $b_h^k \sim \nu_h^k(s_h^k), \forall h \in [H]$ .
end for
Return: $\{(s_h^k, a_h^k, u_h^k)\}_{(h,k)\in[H]\times[K]}$.
```

where $\overline{\mathbf{y}}_h := [\overline{V}_{h+1}(s_{h+1}^1), \cdots, \overline{V}_{h+1}(s_{h+1}^K)]^\top$ and $\underline{\mathbf{y}}_h := [\underline{V}_{h+1}(s_{h+1}^1), \cdots, \underline{V}_{h+1}(s_{h+1}^K)]^\top$. Then, for the bonus term, Players 1 and 2 share the one of the same form, i.e., $\overline{u}_h(z) = \underline{u}_h(z) := u_h(z) = \min\{\beta \cdot w_h(z), H\}$ with

$$w_h(z) = \lambda^{-\frac{1}{2}}[\mathrm{ker}(z,z) - \psi_h(z)^\top(\lambda I + \mathcal{K}_h)^{-1}\psi_h(z)]^{\frac{1}{2}}.$$

5.4.2 Neural Function Approximation

For the neural function approximation, the exploration and planning phases follow Algorithm 8 and 9. In the exploration phase, following the same discussion for the exploration algorithm with kernel function approximation, Algorithm 8 with the neural approximator is intrinsically the same as Algorithm 6. Thus, one can follow the same approaches to construct the neural function approximator $f_h^k(z) = \Pi_{[0,H]}[f(z;W_h^k)]$ and the bonus $u_h^k(z)$ as in (5.6) and (5.7) with only letting $z = (s,a,b)$.

For the planning phase, letting $z = (s,a,b)$, we construct approximation functions separately for Player 1 and Player 2 via solving two regression problems

$$\overline{W}_h = \underset{W\in\mathbb{R}^{2md}}{\mathrm{argmin}} \sum_{\tau=1}^K [\overline{V}_{h+1}(s_{h+1}^\tau) - f(z_h^\tau;W)]^2 + \lambda\|W - W^{(0)}\|_2^2,$$

$$\underline{W}_h = \underset{W\in\mathbb{R}^{2md}}{\mathrm{argmin}} \sum_{\tau=1}^K [\underline{V}_{h+1}(s_{h+1}^\tau) - f(z_h^\tau;W)]^2 + \lambda\|W - W^{(0)}\|_2^2,$$

such that we let $\overline{f}_h(z) = \Pi_{[0,H]}[f(z;\overline{W}_h)]$ and $\underline{f}_h(z) = \Pi_{[0,H]}[f(z;\underline{W}_h)]$. The bonus terms $\overline{u}_h$ and

Algorithm 9 Planning Phase for Zero-Sum Markov Game

```
**Initialize:** Reward function $\{r_h\}_{h\in[H]}$ and exploration data $\{(s_h^k, a_h^k, u_h^k)\}_{(h,k)\in[H]\times[K]}$.
**for** step $h = H, H-1, \ldots, 1$ **do**
    Compute bonus term $\overline{u}_h(\cdot,\cdot,\cdot)$ and $\underline{u}_h(\cdot,\cdot,\cdot)$.
    Compute approximations $\overline{f}_h(\cdot,\cdot,\cdot)$ and$\underline{f}_h(\cdot,\cdot,\cdot)$.
    $\overline{Q}_h(\cdot,\cdot,\cdot) = \Pi_{[0,H]}[(\overline{f}_h + r_h + \overline{u}_h)(\cdot,\cdot,\cdot)]$.
    $\underline{Q}_h(\cdot,\cdot,\cdot) = \Pi_{[0,H]}[(\underline{f}_h + r_h - \underline{u}_h)(\cdot,\cdot,\cdot)]$.
    Let $(\pi_h(\cdot|s), \overline{D}_0(\cdot|s))$ be NE for $\overline{Q}_h(s,\cdot,\cdot)$, $\forall s \in \mathcal{S}$.
    Let $(\underline{D}_0(\cdot|s), \nu_h(\cdot|s))$ be NE for $\underline{Q}_h(s,\cdot,\cdot)$, $\forall s \in \mathcal{S}$.
    $\overline{V}_h(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s), b\sim\overline{D}_0(\cdot|s)}[\overline{Q}_h(s,a,b)]$, $\forall s \in \mathcal{S}$.
    $\underline{V}_h(s) = \mathbb{E}_{a\sim\underline{D}_0(\cdot|s), b\sim\nu_h(\cdot|s)}[\underline{Q}_h(s,a,b)]$, $\forall s \in \mathcal{S}$.
**end for**
**Return:** $\{\pi_h\}_{h\in[H]}$, $\{\nu_h\}_{h\in[H]}$.
```

$\underline{u}_h$ for Players 1 and 2 are $\overline{u}_h(z) := \min\{\beta \cdot \overline{w}_h(z), H\}$ and $\underline{u}_h(z) := \min\{\beta \cdot \underline{w}_h(z), H\}$ with

$$
\begin{aligned}
\overline{w}_h(z) &= [\varphi(z; \overline{W}_h)^\top (\overline{\Lambda}_h)^{-1} \varphi(z; \overline{W}_h)]^{\frac{1}{2}}, \\
\underline{w}_h(z) &= [\varphi(z; \underline{W}_h)^\top (\underline{\Lambda}_h)^{-1} \varphi(z; \underline{W}_h)]^{\frac{1}{2}},
\end{aligned}
$$

where we define the invertible matrices $\overline{\Lambda}_h := \lambda I_{2md} + \sum_{\tau=1}^K \varphi(z_h^\tau; \overline{W}_h)\varphi(z_h^\tau; \overline{W}_h)^\top$ and $\underline{\Lambda}_h := \lambda I_{2md} + \sum_{\tau=1}^K \varphi(z_h^\tau; \underline{W}_h)\varphi(z_h^\tau; \underline{W}_h)^\top$.

5.4.3 Main Results for Markov Game

In this subsection, we present the results for the zero-sum Markov game setting. Particularly, we make the same assumptions as in Section 5.3.3 with only letting $z = (s, a, b)$. Moreover, we also use the same Q-function class $\overline{\mathcal{Q}}$ as (5.8), such that we can see in Algorithms 8 and 9, $Q_h^k \in \overline{\mathcal{Q}}(\mathbf{0}, R, (1 + 1/H)\beta)$ for some R, and $\overline{Q}_h \in \overline{\mathcal{Q}}(r_h, R', \beta)$ for some R'. To characterize the space which $\underline{Q}_h$ lies in, we define a specific Q-function class $\underline{\mathcal{Q}}$ of the form

$$
\underline{\mathcal{Q}}(c, R, B) = \{Q : Q \text{ satisfies the form of } Q^\flat\}, \tag{5.9}
$$

where $Q^\flat(z) = \min\{c(z) + \Pi_{[0,H]}[\langle \mathbf{w}, \phi(z)\rangle_{\mathcal{H}}] - g(z), H\}^+$ for some $\mathbf{w}$ satisfying $\|\mathbf{w}\|_{\mathcal{H}} \le R$ and also $g(z) = B \cdot \max\{\|\phi(z)\|_{\Lambda_{\mathcal{D}}^{-1}}, H/\beta\}^+$. Thus, we have $\underline{Q}_h \in \underline{\mathcal{Q}}(r_h, R', \beta)$. As we show in Section 5.7, $\overline{\mathcal{Q}}(c, R, B)$ and $\underline{\mathcal{Q}}(c, R, B)$ have the same covering number upper bound w.r.t $\|\cdot\|_\infty$. Then, we can use the same notation $\mathcal{N}_\infty$ to denote such upper bound. Thus, we have the following result for kernel approximation.

Theorem 5.6. *Suppose that β satisfies the condition that* $16H^2\big[R_Q^2 + \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) +$

$2\Gamma(K,\lambda;\mathrm{ker}) + 6\log(4KH) + 5\big] \leq \beta^2$. *Under the kernel function approximation setting with a kernel* ker*, letting* $\lambda = 1 + 1/K$*,* $R_K = 2H\sqrt{\Gamma(K,\lambda;\mathrm{ker})}$*, and* $\varsigma^* = H/K$*, with probability at least* $1 - (2K^2H^2)^{-1}$*, the policy pair generated via Algorithm 9 satisfies* $V_1^{\mathrm{br}(\nu),\nu}(s_1,r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1,r) \leq \mathcal{O}(\beta\sqrt{H^4[\Gamma(K,\lambda;\mathrm{ker}) + \log(KH)]}/\sqrt{K})$*, after exploration for* K *episodes with Algorithm 8.*

We further obtain the result for the neural function approximation scenario.

Theorem 5.7. *Suppose that* β *satisfies the condition that* $8H^2[10 + 12\log(4K/\delta) + R_Q^2(1 + \sqrt{\lambda/d})^2 + 4\log\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 4\Gamma(K,\lambda;\mathrm{ker}_m)] \leq \beta^2$ *with* $m = \Omega(K^{19}H^{14}\log^3 m)$*. Under the overparameterized neural function approximation setting, letting* $\lambda = C(1 + 1/K)$ *for some constant* $C \geq 1$*,* $R_K = H\sqrt{K}$*, and* $\varsigma^* = H/K$*, with probability at least* $1 - (2K^2H^2)^{-1} - 4m^{-2}$*, the policy pair generated via Algorithm 9 satisfies* $V_1^{\mathrm{br}(\nu),\nu}(s_1,r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1,r) \leq \mathcal{O}(\beta\sqrt{H^4[\Gamma(K,\lambda;\mathrm{ker}_m) + \log(KH)]}/\sqrt{K} + H^2\beta\iota)$ *with* $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$*, after exploration for* K *episodes with Algorithm 8.*

Following the same discussion as in Section 5.3.3, the above results show that with the kernel function approximation and overparameterized neural function approximation, Algorithms 8 and 9 guarantee an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity to achieve an ε-approximate NE. In particular, when our problem reduces to the Markov game with linear function approximation, the algorithm requires $\widetilde{\mathcal{O}}(H^6\mathfrak{d}^3/\varepsilon^2)$ sample complexity to achieve an ε-approximate NE, where $\mathfrak{d}$ is the feature dimension. This also complements the result of the reward-free RL for the Markov game with the linear function approximation. For the tabular case, Bai and Jin [2020] gives an $\widetilde{\mathcal{O}}(H^5|\mathcal{S}|^2|\mathcal{A}||\mathcal{B}|)$ sample complexity and Liu et al. [2020] gives an $\widetilde{\mathcal{O}}(H^4|\mathcal{S}||\mathcal{A}||\mathcal{B}|)$ sample complexity. Our analysis gives an $\widetilde{\mathcal{O}}(H^6|\mathcal{S}|^3|\mathcal{A}|^3|\mathcal{B}|^3/\varepsilon)$ sample complexity by simply letting $\mathfrak{d} = |\mathcal{S}||\mathcal{A}||\mathcal{B}|$, which matches the existing results in terms of ε. Though the dependence on H, $|\mathcal{S}|$, $|\mathcal{A}|$, $|\mathcal{B}|$ is not as tight as existing results, our work in this chapter presents a more general analysis for the function approximation setting which is not fully studied in previous works.

5.5 Theoretical Analysis

5.5.1 Proof Sketches of Theorem 5.3 and Theorem 5.5

We first show the proof sketches for Theorem 5.3. Our goal is to bound the term $V_1^*(s_1,r) - V_1^\pi(s_1,r)$. By the optimistic updating rule in the planning phase, according to Lemma 5.16, we have $V_1^*(s_1,r) \leq V_1(s_1)$ such that $V_1^*(s_1,r) - V_1^\pi(s_1,r) \leq V_1(s_1) - V_1^\pi(s_1,r)$. Then we only need to consider to upper bound $V_1(s_1) - V_1^\pi(s_1, r)$. Further by this lemma, for any $h \in [H]$, we have

$$\begin{aligned}
&V_h(s) - V_h^\pi(s,r) \\
&\quad \le r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)) - Q_h^\pi(s, \pi_h(s), r) \\
&\quad = \mathbb{P}_h V_{h+1}(s, \pi_h(s)) - \mathbb{P}_h V_{h+1}^\pi(s, \pi_h(s), r) + 2u_h(s, \pi_h(s)).
\end{aligned} \tag{5.10}$$

where we use the fact that $Q_h^\pi(s, \pi_h(s), r) = r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}^\pi(s, \pi_h(s), r)$. Recursively applying the above inequality and also using $V_{H+1}^\pi(s, r) = V_{H+1}(s) = 0$ give

$$V_1(s_1) - V_1^\pi(s_1, r) \le \mathbb{E}_{\mathbb{P}}[\textstyle\sum_{h=1}^H 2u_h(s_h, \pi_h(s_h))|s_1] = 2H \cdot V_1^\pi(s_1, u/H).$$

Moreover, by Lemma 5.17, we build a connection between the exploration and planing phase, which is $V_1^\pi(s_1, u/H) \le K^{-1} \sum_{k=1}^K V_1^*(s_1, r^k)$. Therefore, combining the above results together, we eventually obtain

$$V_1^*(s_1, r) - V_1^\pi(s_1, r) \le 2H/K \cdot \textstyle\sum_{k=1}^K V_1^*(s_1, r^k) \le \mathcal{O}\big(\beta\sqrt{H^4[\Gamma(K, \lambda; \mathrm{ker}) + \log(KH)]}/\sqrt{K}\big),$$

where the last inequality is by Lemma 5.14 and the fact that $\beta \ge H$. This completes the proof of Theorem 5.3. Please see detailed proof in Section 5.8.2.

Next, we show the proof sketches of Theorem 5.5. By Lemma 5.22, we have $V_1^*(s_1, r) \le V_1(s_1) + H\beta\iota$ by optimism, such that $V_1^*(s_1, r) - V_1^\pi(s_1, r) \le V_1(s_1) - V_1^\pi(s_1, r) + H\beta\iota$. Note that different from the proof of Theorem 5.3, there is an extra bias term $H\beta\iota$ introduced by the neural function approximation. Further by Lemma 5.22, and using the same argument as (5.10), we have

$$V_h(s) - V_h^\pi(s, r) \le 2u_h(s, \pi_h(s)) + \beta\iota + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) - \mathbb{P}_h V_{h+1}^\pi(s, \pi_h(s), r),$$

which introducing another bias $\beta\iota$. Recursively applying the above inequality with $V_{H+1}^\pi(s, r) = V_{H+1}(s) = 0$ gives

$$V_1(s_1) - V_1^\pi(s_1, r) = 2H \cdot V_1^\pi(s_1, u/H) + H\beta\iota.$$

Thus, with Lemma 5.23 connecting the exploration and planning such that $V_1^\pi(s_1, u/H) \le K^{-1} \sum_{k=1}^K V_1^*(s_1, r^k) + 2\beta\iota$, combining all the above results eventually yields

$$\begin{aligned}
&V_1^*(s_1, r) - V_1^\pi(s_1, r) \\
&\quad \le 2H/K \cdot \textstyle\sum_{k=1}^K V_1^*(s_1, r^k) + 4H\beta\iota \\
&\quad \le \mathcal{O}(\beta\sqrt{H^4[\Gamma(K, \lambda; \mathrm{ker}_m) + \log(KH)]}/\sqrt{K} + H^2\beta\iota),
\end{aligned}$$

where the second inequality and the last inequality is by Lemma 5.20 and the fact that $\beta \geq H$. This completes the proof. Please see detailed proof in Section 5.9.2.

5.5.2 Proof Sketches of Theorem 5.6 and Theorems 5.7

In the proofs of Theorem 5.6 and Theorems 5.7 and the corresponding lemmas, to simply the notations, we let $\mathbb{E}_{a\sim\pi_h, b\sim\nu_h, s'\sim\mathbb{P}_h}$ to denote the expectation with $a \sim \pi_h(\cdot|s), b \sim \nu_h(\cdot|s), s' \sim \mathbb{P}_h(\cdot|s, a, b)$ given the current state s and arbitrary policies π_h, ν_h at the h-th step.

For the proof sketch of Theorem 5.6, we decompose $V_1^{\mathrm{br}(\nu),\nu}(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r)$ into two terms $V_1^\dagger(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r)$ and $V_1^{\mathrm{br}(\nu),\nu}(s_1, r) - V_1^\dagger(s_1, r)$ and bound them separately. To bound the first term, by Lemma 5.27, we have $V_1^\dagger(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) \leq \overline{V}_1(s_1) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r)$. Note that by the updating rule for $\overline{V}_h$ in Algorithm 9, we have

$$\overline{V}_h(s) = \min_{\nu'} \mathbb{E}_{a\sim\pi_h, b\sim\nu'}[\overline{Q}_h(s, a, b)] \leq \mathbb{E}_{a\sim\pi_h, b\sim\mathrm{br}(\pi)_h}[\overline{Q}_h(s, a, b)],$$

such that further by Lemma 5.27, there is

$$\begin{aligned}
&\overline{V}_h(s_h) - V_h^{\pi,\mathrm{br}(\pi)}(s_h, r) \\
&\qquad \leq \mathbb{E}_{a_h\sim\pi_h, b_h\sim\mathrm{br}(\pi)_h}[(\mathbb{P}_h\overline{V}_{h+1} + r_h + 2u_h)(s_h, a_h, b_h)] - V_h^{\pi,\mathrm{br}(\pi)}(s_h, r) \\
&\qquad = \mathbb{E}_{a_h\sim\pi_h, b_h\sim\mathrm{br}(\pi)_h, s_{h+1}\sim\mathbb{P}_h}[\overline{V}_{h+1}(s_{h+1}) - V_{h+1}^{\pi,\mathrm{br}(\pi)}(s_{h+1}, r) + 2u_h(s_h, a_h, b_h)].
\end{aligned}$$

where the equality uses $V_h^{\pi,\mathrm{br}(\pi)}(s_h, r) = \mathbb{E}_{a_h\sim\pi_h, b_h\sim\mathrm{br}(\pi)_h}[r_h(s_h, a_h, b_h) + \mathbb{P}_h V_{h+1}^{\pi,\mathrm{br}(\pi)}(s_h, a_h, b_h, r)]$. Recursively applying the above inequality yields

$$\begin{aligned}
&\overline{V}_1(s_1) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) \\
&\qquad \leq \mathbb{E}_{\pi,\mathrm{br}(\pi),\mathbb{P}}[\textstyle\sum_{h=1}^H 2u_h(s_h, a_h, b_h)|s_1] \\
&\qquad = 2H \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, u/H).
\end{aligned}$$

Combining the above results eventually gives

$$\begin{aligned}
V_1^\dagger(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) \leq 2H \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, u/H) &\leq \frac{2H}{K}\sum_{k=1}^K V_1^*(s_1, r^k) \\
&\leq \mathcal{O}(\beta\sqrt{H^4[\Gamma(K, \lambda; \mathrm{ker}) + \log(KH)]}/\sqrt{K}),
\end{aligned}$$

where the second inequality is due to Lemma 5.28 and the last inequality is by Lemma 5.25. The upper bound of the difference $V_1^\dagger(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r)$ is also $\mathcal{O}(\beta\sqrt{H^4[\Gamma(K, \lambda; \mathrm{ker}) + \log(KH)]}/\sqrt{K})$ with the similar proof idea. This completes the proof of Theorem 5.6. Please see Section 5.10.2 for details.

The proof of Theorem 5.7 follows the same argument as above. The only difference is that the neural function approximation introduces bias terms depending on ι as we discussed in the proof sketch of Theorem 5.5. Thus, the final bound is $\mathcal{O}(\beta\sqrt{H^4[\Gamma(K,\lambda;\ker_m)+\log(KH)]}/\sqrt{K}+H^2\beta\iota)$. Please see Section 5.11.2 for the detailed proof.

5.6 Conclusion

In this chapter, we study the reward-free RL algorithms with kernel and neural function approximators for both single-agent MDPs and zero-sum Markov games. We propose efficient exploration and planning algorithms incorporating the kernel and neural function approximators. We prove that our methods can achieve $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity for generating an ε-suboptimal policy or ε-approximate NE.

5.7 Discussion of Function Space Complexity

To characterize the function space complexity, we first introduce the notions for the eigenvalues of the RKHS. Define $\mathcal{L}^2(\mathcal{Z})$ as the space of square-integrable functions on $\mathcal{Z}$ w.r.t. Lebesgue measure and define $\langle\cdot,\cdot\rangle_{\mathcal{L}^2}$ as the inner product on the space $\mathcal{L}^2(\mathcal{Z})$. According to Mercer's Theorem [Steinwart and Christmann, 2008], the kernel function $\ker(z,z')$ has a spectral expansion as $\ker(z,z')=\sum_{i=1}^{\infty}\sigma_i\varrho_i(z)\varrho_i(z')$ where $\{\varrho_i\}_{i\geq 1}$ are a set of orthonormal basis on $\mathcal{L}^2(\mathcal{Z})$ and $\{\sigma_i\}_{i\geq 1}$ are positive eigenvalues. In this chapter, we consider two types of eigenvalues' properties and make the following assumptions.

Assumption 5.8. *Assume $\{\sigma_i\}_{i\geq 1}$ satisfies one of the following eigenvalue decay conditions for some constant $\gamma>0$:*

(a) γ-finite spectrum: we have $\sigma_i=0$ for all $i>\gamma$;

(b) γ-exponential spectral decay: there exist constants $C_1>0$ and $C_2>0$ such that $\sigma_i\leq C_1\exp(-C_2\cdot i^{\gamma})$ for all $i\geq 1$.

Covering Numbers. Next, we characterize the upper bound of the covering numbers of the Q-function sets $\overline{\mathcal{Q}}(c,R,B)$ and $\underline{\mathcal{Q}}(c,R,B)$. For any $Q_1,Q_2\in\overline{\mathcal{Q}}(c,R,B)$, we have

$$Q_1(z)=\min\Big\{c(z)+\Pi_{[0,H]}[\langle\mathbf{w}_1,\phi(z)\rangle]+B\cdot\max\{\|\phi(z)\|_{\Lambda_{\mathcal{D}_1}^{-1}},H/\beta\}^+,H\Big\}^+,$$
$$Q_2(z)=\min\Big\{c(z)+\Pi_{[0,H]}[\langle\mathbf{w}_2,\phi(z)\rangle]+B\cdot\max\{\|\phi(z)\|_{\Lambda_{\mathcal{D}_2}^{-1}},H/\beta\}^+,H\Big\}^+,$$

for some $\mathbf{w}_1, \mathbf{w}_2$ satisfying $\|\mathbf{w}_1\|_{\mathcal{H}} \le R$ and $\|\mathbf{w}_2\|_{\mathcal{H}} \le R$. Then, due to the fact that the truncation operator is non-expansive, we have

$$\|Q_1(\cdot) - Q_2(\cdot)\|_\infty \le \sup_z |\langle \mathbf{w}_1 - \mathbf{w}_2, \phi(z)\rangle_{\mathcal{H}}| + B \sup_z \left| \|\phi(z)\|_{\Lambda_{\mathcal{D}_1}^{-1}} - \|\phi(z)\|_{\Lambda_{\mathcal{D}_2}^{-1}} \right|.$$

The above inequality shows that it suffices to bound the covering numbers of of the RKHS norm ball of radius R and the set of functions of the form $\|\phi(z)\|_{\Lambda_{\mathcal{D}}^{-1}}$. Thus, we define the function class $\mathcal{F}_\lambda := \{\|\phi(\cdot)\|_\Upsilon : \|\Upsilon\|_{\mathrm{op}} \le 1/\lambda\}$ since $\|\Lambda_{\mathcal{D}}^{-1}\|_{\mathrm{op}} \le 1/\lambda$ according to the definition of $\Lambda_{\mathcal{D}}$. Let $\overline{\mathcal{N}}_\infty(\epsilon; R, B)$ be the ϵ-covering number of $\overline{\mathcal{Q}}$ w.r.t. $\|\cdot\|_\infty$, $\mathcal{N}_\infty(\epsilon, \mathcal{H}, R)$ be the ϵ-covering number of RKHS norm ball of radius R w.r.t. $\|\cdot\|_\infty$, and $\mathcal{N}_\infty(\epsilon, \mathcal{F}, 1/\lambda)$ be the ϵ-covering number of $\mathcal{F}_\lambda$ w.r.t. $\|\cdot\|_\infty$. Thus, we have

$$\overline{\mathcal{N}}_\infty(\epsilon; R, B) \le \mathcal{N}_\infty(\epsilon/2, \mathcal{H}, R) \cdot \mathcal{N}_\infty(\epsilon/(2B), \mathcal{F}, 1/\lambda).$$

We define the upper bound

$$\mathcal{N}_\infty(\epsilon; R, B) := \mathcal{N}_\infty(\epsilon/2, \mathcal{H}, R) \cdot \mathcal{N}_\infty(\epsilon/(2B), \mathcal{F}, 1/\lambda).$$

Then, we know

$$\log \mathcal{N}_\infty(\epsilon; R, B) = \log \mathcal{N}_\infty(\epsilon/2, \mathcal{H}, R) + \log \mathcal{N}_\infty(\epsilon/(2B), \mathcal{F}, 1/\lambda).$$

Moreover, for any $Q_1, Q_2 \in \underline{\mathcal{Q}}(c, R, B)$, we have

$$\begin{aligned} Q_1(z) &= \min\left\{c(z) + \Pi_{[0,H]}[\langle \mathbf{w}_1, \phi(z)\rangle] - B \cdot \max\{\|\phi(z)\|_{\Lambda_{\mathcal{D}_1}^{-1}}, H/\beta\}^+, H\right\}^+, \\ Q_2(z) &= \min\left\{c(z) + \Pi_{[0,H]}[\langle \mathbf{w}_2, \phi(z)\rangle] - B \cdot \max\{\|\phi(z)\|_{\Lambda_{\mathcal{D}_2}^{-1}}, H/\beta\}^+, H\right\}^+, \end{aligned}$$

which also implies

$$\|Q_1(\cdot) - Q_2(\cdot)\|_\infty \le \sup_z |\langle \mathbf{w}_1 - \mathbf{w}_2, \phi(z)\rangle_{\mathcal{H}}| + B \sup_z \left| \|\phi(z)\|_{\Lambda_{\mathcal{D}_1}^{-1}} - \|\phi(z)\|_{\Lambda_{\mathcal{D}_2}^{-1}} \right|.$$

Thus, we can bound the covering number $\underline{\mathcal{N}}_\infty(\epsilon; R, B)$ of $\underline{\mathcal{Q}}(c, R, B)$ in the same way, i.e., $\underline{\mathcal{N}}_\infty(\epsilon; R, B) \le \mathcal{N}_\infty(\epsilon; R, B)$.

According to Yang et al. [2020], we have the following covering number upper bounds

(a) γ-finite spectrum:

$$\log \mathcal{N}_\infty(\epsilon/2, \mathcal{H}, R) \leq C_3 \gamma [\log(2R/\epsilon) + C_4],$$
$$\log \mathcal{N}_\infty(\epsilon/(2B), \mathcal{F}, 1/\lambda) \leq C_5 \gamma^2 [\log(2B/\epsilon) + C_6];$$

(b) γ-exponential spectral decay:

$$\log \mathcal{N}_\infty(\epsilon/2, \mathcal{H}, R) \leq C_3 [\log(2R/\epsilon) + C_4]^{1+1/\gamma},$$
$$\log \mathcal{N}_\infty(\epsilon/(2B), \mathcal{F}, 1/\lambda) \leq C_5 [\log(2B/\epsilon) + C_6]^{1+2/\gamma}.$$

Maximal Information Gain. Here we give the definition of maximal information gain and discuss its upper bounds based on different kernels.

