CAN REINFORCEMENT LEARNING EFFICIENTLY FIND STACKELBERG-NASH EQUILIBRIA IN GENERAL-SUM MARKOV GAMES WITH MYOPIC FOLLOWERS?

Han Zhong
Peking University
hanzhong@stu.pku.edu.cn

Zhuoran Yang
Princeton University
zy6@princeton.edu

Zhaoran Wang
Northwestern University
zhaoranwang@gmail.com

Michael I. Jordan
UC Berkeley
jordan@cs.berkeley.edu

ABSTRACT

We study multi-player general-sum Markov games with one of the players designated as the leader and the rest regarded as the followers. In particular, we focus on the class of games where the followers are myopic, i.e., the followers aim to maximize the instantaneous rewards. For such a game, our goal is to find the Stackelberg-Nash equilibrium (SNE), which is a policy pair (π^*, ν^*) such that (i) π^* is the optimal policy for the leader when the followers always play their best response, and (ii) ν^* is the best response policy of the followers, which is a Nash equilibrium of the followers' game induced by π^*. We develop sample efficient reinforcement learning (RL) algorithms for solving SNE under both the online and offline settings. Respectively, our algorithms are optimistic and pessimistic variants of least-squares value iteration and are readily able to incorporate function approximation tools for handling large state spaces. Furthermore, for the case with linear function approximation, we prove that our algorithms achieve sublinear regret and suboptimality under online and offline setups respectively. To our best knowledge, we establish the first provably efficient RL algorithms for solving SNE in general-sum Markov games with myopic followers.

1 INTRODUCTION

Reinforcement learning (RL) has achieved striking empirical successes in solving complicated real-world sequential decision-making problems (Mnih et al., 2015; Duan et al., 2016; Silver et al., 2016; 2017; 2018; Agostinelli et al., 2019; Akkaya et al., 2019). Motivated by these successes, multi-agent extensions of RL algorithms recently have gained great popularity in decision-making problems involving multiple interacting agents (Busoniu et al., 2008; Hernandez-Leal et al., 2018; 2019; OroojlooyJadid & Hajinezhad, 2019; Zhang et al., 2019). Multi-agent RL is often modeled as a Markov game (Littman, 1994) where, at each time step, each player (agent) takes an action simultaneously at each state of the environment, observe her own immediate reward, and the environment evolves into a next state. Here both the reward of each player and the state transition depend on the actions of all players. From the perspective of each player, her goal is to find a policy that maximizes her expected total reward in the presence of other agents.

In Markov games, depending on the structure of the reward functions, the relationship among the players can be either collaborative, where each player has the same reward function, or competitive, where the sum of the reward function is equal to zero, or mixed, which corresponds to a general-sum game. While most of the existing theoretical results focus on the collaborative or two-player competitive settings, the mixed setting is oftentimes more pertinent to real-world multi-agent applications.

Moreover, in addition to having diverse reward functions, the players might also have asymmetric roles in the Markov game — the players might be divided into leaders and followers, where the leaders' joint policy determines a general-sum game for the followers. Games with such a leader-follower structure is popular in applications such as mechanism design (Conitzer & Sandholm, 2002; Roughgarden, 2004; Garg & Narahari, 2005; Kang & Wu, 2014), security games (Tambe, 2011; Korzhyk et al., 2011), incentive design (Zheng et al., 1984; Ratliff et al., 2014; Chen et al., 2016; Ratliff & Fiez, 2020), and model-based RL (Rajeswaran et al., 2020). Consider a simplified economic system that consists of a government and a group of companies, where the companies purchase or sell goods, and the government collects taxes from transactions. Such a problem can be viewed as a multi-player general-sum game, where the government serves as the leader and the companies are followers (Zheng et al., 2020). In particular, when the government sets a tax rate, the companies form a general-sum game themselves, whose reward functions depend on the tax rate. Each company aims to maximize their own revenue, and thus ideally they achieve a Nash equilibrium (NE) of the induced game. Whereas the goal of the government might be achieving the social welfare, which can be measured via certain fairness metrics computed by the revenues of the companies.

In multi-player Markov games with such a leader-follower structure, the desired solution concept is the Stackelberg-Nash equilibrium (SNE) (Başar & Olsder, 1998). In the setting where there is a single leader, SNE corresponds to a pair of leader's policy π^* and followers' joint policy ν^* that satisfies the following two properties: (i) when the leader adopts π^*, ν^* is the best-response policy of the followers, i.e., ν^* is a Nash equilibrium of the followers' subgame induced by π^*; and (ii) π^* is the optimal policy of the leader assuming the followers always adopt the best response.

We are interested in finding an SNE in a multi-player Markov game when the reward functions and Markov transition kernel are unknown. In particular, we focus on the setting with a single leader and multiple *myopic* followers. That is, the followers at any step of the game do not take into account the future rewards, but only the rewards in the current step. The formal definition of myopic followers is given in §2.3. This setting is a natural formalization of many real-world problems such as marketing and supply chain management. For example, in a market, the leader is an established firm and the followers are entrants. The entrants are not sure whether the firm is going to exist in the future, so they might just want to maximize instantaneous rewards. See Li & Sethi (2017); Kańska & Wiszniewska-Matyszkiel (2021) and references therein for more examples. For such a game, we are interested in the following question:

> Can we develop sample efficient reinforcement learning methods that provably find Stackelberg-Nash equilibria in general-sum Markov games with myopic followers?

To this end, we consider both online and offline RL settings, where in the former, we learn the SNE in a trial-an-error fashion by interacting with the environment and generating data, and in the latter, we learn the SNE from a given dataset that is collected a priori. For the online setting, as the transition model is unknown, to achieve sample efficiency, the equilibrium-finding algorithm also needs to take the exploration-exploitation tradeoff into consideration. Although the similar challenge has been studied in zero-sum Markov game, it seems unclear how to incorporate popular exploration mechanisms such as optimism in the face of uncertainty (Sutton & Barto, 2018) into SNE finding. Meanwhile, under the offline setting, as the RL agent has no control of data collection, it is ideal to design an RL algorithm with theoretical guarantees for an arbitrary dataset that might not be sufficiently explorative.

Our contributions Our contributions are three-fold. First, for the episodic general-sum Markov game with myopic followers, under the online and offline settings respectively, we propose optimistic and pessimistic variants of the least-squares value iteration (LSVI) algorithm. In particular, in a version of LSVI, we estimate the optimal action-value function of the leader via least-squares regression and construct an estimate of the SNE by solving the SNE of the multi-matrix game for each state, whose payoff matrices are given by the leader's estimated action-value function and the followers' reward functions. Moreover, we add a UCB exploration bonus to the least-squares solution to achieve optimism in the online setting. Whereas in the offline setting, pessimism is achieved by subtracting a penalty function constructed using the offline data, which is equal to the negative bonus function. Moreover, these algorithms are readily able to incorporate function approximators and we showcase the version with linear function approximation. Second, under the online setting, we prove that our optimistic LSVI algorithm achieves a sublinear $\widetilde{\mathcal{O}}(H^2\sqrt{d^3K})$ regret, where K is the number of episodes, H is the horizon, d is the dimension of the feature mapping, and $\widetilde{\mathcal{O}}(\cdot)$ omits logarithmic terms. Finally, under the offline setting, we establish an upper bound on the suboptimality of the proposed algorithm for an arbitrary dataset with K trajectories. Our upper bound yields a sublinear $\widetilde{\mathcal{O}}(H^2\sqrt{d^3/K})$ rate as long as the dataset has sufficient coverage over the trajectory induced by the desired SNE.

Related work. See Appendix A for details.

Notation. We denote by $\|\cdot\|_2$ the ℓ_2-norm of a vector or the spectral norm of a matrix. We also let $\|\cdot\|_{\mathrm{op}}$ denote the matrix operator norm. Furthermore, for a positive definite matrix A, we denote by $\|x\|_A$ the weighted norm $\sqrt{x^\top Ax}$ of a vector x. Also, we denote by $\Delta(\mathcal{A})$ the set of probability distributions on a set $\mathcal{A}$. For some positive integer K, $[K]$ denotes the index set $\{1, 2, \cdots, K\}$.

2 PRELIMINARIES

In this section, we introduce the formulation of the general-sum simultaneous-move Markov games, Stackelberg-Nash equilibrium, and the linear structure we use in this paper.

2.1 GENERAL-SUM SIMULTANEOUS-MOVE MARKOV GAMES

In this setting, two levels of hierarchy in decision making are considered: one leader l and N followers $\{f_i\}_{i\in[N]}$. Specifically, we define an episodic version of general-sum simultaneous-moves Markov game by the tuple $(\mathcal{S}, \mathcal{A}_l, \mathcal{A}_f = \{\mathcal{A}_{f_i}\}_{i\in[N]}, H, r_l, r_f = \{r_{f_i}\}_{i\in[N]}, \mathcal{P})$, where $\mathcal{S}$ is the state space, $\mathcal{A}_l$ and $\mathcal{A}_f$ are the sets of actions of the leader and the followers respectively, H is the number of steps in each episode, $r_l = \{r_{l,h} : \mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f \to [-1,1]\}_{h=1}^H$ and $r_{f_i} = \{r_{f_i,h} : \mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f \to [-1,1]\}_{h=1}^H$ are reward functions of the leader and the followers respectively, and $\mathcal{P} = \{\mathcal{P}_h : \mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f\times\mathcal{S} \to [0,1]\}_{h=1}^H$ is a collection of transition kernels. Here $\mathcal{A}_l\times\mathcal{A}_f = \mathcal{A}_l\times\mathcal{A}_{f_1}\times\cdots\times\mathcal{A}_{f_N}$. Throughout this paper, we also let $\star$ be some element in $\{l, f_1, \cdots, f_N\}$. Moreover, for any $(h,x,a)\in[H]\times\mathcal{S}\times\mathcal{A}_l$ and $b = \{b_i \in \mathcal{A}_{f_i}\}_{i\in[N]}$, we use the shorthands $r_{\star,h}(x,a,b) = r_{\star,h}(x,a,b_1,\cdots,b_N)$ and $\mathcal{P}_h(\cdot\,|\,x,a,b) = \mathcal{P}_h(\cdot\,|\,x,a,b_1,\cdots,b_N)$.

Policy and Value Function. A stochastic policy $\pi = \{\pi_h : \mathcal{S}\to\Delta(\mathcal{A}_l)\}_{h=1}^H$ of the leader is a set of probability distributions over actions given the state. Meanwhile, a stochastic joint policy of the followers is defined by $\nu = \{\nu_{f_i}\}_{i\in[N]}$, where $\nu_{f_i} = \{\nu_{f_i,h} : \mathcal{S}\to\Delta(\mathcal{A}_{f_i})\}_{h=1}^H$. We use the notation $\pi_h(a\,|\,x)$ and $\nu_{f_i,h}(b_i\,|\,x)$ to denote the probability of taking action $a\in\mathcal{A}_l$ or $b_i\in\mathcal{A}_{f_i}$ for state x at step h under policy π, ν_{f_i} respectively. Throughout this paper, for any $\nu = \{\nu_{f_i}\}_{i\in[N]}$ and $b = \{b_i\}_{i\in[N]}$, we use the shorthand $\nu_h(b\,|\,x) = \nu_{f_1,h}(b_1\,|\,x)\times\cdots\times\nu_{f_N,h}(b_N\,|\,x)$.

Given policies $(\pi, \nu = \{\nu_{f_i}\}_{i\in[N]})$, the action-value (Q) and state-value (V) functions for the leader and followers are defined by

$$Q_{\star,h}^{\pi,\nu}(x,a,b) = \mathbb{E}_{\pi,\nu,h,x,a,b}\Bigg[\sum_{t=h}^{H} r_{\star,h}(x_t,a_t,b_t)\Bigg],\quad V_{\star,h}^{\pi,\nu}(x) = \mathbb{E}_{a\sim\pi_h(\cdot\,|\,x),b\sim\nu_h(\cdot\,|\,x)}Q_{\star,h}^{\pi,\nu}(x,a,b), \tag{2.1}$$

where the expectation $\mathbb{E}_{\pi,\nu,h,x,a,b}$ is taken over state-action pairs induced by the policies $(\pi, \nu = \{\nu_{f_i}\}_{i\in[N]})$ and the transition probability, when initializing the process with the triplet $(s, a, b = \{b_i\}_{i\in[N]})$ at step h. For notational simplicity, when h, x, a, b are clear from the context, we omit h, x, a, b from $\mathbb{E}_{\pi,\nu,h,x,a,b}$. By the definition in (2.1), we have the Bellman equation

$$V_{\star,h}^{\pi,\nu} = \langle Q_{\star,h}^{\pi,\nu}, \pi_h\times\nu_h\rangle_{\mathcal{A}_l\times\mathcal{A}_f},\quad Q_{\star,h}^{\pi,\nu} = r_{\star,h} + \mathbb{P}_h V_{\star,h+1}^{\pi,\nu},\quad \forall\star\in\{l, f_1,\cdots,f_N\}, \tag{2.2}$$

where $\pi_h\times\nu_h$ represents $\pi_h\times\nu_{f_1,h}\times\cdots\times\nu_{f_N,h}$. Here $\mathbb{P}_h$ is the operator which is defined by

$$(\mathbb{P}_h f)(x,a,b) = \mathbb{E}[f(x')\,|\,x'\sim\mathcal{P}_h(x'\,|\,x,a,b)] \tag{2.3}$$

for any function $f : \mathcal{S}\to\mathbb{R}$ and $(x,a,b)\in\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$.

2.2 STACKELBERG-NASH EQUILIBRIUM

Given a leader policy π, a Nash equilibrium (Nash, 2016) of the followers is a joint policy $\nu^* = \{\nu_{f_i}^*\}_{i\in[N]}$, such that for any $x\in\mathcal{S}$ and $(i,h)\in[N]\times[H]$

$$V_{f_i,h}^{\pi,\nu^*}(x) \geq V_{f_i,h}^{\pi,\nu_{f_i},\nu_{f_{-i}}^*}(x),\quad \forall\nu_{f_i}. \tag{2.4}$$

Here $-i$ represents all indices in $[N]$ except i. For each leader policy π, we denote the set of best-response policies of the followers by $\mathrm{BR}(\pi)$, which is defined by

$$\mathrm{BR}(\pi) = \{\nu \,|\, \nu \text{ is the NE of the followers given the leader policy } \pi\}. \tag{2.5}$$

Given the best-response set $\mathrm{BR}(\pi)$, we denote $\nu^*(\pi)$ the best-case responses, which break ties in favor of the leader. This is also known as optimistic tie-breaking (Breton et al., 1988; Bucarey et al., 2019a). We will discuss pessimistic tie-breaking (Conitzer & Sandholm, 2006) in §I. Specifically, we define $\nu^*(\pi)$ by

$$\nu^*(\pi) = \{\nu \in \mathrm{BR}(\pi) \,|\, V_{l,h}^{\pi,\nu}(x) \geq V_{l,h}^{\pi,\nu'}(x), \forall x \in \mathcal{S}, h \in [H], \nu' \in \mathrm{BR}(\pi)\}. \tag{2.6}$$

The Stackelberg-Nash equilibrium for the leader is the "best response to the best response". In other words, we want to find a leader's policy π that maximizes her value function under the assumption that the followers always adopt $\nu^*(\pi)$, i.e.,

$$\mathrm{SNE}_l = \{\pi \,|\, V_{l,h}^{\pi,\nu^*(\pi)}(x) \geq V_{l,h}^{\pi',\nu^*(\pi')}(x), \forall x \in \mathcal{S}, h \in [H], \pi'\} \tag{2.7}$$

A Stackelberg-Nash equilibrium of the general-sum game is a policy pair $(\pi^*, \nu^* = \{\nu_{f_i}^*\}_{i\in[N]})$ such that $\nu^* \in \nu^*(\pi^*)$ and $\pi^* \in \mathrm{SNE}_l$.

Our goal is to find the Stackelberg equilibrium: the leader's optimal strategy, under the assumption that the followers always play their best response (Nash equilibrium) to the leader. Equivalently, we need to solve the following optimization problem:

$$\max_{\pi,\nu} V_{l,1}^{\pi,\nu}(x) \qquad \text{s.t. } \nu \in \mathrm{BR}(\pi). \tag{2.8}$$

We study this challenging bilevel optimization problem in both the online setting (Section 3) and the offline setting (Section 4) respectively.

2.3 Linear Markov Games with Myopic Followers

Linear Markov Games. We study the linear Markov games (Xie et al., 2020), where the transition dynamics are linear in a feature map.

Assumption 2.1 (Linear Markov Games)**.** Markov game $(\mathcal{S}, \mathcal{A}_l, \mathcal{A}_f = \{\mathcal{A}_{f_i}\}_{i\in[N]}, H, r_l, r_f = \{r_{f_i}\}_{i\in[N]}, \mathcal{P})$ is a linear Markov game if there exists a feature map $\phi : \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f \to \mathbb{R}^d$ such that

$$\mathcal{P}_h(\cdot \,|\, x, a, b) = \langle \phi(x, a, b), \mu_h(\cdot) \rangle$$

for any $(x, a, b) \in \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$ and $h \in [H]$. Here $\mu_h = (\mu_h^{(1)}, \mu_h^{(2)}, \cdots, \mu_h^{(d)})$ are d unknown signed measures over $\mathcal{S}$. Without loss of generality, we assume that $\|\phi(\cdot,\cdot,\cdot)\|_2 \leq 1$ and $\|\mu_h(\mathcal{S})\| \leq \sqrt{d}$ for all $h \in [H]$.

The linear Markov game above is an extension of linear MDP studied in Jin et al. (2020b) for the single-agent RL. Specifically, when the followers play fixed and known policies, the linear Markov games reduce to the linear MDP. We also remark that Chen et al. (2021b) recently study another variant of linear Markov games. These two variants are incomparable in the sense that one does not imply the other.

Myopic Followers. Throughout this paper, we make the following assumption.

Assumption 2.2 (Myopic followers)**.** We assume that the followers are myopic. Specifically, the followers at any step of the game do not consider the future rewards, but only the instantaneous rewards. Formally, given a leader's policy π, the NE of the myopic followers is defined by the joint policy $\nu^* = \{\nu_{f_i}^*\}_{i\in[N]}$, such that for any $x \in \mathcal{S}$ and $(i, h) \in [N] \times [H]$

$$r_{f_i,h}^{\pi,\nu^*}(x) \geq r_{f_i,h}^{\pi,\nu_{f_i},\nu_{f_{-i}}^*}(x), \quad \forall \nu_{f_i}. \tag{2.9}$$

In other words, at each state x, the followers play a normal form game where the payoff matrices are determined only by the immediate reward functions and the leader's policy. Then, with slight abuse of notation, the best response set of the leader's policy π is

$$\mathrm{BR}(\pi) = \{\nu \,|\, \nu \text{ is the NE of the myopic followers given the leader policy } \pi\}. \tag{2.10}$$

And the best-case response $\nu^*(\pi)$ and Stackelberg-Nash equilibria SNE_l follow the definitions in (2.6) and (2.7).

Leader-Controller Linear Markov Games. A special case of the Markov games with myopic followers is leader-controller Markov games (Filar & Vrieze, 2012; Bucarey et al., 2019a), where the future state only depends on the current state and the leader's action. Such a setting is also well-motivated. One can consider the game where the leader is the government that dictates prices and the followers are companies. This is a leader-controller Markov game because the future state (price) is determined by the current state (price) and the leader (government). Formally, it holds that $\mathcal{P}_h(\cdot \,|\, x, a, b) = \mathcal{P}_h(\cdot \,|\, x, a)$ for any $(x, a, b) \in \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$ and $h \in [H]$. Hence, with slight abuse of notation, it is naturally to make the following assumption.

Assumption 2.3 (Leader-Controller Linear Markov Games)**.** Markov game $(\mathcal{S}, \mathcal{A}_l, \mathcal{A}_f = \{\mathcal{A}_{f_i}\}_{i\in[N]}, H, r_l, r_f = \{r_{f_i}\}_{i\in[N]}, \mathcal{P})$ is a leader-controller linear Markov games if we assume the existence of a feature $\phi : \mathcal{S} \times \mathcal{A}_l \to \mathbb{R}^d$ such that

$$\mathcal{P}_h(\cdot \,|\, x, a, b) = \langle \phi(s, a), \mu_h(\cdot) \rangle, \tag{2.11}$$

for any $(s, a, b) \in \mathcal{A} \times \mathcal{A}_l \times \mathcal{A}_f$, where $\|\phi(\cdot, \cdot)\|_2 \le 1$ and $\|\mu_h(\mathcal{S})\| \le \sqrt{d}$ for all $h \in [H]$.

Notably, Markov games with myopic followers subsume leader-controller Markov games as a special case. Specifically, since the followers' policies cannot affect the following state, then the NE defined in (2.4) is the same as (2.9), which further implies that the best-response set defined in (2.5) is the same as (2.10).

3 Main Results for the Online Setting

In this section, we study the online setting, where a central controller controls one leader l and N followers $\{f_i\}_{i\in[N]}$. Our goal is to learn a Stackelberg-Nash equilibrium. In what follows, we formally describe the setup and learning objectives, and then present our algorithm and provide theoretic guarantees.

3.1 Setup and Learning Objective

We consider the setting where the reward functions r_l and $r_f = \{r_{f_i}\}_{i\in[N]}$ are revealed to the learner before the game. This is reasonable since in practice the reward functions are usually artificially designed. Moreover, we focus on the episodic setting. Specifically, a Markov game is played for K episodes, each of which consists of H timesteps. At the beginning of the k-th episode, the leader and followers determine their policies $(\pi^k, \nu^k = \{\nu^k_{f_i}\}_{i\in[N]})$, and a fixed initial state $x^k_1 = x_1$ is chosen. Here we assume the fixed initial state just for ease of presentation, and our subsequent results can be generalized to the setting where x^k_1 is picked from a fixed distribution. Then the game proceeds as follows. At each step $h \in [H]$, the leader and the followers observe state $x^k_h \in \mathcal{S}$ and pick their own actions $a^k_h \sim \pi^k_h(\cdot \,|\, x^k_h)$ and $b^k_h = \{b^k_{i,h} \sim \nu^k_{f_i,h}(\cdot \,|\, x^k_h)\}_{i\in[N]}$. Subsequently, the environment transitions to the next state $x^k_{h+1} \sim \mathcal{P}_h(\cdot \,|\, x^k_h, a^k_h, b^k_h)$. Each episode terminates after H timesteps.

