Proceedings of Machine Learning Research vol 167:1–35, 2022

Almost Optimal Algorithms for Two-player Zero-Sum Linear Mixture Markov Games

Zixiang Chen CHENZX19@CS.UCLA.EDU
Department of Computer Science, University of California, Los Angeles, CA 90095, USA

Dongruo Zhou DRZHOU@CS.UCLA.EDU
Department of Computer Science, University of California, Los Angeles, CA 90095, USA

Quanquan Gu QGU@CS.UCLA.EDU
Department of Computer Science, University of California, Los Angeles, CA 90095, USA

Abstract

We study reinforcement learning for two-player zero-sum Markov games with simultaneous moves in the finite-horizon setting, where the transition kernel of the underlying Markov games can be parameterized by a linear function over the current state, both players' actions and the next state. In particular, we assume that we can control both players and aim to find the Nash Equilibrium by minimizing the duality gap. We propose an algorithm Nash-UCRL based on the principle "Optimism-in-Face-of-Uncertainty". Our algorithm only needs to find a Coarse Correlated Equilibrium (CCE), which is computationally efficient. Specifically, we show that Nash-UCRL can provably achieve an $\widetilde{O}(dH\sqrt{T})$ regret, where d is the linear function dimension, H is the length of the game and T is the total number of steps in the game. To assess the optimality of our algorithm, we also prove an $\widetilde{\Omega}(dH\sqrt{T})$ lower bound on the regret. Our upper bound matches the lower bound up to logarithmic factors, which suggests the optimality of our algorithm.

Keywords: Markov Games; Reinforcement Learning; Linear Function Approximation.

1. Introduction

Multi-agent reinforcement learning (MARL) has achieved tremendous practical success across a wide range of machine learning tasks, including large-scale strategy games such as GO (Silver et al., 2016), TexasHold'em poker (Brown and Sandholm, 2019), real-time video games such as Starcraft (Vinyals et al., 2019), and autonomous driving (Shalev-Shwartz et al., 2016). Among these models used in MARL, two-player zero-sum Markov games (MG) (Shapley, 1953; Littman, 1994) is probably one of the most widely studied models and can be regarded as a generalization of the Markov Decision Processes (MDP) (Puterman, 2014).

In two-player Markov games, the two players share states, play actions simultaneously and independently, and observe the same reward. One player (i.e., max-player) aims to maximize the return while the other (i.e., min-player) aims to minimize it. A special case of general Markov games (i.e., simultaneous-move games) is turn-based games, where only one player can take action in each step, i.e., the max and min players take turns to play the game. The players aim to find the Nash equilibrium for this game. Most existing results on learning two-player Markov games either assume the access to a generative model that can sample the next state for an arbitrary state-action pair (Jia et al., 2019; Sidford et al., 2020; Cui and Yang, 2020), or a well-explored behavior policy (Lagoudakis and Parr, 2012; Perolat et al., 2015; Pérolat et al., 2016a,b, 2017), and fail to consider the exploration-exploitation tradeoff (Kearns and Singh, 2002).

In order to get rid of the generative model and well-explored behavior policy assumptions, Wei et al. (2017) extended the UCRL2 algorithm (Jaksch et al., 2010) for MDP to zero-sum simultaneous-move Markov games in the average-reward setting, and proposed the UCSG algorithm that achieves a sublinear regret when competing with an arbitrary opponent. Recently, Bai and Jin (2020); Bai et al. (2020); Liu et al. (2020) proposed a series of algorithms for learning tabular episodic two-player zero-sum Markov games (they call it self-play algorithm for competitive reinforcement learning), and proved the upper and lower regret bounds and/or sample complexity. For Markov games with large state and action spaces, it is natural to use linear function approximation. In particular, Xie et al. (2020) proposed the OMNI-VI algorithm for Markov games where the transition kernel and reward function possess a linear structure, and achieved an $\widetilde{O}(\sqrt{d^3H^3T})$ regret bound, with d being the dimension of the linear structure and H being the episode length. However, as we will show in this paper, the information theoretic lower bound for the zero-sum two-player Markov games with linear structures is $\Omega(dH\sqrt{T})$. Therefore, there is still a gap between the upper and lower bounds of existing algorithms for Markov games with linear structures. This raises the following question:

Can we design a minimax optimal algorithm for learning zero-sum Markov games with linear function approximation?

In this paper, we give an affirmative answer to the above question for a class of episodic Markov games in the offline setting[1], where both players are controlled by a central learner. The goal of the central learner is to find an approximate Nash Equilibrium (NE) of the game, with the approximation error measured by a notion of duality gap. In particular, we consider Markov games with a linear mixture structure, where the transition probability kernel is a linear mixture model that is inspired by the linear mixture MDPs studied in (Modi et al., 2020; Jia et al., 2020; Ayoub et al., 2020; Zhou et al., 2021b). We propose the first nearly minimax optimal algorithm based on the principle of "Optimism-in-Face-of-Uncertainty" without assuming the access to the generative model or well-explored behavior policy. We summarize the contributions of our work as follows:

- We propose a Nash-UCRL algorithm for general Markov games (i.e., simultaneous-move game) that can provably achieve an $\widetilde{O}(dH\sqrt{T})$ upper bound on the regret, where d is the dimension of linear mixture structure, H is the length of the game, and T the total number of steps in the Markov game. Our algorithm can be specialized to turn-based games and also achieves $\widetilde{O}(dH\sqrt{T})$ regret.
- To access the optimality of our algorithm Nash-UCRL, we prove an $\Omega(dH\sqrt{T})$ regret lower bound . Our upper bound matches the lower bound up to logarithmic factors, which suggests the optimality of our algorithm. While our lower bound is proved for Markov games with linear mixture structure, we argue that it is also a valid lower bound for Markov games with linear structure (Xie et al., 2020).

Notation We use lower case letters to denote scalars, lower and upper case bold letters to denote vectors and matrices. We use $\|\cdot\|$ to indicate Euclidean norm, and for a semi-positive definite matrix $\mathbf{\Sigma}$ and any vector $\mathbf{x}$, $\|\mathbf{x}\|_{\mathbf{\Sigma}} := \|\mathbf{\Sigma}^{1/2}\mathbf{x}\| = \sqrt{\mathbf{x}^\top\mathbf{\Sigma}\mathbf{x}}$. For a real value x and an interval $[a, b]$, we

1. Here we follow the same terminology "offline setting" as in Xie et al. (2020), which is also called "self-play" in Bai and Jin (2020); Bai et al. (2020)

use $[x]_{[a,b]}$ to indicate the projection of x onto $[a, b]$. We also use the standard O and Ω notations. We say $a_n = O(b_n)$ if and only if $\exists C > 0, N > 0, \forall n > N, a_n \leq Cb_n$; $a_n = \Omega(b_n)$ if and only if $\exists C > 0, N > 0, \forall n > N, a_n \geq Cb_n$. The notation $\widetilde{O}$ is used to hide logarithmic factors.

2. Related Work

Tabular Markov game. Under the tabular setting, Littman and Szepesvári (1996) extended the value iteration and Q-learning algorithms (Watkins, 1989) to zero-sum Markov games. Littman (2001); Greenwald et al. (2003); Hu and Wellman (2003) further extended it to general-sum Markov games with n-player. Hansen et al. (2013) provided the first strong polynomial algorithm for solving two-player turn-based Markov games. Sidford et al. (2018) proposed a variance-reduced variant of the minimax Q-learning algorithm with near-optimal sample complexity. Lagoudakis and Parr (2012); Perolat et al. (2015); Fan et al. (2020) considered value-iteration with function approximation and established finite-time convergence to the NEs of two-player zero-sum Markov games. Their results are based on the framework of fitted value-iteration (Munos and Szepesvári, 2008). Jia et al. (2019) studied turn-based zero-sum Markov games, where the transition model is assumed to be embedded in some d-dimensional feature space. Cui and Yang (2020) proposed an algorithm for turn-based zero-sum Markov games based on plug-in estimator and achieved minimax sample complexity. For the simultaneous-move zero-sum Markov games, Zhang et al. (2020) proposed an algorithm which achieved minimax sample complexity if the algorithm is reward-agnostic. All the above works either assume a generative oracle or a well explored behavioral policy for drawing transitions, therefore bypassing the exploration issue. Bai and Jin (2020) proposed a VI-ULCB algorithm for tabular episodic zero-sum Markov games, which achieves $\widetilde{O}(\sqrt{H^3S^2ABT})$ regret for simultaneous move (i.e., general Markov game) and $O(\sqrt{H^3S^2(A+B)T})$ regret for turn-based game, where A and B are the number of actions for each player, H is the length of the game, and T is the total number of steps played in the game. They also proved an $\Omega(\sqrt{H^2S(A+B)T})$ lower bound. For general Markov game, Bai et al. (2020) proposed an Optimistic Nash Q-learning algorithm with a regret of $\widetilde{O}(\sqrt{H^4SABT})$, and an Optimistic Nash V-learning algorithm with a regret of $\widetilde{O}(\sqrt{H^5S(A+B)T})$, both of which improve the regret in Bai and Jin (2020) in the dependence on S, A, B. The best known regret is achieved by Nash-VI proposed in Liu et al. (2020), which is $\widetilde{O}(\sqrt{H^2SABT})$. As can be seen, without assuming the access to a generative model or a well explored behavioral policy, there is still a gap between the upper and lower regret bounds for existing algorithms, even for the simplest tabular Markov games.

Online RL with linear function approximation. There are several lines of work aiming at providing theoretical guarantees for online RL with function approximation. The first line of work focus on the linear function approximation setting, which assumes that the MDP (e.g., transition probability, reward, or value function) can be represented as a linear function of some given feature mapping. These works proposed algorithms which enjoy sample complexity/regret scaling with the dimension of the feature mapping, rather than the cardinality of state and action spaces. For example, Yang and Wang (2019a); Jin et al. (2020); Wang et al. (2019); Zanette et al. (2020a); He et al. (2021) considered the linear MDP model, where the transition probability function and reward function are linear in some feature mapping over state-action pairs. Zanette et al. (2020b) studied MDPs with low inherent Bellman error, where the value functions are nearly linear w.r.t. the feature mapping. Yang and Wang (2019b); Modi et al. (2020); Jia et al. (2020); Ayoub et al. (2020); Cai et al. (2019); Zhou et al. (2021b); He et al. (2021) studied the linear mixture MDPs, where the transition probability kernel is a linear mixture of a number of basis kernels. Inspired by linear mixture MDPs, we introduce the linear mixture Markov game.

3. Preliminaries

In this section, we introduce the setup of the episodic two-player zero-sum Markov games with simultaneous moves and the linear mixture structure we use in this paper.

3.1. Two-Player Markov Games

The two-player zero-sum Markov game (MG) (Shapley, 1953; Littman, 1994) is a generalization of the standard Markov decision process (MDP) where the max-player seeks to maximize the total return, and the min-player seeks to minimize the total return.

Simultaneous-move MG. Formally, we denote a two-player zero-sum simultaneous-moves episodic Markov Game by a tuple $M(\mathcal{S}, \mathcal{A}_{\max}, \mathcal{A}_{\min}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$. $\mathcal{S}$ is a countable state space, $\mathcal{A}_{\max}, \mathcal{A}_{\min}$ are the finite action spaces of the max-player and the min-player respectively. H is the length of the game/episode. For simplicity, we assume the reward function for the max-player $\{r_h\}_{h=1}^H$ is deterministic and known function $r_h : \mathcal{S} \times \mathcal{A}_{\max} \times \mathcal{A}_{\min} \to [-1, 1]$. $\mathbb{P}_h(s'|s, a, b)$ is the transition probability function which denotes the probability for state s to transit to state s' given players' action pair (a, b) at step h.

Markov Policy and Value Function. We first define the stochastic policies, which give distributions over the actions. A policy $\pi = \{\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}_{\max}}\}_{h=1}^H$ is a collection of functions which map a state $s \in \mathcal{S}$ to a distribution of actions. Here $\Delta_{\mathcal{A}_{\max}}$ is the probability simplex over action set $\mathcal{A}_{\max}$. Similarly, we can define a policy $\nu = \{\nu_h : \mathcal{S} \to \Delta_{\mathcal{A}_{\min}}\}_{h=1}^H$ for the min-player, where $\Delta_{\mathcal{A}_{\min}}$ is the probability simplex over action set $\mathcal{A}_{\min}$. We use the notation $\pi_h(a|s)$ and $\nu_h(b|s)$ to present the probability of taking action a or b for state s at step h under Markov policy π, ν respectively. We define the action-value function (a.k.a., Q function) $Q_h^{\pi,\nu} : \mathcal{S} \times \mathcal{A}_{\max} \times \mathcal{A}_{\min} \to \mathbb{R}$ as follows

$$Q_h^{\pi,\nu}(s, a, b) = \mathbb{E}_{\pi,\nu,h,s,a,b}\bigg[\sum_{h'=h}^{H} r(s_{h'}, a_{h'}, b_{h'})\bigg| s_h = s, a_h = a, b_h = b\bigg],$$

and the value function $V_h^{\pi,\nu} : \mathcal{S} \to \mathbb{R}$ as follows

$$V_h^{\pi,\nu}(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s), b\sim\nu_h(\cdot|s)} Q_h^{\pi,\nu}(s, a, b), \qquad V_{H+1}^{\pi,\nu}(s) = 0.$$

In the definition of $Q_h^{\pi,\nu}$, $\mathbb{E}_{\pi,\nu,h,s,a,b}$ is an expectation over state-action pairs of length $H - h + 1$ induced by the policy (π, ν) and the transition probability of the MG M, when initializing the process with the triplet (s, a, b) at step h. Because $r_h(\cdot, \cdot, \cdot) \in [-1, 1]$, it is easy see that both Q functions and value functions are bounded

$$|Q_h^{\pi,\nu}(\cdot, \cdot, \cdot)| \le H, \qquad |V_h^{\pi,\nu}(\cdot)| \le H.$$

Furthermore, for any joint distribution $\sigma \in \Delta(\mathcal{A}_{\max} \times \mathcal{A}_{\min})$, we denote by $\mathcal{P}_{\max}\sigma$ the marginal distribution for the max-player and by $\mathcal{P}_{\min}\sigma$ the marginal distribution for the min-player.