Definition 5.9 (Maximal Information Gain [Srinivas et al., 2009])**.** For any fixed integer $\mathfrak{C}$ and any $\sigma > 0$, we define the maximal information gain associated with the RKHS $\mathcal{H}$ as

$$\Gamma(\mathfrak{C}, \lambda; \mathrm{ker}) = \sup_{\mathcal{D} \subseteq \mathcal{Z}} \frac{1}{2} \log \det(I + \mathcal{K}_{\mathcal{D}}/\lambda),$$

where the supremum is taken over all discrete subsets of $\mathcal{Z}$ with cardinality no more than $\mathfrak{C}$, and $\mathcal{K}_{\mathcal{D}}$ is the Gram matrix induced by $\mathcal{D} \subseteq \mathcal{Z}$ based on the kernel ker.

According to Theorem 5 in Srinivas et al. [2009], we have the maximal information gain characterized as follows

(a) γ-finite spectrum:

$$\Gamma(K, \lambda; \mathrm{ker}) \leq C_7 \gamma \log K;$$

(b) γ-exponential spectral decay:

$$\Gamma(K, \lambda; \mathrm{ker}) \leq C_7 (\log K)^{1+1/\gamma}.$$

Sample Complexity. Given the above results, for the kernel approximation setting, according to the discussion in the proof of Corollary 4.4 in Yang et al. [2020], under the parameter settings in Theorem 5.3 or Theorem 5.6, we have that for γ-finite spectrum setting,

$$\beta = \mathcal{O}(\gamma H \sqrt{\log(\gamma K H)}), \quad \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}(\gamma^2 \log(\gamma K H)),$$
$$\Gamma(K, \lambda; \mathrm{ker}) = \mathcal{O}(\gamma \log K),$$

which implies after K episodes of exploration, the upper bound in Theorem 5.3 or Theorem 5.6 is

$$\mathcal{O}\left(\sqrt{H^6\gamma^3\log^2(\gamma KH)/K}\right).$$

This result further implies that to obtain an ε-suboptimal policy or ε-approximate NE, it requires $\widetilde{\mathcal{O}}(H^6\gamma^3/\varepsilon^2)$ rounds of exploration. In addition, for the γ-exponential spectral decay setting, we have

$$\beta = \mathcal{O}(H\sqrt{\log(KH)}(\log K)^{1/\gamma}), \quad \log\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}((\log K)^{1+2/\gamma} + (\log\log H)^{1+2/\gamma}),$$
$$\Gamma(K,\lambda;\mathrm{ker}) = \mathcal{O}((\log K)^{1+1/\gamma}),$$

which implies that after K episodes of exploration, the upper bound in Theorem 5.3 or Theorem 5.6 is

$$\mathcal{O}\left(\sqrt{H^6\log^{2+3/\gamma}(KH)/K}\right).$$

Then, to obtain an ε-suboptimal policy or ε-approximate NE, it requires $\mathcal{O}(H^6C_\gamma\log^{4+6/\gamma}(\varepsilon^{-1})/\varepsilon^2) = \widetilde{\mathcal{O}}(H^6C_\gamma/\varepsilon^2)$ episodes of exploration, where C_γ is some constant depending on $1/\gamma$.

The above results also hold for the neural function approximation under both single-agent MDP and Markov game setting if the kernel ker_m satisfies the γ-finite spectrum or γ-exponential spectral decay and the network width m is sufficiently large such that the error term $H^2\beta\iota \leq \varepsilon$. Then, we can similarly obtain the upper bounds in Theorems 5.5 and 5.7.

Linear and Tabular Cases. For the linear function approximation case, we have a feature map $\phi(s) \in \mathbb{R}^{\mathfrak{d}}$, where $\mathfrak{d}$ is the feature dimension. Therefore, the associated kernel can be represented as $\mathrm{ker}(s,s') = \phi(s)^\top\phi(s') = \sum_{i=1}^{\mathfrak{d}}\phi_i(s)\phi_i(s')$. Thus, we know that under the linear setting, the kernel ker has $\mathfrak{d}$-finite spectrum. Thus, letting $\gamma = \mathfrak{d}$ in the γ-finite spectrum case, we have

$$\beta = \mathcal{O}(\mathfrak{d}H\sqrt{\log(\mathfrak{d}KH)}), \quad \log\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}(\mathfrak{d}^2\log(\mathfrak{d}KH)),$$
$$\Gamma(K,\lambda;\mathrm{ker}) = \mathcal{O}(\mathfrak{d}\log K),$$

which further implies that to achieve $V_1^*(s_1,r) - V_1^\pi(s_1,r) \leq \varepsilon$, it requires $\widetilde{\mathcal{O}}(H^6\mathfrak{d}^3/\varepsilon^2)$ rounds of exploration. This is consistent with the result in Wang et al. [2020a] for the single-agent MDP. This result also hold for the Markov game setting.

For the tabular case, since $\phi(z) = \boldsymbol{e}_z$ is the canonical basis in $\mathbb{R}^{|\mathcal{Z}|}$, we have $\gamma = |\mathcal{Z}|$ for the above γ-finite spectrum case. Therefore, for the single-agent MDP setting, we have $|\mathcal{Z}| = |\mathcal{S}||\mathcal{A}|$, which implies

$$\beta = \mathcal{O}(H|\mathcal{S}||\mathcal{A}|\sqrt{\log(|\mathcal{S}||\mathcal{A}|KH)}), \quad \Gamma(K,\lambda;\mathrm{ker}) = \mathcal{O}(|\mathcal{S}||\mathcal{A}|\log K),$$
$$\log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|^2 \log(|\mathcal{S}||\mathcal{A}|KH)).$$

Then, the sample complexity becomes $\widetilde{\mathcal{O}}(H^6|\mathcal{S}|^3|\mathcal{A}|^3/\varepsilon^2)$ to obtain an ε-suboptimal policy. For the two-player Markov game setting, we have $|\mathcal{Z}| = |\mathcal{S}||\mathcal{A}||\mathcal{B}|$, which implies

$$\beta = \mathcal{O}(H|\mathcal{S}||\mathcal{A}||\mathcal{B}|\sqrt{\log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|KH)}), \quad \Gamma(K,\lambda;\mathrm{ker}) = \mathcal{O}(|\mathcal{S}||\mathcal{A}||\mathcal{B}|\log K),$$
$$\log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) = \mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|^2|\mathcal{B}|^2 \log(|\mathcal{S}||\mathcal{A}||\mathcal{B}|KH)).$$

Then, the sample complexity becomes $\widetilde{\mathcal{O}}(H^6|\mathcal{S}|^3|\mathcal{A}|^3|\mathcal{B}|^3/\varepsilon^2)$ to obtain an ε-approximate NE.

5.8 Proofs for Single-Agent MDP with Kernel Function Approximation

5.8.1 Lemmas

Lemma 5.10 (Solution of Kernel Ridge Regression)**.** *The approximation vector $\widehat{f}_h^k \in \mathcal{H}$ is obtained by solving the following kernel ridge regression problem*

$$\underset{f\in\mathcal{H}}{\mathrm{minimize}} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(z_h^\tau)\rangle_{\mathcal{H}}]^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

such that we have

$$\widehat{f}_h^k(z) = \langle \phi(z), \widehat{f}_h^k \rangle_{\mathcal{H}} = \psi_h^k(z)^\top (\lambda \cdot I + \mathcal{K}_h^k)^{-1} \mathbf{y}_h^k,$$

where we define

$$\begin{aligned}
&\psi_h^k(z) := \Phi_h^k \phi(z) = [\mathrm{ker}(z, z_h^1), \cdots, \mathrm{ker}(z, z_h^{k-1})]^\top,\\
&\Phi_h^k = [\phi(z_h^1), \phi(z_h^2), \cdots, \phi(z_h^{k-1})]^\top,\\
&\mathbf{y}_h^k = [V_{h+1}^k(s_{h+1}^1), V_{h+1}^k(s_{h+1}^2), \cdots, V_{h+1}^k(s_{h+1}^{k-1})]^\top,\\
&\mathcal{K}_h^k := \Phi_h^k (\Phi_h^k)^\top = \begin{bmatrix} \mathrm{ker}(z_h^1, z_h^1) & \dots & \mathrm{ker}(z_h^1, z_h^{k-1}) \\ \vdots & \ddots & \vdots \\ \mathrm{ker}(z_h^{k-1}, z_h^1) & \dots & \mathrm{ker}(z_h^{k-1}, z_h^{k-1}) \end{bmatrix},
\end{aligned} \tag{5.11}$$

with denoting $z = (s, a)$ *and* $z_h^\tau = (s_h^\tau, a_h^\tau)$*, and* $\ker(x, y) = \langle \phi(z), \phi(z') \rangle_{\mathcal{H}}, \forall z, z' \in \mathcal{Z} = \mathcal{S} \times \mathcal{A}$*.*

Proof. We seek to solve the following kernel ridge regression problem in the RKHS

$$\widehat{f}_h^k = \operatorname*{argmin}_{f \in \mathcal{H}} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(s_h^\tau, a_h^\tau) \rangle_{\mathcal{H}}]^2 + \lambda \|f\|_{\mathcal{H}}^2,$$

which is equivalent to

$$\widehat{f}_h^k = \operatorname*{argmin}_{f \in \mathcal{H}} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - \langle f, \phi(s_h^\tau, a_h^\tau) \rangle_{\mathcal{H}}]^2 + \lambda \langle f, f \rangle_{\mathcal{H}}.$$

By the first-order optimality condition, the above kernel ridge regression problem admits the following closed-form solution

$$\widehat{f}_h^k = (\Lambda_h^k)^{-1} (\Phi_h^k)^\top \mathbf{y}_h^k, \tag{5.12}$$

where we define

$$\Lambda_h^k = \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) \phi(s_h^\tau, a_h^\tau)^\top + \lambda \cdot I_{\mathcal{H}} = \lambda \cdot I_{\mathcal{H}} + (\Phi_h^k)^\top \Phi_h^k,$$

with $I_{\mathcal{H}}$ being the identity mapping in RKHS. Thus, by (5.12), we have

$$\langle \widehat{f}_h^k, \phi(z) \rangle_{\mathcal{H}} = \langle (\Lambda_h^k)^{-1} (\Phi_h^k)^\top \mathbf{y}_h^k, \phi(s, a) \rangle_{\mathcal{H}}, \quad \forall (z) \in \mathcal{S} \times \mathcal{A},$$

which can be further rewritten in terms of kernel $\ker$ as follows

$$\begin{aligned} \langle \widehat{f}_h^k, \phi(z) \rangle_{\mathcal{H}} &= \langle (\Lambda_h^k)^{-1} (\Phi_h^k)^\top \mathbf{y}_h^k, \phi(z) \rangle_{\mathcal{H}} \\ &= \phi(z)^\top [\lambda \cdot I_{\mathcal{H}} + (\Phi_h^k)^\top \Phi_h^k]^{-1} (\Phi_h^k)^\top \mathbf{y}_h^k \\ &= \phi(z)^\top (\Phi_h^k)^\top [\lambda \cdot I + \Phi_h^k (\Phi_h^k)^\top]^{-1} \mathbf{y}_h^k \\ &= \psi_h^k(z)^\top (\lambda \cdot I + \mathcal{K}_h^k)^{-1} \mathbf{y}_h^k. \end{aligned} \tag{5.13}$$

The third equality is by

$$(\Phi_h^k)^\top [\lambda \cdot I + \Phi_h^k (\Phi_h^k)^\top] = [\lambda \cdot I_{\mathcal{H}} + (\Phi_h^k)^\top \Phi_h^k] (\Phi_h^k)^\top,$$

such that

$$[\lambda \cdot I_{\mathcal{H}} + (\Phi_h^k)^\top \Phi_h^k]^{-1} (\Phi_h^k)^\top = (\Phi_h^k)^\top [\lambda \cdot I + \Phi_h^k (\Phi_h^k)^\top]^{-1},$$

where I is an identity matrix in $\mathbb{R}^{(k-1)\times(k-1)}$. The last equality in (5.13) is by the definitions of $\psi_h^k(z)$ and $\mathcal{K}_h^k$ in (5.11). This completes the proof. $\square$

Lemma 5.11 (Boundedness of Solution)**.** *When $\lambda \geq 1$, for any $(k, h) \in [K] \times [H]$, $\widehat{f}_h^k$ defined in* (5.12) *satisfies*

$$\|\widehat{f}_h^k\|_{\mathcal{H}} \leq H\sqrt{2/\lambda \cdot \log\det(I + \mathcal{K}_h^k/\lambda)} \leq 2H\sqrt{\Gamma(K, \lambda; \mathrm{ker})},$$

where $\mathcal{K}_h^k$ is defined in (5.11) *and* $\Gamma(K, \lambda; \mathrm{ker})$ *is defined in Definition 5.9.*

Proof. For any vector $f \in \mathcal{H}$, we have

$$\begin{aligned} |\langle f, \widehat{f}_h^k\rangle_{\mathcal{H}}| &= |f^\top (\Lambda_h^k)^{-1}(\Phi_h^k)^\top \mathbf{y}_h^k| \\ &= \left| f^\top (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau) V_{h+1}^k(s_{h+1}^\tau)\right| \leq H \sum_{\tau=1}^{k-1} \left| f^\top (\Lambda_h^k)^{-1}\phi(s_h^\tau, a_h^\tau)\right|, \end{aligned}$$

where the last inequality is due to $|V_{h+1}^k(s_{h+1}^\tau)| \leq H$. Then, with Lemma 5.36, the rest of the proof is the same as the proof of Lemma C.5 in Yang et al. [2020], which finishes the proof. $\square$

Lemma 5.12. *With probability at least $1 - \delta'$, we have $\forall (h, k) \in [H] \times [K]$,*

$$\begin{aligned} &\left\| \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau)]\right\|_{(\Lambda_h^k)^{-1}}^2 \\ &\qquad \leq 4H^2\Gamma(K, \lambda; \mathrm{ker}) + 10H^2 + 4H^2 \log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 4H^2\log(K/\delta'), \end{aligned}$$

where we set $\varsigma^* = H/K$ and $\lambda = 1 + 1/K$.

Proof. We first define a value function class as follows

$$\overline{\mathcal{V}}(\mathbf{0}, R, B) = \{V : V(\cdot) = \max_{a\in\mathcal{A}} Q(\cdot, a) \text{ with } Q \in \overline{\mathcal{Q}}(\mathbf{0}, R, B)\},$$

where $\overline{\mathcal{Q}}$ is defined in (5.8). We denote the covering number of $\overline{\mathcal{V}}(\mathbf{0}, R, B)$ w.r.t. the distance dist as $\mathcal{N}_{\mathrm{dist}}^{\overline{\mathcal{V}}}(\epsilon; R, B)$, where the distance dist is defined by $\mathrm{dist}(V_1, V_2) = \sup_{s\in\mathcal{S}} |V_1(s) - V_2(s)|$. Specifically, for any $k \times h \in [K]\times[H]$, we assume that there exist constants R_K and B_K that depend on the number of episodes K such that any $V_h^k \in \overline{\mathcal{V}}(\mathbf{0}, R_K, B_K)$ with $R_K = 2H\sqrt{\Gamma(K, \lambda; \mathrm{ker})}$ and $B_K = (1 + 1/H)\beta$ since $Q_h^k(z) = \Pi_{[0,H]}[(r_h^k + u_h^k + f_h^k)(z)] = \Pi_{[0,H]}[\Pi_{[0,H]}[\langle \widehat{f}_h^k, \phi(z)\rangle_{\mathcal{H}}] + (1 + 1/H)\beta \cdot \min\{\|\phi(z)\|_{(\Lambda_h^k)^{-1}}, H/\beta\}]$ (See the next lemma for the reformulation of the bonus term). By Lemma 5.35 with δ'/K, we have

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h^k)^{-1}}^2 \\
&\quad \leq \sup_{V \in \overline{\mathcal{V}}(\mathbf{0}, R_K, B_K)} \left\| \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)[V(s_{h+1}^\tau) - \mathbb{P}_h V(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h^k)^{-1}}^2 \\
&\quad \leq 2H^2 \log\det(I + \mathcal{K}_k/\lambda) + 2H^2 k(\lambda - 1) + 4H^2 \log(K \mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}(\epsilon; R_K, B_K)/\delta') + 8k^2\epsilon^2/\lambda \\
&\quad \leq 4H^2 \Gamma(K, \lambda; \ker) + 10H^2 + 4H^2 \log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 4H^2 \log(K/\delta'),
\end{aligned}$$

where the last inequality is by setting $\lambda = 1 + 1/K$ and $\epsilon = \varsigma^* = H/K$. Moreover, the last inequality is also due to

$$\begin{aligned}
\text{dist}(V_1, V_2) &= \sup_{s \in \mathcal{S}} |V_1(s) - V_2(s)| = \sup_{s \in \mathcal{S}} \left| \max_{a \in \mathcal{A}} Q_1(s, a) - \max_{a \in \mathcal{A}} Q_2(s, a) \right| \\
&\leq \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |Q_1(s, a) - Q_2(s, a)| = \|Q_1 - Q_2\|_\infty,
\end{aligned}$$

which indicates that $\mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}(\varsigma^*; R_K, B_K)$ upper bounded by the covering number of the class $\overline{\mathcal{Q}}$ w.r.t. $\|\cdot\|_\infty$, such that

$$\mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}(\varsigma^*; R_K, B_K) \leq \mathcal{N}_\infty(\varsigma^*; R_K, B_K).$$

Here $\mathcal{N}_\infty(\epsilon; R, B)$ denotes the upper bound of the covering number of $\overline{\mathcal{Q}}(h, R, B)$ w.r.t. ℓ_∞-norm, which is characterized in Section 5.7. Further by the union bound, we know that the above inequality holds for all $k \in [K]$ with probability at least $1 - \delta'$. This completes the proof. □

Lemma 5.13. *We define the event $\mathcal{E}$ as that the following inequality holds $\forall z = (s, a) \in \mathcal{S} \times \mathcal{A}, \forall (h, k) \in [H] \times [K]$,*

$$|\mathbb{P}_h V_{h+1}^k(z) - f_h^k(z)| \leq u_h^k(z),$$

where $f_h^k(z) = \Pi_{[0,H]}[\widehat{f}_h^k(z)]$ and $u_h^k(z) = \min\{w_h^k(z), H\}$ with $w_h^k(z) = \beta\lambda^{-1/2}[\ker(z, z) - \psi_h^k(z)^\top(\lambda I + \mathcal{K}_h^k)^{-1}\psi_h^k(z)]^{1/2}$. Thus, setting $\beta = B_K/(1 + 1/H)$, if B_K satisfies

$$16H^2\left[R_Q^2 + 2\Gamma(K, \lambda; \ker) + 5 + \log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 2\log(K/\delta')\right] \leq B_K^2, \forall h \in [H],$$

then we have that with probability at least $1 - \delta'$, the event $\mathcal{E}$ happens, i.e.,

$$\Pr(\mathcal{E}) \geq 1 - \delta'.$$

Proof. We assume that $\mathbb{P}_h V_{h+1}^k(s,a) = \langle \widetilde{f}_h^k, \phi(s,a)\rangle_{\mathcal{H}}$ for some $\widetilde{f}_h^k \in \mathcal{H}$. Then, we bound the difference between $f_h^k(z)$ and $\mathbb{P}_h V_{h+1}^k(s,a)$ in the following way

$$\begin{aligned}
&|\mathbb{P}_h V_{h+1}^k(s,a) - f_h^k(s,a)| \\
&\qquad \leq |\langle \widetilde{f}_h^k, \phi(s,a)\rangle_{\mathcal{H}} - \psi_h^k(s,a)^\top(\lambda \cdot I + \mathcal{K}_h^k)^{-1}\mathbf{y}_h^k| \\
&\qquad = |\lambda\phi(s,a)^\top(\Lambda_h^k)^{-1}\widetilde{f}_h^k + \psi_h^k(s,a)^\top(\lambda \cdot I + \mathcal{K}_h^k)^{-1}\Phi_h^k\widetilde{f}_h^k - \psi_h^k(s,a)^\top(\lambda \cdot I + \mathcal{K}_h^k)^{-1}\mathbf{y}_h^k| \\
&\qquad = |\lambda\phi(s,a)^\top(\Lambda_h^k)^{-1}\widetilde{f}_h^k + \psi_h^k(s,a)^\top(\lambda \cdot I + \mathcal{K}_h^k)^{-1}(\Phi_h^k\overline{f}_h^k - \mathbf{y}_h^k)|,
\end{aligned}$$

where the first inequality is due to $0 \leq \mathbb{P}_h V_{h+1}^k(s,a) \leq H$, non-expansiveness of the operator $\Pi_{[0,H]}[\cdot] := \min\{\cdot, H\}^+$, and the definition of $\widehat{f}_h^k(z)$ in Lemma 5.10, and the first equality is due to

$$\begin{aligned}
\phi(s,a) &= (\Lambda_h^k)^{-1}\Lambda_h^k\phi(s,a) = (\Lambda_h^k)^{-1}(\lambda \cdot I + (\Phi_h^k)^\top\Phi_h^k)\phi(s,a) \\
&= \lambda(\Lambda_h^k)^{-1}\phi(s,a) + (\Lambda_h^k)^{-1}(\Phi_h^k)^\top\Phi_h^k\phi(s,a) \\
&= \lambda(\Lambda_h^k)^{-1}\phi(s,a) + (\Phi_h^k)^\top(\lambda \cdot I + \mathcal{K}_h^k)^{-1}\Phi_h^k\phi(s,a) \\
&= \lambda(\Lambda_h^k)^{-1}\phi(s,a) + (\Phi_h^k)^\top(\lambda \cdot I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a).
\end{aligned} \tag{5.14}$$

Thus, we have

$$\begin{aligned}
|\mathbb{P}_h V_{h+1}^k(s,a,r^k) - f_h^k(s,a)| \leq{}& \underbrace{\lambda\|\phi(s,a)^\top(\Lambda_h^k)^{-1}\|_{\mathcal{H}} \cdot \|\widetilde{f}_h^k\|_{\mathcal{H}}}_{\text{Term(I)}} \\
&+ \underbrace{|\psi_h^k(s,a)^\top(\lambda \cdot I + \mathcal{K}_h^k)^{-1}(\Phi_h^k\widetilde{f}_h^k - \mathbf{y}_h^k)|}_{\text{Term(II)}}.
\end{aligned} \tag{5.15}$$

For Term(I), we have

$$\begin{aligned}
\text{Term(I)} &\leq \sqrt{\lambda}R_Q H\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1} \cdot \lambda I \cdot (\Lambda_h^k)^{-1}\phi(s,a)} \\
&\leq \sqrt{\lambda}R_Q H\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1} \cdot \Lambda_h^k \cdot (\Lambda_h^k)^{-1}\phi(s,a)} \\
&\leq \sqrt{\lambda}R_Q H\sqrt{\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)} = \sqrt{\lambda}R_Q H\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}},
\end{aligned} \tag{5.16}$$

where the first inequality is due to Assumption 5.2 and the second inequality is by $\theta^\top(\Phi_h^k)^\top\Phi_h^k\theta = \|\Phi_h^k\theta\|_{\mathcal{H}} \geq 0$ for any $\theta \in \mathcal{H}$.

For Term(II), we have

$$
\begin{aligned}
\text{Term(II)} &= \left|\phi(s,a)^\top(\Lambda_h^k)^{-1}\left\{\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau)-\mathbb{P}_hV_{h+1}^k(s_h^\tau,a_h^\tau)]\right\}\right| \\
&= \left|\phi(s,a)^\top(\Lambda_h^k)^{-1/2}(\Lambda_h^k)^{-1/2}\left\{\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau)-\mathbb{P}_hV_{h+1}^k(s_h^\tau,a_h^\tau)]\right\}\right| \qquad (5.17)\\
&\le \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\left\|\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau)-\mathbb{P}_hV_{h+1}^k(s_h^\tau,a_h^\tau)]\right\|_{(\Lambda_h^k)^{-1}}
\end{aligned}
$$

By Lemma 5.12, we have that with probability at least $1-\delta'$, the following inequality holds for all $k\in[K]$

$$
\begin{aligned}
&\left\|\sum_{\tau=1}^{k-1}\phi(s_h^\tau,a_h^\tau)[V_{h+1}^k(s_{h+1}^\tau)-\mathbb{P}_hV_{h+1}^k(s_h^\tau,a_h^\tau)]\right\|_{(\Lambda_h^k)^{-1}}\\
&\qquad\le [4H^2\Gamma(K,\lambda;\mathrm{ker})+10H^2+4H^2\log\mathcal{N}_\infty(\varsigma^*;R_K,B_K)+4H^2\log(K/\delta')]^{1/2}.
\end{aligned}
$$

Thus, Term(II) can be further bounded as

$$
\text{Term(II)} \le H\big[4\Gamma(K,\lambda;\mathrm{ker})+10+4\log\mathcal{N}_\infty(\varsigma^*;R_K,B_K)+4\log(K/\delta')\big]^{1/2}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.
$$

Plugging the upper bounds of Term(I) and Term(II) into (5.15), we obtain

$$
\begin{aligned}
&|\mathbb{P}_hV_{h+1}^k(s,a,r^k)-f_h^k(s,a)|\\
&\quad\le H\big[\sqrt{\lambda}R_Q+[4\Gamma(K,\lambda;\mathrm{ker})+10+4\log\mathcal{N}_\infty(\varsigma^*;R_K,B_K)+4\log(K/\delta')]^{1/2}\big]\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\\
&\quad\le H\big[2\lambda R_Q^2+8\Gamma(K,\lambda;\mathrm{ker})+20+4\log\mathcal{N}_\infty(\varsigma^*;R_K,B_K)+8\log(K/\delta')\big]^{1/2}\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}\\
&\quad\le \beta\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}=\beta\lambda^{-1/2}[\mathrm{ker}(z,z)-\psi_h^k(s,a)^\top(\lambda I+\mathcal{K}_h^k)^{-1}\psi_h^k(s,a)]^{1/2},
\end{aligned}
$$

where $\varsigma^*=H/K$, and $\lambda=1+1/K$ as in Lemma 5.12. In the last equality, we also use the identity that

$$
\begin{aligned}
\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}^2 &= \lambda^{-1}\phi(s,a)^\top\phi(s,a)-\lambda^{-1}\psi_h^k(s,a)^\top(\lambda\cdot I+\mathcal{K}_h^k)^{-1}\psi_h^k(s,a)\\
&=\lambda^{-1}\mathrm{ker}(z,z)-\lambda^{-1}\psi_h^k(s,a)^\top(\lambda I+\mathcal{K}_h^k)^{-1}\psi_h^k(s,a).