Learning Objective. Let $(\pi^k, \nu^k = \{\nu^k_{f_i}\}_{i\in[N]})$ denote the policies executed by the algorithm in the k-th episode. By the definition of the bilevel optimization problem in (2.8), we expect that $\nu^k \in \text{BR}(\pi^k)$ and that $V^{\pi^*,\nu^*}_{l,1}(x^k_1) - V^{\pi^k,\nu^k}_{l,1}(x^k_1)$ is small for any $k \in [K]$. Hence, we evaluate the suboptimality of our algorithm by the regret, which is defined as

$$\text{Regret}(K) = \sum_{k=1}^{K} V^{\pi^*,\nu^*}_{l,1}(x^k_1) - V^{\pi^k,\nu^k}_{l,1}(x^k_1). \tag{3.1}$$

The goal is to design algorithms with regret that is sublinear in K, and polynomial in d, H. Here K is the number of episodes, d is the dimension of the feature map ϕ, and H is the episode horizon.

3.2 Algorithm

We now present our algorithm, Optimistic Value Iteration to Find Stackelberg-Nash Equilibrium (OVI-SNE), which is given in Algorithm 1.

Algorithm 1 Optimistic Value Iteration to Find Stackelberg-Nash Equilibria

1: Initialize $V_{l,H+1}(\cdot) = V_{f,H+1}(\cdot) = 0$.
2: **for** $k = 1, 2, \cdots, K$ **do**
3: Receive initial state x_1^k.
4: **for** step $h = H, H-1, \cdots, 1$ **do**
5: $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau)\phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top + I$.
6: $w_h^k \leftarrow (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot V_{h+1}^k(x_{h+1}^\tau)$.
7: $\Gamma_h^k(\cdot,\cdot,\cdot) \leftarrow \beta \cdot (\phi(\cdot,\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \phi(\cdot,\cdot,\cdot))^{1/2}$.
8: $Q_h^k(\cdot,\cdot,\cdot) \leftarrow r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{\phi(\cdot,\cdot,\cdot)^\top w_h^k + \Gamma_h^k(\cdot,\cdot,\cdot)\}$.
9: $(\pi_h^k(\cdot\,|\,x), \{\nu_{f_i,h}^k(\cdot\,|\,x)\}_{i\in[N]}) \leftarrow \epsilon$-SNE$(Q_h^k(x,\cdot,\cdot), \{r_{f_i,h}(x,\cdot,\cdot)\}_{i\in[N]})$, $\forall x$. (Alg. 2)
10: $V_h^k(x) \leftarrow \mathbb{E}_{a\sim\pi_h^k(\cdot\,|\,x), b_1\sim\nu_{f_1,h}^k(\cdot\,|\,x), \cdots, b_N\sim\nu_{f_N,h}^k(\cdot\,|\,x)} Q_h^k(x, a, b_1, \cdots, b_N)$, $\forall x$.
11: **end for**
12: **for** $h = 1, 2, \cdot, H$ **do**
13: Sample $a_h^k \sim \pi_h^k(\cdot\,|\,x_h^k), b_{1,h}^k \sim \nu_{f_1,h}^k(\cdot\,|\,x_h^k), \cdots, b_{N,h}^k \sim \nu_{f_N,h}^k(\cdot\,|\,x_h^k)$.
14: Leader takes action a_h^k; Followers take actions $b_h^k = \{b_{i,h}^k\}_{i\in[N]}$.
15: Observe next state x_{h+1}^k.
16: **end for**
17: **end for**

Algorithm 2 ϵ-SNE

1: **Input:** Q_h^k, x, and parameter ϵ.
2: Select $\widetilde{Q}$ from $\mathcal{Q}_{h,\epsilon}^k$ satisfying $\|\widetilde{Q} - Q_h^k\|_\infty \le \epsilon$.
3: For the input state x, let $(\pi_h^k(\cdot\,|\,x), \{\nu_{f_i,h}^k(\cdot\,|\,x)\}_{i\in[N]})$ be the Stackelberg-Nash equilibrium for the matrix game with payoff matrices $(\widetilde{Q}(x,\cdot,\cdot), \{r_{f_i,h}(x,\cdot,\cdot)\}_{i\in[N]})$.
4: **Output:** $(\pi_h^k(\cdot\,|\,x), \{\nu_{f_i,h}^k(\cdot\,|\,x)\}_{i\in[N]})$.

At a high level, in each episode, our algorithm first construct the policies for all players through backward induction with respect to the timestep h (Line 4-11), and then execute the policies to play the game (Line 12-16).

In detail, at h-th step of k-th episode, OVI-SNE estimates leader's Q-function based on the $(k-1)$ historical trajectories. Inspired by previous optimistic least square value iteration (LSVI) algorithms (Jin et al., 2020b), for any $h \in [H]$, we estimate the linear coefficients by solving the following ridge regression problem:

$$
\begin{aligned}
w_h^k &\leftarrow \mathop{\mathrm{argmin}}_{w\in\mathbb{R}^d} \sum_{\tau=1}^{k-1} [V_{h+1}^k(x_{h+1}^\tau) - \phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top w]^2 + \|w\|_2^2, \\
&\text{where } V_{h+1}^k(\cdot) = \langle Q_{h+1}^k(\cdot,\cdot,\cdot), \pi_{h+1}^k(\cdot\,|\,\cdot) \times \nu_{h+1}^k(\cdot\,|\,\cdot)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}.
\end{aligned}
\tag{3.2}
$$

By solving the ridge regression problem in (3.2), we have

$$
\begin{aligned}
w_h^k &= (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot V_{h+1}^k(x_{h+1}^\tau)\Big), \\
\text{where } \Lambda_h^k &= \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau)\phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top + I.
\end{aligned}
\tag{3.3}
$$

To encourage exploration, we additionally adds a bonus function to estimate the leader's Q-function:

$$
\begin{aligned}
&Q_h^k(\cdot,\cdot,\cdot) \leftarrow r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{\phi(\cdot,\cdot,\cdot)^\top w_h^k + \Gamma_h^k(\cdot,\cdot,\cdot)\}, \\
&\text{where } \Gamma_h^k(\cdot,\cdot,\cdot) = \beta \cdot \sqrt{\phi(\cdot,\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \phi(\cdot,\cdot,\cdot)}.
\end{aligned}
\tag{3.4}
$$

Here $\Gamma_h^k : \mathcal{S} \times \mathcal{A}_l \to \mathbb{R}$ is a bonus function and $\beta > 0$ is a parameter which will be specified later. This form of bonus function is common in the literature of linear bandits (Lattimore & Szepesvári, 2020) and linear MDPs (Jin et al., 2020b).

Then, we construct policies for the leader and followers by the subroutine ϵ-SNE (Algorithm 2). Specifically, let $\mathcal{Q}_h^k$ be the class of functions $Q : \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f \to \mathbb{R}$ that takes form

$$Q(\cdot,\cdot,\cdot) = r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{\phi(\cdot,\cdot,\cdot)^\top w + \beta \cdot \big(\phi(\cdot,\cdot,\cdot)^\top \Lambda^{-1}\phi(\cdot,\cdot,\cdot)\big)^{1/2}\}, \quad (3.5)$$

where the parameters $(w, \Lambda) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$ satisfy $\|w\| \le H\sqrt{dk}$ and $\lambda_{\min}(\Lambda) \ge 1$. Moreover, let $\mathcal{Q}_{h,\epsilon}^k$ be a fixed ϵ-covering of $\mathcal{Q}_h^k$ with respect to the ℓ_∞ norm. By Lemma D.10, we have $Q_h^k \in \mathcal{Q}_h^k$, which allows us to pick a $\widetilde{Q} \in \mathcal{Q}_{h,\epsilon}^k$ such that $\|\widetilde{Q} - Q_h^k\|_\infty \le \epsilon$ and calculate policies by

$$(\pi_h^k(\cdot \,|\, x), \{\nu_{f_i,h}^k(\cdot \,|\, x)\}_{i\in[N]}) \leftarrow \text{SNE}(\widetilde{Q}(x,\cdot,\cdot), \{r_{f_i,h}(x,\cdot,\cdot)\}_{i\in[N]}),\ \forall x. \quad (3.6)$$

When there is only one follower, (3.6) requires finding a Stackelberg equilibrium in the matrix game. Such a problem can be transformed to a linear programming (LP) problem (Conitzer & Sandholm, 2006; Von Stengel & Zamir, 2010), and thus can be solved efficiently. For the multi-follower case (i.e., $N \ge 2$), however, solving such a matrix game in general is hard (Conitzer & Sandholm, 2006; Basilico et al., 2017a;b; Coniglio et al., 2020). Given this computational hardness, we focus on the sample complexity and explicitly assume access to the following computational oracle:

Assumption 3.1. We assume access to an oracle that implements Line 3 of Algorithm 2.

Finally, the leader and the followers play the game according to the obtained policies.

Remark 3.2. Due to some technical challenge, the subroutine ϵ-SNE is necessary. See §B.1 for more explanations.

3.3 Theoretical Results

Our main theoretical result is the following bound on the regret incurred by Algorithm 1. Recall that the regret is defined in (3.1) and $T = KH$ is the total number of timesteps.

Theorem 3.3. Under Assumptions 2.1, 2.2, and 3.1, there exists an absolute constant $C > 0$ such that, for any fixed $p \in (0,1)$, by setting $\beta = C\cdot dH\sqrt{\iota}$ with $\iota = \log(2dT/p)$ in Line 7 of Algorithm 1 and $\epsilon = \frac{1}{KH}$ in Algorithm 2, then we have $\nu^k \in \text{BR}(\pi^k)$ for any $k \in [K]$. Meanwhile, with probability at least $1-p$, the regret incurred by OVI-SNE satisfies that

$$\text{Regret}(K) \le \mathcal{O}(\sqrt{d^3H^3T\iota^2}).$$

The proof is deferred to Appendix D. We will discuss the Stackelberg equilibria learning, optimality of the bound, and unknown reward setting in Appendix B.2.

4 Main Results for the Offline Setting

In this section, we study the offline setting, where the central controller aims to find a Stackelberg-Nash equilibrium by an offline dataset. Below we describe the setup and learning objective, followed by our algorithm and theoretical results.

4.1 Setup and Learning Objective

We study the offline setting, where the learner has access to the reward functions $(r_l, r_f = \{r_{f_i}\}_{i=1}^N)$ and a dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, b_h^\tau = \{b_{i,h}^\tau\}_{i=1}^N)\}_{\tau,h=1}^{K,H}$, which is collected a priori by some experimenter. Then we make a minimal assumption for the offline dataset.

Assumption 4.1 (Compliance of Dataset)**.** We assume that the dataset $\mathcal{D}$ is compliant with the underlying Markov game $(\mathcal{S}, \mathcal{A}_l, \mathcal{A}_f, H, r_l, r_f, \mathcal{P})$, that is, for any $x' \in \mathcal{S}$ at step $h \in [H]$ of each trajectory $\tau \in [K]$,

$$P_{\mathcal{D}}(x_{h+1}^\tau = x' \,|\, \{x_h^j, a_h^j, b_h^j, x_{h+1}^j\}_{j=1}^{\tau-1} \cup \{x_h^\tau, a_h^\tau, b_h^\tau\}) = P(x_{h+1} = x' \,|\, x_h = x_h^\tau, a_h = a_h^\tau, b_h = b_h^\tau).$$

Here the probability on the left-hand side is with respect to the joint distribution over dataset $\mathcal{D}$ and the probability on the right-hand side is with respect to the underlying Markov game.

Assumption 4.1 is adopted from Jin et al. (2020c), which indicates the Markov property of the dataset $\mathcal{D}$ and that x_{h+1}^τ is generated by the underlying Markov game conditioned on $(x_h^\tau, a_h^\tau, b_h^\tau)$. As a special case, Assumption 4.1 holds when the experimenter follows fixed behavior policies. More generally, Assumption 4.1 allows the experimenter to choose actions a_h^τ and b_h^τ arbitrarily, even in an adaptive or adversarial manner. In particular, we can assume that a_h^τ and b_h^τ are interdependent across each trajectory $\tau \in [K]$. For instance, the experimenter can sequentially improve the behavior policy using any online algorithm for Markov games.

Learning Objective. Similar to the online setting, we define the following performance metric

$$\mathrm{SubOpt}(\pi, \nu, x) = V_{l,1}^{\pi^*,\nu^*}(x) - V_{l,1}^{\pi,\nu}(x), \tag{4.1}$$

which evaluates the suboptimality of policies $(\pi, \nu = \{\nu_{f_i}\}_{i=1}^N)$ given the initial state $x \in \mathcal{S}$.

4.2 Algorithm

As is known to us, the key challenge of online setting is the the tradeoff between exploration and exploration. In the online setting. by following the "optimism in the face of uncertainty" principle (Sutton & Barto, 2018), we use bonus functions to incentivize exploration and thus achieve sample-efficient. This intrinsic challenge of online setting disappears in the offline setting because we do not need exploration any more. But another challenge arises: we only have access to the limited data. To tackle this challenge, we need add some penalty functions to achieve robustness against the uncertainty due to the finite data. This is also known as pessimism (Yu et al., 2020; Jin et al., 2020c; Liu et al., 2020b; Buckman et al., 2020; Kidambi et al., 2020; Kumar et al., 2020; Rashidinejad et al., 2021). Here we simply flip the sign of bonus functions defined in (3.4) to serve as penalty functions. See Algorithm 3 for details.

4.3 Theoretical Results

Suppose that $(\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H, \widehat{\nu} = \{\widehat{\nu}_{f_i} = \{\nu_{f_i,h}\}_{h=1}^H\}_{i=1}^N)$ are the output policies of Algorithm 3. Then we evaluate the performance of $(\widehat{\pi}, \widehat{\nu})$ by establishing an upper bound for the optimality gap defined in (4.1).

Theorem 4.2. Under Assumptions 2.1, 2.2, 3.1, and 4.1, there exists an absolute constant $C > 0$ such that, for any fixed $p \in (0, 1)$, by setting $\beta' = C \cdot dH\sqrt{\log(2dHK/p)}$ in Line 6 of Algorithm 3 and $\epsilon = \frac{d}{KH}$ in Algorithm 2, then it holds that $\widehat{\nu} \in \mathrm{BR}(\widehat{\pi})$. Meanwhile, with probability at least $1 - p$, we have

$$\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}, x) \le 3\beta' \sum_{h=1}^H \mathbb{E}_{\pi^*,\nu^*,x}\big[\big(\phi(s_h, a_h, b_h)^\top (\Lambda_h)^{-1}\phi(s_h, a_h, b_h)\big)^{1/2}\big], \tag{4.2}$$

where $\mathbb{E}_{\pi^*,\nu^*,x}$ is taken with respect to the trajectory incurred by (π^*, ν^*) in the underlying Markov game when initializing the progress at x. Here Λ_h is defined in Line 4 of Algorithm 3.

The proof of Theorem 4.2 is deferred to §G. To illustrate our theory more, we will provide more comments on Theorem 4.2 in §C.2.

5 Conclusion

In this paper, we investigate the question of can we sample efficiently find SNE in general-sum Markov games with myopic followers and linear function approximation. To the best of our knowledge, in both online and offline settings, we develop the first sample efficient reinforcement learning algorithms for solving SNE. We believe our work opens up many interesting directions for future work. For example, we can ask the following questions: Can we find SNE in general-sum Markov games without the myopic followers assumption? Can we design more computationally efficient algorithms for solving SNE in general-sum Markov games? Can we find SNE in general-sum Markov games with general function approximation?

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, volume 11, pp. 2312–2320, 2011.

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Forest Agostinelli, Stephen McAleer, Alexander Shmakov, and Pierre Baldi. Solving the rubik's cube with deep reinforcement learning and search. *Nature Machine Intelligence*, 1(8):356–363, 2019.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

András Antos, Rémi Munos, and Csaba Szepesvári. Fitted q-iteration in continuous action-space mdps. 2007.

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

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*, pp. 463–474. PMLR, 2020.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3): 325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pp. 551–560. PMLR, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *arXiv preprint arXiv:2006.12007*, 2020.

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of stackelberg equilibria in general-sum games. *arXiv preprint arXiv:2102.11494*, 2021.

Tamer Başar and Geert Jan Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

Nicola Basilico, Stefano Coniglio, and Nicola Gatti. Methods for finding leader–follower equilibria with multiple followers. *arXiv preprint arXiv:1707.02174*, 2017a.

Nicola Basilico, Stefano Coniglio, Nicola Gatti, and Alberto Marchesi. Bilevel programming approaches to the computation of optimistic and pessimistic single-leader-multi-follower equilibria. In *SEA*, volume 75, pp. 1–14. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2017b.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

Avrim Blum, Nika Haghtalab, and Ariel D Procaccia. Learning optimal commitment to overcome insecurity. 2014.

Michele Breton, Abderrahmane Alj, and Alain Haurie. Sequential stackelberg equilibria in two-person games. *Journal of Optimization Theory and Applications*, 59(1):71–97, 1988.

Víctor Bucarey, Eugenio Della Vecchia, Alain Jean-Marie, and Fernando Ordóñez. *Stationary Strong Stackelberg Equilibrium in Discounted Stochastic Games*. PhD thesis, INRIA, 2019a.

Víctor Bucarey, Alain Jean-Marie, Eugenio Della Vecchia, and Fernando Ordóñez. On the value iteration method for dynamic strong stackelberg equilibria. In *ROADEF 2019-20ème congrès annuel de la société Française de Recherche Opérationnelle et d'Aide à la Décision*, 2019b.

Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021a.

Zhuoqun Chen, Yangyang Liu, Bo Zhou, and Meixia Tao. Caching incentive design in wireless d2d networks: A stackelberg game approach. In *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6. IEEE, 2016.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player markov games with linear function approximation. *arXiv preprint arXiv:2102.07404*, 2021b.

Stefano Coniglio, Nicola Gatti, and Alberto Marchesi. Computing a pessimistic stackelberg equilibrium with multiple followers: The mixed-pure case. *Algorithmica*, 82(5):1189–1238, 2020.

Vincent Conitzer and Tuomas Sandholm. Complexity of mechanism design. *arXiv preprint cs/0205075*, 2002.

Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pp. 82–90, 2006.

Qiwen Cui and Lin F Yang. Minimax sample complexity for turn-based stochastic game. *arXiv preprint arXiv:2011.14267*, 2020.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *arXiv preprint arXiv:2101.04233*, 2021.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2016.

Yonathan Efroni, Lior Shani, Aviv Rosenberg, and Shie Mannor. Optimistic policy optimization with bandit feedback. *arXiv preprint arXiv:2002.08243*, 2020.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Amir Massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems*, 2010.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Tanner Fiez, Benjamin Chasnov, and Lillian J Ratliff. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pp. 2052–2062. PMLR, 2019.

Dinesh Garg and Yadati Narahari. Design of incentive compatible mechanisms for stackelberg problems. In *International Workshop on Internet and Network Economics*, pp. 718–727. Springer, 2005.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Amy Greenwald, Keith Hall, and Roberto Serrano. Correlated q-learning. In *ICML*, volume 3, pp. 242–249, 2003.

Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. Is multiagent deep reinforcement learning the answer or the question? a brief survey. *Learning*, 21:22, 2018.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*, 2019.

Zeyu Jia, Lin F Yang, and Mengdi Wang. Feature-based q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *arXiv preprint arXiv:1807.03765*, 2018.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 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*, pp. 2137–2143. PMLR, 2020b.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? *arXiv preprint arXiv:2012.15085*, 2020c.

Xin Kang and Yongdong Wu. Incentive mechanism design for heterogeneous peer-to-peer networks: A stackelberg game approach. *IEEE Transactions on Mobile Computing*, 14(5):1018–1030, 2014.

Katarzyna Kańska and Agnieszka Wiszniewska-Matyszkiel. Dynamic stackelberg duopoly with sticky prices and a myopic follower. *Operational Research*, pp. 1–32, 2021.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

Dmytro Korzhyk, Zhengyu Yin, Christopher Kiekintveld, Vincent Conitzer, and Milind Tambe. Stackelberg vs. Nash in security games: An extended investigation of interchangeability, equivalence, and uniqueness. *Journal of Artificial Intelligence Research*, 41:297–327, 2011.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *arXiv preprint arXiv:1906.00949*, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Michail Lagoudakis and Ron Parr. Value function approximation in zero-sum markov games. *arXiv preprint arXiv:1301.0580*, 2012.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pp. 3652–3661. PMLR, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *International Symposium on Algorithmic Game Theory*, pp. 250–262. Springer, 2009.

Tao Li and Suresh P Sethi. A review of dynamic stackelberg game models. *Discrete & Continuous Dynamical Systems-B*, 22(1):125, 2017.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.

Michael L Littman. Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pp. 322–328, 2001.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *arXiv preprint arXiv:1810.12429*, 2018.

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*, 2020a.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020b.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *arXiv preprint arXiv:2110.05682*, 2021.

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.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

John F Nash. *Non-Cooperative Games*. Princeton University Press, 2016.

Afshin OroojlooyJadid and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning optimal strategies to commit to. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2149–2156, 2019.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning*, pp. 1321–1329. PMLR, 2015.

Aravind Rajeswaran, Igor Mordatch, and Vikash Kumar. A game theoretic framework for model based reinforcement learning. In *International Conference on Machine Learning*, pp. 7953–7963. PMLR, 2020.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021*, 2021.

Lillian J Ratliff and Tanner Fiez. Adaptive incentive design. *IEEE Transactions on Automatic Control*, 2020.

Lillian J Ratliff, Ming Jin, Ioannis C Konstantakopoulos, Costas Spanos, and S Shankar Sastry. Social game for building energy efficiency: Incentive design. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1011–1018. IEEE, 2014.

Tim Roughgarden. Stackelberg scheduling strategies. *SIAM journal on computing*, 33(2):332–350, 2004.

Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of tetris. *J. Mach. Learn. Res.*, 16:1629–1676, 2015.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pp. 2992–3002. PMLR, 2020.

Noah Y Siegel, Jost Tobias Springenberg, Felix Berkenkamp, Abbas Abdolmaleki, Michael Neunert, Thomas Lampe, Roland Hafner, Nicolas Heess, and Martin Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. *arXiv preprint arXiv:2002.08396*, 2020.

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):484–489, 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.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Marwaan Simaan and Jose B Cruz. On the stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications*, 11(5):533–555, 1973.

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pp. 1057–1063. Citeseer, 1999.

Milind Tambe. *Security and game theory: algorithms, deployed systems, lessons learned*. Cambridge university press, 2011.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Provably efficient online agnostic learning in markov games. *arXiv preprint arXiv:2010.15020*, 2020.

Bernhard Von Stengel and Shmuel Zamir. Leadership games with convex strategy sets. *Games and Economic Behavior*, 69(2):446–457, 2010.

Ziyu Wang, Alexander Novikov, Konrad Żołna, Jost Tobias Springenberg, Scott Reed, Bobak Shahriari, Noah Siegel, Josh Merel, Caglar Gulcehre, Nicolas Heess, et al. Critic regularized regression. *arXiv preprint arXiv:2006.15134*, 2020.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. *arXiv preprint arXiv:1712.00579*, 2017.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pp. 3674–3682. PMLR, 2020.

Tengyang Xie and Nan Jiang. $Q^\star$-approximation schemes for batch reinforcement learning: A theoretical comparison. *arXiv preprint arXiv:2003.03924*, 2020.

Lin Yang and Mengdi Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 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*, pp. 10746–10756. PMLR, 2020.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I Jordan. Bridging exploration and general function approximation in reinforcement learning: Provably efficient kernel and neural value iterations. *arXiv preprint arXiv:2011.04622*, 2020.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 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*, pp. 1954–1964. PMLR, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pp. 10978–10989. PMLR, 2020b.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

Kaiqing Zhang, Sham M Kakade, Tamer Başar, and Lin F Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *arXiv preprint arXiv:2007.07461*, 2020a.

Zihan Zhang, Xiangyang Ji, and Simon S Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. *arXiv preprint arXiv:2009.13503*, 2020b.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33, 2020c.

Yulai Zhao, Yuandong Tian, Jason D Lee, and Simon S Du. Provably efficient policy gradient methods for two-player zero-sum markov games. *arXiv preprint arXiv:2102.08903*, 2021.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C Parkes, and Richard Socher. The AI economist: Improving equality and productivity with AI-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

Ying-Ping Zheng, Tamer Basar, and Jose B Cruz. Stackelberg strategies and incentives in multiperson deterministic decision problems. *IEEE transactions on Systems, Man, and Cybernetics*, (1): 10–24, 1984.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. *arXiv preprint arXiv:2012.08507*, 2020.

A RELATED WORK

In the sequel, we discuss the related works.

RL for solving NE in Markov games Our work adds to the vast body of existing literature on RL for finding Nash equilibria in Markov games. In particular, there is a line of works that generalizes single-agent RL algorithms to Markov games under either the generative model (Azar et al., 2013) or offline settings with well-explored datasets (Littman, 2001; Greenwald et al., 2003; Hu & Wellman, 2003; Lagoudakis & Parr, 2012; Hansen et al., 2013; Perolat et al., 2015; Jia et al., 2019; Sidford et al., 2020; Cui & Yang, 2020; Fan et al., 2020; Daskalakis et al., 2021; Zhao et al., 2021). These works all aim to find the Nash equilibrium and their algorithms are generalizations of single-agent RL algorithms. In particular, Littman (2001; 1994); Greenwald et al. (2003); Hu & Wellman (2003) generalize Q-learning (Watkins & Dayan, 1992) to Markov games and establish asymptotic convergence guarantees. Jia et al. (2019); Sidford et al. (2020); Zhang et al. (2020a); Cui & Yang (2020) propose variants of Q-learning or value iteration (Shapley, 1953) algorithms under the generative model setting. Moreover, Perolat et al. (2015); Fan et al. (2020) study the sample efficiency of fitted value iteration (Munos & Szepesvári, 2008) for zero-sum Markov games under the offline setting. They assume the behavior policy is explorative in the sense that the concentrability coefficients (Munos & Szepesvári, 2008) are uniformly bounded. Under similar assumptions, Daskalakis et al. (2021); Zhao et al. (2021) study the sample complexity of policy gradient (Sutton et al., 1999) under the well-explored offline setting. Moreover, under the online setting, there is a recent line of research that proposes provably efficient RL algorithms for zero-sum Markov games. See, e.g., Wei et al. (2017); Bai et al. (2020); Bai & Jin (2020); Liu et al. (2020a); Tian et al. (2020); Xie et al. (2020); Chen et al. (2021b) and the references therein. These works propose optimism-based algorithms and establish sublinear regret guarantees for finding NE. Among these works, our work is particularly related to Xie et al. (2020); Chen et al. (2021b), whose algorithms also incorporate the linear function approximation. Compared with these aforementioned works, we focus on solving the Stackelberg-Nash equilibrium, which involves a bilevel structure and is fundamentally different from the Nash equilibrium. Thus, our work is not directly comparable.

Learning Stackelberg games As for solving Stackelberg-Nash equilibrium, most of the existing results focus on the normal form game, which is equivalent to our Markov game with $H = 1$. Letchford et al. (2009); Blum et al. (2014); Peng et al. (2019) study learning Stackelberg equilibrium with a best response oracle. In addition, Fiez et al. (2019) study the local convergence of first-order methods for finding Stackelberg equilibria in general-sum games with differentiable reward functions, and Ghadimi & Wang (2018); Chen et al. (2021a); Hong et al. (2020) analyze the global convergence of first-order methods for achieving global optimality of bilevel optimization. A more related work is Bai et al. (2021), which studies the matrix Stackelberg game with bandit feedback. This work also studies an RL extension where the leader has a finite action set and the follower is faced with an MDP specified by the leader's action. In comparison, we assume the leader knows the reward functions and the main challenge lies in the unknown transitions. Thus, our setting is different from that in Bai et al. (2021). Furthermore, a more relevant work is (Bucarey et al., 2019b), which establishes the Bellman equation and value iteration algorithm for solving SNE in Markov games. In comparison, we establish modifications of least-squares value iteration that are tailored to online and offline settings.

Learning general-sum Markov games Liu et al. (2020a) present the first result of finding correlated equilibrium (CE) and coarse correlated equilibrium (CCE) in general-sum Markov games. However, their centralized algorithms suffer from the curse of many agents, that is, their sample complexity scales exponentially in the number of agents. Recently, Song et al. (2021); Mao & Başar (2021); Jin et al. (2021) successfully escape the curse of many agents thanks to the decentralized structure of V-learning algorithm (Bai et al., 2020).

Related single-agent RL methods Broadly speaking, our work is also related to the recent line of works that achieve sample efficiency in single-agent RL under the online setting. See, e.g., (Azar et al., 2017; Jin et al., 2018; Yang & Wang, 2019; Zanette & Brunskill, 2019; Jin et al., 2020b; Zhou et al., 2020; Ayoub et al., 2020; Yang & Wang, 2020; Zanette et al., 2020b;a; Zhang et al., 2020c;b; Agarwal et al., 2020) and the references therein. In particular, following the *optimism in the face of uncertainty* principle, these works achieve near-optimal regret under either tabular or function approximation settings. Meanwhile, for offline RL with an arbitrary dataset, various recent works propose to utilize pessimism for achieving robustness. See, e.g., (Yu et al., 2020; Kidambi et al., 2020; Kumar et al., 2020; Jin et al., 2020c; Liu et al., 2020b; Buckman et al., 2020; Rashidinejad et al., 2021) and the references therein. These aforementioned works all focus on the single-agent setting and we prove that optimism and pessimism also play an indispensable role in achieving sample efficiency in finding SNE.

B MISSING PARTS IN SECTION 3

B.1 EXPLANATION OF ϵ-SUBROUTINE

Now we explain the motivation for using the subroutine ϵ-SNE to construct policies instead of solving the matrix games with payoff matrices $(Q_h^k(x,\cdot,\cdot), \{r_{f_i,h}(x,\cdot,\cdot)\}_{i\in[N]})$ directly. By the definition of Q_h^k in (3.4), we know Q_h^k relies on the previous data via the estimated value function V_{h+1}^k and feature maps $\{\phi(x_h^\tau, a_h^\tau, b_h^\tau)\}_{\tau=1}^{k-1}$. Similar to the analysis for linear MDPs (Jin et al., 2020b), we need to use a covering argument to establish uniform concentration bounds for all value V_{h+1}^k. Jin et al. (2020b) directly constructs an ϵ-net for the value functions and establishes a polynomial log-covering number for this ϵ-net. This analysis, however, relies on that the policies executed by the players are greedy (deterministic), which is not valid for our setting. To overcome this technical issue, we construct an ϵ-net for Q-functions and solve an approximate matrix game. Fortunately, by choosing a small enough ϵ, we can handle the errors caused by this approximation. See §D for more details. Moreover, as shown in Xie et al. (2020), this subroutine can be implemented efficiently without explicitly computing the exponentially large ϵ-net.

B.2 MORE DISCUSSIONS ON THOEREM 3.3

Learning Stackelberg Equilibria. When there is only one follower, Stackelberg-Nash equilibrium reduces to the Stackelberg equilibrium (Simaan & Cruz, 1973; Conitzer & Sandholm, 2006; Bai et al., 2021). Thus, we partly answer the open problem in Bai et al. (2021) on how to learn Stackelberg equilibria in general-sum Markov games (with myopic followers).

Optimality of the Bound. Assuming that the action of the follower won't affect the transition kernel and reward function, the linear Markov games reduces to the linear MDP (Jin et al., 2020b). Meanwhile, the lower bound established in Azar et al. (2017); Jin et al. (2018) for tabular MDPs and the lower bound established in Lattimore & Szepesvári (2020) for linear bandits directly imply a lower bound $\Omega(dH\sqrt{T})$ for the linear MDPs, which further yields a lower bound $\Omega(dH\sqrt{T})$ for our setting. Ignoring the logarithmic factors, there is only a gap of $\sqrt{dH}$ between this lower bound and our upper bound. We also point out that, by using the "Bernstein-type" bonus (Azar et al., 2017; Jin et al., 2018; Zhou et al., 2020), we can improve our upper bound by a factor of $\sqrt{H}$. Here we do not apply this technique for the clarity of the analysis.

Unknown Reward Setting. To relax the assumption that the reward is known, we consider the case where the reward functions are unknown. At a high level, we first conduct a reward-free exploration algorithm (Algorithm 4 in §E), a variant of Reward-Free RL-Explore algorithm in Jin et al. (2020a), to obtain estimated reward functions $\{\widehat{r}_l, \widehat{r}_{f_1}, \cdots \widehat{r}_{f_N}\}$. As asserted before, we can use Algorithm 1 to find the SNE with respect to the *known* estimated reward functions $\{\widehat{r}_l, \widehat{r}_{f_1}, \cdots \widehat{r}_{f_N}\}$. Hence, we can obtain the approximate SNE if the value functions of estimated value functions are good approximation of the true value functions. See §E for the detailed algorithm and theoretical guarantees.

C Missing Parts in Section 4

C.1 Algorithm

Algorithm 3 Pessimistic Value Iteration to Find Stackelberg-Nash Equilibria

1: **Input:** $\mathcal{D} = \{x_h^\tau, a_h^\tau, b_h^\tau = \{b_{i,h}^\tau\}_{i\in[N]}\}_{\tau,h=1}^{K,H}$ and reward functions $\{r_l, r_f = \{r_{f_i}\}_{i\in[N]}\}$.
2: Initialize $\widehat{V}_{H+1}(\cdot) = 0$.
3: **for** step $h = H, H-1, \cdots, 1$ **do**
4: $\quad \Lambda_h \leftarrow \sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau, b_h^\tau)\phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top + I$.
5: $\quad w_h \leftarrow (\Lambda_h)^{-1}\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau, b_h^\tau)\cdot \widehat{V}_{h+1}(x_{h+1}^\tau)$.
6: $\quad \Gamma_h(\cdot,\cdot,\cdot) \leftarrow \beta' \cdot (\phi(\cdot,\cdot,\cdot)^\top(\Lambda_h)^{-1}\phi(\cdot,\cdot,\cdot))^{1/2}$.
7: $\quad \widehat{Q}_h(\cdot,\cdot,\cdot) \leftarrow r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{\phi(\cdot,\cdot,\cdot)^\top w_h - \Gamma_h(\cdot,\cdot,\cdot)\}$.
8: $\quad (\widehat{\pi}_h(\cdot\,|\,x), \{\widehat{\nu}_{f_i,h}(\cdot\,|\,x)\}_{i\in[N]}) \leftarrow \epsilon\text{-SNE}(\widehat{Q}_h(x,\cdot,\cdot), \{r_{f_i,h}(x,\cdot,\cdot)\}_{i\in[N]}),\ \forall x$. (Alg. 2)
9: $\quad \widehat{V}_h(x) \leftarrow \mathbb{E}_{a\sim\widehat{\pi}_h(\cdot\,|\,x), b_1\sim\widehat{\nu}_{f_1,h}(\cdot\,|\,x),\cdots,b_N\sim\widehat{\nu}_{f_N,h}(\cdot\,|\,x)}\widehat{Q}_h(x,a,b_1,\cdots,b_N),\ \forall x$.
10: **end for**
11: **Output:** $(\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H, \widehat{\nu} = \{\widehat{\nu}_{f_i} = \{\nu_{f_i,h}\}_{h=1}^H\}_{i=1}^N)$.

C.2 More Comments on Theorem 4.2

Minimal Assumption Requirement: Theorem 4.2 only relies on the compliance of the dataset with linear Markov games. Compared with existing literature on offline RL (Bertsekas & Tsitsiklis, 1996; Antos et al., 2007; 2008; Munos & Szepesvári, 2008; Farahmand et al., 2010; 2016; Scherrer et al., 2015; Liu et al., 2018; Chen & Jiang, 2019; Fan et al., 2020; Xie & Jiang, 2020), we impose no restrictions on the coverage of the dataset. Meanwhile, we need no assumption on the affinity between $(\widehat{\pi}, \widehat{\nu})$ and the behavior policies that induce the dataset, which is often employed as a regularizer (Fujimoto et al., 2019; Laroche et al., 2019; Jaques et al., 2019; Wu et al., 2019; Kumar et al., 2019; Wang et al., 2020; Siegel et al., 2020; Nair et al., 2020; Liu et al., 2020b).

Dataset with Sufficient Coverage: In what follows, we specialize Theorem 4.2 to the setting where we assume the dataset with good "coverage". Note that Λ_h is determined by the offline dataset $\mathcal{D}$ and acts as a fixed matrix in the expectation, that is, the expectation in (4.2) is only taken with the trajectory induced by (π^*, ν^*). As proofed in the following theorem, when the trajectory induced by (π^*, ν^*) is "covered" by the dataset $\mathcal{D}$ sufficiently well, we can establish that the suboptimality incurred by Algorithm 3 diminishes at rate of $\widetilde{\mathcal{O}}(1/\sqrt{K})$.

Corollary C.1. Suppose it holds with probability at least $1 - p/2$ that

$$\Lambda_h \succeq I + c\cdot K\cdot \mathbb{E}_{\pi^*,\nu^*,x}[\phi(s_h,a_h,b_h)\phi(s_h,a_h,b_h)^\top]$$

for all $(x,h)\in\mathcal{S}\times[H]$. Here $c>0$ is an absolute constant and $\mathbb{E}_{\pi^*,\nu^*,x}$ is taken with respect to the trajectory incurred by (π^*,ν^*) in the underlying Markov game when initializing the progress at x. Under Assumptions 2.1, 2.2, 3.1 and 4.1, there exists an absolute constant $C>0$ such that, for any fixed $p\in(0,1)$, by setting $\beta' = C\cdot dH\sqrt{\log(4dHK/p)}$ in Line 6 of Algorithm 3 and $\epsilon = \frac{d}{KH}$ in Algorithm 2, then it holds with probability at least $1-p$ that

$$\text{SubOpt}(\widehat{\pi},\widehat{\nu},x) \le \bar{C}\cdot d^{3/2}H^2\sqrt{\log(4dHK/p)/K}$$

for all $x\in\mathcal{S}$. Here $\bar{C}$ is another absolute constant that only depends on c and C.

Proof. See §H for a detailed proof. □

Note that, unlike the previous literature (Antos et al., 2007; Munos & Szepesvári, 2008; Farahmand et al., 2010; 2016; Scherrer et al., 2015; Liu et al., 2018; Chen & Jiang, 2019; Fan et al., 2020; Xie & Jiang, 2020) which relies on the "uniform coverage" assumption, Corollary C.1 only assumes that the dataset has a good coverage of the trajectory incurred by the policies (π^*, ν^*).

Optimality of the Bound: Fix $x \in \mathcal{S}$. Assuming $r_l = r_{f_i}$ for any $i \in [N]$, we know $(\pi^*, \nu^*) = \mathrm{argmax}_{\pi,\nu} V_{l,1}^{\pi,\nu}(x)$. Then the information-theoretic lower bound for offline single-agent RL (e.g., Theorem 4.7 in Jin et al. (2020c)) can imply the information-theoretic lower bound $\Omega(\sum_{h=1}^H \mathbb{E}_{\pi^*,\nu^*,x}[(\phi(s_h,a_h,b_h)^\top(\Lambda_h)^{-1}\phi(s_h,a_h,b_h))^{1/2}])$ for our setting. In particular, our upper bound established in Theorem 4.2 matches this lower bound up to β' and absolute constants and thus implies that our algorithm is nearly minimax optimal.

D Proof of Theorem 3.3

Proof of Theorem 3.3. By the myopic followers assumption, we have the following lemma.

Lemma D.1. For any $k \in [K]$, we have $\nu^k \in \mathrm{BR}(\pi^k)$. Here $\mathrm{BR}(\cdot)$ is defined in (2.10).

Proof. Combing the definition of (π^k, ν^k) in Line 9 of Algorithm 1 and the definition of the best response set in the Markov games with myopic followers in (2.10), we conclude the proof. □

Then we establish an upper bound for the regret defined in (3.1). Recall the regret takes the following form

$$\mathrm{Regret}(K) = \sum_{k=1}^K V_{l,1}^{\pi^*,\nu^*}(x_1^k) - V_{l,1}^{\pi^k,\nu^k}(x_1^k). \tag{D.1}$$

To facilitate our analysis, for any $(k,h) \in [K] \times [H]$ we define the model prediction error by

$$\delta_h^k = r_{l,h} + \mathbb{P}_h V_{h+1}^k - Q_h^k. \tag{D.2}$$

Moreover, for any $(k,h) \in [K] \times [H]$, we define $\zeta_{k,h}^1$ and $\zeta_{k,h}^2$ as

$$\begin{aligned}\zeta_{k,h}^1 &= [V_h^k(x_h^k) - V_{l,h}^{\pi^k,\nu^k}(x_h^k)] - [Q_h^k(x_h^k,a_h^k,b_h^k) - Q_{l,h}^{\pi^k,\nu^k}(x_h^k,a_h^k,b_h^k)],\\ \zeta_{k,h}^2 &= [(\mathbb{P}_h V_{h+1}^k)(x_h^k,a_h^k,b_h^k) - (\mathbb{P}_h V_{l,h+1}^{\pi^k,\nu^k})(x_h^k,a_h^k,b_h^k)] - [V_{h+1}^k(x_{h+1}^k) - V_{l,h+1}^{\pi^k,\nu^k}(x_{h+1}^k)].\end{aligned} \tag{D.3}$$

Recall that $(\pi^k, \nu^k = \{\nu_{f_i}^k\}_{i\in[N]})$ are the policies executed by the leader and the followers in the k-th episode, which generate a trajectory $\{x_h^k, a_h^k, b_h^k = \{b_{i,h}^k\}_{i\in[N]}\}_{h\in[H]}$. Thus, we know that $\zeta_{k,h}^1$ and $\zeta_{k,h}^2$ characterize the randomness of choosing actions $a_h^k \sim \pi_h^k(\cdot \mid x_h^k)$ and $b_h^k \sim \nu_h^k(\cdot \mid x_h^k)$ and the randomness of drawing the next state $x_{h+1}^k \sim \mathcal{P}_h(\cdot \mid x_h^k, a_h^k, b_h^k)$, respectively.

To establish an upper bound for (D.1), we introduce the following lemma, which decomposes this term into three parts using the notations defined above.

Lemma D.2 (Regret Decomposition)**.** We can decompose (D.1) as follows.

$$\begin{aligned}\mathrm{Regret}(K) = &\underbrace{\sum_{k=1}^K\sum_{h=1}^H \mathbb{E}_{\pi^*,\nu^*}[\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot \mid x_h^k) \times \nu_h^*(\cdot \mid x_h^k) - \pi_h^k(\cdot \mid x_h^k) \times \nu_h^k(\cdot \mid x_h^k)\rangle]}_{(l.1):\ \text{Computational Error}}\\ &+ \underbrace{\sum_{k=1}^K\sum_{h=1}^H (\mathbb{E}_{\pi^*,\nu^*}[\delta_h^k(x_h,a_h,b_h)] - \delta_h^k(x_h^k,a_h^k,b_h^k))}_{(l.2):\ \text{Statistical Error}} + \underbrace{\sum_{k=1}^K\sum_{h=1}^H(\zeta_{k,h}^1 + \zeta_{k,h}^2)}_{(l.3):\ \text{Randomness}},\end{aligned}$$

where $\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot \mid x_h^k) \times \nu_h^*(\cdot \mid x_h^k) - \pi_h^k(\cdot \mid x_h^k) \times \nu_h^k(\cdot \mid x_h^k)\rangle = \langle Q_h^k(x_h^k,\cdot,\cdot,\cdots,\cdot), \pi_h^*(\cdot \mid x_h^k) \times \nu_{f_1,h}^*(\cdot \mid x_h^k) \times \cdots \nu_{f_N,h}^*(\cdot \mid x_h^k) - \pi_h^k(\cdot \mid x_h^k) \times \nu_{f_1,h}^k(\cdot \mid x_h^k) \times \cdots \nu_{f_N,h}^k(\cdot \mid x_h^k)\rangle_{\mathcal{A}_l \times \mathcal{A}_f}$.