Best Response and Bellman Equation. The goal of the max-player is to maximize the total rewards. The goal of the min-player is to minimize the total rewards that the max-player will get because this is a zero-sum game. In other words, the max-player wants to maximize $V_h^{\pi,\nu}(\cdot)$ by choosing a good policy π, while the min-player wants to minimize $V_h^{\pi,\nu}(\cdot)$ by choose a good policy ν. Accordingly, we can define the action-value function and the value function when the max-player gives the best response to a fixed policy ν of the min-player:

$$Q_h^{*,\nu}(s,a,b) = \max_{\pi} Q_h^{\pi,\nu}(s,a,b), \qquad V_h^{*,\nu}(s) = \max_{\pi} V_h^{\pi,\nu}(s).$$

By symmetry, we can also define

$$Q_h^{\pi,*}(s,a,b) = \min_{\nu} Q_h^{\pi,\nu}(s,a,b), \qquad V_h^{\pi,*}(s) = \min_{\nu} V_h^{\pi,\nu}(s).$$

For any function $V : \mathcal{S} \to \mathbb{R}$, we introduce the shorthands:

$$[\mathbb{P}_h V](s,a,b) = \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a,b)} V(s'), \qquad [\mathbb{V}_h V](s,a,b) = [\mathbb{P}_h V^2](s,a,b) - \big([\mathbb{P}_h V](s,a,b)\big)^2,$$

where V^2 stands for the function whose value at s is $V^2(s)$. Using these notation, we have following Bellman equations:

$$Q_h^{\pi,\nu}(s,a,b) = r(s,a,b) + [\mathbb{P}_h V_{h+1}^{\pi,\nu}](s,a,b),$$

and the Bellman optimality equation (Shapley, 1953):

$$Q_h^{\pi,*}(s,a,b) = r(s,a,b) + [\mathbb{P}_h V_{h+1}^{\pi,*}](s,a,b), \qquad V_h^{\pi,*}(s) = \inf_{\sigma \in \Delta_{\min}} \mathbb{E}_{a \sim \pi_h(\cdot|s), b \sim \sigma} Q_h^{\pi,\nu}(s,a,b).$$

Nash Equilibrium. A Nash Equilibrium (NE) of the game is a pair of policies π^*, ν^* such that

$$V_1^{\pi^*,\nu^*}(s) = V_1^{\pi^*,*}(s) = V_1^{*,\nu^*}(s), \text{ for all } s \in \mathcal{S}. \tag{3.1}$$

(3.1) means that (π^*, ν^*) are the best response to each other, so no player can do better by only changing her own policy. Nash equilibrium can also be viewed as "the best response to the best response". For most applications, they are the ultimate solutions we want to pursue. We further abbreviate the value of the Nash equilibrium $V_1^{\pi^*,\nu^*}(s)$ as $V_1^*(s)$. This is because the value of the Nash equilibrium is irrelevant to the choice of (π^*, ν^*) which is a direct corollary of the following weak duality property:

Proposition 3.1 (Weak Duality, Xie et al. 2020) *Given the NE (π^*, ν^*) of a game, for any policy pair (π, ν) we have that*

$$V_1^{*,\nu}(s) \geq V_1^{\pi^*,\nu^*}(s) \geq V_1^{\pi,*}(s), \textit{ for all } s \in \mathcal{S}. \tag{3.2}$$

Learning Objective. The weak duality in Proposition 3.1 suggests that the NE value $V_1^*(s)$ is sandwiched between $V_1^{*,\nu}(s)$ and $V_1^{\pi,*}(s)$. So it is natural to measure the suboptimality of learned policies (π^k, ν^k) at the k-th episode by the gap between their performance and the performance of the optimal strategy (i.e., Nash equilibrium) when playing against the best responses respectively:

$$V_1^{*,\nu^k}(s) - V_1^{\pi^k,*}(s) = \big[V_1^{*,\nu^k}(s) - V_1^*(s)\big] + [V_1^*(s) - V_1^{\pi^k,*}(s)].$$

Accordingly, we aim to design a learning algorithm that outputs a sequence $\{\pi^k, \nu^k\}_k$ based on past information, and minimize the regret over first K episodes defined as follows:

$$\text{Regret}(M, K) = \sum_{k=1}^{K} \Big[V_1^{*,\nu^k}(s_1^k) - V_1^{\pi^k,*}(s_1^k) \Big].$$

This measure has been widely used in previous work that studies the offline learning of two-player game (Bai et al., 2020; Xie et al., 2020; Liu et al., 2020). Following Bai et al. (2020); Xie et al. (2020); Liu et al. (2020), we assume the central controller can choose a joint distribution μ^k for both the max-player and min-player in each episode as their policies, and we set $\pi^k = \mathcal{P}_{\text{max}}\mu^k$ and $\nu^k = \mathcal{P}_{\text{min}}\mu^k$ automatically. In this paper, we focus on proving high probability bounds on the regret Regret(M, K), as well as lower bounds in expectation.

Episodic Linear Mixture Markov Games. In this work, we consider a class of MGs called *linear mixture MGs*, inspired by the linear mixture/kernel MDPs studied in Modi et al. (2020); Jia et al. (2020); Ayoub et al. (2020) for the single-agent RL. Linear mixture MGs assume that at each step h, the transition probability function $\mathbb{P}_h(s'|s, a, b)$ is a linear combination of d feature mappings $\phi_i(s'|s, a, b)$, i.e.,

$$\mathbb{P}_h(s'|s, a, b) = \sum_{i=1}^{d} \theta_{i,h}\phi_i(s'|s, a, b),$$

where each feature mapping $\phi_i(s'|s, a, b)$ is a function defined on the state-action-action-state pair $(s, a, b, s') \in \mathcal{S} \times \mathcal{A}_{\text{max}} \times \mathcal{A}_{\text{max}} \times \mathcal{S}$. For the sake of simplicity, we use a vector function $\boldsymbol{\phi} = [\phi_1, \cdots, \phi_d] \in \mathbb{R}^d$ to denote the collection of ϕ_i. After proper normalization, we assume $\boldsymbol{\phi}$ satisfy that for any bounded function $V : \mathcal{S} \to [-1, 1]$ and any tuple $(s, a, b) \in \mathcal{S} \times \mathcal{A}_{\text{max}} \times \mathcal{A}_{\text{min}}$, we have

$$\|\boldsymbol{\phi}_V(s, a, b)\|_2 \le 1, \tag{3.3}$$

where $\boldsymbol{\phi}_V(s, a, b) = \sum_{s' \in \mathcal{S}} \boldsymbol{\phi}(s'|s, a, b)V(s')$. Formally, we define linear mixture MGs as follows:

Definition 3.2 *$M(\mathcal{S}, \mathcal{A}_{\text{max}}, \mathcal{A}_{\text{min}}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$ is called a time inhomogeneous, episodic B-bounded linear mixture MG if there exist H unknown vectors $\boldsymbol{\theta}_h \in \mathbb{R}^d$ satisfying for any $h \in [H]$, $\|\boldsymbol{\theta}_h\|_2 \le B$, and a known feature mapping $\boldsymbol{\phi}$ satisfying (3.3), such that $\mathbb{P}_h(s'|s, a, b) = \langle \boldsymbol{\phi}(s'|s, a, b), \boldsymbol{\theta}_h \rangle$ for any state-action-action-state triplet (s, a, b, s') and any step h. We denote the linear mixture MG by $M_{\boldsymbol{\theta}}$ for simplicity.*

In this paper, we assume the underlying linear mixture MG is parameterized by $\{\boldsymbol{\theta}_h^*\}_{h=1}^H$, denoted by $M_{\boldsymbol{\theta}^*}$.

Difference between linear and linear mixture MGs. Linear mixture MGs assume that at each step h, the transition probability function $\mathbb{P}_h(s'|s, a, b)$ is a linear combination of d feature mappings $\boldsymbol{\phi}_i(s'|s, a, b)$ for $i = 1, \ldots, d$, i.e., $\mathbb{P}_h(s'|s, a, b) = \langle \boldsymbol{\phi}(s'|s, a, b), \boldsymbol{\theta}_h \rangle$. The linear MG setting considered by Xie et al. (2020), however, assumes $\mathbb{P}_h(s'|s, a, b) = \langle \boldsymbol{\phi}(s, a, b), \boldsymbol{\mu}_h(s') \rangle$, where $\boldsymbol{\mu}_h(\cdot)$ is an unknown vector-valued measure function on S. These two models are different and do not include each other in general. For instance, consider the following MG which is inspired by Zhou et al. (2021b): $\mathcal{S} = \mathbb{Z}$, $\mathcal{A}_{\text{max}} = \mathcal{A}_{\text{min}} = \mathbb{N}$ and $\mathbb{P}_h(s'|s, a, b) = \sum_{i=1}^d \theta_i^h p_i(s'|s, a, b)$, $p_i(s'|s, a, b) = \mathbb{1}(s' >= s)(a + b)^{s'-s} \exp(-(a + b))/(s' - s)!$. This MG is a linear mixture MG but not a linear MG.

4. Algorithm

In this section, we propose our algorithm Nash-UCRL in Algorithm 1. Due to the space limit, we only show the detailed update rules for the max-player in Algorithm 1, and the full algorithm is presented in Algorithm 2 in Appendix D. All the parameters corresponding to the max-player are marked by an overline, while the parameters for the min-player are marked by an underline.

Algorithm 1 Nash-UCRL

Input: Regularization parameter λ, number of episode K, number of horizon H, approximation error ϵ.
For any h, $\overline{\boldsymbol{\Sigma}}_{1,h}^{(i)} \leftarrow \underline{\boldsymbol{\Sigma}}_{1,h}^{(i)} \leftarrow \lambda \mathbf{I}$; $\overline{\mathbf{b}}_{1,h}^{(i)} \leftarrow \underline{\mathbf{b}}_{1,h}^{(i)} \leftarrow \mathbf{0}$; $\overline{\boldsymbol{\theta}}_{1,h}^{(i)} \leftarrow \underline{\boldsymbol{\theta}}_{1,h}^{(i)} \leftarrow \mathbf{0}$, for $i \in \{0, 1\}$.
for $k = 1, \ldots, K$ **do**
  $\overline{V}_{k,H+1}(\cdot) \leftarrow 0$, $\underline{V}_{k,H+1}(\cdot) \leftarrow 0$.
  for $h = H, \ldots, 1$ **do**
    Set $\overline{Q}_{k,h}(\cdot,\cdot,\cdot)$ as in (4.4), and $\underline{Q}_{k,h}(\cdot,\cdot,\cdot)$ in a similar way (See Algorithm 2).
    for $s \in \mathcal{S}$ **do**
      Let $\mu_h^k(\cdot,\cdot|s) = \epsilon\text{-CCE}(\overline{Q}_{k,h}(s,\cdot,\cdot), \underline{Q}_{k,h}(s,\cdot,\cdot))$.
      $\overline{V}_{k,h}(s) = \mathbb{E}_{(a,b)\sim\mu_h^k(\cdot,\cdot|s)} \overline{Q}_{k,h}(s,a,b)$, $\underline{V}_{k,h}(s) = \mathbb{E}_{(a,b)\sim\mu_h^k(\cdot,\cdot|s)} \underline{Q}_{k,h}(s,a,b)$.
      $\pi_h^k(\cdot|s) = \mathcal{P}_{\max}\mu_h^k(\cdot,\cdot|s)$, $\nu_h^k(\cdot|s) = \mathcal{P}_{\min}\mu_h^k(\cdot,\cdot|s)$.
    end for
  end for
  Receives s_1^k
  for $h = 1, \ldots, H$ **do**
    Take action $(a_h^k, b_h^k) \sim \mu_h^k(\cdot,\cdot|s_h^k)$ and central controller receives $s_{h+1}^k \sim \mathbb{P}(\cdot|s_h^k, a_h^k, b_h^k)$.
    Set $\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)$ as in (4.10) and $\overline{E}_{k,h}$ as in (4.13), $\overline{\sigma}_{k,h}$ as in (4.6).
    Set $\overline{\boldsymbol{\Sigma}}_{k+1,h}^{(0)}, \overline{\mathbf{b}}_{k+1,h}^{(0)}$ as in (4.7) and (4.8), $\overline{\boldsymbol{\Sigma}}_{k+1,h}^{(1)}, \overline{\mathbf{b}}_{k+1,h}^{(1)}$ as in (4.11) and (4.12).
    Set $\underline{\boldsymbol{\Sigma}}_{k+1,h}^{(0)}, \underline{\mathbf{b}}_{k+1,h}^{(0)}, \underline{\boldsymbol{\Sigma}}_{k+1,h}^{(1)}, \underline{\mathbf{b}}_{k+1,h}^{(1)}, \mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k), \underline{E}_{k,h}, \underline{\sigma}_{k,h}$ in similar ways (See Algorithm 2).
    Set $\overline{\boldsymbol{\theta}}_{k+1,h}^{(i)} \leftarrow \big[\overline{\boldsymbol{\Sigma}}_{k+1,h}^{(i)}\big]^{-1}\overline{\mathbf{b}}_{k+1,h}^{(i)}$, $\underline{\boldsymbol{\theta}}_{k+1,h}^{(i)} \leftarrow \big[\underline{\boldsymbol{\Sigma}}_{k+1,h}^{(i)}\big]^{-1}\underline{\mathbf{b}}_{k+1,h}^{(i)}$, $i = 0, 1$
  end for
end for

To achieve the near-minimax optimality of solving a linear mixture MG, Nash-UCRL adopts the following three techniques, which we will introduce in sequence.

Value-targeted regression To find the NE of an MG, it suffices to find good estimates of the optimal value functions V_h^{*,ν^k} and $V_h^{\pi^k,*}$. By the Bellman optimality equations and the definition of linear mixture MGs, it is sufficient to estimate the underlying unknown parameter $\boldsymbol{\theta}_h^*$ up to good accuracy. Inspired by the UCRL with "value-targeted regression" (VTR) proposed by Jia et al. (2020); Ayoub et al. (2020), Nash-UCRL uses a supervised learning framework to learn $\boldsymbol{\theta}_h^*$. In the sequel, we introduce how the VTR framework works at episode k and step h. At the beginning of episode k, Nash-UCRL maintains two estimated value functions: optimistic value function $\overline{V}_{k,h+1}$ for the max-player, which overestimates the optimal value function V_h^{*,ν^k}, and optimistic value function $\underline{V}_{k,h+1}$ for the min-player, which underestimates the value function $V_h^{\pi^k,*}$. We focus on the overestimate $\overline{V}_{k,h+1}$ first. Note that the following equation holds due to the definition of linear mixture

MGs:

$$[\mathbb{P}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) = \langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^* \rangle, \tag{4.1}$$

which suggests that $(\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{V}_{k,h+1}(s_{h+1}^k))$ can be regarded as a context and the corresponding targeted value of a linear regression problem with the unknown parameter $\boldsymbol{\theta}_h^*$.