\end{aligned}\qquad (5.18)
$$

This is proved by

$$\begin{aligned}\|\phi(s,a)\|_{\mathcal{H}}^2 &= \phi(s,a)^\top[\lambda(\Lambda_h^k)^{-1}\phi(s,a) + (\Phi_h^k)^\top(\lambda \cdot I + \mathcal{K}_h^k)^{-1}\Phi_h^k\phi(s,a)]\\ &= \lambda\phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a) + \psi_h^k(s,a)^\top(\lambda \cdot I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a),\end{aligned}$$

where the first equality is by (5.14).

According to Lemma 5.11, we know that $\widehat{f}_h^k$ satisfies $\|\widehat{f}_h^k\|_{\mathcal{H}} \leq H\sqrt{2/\lambda \cdot \log\det(I + \mathcal{K}_h^k/\lambda)} \leq 2H\sqrt{\Gamma(K,\lambda;\mathrm{ker})}$. Then, one can set $R_K = 2H\sqrt{\Gamma(K,\lambda;\mathrm{ker})}$. Moreover, as we set $(1+1/H)\beta = B_K$, then $\beta = B_K/(1+1/H)$. Thus, we let

$$\begin{aligned}&\big[2\lambda R_Q^2 H^2 + 8H^2\Gamma(K,\lambda;\mathrm{ker}) + 20H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 8H^2\log(K/\delta')\big]^{1/2}\\ &\qquad \leq \beta = B_K/(1+1/H),\end{aligned}$$

which can be further guaranteed by

$$16H^2\big[R_Q^2 + 2\Gamma(K,\lambda;\mathrm{ker}) + 5 + \log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 2\log(K/\delta')\big] \leq B_K^2$$

as $(1+1/H) \leq 2$ and $\lambda = 1 + 1/K \leq 2$.

According to the above result, letting $w_h^k = \beta\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = \beta\lambda^{-1/2}[\mathrm{ker}(z,z) - \psi_h^k(s,a)^\top(\lambda I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a)]^{1/2}$, we have $-w_h^k \leq \mathbb{P}_h V_{h+1}^k(s,a) - f_h^k(s,a) \leq w_h^k$. Note that we also have $|\mathbb{P}_h V_{h+1}^k(s,a) - f_h^k(s,a)| \leq H$ due to $0 \leq f_h^k(s,a) \leq H$ and $0 \leq \mathbb{P}_h V_{h+1}^k(s,a) \leq H$. Thus, there is $|\mathbb{P}_h V_{h+1}^k(s,a) - f_h^k(s,a)| \leq \min\{w_h^k, H\}$. This completes the proof. □

Lemma 5.14. *Conditioned on the event $\mathcal{E}$ defined in Lemma 5.13, with probability at least $1-\delta'$, we have*

$$\sum_{k=1}^K V_1^*(s_1, r^k) \leq \sum_{k=1}^K V_1^k(s_1) \leq \mathcal{O}\left(\sqrt{H^3K\log(1/\delta')} + \beta\sqrt{H^2K \cdot \Gamma(K,\lambda;\mathrm{ker})}\right).$$

Proof. We first show the first inequality in this lemma, i.e., $\sum_{k=1}^K V_1^*(s_1, r^k) \leq \sum_{k=1}^K V_1^k(s_1)$. To show this inequality holds, it suffices to show $V_h^*(s, r^k) \leq V_h^k(s)$ for all $s \in \mathcal{S}, h \in [H]$. We prove it by induction.

When $h = H+1$, we know $V_{H+1}^*(s, r^k) = 0$ and $V_{H+1}^k(s) = 0$ such that $V_{H+1}^*(s, r^k) = V_{H+1}^k(s_1)$. Now we assume that $V_{h+1}^*(s, r^k) \leq V_{h+1}^k(s)$. Then, conditioned on the event $\mathcal{E}$ defined in Lemma 5.13, for all $s \in \mathcal{S}$, $(h,k) \in [H] \times [K]$, we further have

$$\begin{aligned}
&Q_h^*(s,a,r^k) - Q_h^k(s,a) \\
&\quad = r_h^k(s,a) + \mathbb{P}_h V_{h+1}^*(s,a,r^k) - \min\{r_h^k(s,a) + f_h^k(s,a) + u_h^k(s,a), H\}^+ \\
&\quad \leq \max\{\mathbb{P}_h V_{h+1}^*(s,a,r^k) - f_h^k(s,a) - u_h^k(s,a), 0\} \\
&\quad \leq \max\{\mathbb{P}_h V_{h+1}^k(s,a) - f_h^k(s,a) - u_h^k(s,a), 0\} \\
&\quad \leq 0
\end{aligned} \tag{5.19}$$

where the first inequality is due to $0 \leq r_h^k(s,a) + \mathbb{P}_h V_{h+1}^*(s,a,r^k) \leq H$ and $\min\{x,y\}^+ \geq \min\{x,y\}$, the second inequality is by the assumption that $V_{h+1}^*(s,r^k) \leq V_{h+1}^k(s)$, the last inequality is by Lemma 5.13 such that $\mathbb{P}_h V_{h+1}^k(s,a) - f_h^k(s,a) \leq u_h^k(s,a)$ holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $(k,h) \in [K] \times [H]$. The above inequality (5.19) further leads to

$$V_h^*(s,r^k) = \max_{a \in \mathcal{A}} Q_h^*(s,a,r^k) \leq \max_{a \in \mathcal{A}} Q_h^k(s,a) = V_h^k(s).$$

Therefore, we obtain that conditioned on event $\mathcal{E}$, we have

$$\sum_{k=1}^K V_1^*(s,r^k) \leq \sum_{k=1}^K V_1^k(s).$$

Next, we prove the second inequality in this lemma, namely the upper bound of $\sum_{k=1}^K V_1^k(s_1)$. Specifically, conditioned on $\mathcal{E}$ defined in Lemma 5.13, we have

$$\begin{aligned}
V_h^k(s_h^k) = Q_h^k(s_h^k,a_h^k) &\leq f_h^k(s_h^k,a_h^k) + r_h^k(s_h^k,a_h^k) + u_h^k(s_h^k,a_h^k) \\
&\leq \mathbb{P}_h V_{h+1}^k(s_h^k,a_h^k) + u_h^k(s_h^k,a_h^k) + r_h^k(s_h^k,a_h^k) + u_h^k(s_h^k,a_h^k) \\
&\leq \mathbb{P}_h V_{h+1}^k(s_h^k,a_h^k) + (2+1/H) w_h^k \\
&= \zeta_h^k + V_{h+1}^k(s_{h+1}^k) + (2+1/H)\beta \|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}},
\end{aligned}$$

where the second inequality is due to Lemma 5.13 and in the last equality, we define

$$\zeta_h^k := \mathbb{P}_h V_{h+1}^k(s_h^k,a_h^k) - V_{h+1}^k(s_{h+1}^k).$$

Recursively applying the above inequality gives

$$V_1^k(s_1) \leq \sum_{h=1}^H \zeta_h^k + (2+1/H)\beta \sum_{h=1}^H \|\phi(s_h^k,a_h^k)\|_{(\Lambda_h^k)^{-1}},$$

where we use the fact that $V_{H+1}^k(\cdot) = 0$. Taking summation on both sides of the above inequality, we have

$$\sum_{k=1}^{K} V_1^k(s_1) = \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k + (2+1/H)\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}}.$$

By Azuma-Hoeffding inequality, with probability at least $1-\delta'$, the following inequalities hold

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k \leq \mathcal{O}\left(\sqrt{H^3 K \log \frac{1}{\delta'}}\right).$$

On the other hand, by Lemma 5.36, we have

$$\begin{aligned}\sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} &= \sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)} \\ &\leq \sum_{h=1}^{H} \sqrt{K \sum_{k=1}^{K} \phi(s_h^k, a_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k)} \\ &\leq \sum_{h=1}^{H} \sqrt{2K \log\det(I + \lambda \mathcal{K}_h^K)} = 2H\sqrt{K \cdot \Gamma(K, \lambda; \mathrm{ker})}.\end{aligned}$$

where the first inequality is by Jensen's inequality. Thus, conditioned on event $\mathcal{E}$, we obtain that with probability at least $1-\delta'$, there is

$$\sum_{k=1}^{K} V_1^k(s_1) \leq \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta\sqrt{H^2 K \cdot \Gamma(K, \lambda; \mathrm{ker})}\right),$$

which completes the proof. □

Lemma 5.15. *We define the event $\widetilde{\mathcal{E}}$ as that the following inequality holds $\forall z = (s, a) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H]$,*

$$|\mathbb{P}_h V_{h+1}(z) - f_h(z)| \leq u_h(z),$$

where $u_h(z) = \min\{w_h(z), H\}^+$ with $w_h(z) = \beta\lambda^{-1/2}[\mathrm{ker}(z,z) - \psi_h(z)^\top(\lambda I + \mathcal{K}_h)^{-1}\psi_h(z)]^{1/2}$. Thus, setting $\beta = \widetilde{B}_K$, if $\widetilde{B}_K$ satisfies

$$4H^2\big[R_Q^2 + 2\Gamma(K, \lambda; \mathrm{ker}) + 5 + \log\mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 2\log(K/\delta')\big] \leq \widetilde{B}_K^2, \forall h \in [H],$$

then we have that with probability at least $1-\delta'$*, the event* $\mathcal{E}$ *happens, i.e.,*

$$\Pr(\widetilde{\mathcal{E}}) \geq 1-\delta'.$$

Proof. The proof of this lemma is nearly the same as the proof of Lemma 5.13. We provide the sketch of this proof below.

We assume that the true transition is formulated as $\mathbb{P}_h V_{h+1}(z) = \langle \widetilde{f}_h, \phi(z)\rangle_{\mathcal{H}} =: \widetilde{f}_h(z)$. We have the following definitions

$$\begin{aligned}
&\Phi_h = [\phi(s_h^1, a_h^1), \phi(s_h^2, a_h^2), \cdots, \phi(s_h^K, a_h^K)]^\top,\\
&\Lambda_h = \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda \cdot I_{\mathcal{H}} = \lambda \cdot I_{\mathcal{H}} + (\Phi_h)^\top \Phi_h,\\
&\mathbf{y}_h = [V_{h+1}(s_{h+1}^1), V_{h+1}(s_{h+1}^2), \cdots, V_{h+1}(s_{h+1}^K)]^\top, \quad \mathcal{K}_h = \Phi_h \Phi_h^\top, \quad \psi_h(s,a) = \Phi_h \phi(s,a).
\end{aligned}$$

Then, we bound the following term

$$\begin{aligned}
&|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)|\\
&\qquad \leq |\langle \widetilde{f}_h, \phi(s,a)\rangle_{\mathcal{H}} - \psi_h(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} \mathbf{y}_h|\\
&\qquad = |\lambda \phi(s,a)^\top \Lambda_h^{-1} \widetilde{f}_h + \psi_h(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} \Phi_h \widetilde{f}_h - \psi_h(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} \mathbf{y}_h|\\
&\qquad = |\lambda \phi(s,a)^\top \Lambda_h^{-1} \widetilde{f}_h + \psi_h^k(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} (\Phi_h \widetilde{f}_h - \mathbf{y}_h)|,
\end{aligned}$$

where the first inequality is due to $0 \leq \mathbb{P}_h V_{h+1}(s,a) \leq H$, the non-expansiveness of the operator $\Pi_{[0,H]}$, and the definition of $\widehat{f}_h(s,a)$ in (5.5), and the first equality is by the same reformulation as (5.14) such that

$$\phi(s,a) = \lambda \Lambda_h^{-1} \phi(s,a) + (\Phi_h)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} \psi_h(s,a).$$

Thus, we have

$$\begin{aligned}
|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)| \leq\ & \underbrace{\lambda \|\phi(s,a)^\top \Lambda_h^{-1}\|_{\mathcal{H}} \cdot \|\widetilde{f}_h\|_{\mathcal{H}}}_{\text{Term(I)}}\\
&+ \underbrace{|\psi_h(s,a)^\top (\lambda \cdot I + \mathcal{K}_h)^{-1} (\Phi_h \widetilde{f}_h - \mathbf{y}_h)|}_{\text{Term(II)}}.
\end{aligned} \tag{5.20}$$

Analogous to (5.16), for Term(I) here, we have

$$\text{Term(I)} \leq \sqrt{\lambda} R_Q H \|\phi(s,a)\|_{\Lambda_h^{-1}}.$$

Similar to (5.17), for Term(II), we have

$$\text{Term(II)} \leq \|\phi(s,a)\|_{\Lambda_h^{-1}} \left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)[V_{h+1}(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}(s_h^\tau, a_h^\tau)] \right\|_{\Lambda_h^{-1}}.$$

Then, we need to bound the last factor in the above inequality. Here we apply the similar argument as Lemma 5.12. We have the function class for V_h is

$$\overline{\mathcal{V}}(r_h, \widetilde{R}_K, \widetilde{B}_K) = \{V : V(\cdot) = \max_{a \in \mathcal{A}} Q(\cdot, a) \text{ with } Q \in \overline{\mathcal{Q}}(r_h, \widetilde{R}_K, \widetilde{B}_K)\}.$$

By Lemma 5.35 with δ', we have

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)[V_{h+1}(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h)^{-1}}^2 \\
&\quad \leq \sup_{V \in \overline{\mathcal{V}}(r_h, \widetilde{R}_K, \widetilde{B}_K)} \left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)[V(s_{h+1}^\tau) - \mathbb{P}_h V(s_h^\tau, a_h^\tau)] \right\|_{(\Lambda_h)^{-1}}^2 \\
&\quad \leq 2H^2 \log\det(I + \mathcal{K}/\lambda) + 2H^2 K(\lambda - 1) + 4H^2 \log(\mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}(\epsilon; \widetilde{R}_K, \widetilde{B}_K)/\delta') + 8K^2\epsilon^2/\lambda \\
&\quad \leq 4H^2 \Gamma(K, \lambda; \text{ker}) + 10H^2 + 4H^2 \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 4H^2 \log(1/\delta'),
\end{aligned}$$

where the last inequality is by setting $\lambda = 1 + 1/K$ and $\epsilon = \varsigma^* = H/K$, and also due to

$$\mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) \leq \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K).$$

We have that with probability at least $1 - \delta'$, the following inequality holds for all $k \in [K]$

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau)[V_{h+1}(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau)] \right\|_{\Lambda_h^{-1}} \\
&\qquad \leq [4H^2 \Gamma(K, \lambda; \text{ker}) + 10H^2 + 4H^2 \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 4H^2 \log(K/\delta')]^{1/2}.
\end{aligned}$$

Thus, Term(II) can be further bounded as

$$\text{Term(II)} \leq H\big[4\Gamma(K, \lambda; \text{ker}) + 10 + 4\log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 4\log(K/\delta')\big]^{1/2} \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

Plugging the upper bounds of Term(I) and Term(II) into (5.20), we obtain

$$
\begin{aligned}
&|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)| \\
&\qquad \leq u_h(s,a) \leq \beta \|\phi(s,a)\|_{\Lambda_h^{-1}} = \beta \lambda^{-1/2} [\ker(z,z) - \psi_h(s,a)^\top (\lambda I + \mathcal{K}_h)^{-1} \psi_h(s,a)]^{1/2},
\end{aligned}
$$

where we let $z = (s,a)$, $\varsigma^* = H/K$, and $\lambda = 1 + 1/K$. In the last equality, similar to (5.18), we have

$$
\begin{aligned}
\|\phi(s,a)\|_{\Lambda_h^{-1}}^2 &= \lambda^{-1}\phi(s,a)^\top \phi(s,a) - \lambda^{-1}\phi(s,a)^\top (\Phi_h)^\top [\lambda I + \Phi_h(\Phi_h)^\top]^{-1} \Phi_h \phi(s,a) \\
&= \lambda^{-1} \ker(z,z) - \lambda^{-1} \psi_h(s,a)^\top (\lambda I + \mathcal{K}_h)^{-1} \psi_h(s,a).
\end{aligned} \tag{5.21}
$$

Similar to Lemma 5.11, we know that the function $\widehat{f}_h$ satisfies $\|\widehat{f}_h\|_{\mathcal{H}} \leq H\sqrt{2/\lambda \cdot \log\det(I + \mathcal{K}_h^k/\lambda)} \leq 2H\sqrt{\Gamma(K,\lambda;\ker)}$. Then, one can set $\widetilde{R}_K = 2H\sqrt{\Gamma(K,\lambda;\ker)}$. Moreover, as we set $\beta = \widetilde{B}_K$. Thus, we let

$$
H\big[2\lambda R_Q^2 + 8\Gamma(K,\lambda;\ker) + 20 + 4\log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 8\log(K/\delta')\big]^{1/2} \leq \beta = \widetilde{B}_K,
$$

which can be further guaranteed by

$$
4H^2\big[R_Q^2 + 2\Gamma(K,\lambda;\ker) + 5 + \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 2\log(K/\delta')\big] \leq \widetilde{B}_K^2
$$

as $(1 + 1/H) \leq 2$ and $\lambda = 1 + 1/K \leq 2$. This completes the proof. □

Lemma 5.16. *Conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma 5.15, we have*

$$
V_h^*(s,r) \leq V_h(s) \leq r_h(s,\pi_h(s)) + \mathbb{P}_h V_{h+1}(s,\pi_h(s)) + 2u_h(s,\pi_h(s)), \forall s \in \mathcal{S}, \forall h \in [H],
$$

where $\pi_h(s) = \mathrm{argmax}_{a\in\mathcal{A}} Q_h(s,a)$.

Proof. We first prove the first inequality in this lemma. We prove it by induction. For $h = H+1$, by the planning algorithm, we have $V_{H+1}^*(s,r) = V_{H+1}(s) = 0$ for any $s \in \mathcal{S}$. Then, we assume that $V_{h+1}^*(s,r) \leq V_{h+1}(s)$. Thus, conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma 5.15, we have

$$
\begin{aligned}
&Q_h^*(s,a,r) - Q_h(s,a) \\
&\qquad = r_h(s,a) + \mathbb{P}_h V_{h+1}^*(s,a,r) - \min\{r_h(s,a) + f_h(s,a) + u_h(s,a), H\}^+ \\
&\qquad \leq \max\{\mathbb{P}_h V_{h+1}^*(s,a,r) - f_h(s,a) - u_h(s,a), 0\} \\
&\qquad \leq \max\{\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a) - u_h(s,a), 0\} \\
&\qquad \leq 0
\end{aligned}
$$

where the first inequality is due to $0 \le r_h(s,a) + \mathbb{P}_h V_{h+1}^*(s,a,r) \le H$ and $\min\{x, H\}^+ \ge \min\{x, H\}$, the second inequality is by the assumption that $V_{h+1}^*(s,a,r) \le V_{h+1}(s,a)$, the last inequality is by Lemma 5.15 such that $|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)| \le u_h(s,a)$ holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $(k,h) \in [K] \times [H]$. The above inequality further leads to

$$V_h^*(s,r) = \max_{a \in \mathcal{A}} Q_h^*(s,a,r) \le \max_{a \in \mathcal{A}} Q_h(s,a) = V_h(s).$$

Therefore, we have

$$V_h^*(s,r) \le V_h(s), \forall h \in [H], \forall s \in \mathcal{S}.$$

In addition, we prove the second inequality in this lemma. We have

$$\begin{aligned} Q_h(s,a) &= \min\{r_h(s,a) + f_h(s,a) + u_h(s,a), H\}^+ \\ &\le \min\{r_h(s,a) + \mathbb{P}_h V_{h+1}(s,a) + 2u_h(s,a), H\}^+ \\ &\le r_h(s,a) + \mathbb{P}_h V_{h+1}(s,a) + 2u_h(s,a), \end{aligned}$$

where the first inequality is also by Lemma 5.15 such that $|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)| \le u_h(s,a)$, and the last inequality is because of the non-negativity of $r_h(s,a) + \mathbb{P}_h V_{h+1}(s,a) + 2u_h(s,a)$. Therefore, we have

$$\begin{aligned} V_h(s) &= \max_{a \in \mathcal{A}} Q_h(s,a) = Q_h(s, \pi_h(s)) \\ &\le r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)). \end{aligned}$$

This completes the proof. □

Lemma 5.17. *With the exploration and planning phases, we have the following inequality*

$$K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^K V_1^*(s_1, r^k).$$

Proof. As shown in (5.21), we know that

$$w_h(s,a) = \beta \|\phi(s,a)\|_{\Lambda_h^{-1}} = \beta \sqrt{\phi(s,a)^\top \left[\lambda I_{\mathcal{H}} + \sum_{\tau=1}^K \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\right]^{-1} \phi(s,a)}.$$

On the other hand, by (5.18), we similarly have

$$w_h^k(s,a) = \beta\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = \beta\sqrt{\phi(s,a)^\top \left[\lambda I_{\mathcal{H}} + \sum_{\tau=1}^{k-1}\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top\right]^{-1} \phi(s,a)}.$$

Since $k-1 \leq K$ and $f^\top \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top f = [f^\top \phi(s_h^\tau, a_h^\tau)]^2 \geq 0$ for any τ, then we know that

$$\Lambda_h = \lambda I_{\mathcal{H}} + \sum_{\tau=1}^{K}\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top \succcurlyeq \lambda I_{\mathcal{H}} + \sum_{\tau=1}^{k-1}\phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top = \Lambda_h^k.$$

We use $A \succcurlyeq B$ (or $A \succ B$) to denote $f^\top A f \geq f^\top B f$ (or $f^\top A f > f^\top B f$), $\forall f \in \mathcal{H}$, for two self-adjoint operators A and B. Moreover, if a linear operator A satisfies $A \succ 0$, we say A is a positive operator.

The above relation further implies that $(\Lambda_h^k)^{-1} \succcurlyeq \Lambda_h^{-1}$ such that $\phi(s,a)^\top \Lambda_h^{-1}\phi(s,a) \leq \phi(s,a)^\top(\Lambda_h^k)^{-1}\phi(s,a)$, where $(\Lambda_h^k)^{-1}$ and Λ_h^{-1} are the inverse of Λ_h^k and Λ_h respectively. Here we use the fact that $\Lambda_h \succcurlyeq \Lambda_h^k$ implies $(\Lambda_h^k)^{-1} \succcurlyeq \Lambda_h^{-1}$, which can be proved by extending the standard matrix case to the self-adjoint operator. For completeness, we give a short proof below.

Let $\lambda > 0$ be a fixed constant. Since $\Lambda_h \succcurlyeq \Lambda_h^k \succcurlyeq \lambda I_{\mathcal{H}} \succ 0$, then there exist the inverse Λ_h^{-1}, $(\Lambda_h^k)^{-1}$ and square root $\Lambda_h^{1/2}$, $(\Lambda_h^k)^{1/2}$, which are also positive self-adjoint and invertible operators. We also have $\Lambda_h^{-1/2} := (\Lambda_h^{1/2})^{-1} = (\Lambda_h^{-1})^{1/2}$ and $(\Lambda_h^k)^{-1/2} := [(\Lambda_h^k)^{1/2}]^{-1} = [(\Lambda_h^k)^{-1}]^{1/2}$. Thus, for any $f \in \mathcal{H}$, we have $f^\top f = f^\top \Lambda_h^{-1/2}\Lambda_h^{1/2}\Lambda_h^{1/2}\Lambda_h^{-1/2} f = f^\top \Lambda_h^{-1/2}\Lambda_h\Lambda_h^{-1/2} f \geq f^\top \Lambda_h^{-1/2}\Lambda_h^k\Lambda_h^{-1/2} f$ where the inequality is due to $\Lambda_h \succcurlyeq \Lambda_h^k$ and $\Lambda_h^{-1/2} = (\Lambda_h^{-1/2})^\top$. Then, we further have $f^\top f \geq f^\top \Lambda_h^{-1/2}\Lambda_h^k\Lambda_h^{-1/2} f = f^\top \Lambda_h^{-1/2}(\Lambda_h^k)^{1/2}(\Lambda_h^k)^{1/2}\Lambda_h^{-1/2} f = f^\top A^\top A f$ if we let $A = (\Lambda_h^k)^{1/2}\Lambda_h^{-1/2}$, where we use the fact that $(\Lambda_h^k)^{1/2}$ and $\Lambda_h^{-1/2}$ are self-adjoint operators. Then, we know that $\|f\|_{\mathcal{H}} \geq \|Af\|_{\mathcal{H}}$ holds for all $f \in \mathcal{H}$, indicating that $\|A\|_{\rm op} := \sup_{f\neq \mathbf{0}} \|Af\|_{\mathcal{H}}/\|f\|_{\mathcal{H}} \leq 1$, where $\|\cdot\|_{\rm op}$ denotes the operator norm. Since $\|A\|_{\rm op} = \|A^\top\|_{\rm op}$, we have $\|A^\top\|_{\rm op} \leq 1$ or equivalently $\|f\|_{\mathcal{H}} \geq \|A^\top f\|_{\mathcal{H}}, \forall f \in \mathcal{H}$, which gives $f^\top f \geq f^\top (\Lambda_h^k)^{1/2}\Lambda_h^{-1/2}\Lambda_h^{-1/2}(\Lambda_h^k)^{1/2} f = f^\top (\Lambda_h^k)^{1/2}\Lambda_h^{-1}(\Lambda_h^k)^{1/2} f$. For any $g \in \mathcal{H}$, letting $f = (\Lambda_h^k)^{-1/2} g$, by $f^\top f \geq f^\top (\Lambda_h^k)^{1/2}\Lambda_h^{-1}(\Lambda_h^k)^{1/2} f$, we have $g^\top (\Lambda_h^k)^{-1} g \geq g^\top \Lambda_h^{-1} g$, which gives $(\Lambda_h^k)^{-1} \succcurlyeq \Lambda_h^{-1}$. The above derivation is based on the basic properties of the linear operator, the (self-)adjoint operator, the inverse, and the square root of an operator. See Kreyszig [1978], Schechter [2001], MacCluer [2008] for the details.