Proof. See §D.1 for a detailed proof. □

Remark D.3. Similar regret decomposition results also appear in the single-agent RL literature (Cai et al., 2020; Efroni et al., 2020; Yang et al., 2020), and they can be regarded as the special case of Lemma D.2. Moreover, our regret decomposition lemma is independent of the myopic followers assumption, and thus, can be applied to more general settings.

Lemma D.2 states that the regret has three sources: (i) computational error, which represents the convergence of the algorithm with the known model, (ii) statistical error, that is, the error caused by the inaccurate estimation of the model, and (iii) randomness, as aforementioned, which comes from executing random policies and interaction with random environment.

Returning to the main proof, we only need to characterize these three types of errors, respectively. We first characterize the computational error by the following lemma.

Lemma D.4 (Optimization Error)**.** It holds that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\nu^*}[\langle Q_h^k(x_h^k,\cdot,\cdot),\pi_h^*(\cdot\,|\,x_h^k)\times\nu_h^*(\cdot\,|\,x_h^k)-\pi_h^k(\cdot\,|\,x_h^k)\times\nu_h^k(\cdot\,|\,x_h^k)\rangle]\le\epsilon KH.$$

Proof. See §D.2 for a detailed proof. □

Then, we establish an upper bound for the statistical error. Due to the uncertainty that arises from only observing limited data, the model prediction errors can be possibly large for the triple (x, a, b) that are less visited or even unseen. Fortunately, however, we have the following lemma which characterizes the model prediction errors defined in (D.2).

Lemma D.5 (Optimism)**.** It holds with probability at least $1 - p/2$ that

$$-2\min\{H,\Gamma_h^k(x,a,b)\}\le\delta_h^k(x,a,b)\le 0$$

for any $(k,h)\in[K]\times[H]$ and $(x,a,b)\in\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f$.

Proof. See §D.3 a detailed proof. □

Lemma D.5 states that $\delta_h^k(x,a,b)\le 0$ for any $(x,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{A}$. Combining the definition of model prediction error in (D.2), we obtain

$$Q_h^k(x,a,b)\ge r_{l,h}(x,a,b)+(\mathbb{P}_h V_{h+1}^k)(x,a,b),$$

which further implies that the estimated Q-function $Q_{\star,h}^k$ is "optimistic in the face of uncertainty". Moreover, Lemma D.5 implies that $-\delta_h^k(x,a,b)\le 2\min\{H,\Gamma_h^k(x,a,b)\}$. Thus we only need to establish an upper bound for $2\sum_{k=1}^K\sum_{h=1}^H\min\{H,\Gamma_h^k(x_h^k,a_h^k,b_h^k)\}$, which is the total price paid for the optimism. As shown in the following lemma, we can derive an upper bound for this term by the elliptical potential lemma (Abbasi-Yadkori et al., 2011).

Lemma D.6. For the bonus function Γ_h^k defined in Line 7 of Algorithm 1, it holds that

$$2\sum_{k=1}^{K}\sum_{h=1}^{H}\min\{H,\Gamma_h^k(x_h^k,a_h^k,b_h^k)\}\le\mathcal{O}(\sqrt{d^3H^3T\iota^2}).$$

Here $p\in(0,1)$ and $\iota=\log(2dT/p)$ are defined in Theorem 3.3.

Proof. See §D.4 for a detailed proof. □

It remains to analyze the randomness, which is the purpose of the following lemma.

Lemma D.7. For the $\zeta_{k,h}^1$ and $\zeta_{k,h}^2$ defined in Lemma D.2 and any $p\in(0,1)$, it holds with probability at least $1-p/2$ that

$$\sum_{k=1}^{K}\sum_{h=1}^{H}(\zeta_{k,h}^1+\zeta_{k,h}^2)\le\sqrt{16KH^3\cdot\log(4/p)}.$$

Proof. See §D.5 for a detailed proof. □

Putting above lemmas together, we obtain

$$\text{Regret}(K) \le \mathcal{O}(\sqrt{d^3 H^3 T \iota^2}) \tag{D.4}$$

with probability at least $1-p$, which concludes the proof of Theorem 3.3. □

D.1 Proof of Lemma D.2

First, we establish a more general regret decomposition lemma, which immediately implies Lemma D.2.

Lemma D.8 (General Decomposition for One Episode)**.** Fix $k \in [K]$. Suppose $(\pi^k, \nu^k = \{\nu^k_{f_i}\}_{i\in[N]})$ are the policies executed by the leader l and the followers $\{f_i\}_{i\in[N]}$ in the k-th episode. Moreover, suppose that $Q^k_{\star,h}$ and $V^k_{\star,h} = \langle Q^k_{\star,h}, \pi^k_h \times \nu^k_h\rangle$ are the estimated Q-function and value function for any $\star \in \{l, f_1, \cdots, f_N\}$ at h-th step of k-th episode. Then, for any policies $(\pi, \nu = \{\nu_{f_i}\}_{i\in[N]})$ and $\star \in \{l, f_1, \cdots, f_N\}$, we have

$$\begin{aligned} &V^{\pi,\nu}_{\star,1}(x^k_1) - V^{\pi^k,\nu^k}_{\star,1}(x^k_1) \\ &\quad = \underbrace{\sum_{h=1}^H \mathbb{E}_{\pi,\nu}[\langle Q^k_{\star,h}(x^k_h,\cdot,\cdot), \pi_h(\cdot\,|\,x^k_h)\times\nu_h(\cdot\,|\,x^k_h) - \pi^k_h(\cdot\,|\,x^k_h)\times\nu^k_h(\cdot\,|\,x^k_h)\rangle]}_{\text{Computational Error}} \\ &\qquad + \underbrace{\sum_{h=1}^H \big(\mathbb{E}_{\pi,\nu}[\delta^k_{\star,h}(x_h,a_h,b_h)] - \delta^k_{\star,h}(x^k_h,a^k_h,b^k_h)\big)}_{\text{Statistical Error}} + \underbrace{\sum_{h=1}^H (\zeta^1_{\star,k,h} + \zeta^2_{\star,k,h})}_{\text{Randomness}}, \end{aligned}$$

where $\langle Q^k_{\star,h}, \pi^k_h \times \nu^k_h\rangle = \langle Q^k_{\star,h}, \pi^k_h \times \nu^k_{f_1,h}\times\cdots\times\nu^k_{f_N,h}\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$ and $\langle Q^k_h(x^k_h,\cdot,\cdot), \pi^*_h(\cdot\,|\,x^k_h)\times\nu^*_h(\cdot\,|\,x^k_h) - \pi^k_h(\cdot\,|\,x^k_h)\times\nu^k_h(\cdot\,|\,x^k_h)\rangle = \langle Q^k_h(x^k_h,\cdot,\cdot,\cdots,\cdot), \pi^*_h(\cdot\,|\,x^k_h)\times\nu^*_{f_1,h}(\cdot\,|\,x^k_h)\times\cdots\nu^*_{f_N,h}(\cdot\,|\,x^k_h) - \pi^k_h(\cdot\,|\,x^k_h)\times\nu^k_{f_1,h}(\cdot\,|\,x^k_h)\times\cdots\nu^k_{f_N,h}(\cdot\,|\,x^k_h)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$. Here $\delta^k_{\star,h}$ is the model prediction error defined by

$$\delta^k_{\star,h} = r_{\star,h} + \mathbb{P}_h V^k_{\star,h+1} - Q^k_{\star,h}, \tag{D.5}$$

and $\zeta^1_{\star,k,h}$ and $\zeta^2_{\star,k,h}$ are defined by

$$\begin{aligned} \zeta^1_{\star,k,h} &= [V^k_{\star,h}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h}(x^k_h)] - [Q^k_{\star,h}(x^k_h,a^k_h,b^k_h) - Q^{\pi^k,\nu^k}_{\star,h}(x^k_h,a^k_h,b^k_h)], \\ \zeta^2_{\star,k,h} &= [(\mathbb{P}_h V^k_{\star,h+1})(x^k_h,a^k_h,b^k_h) - (\mathbb{P}_h V^{\pi^k,\nu^k}_{\star,h+1})(x^k_h,a^k_h,b^k_h)] - [V^k_{\star,h+1}(x^k_{h+1}) - V^{\pi^k,\nu^k}_{\star,h+1}(x^k_{h+1})]. \end{aligned} \tag{D.6}$$

Proof of Lemma D.8. To facilitate our analysis, for any $\nu = \{\nu_{f_i}\}_{i\in[N]}$ and $(h,x)\in[H]\times\mathcal{S}$, we denote $\nu_{f_1,h}(\cdot\,|\,x)\times\cdots\nu_{f_N,h}(\cdot\,|\,x)$ by $\nu_h(\cdot\,|\,x)$. Moreover, we define two operators $\mathbb{J}_h$ and $\mathbb{J}_{k,h}$ respectively by

$$\begin{aligned} (\mathbb{J}_h f)(x) &= \langle f(x,\cdot,\cdot), \pi_h(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle, \\ (\mathbb{J}_{k,h} f)(x) &= \langle f(x,\cdot,\cdot), \pi^k_h(\cdot\,|\,x)\times\nu^k_h(\cdot\,|\,x)\rangle \end{aligned} \tag{D.7}$$

for any $h\in[H]$ and any function $f:\mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f\to\mathbb{R}$. Also, we define

$$\begin{aligned} \xi^k_{\star,h}(x) &= (\mathbb{J}_h Q^k_{\star,f})(x) - (\mathbb{J}_{k,h} Q^k_{\star,f})(x) \\ &= \langle Q^k_{\star,h}(x,\cdot,\cdot), \pi_h(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x) - \pi^k_h(\cdot\,|\,x)\times\nu^k_h(\cdot\,|\,x)\rangle \end{aligned} \tag{D.8}$$

for any $(h,x)\in[H]\times\mathcal{S}$ and $\star\in\{l,f_1,\cdots,f_N\}$.

Under the above notations, we decompose the regret at the k-th episode into the following two terms,

$$V^{\pi,\nu}_{\star,1}(x^k_1) - V^{\pi^k,\nu^k}_1(x^k_1) = \underbrace{V^{\pi,\nu}_{\star,1}(x^k_1) - V^k_{\star,1}(x^k_1)}_{\text{(i)}} + \underbrace{V^k_{\star,1}(x^k_1) - V^{\pi^k,\nu^k}_1(x^k_1)}_{\text{(ii)}}. \tag{D.9}$$

Then we characterize these two terms respectively.

Term (i). By the Bellman equation in (2.2) and the definition of the operator $\mathbb{J}_h$ in (D.7), we have $V^{\pi,\nu}_{\star,h} = \mathbb{J}_h Q^{\pi,\nu}_{\star,h}$. Similar, by the definition of $V^k_{\star,h}$ and the definition of the operator $\mathbb{J}_{k,h}$ in (D.7), we have $V^k_{\star,h} = \mathbb{J}_{k,h} Q^k_{\star,h}$. Hence, for any $h \in [H]$, we have

$$\begin{aligned} V^{\pi,\nu}_{\star,h} - V^k_{\star,h} &= \mathbb{J}_h Q^{\pi,\nu}_{\star,h} - \mathbb{J}_{k,h} Q^k_{\star,h} = (\mathbb{J}_h Q^{\pi,\nu}_{\star,h} - \mathbb{J}_h Q^k_{\star,h}) + (\mathbb{J}_h Q^k_{\star,h} - \mathbb{J}_{k,h} Q^k_{\star,h}) \\ &= \mathbb{J}_h (Q^{\pi,\nu}_{\star,h} - Q^k_{\star,h}) + \xi^k_{\star,h}, \end{aligned} \tag{D.10}$$

where the last inequality is obtained by the fact that $\mathbb{J}_h$ is a linear operator and the definition of $\xi^k_{\star,h}$ in (D.8). Meanwhile, by the Bellman equation in (2.2) and the definition of the prediction error $\delta^k_{\star,h}$ in (D.2), we obtain

$$\begin{aligned} Q^{\pi,\nu}_{\star,h} - Q^k_{\star,h} &= (r_{\star,h} + \mathbb{P}_h V^{\pi,\nu}_{\star,h+1}) - (r_{\star,h} + \mathbb{P}_h V^k_{\star,h+1} - \delta^k_{\star,h}) \\ &= \mathbb{P}_h (V^{\pi,\nu}_{\star,h+1} - V^k_{\star,h+1}) + \delta^k_{\star,h}. \end{aligned} \tag{D.11}$$

Putting (D.10) and (D.11) together, we further obtain

$$V^{\pi,\nu}_{\star,h} - V^k_{\star,h} = \mathbb{J}_h \mathbb{P}_h (V^{\pi,\nu}_{\star,h+1} - V^k_{\star,h+1}) + \mathbb{J}_h \delta^k_{\star,h} + \xi^k_{\star,h} \tag{D.12}$$

for any $h \in [H]$ and $\star \in \{l, f_1, \cdots, f_N\}$. By recursively applying (D.12) for all $h \in [H]$, we have

$$\begin{aligned} V^{\pi,\nu}_{\star,1} - V^k_{\star,1} &= \Big(\prod_{h=1}^H \mathbb{J}_h \mathbb{P}_h\Big)(V^{\pi,\nu}_{\star,H+1} - V^{\pi^k,\nu^k}_{\star,H+1}) + \sum_{h=1}^H \Big(\sum_{i=1}^h \mathbb{J}_i \mathbb{P}_i\Big) \mathbb{J}_h \delta^k_{\star,h} + \sum_{h=1}^H \Big(\sum_{i=1}^h \mathbb{J}_i \mathbb{P}_i\Big) \xi^k_{\star,h} \\ &= \sum_{h=1}^H \Big(\sum_{i=1}^h \mathbb{J}_i \mathbb{P}_i\Big) \mathbb{J}_h \delta^k_{\star,h} + \sum_{h=1}^H \Big(\sum_{i=1}^h \mathbb{J}_i \mathbb{P}_i\Big) \xi^k_{\star,h}, \end{aligned} \tag{D.13}$$

where the last equality follows from the fact that $V^{\pi,\nu}_{\star,H+1} = V^{\pi^k,\nu^k}_{\star,H+1} = 0$. Thus, by utilizing the definition of $\xi^k_{\star,h}$ in (D.8), we further obtain

$$\begin{aligned} V^{\pi,\nu}_{\star,1}(x^k_1) - V^k_{\star,1}(x^k_1) &= \mathbb{E}_{\pi,\nu}\Big[\sum_{h=1}^H \langle Q^k_{\star,h}(x^k_h,\cdot,\cdot), \pi_h(\cdot \,|\, x^k_h) \times \nu_h(\cdot \,|\, x^k_h) - \pi^k_h(\cdot \,|\, x^k_h) \times \nu^k_h(\cdot \,|\, x^k_h)\rangle\Big] \\ &\quad + \mathbb{E}_{\pi,\nu}\Big[\sum_{h=1}^H \delta^k_{\star,h}(x_h, a_h, b_h)\Big] \end{aligned} \tag{D.14}$$

for any $k \in [K]$ and $\star \in \{l, f_1, \cdots, f_N\}$.

Term (ii). Recall that we denote $\{b^k_{f_i,h}\}_{i\in[N]}$ by b^k_h for any $h \in [H]$. Then, for any $h \in [H]$ and $\star \in \{l, f_1, \cdots, f_N\}$, by the definition of model prediction error in (D.5), we have

$$\begin{aligned} \delta^k_{\star,h}(x^k_h, a^k_h, b^k_h) &= r^k_{\star,h}(x^k_h, a^k_h, b^k_h) + (\mathbb{P}_h V^k_{\star,h+1})(x^k_h, a^k_h, b^k_h) - Q^k_{\star,h}(x^k_h, a^k_h, b^k_h) \\ &= [r^k_{\star,h}(x^k_h, a^k_h, b^k_h) + (\mathbb{P}_h V^k_{\star,h+1})(x^k_h, a^k_h, b^k_h) - Q^{\pi^k,\nu^k}_{\star,h}(x^k_h, a^k_h, b^k_h)] \\ &\qquad + [Q^{\pi^k,\nu^k}_{\star,h}(x^k_h, a^k_h, b^k_h) - Q^k_{\star,h}(x^k_h, a^k_h, b^k_h)] \\ &= (\mathbb{P}_h V^k_{\star,h+1} - \mathbb{P}_h V^{\pi^k,\nu^k}_{\star,h+1})(x^k_h, a^k_h, b^k_h) + (Q^{\pi^k,\nu^k}_{\star,h} - Q^k_{\star,h})(x^k_h, a^k_h, b^k_h) \end{aligned} \tag{D.15}$$

where the last equation is obtained by the Bellman equation in (2.2). Thus, by (D.15), we have

$$\begin{aligned} &V^k_{\star,h}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h}(x^k_h) \\ &= V^k_{\star,h}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h}(x^k_h) + (Q^{\pi^k,\nu^k}_{\star,h} - Q^k_{\star,h})(x^k_h, a^k_h, b^k_h) \\ &\qquad + (\mathbb{P}_h V^k_{\star,h+1} - \mathbb{P}_h V^{\pi^k,\nu^k}_{\star,h+1})(x^k_h, a^k_h, b^k_h) - \delta^k_{\star,h}(x^k_h, a^k_h, b^k_h) \\ &= V^k_{\star,h}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h}(x^k_h) - (Q^k_{\star,h} - Q^{\pi^k,\nu^k}_{\star,h})(x^k_h, a^k_h, b^k_h) \\ &\qquad + \big(\mathbb{P}_h (V^k_{\star,h+1} - V^{\pi^k,\nu^k}_{\star,h+1})\big)(x^k_h, a^k_h, b^k_h) - (V^k_{\star,h+1} - V^{\pi^k,\nu^k}_{\star,h+1})(x^k_h) \\ &\qquad + (V^k_{\star,h+1} - V^{\pi^k,\nu^k}_{\star,h+1})(x^k_h) - \delta^k_{\star,h}(x^k_h, a^k_h, b^k_h) \end{aligned} \tag{D.16}$$

for any $h \in [H]$ and $\star \in \{l, f_1, \cdots, f_N\}$. By the definitions of $\zeta^1_{\star,k,h}$ and $\zeta^2_{\star,k,h}$ in (D.6), (D.16) can be written as

$$V^k_{\star,h}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h}(x^k_h) = [V^k_{\star,h+1}(x^k_h) - V^{\pi^k,\nu^k}_{\star,h+1}(x^k_h)] + \zeta^1_{\star,k,h} + \zeta^2_{\star,k,h} - \delta^k_{\star,h}(x^k_h, a^k_h, b^k_h). \tag{D.17}$$

For any $\star \in \{l, f_1, \cdots, f_N\}$, recursively expanding (D.17) across $h \in [H]$ yields

$$\begin{aligned} &V^k_{\star,1}(x^k_1) - V^{\pi^k,\nu^k}_{\star,1}(x^k_1) \\ &\quad = V^k_{\star,H+1}(x^k_{H+1}) - V^{\pi^k,\nu^k}_{\star,H+1}(x^k_{H+1}) + \sum_{h=1}^{H}(\zeta^1_{\star,k,h} + \zeta^2_{\star,k,h}) - \sum_{h=1}^{H}\delta^k_{\star,h}(x^k_h, a^k_h, b^k_h) \\ &\quad = \sum_{h=1}^{H}(\zeta^1_{\star,k,h} + \zeta^2_{\star,k,h}) - \sum_{h=1}^{H}\delta^k_{\star,h}(x^k_h, a^k_h, b^k_h), \end{aligned} \tag{D.18}$$

where the last equality follows from the fact that $V^k_{\star,H+1}(x^k_{H+1}) = V^{\pi^k,\nu^k}_{\star,H+1}(x^k_{H+1}) = 0$.

Plugging (D.14) and (D.18) into (D.9), we obtain

$$\begin{aligned} &V^{\pi,\nu}_{\star,1}(x^k_1) - V^{\pi^k,\nu^k}_{\star,1}(x^k_1) \\ &\quad = \underbrace{\sum_{h=1}^{H}\mathbb{E}_{\pi,\nu}[\langle Q^k_{\star,h}(x^k_h,\cdot,\cdot), \pi_h(\cdot\,|\,x^k_h)\times\nu_h(\cdot\,|\,x^k_h) - \pi^k_h(\cdot\,|\,x^k_h)\times\nu^k_h(\cdot\,|\,x^k_h)\rangle]}_{\text{Computational Error}} \\ &\qquad + \underbrace{\sum_{h=1}^{H}\big(\mathbb{E}_{\pi,\nu}[\delta^k_{\star,h}(x_h,a_h,b_h)] - \delta^k_{\star,h}(x^k_h,a^k_h,b^k_h)\big)}_{\text{Statistical Error}} + \underbrace{\sum_{h=1}^{H}(\zeta^1_{\star,k,h} + \zeta^2_{\star,k,h})}_{\text{Randomness}} \end{aligned}$$

for any (π, ν) and $\star \in \{l, f_1, \cdots, f_N\}$. Therefore, we conclude the proof of Lemma D.2. □

Proof of Lemma D.2. For any $k \in [K]$, applying Lemma D.8 with $(\pi, \nu) = (\pi^*, \nu^*)$, we obtain

$$\begin{aligned} &V^{\pi^*,\nu^*}_{l,1}(x^k_1) - V^{\pi^k,\nu^k}_{l,1}(x^k_1) \\ &\quad = \sum_{h=1}^{H}\mathbb{E}_{\pi^*,\nu^*}[\langle Q^k_h(x^k_h,\cdot,\cdot), \pi^*_h(\cdot\,|\,x^k_h)\times\nu^*_h(\cdot\,|\,x^k_h) - \pi^k_h(\cdot\,|\,x^k_h)\times\nu^k_h(\cdot\,|\,x^k_h)\rangle] \\ &\qquad + \sum_{h=1}^{H}\big(\mathbb{E}_{\pi^*,\nu^*}[\delta^k_h(x_h,a_h,b_h)] - \delta^k_h(x^k_h,a^k_h,b^k_h)\big) + \sum_{h=1}^{H}(\zeta^1_{k,h} + \zeta^2_{k,h}). \end{aligned}$$

Taking summation over $k \in [K]$, we decompose (D.1) as desired, which concludes the proof of Lemma D.2.