Therefore, Nash-UCRL constructs $\overline{\boldsymbol{\theta}}_{k,h}^{(0)}$ as the estimator of $\boldsymbol{\theta}_h^*$ based on linear regression on $(\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{V}_{k,h+1}(s_{h+1}^k))$ (the detailed construction of $\overline{\boldsymbol{\theta}}_{k,h}^{(0)}$ will be specified later). Due to the randomness of s_{h+1}^k, $\overline{\boldsymbol{\theta}}_{k,h}^{(0)}$ can not estimate $\boldsymbol{\theta}_h^*$ exactly. Therefore Nash-UCRL also constructs an ellipsoid $\overline{\mathcal{C}}_{k,h}^{(0)}$ centered at $\overline{\boldsymbol{\theta}}_{k,h}^{(0)}$ as the confidence set, which contains $\boldsymbol{\theta}_h^*$ with high probability:

$$\overline{\mathcal{C}}_{k,h}^{(0)} := \bigg\{\boldsymbol{\theta} : \bigg\| \Big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\Big]^{1/2}(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_{k,h}^{(0)}) \bigg\|_2 \le \beta_k^{(0)} \bigg\}. \tag{4.2}$$

Here $\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}$ is the "covariance matrix" of the context $\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)$, and $\beta_k^{(0)}$ is the radius of the confidence set. Both of them will be specified later. Then, to encourage the agent to explore, Nash-UCRL constructs an optimistic action-value function $\overline{Q}_{k,h}$ as follows, following the "optimism-in-the-face-of-uncertainty" principle (Abbasi-Yadkori et al., 2011):

$$\overline{Q}_{k,h} := \Big[r_h + \max_{\boldsymbol{\theta} \in \overline{\mathcal{C}}_{k,h}^{(0)}} \langle \boldsymbol{\theta}_{k,h}, \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \rangle \Big]_{[-H,H]}, \tag{4.3}$$

where the projection onto $[-H, H]$ is because the action-value function of the Markov game lies in $[-H, H]$. The closed-form solution of (4.3) is as follows

$$\overline{Q}_{k,h}(\cdot,\cdot,\cdot) = \Big[r_h(\cdot,\cdot,\cdot) + \langle \overline{\boldsymbol{\theta}}_{k,h}^{(0)}, \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(\cdot,\cdot,\cdot) \rangle + \beta_k^{(0)} \Big\| \Big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\Big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(\cdot,\cdot,\cdot) \Big\|_2 \Big]_{[-H,H]}, \tag{4.4}$$

Similar procedures can be applied to construct the confidence set $\underline{\mathcal{C}}_{k,h}^{(0)}$ and the optimistic action-value function $\underline{Q}_{k,h}$ for the min-player with parameters $\underline{\boldsymbol{\theta}}_{k,h}^{(0)}$, $\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}$. Finally, Nash-UCRL constructs the optimistic value functions $\overline{V}_{k,h}, \underline{V}_{k,h}$ and the policy μ_h^k based on $\overline{Q}_{k,h}, \underline{Q}_{k,h}$ for the current episode and step (which will be specified later).

Coarse Correlated Equilibrium (CCE). Now we introduce how to compute the (π_h^k, ν_h^k) based on the optimistic action-value functions $\overline{Q}_{k,h}, \underline{Q}_{k,h}$. Unlike the single-agent RL, we cannot certify the policy by independently solving max-min problem on $\overline{Q}$ or $\underline{Q}$. This is because $\overline{Q}$ and $\underline{Q}$ are not the estimators of action-value function for the NE but the estimators of action-value function for the best response. Thus we must coordinate both players for their choices of actions. After we get $\overline{Q}_{k,h}(s,\cdot,\cdot)$ for the max-player and $\underline{Q}_{k,h}(s,\cdot,\cdot)$ for the min-player, we solve a general-sum matrix game to find the Coarse Correlated Equilibrium (CCE), following Xie et al. (2020). Here we give the formal definition of CCE as follows:

Definition 4.1 (Moulin and Vial 1978; Aumann 1987) *Given two payoff matrices $Q_{max}, Q_{min} \in \mathbb{R}^{|\mathcal{A}_{max}|\cdot|\mathcal{A}_{min}|}$, we denote the ϵ-Coarse Correlated Equilibrium (ϵ-CCE) as a joint distribution σ over $\mathcal{A}_{max}$ and $\mathcal{A}_{min}$ satisfying that*

$$\mathbb{E}_{(a,b)\sim\sigma} Q_{max}(a,b) \ge \max_{a' \in \mathcal{A}_{max}} \mathbb{E}_{b \sim \mathcal{P}_{min}\sigma} Q_{max}(a', b) - \epsilon,$$

$$\mathbb{E}_{(a,b)\sim\sigma} Q_{min}(a,b) \le \min_{b'\in\mathcal{A}_{min}} \mathbb{E}_{a\sim\mathcal{P}_{max}\sigma} Q_{min}(a,b') + \epsilon.$$

Nash-UCRL computes the distribution $\mu_h^k(\cdot,\cdot|s)$, a ϵ-CCE of $\overline{Q}_{k,h}, \underline{Q}_{k,h}$ for each state s in Line 8. Then Nash-UCRL selects the value functions $\overline{V}_{k,h}, \underline{V}_{k,h}$ as the expectation of $\overline{Q}_{k,h}, \underline{Q}_{k,h}$ over the policies μ_h^k as in Line 9 of Algorithm 1. The difference between CCE and NE is whether the policy of each player is independent of each other. The policy μ_h^k given by ϵ-CCE is correlated for each player because it is found in the class $\Delta_{\mathcal{A}_{\max}\times\mathcal{A}_{\min}}$ rather than $\Delta_{\mathcal{A}_{\max}} \times \Delta_{\mathcal{A}_{\min}}$. After obtaining μ^k, Nash-UCRL sets $\pi_h^k(\cdot|s) = \mathcal{P}_{\max}\mu_h^k(\cdot,\cdot|s)$ and $\nu_h^k(\cdot|s) = \mathcal{P}_{\min}\mu_h^k(\cdot,\cdot|s)$, i.e., the marginal distributions of μ_h^k. Notice that a Nash equilibrium always exists and a Nash equilibrium for a general-sum game is also a CCE. Thus a CCE always exists, so does ϵ-CCE.

Remark 4.2 *Since we assume the action spaces are finite, the constraints for ϵ-CCE can be rewritten as $|\mathcal{A}_{\max}| + |\mathcal{A}_{\min}|$ linear constraints, which can be efficiently solved by linear programming (See e.g., Bai et al. (2020); Liu et al. (2020)).*

Weighted linear regression for value function estimation. Now we specify how to construct the estimators $\overline{\boldsymbol{\theta}}_{k,h}^{(0)}, \underline{\boldsymbol{\theta}}_{k,h}^{(0)}$. For the simplicity, we only show the construction for the max-player and the construction for the min-player is presented in Appendix D. With the linear structure of $\overline{V}_{k,h+1}(s_{h+1}^k)$ in (4.1), it is natural to set the estimator $\overline{\boldsymbol{\theta}}_{k,h}$ as the minimizer to the linear regression problem with square loss over context-target pairs $(\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{V}_{k,h+1}(s_{h+1}^k))$, similar to UCRL (Jia et al., 2020; Ayoub et al., 2020). However, such an estimator is somehow limited since it treats each context-target pair equally and ignore the difference between these pairs. In principle, one should pay *more* attention to the pairs with *less* target variance since they carry more information about the unknown parameter $\boldsymbol{\theta}_h^*$. This observation inspires us to adapt the recently proposed *weighted ridge regression* scheme by Zhou et al. (2021a) to estimate $\boldsymbol{\theta}_h^*$:

$$\overline{\boldsymbol{\theta}}_{k,h}^{(0)} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\mathrm{argmin}}\, \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{j=1}^{k-1} \big[\big\langle \boldsymbol{\phi}_{\overline{V}_{j,h+1}}(s_h^j, a_h^j, b_h^j), \boldsymbol{\theta}\big\rangle - \overline{V}_{j,h+1}(s_{h+1}^j)\big]^2/\overline{\sigma}_{j,h}^2, \tag{4.5}$$

where $\overline{\sigma}_{j,h}^2$ is an appropriate upper bound on the variance of the value function $[\mathbb{V}_h\overline{V}_{j,h+1}](s_h^j, a_h^j, b_h^j)$. In particular, we construct $\overline{\sigma}_{k,h}^2$ as follows

$$\overline{\sigma}_{k,h} = \sqrt{\max\{H^2/d, \mathbb{V}^{\mathrm{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) + \overline{E}_{k,h}\}}, \tag{4.6}$$

where $[\mathbb{V}_{k,h}^{\mathrm{est}}\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)$ is a scalar-valued empirical estimate for the variance of the value function $\overline{V}_{k,h+1}$ under the transition probability $\mathbb{P}_h(\cdot|s_k^h, a_k^h, b_k^h)$, and $\overline{E}_{k,h}$ is an offset term that is used to guarantee that $\overline{\sigma}_{k,h}^2$ upper bounds $[\mathbb{V}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)$ with high probability.

Weighted ridge regression (4.5) has a closed-form solution $\overline{\boldsymbol{\theta}}_{k,h}^{(0)} = \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1}\overline{\mathbf{b}}_{k,h}^{(0)}$, where the covariance matrix $\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}$ can be computed by recursion starting at $\overline{\boldsymbol{\Sigma}}_{1,h}^{(0)} = \lambda\mathbf{I}$:

$$\overline{\boldsymbol{\Sigma}}_{j+1,h}^{(0)} = \overline{\sigma}_{j,h}^{-2}\boldsymbol{\phi}_{\overline{V}_{j,h+1}}(s_h^j, a_h^j, b_h^j)\boldsymbol{\phi}_{\overline{V}_{j,h+1}}(s_h^j, a_h^j, b_h^j)^\top + \overline{\boldsymbol{\Sigma}}_{j,h}^{(0)}, \tag{4.7}$$

and the correlation vector $\overline{\mathbf{b}}_{k,h}^{(0)}$ can be computed by recursion starting at $\overline{\mathbf{b}}_{1,h}^{(0)} = \mathbf{0}$:

$$\overline{\mathbf{b}}_{j+1,h}^{(0)} = \overline{\mathbf{b}}_{j,h}^{(0)} + \overline{\sigma}_{j,h}^{-2}\boldsymbol{\phi}_{\overline{V}_{j,h+1}}(s_h^j, a_h^j, b_h^j)\overline{V}_{j,h+1}(s_{h+1}^j). \tag{4.8}$$

By using a Bernstein-type self-normalized concentration inequality for vector-valued martingales proposed in Zhou et al. (2021a), one can show then that, with high probability, $\boldsymbol{\theta}_h^*$ lies in the ellipsoid $\overline{\mathcal{C}}_{k,h}^{(0)}$ defined in (4.2), where $\beta_k^{(0)}$ is the confidence radius chosen later in Lemma 5.1.

Variance Estimator. It remains to set $\overline{\sigma}_{j,h}^2$. We need to specify how to calculate the empirical variance $[\mathbb{V}_{k,h}^{\text{est}}\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)$ and select $\overline{E}_{k,h}$ to guarantee $\overline{\sigma}_{j,h}^2$ upper bounds $[\mathbb{V}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)$ with high probability. Recall the definition of $[\mathbb{V}_h V](\cdot,\cdot,\cdot)$ as follows:

$$
\begin{aligned}
[\mathbb{V}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) &= [\mathbb{P}_h\overline{V}_{k,h+1}^2](s_h^k, a_h^k, b_h^k) - \big([\mathbb{P}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)\big)^2 \\
&= \underbrace{\big\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^*\big\rangle}_{I_1} - \underbrace{\big[\big\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^*\big\rangle\big]^2}_{I_2}.
\end{aligned}
$$

where the second equality holds due to the definition of linear mixture MGs. Notice that the expectation of $\overline{V}_{k,h+1}^2(s_{h+1}^k)$ over the next state, s_{h+1}^k, is a linear function of $\boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k)$. Therefore, we use $\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_{k,h}^{(1)}\rangle$ to estimate the term I_1 where $\boldsymbol{\theta}_{k,h}^{(1)}$ is the solution to the following ridge regression problem:

$$
\boldsymbol{\theta}_{k,h}^{(1)} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}}\, \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{j=1}^{k-1}\big[\big\langle \boldsymbol{\phi}_{\overline{V}_{j,h+1}^2}(s_h^j, a_h^j, b_h^j), \boldsymbol{\theta}\big\rangle - \overline{V}_{j,h+1}^2(s_{h+1}^j)\big]^2. \tag{4.9}
$$

For term I_2, we can use $\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(0)}\rangle$ to estimate it. Thus we have the following variance estimator,

$$
\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) \leftarrow \big[\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(1)}\rangle\big]_{[0,H^2]} - \big[\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(0)}\rangle\big]_{[-H,H]}^2, \tag{4.10}
$$

where the projection is used to control the range of our variance estimator. Lastly, we can compute $\boldsymbol{\theta}_{k,h}^{(1)}$ in a closed form $\overline{\boldsymbol{\theta}}_{k,h}^{(1)} = \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}\big]^{-1}\overline{\mathbf{b}}_{k,h}^{(1)}$, where the covariance matrix $\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}$ is updated recursively in the following way:

$$
\overline{\boldsymbol{\Sigma}}_{j+1,h}^{(1)} = \overline{\boldsymbol{\Sigma}}_{j,h}^{(1)} + \boldsymbol{\phi}_{\overline{V}_{j,h+1}^2}(s_h^j, a_h^j, b_h^j)\boldsymbol{\phi}_{\overline{V}_{j,h+1}^2}(s_h^j, a_h^j, b_h^j)^\top, \tag{4.11}
$$

and the correlation vector $\overline{\mathbf{b}}_{k,h}^{(1)}$ is updated in the following recursive form:

$$
\overline{\mathbf{b}}_{j+1,h}^{(1)} = \overline{\mathbf{b}}_{j,h}^{(1)} + \boldsymbol{\phi}_{\overline{V}_{j,h+1}^2}(s_h^j, a_h^j, b_h^j)\overline{V}_{j,h+1}^2(s_{h+1}^j). \tag{4.12}
$$

By the standard self-normalized concentration inequality for vector-valued martingales in Abbasi-Yadkori et al. (2011), we can show that, with high probability, $\overline{\sigma}_{j,h}^2$ upper bounds $[\mathbb{V}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)$ if we select $\overline{E}_{k,h}$ as follows

$$
\begin{aligned}
\overline{E}_{k,h} &= \min\big\{H^2, \beta_k^{(1)}\|\big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k)\|_2\big\} \\
&\quad + \min\big\{H^2, 2H\beta_k^{(2)}\|\big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\|_2\big\},
\end{aligned} \tag{4.13}
$$

where $\beta_k^{(1)}, \beta_k^{(2)}$ are constants chosen later in Lemma 5.1.

Remark 4.3 *Our Nash-UCRL is computational efficient for specific feature mapping $\boldsymbol{\phi}$, as Ayoub et al. (2020); Zhou et al. (2021a) suggested. For a special class of $\boldsymbol{\phi}$, where*

$$\boldsymbol{\phi}(s'|s,a,b) = \boldsymbol{\psi}(s') \odot \boldsymbol{\mu}(s,a,b),\ \boldsymbol{\psi}(\cdot): \mathcal{S} \to \mathbb{R}^d,\ \boldsymbol{\mu}(\cdot,\cdot,\cdot): \mathcal{S} \times \mathcal{A}_{max} \times \mathcal{A}_{min} \to \mathbb{R}^d,$$

$\odot$ is the componentwise product, Nash-UCRL can be implemented within poly$(d, |\mathcal{A}_{max}|, |\mathcal{A}_{min}|) \cdot KH$ time complexity with the access to some integration oracle $\mathcal{O}$. The details are deferred to Appendix A.