Thus, by the above result, we have

$$w_h(s,a) \leq w_h^k(s,a).$$

Since $r_h^k = 1/H \cdot u_h^k(s,a) = 1/H \cdot \min\{w_h^k(s,a), H\}$ and $u_h(s,a) = \min\{w_h(s,a), H\}$, then we have

$$u_h(s,a)/H \leq r_h^k(s,a),$$

such that

$$V_1^*(s_1, u/H) \leq V_1^*(s_1, r^k),$$

and thus

$$K \cdot V_1^*(s_1, u/H) \leq \sum_{k=1}^{K} V_1^*(s_1, r^k).$$

This completes the proof. □

5.8.2 Proof of Theorem 5.3

Proof. Conditioned on the event $\mathcal{E}$ defined in Lemma 5.13 and the event $\widetilde{\mathcal{E}}$ defined in Lemma 5.15, we have

$$V_1^*(s_1, r) - V_1^\pi(s_1, r) \leq V_1(s_1) - V_1^\pi(s_1, r), \tag{5.22}$$

where the inequality is by Lemma 5.16. Further by this lemma, we have

$$\begin{aligned} V_h(s) - V_h^\pi(s,r) &\leq r_h(s,\pi_h(s)) + \mathbb{P}_h V_{h+1}(s,\pi_h(s)) + 2u_h(s,\pi_h(s)) - Q_h^\pi(s,\pi_h(s),r) \\ &= r_h(s,\pi_h(s)) + \mathbb{P}_h V_{h+1}(s,\pi_h(s)) + 2u_h(s,\pi_h(s)) \\ &\quad - r_h(s,\pi_h(s)) - \mathbb{P}_h V_{h+1}^\pi(s,\pi_h(s),r) \\ &= \mathbb{P}_h V_{h+1}(s,\pi_h(s)) - \mathbb{P}_h V_{h+1}^\pi(s,\pi_h(s),r) + 2u_h(s,\pi_h(s)). \end{aligned}$$

Recursively applying the above inequality and making use of $V_{H+1}^\pi(s,r) = V_{H+1}(s) = 0$ gives

$$\begin{aligned} V_1(s_1) - V_1^\pi(s_1,r) &\leq \mathbb{E}_{\forall h\in[H]:\, s_{h+1}\sim\mathbb{P}_h(\cdot|s_h,\pi_h(s_h))} \left[\sum_{h=1}^{H} 2u_h(s_h,\pi_h(s_h)) \middle| s_1 \right] \\ &= 2H \cdot V_1^\pi(s_1, u/H). \end{aligned}$$

Combining this inequality with (5.22) gives

$$
\begin{aligned}
V_1^*(s_1, r) - V_1^\pi(s_1, r) &\le 2H \cdot V_1^\pi(s_1, u/H) \le \frac{2H}{K} \sum_{k=1}^{K} V_1^*(s_1, r^k) \\
&\le \frac{2H}{K} \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta\sqrt{H^2 K \cdot \Gamma(K, \lambda; \mathrm{ker})}\right) \\
&= \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta\sqrt{H^4 \cdot \Gamma(K, \lambda; \mathrm{ker})}]/\sqrt{K}\right),
\end{aligned}
$$

where the second inequality is due to Lemma 5.17 and the third inequality is by Lemma 5.14.

By the union bound, we have $P(\mathcal{E} \wedge \widetilde{\mathcal{E}}) \ge 1 - 2\delta'$. Therefore, by setting $\delta' = \delta/2$, we obtain that with probability at least $1 - \delta$

$$
V_1^*(s_1, r) - V_1^\pi(s_1, r) \le \mathcal{O}\left([\sqrt{H^5 \log(2/\delta)} + \beta\sqrt{H^4 \cdot \Gamma(K, \lambda; \mathrm{ker})}]/\sqrt{K}\right).
$$

Note that $\mathcal{E} \wedge \widetilde{\mathcal{E}}$ happens when the following two conditions are satisfied, i.e.,

$$
\begin{aligned}
&4H^2\big[R_Q^2 + 2\Gamma(K, \lambda; \mathrm{ker}) + 5 + \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 2\log(2K/\delta)\big] \le \widetilde{B}_K^2, \\
&16H^2\big[R_Q^2 + 2\Gamma(K, \lambda; \mathrm{ker}) + 5 + \log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 2\log(2K/\delta)\big] \le B_K^2, \forall h \in [H],
\end{aligned}
$$

where $\beta = \widetilde{B}_K$, $(1 + 1/H)\beta = B_K$, $\lambda = 1 + 1/K$, $\widetilde{R}_K = R_K = 2H\sqrt{\Gamma(K, \lambda; \mathrm{ker})}$, and $\varsigma^* = H/K$. The above inequalities hold if we further let β satisfy

$$
16H^2\big[R_Q^2 + 2\Gamma(K, \lambda; \mathrm{ker}) + 5 + \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 2\log(2K/\delta)\big] \le \beta^2, \forall h \in [H],
$$

since $2\beta \ge (1 + 1/H)\beta \ge \beta$ such that $\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) \ge \mathcal{N}_\infty(\varsigma^*; R_K, B_K) \ge \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K)$. Since the above conditions imply that $\beta \ge H$, further setting $\delta = 1/(2K^2H^2)$, we obtain that

$$
V_1^*(s_1, r) - V_1^\pi(s_1, r) \le \mathcal{O}\left(\beta\sqrt{H^4[\Gamma(K, \lambda; \mathrm{ker}) + \log(KH)]}/\sqrt{K}\right),
$$

with further letting

$$
16H^2\big[R_Q^2 + 2\Gamma(K, \lambda; \mathrm{ker}) + 5 + \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 6\log(2KH)\big] \le \beta^2, \forall h \in [H].
$$

This completes the proof. □

5.9 Proofs for Single-Agent MDP with Neural Function Approximation

5.9.1 Lemmas

Lemma 5.18 (Lemma C.7 of Yang et al. [2020]). *With $KH^2 = \mathcal{O}(m \log^{-6} m)$, then there exists a constant $\digamma \geq 1$ such that the following inequalities hold with probability at least $1 - 1/m^2$ for any $z \in \mathcal{S} \times \mathcal{A}$ and any $W \in \{W : \|W - W^{(0)}\|_2 \leq H\sqrt{K/\lambda}\}$,*

$$
\begin{aligned}
&|f(z; W) - \varphi(z; W^{(0)})^\top (W - W^{(0)})| \leq \digamma K^{2/3} H^{4/3} m^{-1/6} \sqrt{\log m}, \\
&\|\varphi(z; W) - \varphi(z; W^{(0)})\|_2 \leq \digamma (KH^2/m)^{1/6} \sqrt{\log m}, \qquad \|\varphi(z; W)\|_2 \leq \digamma.
\end{aligned}
$$

Lemma 5.19. *We define the event $\mathcal{E}$ as that the following inequality holds $\forall (s, a) \in \mathcal{S} \times \mathcal{A}, \forall (h, k) \in [H] \times [K]$,*

$$
\begin{aligned}
&|\mathbb{P}_h V_{h+1}^k(s, a) - f_h^k(s, a)| \leq u_h^k(s, a) + \beta \iota, \\
&\left| \|\varphi(z; W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \right| \leq \iota,
\end{aligned}
$$

where $\iota = 5K^{7/12} H^{1/6} m^{-1/12} \log^{1/4} m$ and we define

$$
\Lambda_h^k = \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W_h^k) \varphi(s_h^\tau, a_h^\tau; W_h^k)^\top + \lambda I, \;\; \widetilde{\Lambda}_h^k = \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W^{(0)}) \varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top + \lambda I.
$$

Setting $(1 + 1/H)\beta = B_K$, $R_K = H\sqrt{K}$, $\varsigma^* = H/K$, and $\lambda = \digamma^2(1 + 1/K)$, $\varsigma^* = H/K$, if we set

$$
\beta^2 \geq H^2[8R_Q^2(1 + \sqrt{\lambda/d})^2 + 32\Gamma(K, \lambda; \mathrm{ker}_m) + 80 + 32 \log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 32 \log(K/\delta')],
$$

and also

$$
m = \Omega(K^{19} H^{14} \log^3 m),
$$

then we have that with probability at least $1 - 2/m^2 - \delta'$, the event $\mathcal{E}$ happens, i.e.,

$$
\Pr(\mathcal{E}) \geq 1 - 2/m^2 - \delta'.
$$

Proof. Recall that we assume $\mathbb{P}_h V_{h+1}$ for any V can be expressed as

$$\mathbb{P}_h V_{h+1}(z) = \int_{\mathbb{R}^d} \mathtt{act}'(\boldsymbol{\omega}^\top z) \cdot z^\top \boldsymbol{\alpha}(\boldsymbol{\omega}) \mathrm{d}p_0(\boldsymbol{\omega}),$$

which thus implies that we have

$$\mathbb{P}_h V_{h+1}^k(z) = \int_{\mathbb{R}^d} \mathtt{act}'(\boldsymbol{\omega}^\top z) \cdot z^\top \boldsymbol{\alpha}_h^k(\boldsymbol{\omega}) \mathrm{d}p_0(\boldsymbol{\omega}),$$

for some $\boldsymbol{\alpha}_h^k(\boldsymbol{\omega})$. Our algorithm suggests to estimate $\mathbb{P}_h V_{h+1}^k(s, a)$ via learning the parameters W_h^k by solving

$$W_h^k = \underset{W}{\mathrm{argmin}} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(s_h^\tau, a_h^\tau; W)]^2 + \lambda \|W - W^{(0)}\|_2^2, \tag{5.23}$$

such that we have the estimate of $\mathbb{P}_h V_{h+1}^k(s, a)$ as $f_h^k(z) = \Pi_{[0,H]}[f(z; W_h^k)]$ with

$$f(z; W_h^k) = \frac{1}{\sqrt{2m}} \sum_{i=1}^{2m} v_i \cdot \mathtt{act}([W_h^k]_i^\top z).$$

Furthermore, we have

$$\begin{aligned}
\|W_h^k - W^{(0)}\|_2^2 &\leq \frac{1}{\lambda} \left(\sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(s_h^\tau, a_h^\tau; W_h^k)]^2 + \lambda \|W_h^k - W^{(0)}\|_2^2 \right) \\
&\leq \frac{1}{\lambda} \left(\sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(s_h^\tau, a_h^\tau; W^{(0)})]^2 + \lambda \|W^{(0)} - W^{(0)}\|_2^2 \right) \\
&= \frac{1}{\lambda} \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau)]^2 \leq H^2 K / \lambda,
\end{aligned}$$

where the second inequality is due to W_h^k is the minimizer of the objective function.

We also define a linearization of the function $f(z; W)$ at the point $W^{(0)}$, which is

$$f_{\mathtt{lin}}(z; W) = f(z; W^{(0)}) + \langle \varphi(z; W^{(0)}), W - W^{(0)} \rangle = \langle \varphi(z; W^{(0)}), W - W^{(0)} \rangle, \tag{5.24}$$

where

$$\varphi(z; W) = \nabla_W f(z; W) = [\nabla_{W_1} f(z; W), \cdots, \nabla_{W_{2m}} f(z; W)].$$

Based on this linearization formulation, we similarly define a parameter matrix $W_{\mathtt{lin},h}^k$ that is generated by solving an optimization problem with the linearied function $f_{\texttt{lin}}$, such that

$$W^k_{\texttt{lin},h} = \underset{W}{\operatorname{argmin}} \sum_{\tau=1}^{k-1} [V^k_{h+1}(s^\tau_{h+1}) - f_{\texttt{lin}}(s^\tau_h, a^\tau_h; W)]^2 + \lambda \|W - W^{(0)}\|_2^2. \tag{5.25}$$

Due to the linear structure of $f_{\texttt{lin}}(z; W)$, one can easily solve the above optimization problem and obtain the closed form of the solution $W^k_{\texttt{lin},h}$, which is

$$W^k_{\texttt{lin},h} = W^{(0)} + (\widetilde{\Lambda}^t_h)^{-1} (\widetilde{\Phi}^k_h)^\top \mathbf{y}^k_h, \tag{5.26}$$

where we define Λ^t_h, Φ^k_h, and $\mathbf{y}^k_h$ as

$$\begin{aligned}
\widetilde{\Phi}^k_h &= [\varphi(s^1_h, a^1_h; W^{(0)}), \cdots, \varphi(s^{k-1}_h, a^{k-1}_h; W^{(0)})]^\top, \\
\widetilde{\Lambda}^k_h &= \sum_{\tau=1}^{k-1} \varphi(s^\tau_h, a^\tau_h; W^{(0)}) \varphi(s^\tau_h, a^\tau_h; W^{(0)})^\top + \lambda \cdot I = \lambda \cdot I + (\widetilde{\Phi}^k_h)^\top \widetilde{\Phi}^k_h, \\
\mathbf{y}^k_h &= [V^k_{h+1}(s^1_{h+1}), V^k_{h+1}(s^2_{h+1}), \cdots, V^k_{h+1}(s^{k-1}_{h+1})]^\top.
\end{aligned}$$

Here we also have the upper bound of $\|W^k_{\texttt{lin},h} - W^{(0)}\|_2$ as

$$\begin{aligned}
\|W^k_{\texttt{lin},h} - W^{(0)}\|_2^2 &\le \frac{1}{\lambda} \left(\sum_{\tau=1}^{k-1} [V^k_{h+1}(s^\tau_{h+1}) - f_{\texttt{lin}}(s^\tau_h, a^\tau_h; W^k_{\texttt{lin},h})]^2 + \lambda \|W^k_{\texttt{lin},h} - W^{(0)}\|_2^2 \right) \\
&\le \frac{1}{\lambda} \left(\sum_{\tau=1}^{k-1} [V^k_{h+1}(s^\tau_{h+1}) - f_{\texttt{lin}}(s^\tau_h, a^\tau_h; W^{(0)})]^2 + \lambda \|W^{(0)} - W^{(0)}\|_2^2 \right) \\
&= \frac{1}{\lambda} \sum_{\tau=1}^{k-1} [V^k_{h+1}(s^\tau_{h+1})]^2 \le H^2 K / \lambda,
\end{aligned}$$

where the second inequality is due to $W^k_{\texttt{lin},h}$ is the minimizer of the objective function. Based on the matrix $W^k_{\texttt{lin},h}$, we define the function

$$f^k_{\texttt{lin},h}(z) := \Pi_{[0,H]}[f_{\texttt{lin}}(z; W^k_{\texttt{lin},h})],$$

where $\Pi_{[0,H]}[\cdot]$ is short for $\min\{\cdot, H\}^+$.

Moreover, we further define an approximation of $\mathbb{P}_h V^k_{h+1}$ as

$$\widetilde{f}(z) = \Pi_{[0,H]} \left[\frac{1}{\sqrt{m}} \sum_{i=1}^m \texttt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i \right],$$

where $\|\boldsymbol{\alpha}_i\| \leq R_Q H/\sqrt{dm}$. According to Gao et al. [2019], we have that with probability at least $1 - 1/m^2$ over the randomness of initialization, for any $(h, k) \in [H] \times [K]$, there exists a constant $C_{\texttt{act}}$ such that $\forall z = (s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\left| \mathbb{P}_h V_{h+1}^k(z) - \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \texttt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i \right| \leq 10 C_{\texttt{act}} R_Q H \sqrt{\log(mKH)/m}.$$

which further implies that

$$|\mathbb{P}_h V_{h+1}^k(z) - \widetilde{f}(z)| \leq 10 C_{\texttt{act}} R_Q H \sqrt{\log(mKH)/m}, \ \forall z = (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{5.27}$$

This indicates that $\widetilde{f}(z)$ is a good estimate of $\mathbb{P}_h V_{h+1}^k(z)$ particularly when m is large, i.e., the estimation error $10 C_{\texttt{act}} R_Q H \sqrt{\log(mKH)/m}$ is small.

Now, based on the above definitions and descriptions, we are ready to present our proof of this lemma. Overall, the basic idea of proving the upper bound of $|P_h V_{h+1}^k(z) - f_h^k(z)|$ is to bound the following difference terms, i.e.,

$$|f_h^k(z) - f_{\texttt{lin},h}^k(z)| \ \text{ and } \ |f_{\texttt{lin},h}^k(z) - \widetilde{f}(z)|. \tag{5.28}$$

As we already have known the upper bound of the term $|\mathbb{P}_h V_{h+1}^h(z) - \widetilde{f}(z)|$ in (5.27), one can immediately obtain the upper bound of $|\mathbb{P}_h V_{h+1}^k(z) - f_h^k(z)|$ by decomposing it into the two aforementioned terms and bounding them separately.

We first bound the first term in (5.28), i.e., $|f_h^k(z) - f_{\texttt{lin}}(z; W_{\texttt{lin},h}^k)|$, in the following way

$$\begin{aligned} &|f_h^k(z) - f_{\texttt{lin},h}^k(z)| \\ &\quad \leq |f(z; W_h^k) - \langle \varphi(z; W^{(0)}), W_{\texttt{lin},h}^k - W^{(0)} \rangle| \\ &\quad \leq |f(z; W_h^k) - \langle \varphi(z; W^{(0)}), W_h^k - W^{(0)} \rangle| + |\langle \varphi(z; W^{(0)}), W_h^k - W_{\texttt{lin},h}^k \rangle| \\ &\quad \leq \digamma K^{2/3} H^{4/3} m^{-1/6} \sqrt{\log m} + \digamma \underbrace{\|W_h^k - W_{\texttt{lin},h}^k\|_2}_{\text{Term(I)}}, \end{aligned} \tag{5.29}$$

where the first inequality is due to the non-expansiveness of projection operation $\Pi_{[0,H]}$, the third inequality is by Lemma 5.18 that holds with probability at least $1 - m^{-2}$. Then, we need to bound Term(I) in the above inequality. Specifically, by the first order optimality condition for the objectives in (5.23) and (5.25), we have

$$
\begin{aligned}
\lambda(W_h^k - W^{(0)}) &= \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - f(z_h^\tau; W_h^k)]\varphi(z_h^\tau; W_h^k) = (\Phi_h^k)^\top(\mathbf{y}_h^k - \mathbf{f}_h^k), \\
\lambda(W_{\mathtt{lin},h}^k - W^{(0)}) &= \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - \langle \varphi(z_h^\tau; W^{(0)}), W_{\mathtt{lin},h}^k - W^{(0)} \rangle]\varphi(z_h^\tau; W^{(0)}) \\
&= (\widetilde{\Phi}_h^k)^\top \mathbf{y}_h^k - (\widetilde{\Phi}_h^k)^\top \widetilde{\Phi}_h^k (W_{\mathtt{lin},h}^k - W^{(0)}),
\end{aligned}
$$

where we define

$$
\begin{aligned}
&\Phi_h^k = [\varphi(s_h^1, a_h^1; W_h^k), \cdots, \varphi(s_h^{k-1}, a_h^{k-1}; W_h^k)]^\top, \\
&\Lambda_h^k = \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W_h^k)\varphi(s_h^\tau, a_h^\tau; W_h^k)^\top + \lambda \cdot I = \lambda \cdot I + (\Phi_h^k)^\top \Phi_h^k, \\
&\mathbf{f}_h^k = [f(z_h^1; W_h^k), f(z_h^2; W_h^k), \cdots, f(z_h^{k-1}; W_h^k)]^\top.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
\text{Term(I)} &= \lambda^{-1} \| (\Phi_h^k)^\top (\mathbf{y}_h^k - \mathbf{f}_h^k) - (\widetilde{\Phi}_h^k)^\top \mathbf{y}_h^k + (\widetilde{\Phi}_h^k)^\top \widetilde{\Phi}_h^k (W_{\mathtt{lin},h}^k - W^{(0)}) \|_2 \\
&= \lambda^{-1} \| (\Phi_h^k)^\top (\mathbf{y}_h^k - \mathbf{f}_h^k) - (\widetilde{\Phi}_h^k)^\top \mathbf{y}_h^k + (\widetilde{\Phi}_h^k)^\top \widetilde{\Phi}_h^k (W_{\mathtt{lin},h}^k - W^{(0)}) \|_2 \\
&\leq \lambda^{-1} \| ((\Phi_h^k)^\top - (\widetilde{\Phi}_h^k)^\top) \mathbf{y}_h^k \| + \lambda^{-1} \| (\Phi_h^k)^\top [\mathbf{f}_h^k - \widetilde{\Phi}_h^k (W_{\mathtt{lin},h}^k - W^{(0)})] \|_2 \\
&\quad + \lambda^{-1} \| ((\Phi_h^k)^\top - (\widetilde{\Phi}_h^k)^\top) \widetilde{\Phi}_h^k (W_{\mathtt{lin},h}^k - W^{(0)}) \|_2.
\end{aligned}
$$

According to Lemma 5.18, we can bound the last three terms in the above inequality separately as follows

$$
\begin{aligned}
\lambda^{-1} \| ((\Phi_h^k)^\top - (\widetilde{\Phi}_h^k)^\top) \mathbf{y}_h^k \|_2 &\leq \lambda^{-1} K \max_{\tau \in [k-1]} | [\varphi(z_h^\tau; W_h^k) - \varphi(z_h^\tau; W^{(0)})] \cdot [\mathbf{y}_h^k]_\tau | \\
&\leq \digamma \lambda^{-1} K^{7/6} H^{4/3} m^{-1/6} \sqrt{\log m},
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
&\lambda^{-1} \| (\Phi_h^k)^\top [\mathbf{f}_h^k - \widetilde{\Phi}_h^k (W_{\mathtt{lin},h}^k - W^{(0)})] \|_2 \leq \lambda^{-1} \digamma^2 K^{5/3} H^{4/3} m^{-1/6} \sqrt{\log m}, \\
&\lambda^{-1} \| ((\Phi_h^k)^\top - (\widetilde{\Phi}_h^k)^\top) \widetilde{\Phi}_h^k (W_{\mathtt{lin},h}^k - W^{(0)}) \|_2 \leq \lambda^{-3/2} \digamma^2 K^{5/3} H^{4/3} m^{-1/6} \sqrt{\log m}.
\end{aligned}
$$

Thus, we have

$$\text{Term(I)} \leq \lambda^{-1}(\digamma K^{7/6} + 2\digamma^2 K^{5/3}) H^{4/3} m^{-1/6}\sqrt{\log m} \leq 3K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m}.$$

where we set $\lambda = \digamma^2(1 + 1/K)$, and use the fact that $\lambda \geq 1$ as $\digamma \geq 1$ as well as $\digamma^2/\lambda \in [1/2, 1]$ and $\digamma/\lambda \in [1/2, 1]$. Combining the above upper bound of Term(I) with (5.29), we obtain

$$|f_h^k(z) - f_{\mathtt{lin},h}^k(z)| \leq 4\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m}. \tag{5.30}$$

Next, we bound the second term in (5.28), namely $|f_{\mathtt{lin},h}^k(z) - \widetilde{f}(z)|$. Note that we have

$$\begin{aligned}
&\frac{1}{\sqrt{m}}\sum_{i=1}^{m} \mathtt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i \\
&\quad = \frac{1}{\sqrt{2m}}\sum_{i=1}^{m} \frac{(v_i^{(0)})^2}{\sqrt{2}} \mathtt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i + \frac{1}{\sqrt{2m}}\sum_{i=1}^{m} \frac{(v_i^{(0)})^2}{\sqrt{2}} \mathtt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i \\
&\quad = \frac{1}{\sqrt{2m}}\sum_{i=1}^{m} \frac{(v_i^{(0)})^2}{\sqrt{2}} \mathtt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i + \frac{1}{\sqrt{2m}}\sum_{i=m+1}^{2m} \frac{(v_{i-m}^{(0)})^2}{\sqrt{2}} \mathtt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_{i-m} \\
&\quad = \frac{1}{\sqrt{2m}}\sum_{i=1}^{m} \frac{(v_i^{(0)})^2}{\sqrt{2}} \mathtt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i + \frac{1}{\sqrt{2m}}\sum_{i=m+1}^{2m} \frac{(v_i^{(0)})^2}{\sqrt{2}} \mathtt{act}'(W_i^{(0)\top} z) z^\top \boldsymbol{\alpha}_i \\
&\quad = \frac{1}{\sqrt{2m}}\sum_{i=1}^{2m} v_i^{(0)} \mathtt{act}'(W_i^{(0)\top} z) z^\top (\widetilde{W}_i - W_i^{(0)}) = \langle \varphi(z; W^{(0)}), \widetilde{W} - W^{(0)} \rangle,
\end{aligned}$$

where we define

$$\widetilde{W}_i = \begin{cases} W_i^{(0)} + \frac{v_i^{(0)}}{\sqrt{2}}\boldsymbol{\alpha}_i, \text{ if } 1 \leq i \leq m, \\ W_i^{(0)} + \frac{v_i^{(0)}}{\sqrt{2}}\boldsymbol{\alpha}_{i-m}, \text{ if } m+1 \leq i \leq 2m. \end{cases}$$

Then, we can reformulate $\widetilde{f}(z)$ as follows

$$\widetilde{f}(z) = \Pi_{[0,H]}[\langle \varphi(z; W^{(0)}), \widetilde{W} - W^{(0)} \rangle].$$

Since $\|\boldsymbol{\alpha}_i\|_2 \leq R_Q H/\sqrt{d}$, then there is $\|\widetilde{W} - W^{(0)}\|_2 \leq R_Q H/\sqrt{d}$. Equivalently, we further have

$$\begin{aligned}
\langle \varphi(z; W^{(0)}), \widetilde{W} - W^{(0)} \rangle &= \langle \varphi(z; W^{(0)}), (\widetilde{\Lambda}_h^k)^{-1} \widetilde{\Lambda}_h^k (\widetilde{W} - W^{(0)}) \rangle \\
&= \langle \varphi(z; W^{(0)}), \lambda (\widetilde{\Lambda}_h^k)^{-1} (\widetilde{W} - W^{(0)}) \rangle \\
&\quad + \langle \varphi(z; W^{(0)}), (\widetilde{\Lambda}_h^k)^{-1} (\widetilde{\Phi}_h^k)^\top \widetilde{\Phi}_h^k (\widetilde{W} - W^{(0)}) \rangle,
\end{aligned} \tag{5.31}$$

since $\Lambda_h^k = \lambda I + (\widetilde{\Phi}_h^k)^\top \widetilde{\Phi}_h^k$. Thus, by the above equivalent form of $\widetilde{f}(z)$ in (5.31), and further with the formulation of $f_{\mathtt{lin},h}^k(z)$ according to (5.24) and (5.26), we have

$$
\begin{aligned}
&|f_{\mathtt{lin},h}^k(z) - \widetilde{f}(z)| \\
&\qquad \le |\langle \varphi(z; W^{(0)}), W_{\mathtt{lin},h}^k - \widetilde{W} \rangle| \\
&\qquad \le \underbrace{|\langle \varphi(z; W^{(0)}), \lambda (\widetilde{\Lambda}_h^k)^{-1} (\widetilde{W} - W^{(0)}) \rangle|}_{\text{Term(II)}} \\
&\qquad\quad + \underbrace{|\langle \varphi(z; W^{(0)}), (\widetilde{\Lambda}_h^t)^{-1} (\widetilde{\Phi}_h^k)^\top [\mathbf{y}_h^k - \widetilde{\Phi}_h^k (\widetilde{W} - W^{(0)})] \rangle|}_{\text{Term(III)}}.
\end{aligned}
$$

The first term Term(II) can be bounded as

$$
\begin{aligned}
\text{Term(II)} &= |\langle \varphi(z; W^{(0)}), \lambda (\widetilde{\Lambda}_h^k)^{-1} (\widetilde{W} - W^{(0)}) \rangle| \\
&\le \lambda \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \|\widetilde{W} - W^{(0)}\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\le \sqrt{\lambda} \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \|\widetilde{W} - W^{(0)}\|_2 \\
&\le \sqrt{\lambda} R_Q H / \sqrt{d} \cdot \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}},
\end{aligned}
$$

where the first inequality is by $\|\widetilde{W} - W^{(0)}\|_{(\widetilde{\Lambda}_h^k)^{-1}} = \sqrt{(\widetilde{W} - W^{(0)})^\top (\widetilde{\Lambda}_h^k)^{-1} (\widetilde{W} - W^{(0)})} \le 1/\sqrt{\lambda} \|\widetilde{W} - W^{(0)}\|_2$ since $(\widetilde{\Lambda}_h^k)^{-1} \preccurlyeq 1/\lambda \cdot I$ and the last inequality is due to $\|\widetilde{W} - W^{(0)}\|_2 \le R_Q H / \sqrt{d}$.