□

D.2 Proof of Lemma D.4

Proof of Lemma D.4. By the myopic followers assumption, we have that, for the matrix game with payoff matrices $(\widetilde{Q}(x^k_h,\cdot,\cdot), \{r^k_{f_i,h}(x^k_h,\cdot,\cdot)\}_{i\in[N]})$, $\nu^*_h(\cdot\,|\,x^k_h)$ belongs to the best response set of $\pi^*_h(\cdot\,|\,x^k_h)$. Moreover, we define $\widetilde{\nu}^*_h(\cdot\,|\,x^k_h)$ as the policy belongs to the best response set of $\pi^*_h(\cdot\,|\,x^k_h)$ and breaks ties in favor of the leader.

Recall that $(\pi^k_h(\cdot\,|\,x^k_h), \nu^k_h(\cdot\,|\,x^k_h) = \{\nu^k_{f_i,h}(\cdot\,|\,x^k_h)\}_{i\in[N]})$ is the Stackelberg-Nash equilibrium of the matrix game with payoff matrices $(\widetilde{Q}(x^k_h,\cdot,\cdot,\cdot), \{r^k_{f_i,h}(x^k_h,\cdot,\cdot,\cdot)\}_{i\in[N]})$, which implies that $\pi^k_h(\cdot\,|\,x^k_h)$ is the "best response to the best response", which further implies that

$$\langle \widetilde{Q}(x^k_h,\cdot,\cdot), \pi^*_h(\cdot\,|\,x^k_h)\times\widetilde{\nu}^*_h(\cdot\,|\,x^k_h) - \pi^k_h(\cdot\,|\,x^k_h)\times\nu^k_h(\cdot\,|\,x^k_h)\rangle \le 0 \tag{D.19}$$

for any $(k,h) \in [K] \times [H]$. Thus, for any $(k,h) \in [K] \times [H]$, we have

$$\begin{aligned}
&\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot \,|\, x_h^k) \times \nu_h^*(\cdot \,|\, x_h^k) - \pi_h^k(\cdot \,|\, x_h^k) \times \nu_h^k(\cdot \,|\, x_h^k)\rangle \\
&\quad = \langle \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot \,|\, x_h^k) \times \nu_h^*(\cdot \,|\, x_h^k) - \pi_h^k(\cdot \,|\, x_h^k) \times \nu_h^k(\cdot \,|\, x_h^k)\rangle \\
&\qquad + \langle Q_h^k(x_h^k,\cdot,\cdot) - \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot \,|\, x_h^k) \times \nu_h^*(\cdot \,|\, x_h^k) - \pi_h^k(\cdot \,|\, x_h^k) \times \nu_h^k(\cdot \,|\, x_h^k)\rangle \qquad \text{(D.20)} \\
&\quad \le \langle \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot \,|\, x_h^k) \times \widetilde{\nu}_h^*(\cdot \,|\, x_h^k) - \pi_h^k(\cdot \,|\, x_h^k) \times \nu_h^k(\cdot \,|\, x_h^k)\rangle \\
&\qquad + \langle Q_h^k(x_h^k,\cdot,\cdot) - \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot \,|\, x_h^k) \times \nu_h^*(\cdot \,|\, x_h^k) - \pi_h^k(\cdot \,|\, x_h^k) \times \nu_h^k(\cdot \,|\, x_h^k)\rangle \\
&\quad \le \epsilon, \qquad \text{(D.21)}
\end{aligned}$$

where the first inequality follows from the definition of $\widetilde{\nu}_h^k(\cdot \,|\, x_h^k)$ and the last inequality uses (D.19) and the fact that $\|Q_h^k - \widetilde{Q}\|_\infty \le \epsilon$. By taking summation over $(k,h) \in [K] \times [H]$, we conclude the proof of Lemma D.4.

□

D.3 Proof of Lemma D.5

Proof of Lemma D.5. Recall that the estimated Q-function Q_h^k defined in Line 8 of Algorithm 1 takes the following form:

$$\begin{aligned}
&Q_h^k(\cdot,\cdot,\cdot) \leftarrow r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{\phi(\cdot,\cdot,\cdot)^\top w_h^k + \Gamma_h^k(\cdot,\cdot,\cdot)\}, \\
&\text{where } w_h^k = (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot V_{h+1}^k(x_{h+1}^\tau)\Big).
\end{aligned} \qquad \text{(D.22)}$$

Here Λ_h^k and Γ_h^k are defined in Lines 5 and 7 of Algorithm 1, respectively. Meanwhile, by Assumption 2.1, we have

$$\begin{aligned}
(\mathbb{P}_h V_{h+1}^k)(x,a,b) &= \phi(x,a,b)^\top \langle \mu_h, V_{h+1}^k\rangle \\
&= \phi(x,a,b)^\top (\Lambda_h^k)^{-1}\Lambda_h^k \langle \mu_h, V_{h+1}^k\rangle
\end{aligned} \qquad \text{(D.23)}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$. Here $\langle \mu_h, V_{h+1}^k\rangle = \int_{\mathcal{S}} V_{h+1}^k(x')\mathrm{d}\mu_h(x')$. Together with the definition of Λ_h^k in Line 5 of Algorithm 1, we further obtain

$$\begin{aligned}
(\mathbb{P}_h V_{h+1}^k)(x,a,b) &= \phi(x,a,b)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau)\phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top \langle \mu_h, V_{h+1}^k\rangle + \langle \mu_h, V_{h+1}^k\rangle\Big) \\
&= \phi(x,a,b)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot (\mathbb{P}_h V_{h+1}^k)(x_h^\tau, a_h^\tau, b_h^\tau) + \langle \mu_h, V_{h+1}^k\rangle\Big),
\end{aligned} \qquad \text{(D.24)}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$. Here the last equality uses (D.23). Putting (D.22) and (D.24) together, we have

$$\begin{aligned}
&\phi(x,a,b)^\top w_h^k - (\mathbb{P}_h V_{h+1}^k)(x,a,b) \\
&\quad = \underbrace{\phi(x,a,b)^\top (\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot \big(V_{h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(x_h^\tau, a_h^\tau, b_h^\tau)\big)\Big)}_{\text{(i)}} \qquad \text{(D.25)} \\
&\qquad - \underbrace{\phi(x,a,b)^\top (\Lambda_h^k)^{-1} \langle \mu_h, V_{h+1}^k\rangle}_{\text{(ii)}}
\end{aligned}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$. Then we upper bound these two terms respectively.

Term (i). By Cauchy-Schwarz inequality, we have

$$|\text{(i)}| \le \|\phi(x,a,b)\|_{(\Lambda_h^k)^{-1}} \cdot \Big\| \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot \big(V_{h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(x_h^\tau, a_h^\tau, b_h^\tau)\big)\Big\|_{(\Lambda_h^k)^{-1}} \qquad \text{(D.26)}$$

for any $(k,h,x,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l$. Under the event $\mathcal{E}$ defined in Lemma D.9, we further have

$$|(\mathrm{i})| \le C' dH \sqrt{\log(2dT/p)} \cdot \|\phi(x,a)\|_{(\Lambda_h^k)^{-1}} \tag{D.27}$$

for any $(k,h,x,a) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l$.

Term (ii). Similarly, by Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} |(\mathrm{ii})| &\le \|\phi(x,a,b)\|_{(\Lambda_h^k)^{-1}} \cdot \|\langle \mu_h, V_{h+1}^k \rangle\|_{(\Lambda_h^k)^{-1}} \\ &\le \|\phi(x,a,b)\|_{(\Lambda_h^k)^{-1}} \cdot \|\langle \mu_h, V_{h+1}^k \rangle\|_2 \le \sqrt{d}H \cdot \|\phi(x,a,b)\|_{(\Lambda_h^k)^{-1}} \end{aligned} \tag{D.28}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$. Here the second inequality follows from the fact that $\Lambda_h^k \succeq I$ and the last inequality is obtained by

$$\|\langle \mu_h, V_{h+1}^k \rangle\|_2 \le \|\mu_h(\mathcal{S})\|_2 \cdot \|V_{h+1}^k\|_\infty \le \sqrt{d}H.$$

Here we use the fact that $\|V_{h+1}^k\|_\infty \le H$ and Assumption 2.1, which assumes $\|\mu_h(\mathcal{S})\|_2 \le \sqrt{d}$. Plugging (D.27) and (D.28) into (D.25), we obtain that

$$|\phi(x,a,b)^\top w_h^k - (\mathbb{P}_h V_{h+1}^k)(x,a,b)| \le CdH\sqrt{\log(2dT/p)} \cdot \|\phi(x,a,b)\|_{(\Lambda_h^k)^{-1}} \tag{D.29}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$ under the event $\mathcal{E}$. Here $C > 0$ is a constant. By setting

$$\beta = CdH\sqrt{\log(2dT/p)} \tag{D.30}$$

in Line 7 of Algorithm 1, (D.29) gives

$$|\phi(x,a,b)^\top w_h^k - (\mathbb{P}_h V_{h+1}^k)(x,a,b)| \le \Gamma_h^k(x,a,b) \tag{D.31}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$ under the event $\mathcal{E}$. Meanwhile, by the truncation in Line 8 of Algorithm 1 and the fact that $r_{l,h} \in [-1,1]$, we have $Q_h^k \in [-(H-h+1), H-h+1]$, which further implies that

$$V_h^k \in [-(H-h+1), H-h+1] \tag{D.32}$$

for any $(k,h) \in [K] \times [H]$. Hence, by (D.31), we have

$$\phi(x,a,b)^\top w_h^k + \Gamma_h^k(x,a,b) \ge (\mathbb{P}_h V_{h+1}^k)(x,a,b) \ge H-h \tag{D.33}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$ under the event $\mathcal{E}$, where the last inequality is obtained by (D.32). Thus, for the model prediction error defined in (D.2), we have

$$\begin{aligned} -\delta_h^k(x,a,b) &= Q_h^k(x,a,b) - r_{l,h}(x,a,b) - \mathbb{P}_h V_{h+1}^k(x,a,b) \\ &\le \phi(x,a,b)^\top w_h^k + \Gamma_h^k(x,a,b) - \mathbb{P}_h V_{h+1}^k(x,a,b) \\ &\le 2\Gamma_h^k(x,a,b) \end{aligned} \tag{D.34}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$ under the event $\mathcal{E}$. Moreover, by the definition of the model prediction error, we have $-\delta_h^k(\cdot,\cdot,\cdot) \le 2H$. Together with (D.34), we have

$$-\delta_h^k(x,a,b) \le 2\min\{H, \Gamma_h^k(x,a,b)\} \tag{D.35}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$ under the event $\mathcal{E}$. On the other hand, by (3.4), we have

$$\begin{aligned} \delta_h^k(x,a,b) &= r_{l,h}(x,a,b) + \mathbb{P}_h V_{h+1}^k(x,a,b) - Q_h^k(x,a,b) \\ &\le \mathbb{P}_h V_{h+1}^k(x,a,b) - \min\{\phi(x,a,b)^\top w_h^k + \Gamma_h^k(x,a,b), H-h\} \\ &= \max\{\mathbb{P}_h V_{h+1}^k(x,a,b) - \phi(x,a,b)^\top w_h^k - \Gamma_h^k(x,a,b), \mathbb{P}_h V_{h+1}^k(x,a,b) - (H-h)\} \\ &\le 0 \end{aligned} \tag{D.36}$$

for any $(k,h,x,a,b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$ under the event $\mathcal{E}$. Here the last inequality follows from (D.31) and the fact that $V_{h+1}^k \le H-h$. Combining (D.35) and (D.36), we conclude the proof of Lemma D.5. $\square$

Lemma D.9. For any $p \in (0,1]$, the event $\mathcal{E}$ that, for any $(k,h) \in [K] \times [H]$,

$$\Big\| \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot \big(V_{h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(x_h^\tau, a_h^\tau, b_h^\tau)\big) \Big\|_{(\Lambda_h^k)^{-1}} \le C' dH \sqrt{\log(2dT/p)}$$

happens with probability at least $1 - p/2$, where $C' > 0$ is an absolute constant.

Proof of Lemma D.9. Fix $(k,h) \in [K] \times [H]$. By Lemma D.10, we have $w_{h+1}^k \le H\sqrt{dk}$, which implies that $Q_{h+1}^k \in \mathcal{Q}_{h+1,\epsilon}^k$. Here $\mathcal{Q}_{h+1,\epsilon}^k$ is defined in (3.5). Moreover, as shown in Algorithm 2, we find a $\widetilde{Q}$ in the ϵ-net $\mathcal{Q}_{h+1,\epsilon}^k$ such that $\|Q_{h+1}^k - \widetilde{Q}\|_\infty \le \epsilon$. For any $x \in \mathcal{S}$, let $(\widetilde{\pi}(\cdot \,|\, x), \widetilde{\nu} = \{\nu_{f_i}\}_{i=1}^N)$ be the Stackelberg-Nash equilibrium of the matrix game with payoff matrices $(\widetilde{Q}(x,\cdot,\cdot), \{r_{f_i,h+1}(x,\cdot,\cdot)\}_{i=1}^N)$. Moreover, we define $\widetilde{V}(x) = \mathbb{E}_{a\sim\widetilde{\pi}(\cdot\,|\,x), b\sim\widetilde{\nu}(\cdot\,|\,x)}[\widetilde{Q}(x,a,b)]$ for any $x \in \mathcal{S}$. Then, we have

$$\begin{aligned}
&\Big\| \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot \big(V_{h+1}^k(x_{h+1}^\tau) - (\mathbb{P}_h V_{h+1}^k)(x_h^\tau, a_h^\tau, b_h^\tau)\big) \Big\|_{(\Lambda_h^k)^{-1}} \\
&\quad \le \underbrace{\Big\| \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot \big(\widetilde{V}(x_{h+1}^\tau) - (\mathbb{P}_h \widetilde{V})(x_h^\tau, a_h^\tau, b_h^\tau)\big) \Big\|_{(\Lambda_h^k)^{-1}}}_{\text{(i)}} \qquad \text{(D.37)} \\
&\qquad + \underbrace{\Big\| \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot \big([V_{h+1}^k(x_{h+1}^\tau) - \widetilde{V}(x_{h+1}^\tau)] - \big(\mathbb{P}_h(V_{h+1}^k - \widetilde{V})\big)(x_h^\tau, a_h^\tau, b_h^\tau)\big) \Big\|_{(\Lambda_h^k)^{-1}}}_{\text{(ii)}}.
\end{aligned}$$

By Lemma J.2 and a union bound argument, it holds for any $\widetilde{Q} \in \mathcal{Q}_{h+1,\epsilon}^k$ with probability at least $1 - p/2$ that

$$|\text{(i)}| \le 4H^2 \Big(\frac{d}{2}\log(k+1) + \log\frac{2\mathcal{N}_\epsilon}{p}\Big), \qquad \text{(D.38)}$$

where $\mathcal{N}_\epsilon$ is the covering number of $Q_{h+1,\epsilon}$. Meanwhile, by applying Lemma J.4 with $L = H\sqrt{dk}$ and $\lambda = 1$, (D.38) gives that

$$|\text{(i)}| \le C' dH \sqrt{\log(dT/p)}, \qquad \text{(D.39)}$$

with probability at least $1 - p/2$. Here C' is a constant. Meanwhile, by the definition of V_{h+1}^k in Line 10 of Algorithm 1, we have $V_{h+1}^k(x) = \mathbb{E}_{a\sim\widehat{\pi}(\cdot\,|\,x), b\sim\widehat{\nu}(\cdot\,|\,x)}[Q_{h+1}^k(x,a,b)]$, which yields that

$$\begin{aligned}
|V_{h+1}^k(x) - \widetilde{V}(x)| &= \big|\mathbb{E}_{a\sim\widehat{\pi}(\cdot\,|\,x), b\sim\widehat{\nu}(\cdot\,|\,x)}[Q_{h+1}^k(x,a,b) - \widetilde{Q}(x,a,b)]\big| \\
&\le \mathbb{E}_{a\sim\widehat{\pi}(\cdot\,|\,x), b\sim\widehat{\nu}(\cdot\,|\,x)}|Q_{h+1}^k(x,a,b) - \widetilde{Q}(x,a,b)| \le \epsilon
\end{aligned}$$

for any $x \in \mathcal{S}$, which further implies that

$$|\text{(ii)}| \le \epsilon \cdot \sum_{\tau=1}^{k-1} \|\phi(x_h^\tau, a_h^\tau, b_h^\tau)\|_{(\Lambda_h^k)^{-1}} \le \epsilon k, \qquad \text{(D.40)}$$

where the last inequality follows from the fact that $\|\phi(\cdot,\cdot,\cdot)\|_{(\Lambda_h^k)^{-1}} \le \|\phi(\cdot,\cdot,\cdot)\|_2 \le 1$ for any $(k,h) \in [K] \times [H]$. Plugging (D.39) and (D.40) into (D.37), together with the fact that $\epsilon = 1/KH$, we conclude the proof of Lemma D.9. ☐

Lemma D.10 (Bounded Weight of Value Functions)**.** For all $(k,h) \in [K] \times [H]$, the linear coefficient w_h^k defined in (3.3) satisfies $\|w_h^k\| \le H\sqrt{kd}$.

Proof of Lemma D.10. By the definition of w_h^k in (3.3) and triangle inequality, we have

$$\|w_h^k\| = \Big\|(\Lambda_h^k)^{-1}\Big(\sum_{\tau=1}^{k-1}\phi(x_h^\tau, a_h^\tau, b_h^\tau)\cdot V_{h+1}^k(x_{h+1}^\tau)\Big)\Big\|$$
$$\le \sum_{\tau=1}^{k-1}\|(\Lambda_h^k)^{-1}\phi(x_h^\tau, a_h^\tau, b_h^\tau)\cdot V_{h+1}^k(x_{h+1}^\tau)\|. \tag{D.41}$$

Together with the fact that $|V_h^k(\cdot)| \le H$ for any $(k,h)\in[K]\times[H]$, (D.41) gives

$$\|w_h^k\| \le H\cdot\sum_{\tau=1}^{k-1}\|(\Lambda_h^k)^{-1}\phi(x_h^\tau, a_h^\tau, b_h^\tau)\|$$
$$\le H\cdot\sum_{\tau=1}^{k-1}\|(\Lambda_h^k)^{-1/2}\|\cdot\|\phi(x_h^\tau, a_h^\tau, b_h^\tau)\|_{(\Lambda_h^k)^{-1}}$$
$$\le H\cdot\sum_{\tau=1}^{k-1}\|\phi(x_h^\tau, a_h^\tau, b_h^\tau)\|_{(\Lambda_h^k)^{-1}}, \tag{D.42}$$

where the second inequality uses Cauchy-Schwarz inequality and the last inequality follows from the fact that $\Lambda_h^k \succeq I$ for any $(k,h)\in[K]\times[H]$. Then, by Cauchy-Schwarz inequality, we obtain

$$\sum_{\tau=1}^{k-1}\|\phi(x_h^\tau, a_h^\tau, b_h^\tau)\|_{(\Lambda_h^k)^{-1}} \le \sqrt{k}\cdot\Big(\sum_{\tau=1}^{k-1}\phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top(\Lambda_h^k)^{-1}\phi(x_h^\tau, a_h^\tau, b_h^\tau)\Big)^{1/2}$$
$$= \sqrt{k}\cdot\Big(\sum_{\tau=1}^{k-1}\mathrm{Tr}\big(\phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top(\Lambda_h^k)^{-1}\phi(x_h^\tau, a_h^\tau, b_h^\tau)\big)\Big)^{1/2}$$
$$= \sqrt{k}\cdot\Big(\mathrm{Tr}\big((\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi(x_h^\tau, a_h^\tau, b_h^\tau)\phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top\big)\Big)^{1/2}. \tag{D.43}$$

Meanwhile, recall that $\Lambda_h^k = \sum_{\tau=1}^{k-1}\phi(x_h^\tau, a_h^\tau, b_h^\tau)\phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top + I$, we have

$$\mathrm{Tr}\Big((\Lambda_h^k)^{-1}\sum_{\tau=1}^{k-1}\phi(x_h^\tau, a_h^\tau, b_h^\tau)\phi(x_h^\tau, a_h^\tau, b_h^\tau)^\top\Big) \le \mathrm{Tr}(I) = d. \tag{D.44}$$

Plugging (D.43) and (D.44) into (D.42), we conclude the proof of Lemma D.10. □

D.4 Proof of Lemma D.6

Proof of Lemma D.6. Recall the definition of Γ_h^k in Line 7 of Algorithm 1, we have

$$2\sum_{k=1}^K\sum_{h=1}^H\min\{H, \Gamma_h^k(x_h^k, a_h^k, b_h^k)\} = 2\beta\cdot\sum_{k=1}^K\sum_{h=1}^H\min\{H/\beta, \|\phi(x_h^k, a_h^k, b_h^k)\|_{(\Lambda_h^k)^{-1}}\}$$
$$\le 2\beta\cdot\sum_{k=1}^K\sum_{h=1}^H\min\{1, \|\phi(x_h^k, a_h^k, b_h^k)\|_{(\Lambda_h^k)^{-1}}\}. \tag{D.45}$$