Difference between Nash-UCRL and previous algorithms Here we compare our Nash-UCRL with the OMNI-V proposed by Xie et al. (2020). First, Xie et al. (2020) studied the linear MGs while we study the linear mixture MGs. Second, due to the difference between the studied models, OMNI-V needs to maintain a covering set of the estimated Q functions (Eq. (5), Xie et al. 2020), which makes its space complexity exponential in d. In sharp contrast, our Nash-UCRL relies on the value targeted regression (Jia et al., 2020; Ayoub et al., 2020) and does not need to maintain such a cover set.

5. Main Results

In this section, we present the main theoretical results. We first show that under a specific parameter choice, our constructed confidence sets $\overline{\mathcal{C}}_{k,h}^{(0)}$ and $\underline{\mathcal{C}}_{k,h}^{(0)}$ include $\boldsymbol{\theta}_h^*$ with high probability, and the estimated variances $\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)$ and $\mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)$ deviate from the true variances by at most the offset terms $\overline{E}_{k,h}, \underline{E}_{k,h}$.

Lemma 5.1 *Setting $\beta_k^{(0)}$ in (4.2) and $\beta_k^{(1)}, \beta_k^{(2)}$ in (4.13) to*

$$\begin{aligned}
\beta_k^{(0)} &= 16\sqrt{d\log(1+k/\lambda)\log(4k^2H/\delta)} + 8\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B \\
\beta_k^{(1)} &= 16\sqrt{dH^4\log(1+KH^4/(d\lambda))\log(4k^2H/d\delta)} + 8H^2\log(4k^2H/\delta) + \sqrt{\lambda}B \\
\beta_k^{(2)} &= 16d\sqrt{\log(1+k/\lambda)\log(4k^2H/\delta)} + 8\sqrt{d}\log(4k^2H/\delta) + \sqrt{\lambda}B,
\end{aligned}$$

then with probability at least $1-3\delta$, we have $\boldsymbol{\theta}_h^* \in \overline{\mathcal{C}}_{k,h}^{(0)} \cap \underline{\mathcal{C}}_{k,h}^{(0)}$. In addition, we have

$$\begin{aligned}
|\mathbb{V}^{est}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) - \mathbb{V}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)| &\le \overline{E}_{k,h} \\
|\mathbb{V}^{est}\underline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) - \mathbb{V}\underline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)| &\le \underline{E}_{k,h}
\end{aligned}$$

Next, we present the regret of Nash-UCRL.

Theorem 5.2 *Setting $\lambda = 1/B^2$, $\epsilon = O(HT^{-1/2})$, then with probability at least $1-5\delta$, the regret of Algorithm 1 Regret$(M_{\boldsymbol{\theta}^*}, K)$ is bounded by*

$$\widetilde{\mathcal{O}}\big(\sqrt{d^2H^2 + dH^3}\sqrt{T} + d^2H^3 + d^3H^2\big),$$

where $T = KH$.

Theorem 5.2 suggest that when $d \ge H$ and $T \ge d^4H^2$, the regret of Nash-UCRL is bounded by $\widetilde{O}(dH\sqrt{T})$.

Remark 5.3 *Our Nash-UCRL also enjoys a finite sample complexity. By the standard online-to-batch conversion, we can show that Nash-UCRL is guaranteed to find an ϵ-approximate NE, i.e., (π, ν) satisfying $V_1^{*,\nu} - V_1^{\pi,*} \le \epsilon$, within $\widetilde{O}((d^2H^3 + dH^4)/\epsilon^2)$ episodes.*

Remark 5.4 *We can apply our algorithm to tabular MGs and our results can be reduced to the setting with $|S| = S, |\mathcal{A}_{\max}| = A, |\mathcal{A}_{\min}| = B$ by choosing $\phi(s'|s, a, b)$ as the one-hot representation of $\mathbb{P}(s'|s, a, b)$. It is easy to verify that (3.3) holds and $d = S^2AB$. Thus the regret bound given in Theorem 5.2 reduces to $\widetilde{O}(\sqrt{S^4H^2A^2B^2T})$, which does not match the lower bound of tabular MGs in Bai and Jin (2020). We would like to point out that by using some techniques specialized to the tabular setting, the regret bound of our algorithm for tabular MGs can be improved, which is beyond the scope of this work.*

Here, we present a lower bound for linear mixture MGs. It has been shown in Zhou et al. (2021a) that the regret lower bound for learning linear mixture MDPs is $\Omega(dH\sqrt{T})$, from which we can prove a lower bound for learning linear mixture MGs, since MDPs can be regarded as a special case of MGs with one dummy player, i.e., $\mathbb{P}_h(s'|s, a, b) = \mathbb{P}_h(s'|s, a)$ and $r_h(s, a, b) = r_h(s, a)$. Formally, we have the following lower bound:

Theorem 5.5 (Regret lower bound) *Let $B > 1$ and $K \ge \max\{(d-1)^2H/2, (d-1)/(32H(B-1))\}$, $d \ge 4$, $H \ge 3$. Then for any algorithm there exists an episodic, B-bounded linear mixture MG $M_{\boldsymbol{\theta}^*}$ such that the expected regret of first T rounds is lower bounded as follows:*

$$\mathbb{E}[\mathit{Regret}(M_{\boldsymbol{\theta}^*}, K)] \ge \Omega\big(dH\sqrt{T}\big),$$

where $T = KH$.

Remark 5.6 *When $d \ge H$ and $T \ge d^4H^2$, the regret of Nash-UCRL matches the lower bound up to logarithmic factors. Therefore, Nash-UCRL is nearly minimax optimal.*

Remark 5.7 *Based on a similar argument made in Zhou et al. (2021a), we can show that the same lower bound holds for the Markov games with linear structures. Recall that the best-known algorithm for learning MGs with linear structures is OMNI-VI (Xie et al., 2020), which has an $\widetilde{O}(\sqrt{d^3H^3T})$ regret. This suggests that there is still a gap that needs to be closed for learning MGs with linear structure. Please see the appendix for more details.*

Turn-based linear mixture MG can be regarded as a special case of linear mixture simultaneous-move MG. Therefore, we can still use Algorithm 1 to find the Nash equilibrium and then by Theorem 5.2, we can further show that the regret of our turn-based algorithm is also bounded by $\widetilde{O}(dH\sqrt{T})$. Notice that for the turn-based game, at each step only one player can take action. Thus, the ϵ-CCE routine in Line 8 of Algorithm 1 needs be replaced by two separate subroutines: taking π_h^k and ν_h^k as greedy policies w.r.t. $\overline{Q}_{k,h}$ and $\underline{Q}_{k,h}$. For completeness, we present the turn-based version of Algorithm 1 as Algorithm 3 in Appendix E.

6. Conclusions and Future Work

In this paper, we proposed the first provably optimal algorithm for learning two-player zero-sum Markov games with linear function approximation and without assuming access to the generative model. Specifically, we show that Nash-UCRL can provably achieve an $\widetilde{O}(dH\sqrt{T})$ regret, where d is the linear function dimension, H is the length of the game/episode, and T is the total number of steps in the Markov game. We also prove an $\widetilde{\Omega}(dH\sqrt{T})$ lower bound on the regret. Our upper bound matches the lower bound up to logarithmic factors, which suggests the optimality of our algorithm.

There are several important future directions. First, in the current linear mixture MG, the feature mapping encodes the information of both players. To reproduce the difference between $(A + B)$ and AB in the tabular setting, we may need to construct a separate feature mapping for each player (Bai et al., 2020). Second, while our algorithms can be extended to the decentralized setting, it is not clear if the minimax-optimal regret can still be obtained because of the adversarial policy. How to achieve a near-optimal decentralized algorithm is another important future work.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. Part of this work was done when ZC, DZ and QG participated the Theory of Reinforcement Learning program at the Simons Institute for the Theory of Computing in Fall 2020. ZC, DZ and QG are partially supported by the National Science Foundation IIS-2008981, CAREER Award 1906169, and AWS Machine Learning Research Award. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Robert J Aumann. Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society*, pages 1–18, 1987.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin F Yang. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

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

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, 33, 2020.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019.

Qi Cai, Zhuoran Yang, Chi Jin, and Zhaoran Wang. Provably efficient exploration in policy optimization. *arXiv preprint arXiv:1912.05830*, 2019.

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

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

Amy Greenwald, Keith Hall, and Roberto Serrano. Correlated q-learning. In *ICML*, volume 3, pages 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.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

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

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

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

Zeyu Jia, Lin Yang, Csaba Szepesvari, and Mengdi Wang. Model-based reinforcement learning with value-targeted regression. In *L4DC*, 2020.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4868–4878, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143, 2020.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

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

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

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

Michael L Littman and Csaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *ICML*, volume 96, pages 310–318. Citeseer, 1996.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

Aditya Modi, Nan Jiang, Ambuj Tewari, and Satinder Singh. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pages 2010–2020. PMLR, 2020.

Hervé Moulin and J-P Vial. Strategically zero-sum games: the class of games whose completely mixed equilibria cannot be improved upon. *International Journal of Game Theory*, 7(3-4):201–221, 1978.

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

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*, pages 1321–1329. PMLR, 2015.

Julien Pérolat, Bilal Piot, Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. Softened approximate policy iteration for markov games. In *International Conference on Machine Learning*, pages 1860–1868. PMLR, 2016a.

Julien Pérolat, Bilal Piot, Bruno Scherrer, and Olivier Pietquin. On the use of non-stationary strategies for solving two-player zero-sum markov games. In *Artificial Intelligence and Statistics*, pages 893–901. PMLR, 2016b.

Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning nash equilibrium for general-sum markov games from batch data. In *Artificial Intelligence and Statistics*, pages 232–241. PMLR, 2017.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

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

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for for solving discounted Markov decision process with a generative model. *arXiv preprint arXiv:1806.01492*, 2018.

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*, pages 2992–3002. PMLR, 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, 2016.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

Christopher John Cornish Hellaby Watkins. Learning from delayed rewards. 1989.

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

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. *arXiv preprint arXiv:2002.07066*, 2020.

Lin Yang and Mengdi Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004, 2019a.

Lin F Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. *arXiv preprint arXiv:1905.10389*, 2019b.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964, 2020a.

Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent Bellman error. *arXiv preprint arXiv:2003.00153*, 2020b.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Model-free reinforcement learning: from clipped pseudo-regret to sample complexity. *arXiv preprint arXiv:2006.03864*, 2020.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021b.

Appendix A. Computational Efficiency of Nash-UCRL

As Ayoub et al. (2020); Zhou et al. (2021a) suggested, the computational efficiency of Nash-UCRL will depend on the feature mapping $\boldsymbol{\phi}(s'|s, a, b)$. In this section we show that for a specific family of

$\boldsymbol{\phi}$ with the access to some integration oracle $\mathcal{O}$, Nash-UCRL can be implemented within polynomial computational complexity. We consider a special class of $\boldsymbol{\phi}$, where

$$\boldsymbol{\phi}(s'|s,a,b) = \boldsymbol{\psi}(s') \odot \boldsymbol{\mu}(s,a,b),\ \boldsymbol{\psi}(\cdot): \mathcal{S} \to \mathbb{R}^d,\ \boldsymbol{\mu}(\cdot,\cdot,\cdot): \mathcal{S} \times \mathcal{A}_{\text{max}} \times \mathcal{A}_{\text{min}} \to \mathbb{R}^d,$$

$\odot$ is the componentwise product. Meanwhile, we assume that there exists an oracle $\mathcal{O}$ such that for any function $V: \mathcal{S} \to \mathbb{R}$, the summation $\sum_s \boldsymbol{\psi}(s)V(s)$ can be evaluated by considering at most $p(d)$ number of states $s \in \mathcal{S}$. From now on we show how to compute each key step in Nash-UCRL. First, to compute $\overline{Q}_{k,h}$, note that $\overline{Q}_{k,h}$ can be parameterized by $\widehat{\boldsymbol{\theta}}_{k,h}^{(0)}$ and $\widehat{\boldsymbol{\Sigma}}_{k,h}^{(0)}$ as follows:

$$\begin{aligned}
&\overline{Q}_{k,h}(\cdot,\cdot,\cdot) \\
&= \Big[r_h(\cdot,\cdot,\cdot) + \langle \overline{\boldsymbol{\theta}}_{k,h}^{(0)}, \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(\cdot,\cdot,\cdot)\rangle + \beta_k^{(0)} \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(\cdot,\cdot,\cdot)\Big\|_2\Big]_{[-H,H]} \\
&= \Big[r_h(\cdot,\cdot,\cdot) + \langle \underbrace{\overline{\boldsymbol{\theta}}_{k,h}^{(0)} \odot \Big(\sum_{s'} \boldsymbol{\psi}(s')\overline{V}_{k,h+1}(s')\Big)}_{\widehat{\boldsymbol{\theta}}_{k,h}^{(0)}}, \boldsymbol{\mu}(\cdot,\cdot,\cdot)\rangle + \beta_k^{(0)} \Big\| \widehat{\boldsymbol{\Sigma}}_{k,h}^{(0)} \boldsymbol{\mu}(\cdot,\cdot,\cdot)\Big\|_2\Big]_{[-H,H]},
\end{aligned} \tag{A.1}$$

where the (i,j)-th entry of $\widehat{\boldsymbol{\Sigma}}_{k,h}^{(0)}$ is $\big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]_{i,j}^{-1/2}[\sum_{s'} \psi_j(s')\overline{V}_{k,h+1}(s')]$. Given $\widehat{\boldsymbol{\theta}}_{k,h+1}^{(0)}$ and $\widehat{\boldsymbol{\Sigma}}_{k,h+1}^{(0)}$, we need $O(d^2)$ to compute $\overline{Q}_{k,h}$, which is the same as computing $\underline{Q}_{k,h}$. Then, for each s, we need $O(|\mathcal{A}_{\text{max}}||\mathcal{A}_{\text{min}}|)$ complexity to compute $\mu_h^k(\cdot,\cdot|s)$ and $O(d^2|\mathcal{A}_{\text{max}}||\mathcal{A}_{\text{min}}|)$ complexity to compute $\overline{V}_{k,h}$ and $\underline{V}_{k,h}$. Finally, to obtain $\widehat{\boldsymbol{\theta}}_{k,h}^{(0)}$ and $\widehat{\boldsymbol{\Sigma}}_{k,h}^{(0)}$, we need to evaluate $\overline{V}_{k,h}$ over $p(d)$ states, which by (A.1), requires $O(p(d)d^2|\mathcal{A}_{\text{max}}||\mathcal{A}_{\text{min}}|)$ complexity in total. Therefore, we need poly$(d, |\mathcal{A}_{\text{max}}|, |\mathcal{A}_{\text{min}}|)$ complexity to compute one $\overline{Q}_{k,h}$, and we need poly$(d, |\mathcal{A}_{\text{max}}|, |\mathcal{A}_{\text{min}}|) \cdot KH$ complexity for implementing Nash-UCRL, given all $\overline{\boldsymbol{\theta}}_{k,h}^{(0)}$, $\underline{\boldsymbol{\theta}}_{k,h}^{(0)}$. The complexity of computing $\overline{\boldsymbol{\theta}}_{k,h}^{(0)}$ includes the complexity to solve the regression problem (4.5), (4.9) and to compute the variance estimator $\overline{\sigma}_{k,h}$, $\underline{\sigma}_{k,h}$, which again is at most poly$(d, |\mathcal{A}_{\text{max}}|, |\mathcal{A}_{\text{min}}|) \cdot KH$ according to previous analysis. Therefore, the total complexity of implementing Nash-UCRL is poly$(d, |\mathcal{A}_{\text{max}}|, |\mathcal{A}_{\text{min}}|) \cdot KH$.