Next, we prove the bound of Term(III) in the following way

$$
\begin{aligned}
\text{Term(III)} &= |\langle \varphi(z; W^{(0)}), (\widetilde{\Lambda}_h^t)^{-1} (\widetilde{\Phi}_h^k)^\top [\mathbf{y}_h^k - \widetilde{\Phi}_h^k (\widetilde{W} - W^{(0)})] \rangle| \\
&\le |\langle \varphi(z; W^{(0)}), (\widetilde{\Lambda}_h^t)^{-1} (\widetilde{\Phi}_h^k)^\top [\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k (\widetilde{W} - W^{(0)})] \rangle| \\
&\quad + |\langle \varphi(z; W^{(0)}), (\widetilde{\Lambda}_h^t)^{-1} (\widetilde{\Phi}_h^k)^\top [\mathbf{y}_h^k - \widetilde{\mathbf{y}}_h^k] \rangle| \\
&\le \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \cdot \|(\widetilde{\Phi}_h^k)^\top [\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k (\widetilde{W} - W^{(0)})]\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\quad + \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \cdot \|(\Phi_h^k)^\top [\mathbf{y}_h^k - \widetilde{\mathbf{y}}_h^k]\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\le 10 C_{\mathtt{act}} R_Q H \sqrt{K \log(mKH)/m} \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\quad + \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \cdot \underbrace{\|(\widetilde{\Phi}_h^k)^\top [\mathbf{y}_h^k - \widetilde{\mathbf{y}}_h^k]\|_{(\widetilde{\Lambda}_h^k)^{-1}}}_{\text{Term(IV)}},
\end{aligned}
$$

where we define $\widetilde{\mathbf{y}}_h^k = [\mathbb{P}_h V_{h+1}^k (s_{h+1}^1), \mathbb{P}_h V_{h+1}^k (s_{h+1}^2), \cdots, \mathbb{P}_h V_{h+1}^k (s_{h+1}^{k-1})]^\top$. Here, the last inequality is by

$$
\begin{aligned}
&\|(\Phi_h^k)^\top[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\qquad = \sqrt{[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]^\top \widetilde{\Phi}_h^k[\lambda I + (\widetilde{\Phi}_h^k)^\top \widetilde{\Phi}_h^k]^{-1}(\widetilde{\Phi}_h^k)^\top[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]} \\
&\qquad = \sqrt{[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]^\top \widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top[\lambda I + \widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top]^{-1}[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]} \\
&\qquad \le \sqrt{[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]^\top [\lambda I + \widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top][\lambda I + \widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top]^{-1}[\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})]} \\
&\qquad = \|\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})\|_2 \le 10 C_{\mathtt{act}} R_Q H \sqrt{K \log(mKH)/m},
\end{aligned}
$$

where the second equality is by Woodbury matrix identity, the first inequality is due to $[\lambda I + \widetilde{\Phi}_h^k(\widetilde{\Phi}_h^k)^\top]^{-1} \succ 0$, and the second inequality is by (5.27) such that

$$
\begin{aligned}
\|\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})\|_2 &\le \sqrt{k-1}\|\widetilde{\mathbf{y}}_h^k - \widetilde{\Phi}_h^k(\widetilde{W} - W^{(0)})\|_\infty \\
&= \sqrt{k-1} \sup_{\tau \in [k-1]} |\mathbb{P}_h V_{h+1}^k(s_h^\tau, a_h^\tau) - \widetilde{f}(s_h^\tau, a_h^\tau)| \\
&\le 10 C_{\mathtt{act}} R_Q H \sqrt{K \log(mKH)/m}.
\end{aligned}
$$

In order to further bound Term(IV), we define a new Q-function based on $W_{\mathtt{lin},h}^k$, which is

$$
Q_{\mathtt{lin},h}^k(z) := \Pi_{[0,H]}[r_{\mathtt{lin},h}^k(z) + f_{\mathtt{lin},h}^k(z) + u_{\mathtt{lin},h}^k(z)],
$$

where $r_{\mathtt{lin},h}(s,a) = u_{\mathtt{lin},h}^k(z)/H$, and $u_{\mathtt{lin},h}^k(z) = \min\{\beta\|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}, H\}$. This Q-function can be equivalently reformulated with a normalized representation $\vartheta = \varphi/\digamma$ as follows

$$
\begin{aligned}
Q_{\mathtt{lin},h}^k(z) = \min\{&\Pi_{[0,H]}[\langle \vartheta(z; W^{(0)}), \digamma \cdot (W_{\mathtt{lin},h}^k - W^{(0)})\rangle] \\
&+ (1 + 1/H) \cdot \min\{\beta\|\vartheta(z; W^{(0)})\|_{(\Xi_h^k)^{-1}}\}, H\}^+,
\end{aligned}
\tag{5.32}
$$

where we have

$$
\Xi_h^k := \lambda/\digamma^2 \cdot I + (\Theta_h^k)^\top \Theta_h^k, \qquad \Theta_h^k := \Phi_h^k/\digamma.
$$

Note that $\digamma\|W_{\mathtt{lin},h}^k - W^{(0)}\|_2 \le \digamma H\sqrt{K/\lambda} \le H\sqrt{K}$ since $\lambda = \digamma^2(1 + 1/K)$. Thus, we can see that this new Q-function lies in the space $\overline{\mathcal{Q}}(\mathbf{0}, R_K, B_K)$ as in (5.8), with $R_K = H\sqrt{K}$ and $B_K = (1 + 1/H)\beta$ with the kernel function defined as $\widetilde{\ker}_m(z, z') := \langle \vartheta(z), \vartheta(z')\rangle$.

Now we try to bound the difference between the Q-function $Q_h^k(z)$ in the exploration algorithm and the one $Q_{\mathtt{lin},h}^k(z)$, which is

$$\begin{aligned}&|Q_h^k(z) - Q_{\mathtt{lin},h}^k(z)|\\&\qquad \le |f_h^k(z) - f_{\mathtt{lin},h}^k(z)| + (1+1/H)\beta \left| \|\varphi(z;W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \right|,\end{aligned}$$

where the inequality is by the contraction of the operator $\min\{\cdot, H\}^+$. The upper bound of the term $|f_h^k(z) - f_{\mathtt{lin},h}^k(z)|$ has already been studied in (5.30). Then, we focus on bounding the last term. Thus, we have

$$\begin{aligned}&\left| \|\varphi(z;W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \right|\\&\le \sqrt{\left|\varphi(z;W_h^k)^\top(\Lambda_h^k)^{-1}\varphi(z;W_h^k) - \varphi(z;W^{(0)})^\top(\widetilde{\Lambda}_h^k)^{-1}\varphi(z;W^{(0)})\right|}\\&\le \sqrt{\left|[\varphi(z;W_h^k) - \varphi(z;W^{(0)})]^\top(\Lambda_h^k)^{-1}\varphi(z;W_h^k)\right|} + \sqrt{\left|\varphi(z;W^{(0)})^\top((\Lambda_h^k)^{-1} - (\widetilde{\Lambda}_h^k)^{-1})\varphi(z;W_h^k)\right|}\\&\qquad + \sqrt{\left|\varphi(z;W^{(0)})^\top(\widetilde{\Lambda}_h^k)^{-1}[\varphi(z;W_h^k) - \varphi(z;W^{(0)})]\right|}.\end{aligned}$$

Conditioned on the event that all the inequalities in Lemma 5.18 hold, we can bound the last three terms above as follows

$$\begin{aligned}&\left|[\varphi(z;W_h^k) - \varphi(z;W^{(0)})]^\top(\Lambda_h^k)^{-1}\varphi(z;W_h^k)\right|\\&\qquad \le \|\varphi(z;W_h^k) - \varphi(z;W^{(0)})\|_2\|(\Lambda_h^k)^{-1}\|_2\|\varphi(z;W_h^k)\|_2 \le \lambda^{-1}\digamma^2(KH^2/m)^{1/6}\sqrt{\log m},\\&\left|\varphi(z;W^{(0)})^\top(\widetilde{\Lambda}_h^k)^{-1}[\varphi(z;W_h^k) - \varphi(z;W^{(0)})]\right| \le \lambda^{-1}\digamma^2(KH^2/m)^{1/6}\sqrt{\log m},\\&\left|\varphi(z;W^{(0)})^\top((\Lambda_h^k)^{-1} - (\widetilde{\Lambda}_h^k)^{-1})\varphi(z;W_h^k)\right|\\&\qquad \le \|\varphi(z;W^{(0)})\|_2\|(\Lambda_h^k)^{-1}(\Lambda_h^k - \widetilde{\Lambda}_h^k)(\widetilde{\Lambda}_h^k)^{-1}\|_2\|\varphi(z;W_h^k)\|_2\\&\qquad \le \lambda^{-2}\digamma^2\|(\Phi_h^k)^\top\Phi_h^k - (\widetilde{\Phi}_h^k)^\top\widetilde{\Phi}_h^k\|_{\mathrm{fro}} \le \lambda^{-2}\digamma^2(\|(\Phi_h^k - \widetilde{\Phi}_h^k)^\top\Phi_h^k\|_{\mathrm{fro}} + \|(\widetilde{\Phi}_h^k)^\top(\Phi_h^k - \widetilde{\Phi}_h^k)\|_{\mathrm{fro}})\\&\qquad \le \lambda^{-2}\digamma^4K^{7/6}H^{1/3}m^{-1/6}\sqrt{\log m},\end{aligned}$$

which thus lead to

$$\left| \|\varphi(z;W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \right| \le 3K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m, \tag{5.33}$$

and thus

$$|Q_h^k(z) - Q_{\mathtt{lin},h}^k(z)| \le 4\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 3(1+1/H)\beta K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m,$$

where we use the fact that $\lambda = \digamma^2(1+1/K) \in [\digamma^2, 2\digamma^2]$. This further implies that we have the same bound for $|V_h^k(s) - V_{\texttt{lin},h}^k(s)|$, .i.e.,

$$
\begin{aligned}
|V_h^k(s) - V_{\texttt{lin},h}^k(s)| &\le \max_{a\in\mathcal{A}} |Q_h^k(s,a) - Q_{\texttt{lin},h}^k(s,a)| \\
&\le 4\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 3(1+1/H)\beta K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m,
\end{aligned}
\tag{5.34}
$$

where we define $V_{\texttt{lin},h}^k(s) = \max_{a\in\mathcal{A}} Q_{\texttt{lin},h}^k(s,a)$.

Now, we are ready to bound Term(IV). With probability at least $1-\delta'$, we have

$$
\begin{aligned}
&\text{Term(IV)} \\
&\quad = \Bigg\| \sum_{\tau=1}^{k-1} [V_{h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{h+1}^k(z_h^\tau)]\varphi(z_h^\tau; W^{(0)}) \Bigg\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\quad \le \Bigg\| \sum_{\tau=1}^{k-1} [V_{\texttt{lin},h+1}^k(s_{h+1}^\tau) - \mathbb{P}_h V_{\texttt{lin},h+1}^k(z_h^\tau)]\varphi(z_h^\tau; W^{(0)}) \Bigg\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\qquad + \Bigg\| \sum_{\tau=1}^{k-1} \{[V_{h+1}^k(s_{h+1}^\tau) - V_{\texttt{lin},h+1}^k(s_{h+1}^\tau)] - \mathbb{P}_h[V_{h+1}^k - V_{\texttt{lin},h+1}^k(s_{h+1}^\tau)]\}\varphi(z_h^\tau; W^{(0)}) \Bigg\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\quad \le [4H^2\Gamma(K,\lambda'; \widetilde{\ker}_m) + 10H^2 + 4H^2\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 4H^2\log(K/\delta')]^{1/2} \\
&\qquad + 8\digamma K^{8/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 12\beta K^{19/12}H^{1/6}m^{-1/12}\log^{1/4} m.
\end{aligned}
$$

Here we set $\lambda' = \lambda/\digamma^2 = (1+1/K)$, $\varsigma^* = H/K$, $R_K = H\sqrt{K}$, $B_K = (1+1/H)\beta$, and $\widetilde{\ker}_m(z,z') = \langle \vartheta(z), \vartheta(z')\rangle$. Here the second inequality is by (5.32), and also follows the similar proof of Lemma 5.12. The last inequality is by (5.34) and Lemma 5.18, which lead to

$$
\begin{aligned}
&\Bigg\| \sum_{\tau=1}^{k-1} \{[V_{h+1}^k(s_{h+1}^\tau) - V_{\texttt{lin},h+1}^k(s_{h+1}^\tau)] - \mathbb{P}_h[V_{h+1}^k - V_{\texttt{lin},h+1}^k(s_{h+1}^\tau)]\}\varphi(z_h^\tau; W^{(0)}) \Bigg\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\quad \le \sum_{\tau=1}^{k-1} [8\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 12\beta K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m] \|\varphi(z_h^\tau; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\quad \le K\digamma/\sqrt{\lambda}[8\digamma K^{5/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 12\beta K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m] \\
&\quad \le 8\digamma K^{8/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 12\beta K^{19/12}H^{1/6}m^{-1/12}\log^{1/4} m,
\end{aligned}
$$

where we use $\digamma^2/\lambda = 1/(1+1/K) \le 1$ and $(1+1/H) \le 2$ due to $H \ge 1$. Now we let β satisfy

$$
\begin{aligned}
&\sqrt{\lambda}R_Q H/\sqrt{d} + 10C_{\texttt{act}}R_Q H\sqrt{K\log(mKH)/m} + H[4\Gamma(K,\lambda'; \widetilde{\ker}_m) + 4\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) \\
&\quad + 10 + 4\log(K/\delta')]^{1/2} + 8\digamma K^{8/3}H^{4/3}m^{-1/6}\sqrt{\log m} + 12\beta K^{19/12}H^{1/6}m^{-1/12}\log^{1/4} m \le \beta.
\end{aligned}
$$

To obtain the above relation, it suffices to set

$$m = \Omega(K^{19} H^{14} \log^3 m)$$

such that m is sufficient large which results in

$$\begin{aligned}10 C_{\mathtt{act}} R_Q H \sqrt{K \log(mKH)/m} + 8\digamma K^{8/3} H^{4/3} m^{-1/6} \sqrt{\log m} \\ + 12\beta K^{19/12} H^{1/6} m^{-1/12} \log^{1/4} m \leq R_Q H + \beta/2.\end{aligned}$$

Then, there is

$$\begin{aligned}&\sqrt{\lambda} R_Q H/\sqrt{d} + R_Q H + \beta/2 \\ &\qquad + 2H[\Gamma(K,\lambda;\ker_m) + 5/2 + \log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + \log(K/\delta')]^{1/2} \leq \beta,\end{aligned}$$

where $\Gamma(K,\lambda;\ker_m) = \Gamma(K,\lambda';\widetilde{\ker}_m)$ with $\ker_m := \langle \varphi(z;W^{(0)}), \varphi(z';W^{(0)})\rangle$. This inequality can be satisfied if we set β as

$$\beta^2 \geq H^2[8R_Q^2(1+\sqrt{\lambda/d})^2 + 32\Gamma(K,\lambda;\ker_m) + 80 + 32\log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 32\log(K/\delta')].$$

If the above conditions hold, we have

$$|f_{\mathtt{lin},h}^k(z) - \widetilde{f}(z)| \leq \beta \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \leq w_h^k + \beta(3K^{7/12} H^{1/6} m^{-1/12} \log^{1/4} m),$$

where the inequality is due to (5.33). Since $f_{\mathtt{lin},h}^k(z) \in [0,H]$ and $\widetilde{f}(z) \in [0,H]$, thus we have $|f_{\mathtt{lin},h}^k(z) - \widetilde{f}(z)| \leq H$, which further gives

$$\begin{aligned}|f_{\mathtt{lin},h}^k(z) - \widetilde{f}(z)| &\leq \min\{w_h^k, H\} + \beta(3K^{7/12} H^{1/6} m^{-1/12} \log^{1/4} m) \\ &= u_h^k + \beta(3K^{7/12} H^{1/6} m^{-1/12} \log^{1/4} m).\end{aligned} \tag{5.35}$$

Now we combine (5.30) and (5.35) as well as (5.27) and obtain

$$\begin{aligned}&|\mathbb{P}_h V_{h+1}^k(z) - f_h^k(z)| \\ &\qquad \leq |\mathbb{P}_h V_{h+1}^k(z) - \widetilde{f}(z)| + |f_h^k(z) - f_{\mathtt{lin},h}^k(z)| + |f_{\mathtt{lin},h}^k(z) - \widetilde{f}(z)| \\ &\qquad \leq 10 C_{\mathtt{act}} R_Q H \sqrt{\log(mKH)/m} + 4\digamma K^{5/3} H^{4/3} m^{-1/6} \sqrt{\log m} \\ &\qquad\quad + u_h^k + \beta(3K^{7/12} H^{1/6} m^{-1/12} \log^{1/4} m) \\ &\qquad \leq u_h^k + \beta(5K^{7/12} H^{1/6} m^{-1/12} \log^{1/4} m),\end{aligned}$$

with m are sufficiently. We also have $\Big|\|\varphi(z;W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}\Big| \leq \iota$ according to (5.33). The above inequalities hold with probability at least $1 - 2/m^2 - \delta'$ by the union bound. This completes the proof. □

Lemma 5.20. *Conditioned on the event $\mathcal{E}$ defined in Lemma 5.19, with probability at least $1-\delta'$, we have*

$$
\begin{aligned}
&\sum_{k=1}^K V_1^*(s_1, r^k) \leq \sum_{k=1}^K V_1^k(s_1) + \beta H K \iota,\\
&\sum_{k=1}^K V_1^k(s_1) \leq \mathcal{O}\Big(\sqrt{H^3 K \log(1/\delta')} + \beta\sqrt{H^2 K \cdot \Gamma(K,\lambda;\ker_m)}\Big) + \beta H K \iota,
\end{aligned}
$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$.

Proof. We first show the first inequality in this lemma. We prove $V_h^*(s, r^k) \leq V_h^k(s) + (H+1-h)\iota$ for all $s \in \mathcal{S}, h \in [H]$ by induction. When $h = H+1$, we know $V_{H+1}^*(s, r^k) = 0$ and $V_{H+1}^k(s) = 0$ such that $V_{H+1}^*(s, r^k) \leq V_{H+1}^k(s_1)$. Now we assume that $V_{h+1}^*(s, r^k) \leq V_{h+1}^k(s) + (H-h)\beta\iota$. Then, conditioned on the event $\mathcal{E}$ defined in Lemma 5.13, for all $s \in \mathcal{S}$, $(h,k) \in [H] \times [K]$, we further have

$$
\begin{aligned}
&Q_h^*(s,a,r^k) - Q_h^k(s,a)\\
&\quad = r_h^k(s,a) + \mathbb{P}_h V_{h+1}^*(s,a,r^k) - \min\{r_h^k(s,a) + f_h^k(s,a) + u_h^k(s,a), H\}^+\\
&\quad \leq \max\{\mathbb{P}_h V_{h+1}^*(s,a,r^k) - f_h^k(s,a) - u_h^k(s,a), 0\}\\
&\quad \leq \max\{\mathbb{P}_h V_{h+1}^k(s,a) + \beta(H-h)\iota - f_h^k(s,a) - u_h^k(s,a), 0\}\\
&\quad \leq \beta(H+1-h)\iota,
\end{aligned}
\tag{5.36}
$$

where the first inequality is due to $0 \leq r_h^k(s,a) + \mathbb{P}_h V_{h+1}^*(s,a,r^k) \leq H$ and $\min\{x,y\}^+ \geq \min\{x,y\}$, the second inequality is by the assumption that $V_{h+1}^*(s,r^k) \leq V_{h+1}^k(s) + (H-h)\beta\iota$, the last inequality is by Lemma 5.19 such that $|\mathbb{P}_h V_{h+1}^k(s,a) - f_h^k(s,a)| \leq u_h^k(s,a) + \beta\iota$ holds for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $(k,h) \in [K] \times [H]$. The above inequality (5.36) further leads to

$$
V_h^*(s,r^k) = \max_{a\in\mathcal{A}} Q_h^*(s,a,r^k) \leq \max_{a\in\mathcal{A}} Q_h^k(s,a) = V_h^k(s) + \beta(H+1-h)\iota.
$$

Therefore, we obtain that conditioned on event $\mathcal{E}$, we have

$$
\sum_{k=1}^K V_1^*(s,r^k) \leq \sum_{k=1}^K V_1^k(s) + \beta H K \iota.
$$

Next, we prove the second inequality in this lemma. Conditioned on $\mathcal{E}$ defined in Lemma 5.19, we have

$$\begin{aligned} V_h^k(s_h^k) = Q_h^k(s_h^k, a_h^k) &\leq \max\{0, f_h^k(s_h^k, a_h^k) + r_h^k(s_h^k, a_h^k) + u_h^k(s_h^k, a_h^k)\} \\ &\leq \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k) + u_h^k(s_h^k, a_h^k) + r_h^k(s_h^k, a_h^k) + u_h^k(s_h^k, a_h^k) \\ &\leq \zeta_h^k + V_{h+1}^k(s_{h+1}^k) + (2 + 1/H)\beta \|\varphi(s_h^k, a_h^k; W_h^k)\|_{(\Lambda_h^k)^{-1}}, \end{aligned}$$

where we define

$$\zeta_h^k := \mathbb{P}_h V_{h+1}^k(s_h^k, a_h^k) - V_{h+1}^k(s_{h+1}^k).$$

Recursively applying the above inequality gives

$$V_1^k(s_1) \leq \sum_{h=1}^{H} \zeta_h^k + (2 + 1/H)\beta \sum_{h=1}^{H} \|\varphi(s_h^k, a_h^k; W_h^k)\|_{(\Lambda_h^k)^{-1}},$$

where we use the fact that $V_{H+1}^k(\cdot) = 0$. Taking summation on both sides of the above inequality, we have

$$\sum_{k=1}^{K} V_1^k(s_1) = \sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k + (2 + 1/H)\beta \sum_{k=1}^{K}\sum_{h=1}^{H} \|\varphi(s_h^k, a_h^k; W_h^k)\|_{(\Lambda_h^k)^{-1}}.$$

By Azuma-Hoeffding inequality, with probability at least $1 - \delta'$, the following inequalities hold

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \zeta_h^k \leq \mathcal{O}\left(\sqrt{H^3 K \log \frac{1}{\delta'}}\right).$$

On the other hand, by Lemma 5.36, we have

$$\begin{aligned} \sum_{k=1}^{K}\sum_{h=1}^{H} \|\phi(s_h^k, a_h^k)\|_{(\Lambda_h^k)^{-1}} &= \sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\varphi(s_h^k, a_h^k; W_h^k)^\top (\Lambda_h^k)^{-1} \phi(s_h^k, a_h^k; W_h^k)} \\ &\leq \sum_{k=1}^{K}\sum_{h=1}^{H} \sqrt{\varphi(s_h^k, a_h^k; W^{(0)})^\top (\widetilde{\Lambda}_h^k)^{-1} \varphi(s_h^k, a_h^k; W^{(0)})} + HK\iota \\ &\leq \sum_{h=1}^{H} \sqrt{K \sum_{k=1}^{K} \varphi(s_h^k, a_h^k; W^{(0)})^\top (\widetilde{\Lambda}_h^k)^{-1} \varphi(s_h^k, a_h^k; W^{(0)}))} + HK\iota \\ &= 2H\sqrt{K \cdot \Gamma(K, \lambda; \ker_m)} + HK\iota. \end{aligned}$$

where the first inequality is due to Lemma 5.19, the second inequality is by Jensen's inequality.

Thus, conditioned on event $\mathcal{E}$, we obtain that with probability at least $1-\delta'$, there is

$$\sum_{k=1}^{K} V_1^k(s_1) \leq \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta\sqrt{H^2 K \cdot \Gamma(K,\lambda;\mathrm{ker})}\right) + \beta H K \iota,$$

which completes the proof. □

Lemma 5.21. *We define the event $\widetilde{\mathcal{E}}$ as that the following inequality holds $\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \forall h \in [H]$,*

$$\begin{aligned}
&|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)| \leq u_h(s,a) + \beta\iota,\\
&\left|\|\varphi(z;W_h)\|_{(\Lambda_h)^{-1}} - \|\varphi(z;W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}}\right| \leq \iota,
\end{aligned}$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$ and we define

$$\Lambda_h = \sum_{\tau=1}^{K} \varphi(z_h^\tau; W_h)\varphi(z_h^\tau; W_h)^\top + \lambda \cdot I, \quad \widetilde{\Lambda}_h = \sum_{\tau=1}^{K} \varphi(z_h^\tau; W^{(0)})\varphi(z_h^\tau; W^{(0)})^\top + \lambda \cdot I.$$

Setting $\beta = \widetilde{B}_K$, $\widetilde{R}_K = H\sqrt{K}$, $\varsigma^* = H/K$, and $\lambda = \digamma^2(1+1/K)$, $\varsigma^* = H/K$, if we set

$$\beta^2 \geq H^2[8R_Q^2(1+\sqrt{\lambda/d})^2 + 32\Gamma(K,\lambda;\mathrm{ker}_m) + 80 + 32\log\mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K) + 32\log(K/\delta')],$$

and also

$$m = \Omega(K^{19}H^{14}\log^3 m),$$

then we have that with probability at least $1-2/m^2-\delta'$, the event $\widetilde{\mathcal{E}}$ happens, i.e.,

$$\Pr(\widetilde{\mathcal{E}}) \geq 1 - 2/m^2 - \delta'.$$

Proof. The proof of this lemma exactly follows our proof of Lemma 5.19. There are several minor differences here. In the proof of this lemma, we set $\widetilde{B}_K = \beta$ instead of $(1+1/H)\beta$ due to the structure of the planning phase. Moreover, we use $\mathcal{N}_\infty(\epsilon; R_K, B_K)$ to denote covering number of the Q-function class $\overline{\mathcal{Q}}(r_h, R_K, B_K)$. Since the covering numbers of $\overline{\mathcal{Q}}(r_h, R_K, B_K)$ and $\overline{\mathcal{Q}}(\mathbf{0}, R_K, B_K)$ are the same where the former one only has an extra bias r_h, we use the same notation $\mathcal{N}_\infty(\epsilon; R_K, B_K)$ to denote their covering number. Then, the rest of this proof can be completed by using the same argument as the proof of Lemma 5.19. □

Lemma 5.22. *Conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma 5.21, we have*

$$\begin{aligned}
&V_h^*(s,r) \leq V_h(s) + (H+1-h)\beta\iota, \forall s \in \mathcal{S}, \forall h \in [H],\\
&V_h(s) \leq r_h(s,\pi_h(s)) + \mathbb{P}_h V_{h+1}(s,\pi_h(s)) + 2u_h(s,\pi_h(s)) + \beta\iota, \forall s \in \mathcal{S}, \forall h \in [H],
\end{aligned}$$

where $\pi_h(s) = \mathrm{argmax}_{a\in\mathcal{A}} Q_h(s,a)$.