Here the last inequality uses the fact that $\beta = CdH\sqrt{\log(2dT/p)}$, where $C > 1$ is a constant. By Cauchy-Schwarz inequality, we further obtain that

$$\sum_{k=1}^K\sum_{h=1}^H\min\{1, \|\phi(x_h^k, a_h^k, b_h^k)\|_{(\Lambda_h^k)^{-1}}\} \le \sum_{h=1}^H\Big(K\cdot\sum_{k=1}^K\min\{1, \|\phi(x_h^k, a_h^k, b_h^k)\|^2_{(\Lambda_h^k)^{-1}}\}\Big)$$
$$\le \sum_{h=1}^H\sqrt{K}\cdot\Big(2\log\Big(\frac{\det(\Lambda_h^{K+1})}{\det(\Lambda_h^1)}\Big)\Big)^{1/2}, \tag{D.46}$$

where the last inequality follows from Lemma J.1. Moreover, Assumption 2.1 gives that

$$\|\phi(x, a, b)\|_2 \leq 1$$

for any $(k, h, x, a, b) \in [K] \times [H] \times \mathcal{S} \times \mathcal{A}$, which further implies that

$$\Lambda_h^{K+1} = \sum_{k=1}^{K} \phi(x_h^k, a_h^k, b_h^k)\phi(x_h^k, a_h^k, b_h^k)^\top + I \preceq (K+1) \cdot I \tag{D.47}$$

for any $h \in [H]$. Combining (D.47) and the fact that $\Lambda_h^1 = I$, we obtain

$$2 \log\Big(\frac{\det(\Lambda_h^{K+1})}{\det(\Lambda_h^1)}\Big) \leq 2d \cdot \log(K+1) \leq 4d \cdot \log(K). \tag{D.48}$$

Combining (D.45), (D.46), (D.47) and (D.48), it holds that

$$2\sum_{k=1}^{K}\sum_{h=1}^{H} \min\{H, \Gamma_h^k(x_h^k, a_h^k, b_h^k)\} \leq 2\beta\sqrt{dHT \cdot \log(K)} \leq \mathcal{O}(\sqrt{d^3H^3T\iota^2}),$$

where $\iota = \log(2dT/p)$. Therefore, we conclude the proof of Lemma D.6. □

D.5 Proof of Lemma D.7

Proof of Lemma D.7. First, we show that $\{\zeta_{k,h}^1, \zeta_{k,h}^2\}_{(k,h)\in[K]\times[H]}$ can be written as a bounded martingale difference with respect to a filtration. Similar to Cai et al. (2020), we construct the following filtration. For any $(k, h) \in [K] \times [H]$, we define σ-algebras $\mathcal{F}_{k,h}^1$ and $\mathcal{F}_{k,h}^2$ as follows:

$$\mathcal{F}_{k,h}^2 = \sigma\big(\{x_i^\tau, a_i^\tau, b_{1,i}^\tau, \cdots, b_{N,i}^\tau\}_{(\tau,i)\in[k-1]\times[h]} \cup \{x_i^k, a_i^k, b_{1,i}^k, \cdots, b_{N,i}^k\}_{i\in[h]}\big),$$
$$\mathcal{F}_{k,h}^2 = \sigma\big(\{x_i^\tau, a_i^\tau, b_{1,i}^\tau, \cdots, b_{N,i}^\tau\}_{(\tau,i)\in[k-1]\times[h]} \cup \{x_i^k, a_i^k, b_{1,i}^k, \cdots, b_{N,i}^k\}_{i\in[h]} \cup \{x_{h+1}^k\}\big), \tag{D.49}$$

where x_{H+1} is a null state for any $k \in [K]$. Here $\sigma(\cdot)$ denotes the σ-algebra generated by a finite set. Moreover, for any $(k, h, m) \in [K] \times [H] \times [2]$, we define the timestep index $t(k, h, m)$ as

$$t(k, h, m) = (k-1) \cdot 2H + (h-1) \cdot 2 + m. \tag{D.50}$$

By the definitions of σ-algebras in (D.49), we have $\mathcal{F}_{k,h}^m \subset \mathcal{F}_{k',h'}^{m'}$ for any $t(k, h, m) \leq t(k', m', h')$, which implies that the σ-algebra sequence $\{\mathcal{F}_{k,h}^m\}_{(k,h,m)\in[K]\times[H]\times[2]}$ is a filtration. Moreover, by the definitions of $\{\zeta_{k,h}^1, \zeta_{k,h}^2\}_{(k,h)\in[K]\times[H]}$ in (D.3), we have

$$\zeta_{k,h}^1 \in \mathcal{F}_{k,h}^1, \quad \zeta_{k,h}^2 \in \mathcal{F}_{k,h}^2, \quad \mathbb{E}[\zeta_{k,h}^1 \,|\, \mathcal{F}_{k,h-1}^2] = 0, \quad \mathbb{E}[\zeta_{k,h}^2 \,|\, \mathcal{F}_{k,h}^1] = 0 \tag{D.51}$$

for any $(k, h) \in [K] \times [H]$. Here we identify $\mathcal{F}_{k,0}^2$ with $\mathcal{F}_{k-1,H}^2$ for any $k \geq 2$ and define $\mathcal{F}_{1,0,2}$ be the empty set. Hence, we can define the martingale

$$\mathcal{M}_{k,h}^m = \Big\{\sum_{k',h',m'} \zeta_{k',h'}^{m'} : t(k', h', m') \leq t(k, h, m)\Big\}. \tag{D.52}$$

Such a martingale is adaptive to the filtration $\{\mathcal{F}_{k,h}^m\}_{(k,h,m)\in[K]\times[H]\times[2]}$. In particular, we have

$$\mathcal{M}_{K,H}^2 = \sum_{k=1}^{K}\sum_{h=1}^{H}(\zeta_{k,h}^1 + \zeta_{k,h}^2). \tag{D.53}$$

Moreover, note the fact that $V_h^k, Q_h^k, V_{l,h}^{\pi^k,\nu^k}, Q_{l,h}^{\pi^k,\nu^k} \in [-H, H]$, we further obtain $|\zeta_{k,h}^m| \leq 2H$, for any $(k, h, m) \in [K] \times [H] \times [2]$. Finally, by applying the Azuma-Hoeffding inequality to $\mathcal{M}_{K,H}^2$ defined in (D.53), we have

$$\sum_{k=1}^{K}\sum_{h=1}^{H}(\zeta_{k,h}^1 + \zeta_{k,h}^2) \leq \sqrt{16H^3K \cdot \log(4/p)}$$

with probability at least $1 - p/2$, which concludes the proof of Lemma D.7. □

E UNKNOWN REWARD SETTING

We focus on the tabular case for simplicity, and the extension to linear case is left as future work. We assume that $S = |\mathcal{S}|$, $A_l = |\mathcal{A}_l|$ and $A_f = |\mathcal{A}_f| = |\mathcal{A}_{f_1} \times \cdots \times \mathcal{A}_{f_N}|$. For simplicity, we use the shorthand $V_\star^{\pi,\nu} = V_{\star,1}^{\pi,\nu}(x_1)$, where $x_1 \in \mathcal{S}$ is the fixed initial state.

E.1 ALGORITHM

We present the pseudocode of Reward-Free Explore algorithm (Jin et al., 2020a) below.

Algorithm 4 Reward-Free Explore

Input: iteration number K_0 and K.
Let policy class $\Phi = \varnothing$.
for $(x, h) \in \mathcal{S} \times [H]$ **do**
  $r_{h'}(x', a', b') \leftarrow \mathbf{1}\left[x' = x \text{ and } h' = h\right]$ for all $(x', a', b', h') \in \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f \times [H]$.
  $\Phi^{(x,h)} \leftarrow \text{EULER}\,(r, K_0)$. [1]
  $\pi_h(\cdot \,|\, x) \leftarrow \text{Uniform}(\mathcal{A}_l)$ and $\nu_h(\cdot \,|\, x) \leftarrow \text{Uniform}(\mathcal{A}_f)$ for all $(\pi, \nu) \in \Phi^{(x,h)}$.
  $\Psi \leftarrow \Psi \cup \Phi^{(x,h)}$.
end for
for $k = 1, \cdots, K$ **do**
  Sample policy $(\pi, \nu) \sim \text{Uniform}(\Psi)$.
  Play the game $\mathcal{M}$ using policy π and ν, and observe the trajectory $\{x_h^k, a_h^k, b_h^k\}_{h\in[H]}$ and rewards $\{r_{\star,h}(x_h^k, a_h^k, b_h^k)\}_{h\in[H]}$.
end for
Calculate the empirical reward as

$$\widehat{r}_{\star,h}(x, a, b) = \frac{\sum_{k=1}^K r_{\star,h}(x, a, b) \cdot \mathbf{1}[x_h^k = x, a_h^k = a, x_{h+1}^k = x']}{\sum_{k=1}^K \mathbf{1}[x_h^k = x, a_h^k = a, x_{h+1}^k = x']}.$$

Lemma E.1. Fix $\varepsilon, p > 0$. If we set $K_0 \geq \Omega(H^7 S^4 A_l/\varepsilon)$ and $K \geq \Omega(H^3 S^2 A_l A_f/\varepsilon^2)$ in Algorithm 4, then we have the empirical rewards $\{\widehat{r}_l, \widehat{r}_{f_1}, \cdots \widehat{r}_{f_N}\}$ and corresponding value functions $(\widehat{V}_l, \widehat{V}_{f_1}, \cdots, \widehat{V}_{f_N})$ satisfying that

$$\sup_{\pi,\nu} |\widehat{V}_\star^{\pi,\nu} - V_\star^{\pi,\nu}| \leq \varepsilon$$

with probability at least $1 - p$. Here $\Omega(\cdot)$ hides some logarithmic factors.

Proof. This lemma is a simple extension of Lemma D.1 in Bai et al. (2021). They focus on the MDP setting and we consider the more complex Markov games. For completeness, we present the detailed proof in §E.2. ☐

Lemma E.1 states that we can obtain estimated reward functions and the associated value functions is an ε-approximation of the true value functions, which further implies that the SNE with respect to the estimated reward functions is a good approximation of the SNE in the original problem. We also remark that if we consider the Markov games with only one follower and aim to find the Stackelberg equilibria, we can provide a more refined analysis. See §F for more details.

E.2 PROOF OF LEMMA E.1

Before our proof, we present a useful lemma.

Lemma E.2. We define the set of δ-significant states as

$$\mathcal{S}_h^\delta = \{s : \max_{\pi,\nu} \mathbb{P}_h^{\pi,\nu}(x) \geq \delta\}, \tag{E.1}$$

[2]Here EULER is a single-agent RL algorithm proposed in Zanette & Brunskill (2019).

where $\mathbb{P}_h^{\pi,\nu}(x)$ is the probability of visiting x at h-th step under policies (π, ν). Then, We have

$$\max_{\pi,\nu} \frac{\mathbb{P}_h^{\pi,\nu}(x)}{\frac{1}{K_0}\sum_{(\pi,\nu)\in\Phi^{(x,h)}} \mathbb{P}_h^{\pi,\nu}(x)} \leq 2$$

for any $s \in \mathcal{S}_h^\delta$. Here $\mathbb{P}_h^\pi(x, a)$ is the probability of visiting (x, a) at h-th step under policy π.

Proof. See the proof of Theorem 3.3 in Jin et al. (2020a) for more details. □

Now, we are ready to proof Lemma E.1.

Proof of Lemma E.1. For any (π, ν), we denote $\mathbb{P}_h^{\pi,\nu}(x, a, b)$ as the probability of visiting (x, a, b) at h-th step under policies (π, ν). Under this notion, by Lemma E.2 and the fact that all policies in $\Phi^{(x,h)}$ are uniform at (x, h), we have

$$\max_{\pi,a,b} \frac{\mathbb{P}_h^{\pi,\nu}(x,a,b)}{\frac{1}{K_0}\sum_{\pi\in\Phi^{(x,h)}} \mathbb{P}_h^{\pi,\nu}(x,a,b)} \leq 2A_l A_f,$$

where $|\mathcal{A}_l| = A_l$ and $A_f = |\mathcal{A}_f| = |\mathcal{A}_{f_1} \times \cdots \times \mathcal{A}_{f_N}|$. Thus, for any δ-significant (x, h), we have

$$\max_{\pi,\nu,a,b} \frac{\mathbb{P}_h^{\pi,\nu}(x,a,b)}{\frac{1}{K_0 SH}\sum_{(\pi,\nu)\in\cup\{\Phi^{(x,h)}\}_{(x,h)}} \mathbb{P}_h^{\pi,\nu}(x,a,b)} \leq 2SA_l A_f H.$$

Then the data obtained from Algorithm 4 is sampled i.i.d. from some distribution ζ_h, such that

$$\max_{\pi,\nu,a,b} \frac{\mathbb{P}_h^{\pi,\nu}(x,a,b)}{\zeta_h(x,a,b)} \leq 2SA_l A_f H. \tag{E.2}$$

for any $s \in \mathcal{S}_h^\delta$. Back to our proof, we have

$$\begin{aligned}
|\widehat{V}_\star^{\pi,\nu} - V_\star^{\pi,\nu}| &= \Big|\sum_{h=1}^H \sum_{x,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)\Big| \\
&= \Big|\sum_{h=1}^H \sum_{x,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)\Big| \\
&\leq \underbrace{\Big|\sum_{h=1}^H \sum_{x\notin\mathcal{S}_h^\delta,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)\Big|}_{\text{(i)}} \\
&\quad + \underbrace{\Big|\sum_{h=1}^H \sum_{x\in\mathcal{S}_h^\delta,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)\Big|}_{\text{(ii)}}.
\end{aligned} \tag{E.3}$$

Clearly,

$$\text{(i)} \leq \sum_{h=1}^H \sum_{x\notin\mathcal{S}_h^\delta,a,b} \mathbb{P}_h^{\pi,\nu}(s,a,b) = \sum_{h=1}^H \sum_{x\notin\mathcal{S}_h^\delta} \mathbb{P}_h^\pi(x) \leq HS\delta \leq \varepsilon/2, \tag{E.4}$$

where the second inequality uses the definition of δ-significant set in (E.1) and the last inequality is implied by the fact that $\delta = \varepsilon/2H^2S$. Meanwhile, we have

$$\begin{aligned}
\text{(ii)} &\leq \sum_{h=1}^H \Big|\sum_{x\in\mathcal{S}_h^\delta,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)\Big| \\
&\leq \sum_{h=1}^H \underbrace{\Big(\sum_{x\in\mathcal{S}_h^\delta,a,b} \mathbb{P}_h^{\pi,\nu}(x,a,b)\cdot\big(\widehat{r}_{\star,h}(x,a,b) - r_{\star,h}(x,a,b)\big)^2\Big)^{1/2}}_{\Delta_h}.
\end{aligned} \tag{E.5}$$

Note that $\mathbb{P}_h^{\pi,\nu}(x,a,b) = \mathbb{P}_h^{\pi,\nu}(x)\cdot\pi_h(a\,|\,x)\cdot\nu_h(b\,|\,x)$, together with Cauchy-Schwarz inequality, we further have

$$\begin{aligned}\Delta_h &\le \max_{\pi':\mathcal{S}\to\mathcal{A}_l,\nu':\mathcal{S}\to\mathcal{A}_f}\Big(\sum_{x\in\mathcal{S}_h^\delta,a,b}\mathbb{P}_h^{\pi}(x)\cdot\big(\widehat{r}_{\star,h}(x,a,b)-r_{\star,h}(x,a,b)\big)^2\mathbf{1}[a=\pi'(s),b=\nu'(s)]\Big)\Big)^{1/2}\\ &\le \max_{\pi':\mathcal{S}\to\mathcal{A}_l,\nu':\mathcal{S}\to\mathcal{A}_f}\Big(\sum_{x\in\mathcal{S}_h^\delta,a,b}\mathbb{P}_h^{\pi}(x)\cdot\big(\widehat{r}_{\star,h}(x,a,b)-r_{\star,h}(x,a,b)\big)^2\mathbf{1}[a=\pi'(s),b=\nu'(s)]\Big)\Big)^{1/2}\\ &\le \max_{\pi':\mathcal{S}\to\mathcal{A}_l,\nu':\mathcal{S}\to\mathcal{A}_f}(2SA_lA_fH)^{1/2}\\ &\quad\times\Big(\sum_{x\in\mathcal{S}_h^\delta,a,b}\zeta_h(x,a,b)\cdot\big(\widehat{r}_{\star,h}(x,a,b)-r_{\star,h}(x,a,b)\big)^2\mathbf{1}[a=\pi'(s),b=\nu'(s)]\Big)\Big)^{1/2},\end{aligned}\tag{E.6}$$

where the last inequality follows from (E.2). Meanwhile, by Hoeffding inequality and a union bound for the reward estimations we have

$$\begin{aligned}&\Big(\sum_{x\in\mathcal{S}_h^\delta,a,b}\zeta_h(x,a,b)\cdot\big(\widehat{r}_{\star,h}(x,a,b)-r_{\star,h}(x,a,b)\big)^2\mathbf{1}[a=\pi'(s),b=\nu'(s)]\Big)\Big)^{1/2}\\ &\qquad\le\Big(\sum_{x\in\mathcal{S}_h^\delta,a,b}\zeta_h(x,a,b)\cdot\widetilde{\mathcal{O}}\big(\frac{1}{N_h(s,a,b)}\big)\mathbf{1}[a=\pi'(s),b=\nu'(s)]\Big)\Big)^{1/2}.\end{aligned}\tag{E.7}$$

Choose $\delta=\varepsilon/2H^2S$. Together with (E.2), we have $\zeta_h(s,a,b)\ge\varepsilon/4H^3S^2A_lA_f$ for any $s\in\mathcal{S}_h^\delta$. Hence, we have $K\ge\Omega(H^3S^2A_lA_f/\varepsilon)\ge\Omega(1/\min_{s,a,b}\zeta_h(s,a,b))$. Applying multiplicative Chernoff bound for the counter $N_h(s,a,b)\sim\mathrm{Bin}(K,\zeta_h(s,a,b))$, we have

$$\begin{aligned}&\Big(\sum_{x\in\mathcal{S}_h^\delta,a,b}\zeta_h(x,a,b)\cdot\widetilde{\mathcal{O}}\big(\frac{1}{N_h(s,a,b)}\big)\mathbf{1}[a=\pi'(s),b=\nu'(s)]\Big)\Big)^{1/2}\\ &\qquad\le\Big(\sum_{x\in\mathcal{S}_h^\delta,a,b}\zeta_h(x,a,b)\cdot\widetilde{\mathcal{O}}\big(\frac{1}{K\zeta_h(s,a,b)}\big)\mathbf{1}[a=\pi'(s),b=\nu'(s)]\Big)\Big)^{1/2}\\ &\qquad=\widetilde{\mathcal{O}}\Big(\sqrt{\frac{S}{K}}\Big).\end{aligned}\tag{E.8}$$

Plugging (E.6), (E.7), and (E.7) into (E.5), we have

$$\text{(ii)}\le\widetilde{\mathcal{O}}\Big(\sqrt{\frac{H^3S^2A_lA_f}{K}}\Big)\le\varepsilon/2,\tag{E.9}$$

where the last inequality follows form our choice that $K\ge\Omega(H^3S^2A_lA_f/\varepsilon^2)$. Combining (E.3), (E.4) and (E.9), we have $|\widehat{V}_\star^{\pi,\nu}-V_\star^{\pi,\nu}|\le\varepsilon$ for any (π,ν), which concludes the proof of Lemma E.1. □

F Learning Stackelberg Equilibria

In this section, we analyze the sample-efficiency of learning Stackelberg equilibria in two-player tabular Markov games without the known reward assumption.

For simplicity, we use the shorthands $f=f_1$ and $V_\star^{\pi,\nu}=V_{\star,1}^{\pi,\nu}(x_1)$, where $x_1\in\mathcal{S}$ is the fixed initial state. Meanwhile, for any $\varepsilon>0$, we define the ε-approximate value of best-case response by

$$\begin{aligned}V_\varepsilon^\pi &= \max_{\nu\in\mathrm{BR}_\varepsilon(\pi)}V_l^{\pi,\nu},\\ \mathrm{BR}_\varepsilon(\pi)&=\{\nu: V_f^{\pi,\nu}\ge\max_{\nu'}V_f^{\pi,\nu'}-\varepsilon\}.\end{aligned}$$

By the above definitions, we can immediately obtain that $\mathrm{BR}(\pi) \subseteq \mathrm{BR}_\varepsilon(\pi)$, which further implies $V_\varepsilon^\pi \geq V_l^{\pi,\nu^*(\pi)}$. Then we can define the gap

$$\begin{aligned} \mathrm{gap}_\varepsilon &= \max_{\pi \in \Pi_\varepsilon} [V_\varepsilon^\pi - V_l^{\pi,\nu^*(\pi)}], \\ \Pi_\varepsilon &= \{\pi : V_\varepsilon^\pi \geq V^{\pi^*,\nu^*} - \varepsilon\}. \end{aligned} \tag{F.1}$$

F.1 Algorithm

As stated before, we first conduct a Reward-Free Explore algorithm (Algorithm 4) to obtain the estimated rewards $(\widehat{r}_l, \widehat{r}_f)$. We also define $(\widehat{V}_l, \widehat{V}_f)$ as the corresponding value functions. Then we use Algorithm 1 to solve the SNE with respect to the *known* reward functions $(\widehat{r}_l, \widehat{r}_f)$. Specifically, we consider the following optimization problem of finding approximation Stackelberg equilibria with respect to the empirical rewards $(\widehat{r}_l, \widehat{r}_f)$.

$$\begin{aligned} &\operatorname*{argmax}_\pi \widehat{V}_{3\varepsilon/4}(\pi) = \operatorname*{argmax}_\pi \widehat{V}_l^{\pi,\nu(\pi)}, \\ &\nu(\pi) = \operatorname*{argmax}_{\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)} \widehat{V}_l^{\pi,\nu}, \\ &\widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi) = \big\{\nu : \widehat{V}_f^{\pi,\nu} \geq \max_\nu \widehat{V}_f^{\pi,\nu} - 3\varepsilon/4\big\}. \end{aligned} \tag{F.2}$$

Since $(\widehat{r}_l, \widehat{r}_f)$ are known to us, we can use Algorithm 1 to obtain the solution $(\widehat{\pi}, \widehat{\nu} = \nu(\widehat{\pi}))$, which is our approximate solution. See Algorithm 5 for more details.

Algorithm 5 Reward-Free Explore then Commit

1: **Input:** Accuracy coefficient $\varepsilon > 0$.
2: Run the Reward-Free Explore algorithm (Algorithm 4) with $K_0 \geq \Omega(H^7 S^4 A_l/\varepsilon)$ and $K \geq \Omega(H^3 S^2 A_l A_f/\varepsilon^2)$, and obtain empirical rewards $(\widehat{r}_l, \widehat{r}_f)$.
3: Use Algorithm 1 as an oracle to solve the problem defined in (F.2) and obtain the solution $(\widehat{\pi}, \widehat{\nu} = \nu(\pi))$.
4: **Output:** $(\widehat{\pi}, \widehat{\nu})$.