Appendix B. Proof of Results in Section 5

We let $\mathbb{P}$ be the distribution over $(\mathcal{S} \times \mathcal{A}_{\max} \times \mathcal{A}_{\min})^{\mathbb{N}}$ induced by the episodic MG M, and further denote the sample space $\Omega = (\mathcal{S} \times \mathcal{A}_{\max} \times \mathcal{A}_{\min})^{\mathbb{N}}$. Thus, we work with the probability space given by the triplet $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is the product σ-algebra generated by the discrete σ-algebras underlying $\mathcal{S}$, $\mathcal{A}_{\max}$ and $\mathcal{A}_{\min}$.

For $1 \le k \le K$, $1 \le h \le H$, let $\mathcal{F}_{k,h}$ be the σ-algebra generated by the random variables representing the state-action-action pairs up to and including those that appear stage h of episode k. That is, $\mathcal{F}_{k,h}$ is generated by

$$\begin{gathered}
s_1^1, a_1^1, b_1^1, \ldots, s_h^1, a_h^1, b_h^1, \ldots, s_H^1, a_H^1, b_H^1\,, \\
s_1^2, a_1^2, b_1^2, \ldots, s_h^2, a_h^2, b_h^2, \ldots, s_H^2, a_H^2, b_H^2\,, \\
\vdots \\
s_1^k, a_1^k, b_1^k, \ldots, s_h^k, a_h^k, b_h^k\,.
\end{gathered}$$

B.1. Proof of Lemma 5.1

For simplicity we denote the following confident sets:

$$\overline{\mathcal{C}}_{k,h}^{(0)} = \left\{\boldsymbol{\theta} : \left\|[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{1/2}(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_{k,h}^{(0)})\right\|_2 \le \beta_k^{(0)}\right\}, \underline{\mathcal{C}}_{k,h}^{(0)} = \left\{\boldsymbol{\theta} : \left\|[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{1/2}(\boldsymbol{\theta} - \underline{\boldsymbol{\theta}}_{k,h}^{(0)})\right\|_2 \le \beta_k^{(0)}\right\},$$
$$\overline{\mathcal{C}}_{k,h}^{(1)} = \left\{\boldsymbol{\theta} : \left\|[\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}]^{1/2}(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_{k,h}^{(1)})\right\|_2 \le \beta_k^{(1)}\right\}, \underline{\mathcal{C}}_{k,h}^{(1)} = \left\{\boldsymbol{\theta} : \left\|[\underline{\boldsymbol{\Sigma}}_{k,h}^{(1)}]^{1/2}(\boldsymbol{\theta} - \underline{\boldsymbol{\theta}}_{k,h}^{(1)})\right\|_2 \le \beta_k^{(1)}\right\},$$
$$\overline{\mathcal{C}}_{k,h}^{(2)} = \left\{\boldsymbol{\theta} : \left\|[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{1/2}(\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}_{k,h}^{(0)})\right\|_2 \le \beta_k^{(2)}\right\}, \underline{\mathcal{C}}_{k,h}^{(2)} = \left\{\boldsymbol{\theta} : \left\|[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{1/2}(\boldsymbol{\theta} - \underline{\boldsymbol{\theta}}_{k,h}^{(0)})\right\|_2 \le \beta_k^{(2)}\right\}.$$

By the selection $\beta_k^{(0)} < \beta_k^{(2)}$ in Lemma 5.1, we have that $\overline{\mathcal{C}}_{k,h}^{(0)} \subset \overline{\mathcal{C}}_{k,h}^{(2)}$ and $\underline{\mathcal{C}}_{k,h}^{(0)} \subset \underline{\mathcal{C}}_{k,h}^{(2)}$. We first use standard self-normalized tail inequality to show that $\boldsymbol{\theta}_h^*$ is included in $\overline{\mathcal{C}}_{k,h}^{(1)} \cap \overline{\mathcal{C}}_{k,h}^{(2)}$ with high probability. Based on that we can further decrease $\beta_k^{(2)}$ to $\beta_k^{(1)}$ without significantly increasing the probability of the bad event when $\boldsymbol{\theta}_h^* \notin \overline{\mathcal{C}}_{k,h}^{(0)}$ or $\boldsymbol{\theta}_h^* \notin \underline{\mathcal{C}}_{k,h}^{(0)}$.

We start with the following Bernstein-type self-normalized concentration inequality.

Lemma B.1 (Theorem 2, Zhou et al. 2021a) *Let $\{\mathcal{G}_t\}_{t=1}^\infty$ be a filtration, $\{\mathbf{x}_t, \eta_t\}_{t\ge 1}$ a stochastic process so that $\mathbf{x}_t \in \mathbb{R}^d$ is $\mathcal{G}_t$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{G}_{t+1}$-measurable. Fix $R, L, \sigma, \lambda > 0$, $\boldsymbol{\mu}^* \in \mathbb{R}^d$. For $t \ge 1$ let $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t\rangle + \eta_t$ and suppose that $\eta_t, \mathbf{x}_t$ also satisfy*

$$|\eta_t| \le R,\ \mathbb{E}[\eta_t|\mathcal{G}_t] = 0,\ \mathbb{E}[\eta_t^2|\mathcal{G}_t] \le \sigma^2,\ \|\mathbf{x}_t\|_2 \le L.$$

Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ we have

$$\forall t > 0,\ \left\|\sum_{i=1}^t \mathbf{x}_i \eta_i\right\|_{\mathbf{Z}_t^{-1}} \le \beta_t,\ \|\boldsymbol{\mu}_t - \boldsymbol{\mu}^*\|_{\mathbf{Z}_t} \le \beta_t + \sqrt{\lambda}\|\boldsymbol{\mu}^*\|_2, \tag{B.1}$$

where for $t \ge 1$, $\boldsymbol{\mu}_t = \mathbf{Z}_t^{-1}\mathbf{b}_t$, $\mathbf{Z}_t = \lambda\mathbf{I} + \sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^\top$, $\mathbf{b}_t = \sum_{i=1}^t y_i\mathbf{x}_i$ and

$$\beta_t = 8\sigma\sqrt{d\log(1 + tL^2/(d\lambda))\log(4t^2/\delta)} + 4R\log(4t^2/\delta)\,.$$

Lemma B.2 *For every $1 \le k \le K$ and $1 \le h \le H$, we have*

$$\begin{aligned}
&|\mathbb{V}^{est}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) - \mathbb{V}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)| \\
&\le \min\left\{H^2, \left\|[\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k)\right\|_2\left\|[\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(1)} - \boldsymbol{\theta}_h^*)\right\|_2\right\} \\
&\quad + \min\left\{H^2, 2H\left\|[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k)\right\|_2\left\|[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*)\right\|_2\right\},
\end{aligned}$$

and

$$\begin{aligned}
&|\mathbb{V}^{est}\underline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) - \mathbb{V}\underline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)| \\
&\le \min\left\{H^2, \left\|[\underline{\boldsymbol{\Sigma}}_{k,h}^{(1)}]^{-1/2}\boldsymbol{\phi}_{\underline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k)\right\|_2\left\|[\underline{\boldsymbol{\Sigma}}_{k,h}^{(1)}]^{1/2}(\underline{\boldsymbol{\theta}}_{k,h}^{(1)} - \boldsymbol{\theta}_h^*)\right\|_2\right\} \\
&\quad + \min\left\{H^2, 2H\left\|[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2}\boldsymbol{\phi}_{\underline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k)\right\|_2\left\|[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{1/2}(\underline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*)\right\|_2\right\}.
\end{aligned}$$

Proof [Proof of Lemma 5.1] For simplicity, we only prove the results for the max-player. Fix $h \in [H]$.

We first show that with probability at least $1-\delta/(2H)$, $\Big\|\big[\overline{\boldsymbol{\Sigma}}^{(0)}_{k,h}\big]^{1/2}(\overline{\boldsymbol{\theta}}^{(0)}_{k,h}-\boldsymbol{\theta}^*_h)\Big\|_2 \le \beta^{(2)}_k$.To show this, we apply Lemma B.1. Let $\mathbf{x}_i = \overline{\sigma}^{-1}_{i,h}\boldsymbol{\phi}_{\overline{V}_{i,h+1}}(s^i_h,a^i_h,b^i_h)$ and $\eta_i = \overline{\sigma}^{-1}_{i,h}\overline{V}_{i,h+1}(s^i_{h+1}) - \overline{\sigma}^{-1}_{i,h}\langle\boldsymbol{\phi}_{\overline{V}_{i,h+1}}(s^i_h,a^i_h,b^i_h),\boldsymbol{\theta}^*_h\rangle$, $\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*_h$, $y_i = \langle\boldsymbol{\mu}^*,\mathbf{x}_i\rangle + \eta_i$, $\mathbf{Z}_i = \lambda\mathbf{I} + \sum_{i'=1}^i \mathbf{x}_{i'}\mathbf{x}_{i'}^\top$, $\mathbf{b}_i = \sum_{i'=1}^i \mathbf{x}_{i'}y_{i'}$ and $\boldsymbol{\mu}_i = \mathbf{Z}_i^{-1}\mathbf{b}_i$. Then it can be verified that $y_i = \overline{\sigma}^{-1}_{i,h}\overline{V}_{i,h+1}(s^i_{h+1})$ and $\boldsymbol{\mu}_i = \overline{\boldsymbol{\theta}}^{(0)}_{i+1,h}$. Moreover, we have that

$$\|\mathbf{x}_i\|_2 \le \overline{\sigma}^{-1}_{i,h}H \le \sqrt{d},\ \ |\eta_i| \le \overline{\sigma}^{-1}_{i,h}2H \le 2\sqrt{d},\ \ \mathbb{E}[\eta_i|\mathcal{G}_i]=0,\ \ \mathbb{E}[\eta_i^2|\mathcal{G}_i] \le 4d\,,$$

where we apply $\|\boldsymbol{\phi}_{\overline{V}_{i,h+1}}(\cdot,\cdot,\cdot)\|_2 \le H, \overline{V}_{i,h+1} \in [-H,H]$ and $\overline{\sigma}_{i,h} \ge H/\sqrt{d}$. Since we also have that $\boldsymbol{x}_i$ is $\mathcal{G}_i$ measurable and η_i is $\mathcal{G}_{i+1}$ measurable, by Lemma B.1, we obtain that with probability at least $1-\delta/(2H)$, for all $k \le K$, $\Big\|\big[\overline{\boldsymbol{\Sigma}}^{(0)}_{k,h}\big]^{1/2}(\overline{\boldsymbol{\theta}}^{(0)}_{k,h}-\boldsymbol{\theta}^*_h)\Big\|_2$ is bounded by

$$16d\sqrt{\log(1+k/\lambda)\log(8k^2H/\delta)} + 8\sqrt{d}\log(8k^2H/\delta) + \sqrt{\lambda}B = \beta^{(2)}_k, \tag{B.2}$$

implying that with probability at least $1-\delta/(2H)$, for any $k \le K$, $\boldsymbol{\theta}^*_h \in \overline{\mathcal{C}}^{(2)}_{k,h}$.

An argument, which is analogous to the one just used (except that now the range of the "noise" matches the range of "squared values" and is thus bounded by H^2, rather than being bounded by $\sqrt{d}$) gives that with probability at least $1-\delta/(2H)$, for any $k \le K$ we have $\Big\|\big[\overline{\boldsymbol{\Sigma}}^{(1)}_{k,h}\big]^{1/2}(\overline{\boldsymbol{\theta}}^{(1)}_{k,h}-\boldsymbol{\theta}^*_h)\Big\|_2$ bounded by

$$16\sqrt{dH^4\log(1+kH^4/(d\lambda))\log(8k^2H/\delta)} + 8H^2\log(8k^2H/\delta) + \sqrt{\lambda}B = \beta^{(1)}_k, \tag{B.3}$$

implying that with probability at least $1-\delta/(2H)$, for any $k \le K$, $\boldsymbol{\theta}^*_h \in \overline{\mathcal{C}}^{(1)}_{k,h}$.