Proof. We first prove the first inequality in this lemma by induction. For $h = H+1$, we have $V_{H+1}^*(s,r) = V_{H+1}(s) = 0$ for any $s \in \mathcal{S}$. Then, we assume that $V_{h+1}^*(s,r) \leq V_{h+1}(s) + (H-h)\beta\iota$. Thus, conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma 5.21, we have

$$\begin{aligned}
&Q_h^*(s,a,r) - Q_h(s,a)\\
&\quad = r_h(s,a) + \mathbb{P}_h V_{h+1}^*(s,a,r) - \min\{r_h(s,a) + f_h(s,a) + u_h(s,a), H\}^+\\
&\quad \leq \max\{\mathbb{P}_h V_{h+1}^*(s,a,r) - f_h(s,a) - u_h(s,a), 0\}\\
&\quad \leq \max\{\mathbb{P}_h V_{h+1}(s,a) + (H-h)\beta\iota - f_h(s,a) - u_h(s,a), 0\}\\
&\quad \leq (H+1-h)\beta\iota,
\end{aligned}$$

where the first inequality is due to $0 \leq r_h(s,a) + \mathbb{P}_h V_{h+1}^*(s,a,r) \leq H$ and $\min\{x,y\}^+ \geq \min\{x,y\}$, the second inequality is by the assumption that $V_{h+1}^*(s,a,r) \leq V_{h+1}(s,a) + (H-h)\beta\iota$, the last inequality is by Lemma 5.21 such that $|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)| \leq u_h(s,a) + \beta\iota$ holds for any $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $(k,h) \in [K]\times[H]$. The above inequality further leads to

$$V_h^*(s,r) = \max_{a\in\mathcal{A}} Q_h^*(s,a,r) \leq \max_{a\in\mathcal{A}} Q_h(s,a) + (H+1-h)\beta\iota = V_h(s) + (H+1-h)\beta\iota.$$

Therefore, we have

$$V_h^*(s,r) \leq V_h(s) + (H+1-h)\beta\iota, \forall h \in [H], \forall s \in \mathcal{S}.$$

We further prove the second inequality in this lemma. We have

$$\begin{aligned}
Q_h(s,a) &= \min\{r_h(s,a) + f_h(s,a) + u_h(s,a), H\}^+\\
&\leq \min\{r_h(s,a) + \mathbb{P}_h V_{h+1}(s,a) + 2u_h(s,a) + \beta\iota, H\}^+\\
&\leq r_h(s,a) + \mathbb{P}_h V_{h+1}(s,a) + 2u_h(s,a) + \beta\iota,
\end{aligned}$$

where the first inequality is also by Lemma 5.21 such that $|\mathbb{P}_h V_{h+1}(s,a) - f_h(s,a)| \leq u_h(s,a) + \beta\iota$, and the last inequality is because of the non-negativity of $r_h(s,a) + \mathbb{P}_h V_{h+1}(s,a) + 2u_h(s,a) + \beta\iota$.

Therefore, we have

$$V_h(s) = \max_{a \in \mathcal{A}} Q_h(s,a) = Q_h(s, \pi_h(s)) \le r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)) + \beta\iota.$$

This completes the proof. □

Lemma 5.23. *With the exploration and planning phases, conditioned on events $\mathcal{E}$ and $\widetilde{\mathcal{E}}$, we have the following inequality*

$$K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota,$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$.

Proof. The bonus for the planning phase is $u_h(s,a) = \min\{\beta w_h(s,a), H\}$ where $w_h(s,a) = \|\varphi(s,a;W_h)\|_{\Lambda_h^{-1}}$. We also have $H \cdot r_h^k(s,a) = u_h^k(s,a) = \min\{\beta w_h^k(s,a), H\}$ where $w_h^k(s,a) = \|\varphi(s,a;W_h^k)\|_{(\Lambda_h^k)^{-1}}$. Conditioned on events $\mathcal{E}$ and $\widetilde{\mathcal{E}}$, according to Lemmas 5.19 and 5.21, we have

$$\begin{aligned}
\left| \|\varphi(s,a;W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \right| &\le \iota, \\
\left| \|\varphi(s,a;W_h)\|_{(\Lambda_h)^{-1}} - \|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}} \right| &\le \iota,
\end{aligned}$$

such that

$$\begin{aligned}
\beta w_h(s,a) &\le \beta \|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}} + \beta\iota, \\
\beta\iota + \beta w_h^k(s,a) &\ge \beta \|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}}.
\end{aligned}$$

Moreover, we know

$$\begin{aligned}
&\|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}} \\
&\quad = \sqrt{\varphi(s,a;W^{(0)})^\top \left[\lambda I + \sum_{\tau=1}^K \varphi(s_h^\tau, a_h^\tau; W^{(0)}) \varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top \right]^{-1} \varphi(s,a;W^{(0)})},
\end{aligned}$$

and also

$$\begin{aligned}
&\|\varphi(s,a;W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \\
&\quad = \sqrt{\varphi(s,a;W^{(0)})^\top \left[\lambda I + \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W^{(0)}) \varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top \right]^{-1} \varphi(s,a;W^{(0)})}.
\end{aligned}$$

Since $k-1 \leq K$ and $x^\top \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top x = [x^\top \phi(s_h^\tau, a_h^\tau)]^2 \geq 0, \forall x$, then we know that

$$\begin{aligned} \widetilde{\Lambda}_h &= \lambda I + \sum_{\tau=1}^{K} \varphi(s_h^\tau, a_h^\tau; W^{(0)}) \varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top \\ &\succcurlyeq \lambda I + \sum_{\tau=1}^{k-1} \varphi(s_h^\tau, a_h^\tau; W^{(0)}) \varphi(s_h^\tau, a_h^\tau; W^{(0)})^\top = \widetilde{\Lambda}_h^k. \end{aligned}$$

The above relation further implies that $\widetilde{\Lambda}_h^{-1} \preccurlyeq (\widetilde{\Lambda}_h^k)^{-1}$ such that

$$\varphi(s,a;W^{(0)})^\top \widetilde{\Lambda}_h^{-1} \varphi(s,a;W^{(0)}) \leq \varphi(s,a;W^{(0)})^\top (\widetilde{\Lambda}_h^k)^{-1} \varphi(s,a;W^{(0)}).$$

Thus, we have

$$\beta w_h(s,a) \leq \beta w_h^k(s,a) + 2\beta\iota,$$

such that

$$\min\{\beta w_h(s,a), H\} \leq \min\{\beta w_h^k(s,a) + 2\beta\iota, H\} \leq \min\{\beta w_h^k(s,a), H\} + 2\beta\iota,$$

which further implies that

$$u_h(s,a) \leq u_h^k(s,a) + 2\beta\iota = H \cdot r_h^k(s,a) + 2\beta\iota.$$

Then, by the definition of the value function, we have

$$V_1^*(s_1, u/H) \leq V_1^*(s_1, r^k) + 2\beta\iota,$$

which thus gives

$$K \cdot V_1^*(s_1, u/H) \leq \sum_{k=1}^{K} V_1^*(s_1, r^k) + 2K\beta\iota.$$

This completes the proof. □

5.9.2 Proof of Theorem 5.5

Proof. Conditioned on the event $\mathcal{E}$ in Lemma 5.19 and the event $\widetilde{\mathcal{E}}$ in Lemma 5.21, we have

$$V_1^*(s_1, r) - V_1^\pi(s_1, r) \leq V_1(s_1) - V_1^\pi(s_1, r) + H\beta\iota, \tag{5.37}$$

where the inequality is by Lemma 5.22. Further by this lemma, we have

$$\begin{aligned} V_h(s) - V_h^\pi(s, r) &\le r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)) - Q_h^\pi(s, \pi_h(s), r) + \beta\iota \\ &= r_h(s, \pi_h(s)) + \mathbb{P}_h V_{h+1}(s, \pi_h(s)) + 2u_h(s, \pi_h(s)) - r_h(s, \pi_h(s)) \\ &\quad - \mathbb{P}_h V_{h+1}^\pi(s, \pi_h(s), r) + \beta\iota \\ &= \mathbb{P}_h V_{h+1}(s, \pi_h(s)) - \mathbb{P}_h V_{h+1}^\pi(s, \pi_h(s), r) + 2u_h(s, \pi_h(s)) + \beta\iota. \end{aligned}$$

Recursively applying the above inequality and making use of $V_{H+1}^\pi(s, r) = V_{H+1}(s) = 0$ gives

$$\begin{aligned} V_1(s_1) - V_1^\pi(s_1, r) &\le \mathbb{E}_{\forall h \in [H]:\, s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, \pi_h(s_h))} \left[\sum_{h=1}^H 2u_h(s_h, \pi_h(s_h)) \middle| s_1 \right] + H\beta\iota \\ &= 2H \cdot V_1^\pi(s_1, u/H) + H\beta\iota. \end{aligned}$$

Combining with (5.37) gives

$$\begin{aligned} V_1^*(s_1, r) - V_1^\pi(s_1, r) &\le 2H \cdot V_1^\pi(s_1, u/H) + 2H\beta\iota \le \frac{2H}{K} \sum_{k=1}^K V_1^*(s_1, r^k) + 4H\beta\iota \\ &\le \frac{2H}{K} \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta\sqrt{H^2 K \cdot \Gamma(K, \lambda; \mathrm{ker}_m)} \right) + H\beta\iota(H+4) \\ &\le \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta\sqrt{H^4 \cdot \Gamma(K, \lambda; \mathrm{ker}_m)}]/\sqrt{K} + H^2\beta\iota \right), \end{aligned}$$

where the second inequality is due to Lemma 5.23 and the third inequality is by Lemma 5.20.

By the union bound, we have $P(\mathcal{E} \wedge \widetilde{\mathcal{E}}) \ge 1 - 2\delta' - 4/m^2$. Therefore, by setting $\delta' = 1/(4K^2H^2)$, we obtain that with probability at least $1 - 1/(2K^2H^2) - 4/m^2$

$$\begin{aligned} V_1^*(s_1, r) - V_1^\pi(s_1, r) &\le \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta\sqrt{H^4 \cdot \Gamma(K, \lambda; \mathrm{ker}_m)}]/\sqrt{K} + H^2\beta\iota \right) \\ &\le \mathcal{O}\left(\beta\sqrt{H^4[\Gamma(K, \lambda; \mathrm{ker}_m) + \log(KH)]}/\sqrt{K} + H^2\beta\iota \right), \end{aligned}$$

where the last inequality is due to $\beta \ge H$. Note that $\mathcal{E} \wedge \widetilde{\mathcal{E}}$ happens when the following two conditions are satisfied, i.e.,

$$\begin{aligned} &\beta^2 \ge H^2[8R_Q^2(1 + \sqrt{\lambda/d})^2 + 32\Gamma(K, \lambda; \mathrm{ker}_m) + 80 + 32\log\mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 32\log(K/\delta')], \\ &\beta^2 \ge H^2[8R_Q^2(1 + \sqrt{\lambda/d})^2 + 32\Gamma(K, \lambda; \mathrm{ker}_m) + 80 + 32\log\mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 32\log(K/\delta')], \end{aligned}$$

where $\beta = \widetilde{B}_K$,$(1 + 1/H)\beta = B_K$, $\lambda = \varGamma(1 + 1/K)$, $\widetilde{R}_K = R_K = H\sqrt{K}$, and $\varsigma^* = H/K$. The above inequalities hold if we further let β satisfy

$$\beta^2 \geq H^2[8R_Q^2(1+\sqrt{\lambda/d})^2 + 32\Gamma(K,\lambda;\ker_m) + 80 + 32\log\mathcal{N}_\infty(\varsigma^*;R_K,2\beta) + 96\log(2KH)],$$

since $2\beta \geq (1+1/H)\beta \geq \beta$ such that $\mathcal{N}_\infty(\varsigma^*;R_K,2\beta) \geq \mathcal{N}_\infty(\varsigma^*;R_K,B_K) \geq \mathcal{N}_\infty(\varsigma^*;\widetilde{R}_K,\widetilde{B}_K)$. This completes the proof. □

5.10 Proofs for Markov Game with Kernel Function Approximation

5.10.1 Lemmas

Lemma 5.24. *We define the event $\mathcal{E}$ as that the following inequality holds $\forall(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}, \forall(h,k)\in[H]\times[K]$,*

$$|\mathbb{P}_h V_{h+1}^k(s,a,b) - f_h^k(s,a,b)| \leq u_h^k(s,a,b),$$

where $u_h^k(s,a,b) = \min\{w_h^k(s,a,b),H\}$, $w_h^k(s,a,b) = \beta\lambda^{-1/2}[\ker(z,z) - \psi_h^k(s,a,b)^\top(\lambda I + \mathcal{K}_h^k)^{-1}\psi_h^k(s,a,b)]^{1/2}$ with $z=(s,a,b)$, and $f_h^k(z) = \Pi_{[0,H]}[\psi_h^k(z)^\top(\lambda\cdot I + \mathcal{K}_h^k)^{-1}\mathbf{y}_h^k]$ with

$$\begin{aligned}
&\psi_h^k(z) = \Phi_h^k\phi(z) = [\ker(z,z_h^1),\cdots,\ker(z,z_h^{k-1})]^\top,\\
&\Phi_h^k = [\phi(z_h^1),\phi(z_h^2),\cdots,\phi(z_h^{k-1})]^\top,\\
&\mathbf{y}_h^k = [V_{h+1}^k(s_{h+1}^1),V_{h+1}^k(s_{h+1}^2),\cdots,V_{h+1}^k(s_{h+1}^{k-1})]^\top,\\
&\mathcal{K}_h^k = \Phi_h^k(\Phi_h^k)^\top = \begin{bmatrix} \ker(z_h^1,z_h^1) & \dots & \ker(z_h^1,z_h^{k-1})\\ \vdots & \ddots & \vdots \\ \ker(z_h^{k-1},z_h^1) & \dots & \ker(z_h^{k-1},z_h^{k-1})\end{bmatrix},
\end{aligned}$$

Thus, setting $\beta = B_K/(1+1/H)$, if B_K satisfies

$$16H^2\big[R_Q^2 + 2\Gamma(K,\lambda;\ker) + 5 + \log\mathcal{N}_\infty(\varsigma^*;R_K,B_K) + 2\log(K/\delta')\big] \leq B_K^2, \forall h\in[H],$$

then we have that with probability at least $1-\delta'$, the event $\mathcal{E}$ happens, i.e.,

$$\Pr(\mathcal{E}) \geq 1-\delta'.$$

Proof. According to the exploration algorithm for the game, we can see that by letting $\boldsymbol{a}=(a,b)$ be an action in the space $\mathcal{A}\times\mathcal{B}$, Algorithm 8 reduces to Algorithm 6 with the action space $\mathcal{A}\times\mathcal{B}$

and state space $\mathcal{S}$. Now, we also have a transition in the form of $\mathbb{P}_h(s|\boldsymbol{a})$ and a product policy $(\pi_h^k \otimes \nu_h^k)(s)$ such that $\boldsymbol{a} \sim (\pi_h^k \otimes \nu_h^k)(s)$ at state $s \in \mathcal{S}$ for all $(h,k) \in [H] \times [K]$. Similarly, we have $Q_h^k(s,a,b) = Q_h^k(s,\boldsymbol{a})$ and $V_h^k(s,a,b) = V_h^k(s,\boldsymbol{a})$ as well as $u_h^k(s,a,b) = u_h^k(s,\boldsymbol{a})$ and $u_h^k(s,a,b) = u_h^k(s,\boldsymbol{a})$ and $r_h^k(s,a,b) = r_h^k(s,\boldsymbol{a})$. Thus, we can simply apply the proof of Lemma 5.13 and obtain the proof for this lemma. This completes the proof. □

Lemma 5.25. *Conditioned on the event $\mathcal{E}$ defined in Lemma 5.24, with probability at least $1-\delta'$, we have*

$$\sum_{k=1}^{K} V_1^*(s_1, r^k) \le \sum_{k=1}^{K} V_1^k(s_1) \le \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta\sqrt{H^2 K \cdot \Gamma(K,\lambda;\ker)}\right).$$

Proof. By the reduction of Algorithm 8 to Algorithm 6, we can apply the same proof as the one for Lemma 5.14, which completes the proof. □

Lemma 5.26. *We define the event $\widetilde{\mathcal{E}}$ as that the following inequality holds $\forall (s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}, \forall h \in [H]$,*

$$|\mathbb{P}_h \overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b)| \le u_h(s,a,b), \tag{5.38}$$

$$|\mathbb{P}_h \underline{V}_{h+1}(s,a,b) - \underline{f}_h(s,a,b)| \le u_h(s,a,b), \tag{5.39}$$

where $u_h(s,a,b) = \overline{u}_h(s,a,b) = \underline{u}_h(s,a,b) = \min\{w_h(s,a,b), H\}$, $w_h(s,a,b) = \beta\lambda^{-1/2}[\ker(z,z) - \psi_h(s,a,b)^\top(\lambda I + \mathcal{K}_h)^{-1}\psi_h(s,a,b)]^{1/2}$ *with* $z = (s,a,b)$, $\mathcal{K}_h = \Phi_h \Phi_h^\top$, *and* $\psi_h(s,a,b) = \Phi_h \phi(s,a,b)$ *with* $\Phi_h = [\phi(z_h^1), \phi(z_h^2), \cdots, \phi(z_h^K)]^\top$. *Moreover, we have*

$$\overline{f}_h(s,a,b) = \Pi_{[0,H]}[\psi_h(s,a,b)^\top(\lambda \cdot I + \mathcal{K}_h)^{-1}\overline{\mathbf{y}}_h],$$

$$\underline{f}_h(s,a,b) = \Pi_{[0,H]}[\psi_h(s,a,b)^\top(\lambda \cdot I + \mathcal{K}_h)^{-1}\underline{\mathbf{y}}_h],$$

where $\overline{\mathbf{y}}_h := [\overline{V}_{h+1}(s_{h+1}^1), \cdots, \overline{V}_{h+1}(s_{h+1}^K)]^\top$ *and* $\underline{\mathbf{y}}_h := [\underline{V}_{h+1}(s_{h+1}^1), \cdots, \underline{V}_{h+1}(s_{h+1}^K)]^\top$.

Thus, setting $\beta = \widetilde{B}_K$, *if* $\widetilde{B}_K$ *satisfies*

$$4H^2\big[R_Q^2 + 2\Gamma(K,\lambda;\ker) + 5 + \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 2\log(2K/\delta')\big] \le \widetilde{B}_K^2, \forall h \in [H],$$

then we have that with probability at least $1-\delta'$, the event $\mathcal{E}$ happens, i.e.,

$$\Pr(\widetilde{\mathcal{E}}) \ge 1 - \delta'.$$

Proof. According to the construction of u_h and $\overline{f}_h$, the proof for the the first inequality in this lemma is nearly the same as the proof of Lemma 5.15 but one difference for computing the covering number of the value function space. Specifically, we have the function class for $\overline{V}_h$ which is

$$\overline{\mathcal{V}}(r_h, \widetilde{R}_K, \widetilde{B}_K) = \{V : V(\cdot) = \max_{a\sim\pi'} \min_{b\sim\nu'} \mathbb{E}_{\pi',\nu'} Q(\cdot, a, b) \text{ with } Q \in \overline{\mathcal{Q}}(r_h, \widetilde{R}_K, \widetilde{B}_K)\}.$$

By Lemma 5.35 with $\delta'/2$, we have

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau, b_h^\tau)[\overline{V}_{h+1}(s_{h+1}^\tau) - \mathbb{P}_h \overline{V}_{h+1}(s_h^\tau, a_h^\tau, b_h^\tau)] \right\|_{(\Lambda_h)^{-1}}^2 \\
&\qquad \le \sup_{V \in \overline{\mathcal{V}}(r_h, \widetilde{R}_K, \widetilde{B}_K)} \left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau, b_h^\tau)[\overline{V}(s_{h+1}^\tau) - \mathbb{P}_h \overline{V}(s_h^\tau, a_h^\tau, b_h^\tau)] \right\|_{(\Lambda_h)^{-1}}^2 \\
&\qquad \le 2H^2 \log\det(I + \mathcal{K}/\lambda) + 2H^2 K(\lambda - 1) + 4H^2 \log(\mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}(\epsilon; \widetilde{R}_K, \widetilde{B}_K)/\delta') + 8K^2\epsilon^2/\lambda \\
&\qquad \le 4H^2\Gamma(K, \lambda; \ker) + 10H^2 + 4H^2 \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 4H^2 \log(2/\delta'),
\end{aligned}$$

where the last inequality is by setting $\lambda = 1 + 1/K$ and $\epsilon = \varsigma^* = H/K$. Here $\mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}$ is the covering number of the function space $\overline{\mathcal{V}}$ w.r.t. the distance $\text{dist}(V_1, V_2) = \sup_s |V_1(s) - V_2(s)|$, and $\mathcal{N}_\infty$ is the covering number for the function space $\overline{\mathcal{Q}}$ w.r.t. the infinity norm. In the last inequality, we also use

$$\mathcal{N}_{\text{dist}}^{\overline{\mathcal{V}}}(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) \le \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K),$$

which is in particular due to

$$\begin{aligned}
\text{dist}(V_1, V_2) &= \sup_{s\in\mathcal{S}} |V_1(s) - V_2(s)| \\
&= \sup_{s\in\mathcal{S}} |\max_{\pi'} \min_{\nu'} \mathbb{E}_{a\sim\pi', b\sim\nu'}[Q_1(s, a, b)] - \max_{\pi''} \min_{\nu''} \mathbb{E}_{a\sim\pi'', b\sim\nu''}[Q_2(s, a, b)]| \\
&\le \sup_{s\in\mathcal{S}} \sup_{a\in\mathcal{A}} \sup_{b\in\mathcal{B}} |Q_1(s, a, b) - Q_2(s, a, b)| \\
&= \|Q_1(\cdot,\cdot,\cdot) - Q_2(\cdot,\cdot,\cdot)\|_\infty,
\end{aligned} \tag{5.40}$$

where we use the fact that max-min operator is non-expansive. Thus, we have that with probability at least $1 - \delta'/2$, the following inequality holds for all $k \in [K]$

$$\begin{aligned}
&\left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau, b_h^\tau)[\overline{V}_{h+1}(s_{h+1}^\tau) - \mathbb{P}_h \overline{V}_{h+1}(s_h^\tau, a_h^\tau, b_h^\tau)] \right\|_{\Lambda_h^{-1}} \\
&\qquad \le [4H^2\Gamma(K, \lambda; \ker) + 10H^2 + 4H^2 \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 4H^2 \log(2K/\delta')]^{1/2}.
\end{aligned}$$

Then, the rest of the proof for (5.38) follows the proof of Lemma 5.15.

Next, we give the proof of (5.39). We define another function class for $\underline{V}_h$ as

$$\underline{\mathcal{V}}(r_h, \widetilde{R}_K, \widetilde{B}_K) = \{V : V(\cdot) = \max_{a\sim\pi'} \min_{b\sim\nu'} \mathbb{E}_{\pi',\nu'} Q(\cdot, a, b) \text{ with } Q \in \underline{\mathcal{Q}}(r_h, \widetilde{R}_K, \widetilde{B}_K)\}.$$

Note that as we can show in the covering number for the function spaces $\underline{\mathcal{Q}}$ and $\overline{\mathcal{Q}}$ have the same covering number upper bound. Therefore, we use the same notation $\mathcal{N}_\infty$ for their upper bound. Thus, by the similar argument as (5.40), we have that with probability at least $1-\delta'/2$, the following inequality holds for all $k \in [K]$

$$\left\| \sum_{\tau=1}^{K} \phi(s_h^\tau, a_h^\tau, b_h^\tau)[\underline{V}_{h+1}(s_{h+1}^\tau) - \mathbb{P}_h \underline{V}_{h+1}(s_h^\tau, a_h^\tau, b_h^\tau)] \right\|_{\Lambda_h^{-1}}$$
$$\leq [4H^2\Gamma(K, \lambda; \mathrm{ker}) + 10H^2 + 4H^2 \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 4H^2 \log(2K/\delta')]^{1/2},$$

where we use the fact that

$$\mathcal{N}_{\mathrm{dist}}^{\underline{\mathcal{V}}}(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) \leq \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K).$$

The rest of the proof are exactly the same as the proof of Lemma 5.15.

In this lemma, we let

$$H\big[2\lambda R_Q^2 + 8\Gamma(K, \lambda; \mathrm{ker}) + 20 + 4 \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 8 \log(2K/\delta')\big]^{1/2} \leq \beta = \widetilde{B}_K,$$

which can be further guaranteed by

$$4H^2\big[R_Q^2 + 2\Gamma(K, \lambda; \mathrm{ker}) + 5 + \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 2 \log(2K/\delta')\big] \leq \widetilde{B}_K^2$$

as $(1 + 1/H) \leq 2$ and $\lambda = 1 + 1/K \leq 2$. This completes the proof. □

Lemma 5.27. *Conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma 5.26, we have*

$$V_h^\dagger(s, r) \leq \overline{V}_h(s) \leq \mathbb{E}_{a\sim\pi_h, b\sim \mathrm{br}(\pi)_h}[(\mathbb{P}_h \overline{V}_{h+1} + r_h + 2u_h)(s, a, b)], \forall s \in \mathcal{S}, \forall h \in [H], \quad (5.41)$$
$$V_h^\dagger(s, r) \geq \underline{V}_h(s) \geq \mathbb{E}_{a\sim \mathrm{br}(\nu)_h, b\sim\nu_h}[(\mathbb{P}_h \underline{V}_{h+1} - r_h - 2u_h)(s, a, b)], \forall s \in \mathcal{S}, \forall h \in [H]. \quad (5.42)$$

Proof. For the first inequality of (5.41), we can prove it by induction. We first prove the first inequality in this lemma. We prove it by induction. For $h = H + 1$, by the planning algorithm, we have $V_{H+1}^\dagger(s, r) = V_{H+1}(s) = 0$ for any $s \in \mathcal{S}$. Then, we assume that $V_{h+1}^\dagger(s, r) \leq \overline{V}_{h+1}(s)$.