F.2 Theoretical Results

The performance of Algorithm 5 is guaranteed by the following theorem.

Theorem F.1. Suppose Algorithm 5 outputs $(\widehat{\pi}, \widehat{\nu})$. Then it holds with probability at least $1-p$ that

$$V_l^{\widehat{\pi},\nu^*(\widehat{\pi})} \geq V_l^{\pi^*,\nu^*} - \mathrm{gap}_\varepsilon - \varepsilon, \qquad V_f^{\widehat{\pi},\widehat{\nu}} \geq V_f^{\widehat{\pi},\nu^*(\widehat{\pi})} - \varepsilon.$$

Proof. Similar analysis also appears in Bai et al. (2021). As stated before, however, their setting is different with ours. For completeness, we provide a detailed proof here. First, we show that

$$\mathrm{BR}_{\varepsilon/2}(\pi) \subseteq \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi) \subseteq \mathrm{BR}_\varepsilon(\pi). \tag{F.3}$$

By choosing a large absolute constant in K, together with Lemma E.1, it holds for any $\star \in \{l, f\}$ that

$$\sup_{\pi,\nu} |\widehat{V}_\star^{\pi,\nu} - V_\star^{\pi,\nu}| \leq \varepsilon/8. \tag{F.4}$$

Meanwhile, for the empirical rewards $(\widehat{r}_l, \widehat{r}_f)$, we define the best response of leader's policy π as $\widehat{\nu^*(\pi)}$. Under this notation, for any $\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)$, we have

$$\begin{aligned} &V_f^{\pi,\nu^*(\pi)} - V_f^{\pi,\nu} \\ &\quad = \underbrace{(V_f^{\pi,\nu^*(\pi)} - \widehat{V}_f^{\pi,\nu^*(\pi)})}_{\text{(i)}} + \underbrace{(\widehat{V}_f^{\pi,\nu^*(\pi)} - \widehat{V}_f^{\pi,\widehat{\nu^*(\pi)}})}_{\text{(ii)}} + \underbrace{(\widehat{V}_f^{\pi,\widehat{\nu^*(\pi)}} - \widehat{V}_f^{\pi,\nu})}_{\text{(iii)}} + \underbrace{(\widehat{V}_{f_i}^{\pi,\nu} - V_f^{\pi,\nu})}_{\text{(iv)}} \\ &\quad \leq \varepsilon/8 + 0 + 3\varepsilon/4 + \varepsilon/8 \leq \varepsilon. \end{aligned} \tag{F.5}$$

where (i) $\leq \varepsilon/8$ and (iv) $\leq \varepsilon/8$ is implied by the uniform convergence in (F.4), (ii) ≤ 0 uses the definition of $\widehat{\nu^*(\pi)}$, and (iii) ≤ 0 follows from the fact that $\nu \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)$.

Similarly, for any $\nu \in \mathrm{BR}_{\varepsilon/2}(\pi)$, we can show that

$$\widehat{V}_f^{\pi,\widehat{\nu^*(\pi)}} - \widehat{V}_f^{\pi,\nu}$$
$$= (\widehat{V}_f^{\pi,\widehat{\nu^*(\pi)}} - V_f^{\pi,\widehat{\nu^*(\pi)}}) + (V_f^{\pi,\widehat{\nu^*(\pi)}} - V_f^{\pi,\nu^*(\pi)}) + (V_f^{\pi,\nu^*(\pi)} - V_f^{\pi,\nu}) + (V_f^{\pi,\nu} - \widehat{V}_f^{\pi,\nu})$$
$$\leq \varepsilon/8 + 0 + \varepsilon/2 + \varepsilon/8 = 3\varepsilon/4. \tag{F.6}$$

Combining (F.5) and (F.6), we obtain $\mathrm{BR}_{\varepsilon/2}(\pi) \subseteq \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi) \subseteq \mathrm{BR}_\varepsilon(\pi)$ as desired.

Back to our proof, by the fact that $\widehat{\pi}$ maximizes $\widehat{V}^\pi_{3\varepsilon/4} = \max_{\nu\in\widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)} \widehat{V}_l(\pi,\nu)$, we have

$$\max_{\nu\in\widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi})} \widehat{V}_l^{\widehat{\pi},\nu} = \widehat{V}^{\widehat{\pi}}_{3\varepsilon/4} \geq \widehat{V}^{\pi}_{3\varepsilon/4} = \max_{\nu\in\widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)} V_l^{\pi,\nu} \geq \max_{\nu\in\mathrm{BR}_{\varepsilon/2}(\pi)} \widehat{V}_l^{\pi,\nu}, \tag{F.7}$$

for any π. Here the last inequality uses the fact $\mathrm{BR}_{\varepsilon/2}(\pi) \subseteq \widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi)$ in (F.3). Together with the uniform convergence in (F.4), (F.7) yields

$$\max_{\nu\in\widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi})} V_l^{\widehat{\pi},\nu} \geq \min_{\nu\in\mathrm{BR}_{\varepsilon/2}(\pi)} V_l^{\pi,\nu} - \varepsilon/8 \geq V^\pi_{\varepsilon/2} - \varepsilon. \tag{F.8}$$

Meanwhile, by the fact $\widehat{\mathrm{BR}}_{3\varepsilon/4}(\pi) \subseteq \mathrm{BR}_\varepsilon(\pi)$ in (F.4), we have

$$V^{\widehat{\pi}}_\varepsilon = \min_{\nu\in\mathrm{BR}_\varepsilon(\widehat{\pi})} V_l^{\widehat{\pi},\nu} \geq \min_{\nu\in\widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi})} V_l^{\widehat{\pi},\nu}. \tag{F.9}$$

Combining (F.8) and (F.9), we have

$$V^{\widehat{\pi}}_\varepsilon \geq \max_\pi V^\pi_{\varepsilon/2} - \varepsilon \geq \max_\pi V_l^{\pi,\nu^*(\pi)} - \varepsilon, \tag{F.10}$$

which implies that $\widehat{\pi} \in \Pi_\varepsilon$. Furthermore, (F.10) is equivalent to

$$V_l^{\widehat{\pi},\nu^*(\widehat{\pi})} \geq V_l^{\pi^*,\nu^*} - [V^{\widehat{\pi}}_\varepsilon - V_l^{\widehat{\pi},\nu^*(\widehat{\pi})}] - \varepsilon \geq V_l^{\pi^*,\nu^*} - \mathrm{gap}_\varepsilon - \varepsilon,$$

where the equality uses the definition of gap_ε in (F.1). as desired. Meanwhile, by the facts that $\widehat{\nu} \in \widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi})$ and $\widehat{\mathrm{BR}}_{3\varepsilon/4}(\widehat{\pi}) \subseteq \mathrm{BR}_\varepsilon(\widehat{\pi})$, we have

$$V_f^{\widehat{\pi},\widehat{\nu}} \geq V_f^{\widehat{\pi},\nu^*(\widehat{\pi})} - \varepsilon.$$

Therefore, we conclude the proof of Theorem F.1. □

G PROOF OF THEOREM 4.2

To facilitate our analysis, we first define the prediction error

$$\delta_h = r_{l,h} + \widehat{Q}_h - \mathbb{P}_h \widehat{V}_h \tag{G.1}$$

for any $h \in [H]$. Then we show the proof of Theorem 4.2.

Proof of Theorem 4.2. Similar to Lemma D.1, we have the following lemma.

Lemma G.1. It holds that $\widehat{\nu} \in \mathrm{BR}(\widehat{\pi})$. Here $\mathrm{BR}(\cdot)$ is defined in (2.10).

Proof. This is implied by the definitions of $(\widehat{\pi}, \widehat{\nu})$ and the assumption that the followers are myopic. □

Recall that the definition of optimality gap defined in (4.1) takes the following form

$$\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}, x) = V_{l,1}^{\pi^*,\nu^*}(x) - V_{l,1}^{\widehat{\pi},\widehat{\nu}}(x). \tag{G.2}$$

In that follows, we decompose it by the following lemma.

Lemma G.2. For the $\widehat{V}_1$ defined in Line 10 of Algorithm 3 and any (π, ν), it holds that

$$V_{l,1}^{\pi,\nu}(x) - \widehat{V}_1(x) = \mathbb{E}_{\pi,\nu}\Big[\sum_{h=1}^{H} \langle \widehat{Q}_h(x_h,\cdot,\cdot), \pi_h(\cdot\,|\,x_h) \times \nu_h(\cdot\,|\,x_h) - \widehat{\pi}_h(\cdot\,|\,x_h) \times \widehat{\nu}_h(\cdot\,|\,x_h)\rangle\Big] + \mathbb{E}_{\pi,\nu}\Big[\sum_{h=1}^{H} \delta_h(x_h,a_h,b_h)\Big].$$

Proof. This proof is the same as the proof of (D.14), and we omit it to avoid repetition. □

Applying Lemma G.2 with $(\pi, \nu) = (\pi^*, \nu^*)$, we have

$$V_{l,1}^{\pi^*,\nu^*}(x) - \widehat{V}_1(x) = \mathbb{E}_{\pi^*,\nu^*}\Big[\sum_{h=1}^{H} \langle \widehat{Q}_h(x_h,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h) \times \nu_h^*(\cdot\,|\,x_h) - \widehat{\pi}_h(\cdot\,|\,x_h) \times \widehat{\nu}_h(\cdot\,|\,x_h)\rangle\Big] + \mathbb{E}_{\pi^*,\nu^*}\Big[\sum_{h=1}^{H} \delta_h(x_h,a_h,b_h)\Big]. \tag{G.3}$$

Similarly, applying Lemma G.2 with $(\pi, \nu) = (\widehat{\pi}, \widehat{\nu})$ gives that

$$\widehat{V}_1(x) - V_{l,1}^{\widehat{\pi},\widehat{\nu}}(x) = -\mathbb{E}_{\widehat{\pi},\widehat{\nu}}\Big[\sum_{h=1}^{H} \delta_h(x_h,a_h,b_h)\Big]. \tag{G.4}$$

Combining (G.3) and (G.4), we obtain

$$V_{l,1}^{\pi^*,\nu^*}(x) - V_{l,1}^{\widehat{\pi},\widehat{\nu}}(x) = \mathbb{E}_{\pi^*,\nu^*}\Big[\sum_{h=1}^{H} \langle \widehat{Q}_h(x_h,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h) \times \nu_h^*(\cdot\,|\,x_h) - \widehat{\pi}_h(\cdot\,|\,x_h) \times \widehat{\nu}_h(\cdot\,|\,x_h)\rangle\Big] + \mathbb{E}_{\pi^*,\nu^*}\Big[\sum_{h=1}^{H} \delta_h(x_h,a_h,b_h)\Big] - \mathbb{E}_{\widehat{\pi},\widehat{\nu}}\Big[\sum_{h=1}^{H} \delta_h(x_h,a_h,b_h)\Big]. \tag{G.5}$$

As stated in §D, these two terms characterize the optimization error and the statistical error, respectively. Similar to Lemmas D.4 and D.5, we introduce the following two lemmas to analyze these two errors.

Lemma G.3. It holds that

$$\mathbb{E}_{\pi^*,\nu^*}\Big[\sum_{h=1}^{H} \langle \widehat{Q}_h(x_h,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h) \times \nu_h^*(\cdot\,|\,x_h) - \widehat{\pi}_h(\cdot\,|\,x_h) \times \widehat{\nu}_h(\cdot\,|\,x_h)\rangle\Big] \le \epsilon H.$$

Proof. This proof is similar to the proof of Lemma D.4, and we omit it to avoid repetition. □

Lemma G.4. It holds with probability at least $1 - p/2$ that

$$0 \le \delta_h(x,a,b) \le 2\Gamma_h(x,a,b)$$

for any $h \in [H]$ and $(x,a,b) \in \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$.

Proof. See §G.1 for a detailed proof. □

Combining (G.5) and Lemmas G.3 and G.4, we further obtain that

$$\begin{aligned} V_{l,1}^{\pi^*,\nu^*}(x) - V_{l,1}^{\widehat{\pi},\widehat{\nu}}(x) &\le \epsilon H + 2\mathbb{E}_{\pi^*,\nu^*,x}\Big[\sum_{h=1}^{H} \Gamma_h(x_h,a_h,b_h)\Big] \\ &\le 3\beta' \sum_{h=1}^{H} \mathbb{E}_{\pi^*,\nu^*,x}\Big[\big(\phi(s_h,a_h,b_h)^\top (\Lambda_h)^{-1} \phi(s_h,a_h,b_h)\big)^{1/2}\Big], \end{aligned} \tag{G.6}$$

where the last inequality is obtained by the definition of Γ_h in Line 6 of Algorithm 3 and the fact that $\epsilon = d/KH$. Therefore, we conclude the proof of Theorem 4.2. □

G.1 Proof of Lemma G.4

Proof of Lemma G.4. Similar to (D.31), it holds with probability at least $1 - p/2$ that

$$|\phi(x,a,b)^\top w_h - (\mathbb{P}_h \widehat{V}_{h+1})(x,a,b)| \le \Gamma_h(x,a,b) \tag{G.7}$$

for any $h \in [H]$. The only exception is that we use Lemma J.3 instead of the classical concentration lemma (Lemma J.2) for the self-normalized process. Here we omit the detailed proof to avoid repetition.

By (G.7) and the fact that $\widehat{V}_{h+1}(\cdot) \le H - h$, we obtain

$$\phi(x,a,b)^\top w_h - \Gamma_h(x,a,b) \le (\mathbb{P}_h \widehat{V}_{h+1})(x,a,b) \le H - h. \tag{G.8}$$

Thus, we have $\widehat{Q}_h \ge \phi^\top w_h - \Gamma_h$, which further implies that

$$\begin{aligned}\delta_h(x,a,b) &= r_{l,h}(x,a,b) + \mathbb{P}_h \widehat{V}_{h+1}(x,a,b) - \widehat{Q}_h(x,a,b)\\ &\le \mathbb{P}_h \widehat{V}_{h+1}(x,a,b) - \phi(x,a,b)^\top w_h + \Gamma_h(x,a,b)\\ &\le 2\Gamma_h(x,a,b),\end{aligned} \tag{G.9}$$

where the last inequality uses (G.7). Meanwhile, it holds that

$$\begin{aligned}\delta_h(x,a,b) &= r_{l,h}(x,a,b) + \mathbb{P}_h \widehat{V}_{h+1}(x,a,b) - \widehat{Q}_h(x,a,b)\\ &\ge \mathbb{P}_h \widehat{V}_{h+1}(x,a,b) - \max\{\phi(x,a,b)^\top w_h - \Gamma_h^k(x,a,b), -(H-h)\}\\ &= \min\{\mathbb{P}_h V_{h+1}^k(x,a,b) - \phi(x,a,b)^\top w_h^k + \Gamma_h^k(x,a,b), \mathbb{P}_h V_{h+1}^k(x,a,b) + (H-h)\}\\ &\ge 0,\end{aligned} \tag{G.10}$$

where the last inequality follows from (G.7). Combining (G.9) and (G.10), we conclude the proof of Lemma G.4. ☐

H Proof of Corollary C.1

Proof of Corollary C.1. The proof is an extension of Corollary 4.5 in Jin et al. (2020c). For notational simplicity, we define

$$\Sigma_h(x) = \mathbb{E}_{\pi^*,\nu^*,x}[\phi(s_h,a_h,b_h)\phi(s_h,a_h,b_h)^\top]$$

for all $x \in \mathcal{S}$ and $h \in [H]$. With this notation and Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}_{\pi^*,\nu^*,x}\Big[\sqrt{\phi(s_h,a_h,b_h)^\top \Lambda_h^{-1}\phi(s_h,a_h.b_h)}\Big] &= \mathbb{E}_{\pi^*,\nu^*,x}\Big[\sqrt{\mathrm{Tr}\big(\phi(s_h,a_h,b_h)^\top \Lambda_h^{-1}\phi(s_h,a_h,b_h)\big)}\Big]\\ &= \mathbb{E}_{\pi^*,\nu^*,x}\Big[\sqrt{\mathrm{Tr}\big(\phi(s_h,a_h.b_h)\phi(s_h,a_h,b_h)^\top \Lambda_h^{-1}\big)}\Big]\\ &= \mathbb{E}_{\pi^*,\nu^*}\Big[\sqrt{\mathrm{Tr}\big(\Sigma_h(x)\Lambda_h^{-1}\big)}\Big].\end{aligned} \tag{H.1}$$

Plugging (H.1) into Theorem 4.2, together with the assumption that $\Lambda_h \succeq I + c \cdot K \cdot \mathbb{E}_{\pi^*,\nu^*,x}[\phi(s_h,a_h,b_h)\phi(s_h,a_h,b_h)^\top]$ with probability at least $1-p/2$ and a union bound argument, we further with probability at least $1-p$ have

$$\begin{aligned}\mathrm{SubOpt}(\widehat{\pi},\widehat{\nu},x) &\le 3\beta' \sum_{h=1}^H \mathbb{E}_{\pi^*,\nu^*}\bigg[\sqrt{\mathrm{Tr}\Big(\Sigma_h(x)\big(I + c\cdot K\cdot \Sigma_h(x)\big)^{-1}\Big)}\bigg]\\ &= 3\beta' \sum_{h=1}^H \sqrt{\sum_{j=1}^d \frac{\lambda_{h,j}(x)}{1 + cK\lambda_{h,j}(x)}}\end{aligned} \tag{H.2}$$

for all $x \in \mathcal{S}$. Here $\{\lambda_{h,j}(x)\}_{j=1}^d$ are the eigenvalues of $\Sigma_h(x)$. Meanwhile, by Jensen's inequality, we obtain

$$\|\Sigma_h(x)\|_{\mathrm{op}} \le \mathbb{E}_{\pi^*,\nu^*,x}[\|\phi(s_h,a_h,b_h)\phi(s_h,a_h,b_h)^\top\|_{\mathrm{op}}] \le 1, \tag{H.3}$$

where the last inequality follows from the fact that $\|\phi(\cdot,\cdot,\cdot)\|_2 \le 1$. Combining (H.2) and (H.3), it holds with probability at least $1-p$ that

$$\begin{aligned}\text{SubOpt}(\widehat{\pi}, \widehat{\nu}, x) &\le 3\beta' \sum_{h=1}^{H} \sqrt{\sum_{j=1}^{d} \frac{1}{1+cK}} \\ &\le \bar{C} \cdot d^{3/2} H^2 \sqrt{\log(4dHK/p)/K},\end{aligned}$$

where $\bar{C} = 3C/\sqrt{c}$, which concludes the proof of Corollary C.1. □

I Results with Pessimistic Tie-breaking

I.1 Stackelberg-Nash Equilibria in Pessimistic Tie-breaking Setup

For any leader policy π, we can define

$$\nu^{\dagger}(\pi) = \{\nu \in \text{BR}(\pi) \,|\, V_{l,h}^{\pi,\nu}(x) \le V_{l,h}^{\pi,\nu'}(x), \forall x \in \mathcal{S}, h \in [H], \nu' \in \text{BR}(\pi)\}, \tag{I.1}$$

where $\text{BR}(\pi)$ is the best-response set defined in (2.10). That is, $\nu^{\dagger}(\pi)$ is the worst-case response in the set $\text{BR}(\pi)$. Then we define the Stackelberg-Nash equilibria by

$$\text{SNE}_l^{\dagger} = \{\pi \,|\, V_{l,h}^{\pi,\nu^*(\pi)}(x) \ge V_{l,h}^{\pi',\nu^{\dagger}(\pi')}(x), \forall x \in \mathcal{S}, h \in [H], \pi'\}. \tag{I.2}$$

We point out that finding the Stackelberg-Nash equilibria in the pessimistic tie-breaking setting is harder. Specifically, compared with optimistic tie-breaking setting (cf. (2.8)), we need to solve a more complicated constrained max-min optimization problem:

$$\max_{\pi} \min_{\nu} V_{l,1}^{\pi,\nu}(x) \qquad \text{s.t.}\, \nu \in \text{BR}(\pi).$$

Under this more challenging setting, we focus on the leader-controller linear Markov games setting (Assumption 2.3). Similar to Theorems 3.3 and 4.2, we can have the following two theorems in the online and offline settings.

I.2 Main Results for the Online Setting

Theorem I.1. Under Assumptions 2.1, 2.3, and 3.1, there exists an absolute constant $C > 0$ such that, for any fixed $p \in (0,1)$, by setting $\beta = C \cdot dH\sqrt{\iota}$ with $\iota = \log(2dT/p)$ in Line 7 of Algorithm 6 and $\epsilon = \frac{1}{KH}$ in Algorithm 7, then have $\nu^k = \nu^{\dagger}(\pi^k)$ for any $k \in [K]$. Meanwhile, with probability at least $1-p$, the regret incurred by Algorithm 6 satisfies that

$$\text{Regret}(K) = \sum_{k=1}^{K} V_{l,1}^{\pi^*,\nu^*}(x_1^k) - V_{l,1}^{\pi^k,\nu^k}(x_1^k) \le \mathcal{O}(\sqrt{d^3 H^3 T \iota^2}).$$

Proof. See §I.4 for a detailed proof. □

Misspecification. When the transitions do not ideally satisfy the leader-controller assumption, we can potentially consider cases that transitions satisfy, for instance, $|\mathcal{P}_h(\cdot\,|\,x,a,b) - \mathcal{P}_h(\cdot\,|\,x,a)\|_\infty \le \varrho$ for any $(h,x,a,b) \in [H] \times \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f$, Here ϱ is the misspecification error. We can still follow the above method to tackle the misspecified cases. However, because of the misspecification error cumulated during T steps, an extra term $\mathcal{O}(\varrho T)$ will appear in the final result. In particular, When ϱ is small, that is the Markov games have approximately leader-controller transitions, the extra term $\mathcal{O}(\varrho T)$ should be small, which further indicates that we can find SNEs efficiently in some misspecified general-sum Markov games.