We now show that $\boldsymbol{\theta}^*_h \in \overline{\mathcal{C}}^{(0)}_{k,h}$ with high probability. We again apply Lemma B.1. Let $\mathbf{x}_i = \overline{\sigma}^{-1}_{i,h}\boldsymbol{\phi}_{\overline{V}_{i,h+1}}(s^i_h,a^i_h,b^i_h)$ and

$$\eta_i = \overline{\sigma}^{-1}_{i,h}\,\mathbb{1}\{\boldsymbol{\theta}^*_h \in \overline{\mathcal{C}}^{(1)}_{i,h}\cap\overline{\mathcal{C}}^{(2)}_{i,h}\}\big[\overline{V}_{i,h+1}(s^i_{h+1}) - \langle\boldsymbol{\phi}_{\overline{V}_{i,h+1}}(s^i_h,a^i_h,b^i_h),\boldsymbol{\theta}^*_h\rangle\big],$$

$\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}^*_h$, $y_i = \langle\boldsymbol{\mu}^*,\mathbf{x}_i\rangle + \eta_i$, $\mathbf{Z}_i = \lambda\mathbf{I} + \sum_{i'=1}^i \mathbf{x}_{i'}\mathbf{x}_{i'}^\top$, $\mathbf{b}_i = \sum_{i'=1}^i \mathbf{x}_{i'}y_{i'}$ and $\boldsymbol{\mu}_i = \mathbf{Z}_i^{-1}\mathbf{b}_i$. Still we have that $\|\mathbf{x}_i\|_2 \le \overline{\sigma}^{-1}_{i,h}H \le \sqrt{d}$.Because $\mathbb{1}\{\boldsymbol{\theta}^*_h \in \overline{\mathcal{C}}^{(1)}_{i,h}\cap\overline{\mathcal{C}}^{(2)}_{i,h}\}$ is $\mathcal{G}_i$-measurable, we have $\mathbb{E}[\eta_i|\mathcal{G}_i]=0$. We also have $|\eta_i| \le \overline{\sigma}^{-1}_{i,h}2H \le 2\sqrt{d}$ since $|\overline{V}_{i,h+1}(\cdot)| \le H$ and $\overline{\sigma}_{i,h} \ge H/\sqrt{d}$. To get better bound $\beta^{(0)}_k$ rather than $\beta^{(2)}_k$ in (B.2), we need more careful computation of $\mathbb{E}[\eta_i^2|\mathcal{G}_i]$ as follows,

$$\begin{aligned}
\mathbb{E}[\eta_i^2|\mathcal{G}_i] &= \overline{\sigma}^{-2}_{i,h}\,\mathbb{1}\{\boldsymbol{\theta}^*_h \in \overline{\mathcal{C}}^{(1)}_{i,h}\cap\overline{\mathcal{C}}^{(2)}_{i,h}\}[\mathbb{V}_h\overline{V}_{i,h+1}](s^i_h,a^i_h,b^i_h)\\
&\le \overline{\sigma}^{-2}_{i,h}\,\mathbb{1}\{\boldsymbol{\theta}^*_h \in \overline{\mathcal{C}}^{(1)}_{i,h}\cap\overline{\mathcal{C}}^{(2)}_{i,h}\}\bigg[[\mathbb{V}^{\text{est}}_{i,h}\overline{V}_{i,h+1}](s^i_h,a^i_h,b^i_h)\\
&\quad + \min\Big\{H^2, \Big\|\big[\overline{\boldsymbol{\Sigma}}^{(1)}_{i,h}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}^2_{i,h+1}}(s^i_h,a^i_h,b^i_h)\Big\|_2\Big\|\big[\overline{\boldsymbol{\Sigma}}^{(1)}_{i,h}\big]^{1/2}(\overline{\boldsymbol{\theta}}^{(1)}_{i,h}-\boldsymbol{\theta}^*_h)\Big\|_2\Big\}
\end{aligned}$$

$$+ \min\Big\{H^2, 2H\Big\|[\overline{\boldsymbol{\Sigma}}_{i,h}^{(0)}]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{i,h+1}}(s_h^i, a_h^i, b_h^i)\Big\|_2\Big\|[\overline{\boldsymbol{\Sigma}}_{i,h}^{(0)}]^{1/2}(\overline{\boldsymbol{\theta}}_{i,h}^{(0)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\}\Big]$$

$$\leq \overline{\sigma}_{i,h}^{-2}\Big[[\mathbb{V}_{i,h}^{\text{est}}\overline{V}_{i,h+1}](s_h^i, a_h^i, b_h^i) + \min\Big\{H^2, \beta_i^{(1)}\Big\|[\boldsymbol{\Sigma}_{i,h}^{(1)}]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{i,h+1}^2}(s_h^i, a_h^i, b_h^i)\Big\|_2\Big\}$$

$$+ \min\Big\{H^2, 2H\beta_i^{(2)}\Big\|[\boldsymbol{\Sigma}_{i,h}^{(0)}]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{i,h+1}}(s_h^i, a_h^i, b_h^i)\Big\|_2\Big\}\Big]$$

$$= 1,$$

where the first inequality holds due to Lemma B.2, the second inequality holds due to the indicator function, the last equality holds due to the definition of $\overline{\sigma}_{i,h}$. Then, by Lemma B.1, with probability at least $1 - \delta/(2H)$, $\forall k \leq K$,

$$\|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_i} \leq 16\sqrt{d\log(1 + k/\lambda)\log(8k^2H/\delta)} + 8\sqrt{d}\log(8k^2H/\delta) + \sqrt{\lambda}B = \beta_k^{(0)}, \quad \text{(B.4)}$$

where the equality uses the definition of $\beta_k^{(0)}$. Let $\mathcal{E}'$ be the event when $\boldsymbol{\theta}_h^* \in \cap_{k\leq K}\overline{\mathcal{C}}_{k,h}^{(1)} \cap \overline{\mathcal{C}}_{k,h}^{(2)}$ and (B.4) hold. By the union bound, $\mathbb{P}(\mathcal{E}') \geq 1 - 3\delta/(2H)$.

We now show that $\boldsymbol{\theta}_h^* \in \overline{\mathcal{C}}_{k,h}^{(0)}$ holds on $\mathcal{E}'$. For this note that on $\mathcal{E}'$, for any $k \leq K$, $\boldsymbol{\mu}_k = \overline{\boldsymbol{\theta}}_{k+1,h}^{(0)}$ and for any $i \leq K$,

$$y_i = \overline{\sigma}_{i,h}^{-1}\big(\langle\boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\overline{V}_{i,h+1}}(s_h^i, a_h^i, b_h^i)\rangle + \mathbb{1}\{\boldsymbol{\theta}_h^* \in \overline{\mathcal{C}}_{i,h}^{(1)} \cap \overline{\mathcal{C}}_{i,h}^{(2)}\}\big[\overline{V}_{i,h+1}(s_{h+1}^i)$$
$$- \langle\boldsymbol{\phi}_{\overline{V}_{i,h+1}}(s_h^i, a_h^i, b_h^i), \boldsymbol{\theta}^*\rangle\big]\big)$$
$$= \overline{\sigma}_{i,h}^{-1}\overline{V}_{i,h+1}(s_{h+1}^i),$$

which implies the claim. Therefore, by the definition of $\mathcal{C}_{k,h}^{(0)}$, we get that on $\mathcal{E}'$, $\boldsymbol{\theta}_h^* \in \cap_{k\leq K}\overline{\mathcal{C}}_{k,h}^{(0)} \cap \overline{\mathcal{C}}_{k,h}^{(1)}$. Moreover, $\mathbb{P}(\mathcal{E}') \geq 1 - 3\delta/(2H)$. Finally, taking union bound over h shows that with probability at least $1 - 3\delta/2$, for all $h \in [H]$,

$$\boldsymbol{\theta}_h^* \in \cap_{k\leq K}\overline{\mathcal{C}}_{k,h}^{(1)} \cap \overline{\mathcal{C}}_{k,h}^{(2)} \quad \text{(B.5)}$$

To finish our proof, it is thus sufficient to show that on the event when (B.5) holds, it also holds that

$$\big|[\mathbb{V}_{k,h}^{\text{est}}\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{V}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)\big| \leq \overline{E}_{k,h}.$$

However, by the definition of $\overline{E}_{k,h}$, this is immediate from substituting (B.2), (B.3) into Lemma B.2. ■

B.2. Proof of Theorem 5.2

Let the event $\mathcal{E}$ denote the event when the conclusion of Lemma 5.1 holds. Then Lemma 5.1 suggests that $\mathbb{P}(\mathcal{E}) \geq 1 - 3\delta$. We introduce another two events in the following lemma.

Lemma B.3 *Denote events $\mathcal{E}_1$ and $\mathcal{E}_2$ as follows*

$$\mathcal{E}_1 = \Big\{\forall h' \in [H], \sum_{k=1}^{K}\sum_{h=h'}^{H}\Big[[\mathbb{P}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h\underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)$$

$$
\begin{aligned}
&\quad - \overline{V}_{k,h+1}(s_{h+1}^k) + \underline{V}_{k,h+1}(s_{h+1}^k)\Big] \le 8H\sqrt{2T\log(H/\delta)}\Big\} \\
\mathcal{E}_2 &= \Big\{\sum_{k=1}^K\sum_{h=1}^H \mathbb{V}_h V_{h+1}^{\mu^k}(s_h^k, a_h^k, b_h^k) \le 3(HT + H^3\log(1/\delta))\Big\}.
\end{aligned}
$$

Then we have $\mathbb{P}(\mathcal{E}_1) \ge 1-\delta$ *and* $\mathbb{P}(\mathcal{E}_2) \ge 1-\delta$.

We now present three lemmas based on $\mathcal{E}, \mathcal{E}_1, \mathcal{E}_2$. The following lemma shows that $\overline{Q}$ and $\overline{V}$ provide the good UCB for the best response of the max-player and $\underline{Q}$ and $\underline{V}$ provide the good LCB for the best response of the min-player.

Lemma B.4 *Suppose the event $\mathcal{E}$ hold, then we have for any s,a,b,k,h following inequalities hold,*

$$
\underline{Q}_{k,h}(s,a,b) - (H-h+1)\epsilon \le Q_h^{\pi^k,*}(s,a,b) \le Q_h^{*,\nu^k}(s,a,b) \le \overline{Q}_{k,h}(s,a,b) + (H-h+1)\epsilon,
$$

and

$$
\underline{V}_{k,h}(s) - (H-h+2)\epsilon \le V_h^{\pi^k,*}(s) \le V_h^{*,\nu^k}(s) \le \overline{V}_{k,h}(s) + (H-h+2)\epsilon.
$$

Lemma B.5 *Suppose the events $\mathcal{E} \cap \mathcal{E}_1$ hold, then we have*

$$
\begin{aligned}
\sum_{k=1}^K [\overline{V}_{k,1}(s_{k,1}) - \underline{V}_{k,1}(s_{k,1})] \le 4\beta_K^{(0)}&\sqrt{\sum_{k=1}^K\sum_{h=1}^H \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)} \\
&+ 8H\sqrt{2T\log(H/\delta)},
\end{aligned}
$$

$$
\begin{aligned}
\sum_{k=1}^K\sum_{h=1}^H \mathbb{P}_h[\overline{V}_{k,h+1} - \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) \le 4\beta_K^{(0)}&\sqrt{\sum_{k=1}^K\sum_{h=1}^H \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2}\sqrt{2H^3d\log(1+K/\lambda)} \\
&+ 8H^2\sqrt{2T\log(H/\delta)},
\end{aligned}
$$

Lemma B.6 *Suppose the events $\mathcal{E} \cap \mathcal{E}_2$ hold, then we have*

$$
\begin{aligned}
\sum_{k=1}^K\sum_{h=1}^H \overline{\sigma}_{k,h}^2 &\le H^2T/d + 3(HT + H^3\log(1/\delta)) + 4H\sum_{k=1}^K\sum_{h=1}^H \mathbb{P}_h[\overline{V}_{k,h+1} - V_{h+1}^{\mu^k}] \\
&\quad + 2\beta_K^{(2)}\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 7\beta_K^{(1)}H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)} \\
\sum_{k=1}^K\sum_{h=1}^H \underline{\sigma}_{k,h}^2 &\le H^2T/d + 3(HT + H^3\log(1/\delta)) + 4H\sum_{k=1}^K\sum_{h=1}^H \mathbb{P}_h[V_{h+1}^{\mu^k} - \underline{V}_{k,h+1}] \\
&\quad + 2\beta_K^{(2)}\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 7\beta_K^{(1)}H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)}
\end{aligned}
$$

With all these lemmas, we can now give the proof of Theorem 5.2.

Proof [Proof of Theorem 5.2] By definition of Regret we have that

$$
\begin{aligned}
\text{Regret}(K) &= \sum_{k=1}^{K} V_1^{*,\nu^k}(s_1^k) - \sum_{k=1}^{K} V_1^{\pi^k,*}(s_1^k) \\
&\le \sum_{k=1}^{K} \overline{V}_{k,1}(s_{k,1}) - \sum_{k=1}^{K} \underline{V}_{k,1}(s_{k,1}) + 4KH\epsilon \\
&\le 4\beta_K^{(0)} \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2} \sqrt{2Hd\log(1+K/\lambda)} + 8H\sqrt{2T\log(H/\delta)} + 4KH\epsilon \\
&= \widetilde{O}\bigg(d\sqrt{H}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2} + H\sqrt{T}\bigg), \qquad \text{(B.6)}
\end{aligned}
$$

where the first inequality is by Lemma B.4, the second inequality is by the bound of accumulated difference between the UCB and LCB in Lemma B.5, the last inequality is due to $\epsilon = O(H/\sqrt{T})$, $\lambda = 1/B^2$ and the choice of $\beta_K^{(0)} = \widetilde{O}(\sqrt{d})$ in Lemma 5.1.

Now we bound $\sum_{k=1}^{K}\sum_{h=1}^{H} \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2$,

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H} \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2 \qquad \text{(B.7)} \\
&\le 2H^2T/d + 6(HT + H^3\log(1/\delta)) + 4H\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{P}_h[\overline{V}_{k,h+1} - \underline{V}_{k,h+1}] \\
&\quad + 4\beta_K^{(2)}\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 14\beta_K^{(1)}H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)} \\
&\le 2H^2T/d + 6(HT + H^3\log(1/\delta)) \\
&\quad + 4H\bigg(4\beta_K^{(0)}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2}\sqrt{2H^3d\log(1+K/\lambda)} + 8H^2\sqrt{2T\log(H/\delta)}\bigg) \\
&\quad + 4\beta_K^{(2)}\sqrt{T}\sqrt{2dH\log(1+KH^4/(d\lambda))} + 14\beta_K^{(1)}H^2\sqrt{T}\sqrt{2dH\log(1+K/\lambda)} \\
&= \widetilde{O}\bigg(\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H} \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2}\sqrt{d^2H^5} + H^2T/d + TH + \sqrt{T}d^{1.5}H^{2.5} + H^3\sqrt{T}\bigg) \qquad \text{(B.8)}
\end{aligned}
$$

where the first inequality is by Lemma B.6, the second inequality is by Lemma B.5 and the last inequality is due to the choice of $\beta_K^{(0)} = \widetilde{O}(\sqrt{d})$ in Lemma 5.1, $\lambda = 1/B^2$,

$$
\begin{aligned}
\beta_K^{(1)} &= 16\sqrt{dH^4\log(1+kH^4/d\lambda)\log(8k^2H/\delta)} + 8H^2\log(8k^2H/\delta) + \sqrt{\lambda}B = \widetilde{O}(dH^2) \\
\beta_K^{(2)} &= 16d\sqrt{\log(1+k/\lambda)\log(8k^2H/\delta)} + 8\sqrt{d}\log(8k^2H/\delta) + \sqrt{\lambda}B = \widetilde{O}(d).
\end{aligned}
$$