Thus, conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma 5.26, we have

$$\begin{aligned}
&Q_h^\dagger(s,a,b,r) - \overline{Q}_h(s,a,b) \\
&\qquad = r_h(s,a,b) + \mathbb{P}_h V_{h+1}^\dagger(s,a,b,r) - \min\{r_h(s,a,b) + \overline{f}_h(s,a,b) + u_h(s,a,b), H\}^+ \\
&\qquad \le \max\{\mathbb{P}_h V_{h+1}^\dagger(s,a,b,r) - \overline{f}_h(s,a,b) - u_h(s,a,b), 0\} \\
&\qquad \le \max\{\mathbb{P}_h \overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b) - u_h(s,a,b), 0\} \le 0,
\end{aligned}$$

where the first inequality is due to $0 \le r_h(s,a,b) + \mathbb{P}_h V_{h+1}^\dagger(s,a,b,r) \le H$ and $\min\{x,y\}^+ \ge \min\{x,y\}$, the second inequality is by the assumption that $V_{h+1}^\dagger(s,a,b,r) \le \overline{V}_{h+1}(s,a,b)$, the last inequality is by Lemma 5.26 such that $|\mathbb{P}_h \overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b)| \le u_h(s,a,b)$ holds for any $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and $(k,h) \in [K] \times [H]$. Thus, the above inequality leads to

$$V_h^\dagger(s,r) = \max_{\pi'_h} \min_{\nu'_h} \mathbb{E}_{a\sim\pi'_h, b\sim\nu'_h}[Q_h^\dagger(s,a,b,r)] \le \max_{\pi'_h} \min_{\nu'_h} \mathbb{E}_{a\sim\pi'_h, b\sim\nu'_h}[\overline{Q}_h(s,a,b)] = \overline{V}_h(s),$$

which eventually gives

$$V_h^*(s,r) \le V_h(s), \forall h \in [H], \forall s \in \mathcal{S}.$$

To prove the second inequality of (5.41), we have

$$\begin{aligned}
\overline{V}_h(s) &= \min_{\nu'} \mathbb{E}_{a\sim\pi_h, b\sim\nu'} \overline{Q}_h(s,a,b) \\
&\le \mathbb{E}_{a\sim\pi_h, b\sim\mathrm{br}(\pi)_h} \overline{Q}_h(s,a,b) \\
&= \mathbb{E}_{a\sim\pi_h, b\sim\mathrm{br}(\pi)_h} \min\{(\overline{f}_h + r_h + u_h)(s,a,b), H\}^+ \\
&\le \mathbb{E}_{a\sim\pi_h, b\sim\mathrm{br}(\pi)_h} \min\{(\mathbb{P}_h \overline{V}_{h+1} + r_h + 2u_h)(s,a,b), H\}^+ \\
&\le \mathbb{E}_{a\sim\pi_h, b\sim\mathrm{br}(\pi)_h} [(\mathbb{P}_h \overline{V}_{h+1} + r_h + 2u_h)(s,a,b)],
\end{aligned}$$

where the first and the second equality is by the iterations in Algorithm 9, the second inequality is by Lemma 5.26, and the last inequality is due to the non-negativity of $(\mathbb{P}_h \overline{V}_{h+1} + r_h + 2u_h)(s,a,b)$.

For the inequalities in (5.42), one can similarly adopt the argument above to give the proof. From the perspective of Player 2, this player is trying to find a policy to maximize the cumulative rewards w.r.t. a reward function $\{-r_h(s,a,b)\}_{h\in[H]}$. Thus, the proof of (5.42) follows the proof of (5.41). This completes the proof. □

Lemma 5.28. *With the exploration and planning phases, we have the following inequalities*

$$K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k), \quad K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k).$$

Proof. First, we have $K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, u/H) \le K \cdot V_1^*(s_1, u/H)$, as well as $K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, u/H) \le K \cdot V_1^*(s_1, u/H)$ due to the definition of $V_1^*(\cdot, u/H)$. Thus, to prove this lemma, we only need to show

$$K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k).$$

Since the constructions of u_h and r_h^k are the same as the ones for the single-agent case, similar to the proof of Lemma 5.17, we have

$$u_h(s, a)/H \le r_h^k(s, a),$$

such that

$$V_1^*(s_1, u/H) \le V_1^*(s_1, r^k),$$

and thus

$$K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k).$$

Therefore, we eventually obtain

$$K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, u/H) \le K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k),$$
$$K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, u/H) \le K \cdot V_1^*(s_1, u/H) \le \sum_{k=1}^{K} V_1^*(s_1, r^k).$$

This completes the proof. □

5.10.2 Proof of Theorem 5.6

Proof. Conditioned on the event $\mathcal{E}$ defined in Lemma 5.24 and the event $\widetilde{\mathcal{E}}$ defined in Lemma 5.26, we have

$$V_1^{\dagger}(s_1, r) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) \leq \overline{V}_1(s_1) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r), \tag{5.43}$$

where the inequality is by Lemma 5.27. Further by this lemma, we have

$$\begin{aligned}
&\overline{V}_h(s_h) - V_h^{\pi, \mathrm{br}(\pi)}(s_h, r) \\
&\quad \leq \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h}[(\mathbb{P}_h \overline{V}_{h+1} + r_h + 2u_h)(s_h, a_h, b_h)] - V_h^{\pi, \mathrm{br}(\pi)}(s_h, r) \\
&\quad = \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h}[(r_h + \mathbb{P}_h \overline{V}_{h+1} + 2u_h)(s_h, a_h, b_h) - r_h(s_h, a_h, b_h) - \mathbb{P}_h V_{h+1}^{\pi, \mathrm{br}(\pi)}(s_h, a_h, b_h, r)] \\
&\quad = \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h}[\mathbb{P}_h \overline{V}_{h+1}(s_h, a_h, b_h) - \mathbb{P}_h V_{h+1}^{\pi, \mathrm{br}(\pi)}(s_h, a_h, b_h, r) + 2u_h(s_h, a_h, b_h)] \\
&\quad = \mathbb{E}_{a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h, s_{h+1} \sim \mathbb{P}_h}[\overline{V}_{h+1}(s_{h+1}) - V_{h+1}^{\pi, \mathrm{br}(\pi)}(s_{h+1}, r) + 2u_h(s_h, a_h, b_h)].
\end{aligned}$$

Recursively applying the above inequality and making use of $\overline{V}_{H+1}(s) = V_{H+1}^{\pi, \mathrm{br}(\pi)}(s, r) = 0$ yield

$$\begin{aligned}
&\overline{V}_1(s_1) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) \\
&\quad \leq \mathbb{E}_{\forall h \in [H]:\ a_h \sim \pi_h, b_h \sim \mathrm{br}(\pi)_h, s_{h+1} \sim \mathbb{P}_h} \left[\sum_{h=1}^{H} 2u_h(s_h, a_h, b_h) \middle| s_1 \right] \\
&\quad = 2H \cdot V_1^{\pi, \mathrm{br}(\pi)}(s_1, u/H).
\end{aligned}$$

Combining this inequality with (5.43) gives

$$\begin{aligned}
V_1^{\dagger}(s_1, r) - V_1^{\pi, \mathrm{br}(\pi)}(s_1, r) &\leq 2H \cdot V_1^{\pi, \mathrm{br}(\pi)}(s_1, u/H) \leq \frac{2H}{K} \sum_{k=1}^{K} V_1^*(s_1, r^k) \\
&\leq \frac{2H}{K} \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta \sqrt{H^2 K \cdot \Gamma(K, \lambda; \mathrm{ker})}\right) \\
&\leq \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta \sqrt{H^4 \cdot \Gamma(K, \lambda; \mathrm{ker})}]/\sqrt{K}\right),
\end{aligned}$$

where the second inequality is due to Lemma 5.28 and the third inequality is by Lemma 5.25.

Next, we prove the upper bound of the term $V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^{\dagger}(s_1, r)$. Conditioned on the event $\mathcal{E}$ defined in Lemma 5.24 and the event $\widetilde{\mathcal{E}}$ defined in Lemma 5.26, we have

$$V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - V_1^{\dagger}(s_1, r) \leq V_1^{\mathrm{br}(\nu), \nu}(s_1, r) - \underline{V}_1(s_1, r), \tag{5.44}$$

where the inequality is by Lemma 5.27. Further by Lemma 5.27, we have

$$\begin{aligned}
&V_h^{\mathrm{br}(\nu),\nu}(s_h, r) - \underline{V}_h(s_h)\\
&\quad\le V_h^{\mathrm{br}(\nu),\nu}(s_h, r) - \mathbb{E}_{a\sim \mathrm{br}(\nu)_h, b\sim\nu_h}[(\mathbb{P}_h\underline{V}_{h+1} - r_h - 2u_h)(s_h, a_h, b_h)]\\
&\quad= \mathbb{E}_{a_h\sim \mathrm{br}(\nu)_h, b_h\sim\nu_h}[\mathbb{P}_h V_{h+1}^{\mathrm{br}(\nu),\nu}(s_h, a_h, b_h, r) - \mathbb{P}_h\underline{V}_{h+1}(s_h, a_h, b_h) + 2u_h(s_h, a_h, b_h)]\\
&\quad= \mathbb{E}_{a_h\sim \mathrm{br}(\nu)_h, b_h\sim\nu_h, s_{h+1}\sim\mathbb{P}_h}[V_{h+1}^{\mathrm{br}(\nu),\nu}(s_{h+1}, r) - \mathbb{P}_h\underline{V}_{h+1}(s_{h+1}) + 2u_h(s_h, a_h, b_h)].
\end{aligned}$$

Recursively applying the above inequality yields

$$V_1^{\mathrm{br}(\nu),\nu}(s_1, r) - \underline{V}_1(s_h, r) \le 2H \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, u/H).$$

Combining this inequality with (5.44) gives

$$\begin{aligned}
V_1^{\mathrm{br}(\nu),\nu}(s_1, r) - V_1^{\dagger}(s_1, r) \le 2H \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, u/H) &\le \frac{2H}{K}\sum_{k=1}^{K} V_1^{*}(s_1, r^k)\\
&\le \frac{2H}{K}\mathcal{O}\left(\sqrt{H^3K\log(1/\delta')} + \beta\sqrt{H^2K\cdot\Gamma(K,\lambda;\mathrm{ker})}\right)\\
&\le \mathcal{O}\left([\sqrt{H^5\log(1/\delta')} + \beta\sqrt{H^4\cdot\Gamma(K,\lambda;\mathrm{ker})}]/\sqrt{K}\right),
\end{aligned}$$

where the second inequality is due to Lemma 5.28 and the third inequality is by Lemma 5.25.

Since $\Pr(\mathcal{E}\wedge\widetilde{\mathcal{E}}) \ge 1-2\delta'$ by the union bound, by setting $\delta' = 1/(4H^2K^2)$, we obtain that with probability at least $1 - 1/(2H^2K^2)$

$$\begin{aligned}
V_1^{\dagger}(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) &\le \mathcal{O}\left([\sqrt{2H^5\log(2HK)} + \beta\sqrt{H^4\cdot\Gamma(K,\lambda;\mathrm{ker})}]/\sqrt{K}\right),\\
V_1^{\mathrm{br}(\nu),\nu}(s_1, r) - V_1^{\dagger}(s_1, r) &\le \mathcal{O}\left([\sqrt{2H^5\log(2HK)} + \beta\sqrt{H^4\cdot\Gamma(K,\lambda;\mathrm{ker})}]/\sqrt{K}\right),
\end{aligned}$$

such that

$$\begin{aligned}
V_1^{\mathrm{br}(\nu),\nu}(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) &\le \mathcal{O}\left([\sqrt{2H^5\log(2HK)} + \beta\sqrt{H^4\cdot\Gamma(K,\lambda;\mathrm{ker})}]/\sqrt{K}\right)\\
&\le \mathcal{O}\left(\beta\sqrt{H^4[\Gamma(K,\lambda;\mathrm{ker}) + \log(HK)]}/\sqrt{K}\right),
\end{aligned}$$

where the last inequality is due to $\beta \ge H$. The event $\mathcal{E}\wedge\widetilde{\mathcal{E}}$ happens if we further let β satisfy

$$16H^2\big[R_Q^2 + 2\Gamma(K,\lambda;\mathrm{ker}) + 5 + \log\mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 6\log(2HK)\big] \le \beta^2, \forall h\in[H],$$

where $\lambda = 1 + 1/K$, $\widetilde{R}_K = R_K = 2H\sqrt{\Gamma(K,\lambda;\mathrm{ker})}$, and $\varsigma^* = H/K$. This completes the proof. □

5.11 Proofs for Markov Game with Neural Function Approximation

5.11.1 Lemmas

Lemma 5.29 (Lemma C.7 of Yang et al. [2020]). *With $TH^2 = \mathcal{O}(m \log^{-6} m)$, then there exists a constant $\digamma$ such that the following inequalities hold with probability at least $1 - 1/m^2$ for any $z \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and any $W \in \{W : \|W - W^{(0)}\|_2 \leq H\sqrt{K/\lambda}\}$,*

$$|f(z; W) - \varphi(z; W^{(0)})^\top (W - W^{(0)})| \leq \digamma K^{2/3} H^{4/3} m^{-1/6} \sqrt{\log m},$$
$$\|\varphi(z; W) - \varphi(z; W^{(0)})\|_2 \leq \digamma (KH^2/m)^{1/6} \sqrt{\log m}, \qquad \|\varphi(z; W)\|_2 \leq \digamma,$$

with $\digamma \geq 1$.

Lemma 5.30. *We define the event $\mathcal{E}$ as that the following inequality holds $\forall z = (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}, \forall (h, k) \in [H] \times [K]$,*

$$|\mathbb{P}_h V_{h+1}^k(s, a, b) - f_h^k(s, a, b)| \leq u_h^k(s, a, b) + \beta \iota,$$
$$\Big| \|\varphi(z; W_h^k)\|_{(\Lambda_h^k)^{-1}} - \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h^k)^{-1}} \Big| \leq \iota,$$

where $\iota = 5K^{7/12} H^{1/6} m^{-1/12} \log^{1/4} m$ and we define

$$\Lambda_h^k = \sum_{\tau=1}^{k-1} \varphi(z_h^\tau; W_h^k) \varphi(z_h^\tau; W_h^k)^\top + \lambda \cdot I, \quad \widetilde{\Lambda}_h^k = \sum_{\tau=1}^{k-1} \varphi(z_h^\tau; W^{(0)}) \varphi(z_h^\tau; W^{(0)})^\top + \lambda \cdot I.$$

Setting $(1 + 1/H)\beta = B_K$, $R_K = H\sqrt{K}$, $\varsigma^* = H/K$, and $\lambda = \digamma^2(1 + 1/K)$, $\varsigma^* = H/K$, if we let

$$\beta^2 \geq 8R_Q^2 H^2 (1 + \sqrt{\lambda/d})^2 + 32H^2 \Gamma(K, \lambda; \ker_m) + 80H^2$$
$$+ 32H^2 \log \mathcal{N}_\infty(\varsigma^*; R_K, B_K) + 32H^2 \log(K/\delta'),$$

and also

$$m = \Omega(K^{19} H^{14} \log^3 m),$$

then we have that with probability at least $1 - 2/m^2 - \delta'$, the event $\mathcal{E}$ happens, i.e.,

$$\Pr(\mathcal{E}) \geq 1 - 2/m^2 - \delta'.$$

Proof. By letting $\boldsymbol{a} = (a, b)$ be an action in the space $\mathcal{A} \times \mathcal{B}$, Algorithm 8 reduces to Algorithm 6 with the action space $\mathcal{A} \times \mathcal{B}$ and state space $\mathcal{S}$. We have $Q_h^k(s, a, b) = Q_h^k(s, \boldsymbol{a})$, $V_h^k(s, a, b) = V_h^k(s, \boldsymbol{a})$, $u_h^k(s, a, b) = u_h^k(s, \boldsymbol{a})$, $u_h^k(s, a, b) = u_h^k(s, \boldsymbol{a})$ and $r_h^k(s, a, b) = r_h^k(s, \boldsymbol{a})$. Simply applying the proof of Lemma 5.19, we have the proof of this lemma. □

Lemma 5.31. *Conditioned on the event $\mathcal{E}$ defined in Lemma 5.30, with probability at least $1 - \delta'$, we have*

$$\sum_{k=1}^{K} V_1^*(s_1, r^k) \le \sum_{k=1}^{K} V_1^k(s_1) + \beta H K \iota,$$
$$\sum_{k=1}^{K} V_1^k(s_1) \le \mathcal{O}\left(\sqrt{H^3 K \log(1/\delta')} + \beta\sqrt{H^2 K \cdot \Gamma(K, \lambda; \mathrm{ker}_m)}\right) + \beta H K \iota,$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$.

Proof. By the reduction of Algorithm 8 to Algorithm 6, we can apply the same proof for Lemma 5.20, which completes the proof. □

Lemma 5.32. *We define the event $\widetilde{\mathcal{E}}$ as that the following inequality holds $\forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}, \forall h \in [H]$,*

$$|\mathbb{P}_h \overline{V}_{h+1}(s, a, b) - \overline{f}_h(s, a, b)| \le \overline{u}_h(s, a) + \beta\iota,$$
$$|\mathbb{P}_h \underline{V}_{h+1}(s, a, b) - \underline{f}_h(s, a, b)| \le \underline{u}_h(s, a) + \beta\iota,$$
$$\left| \|\varphi(z; \overline{W}_h)\|_{(\overline{\Lambda}_h)^{-1}} - \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}} \right| \le \iota,$$
$$\left| \|\varphi(z; \underline{W}_h)\|_{(\underline{\Lambda}_h)^{-1}} - \|\varphi(z; W^{(0)})\|_{(\widetilde{\Lambda}_h)^{-1}} \right| \le \iota.$$

where $\iota = 5K^{7/12}H^{1/6}m^{-1/12}\log^{1/4} m$, and we define $\overline{f}_h(z) = \Pi_{[0,H]}[f(z; \overline{W}_h)]$ and $\underline{f}_h(z) = \Pi_{[0,H]}[f(z; \underline{W}_h)]$ as well as

$$\overline{\Lambda}_h = \sum_{\tau=1}^{K} \varphi(z_h^\tau; \overline{W}_h)\varphi(z_h^\tau; \overline{W}_h)^\top + \lambda \cdot I, \quad \underline{\Lambda}_h = \sum_{\tau=1}^{K} \varphi(z_h^\tau; \underline{W}_h)\varphi(z_h^\tau; \underline{W}_h)^\top + \lambda \cdot I,$$
$$\widetilde{\Lambda}_h = \sum_{\tau=1}^{K} \varphi(z_h^\tau; W^{(0)})\varphi(z_h^\tau; W^{(0)})^\top + \lambda \cdot I.$$

Setting $\beta = \widetilde{B}_K$, $\widetilde{R}_K = H\sqrt{K}$, $\varsigma^* = H/K$, and $\lambda = \digamma^2(1 + 1/K)$, $\varsigma^* = H/K$, if we set

$$\beta^2 \ge 8R_Q^2 H^2(1 + \sqrt{\lambda/d})^2 + 32H^2 + \Gamma(K, \lambda; \mathrm{ker}_m)$$
$$+ 80H^2 + 32H^2 \log \mathcal{N}_\infty(\varsigma^*; \widetilde{R}_K, \widetilde{B}_K) + 32H^2 \log(2K/\delta'),$$

and also

$$m = \Omega(K^{19} H^{14} \log^3 m),$$

then we have that with probability at least $1 - 2/m^2 - \delta'$*, the event* $\widetilde{\mathcal{E}}$ *happens, i.e.,*

$$\Pr(\widetilde{\mathcal{E}}) \geq 1 - 2/m^2 - \delta'.$$

Proof. The proof of this lemma follows our proof of Lemmas 5.19 and 5.21 and apply some similar ideas from the proof of Lemma 5.26. Particularly, to deal with the upper bounds of the estimation errors of $\mathbb{P}_h \overline{V}_{h+1}$ and $\mathbb{P}_h \underline{V}_{h+1}$, we define the two value function space $\overline{\mathcal{V}}$ and $\underline{\mathcal{V}}$ and show their covering numbers similar to the proof of Lemma 5.26. Then, we further use the proof of Lemma 5.21, which is derived from the proof of Lemma 5.19, to show the eventual results in this lemma. In the proof of this lemma, we set $\widetilde{B}_K = \beta$ instead of $(1 + 1/H)\beta$ due to the structure of the planning phase. This completes the proof. □

Lemma 5.33. *Conditioned on the event* $\widetilde{\mathcal{E}}$ *as defined in Lemma 5.32, we have*

$$\begin{aligned} &V_h^\dagger(s, r) \leq \overline{V}_h(s) + (H + 1 - h)\beta\iota, \forall s \in \mathcal{S}, \forall h \in [H], \\ &\overline{V}_h(s) \leq \mathbb{E}_{a \sim \pi_h, b \sim \mathrm{br}(\pi)_h}[(\mathbb{P}_h \overline{V}_{h+1} + r_h + 2\overline{u}_h)(s, a, b)] + \beta\iota, \forall s \in \mathcal{S}, \forall h \in [H], \end{aligned} \tag{5.45}$$

$$\begin{aligned} &V_h^\dagger(s, r) \geq \underline{V}_h(s) - (H + 1 - h)\beta\iota, \forall s \in \mathcal{S}, \forall h \in [H], \\ &\underline{V}_h(s) \geq \mathbb{E}_{a \sim \mathrm{br}(\nu)_h, b \sim \nu_h}[(\mathbb{P}_h \underline{V}_{h+1} - r_h - 2\underline{u}_h)(s, a, b)] - \beta\iota, \forall s \in \mathcal{S}, \forall h \in [H]. \end{aligned} \tag{5.46}$$

Proof. We prove the first inequality in (5.45) by induction. For $h = H + 1$, we have $V_{H+1}^\dagger(s, r) = \overline{V}_{H+1}(s) = 0$ for any $s \in \mathcal{S}$. Then, we assume that $V_{h+1}^\dagger(s, r) \leq \overline{V}_{h+1}(s) + (H - h)\beta\iota$. Thus, conditioned on the event $\widetilde{\mathcal{E}}$ as defined in Lemma 5.32, we have

$$\begin{aligned} &Q_h^\dagger(s, a, b, r) - \overline{Q}_h(s, a, b) \\ &\qquad = r_h(s, a, b) + \mathbb{P}_h V_{h+1}^\dagger(s, a, b, r) - \min\{[r_h(s, a, b) + \overline{f}_h(s, a, b) + u_h(s, a, b)], H\}^+ \\ &\qquad \leq \max\{[\mathbb{P}_h V_{h+1}^\dagger(s, a, b, r) - \overline{f}_h(s, a, b) - \overline{u}_h(s, a, b)], 0\} \\ &\qquad \leq \max\{[\mathbb{P}_h V_{h+1}(s, a, b) + (H - h)\beta\iota - f_h(s, a, b) - \overline{u}_h(s, a, b)], 0\} \\ &\qquad \leq (H + 1 - h)\beta\iota, \end{aligned}$$

where the first inequality is due to $0 \leq r_h(s, a, b) + \mathbb{P}_h V_{h+1}^\dagger(s, a, b, r) \leq H$ and $\min\{x, y\}^+ \geq \min\{x, y\}$, the second inequality is by the assumption that $V_{h+1}^\dagger(s, a, b, r) \leq \overline{V}_{h+1}(s, a, b) + (H - h)\beta\iota$, the last inequality is by Lemma 5.32 such that $|\mathbb{P}_h \overline{V}_{h+1}(s, a, b) - \overline{f}_h(s, a, b)| \leq \overline{u}_h(s, a, b) +$

$\beta\iota$ holds for any $(s,a,b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and $(k,h) \in [K] \times [H]$. The above inequality leads to

$$\begin{aligned}
V_h^\dagger(s,r) &= \max_{\pi'_h} \min_{\nu'_h} \mathbb{E}_{a\sim\pi'_h, b\sim\nu'_h}[Q_h^\dagger(s,a,b,r)] \\
&\leq \max_{\pi'_h} \min_{\nu'_h} \mathbb{E}_{a\sim\pi'_h, b\sim\nu'_h}[\overline{Q}_h(s,a,b)] + (H+1-h)\beta\iota \\
&= \overline{V}_h(s) + (H+1-h)\beta\iota.
\end{aligned}$$

Therefore, we have

$$V_h^\dagger(s,r) \leq \overline{V}_h(s) + (H+1-h)\beta\iota, \forall h \in [H], \forall s \in \mathcal{S}.$$

We further prove the second inequality in (5.45). We have

$$\begin{aligned}
\overline{Q}_h(s,a,b) &= \min\{[r_h(s,a,b) + \overline{f}_h(s,a,b) + \overline{u}_h(s,a,b)], H\}^+ \\
&\leq \min\{[r_h(s,a,b) + \mathbb{P}_h\overline{V}_{h+1}(s,a,b) + 2\overline{u}_h(s,a,b) + \beta\iota], H\}^+ \\
&\leq r_h(s,a,b) + \mathbb{P}_h\overline{V}_{h+1}(s,a,b) + 2\overline{u}_h(s,a,b) + \beta\iota,
\end{aligned}$$

where the first inequality is also by Lemma 5.32 such that $|\mathbb{P}_h\overline{V}_{h+1}(s,a,b) - \overline{f}_h(s,a,b)| \leq \overline{u}_h(s,a,b) + \beta\iota$, and the last inequality is because of the non-negativity of $r_h(s,a,b) + \mathbb{P}_h V_{h+1}(s,a,b) + 2\overline{u}_h(s,a,b) + \beta\iota$. Therefore, we have

$$\begin{aligned}
\overline{V}_h(s) &= \min_{\nu'} \mathbb{E}_{a\sim\pi_h, b\sim\nu'}\overline{Q}_h(s,a,b) \\
&\leq \mathbb{E}_{a\sim\pi_h, b\sim\mathrm{br}(\pi)_h}\overline{Q}_h(s,a,b) \\
&\leq \mathbb{E}_{a\sim\pi_h, b\sim\mathrm{br}(\pi)_h}[r_h(s,a,b) + \mathbb{P}_h\overline{V}_{h+1}(s,a,b) + 2\overline{u}_h(s,a,b)] + \beta\iota.
\end{aligned}$$

For the inequalities in (5.46), we can prove them in the same way to proving (5.45). From the perspective of Player 2, this player is trying to find a policy to maximize the cumulative rewards w.r.t. a reward function $\{-r_h(s,a,b)\}_{h\in[H]}$. Thus, one can further use the proof technique for (5.45) to prove (5.46). This completes the proof. □

Lemma 5.34. *With the exploration and planning phases, conditioned on the event $\mathcal{E}$ defined in Lemma 5.30 and the event $\widetilde{\mathcal{E}}$ defined in Lemma 5.32, we have the following inequalities*

$$\begin{aligned}
K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, \overline{u}/H) &\leq \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota, \\
K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, \underline{u}/H) &\leq \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota.