Algorithm 6 Optimistic Value Iteration to Find Stackelberg-Nash Equilibria (pessimistic tie-breaking version)

Initialize $V_{l,H+1}(\cdot) = V_{f,H+1}(\cdot) = 0$.
for $k = 1, 2, \cdots, K$ **do**
  Receive initial state x_1^k.
  for step $h = H, H-1, \cdots, 1$ **do**
    $\Lambda_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + I$.
    $w_h^k \leftarrow (\Lambda_h^k)^{-1} \sum_{\tau=1}^{k-1} \phi(x_h^\tau, a_h^\tau) \cdot V_{h+1}^k(x_{h+1}^\tau)$.
    $\Gamma_h^k(\cdot,\cdot,\cdot) \leftarrow \beta \cdot (\phi(\cdot,\cdot)^\top (\Lambda_h^k)^{-1} \phi(\cdot,\cdot))^{1/2}$.
    $Q_h^k(\cdot,\cdot,\cdot) \leftarrow r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{\phi(\cdot,\cdot)^\top w_h^k + \Gamma_h^k(\cdot,\cdot)\}$.
    $(\pi_h^k(\cdot \,|\, x), \{\nu_{f_i,h}^k(\cdot \,|\, x)\}_{i\in[N]}) \leftarrow \epsilon$-SNE$(Q_h^k(x,\cdot,\cdot), \{r_{f_i,h}(x,\cdot,\cdot)\}_{i\in[N]})$, $\forall x$. (Alg. 7)
    $V_h^k(x) \leftarrow \mathbb{E}_{a\sim\pi_h^k(\cdot \,|\, x), b_1\sim\nu_{f_1,h}^k(\cdot \,|\, x), \cdots, b_N\sim\nu_{f_N,h}^k(\cdot \,|\, x)} Q_h^k(x, a, b_1, \cdots, b_N)$, $\forall x$.
  end for
  for $h = 1, 2, \cdot, H$ **do**
    Sample $a_h^k \sim \pi_h^k(\cdot \,|\, x_h^k), b_{1,h}^k \sim \nu_{f_1,h}^k(\cdot \,|\, x_h^k), \cdots, b_{N,h}^k \sim \nu_{f_N,h}^k(\cdot \,|\, x_h^k)$.
    Leader takes action a_h^k; Followers take actions $b_h^k = \{b_{i,h}^k\}_{i\in[N]}$.
    Observe next state x_{h+1}^k.
  end for
end for

Algorithm 7 ϵ-SNE (pessimistic tie-breaking version)

Input: Q_h^k, x, and parameter ϵ.
Select $\widetilde{Q}$ from $\mathcal{Q}_{h,\epsilon}^k$ satisfying $\|\widetilde{Q} - Q_h^k\|_\infty \le \epsilon$.
For the input state x, let $(\pi_h^k(\cdot \,|\, x), \{\nu_{f_i,h}^k(\cdot \,|\, x)\}_{i\in[N]})$ be the Stackelberg-Nash equilibrium for the matrix game with payoff matrices $(\widetilde{Q}(x,\cdot,\cdot), \{r_{f_i,h}(x,\cdot,\cdot)\}_{i\in[N]})$ in the pessimistic tie-breaking setting.
Output: $(\pi_h^k(\cdot \,|\, x), \{\nu_{f_i,h}^k(\cdot \,|\, x)\}_{i\in[N]})$.

I.3 Main Results for the Offline Setting

Theorem I.2. Under Assumptions 2.1, 2.3, 3.1, and 4.1, there exists an absolute constant $C > 0$ such that, for any fixed $p \in (0,1)$, by setting $\beta' = C \cdot dH\sqrt{\log(2dHK/p)}$ in Line 6 of Algorithm 8 and $\epsilon = \frac{d}{KH}$ in Algorithm 7, then we have $\widehat{\nu} = \nu^\dagger(\widehat{\pi})$. Meanwhile, with probability at least $1-p$, we have

$$\mathrm{SubOpt}(\widehat{\pi}, \widehat{\nu}, x) = V_{l,1}^{\pi^*,\nu^*}(x) - V_{l,1}^{\widehat{\pi},\widehat{\nu}}(x) \le 3\beta' \sum_{h=1}^H \mathbb{E}_{\pi^*,x}\big[\big(\phi(s_h,a_h)^\top (\Lambda_h)^{-1}\phi(s_h,a_h)\big)^{1/2}\big],$$

where $\mathbb{E}_{\pi^*,x}$ is taken with respect to the trajectory incurred by π^* in the underlying leader-controller Markov game when initializing the progress at x. Here Λ_h is defined in Line 4 of Algorithm 8.

Proof. Combining the proofs of Theorems 4.2 and I.1, we can conclude the proof of Theorem I.2. To avoid repetition, we omit the detailed proof here. □

Optimality of the Bound: Assuming the dummy followers, that is, the actions taken by the followers won't affect the reward functions and transition kernels, the Markov games reduces to the linear MDP (Jin et al., 2020b). Together with the information-theoretic lower bound $\Omega(\sum_{h=1}^H \mathbb{E}_{\pi^*,x}[(\phi(s_h,a_h)^\top(\Lambda_h)^{-1}\phi(s_h,a_h))^{1/2}])$ established in Jin et al. (2020c) for linear MDPs, we immediately obtain the same lower bound for our setting. In particular, our upper bound established in Theorem I.2 matches this lower bound up to β' and absolute constants and thus implies that our algorithm is nearly minimax optimal.

Algorithm 8 Pessimistic Value Iteration to Find Stackelberg-Nash Equilibria (pessimistic tie-breaking version)

Input: $\mathcal{D} = \{x_h^\tau, a_h^\tau, b_h^\tau = \{b_{i,h}^\tau\}_{i\in[N]}\}_{\tau,h=1}^{K,H}$ and reward functions $\{r_l, r_f = \{r_{f_i}\}_{i\in[N]}\}$.
Initialize $\widehat{V}_{H+1}(\cdot) = 0$.
for step $h = H, H-1, \cdots, 1$ **do**
  $\Lambda_h \leftarrow \sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + I$.
  $w_h \leftarrow (\Lambda_h)^{-1}\sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\cdot \widehat{V}_{h+1}(x_{h+1}^\tau)$.
  $\Gamma_h(\cdot,\cdot) \leftarrow \beta' \cdot (\phi(\cdot,\cdot)^\top(\Lambda_h)^{-1}\phi(\cdot,\cdot))^{1/2}$.
  $\widehat{Q}_h(\cdot,\cdot,\cdot) \leftarrow r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{\phi(\cdot,\cdot)^\top w_h - \Gamma_h(\cdot,\cdot)\}$.
  $(\widehat{\pi}_h(\cdot\,|\,x), \{\widehat{\nu}_{f_i,h}(\cdot\,|\,x)\}_{i\in[N]}) \leftarrow \epsilon\text{-SNE}(\widehat{Q}_h(x,\cdot,\cdot), \{r_{f_i,h}(x,\cdot,\cdot)\}_{i\in[N]}),\ \forall x$. (Alg. 7)
  $\widehat{V}_h(x) \leftarrow \mathbb{E}_{a\sim\widehat{\pi}_h(\cdot\,|\,x), b_1\sim\widehat{\nu}_{f_1,h}(\cdot\,|\,x),\cdots,b_N\sim\widehat{\nu}_{f_N,h}(\cdot\,|\,x)}\widehat{Q}_h(x,a,b_1,\cdots,b_N),\ \forall x$.
end for
Output: $(\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H, \widehat{\nu} = \{\widehat{\nu}_{f_i} = \{\nu_{f_i,h}\}_{h=1}^H\}_{i=1}^N)$.

I.4 PROOF OF THEOREM I.1

Proof of Theorem I.1. For leader-controller Markov games, we have a stronger version of Lemma D.1.

Lemma I.3. For any $k \in [K]$, we have $\nu^k = \nu^\dagger(\pi^k)$. Here $\nu^\dagger(\cdot)$ is defined in (I.1).

Proof. Fix $k \in [K]$, by the definition of the best response in (2.5), we have

$$\begin{aligned}\mathrm{BR}(\pi^k) &= \{\nu = \{\nu_{f_i}\}_{i\in[N]}\,|\,\nu \text{ is the NE of the followers given the leader policy } \pi^k\}\\ &= \{\nu = \{\nu_{f_i}\}_{i\in[N]}\,|\,\nu \text{ is the NE of } \{V_{f_i,h}^{\pi^k,\nu}(x)\}_{i\in[N]},\ \forall h\in[H] \text{ and } x\in\mathcal{S}\}\\ &= \{\nu = \{\nu_{f_i}\}_{i\in[N]}\,|\,\nu \text{ is the NE of } \{r_{f_i,h}^{\pi^k,\nu}(x)\}_{i\in[N]},\ \forall h\in[H] \text{ and } x\in\mathcal{S}\},\end{aligned} \tag{I.3}$$

where $r_{f_i,h}^{\pi^k,\nu}(x) = \langle r_{f_i,h}(x,\cdot,\cdot,\cdots,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_{f_1,h}(\cdot\,|\,x)\times\cdots\times\nu_{f_N,h}(\cdot\,|\,x)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$. Here the last inequality uses Bellman equality (2.2) and the leader-controller assumption. Moreover, by the definition of $\nu^\dagger(\pi^k)$ defined in (2.6), we have that

$$\nu_h^\dagger(\pi^k) = \{\nu_{f_i,h}^\dagger(\pi^k)\}_{i\in[N]} \in \operatorname*{argmin}_{\nu\in\mathrm{BR}(\pi^k)} V_{l,h}^{\pi^k,\nu}(x) = \operatorname*{argmin}_{\nu\in\mathrm{BR}(\pi^k)} r_{l,h}^{\pi^k,\nu}(x), \tag{I.4}$$

where $r_{l,h}^{\pi^k,\nu}(x) = \langle r_{l,h}(x,\cdot,\cdot,\cdots,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_{f_1,h}(\cdot\,|\,x)\times\cdots\times\nu_{f_N,h}(\cdot\,|\,x)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$. Here the last equality uses the single-controller assumption.

Recall that, in the subroutine ϵ-SNE (Algorithm 2), we pick the function $\widetilde{Q} \in \mathcal{Q}_{h,\epsilon}^k$ such that $\|Q_h^k - \widetilde{Q}\|_\infty \le \epsilon$ and solve the matrix game defined in (3.6). Here $\mathcal{Q}_{h,\epsilon}^k$ is the class of functions $Q : \mathcal{S}\times\mathcal{A}_l\times\mathcal{A}_f \to \mathbb{R}$ that takes form

$$Q(\cdot,\cdot,\cdot) = r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{\phi(\cdot,\cdot)^\top w + \beta\cdot\big(\phi(\cdot,\cdot)^\top\Lambda^{-1}\phi(\cdot,\cdot)\big)^{1/2}\}, \tag{I.5}$$

where $\|w\|_2 \le H\sqrt{dk}$ and $\lambda_{\min}(\Lambda) \ge 1$. Thus, given the leader policy π^k, the best response of the followers for the matrix game defined in (3.6) takes the form

$$\begin{aligned}\mathrm{BR}'(\pi^k) &= \{\nu\,|\,\nu \text{ is the NE of } \{\langle r_{f_i,h}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle\}_{i\in[N]}, \forall h\in[H] \text{ and } x\in\mathcal{S}\}\\ &= \mathrm{BR}(\pi^k)\end{aligned} \tag{I.6}$$

where $\langle r_{f_i,h}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle$ is the shorthand of $\langle r_{f_i,h}(x,\cdot,\cdot,\cdots,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_{f_1,h}(\cdot\,|\,x)\times\cdots\times\nu_{f_N,h}(\cdot\,|\,x)\rangle_{\mathcal{A}_l\times\mathcal{A}_f}$. Here the last equality uses (I.3). Similarly, by the definition of $\mathcal{Q}_{h,\epsilon}^k$ in (I.5), we can obtain that

$$\operatorname*{argmin}_{\nu_h}\langle\widetilde{Q}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle = \operatorname*{argmin}_{\nu_h}\langle r_{l,h}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x)\times\nu_h(\cdot\,|\,x)\rangle, \tag{I.7}$$

where $\langle r_{l,h}(x,\cdot,\cdot), \pi_h^k(\cdot\,|\,x) \times \nu_h(\cdot\,|\,x)\rangle$ is the abbreviation of $\langle r_{f_i,h}(x,\cdot,\cdot,\cdots,\cdot), \pi_h^k(\cdot\,|\,x) \times \nu_{f_1,h}(\cdot\,|\,x) \times \cdots \times \nu_{f_N,h}(\cdot\,|\,x)\rangle_{\mathcal{A}_l \times \mathcal{A}_f}$ Together with (I.4) and (I.6), we have that, for the matrix game with payoff matrices $(\widetilde{Q}(x_h^k,\cdot,\cdot), \{r_{f_i,h}^k(x_h^k,\cdot,\cdot)\}_{i\in[N]})$, the policy $\nu_h^k(\cdot\,|\,x_h^k) = \{\nu_{f_i,h}^k(\cdot\,|\,x_h^k)\}_{i\in[N]}$ is also the best response of $\pi_h^k(\cdot\,|\,x_h^k)$ and breaks ties against favor of the leader. Therefore, we have $\nu^k = \nu^\dagger(\pi^k)$ for any $k \in [K]$, which concludes the proof of Lemma I.3. □

Then we only need to bound the quantity $\sum_{k=1}^K \sum_{k=1}^K V_{l,1}^{\pi^*,\nu^*}(x_1^k) - V_{l,1}^{\pi^k,\nu^k}(x_1^k)$. By Lemma D.2, we have

$$\begin{aligned}\text{Regret}(K) &= \underbrace{\sum_{k=1}^K \sum_{h=1}^H \mathbb{E}_{\pi^*,\nu^*}[\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle]}_{(l.1)\text{: Computational Error}} \\ &\quad + \underbrace{\sum_{k=1}^K \sum_{h=1}^H \big(\mathbb{E}_{\pi^*,\nu^*}[\delta_h^k(x_h,a_h,b_h)] - \delta_h^k(x_h^k,a_h^k,b_h^k)\big)}_{(l.2)\text{: Statistical Error}} + \underbrace{\sum_{k=1}^K \sum_{h=1}^H (\zeta_{k,h}^1 + \zeta_{k,h}^2)}_{(l.3)\text{: Randomness}},\end{aligned}$$

where $\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle = \langle Q_h^k(x_h^k,\cdot,\cdot,\cdots,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_{f_1,h}^*(\cdot\,|\,x_h^k) \times \cdots \nu_{f_N,h}^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_{f_1,h}^k(\cdot\,|\,x_h^k) \times \cdots \nu_{f_N,h}^k(\cdot\,|\,x_h^k)\rangle_{\mathcal{A}_l \times \mathcal{A}_f}$.

By the same argument of Lemma I.3, we have that, for the matrix game with payoff matrices $(\widetilde{Q}(x_h^k,\cdot,\cdot), \{r_{f_i,h}^k(x_h^k,\cdot,\cdot)\}_{i\in[N]})$, $\nu_h^*(\cdot\,|\,x_h^k)$ belongs to the best response set of $\pi_h^*(\cdot\,|\,x_h^k)$ and breaks ties against favor of the leader. Recall that $(\pi_h^k(\cdot\,|\,x_h^k), \nu_h^k(\cdot\,|\,x_h^k) = \{\nu_{f_i,h}^k(\cdot\,|\,x_h^k)\}_{i\in[N]})$ is the Stackelberg-Nash equilibrium of the matrix game with payoff matrices $(\widetilde{Q}(x_h^k,\cdot,\cdot,\cdot), \{r_{f_i,h}^k(x_h^k,\cdot,\cdot,\cdot)\}_{i\in[N]})$ in the pessimistic tie-breaking setting, which implies that $\pi_h^k(\cdot\,|\,x_h^k)$ is the "worst response to the best response", which further implies that

$$\langle \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle \le 0 \tag{I.8}$$

for any $(k,h) \in [K] \times [H]$. Thus, for any $(k,h) \in [K] \times [H]$, we have

$$\begin{aligned}&\langle Q_h^k(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle \\ &\quad = \langle \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle \\ &\qquad + \langle Q_h^k(x_h^k,\cdot,\cdot) - \widetilde{Q}(x_h^k,\cdot,\cdot), \pi_h^*(\cdot\,|\,x_h^k) \times \nu_h^*(\cdot\,|\,x_h^k) - \pi_h^k(\cdot\,|\,x_h^k) \times \nu_h^k(\cdot\,|\,x_h^k)\rangle \\ &\quad \le \epsilon,\end{aligned} \tag{I.9}$$

where the last inequality uses (D.19) and the fact that $\|Q_h^k - \widetilde{Q}\|_\infty \le \epsilon$. By taking summation over $(k,h) \in [K] \times [H]$, we bound the computational error as desired. Moreover, we can characterize statistical error by Lemmas D.5 and D.6. The remaining randomness term can be bounded by Lemma D.7. Putting these together, we have $\text{Regret}(K) \le \mathcal{O}(\sqrt{d^3H^3T\iota^2})$, which concludes the proof of Theorem I.1. □

J Supporting Lemmas

Lemma J.1 (Elliptical Potential Lemma (Dani et al., 2008; Abbasi-Yadkori et al., 2011; Jin et al., 2020b; Cai et al., 2020)). Let $\{\phi_t\}_{t=1}^\infty$ be an $\mathbb{R}^d$-valued sequence. Meanwhile, let $\Lambda_0 \in \mathbb{R}^{d\times d}$ be a positive-definite matrix and $\Lambda_t = \Lambda_0 + \sum_{j=1}^{t-1} \phi_j\phi_j^\top$. It holds for any $t \in \mathbb{Z}_+$ that

$$\sum_{j=1}^t \min\{1, \|\phi_j\|_{\Lambda_j^{-1}}^2\} \le 2\log\bigg(\frac{\det(\Lambda_{t+1})}{\det(\Lambda_1)}\bigg).$$

Proof. See Lemma 11 of Abbasi-Yadkori et al. (2011) for a detailed proof. □

Lemma J.2 (Concentration of Self-Normalized Process (Abbasi-Yadkori et al., 2011))**.** Let $\{\widetilde{\mathcal{F}}_t\}_{t=0}^{\infty}$ be a filtration and $\{\eta_t\}_{t=1}^{\infty}$ be an $\mathbb{R}$-valued stochastic process such that η_t is $\widetilde{\mathcal{F}}_t$-measurable for any $t \geq 0$. We also assume that, for any $t \geq 0$, conditioning on $\widetilde{\mathcal{F}}_t$, η_t is a zero-mean and σ-sub-Gaussian random variable, that is,

$$\mathbb{E}[\eta_t \,|\, \widetilde{\mathcal{F}}_t] = 0, \qquad \mathbb{E}[e^{\lambda \eta_t} \,|\, \widetilde{\mathcal{F}}_t] \leq e^{\lambda^2 \sigma^2/2} \tag{J.1}$$

for any $\lambda \in \mathbb{R}$. Let $\{X_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued stochastic process such that X_t is $\widetilde{\mathcal{F}}_t$-measurable for any $t \geq 0$. Also, let $Y \in \mathbb{R}^{d\times d}$ be a deterministic and positive-definite matrix. For any $t \geq 0$, we define

$$\overline{Y}_t = Y + \sum_{s=1}^{t} X_s X_s^\top, \quad S_t = \sum_{s=1}^{t} \eta_s \cdot X_s.$$

For any $\delta > 0$ and $t \geq 0$., it holds with probability at least $1 - \delta$ that

$$\|S_t\|^2_{\overline{Y}_t^{-1}} \leq 2\sigma^2 \cdot \log\biggl(\frac{\det(\overline{Y}_t)^{1/2} \det(Y)^{-1/2}}{\delta}\biggr).$$

Proof. See Theorem 1 of Abbasi-Yadkori et al. (2011) for a detailed proof. □

Lemma J.3. For any fixed $h \in [H]$, let $V : \mathcal{S} \to [0, H]$ be any fixed value function. Under Assumption 4.1, for any fixed $\delta > 0$, we have

$$P_{\mathcal{D}}\biggl(\Bigl\| \sum_{k=1}^{K} \phi(x_h^\tau, a_h^\tau, b_h^\tau) \cdot \bigl(V(x_{h+1}^\tau) - \mathbb{P}_h V(x_h^\tau, a_h^\tau, b_h^\tau)\bigr) \Bigr\|_{\Lambda_h^{-1}} > H^2 \cdot \bigl(2\log(1/\delta) + d \cdot \log(1+K)\bigr)\biggr) \leq \delta.$$

Proof. See Lemma B.2 of Jin et al. (2020c) for a detailed proof. □

Lemma J.4 (Covering)**.** Let $\mathcal{Q}_h$ be the class of value functions $Q : \mathcal{S} \times \mathcal{A}_l \times \mathcal{A}_f \to \mathbb{R}$ that takes the form

$$Q(\cdot,\cdot,\cdot) = r_{l,h}(\cdot,\cdot,\cdot) + \Pi_{H-h}\{(\phi(\cdot,\cdot,\cdot)^\top w + \beta \cdot \bigl(\phi(\cdot,\cdot,\cdot)^\top \Lambda^{-1} \phi(\cdot,\cdot,\cdot)\bigr)^{1/2}\},$$

which are parameterized by $(w, \Lambda) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$ such that $\|w\| \leq L$ and $\lambda_{\min}(\Lambda) \geq \lambda$. We assume that β is fixed and satisfy that $\beta \in [0, B]$, and the feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ satisfies that $\|\phi(\cdot,\cdot)\|_2 \leq 1$. We have that, for any $L, B, \epsilon > 0$, there exists an ϵ-covering of $\mathcal{Q}_h$ with respect to the ℓ_∞ norm such that the covering number $\mathcal{N}_\epsilon$ satisfies

$$\log \mathcal{N}_\epsilon \leq d \cdot \log(1 + 4L/\epsilon) + d^2 \cdot \log\bigl(1 + 8B^2\sqrt{d}/(\epsilon^2\lambda)\bigr).$$

Proof. See Jin et al. (2020b) for a detailed proof. □