Therefore by the fact that $x \le a\sqrt{x} + b \Rightarrow x \le 2a^2 + b$, (B.7) suggests that

$$
\sum_{k=1}^{K}\sum_{h=1}^{H} \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2 = \widetilde{O}(d^2H^5 + H^2T/d + TH + \sqrt{T}d^{1.5}H^{2.5} + H^3\sqrt{T})
$$

$$= \widetilde{O}(d^2H^5 + d^4H^3 + TH + H^2T/d), \tag{B.9}$$

where the inequality holds by $\sqrt{T}d^{1.5}H^{2.5} \leq (TH^2/4d + d^4H^3)/2$ and $H^3\sqrt{T} \leq (d^2H^5 + H^2T/d)/2$. Plugging (B.9) into (B.6) we have

$$\text{Regret}(M_{\boldsymbol{\theta}^*}, K) = \widetilde{\mathcal{O}}(\sqrt{d^2H^2 + dH^3}\sqrt{T} + d^2H^3 + d^3H^2),$$

which finishes the proof. ■

B.3. Proof of Theorem 5.5

Proof [Proof of Theorem 5.5] For any algorithm, we need to construct a hard-to-learn episodic,B-bounded linear mixture Markov game. We make the min-player dummy: the action of the min-player won't affect the transition ability or reward function. So there exists $\widetilde{\mathbb{P}}_h(\cdot|\cdot,\cdot)$ and $\widetilde{r}_h(\cdot,\cdot)$ such that for any state-action-action-state pair s', a, b, s we have that $\mathbb{P}_h(s'|s,a,b) = \widetilde{\mathbb{P}}_h(s'|s,a)$ and $r_h(s,a,b) = \widetilde{r}_h(s,a)$. Thus we can get a new MDP $\widetilde{M}(\mathcal{S}, \mathcal{A}_{\max}, H, \{\widetilde{r}_h\}, \{\widetilde{\mathbb{P}}_h\})$. We further have $V_h^{\pi,*}(s) = \widetilde{V}_h^{\pi}(s)$ and $V_h^{*,\nu}(s) = \widetilde{V}_h^{*}(s)$. The regret of two-player game can be reduced to the standard regret for single agent reinforcement learning setting. In particular,

$$\begin{aligned}
\text{Regret}(M_{\boldsymbol{\theta}^*}, K) &= \sum_{k=1}^{K} V_1^{*,\nu^k}(s_1^k) - \sum_{k=1}^{K} V_1^{\pi^k,*}(s_1^k) \\
&= \sum_{k=1}^{K} \widetilde{V}_1^{*}(s_1^k) - \sum_{k=1}^{K} \widetilde{V}_1^{\pi^k}(s_1^k).
\end{aligned}$$

Notice that $\widetilde{r}_h \in [-1, 1]$ rather than $[0, 1]$, we can shift the reward by $(1 + \widetilde{r}_h)/2$ to make it standard if necessary. Now recall the Theorem 5.6 in Zhou et al. (2021a)], there exists an episodic, B-bounded linear mixture MDP $\widetilde{M}(\mathcal{S}, \mathcal{A}_{\max}, H, \{\widetilde{r}_h\}, \{\widetilde{\mathbb{P}}_h\})$ with feature $\widetilde{\boldsymbol{\phi}}(\cdot,\cdot)$ parameterized by $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H)$ such that the expected regret is lower bounded as follows:

$$\mathbb{E}_{\boldsymbol{\Theta}}\text{Regret}\big(\widetilde{M}_{\boldsymbol{\Theta}}, K\big) \geq \Omega\big(dH\sqrt{T}\big),$$

where $T = KH$ and $\mathbb{E}_{\boldsymbol{\Theta}}$ denotes the expectation over the probability distribution generated by the interconnection of the algorithm and the MDP.

Now we only need to extend the MDP feature $\widetilde{\boldsymbol{\phi}}(\cdot|\cdot,\cdot)$ to the Markov game feature $\boldsymbol{\phi}(\cdot|\cdot,\cdot,\cdot)$. In particular, we set

$$\boldsymbol{\phi}(s'|s,a,b) = \widetilde{\boldsymbol{\phi}}(s'|s,a), \forall s' \in \mathcal{S}, s \in \mathcal{S}, a \in \mathcal{A}_{\max}, b \in \mathcal{A}_{\min},$$

then we know that $\boldsymbol{\phi}(\cdot|\cdot,\cdot,\cdot)$ satisfies (3.3) because by the definition of linear mixture MDP in Zhou et al. (2021a), we know that $\widetilde{\boldsymbol{\phi}}(\cdot|\cdot,\cdot)$ satisfies for any bounded function $V : \mathcal{S} \to [0,1]$,

$$\|\widetilde{\boldsymbol{\phi}}_V(s,a)\|_2 \leq 1,$$

where $\widetilde{\boldsymbol{\phi}}_V(s,a) = \sum_{s' \in \mathcal{S}} \widetilde{\boldsymbol{\phi}}(s'|s,a)V(s')$. ■

Appendix C. Proof of Lemmas in Appendix B

C.1. Proof of Lemma B.2

Proof [Proof of Lemma B.2] For simplicity, we only prove the results for the max-player.

By the triangle inequality we have that

$$\begin{aligned}
&|\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) - \mathbb{V}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)| \\
&\leq \underbrace{\Big|\langle \boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^*\rangle - \big[\langle \boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(1)}\rangle\big]_{[0,H^2]}\Big|}_{I_1} \\
&\quad + \underbrace{\Big|(\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^*\rangle)^2 - \big[\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(0)}\rangle\big]^2_{[-H,H]}\Big|}_{I_2}. \qquad \text{(C.1)}
\end{aligned}$$

We first bound I_1. Because $\langle \boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^*\rangle \in [0, H^2]$, we have that

$$\begin{aligned}
I_1 &\leq \Big|\langle \boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^*\rangle - \langle \boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(1)}\rangle\Big| \\
&\leq \Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\Big\|_2\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(1)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(1)} - \boldsymbol{\theta}_h^*)\Big\|_2,
\end{aligned}$$

where the first inequality is by the property of projection, the second inequality holds due to Cauchy-Schwarz. We also have that $I_1 \leq H^2$ since both terms in I_1 belongs to the interval $[0, H^2]$, so we have that

$$I_1 \leq \min\Big\{H^2, \Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\Big\|_2\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(1)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(1)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\}, \qquad \text{(C.2)}$$

For the term I_2,

$$\begin{aligned}
I_2 &= \Big|\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^*\rangle - \big[\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(0)}\rangle\big]_{[-H,H]}\Big| \\
&\qquad \cdot \Big|\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^*\rangle + \big[\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(0)}\rangle\big]_{[-H,H]}\Big| \\
&\leq 2H\Big|\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \boldsymbol{\theta}_h^*\rangle - \langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(0)}\rangle\Big| \\
&\leq 2H\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\Big\|_2\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*)\Big\|_2,
\end{aligned}$$

where the first inequality holds since both terms in this line lies in $[-H, H]$, the second inequality holds since the Cauchy-Schwarz inequality. We also have that $I_2 \leq H^2$, so we have that

$$I_2 \leq \min\Big\{H^2, 2H\Big\|\boldsymbol{\Sigma}_{k,h}^{(0)-1/2}\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\Big\|_2\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\}. \qquad \text{(C.3)}$$

Plugging (C.3) and (C.2) into (C.1) gets

$$\begin{aligned}
&|\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) - \mathbb{V}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)| \\
&\leq \min\Big\{H^2, \Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\Big\|_2\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(1)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(1)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\} \\
&\quad + \min\Big\{H^2, 2H\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\Big\|_2\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\}.
\end{aligned}$$

■

C.2. Proof of Lemma B.3

We first present the Azuma-Hoeffding inequality:

Lemma C.1 (Azuma-Hoeffding inequality, Azuma 1967) *Let $M > 0$ be a constant. Let $\{x_i\}_{i=1}^n$ be a martingale difference sequence with respect to a filtration $\{\mathcal{G}_i\}_i$ ($\mathbb{E}[x_i|\mathcal{G}_i] = 0$ a.s. and x_i is $\mathcal{G}_{i+1}$-measurable) such that for all $i \in [n]$, $|x_i| \le M$ holds almost surely. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, we have*

$$\sum_{i=1}^{n} x_i \le M\sqrt{2n\log(1/\delta)}.$$

Proof [Proof of Lemma B.3] To prove $\mathbb{P}(\mathcal{E}_1) \ge 1 - \delta$, we apply the Azuma-Hoeffding inequality (Lemma C.1). Fix $h' \in H$, set $x_{k,h} = [\mathbb{P}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h\underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\overline{V}_{k,h+1}(s_{h+1}^k) - \underline{V}_{k,h+1}(s_{h+1}^k)]$. $x_{1,h'}, \dots, x_{1,H}, x_{2,h'}, \dots, x_{2,H}, \dots, x_{K,h'}, \dots, x_{K,H}$ forms a martingale difference sequence of which the absolute value is bounded by 8H and length no greater than $T = KH$. Thus with probability at least $1 - \delta/H$, we have

$$\sum_{k=1}^{K}\sum_{h=h'}^{H}\Big[[\mathbb{P}_h\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h\underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - \overline{V}_{k,h+1}(s_{h+1}^k) + \underline{V}_{k,h+1}(s_{h+1}^k)\Big]$$
$$\le 8H\sqrt{2T\log(H/\delta)}.$$

Take union bound for $h' \in [H]$, we get $\mathbb{P}(\mathcal{E}_1) \ge 1 - \delta$.

$\mathbb{P}(\mathcal{E}_2) \ge 1 - \delta$ holds due to the Lemma C.5 in Jin et al. (2018) or Lemma 8 in Azar et al. (2017).

■

C.3. Proof of Lemma B.4

Following Lemma directly from the definition of ϵ-CCE,

Lemma C.2 *For each (k, h, s), $\mu_h^k(\cdot, \cdot|s), \pi_h^k(\cdot|s), \nu_h^k(\cdot|s)$ satisfy that*

$$\mathbb{E}_{(a,b)\sim\mu_h^k(\cdot,\cdot|s)}\big[\overline{Q}_{k,h}(s, a, b)\big] \ge \mathbb{E}_{b\sim\nu_h^k(s)}\big[\overline{Q}_{k,h}(s, a', b)\big] - \epsilon, \forall a' \in \mathcal{A}_{\max}$$
$$\mathbb{E}_{(a,b)\sim\mu_h^k(\cdot,\cdot|s)}\big[\underline{Q}_{k,h}(s, a, b)\big] \le \mathbb{E}_{a\sim\pi_h^k(s)}\big[\underline{Q}_{k,h}(s, a, b')\big] - \epsilon, \forall b' \in \mathcal{A}_{\min}$$

Proof [Proof of Lemma B.4] For simplicity, we only prove the following UCB by induction,

$$Q_h^{*,\nu^k}(s, a, b) \le \overline{Q}_{k,h}(s, a, b) + (H - h + 1)\epsilon, V_h^{*,\nu^k}(s) \le \overline{V}_{k,h}(s) + (H - h + 2)\epsilon. \tag{C.4}$$

The base case $h = H + 1$ holds trivially since the terminal cost is zero. Now we assume that the bounds (C.4) holds for step $h + 1$. That is,

$$Q_{h+1}^{*,\nu^k}(s, a, b) \le \overline{Q}_{k,h+1}(s, a, b) + (H - h)\epsilon, V_{h+1}^{*,\nu^k}(s) \le \overline{V}_{k,h+1}(s) + (H - h + 1)\epsilon. \tag{C.5}$$

If $\overline{Q}_{k,h}(s, a, b) \ge H$, then it is obvious to have $Q_h^{*,\nu^k}(s, a, b) \le \overline{Q}_{k,h}(s, a, b) + (H - h)\epsilon$, otherwise we have that

$$\overline{Q}_{k,h}(s, a, b) - Q_h^{*,\nu^k}(s, a, b)$$

$$
\begin{aligned}
&= \langle \overline{\boldsymbol{\theta}}_{k,h}^{(0)}, \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \rangle + \beta_k^{(0)} \Big\| \big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 - \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \rangle \\
&\quad + \mathbb{P}_h \overline{V}_{k,h+1}(s,a,b) - \mathbb{P}_h V_{h+1}^{*,\nu^k}(s) \\
&\geq \beta_k^{(0)} \Big\| \big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 - \Big\| \big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{1/2} (\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*) \Big\|_2 \Big\| \big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 \\
&\quad + \mathbb{P}_h \overline{V}_{k,h+1}(s) - \mathbb{P}_h V_{h+1}^{*,\nu^k}(s) \\
&\geq \mathbb{P}_h \overline{V}_{k,h+1}(s) - \mathbb{P}_h V_{h+1}^{*,\nu^k}(s) \\
&\geq -(H-h+1)\epsilon,
\end{aligned} \tag{C.6}
$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds since the assumption that $\theta_h^* \in \overline{\mathcal{C}}_{k,h}^{(0)}$ on event $\mathcal{E}$, the third inequality holds by the induction assumption. Finally, let $\mathrm{br}(\nu_h^k(\cdot|s))$ denote the best response to $\nu_h^k(\cdot|s)$ with respect to $Q_h^{*,\nu^k}(s,\cdot,\cdot)$ such that

$$
\mathrm{br}(\nu_h^k(\cdot|s)) = \operatorname*{argmax}_{\sigma \in \Delta_{\mathcal{A}_{\max}}} \mathbb{E}_{a\sim\sigma, b\sim\nu^k(\cdot|s)} Q_h^{*,\nu^k}(s,a,b).
$$

Then we have that

$$
\begin{aligned}
\overline{V}_{k,h}(s) &= \mathbb{E}_{(a,b)\sim\mu_h^k(\cdot,\cdot|s)} \big[\overline{Q}_{k,h}(s,a,b)\big] \\
&\geq \mathbb{E}_{a'\sim \mathrm{br}(v_h^k(\cdot|s)), b\sim\nu_h^k(\cdot|s)} \big[\overline{Q}_{k,h}(s,a',b)\big] - \epsilon \\
&\geq \mathbb{E}_{a'\sim \mathrm{br}(v_h^k(\cdot|s)), b\sim\nu_h^k(\cdot|s)} \big[Q_h^{*,\nu^k}(s,a',b)\big] - (H-h+2)\epsilon \\
&= V_h^{*,\nu^k}(s) - (H-h+2)\epsilon,
\end{aligned}
$$

where the the first equality is by the property of ϵ-CCE in Lemma C.2, the second inequality is by (C.6), the last inequality is due to the Bellman equation.Therefore, our proof ends. ■

C.4. Proof of Lemma B.5

Proof [Proof of Lemma B.5]

$$
\begin{aligned}
&\overline{V}_{k,h}(s_h^k) - \underline{V}_{k,h}(s_h^k) \\
&= \langle \overline{\boldsymbol{\theta}}_{k,h}^{(0)}, \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \rangle + \beta_k^{(0)} \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 - \langle \underline{\boldsymbol{\theta}}_{k,h}^{(0)}, \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \rangle + \beta_k^{(0)} \Big\| \big[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \Big\|_2 \\
&= \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \rangle + \langle \overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \rangle + \beta_k^{(0)} \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 \\
&\quad - \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \rangle - \langle \underline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \rangle + \beta_k^{(0)} \Big\| \big[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \Big\|_2 \\
&\leq \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \rangle + \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{1/2} (\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*) \Big\|_2 \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 + \beta_k^{(0)} \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 \\
&\quad - \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \rangle + \Big\| \big[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{1/2} (\underline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*) \Big\|_2 \Big\| \big[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \Big\|_2 \\
&\quad + \beta_k^{(0)} \Big\| \big[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \Big\|_2 \\
&\leq [\mathbb{P}_h \overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) + 2\beta_k^{(0)} \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2
\end{aligned}
$$

$$- [\mathbb{P}_h \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) + 2\beta_k^{(0)} \Big\| [\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \Big\|_2,$$

where the first equation is by the definition of $\overline{V}_{k,h}(s_h^k), \underline{V}_{k,h}(s_h^k)$ and the second inequality is due to Cauchy-Schwarz inequality, the last inequality is by $\boldsymbol{\theta}_h^* \in \overline{\mathcal{C}}_{k,h}^{(0)} \cap \underline{\mathcal{C}}_{k,h}^{(0)}$ on the event $\mathcal{E}$.