\end{aligned}$$

Proof. First, we have $K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, \overline{u}/H) \le K \cdot V_1^*(s_1, \overline{u}/H)$ as well as $K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, \underline{u}/H) \le K \cdot V_1^*(s_1, \underline{u}/H)$ according to the definition of V_1^*. Thus, to prove this lemma, we only need to show

$$K \cdot V_1^*(s_1, \overline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota,$$
$$K \cdot V_1^*(s_1, \underline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota.$$

Because the constructions of the planning bonus $\overline{u}_h$ and the exploration reward r_h^k are the same as the ones for the single-agent case, similar to the proof of Lemma 5.23, and according to Lemmas 5.30 and 5.32, we have the following results

$$\overline{u}_h(s,a,b) \le H \cdot r_h^k(s,a,b) + 2\beta\iota, \quad \underline{u}_h(s,a,b) \le H \cdot r_h^k(s,a,b) + 2\beta\iota$$

such that

$$V_1^*(s_1, \overline{u}/H) \le V_1^*(s_1, r^k) + 2\beta\iota, \quad V_1^*(s_1, \underline{u}/H) \le V_1^*(s_1, r^k) + 2\beta\iota,$$

Therefore, we eventually obtain

$$K \cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, \overline{u}/H) \le K \cdot V_1^*(s_1, \overline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota,$$
$$K \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1, \underline{u}/H) \le K \cdot V_1^*(s_1, \underline{u}/H) \le \sum_{k=1}^K V_1^*(s_1, r^k) + 2K\beta\iota.$$

This completes the proof. □

5.11.2 Proof of Theorem 5.7

Proof. Conditioned on the events $\mathcal{E}$ and $\widetilde{\mathcal{E}}$ defined in Lemmas 5.30 and 5.32, we have

$$V_1^\dagger(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) \le \overline{V}_1(s_1) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) + H\beta\iota, \tag{5.47}$$

where the inequality is by Lemma 5.33. Further by this lemma, we have

$$\begin{aligned}
&\overline{V}_h(s_h) - V_h^{\pi,\mathrm{br}(\pi)}(s_h, r) \\
&\quad\le \mathbb{E}_{a_h\sim\pi_h, b_h\sim\mathrm{br}(\pi)_h}[(\mathbb{P}_h\overline{V}_{h+1} + r_h + 2\overline{u}_h)(s_h, a_h, b_h)] - V_h^{\pi,\mathrm{br}(\pi)}(s_h, r) + \beta\iota \\
&\quad= \mathbb{E}_{a_h\sim\pi_h, b_h\sim\mathrm{br}(\pi)_h}[(r_h + \mathbb{P}_h\overline{V}_{h+1} + 2\overline{u}_h)(s_h, a_h, b_h) - r_h(s_h, a_h, b_h) \\
&\qquad - \mathbb{P}_h V_{h+1}^{\pi,\mathrm{br}(\pi)}(s_h, a_h, b_h, r)] + \beta\iota \\
&\quad= \mathbb{E}_{a_h\sim\pi_h, b_h\sim\mathrm{br}(\pi)_h}[\mathbb{P}_h\overline{V}_{h+1}(s_h, a_h, b_h) - \mathbb{P}_h V_{h+1}^{\pi,\mathrm{br}(\pi)}(s_h, a_h, b_h, r) + 2\overline{u}_h(s_h, a_h, b_h)] + \beta\iota \\
&\quad= \mathbb{E}_{a_h\sim\pi_h, b_h\sim\mathrm{br}(\pi)_h, s_{h+1}\sim\mathbb{P}_h}[\overline{V}_{h+1}(s_{h+1}) - V_{h+1}^{\pi,\mathrm{br}(\pi)}(s_{h+1}, r) + 2\overline{u}_h(s_h, a_h, b_h)] + \beta\iota.
\end{aligned}$$

Recursively applying the above inequality and making use of $\overline{V}_{H+1}(s, r) = V_{H+1}^{\pi,\mathrm{br}(\pi)}(s) = 0$ gives

$$\begin{aligned}
\overline{V}_1(s_1) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) &\le \mathbb{E}_{\forall h\in[H]:\, a_h\sim\pi_h, b_h\sim\mathrm{br}(\pi)_h, s_{h+1}\sim\mathbb{P}_h}\left[\sum_{h=1}^{H} 2\overline{u}_h(s_h, a_h, b_h)\middle| s_1\right] \\
&= 2H\cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, \overline{u}/H) + H\beta\iota.
\end{aligned}$$

Combining with (5.47) gives

$$\begin{aligned}
&V_1^{\dagger}(s_1, r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1, r) \\
&\quad\le 2H\cdot V_1^{\pi,\mathrm{br}(\pi)}(s_1, \overline{u}/H) + 2H\beta\iota \le \frac{2H}{K}\sum_{k=1}^{K} V_1^{*}(s_1, r^k) + 4H\beta\iota \\
&\quad\le \frac{2H}{K}\mathcal{O}\left(\sqrt{H^3K\log(1/\delta')} + \beta\sqrt{H^2K\cdot\Gamma(K,\lambda;\ker_m)}\right) + (H+4)H\beta\iota \\
&\quad\le \mathcal{O}\left([\sqrt{H^5\log(1/\delta')} + \beta\sqrt{H^4\cdot\Gamma(K,\lambda;\ker_m)}]/\sqrt{K} + H^2\beta\iota\right),
\end{aligned}$$

where the second inequality is due to Lemma 5.34 and the third inequality is by Lemma 5.31.

Next, we give the upper bound of $V_1^{\mathrm{br}(\nu),\nu}(s_1, r) - V_1^{\dagger}(s_1, r)$. Conditioned on the event $\mathcal{E}$ defined in Lemma 5.30 and the event $\widetilde{\mathcal{E}}$ defined in Lemma 5.32, we have

$$V_1^{\mathrm{br}(\nu),\nu}(s_1, r) - V_1^{\dagger}(s_1, r) \le V_1^{\mathrm{br}(\nu),\nu}(s_1, r) - \underline{V}_1(s_1) + H\beta\iota, \tag{5.48}$$

where the inequality is by Lemma 5.33. Further by this lemma, we have

$$\begin{aligned}
&V_h^{\mathrm{br}(\nu),\nu}(s_h, r) - \underline{V}_h(s_h) \\
&\quad\le V_h^{\mathrm{br}(\nu),\nu}(s_h, r) - \mathbb{E}_{a\sim\mathrm{br}(\nu)_h, b\sim\nu_h}[(\mathbb{P}_h\underline{V}_{h+1} - r_h - 2\underline{u}_h)(s_h, a_h, b_h)] + \beta\iota \\
&\quad= \mathbb{E}_{a_h\sim\mathrm{br}(\nu)_h, b_h\sim\nu_h}[\mathbb{P}_h V_{h+1}^{\mathrm{br}(\nu),\nu}(s_h, a_h, b_h, r) - \mathbb{P}_h\underline{V}_{h+1}(s_h, a_h, b_h) + 2\underline{u}_h(s_h, a_h, b_h)] + \beta\iota \\
&\quad= \mathbb{E}_{a_h\sim\mathrm{br}(\nu)_h, b_h\sim\nu_h, s_{h+1}\sim\mathbb{P}_h}[V_{h+1}^{\mathrm{br}(\nu),\nu}(s_{h+1}, r) - \mathbb{P}_h\underline{V}_{h+1}(s_{h+1}) + 2\underline{u}_h(s_h, a_h, b_h)] + \beta\iota.
\end{aligned}$$

Recursively applying the above inequality gives

$$V_1^{\mathrm{br}(\nu),\nu}(s_1,r) - \underline{V}_1(s_1) \le 2H \cdot V_1^{\mathrm{br}(\nu),\nu}(s_1,\underline{u}/H) + H\beta\iota.$$

Combining with (5.48) gives

$$\begin{aligned} V_1^{\mathrm{br}(\nu),\nu}(s_1,r) - V_1^{\dagger}(s_1,r) &\le 2H \cdot V_1^{\pi}(s_1,\underline{u}/H) + 2H\beta\iota \le \frac{2H}{K}\sum_{k=1}^{K} V_1^{*}(s_1,r^k) + 4H\beta\iota \\ &\le \mathcal{O}\left([\sqrt{H^5 \log(1/\delta')} + \beta\sqrt{H^4 \cdot \Gamma(K,\lambda;\mathrm{ker}_m)}]/\sqrt{K} + H^2\beta\iota \right), \end{aligned}$$

where the second inequality is due to Lemma 5.34 and the last inequality is by Lemma 5.31. Thus, we eventually have

$$\begin{aligned} &V_1^{\mathrm{br}(\nu),\nu}(s_1,r) - V_1^{\pi,\mathrm{br}(\pi)}(s_1,r) \\ &\quad \le \mathcal{O}\left(\left[\sqrt{H^5 \log(1/\delta')} + \beta\sqrt{H^4 \cdot \Gamma(K,\lambda;\mathrm{ker}_m)}\right]/\sqrt{K} + H^2\beta\iota \right). \end{aligned}$$

Moreover, we also have $P(\mathcal{E} \wedge \widetilde{\mathcal{E}}) \ge 1 - 2\delta' - 4/m^2$ by the union bound. Therefore, since $\beta \ge H$ as shown in Lemmas 5.30 and 5.32, setting $\delta' = 1/(4K^2H^2)$, we obtain that with probability at least $1 - 1/(2K^2H^2) - 4/m^2$,

$$V_1^{*}(s_1,r) - V_1^{\pi}(s_1,r) \le \mathcal{O}\left(\beta\sqrt{H^4[\Gamma(K,\lambda;\mathrm{ker}_m) + \log(KH)]}/\sqrt{K} + H^2\beta\iota \right).$$

The event $\mathcal{E} \wedge \widetilde{\mathcal{E}}$ happens if we further let β satisfy

$$\begin{aligned} \beta^2 \ge\ & 8R_Q^2H^2(1+\sqrt{\lambda/d})^2 + 32H^2\Gamma(K,\lambda;\mathrm{ker}_m) + 80H^2 \\ & + 32H^2 \log \mathcal{N}_\infty(\varsigma^*; R_K, 2\beta) + 96H^2\log(2KH). \end{aligned}$$

where guarantees the conditions in Lemmas 5.30 and 5.32 hold. This completes the proof. □

5.12 Other Supporting Lemmas

Lemma 5.35 (Lemma E.2 of Yang et al. [2020]). *Let $\{s_\tau\}_{\tau=1}^{\infty}$ and $\{\phi_\tau\}_{\tau=1}^{\infty}$ be $\mathcal{S}$-valued and $\mathcal{H}$-valued stochastic processes adapted to filtration $\{\mathcal{F}_\tau\}_{\tau=0}^{\infty}$, respectively, where we assume that $\|\phi_\tau\| \le 1$ for all $\tau \ge 1$. Moreover, for any $t \ge 1$, we let $\mathcal{K}_t \in \mathbb{R}^{t\times t}$ be the Gram matrix of $\{\phi_\tau\}_{\tau\in[t]}$ and define an operator $\Lambda_t : \mathcal{H} \mapsto \mathcal{H}$ as $\Lambda_t = \lambda I + \sum_{\tau=1}^{t} \phi_\tau\phi_\tau^\top$ with $\lambda > 1$. Let $\mathcal{V} \subseteq \{V : \mathcal{S} \mapsto [0,H]\}$ be a class of bounded functions on $\mathcal{S}$. Then for any $\delta \in (0,1)$, with* *probability at least* $1 - \delta$*, we have simultaneously for all* $t \geq 1$ *that*

$$\sup_{V \in \mathcal{V}} \left\| \sum_{\tau=1}^{t} \phi_\tau \{V(s_\tau) - \mathbb{E}[V(s_\tau)|\mathcal{F}_{\tau-1}]\} \right\|_{\Lambda_t^{-1}}^2$$
$$\leq 2H^2 \log\det(I + \mathcal{K}_t/\lambda) + 2H^2 t(\lambda - 1) + 4H^2 \log(\mathcal{N}_\epsilon/\delta) + 8t^2\epsilon^2/\lambda,$$

where $\mathcal{N}_\epsilon$ *is the* ϵ*-covering number of* $\mathcal{V}$ *with respect to the distance* $\mathrm{dist}(\cdot,\cdot) := \sup_{\mathcal{S}} |V_1(s) - V_2(s)|$.

Lemma 5.36 (Lemma E.3 of Yang et al. [2020])**.** *Let* $\{\phi_t\}_{t\geq 1}$ *be a sequence in the RKHS* $\mathcal{H}$*. Let* $\Lambda_0 : \mathcal{H} \mapsto \mathcal{H}$ *be defined as* λI *where* $\lambda \geq 1$ *and* I *is the identity mapping on* $\mathcal{H}$*. For any* $t \geq 1$*, we define a self-adjoint and positive-definite operator* Λ_t *by letting* $\Lambda_t = \Lambda_0 + \sum_{j=1}^{t} \phi_j \phi_j^\top$*. Then, for any* $t \geq 1$*, we have*

$$\sum_{j=1}^{t} \min\{1, \phi_j \Lambda_{j-1}^{-1} \phi_j^\top\} \leq 2 \log\det(I + \mathcal{K}_t/\lambda),$$

where $\mathcal{K}_t \in \mathbb{R}^{t\times t}$ *is the Gram matrix obtained from* $\{\phi_j\}_{j\in[t]}$*, i.e., for any* $j, j' \in [t]$*, the* (j, j')*-th entry of* $\mathcal{K}_t$ *is* $\langle \phi_j, \phi_j \rangle_{\mathcal{H}}$*. Moreover, if we further have* $\sup_{t\geq 0}\{\|\phi_t\|_{\mathcal{H}}\} \leq 1$*, then it holds that*

$$\log\det(I + \mathcal{K}_t/\lambda) \leq \sum_{j=1}^{t} \phi_j^\top \Lambda_{j-1}^{-1} \phi_j \leq 2\log\det(I + \mathcal{K}_t/\lambda).$$

CHAPTER 6

Conclusion

Due to the huge empirical successes of RL in solving real-world decision-making problems, there have been a large number of works studying the theoretical understandings of the RL algorithms. Along such a research direction, this thesis focuses on two classes of RL methods, i.e., reward-based online RL and reward-free RL, for both single-agent MDPs and Markov games. For the reward-based online RL, this thesis investigates two concrete problems, namely online learning for constrained MDPs and policy optimization for two-player zero-sum Markov games. Moreover, within the framework of reward-free RL, this thesis proposes and analyzes novel algorithms for both single-agent MDPs and Markov games incorporating the powerful nonlinear function approximations. Specifically, the main contributions of this thesis are concluded as follows:

Online Learning for Constrained MDPs. Chapter 3 proposes a new upper confidence primal-dual algorithm for constrained MDP online learning problems. The proposed algorithm estimates the unknown transition model based on the trajectories and maintains a confidence set inspired by the idea of UCB. It incorporates the confidence set into the online primal-dual method for learning the policies. The proposed algorithm is proved to achieve $\widetilde{\mathcal{O}}(\sqrt{K})$ upper bounds for the regret and the constraint violation simultaneously. Moreover, the regret bound nearly matches the lower bound of the regret for learning MDPs. The analysis incorporates a new high-probability drift analysis of Lagrange multiplier processes into the regret and constraint violation proofs for the proposed upper confidence algorithm.

Policy Optimization for Zero-Sum Markov Games. Chapter 4 proposes and analyzes new provable optimistic PO algorithms for two-player zero-sum Markov games with two non-trivial special transition structures, namely the factored independent transition and the single-controller transition. The proposed algorithms feature a combination of UCB-type optimism and policy optimization updating rules adapted to the structured transitions in a multi-agent non-stationary environment. In order to handle the non-stationarity resulting from the opponent's varying state, both players in the factored independent transition setting and Player 2 in the single-controller setting demand to make an estimation of the opponent's state reaching probability. For both transition structures, this thesis provides $\widetilde{\mathcal{O}}(\sqrt{K})$ regret bounds after K episodes. The $\widetilde{\mathcal{O}}(\sqrt{K})$ regret bounds in this thesis match the regrets of the value-based methods when translating their results in terms of the regret definition here. This thesis also proposes novel value difference decomposition by taking the transition structures and the state reaching probability estimation error into consideration.

Reward-Free RL with Kernel and Neural Function Approximations. Chapter 5 first proposes sample- and computationally efficient reward-free RL algorithms with kernel and neural function approximations for single-agent MDPs. The proposed exploration algorithm is an optimistic variant of the least-square value iteration algorithm incorporating kernel and neural function approximators. Further with the planning phase, which is a single-episode optimistic value iteration algorithm, the proposed method achieves an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity to generate an ε-suboptimal policy for an arbitrary extrinsic reward function. Moreover, this thesis extends the proposed method from the single-agent scenario to the two-player zero-sum Markov games, which can achieve an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity to generate an ε-approximate Nash equilibrium. Particularly, in the planning phase for Markov games, the proposed algorithm only involves finding the Nash equilibrium of matrix games formed by Q-function that can be solved efficiently, which is of independent interest. The above sample complexities match the $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ results in existing works for tabular or linear function approximation settings.

6.1 Future Direction

There are plenty of topics remaining to explore in the area of theoretical RL. This section provides several lines of future research directions specific to the main problems studied in this thesis.

Extension to General Function Approximations. The problems of online learning for constrained MDPs and policy optimization for Markov games are only investigated in the tabular case. However, when the state and action spaces are large, it is necessary to analyze the function approximation scenario, especially the general (nonlinear) function approximation. On the other hand, this thesis studies two specific nonlinear function approximators, i.e., kernel function and 1-layer neural network, for the reward-free RL problem. It is challenging to further investigate the reward-free RL with a general nonlinear function approximator or a multi-layer neural network approximator beyond the neural tangent kernel modeling.

Extension to General Multi-Agent Scenarios. While this thesis studies constrained single-agent MDPs, it is interesting to see the exploration of new provable algorithms for multi-agent RL with constraints, e.g., constrained Markov games. In addition, this thesis analyzes the policy optimization algorithms for two-player zero-sum Markov games with special transition structures. However, whether the policy optimization for Markov games with a general transition can attain an

$\mathcal{O}(\sqrt{K})$ remains a challenging problem. Moreover, it is worthwhile to investigate how to extend such analysis to multi-player general-sum games. For the reward-free RL, it is appealing to study the extension of the analysis for two-player Markov games based on a joint environment exploration to the multi-player game scenario where each player can explore the environment separately.

Toward Tighter Bounds. Recently, there have been a lot of works investigating how to sharpen the upper bounds of the regrets or the sample complexities such that one can obtain tighter results matching the lower bounds. It has been shown that employing the bonus terms based on Bernstein's inequality instead of Hoeffding's inequality could lead to a better dependence on the episode length H and the sizes of the action/state spaces $|\mathcal{S}|, |\mathcal{A}|, |\mathcal{B}|$ or the feature dimension for the online value-based RL algorithms in the tabular case or with linear function approximation. Thus, it is an interesting research question that whether we can adopt such a technique to the constrained RL, policy optimization methods, and reward-free RL or even these three aspects generally with nonlinear function approximations under the multi-agent setting.

BIBLIOGRAPHY

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702, 2019a.

Yasin Abbasi-Yadkori, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. Exploration-enhanced politex. *arXiv preprint arXiv:1908.10479*, 2019b.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 22–31. JMLR. org, 2017.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. *arXiv preprint arXiv:1908.00261*, 2019.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*, 2018.

Eitan Altman. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Eitan Altman, Konstantin Avrachenkov, Richard Marquez, and Gregory Miller. Zero-sum constrained stochastic games with independent state processes. *Mathematical Methods of Operations Research*, 62(3):375–386, 2005.

Eitan Altman, Konstantin Avrachenkov, Nicolas Bonneau, Merouane Debbah, Rachid El-Azouzi, and Daniel Sadoc Menasche. Constrained cost-coupled stochastic games with independent state processes. *Operations Research Letters*, 36(2):160–164, 2008.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pages 551–560. PMLR, 2020.

Yu Bai and Jason D Lee. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007*, 2020.

Jonathan Baxter and Peter L Bartlett. Direct gradient-based reinforcement learning. In *2000 IEEE International Symposium on Circuits and Systems. Emerging Technologies for the 21st Century. Proceedings (IEEE Cat No. 00CH36353)*, volume 3, pages 271–274. IEEE, 2000.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Dimitri P Bertsekas. *Convex optimization theory*. Athena Scientific Belmont, 2009.

Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

Lucian Bu, Robert Babu, Bart De Schutter, et al. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *arXiv preprint arXiv:2101.04233*, 2021.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*, 2019.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pages 1329–1338. PMLR, 2016.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pages 12203–12213, 2019.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020a.

Yonathan Efroni, Lior Shani, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020b.

AbdelRahman Eldosouky, Walid Saad, and Dusit Niyato. Single controller stochastic games for optimized moving target defense. In *2016 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2016.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. *arXiv preprint arXiv:1801.05039*, 2018.

Jerzy A Filar and TES Raghavan. A matrix game solution of the single-controller stochastic game. *Mathematics of Operations Research*, 9(3):356–362, 1984.

János Flesch, Gijs Schoenmakers, and Koos Vrieze. Stochastic games on a product state space. *Mathematics of Operations Research*, 33(2):403–420, 2008.

Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in neural information processing systems*, pages 2137–2145, 2016.

Bennett Fox. Markov renewal programming by linear fractional programming. *SIAM Journal on Applied Mathematics*, 14(6):1418–1432, 1966.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1578–1586. PMLR, 2018.

Drew Fudenberg and David Levine. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 1995.

Ruiqi Gao, Tianle Cai, Haochuan Li, Liwei Wang, Cho-Jui Hsieh, and Jason D Lee. Convergence of adversarial training in overparametrized neural networks. *arXiv preprint arXiv:1906.07916*, 2019.

Marek Grzes. Reward shaping in episodic reinforcement learning. 2017.

Peng Guan, Maxim Raginsky, Rebecca Willett, and Daphney-Stavroula Zois. Regret minimization algorithms for single-controller zero-sum stochastic games. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 7075–7080. IEEE, 2016.

András György, Tamás Linder, Gábor Lugosi, and György Ottucsák. The on-line shortest path problem under partial monitoring. *Journal of Machine Learning Research*, 8(Oct):2369–2403, 2007.

Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *International Conference on Machine Learning*, pages 805–813, 2015.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition. *arXiv preprint arXiv:1912.01192*, 2019.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pages 4870–4879. PMLR, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020b.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. *arXiv preprint arXiv:2006.06294*, 2020.

Daphne Koller, Nimrod Megiddo, and Bernhard Von Stengel. Fast algorithms for finding randomized strategies in game trees. In *Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 750–759, 1994.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

Erwin Kreyszig. *Introductory functional analysis with applications*, volume 1. wiley New York, 1978.

Adam Daniel Laud. Theory and application of reward shaping in reinforcement learning. Technical report, 2004.

Hoang M Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*, 2017.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, 2019.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

Barbara MacCluer. *Elementary functional analysis*, volume 253. Springer Science & Business Media, 2008.

Shie Mannor, John N Tsitsiklis, and Jia Yuan Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(Mar):569–590, 2009.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. *arXiv preprint arXiv:2007.13442*, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Angelia Nedić and Asuman Ozdaglar. Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization*, 19(4):1757–1780, 2009.

Michael J Neely. Dynamic optimization and learning for renewal systems. *IEEE Transactions on Automatic Control*, 58(1):32–46, 2012.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009.

Gergely Neu, András György, and Csaba Szepesvári. The online loop-free stochastic shortest-path problem. In *COLT*, volume 2010, pages 231–243. Citeseer, 2010.

Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813. PMLR, 2012.

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*, 2018.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, pages 1333–1342, 2017.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1806.05618*, 2018.

Thiruvenkatachari Parthasarathy and TES Raghavan. An orderfield property for stochastic games when one player controls transition probabilities. *Journal of Optimization Theory and Applications*, 33(3):375–392, 1981.

Julien Perolat, Bilal Piot, and Olivier Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, pages 919–928. PMLR, 2018.

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From poincar\'e recurrence to convergence in imperfect information games: Finding equilibrium via regularization. *arXiv preprint arXiv:2002.08456*, 2020.

Shuang Qiu, Xiaohan Wei, Zhuoran Yang, Jieping Ye, and Zhaoran Wang. Upper confidence primal-dual reinforcement learning for cmdp with adversarial loss. In *Advances in Neural Information Processing Systems*, 2020.

Shuang Qiu, Xiaohan Wei, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. Provably efficient fictitious play policy optimization for zero-sum markov games with structured transitions. In *International Conference on Machine Learning*, pages 8715–8725. PMLR, 2021a.

Shuang Qiu, Jieping Ye, Zhaoran Wang, and Zhuoran Yang. On reward-free rl with kernel and neural function approximations: Single-agent mdp and markov game. In *International Conference on Machine Learning*, pages 8737–8747. PMLR, 2021b.

Julia Robinson. An iterative method of solving a game. *Annals of mathematics*, pages 296–301, 1951.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. *arXiv preprint arXiv:1905.07773*, 2019a.

Aviv Rosenberg and Yishay Mansour. Online stochastic shortest path with bandit feedback and unknown transition function. In *Advances in Neural Information Processing Systems*, pages 2209–2218, 2019b.

Dinah Rosenberg, Eilon Solan, and Nicolas Vieille. Stochastic games with a single controller and incomplete information. *SIAM journal on control and optimization*, 43(1):86–110, 2004.

Martin Schechter. *Principles of functional analysis*. Number 36. American Mathematical Soc., 2001.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Vikas Vikram Singh and N Hemachandra. A characterization of stationary nash equilibria of constrained stochastic games with independent state processes. *Operations Research Letters*, 42 (1):48–52, 2014.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Provably efficient online agnostic learning in markov games. *arXiv preprint arXiv:2010.15020*, 2020.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 1, 2008.

Rahul Urgaonkar, Shiqiang Wang, Ting He, Murtaza Zafer, Kevin Chan, and Kin K Leung. Dynamic service migration and workload scheduling in edge-clouds. *Performance Evaluation*, 91: 205–228, 2015.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019a.

Ruosong Wang, Simon S Du, Lin F Yang, and Ruslan Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020a.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F Yang. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020b.

Shiqiang Wang, Rahul Urgaonkar, Murtaza Zafer, Ting He, Kevin Chan, and Kin K Leung. Dynamic service migration in mobile edge-clouds. In *2015 IFIP Networking Conference (IFIP Networking)*, pages 1–9. IEEE, 2015.

William Yang Wang, Jiwei Li, and Xiaodong He. Deep reinforcement learning for nlp. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 19–21, 2018.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019b.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pages 4987–4997, 2017.

Chen-Yu Wei, Mehdi Jafarnia Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR, 2020.

Xiaohan Wei, Hao Yu, and Michael J Neely. Online learning in weakly coupled markov decision processes: A convergence time study. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1):12, 2018.

Xiaohan Wei, Hao Yu, and Michael J Neely. Online primal-dual mirror descent under stochastic constraints. *arXiv preprint arXiv:1908.00305*, 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. *arXiv preprint arXiv:2002.07066*, 2020.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael Jordan. Provably efficient reinforcement learning with kernel and neural function approximations. *Advances in Neural Information Processing Systems*, 33, 2020.

Hao Yu, Michael Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, pages 1428–1438, 2017.

Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov decision processes with arbitrary reward processes. *Mathematics of Operations Research*, 34(3):737–757, 2009.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel J Kochenderfer, and Emma Brunskill. Provably efficient reward-agnostic navigation with linear value iteration. *arXiv preprint arXiv:2008.07737*, 2020b.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019a.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, pages 11598–11610, 2019b.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, pages 2823–2832, 2019.

Zihan Zhang, Simon S Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020.

Liyuan Zheng and Lillian J Ratliff. Constrained upper confidence reinforcement learning. *arXiv preprint arXiv:2001.09377*, 2020.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. *arXiv preprint arXiv:2006.13165*, 2020.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Advances in neural information processing systems*, pages 1583–1591, 2013.