Meanwhile, since $\overline{V}_{k,h}(s_h^k) - \underline{V}_{k,h}(s_h^k) \le 2H$, we have that

$$\begin{aligned}
\overline{V}_{k,h}(s_h^k) - \underline{V}_{k,h}(s_h^k) &\le \min\Big\{2H, [\mathbb{P}_h \overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) + 2\beta_k^{(0)} \Big\| [\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 \\
&\qquad - [\mathbb{P}_h \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) + 2\beta_k^{(0)} \Big\| [\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \Big\|_2 \Big\} \\
&\le \min\Big\{4H, 2\beta_k^{(0)} \Big\| [\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 + 2\beta_k^{(0)} \Big\| [\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \Big\|_2 \Big\} \\
&\qquad + [\mathbb{P}_h \overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) \\
&\le \min\Big\{4H, 2\beta_k^{(0)} \Big\| [\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 \Big\} \\
&\qquad + \min\Big\{4H, 2\beta_k^{(0)} \Big\| [\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} \Big\|_2 \Big\} \\
&\qquad + [\mathbb{P}_h \overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) \\
&\le 2\beta_k^{(0)} \overline{\sigma}_{k,h} \min\Big\{1, \Big\| [\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} / \overline{\sigma}_{k,h} \Big\|_2 \Big\} \\
&\qquad + 2\beta_k^{(0)} \underline{\sigma}_{k,h} \min\Big\{1, \Big\| [\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} / \underline{\sigma}_{k,h} \Big\|_2 \Big\} \\
&\qquad + [\mathbb{P}_h \overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k),
\end{aligned}$$

where the second inequality holds because $[\mathbb{P}_h \overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) \ge -2H$, the last inequality holds since $\beta_k^{(0)} \overline{\sigma}_{k,h} \ge 2H$, $\beta_k^{(0)} \underline{\sigma}_{k,h} \ge 2H$. Subtracting $\overline{V}_{k,h+1}(s_{h+1}^k) - \underline{V}_{k,h+1}(s_{h+1}^k)$ from the both side, we can further get,

$$\begin{aligned}
&\overline{V}_{k,h}(s_h^k) - \underline{V}_{k,h}(s_h^k) - [\overline{V}_{k,h+1}(s_h^k) - \underline{V}_{k,h+1}(s_h^k)] \\
&\le 2\beta_k^{(0)} \overline{\sigma}_{k,h} \min\Big\{1, \Big\| [\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} / \overline{\sigma}_{k,h} \Big\|_2 \Big\} \\
&\qquad + 2\beta_k^{(0)} \underline{\sigma}_{k,h} \min\Big\{1, \Big\| [\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} / \underline{\sigma}_{k,h} \Big\|_2 \Big\} \\
&\qquad + [\mathbb{P}_h \overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) \\
&\qquad - [\overline{V}_{k,h+1}(s_{h+1}^k) - \underline{V}_{k,h+1}(s_{h+1}^k)],
\end{aligned} \tag{C.7}$$

Taking summation of (C.7) from $k = 1 \dots K$ and $h = h' \dots H$, we have following inequality holds

$$\begin{aligned}
&\sum_{k=1}^K [\overline{V}_{k,h'}(s_{h'}^k) - \underline{V}_{k,h'}(s_{h'}^k)] \\
&\le 2\beta_k^{(0)} \sum_{k=1}^K \sum_{h=h'}^H \overline{\sigma}_{k,h} \min\Big\{1, \Big\| [\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} / \overline{\sigma}_{k,h} \Big\|_2 \Big\} \\
&\qquad + 2\beta_k^{(0)} \sum_{k=1}^K \sum_{h=h'}^H \underline{\sigma}_{k,h} \min\Big\{1, \Big\| [\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} / \underline{\sigma}_{k,h} \Big\|_2 \Big\}
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{k=1}^{K}\sum_{h=h'}^{H}\Big[[\mathbb{P}_h\overline{V}_{k,h+1}](s_h^k,a_h^k,b_h^k)-[\mathbb{P}_h\underline{V}_{k,h+1}](s_h^k,a_h^k,b_h^k)\\
&-[\overline{V}_{k,h+1}(s_{h+1}^k)-\underline{V}_{k,h+1}(s_{h+1}^k)]\Big]\\
\le\,& 2\beta_k^{(0)}\sum_{k=1}^{K}\sum_{h=1}^{H}\overline{\sigma}_{k,h}\min\Big\{1,\Big\|[\overline{\mathbf{\Sigma}}_{k,h}^{(0)}]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}/\overline{\sigma}_{k,h}\Big\|_2\Big\}\\
&+2\beta_k^{(0)}\sum_{k=1}^{K}\sum_{h=1}^{H}\underline{\sigma}_{k,h}\min\Big\{1,\Big\|[\underline{\mathbf{\Sigma}}_{k,h}^{(0)}]^{-1/2}\boldsymbol{\phi}_{\underline{V}_{k,h+1}}/\underline{\sigma}_{k,h}\Big\|_2\Big\}+8H\sqrt{2T\log(H/\delta)}\\
\le\,& 2\beta_k^{(0)}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\overline{\sigma}_{k,h}^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\Big\{1,\Big\|[\overline{\mathbf{\Sigma}}_{k,h}^{(0)}]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}/\overline{\sigma}_{k,h}\Big\|_2^2\Big\}}\\
&+2\beta_k^{(0)}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\underline{\sigma}_{k,h}^2}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\min\Big\{1,\Big\|[\underline{\mathbf{\Sigma}}_{k,h}^{(0)}]^{-1/2}\boldsymbol{\phi}_{\underline{V}_{k,h+1}}/\underline{\sigma}_{k,h}\Big\|_2^2\Big\}}\\
&+8H\sqrt{2T\log(H/\delta)}\\
\le\,& 2\beta_K^{(0)}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\overline{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)}\\
&+2\beta_K^{(0)}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\underline{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)}+8H\sqrt{2T\log(H/\delta)}\\
\le\,& 4\beta_K^{(0)}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\overline{\sigma}_{k,h}^2+\underline{\sigma}_{k,h}^2}\sqrt{2Hd\log(1+K/\lambda)}+8H\sqrt{2T\log(H/\delta)},
\end{aligned}
\tag{C.8}
$$

where the first inequality holds since $\overline{V}_{k,H+1}=\underline{V}_{k,H+1}=0$, the second inequality holds on event $\mathcal{E}_1$, the third inequality holds due to Cauchy-Schwarz inequality, the fourth inequality holds due to Azuma Hoeffding inequality with the fact that $\Big\|\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k,a_h^k,b_h^k)/\overline{\sigma}_{k,h}\Big\|_2\le\Big\|\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k,a_h^k,b_h^k)\Big\|_2\cdot\sqrt{d}/H\le\sqrt{d}$, $\Big\|\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s_h^k,a_h^k,b_h^k)/\underline{\sigma}_{k,h}\Big\|_2\le\Big\|\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s_h^k,a_h^k,b_h^k)\Big\|_2\cdot\sqrt{d}/H\le\sqrt{d}$, the last inequality is by the fact that $\sqrt{a}+\sqrt{b}\le2\sqrt{a+b}$. (C.8) holds for any h', then we have following inequality holds

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{P}_h[\overline{V}_{k,h+1}-\underline{V}_{k,h+1}](s_h^k,a_h^k,b_h^k)\\
&=\sum_{k=1}^{K}\sum_{h=1}^{H}[\overline{V}_{k,h}-\underline{V}_{k,h}](s_h^k)+\sum_{k=1}^{K}\sum_{h=1}^{H}\Big[[\mathbb{P}_h\overline{V}_{k,h+1}](s_h^k,a_h^k,b_h^k)-[\mathbb{P}_h\underline{V}_{k,h+1}](s_h^k,a_h^k,b_h^k)\\
&\quad-[\overline{V}_{k,h+1}(s_{h+1}^k)-\underline{V}_{k,h+1}(s_{h+1}^k)]\Big]
\end{aligned}
$$

$$\leq 4\beta_K^{(0)}\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\overline{\sigma}_{k,h}^2+\underline{\sigma}_{k,h}^2}\sqrt{2H^3d\log(1+K/\lambda)}+8H^2\sqrt{2T\log(H/\delta)},$$

where the inequality holds due to (C.8) and on event $\mathcal{E}_1$. ∎

C.5. Proof of Lemma B.6

To estimate the variance in weighted ridge regression we need the following lemma, which is similar to the Lemma B.4 but without tolerant error ϵ.

Lemma C.3 *Suppose the event $\mathcal{E}$ hold. Then we have for any s, a, b, k, h, the following inequalities hold,*

$$\underline{Q}_{k,h}(s,a,b)\leq Q_h^{\mu^k}(s,a,b)\leq\overline{Q}_{k,h}(s,a,b),$$

and

$$\underline{V}_{k,h}(s)\leq V_h^{\mu^k}(s)\leq\overline{V}_{k,h}(s).$$

Proof For simplicity, we only prove the following UCB by induction,

$$Q_h^{\mu^k}(s,a,b)\leq\overline{Q}_{k,h}(s,a,b), V_h^{\mu^k}(s)\leq\overline{V}_{k,h}(s).$$

The base case $h = H + 1$ holds trivially since the terminal cost is zero. Now we assume that the bounds (C.4) holds for step $h + 1$. That is,

$$Q_{h+1}^{\mu^k}(s,a,b)\leq\overline{Q}_{k,h+1}(s,a,b), V_{h+1}^{\mu^k}(s)\leq\overline{V}_{k,h+1}(s).$$

If $\overline{Q}_{k,h}(s,a,b)\geq H$, then it is obvious to have $Q_h^{\mu^k}(s,a,b)\leq\overline{Q}_{k,h}(s,a,b)$, otherwise we have that

$$\begin{aligned}
&\overline{Q}_{k,h}(s,a,b)-Q_h^{\mu^k}(s,a,b)\\
&=\langle\overline{\boldsymbol{\theta}}_{k,h}^{(0)},\boldsymbol{\phi}_{\overline{V}_{k,h+1}}\rangle+\beta_k^{(0)}\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}\Big\|_2-\langle\boldsymbol{\theta}_h^*,\boldsymbol{\phi}_{\overline{V}_{k,h+1}}\rangle\\
&\quad+\mathbb{P}_h\overline{V}_{k,h+1}(s,a,b)-\mathbb{P}_hV_{h+1}^{*,\nu^k}(s)\\
&\geq\beta_k^{(0)}\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}\Big\|_2-\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(0)}-\boldsymbol{\theta}_h^*)\Big\|_2\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}\Big\|_2\\
&\quad+\mathbb{P}_h\overline{V}_{k,h+1}(s)-\mathbb{P}_hV_{h+1}^{\mu^k}(s)\\
&\geq\mathbb{P}_h\overline{V}_{k,h+1}(s)-\mathbb{P}_hV_{h+1}^{\mu^k}(s)\\
&\geq 0,
\end{aligned}\tag{C.9}$$

where the first inequality holds due to Cauchy-Schwarz inequality, the second inequality holds since the assumption that $\theta_h^*\in\overline{\mathcal{C}}_{k,h}^{(0)}$ in event $\mathcal{E}$, the third inequality holds by the induction assumption.

Then we have that

$$\overline{V}_{k,h}(s)=\mathbb{E}_{(a,b)\sim\mu_h^k(\cdot,\cdot|s)}\big[\overline{Q}_{k,h}(s,a,b)\big]$$

$$\begin{aligned}&\geq \mathbb{E}_{(a,b)\sim\mu_h^k(\cdot,\cdot|s)}\big[Q_h^{\mu^k}(s,a',b)\big]\\&= V_h^{\mu^k}(s),\end{aligned}$$

where the inequality is by (C.9), the last inequality is due to the Bellman equation. Therefore, our proof is completed. ■

Lemma C.4 *(Lemma 11, Abbasi-Yadkori et al. 2011). For any $\{\mathbf{x}_t\}_{t=1}^T \subset \mathbb{R}^d$ satisfying that $\|\mathbf{x}_t\|_2 \leq L$, let $\mathbf{A}_0 = \lambda\mathbf{I}$ and $\mathbf{A}_t = \mathbf{A}_0 + \sum_{i=1}^t \mathbf{x}_i\mathbf{x}_i^\top$, then we have*

$$\sum_{t=1}^T \min\{1, \|\mathbf{x}_t\|_{\mathbf{A}_{t-1}^{-1}}^2\} \leq 2d\log\frac{d\lambda + TL^2}{d\lambda}.$$

Proof [Proof of Lemma B.6] Suppose the event in Lemma 9.1 holds, we have the following results:

$$\begin{aligned}\sum_{k=1}^K\sum_{h=1}^H \overline{\sigma}_{k,h}^2 &= \sum_{k=1}^K\sum_{h=1}^H \big[H^2/d + \mathbb{V}_{k,h}^{\text{est}}\overline{V}_{k,h+1}(s_h^k,a_h^k,b_h^k) + \overline{E}_{k,h}\big]\\
&= H^2T/d + \underbrace{\sum_{k=1}^K\sum_{h=1}^H \big[\mathbb{V}_h\overline{V}_{k,h+1}(s_h^k,a_h^k,b_h^k) - \mathbb{V}_hV_{h+1}^{\mu^k}(s_h^k,a_h^k,b_h^k)\big]}_{I_1}\\
&\quad + \underbrace{2\sum_{k=1}^K\sum_{h=1}^H \overline{E}_{k,h}}_{I_2} + \underbrace{\sum_{k=1}^K\sum_{h=1}^H \mathbb{V}_hV_{h+1}^{\mu^k}(s_h^k,a_h^k,b_h^k)}_{I_3}\\
&\quad + \underbrace{\sum_{k=1}^K\sum_{h=1}^H \big[\mathbb{V}_h^{\text{est}}\overline{V}_{k,h+1}(s_h^k,a_h^k,b_h^k) - \mathbb{V}_h\overline{V}_{k,h+1}(s_h^k,a_h^k,b_h^k) - \overline{E}_{k,h}\big]}_{I_4},\end{aligned} \tag{C.10}$$

where the first equation is by the definition of $\overline{\sigma}_{k,h}$. To bound I_1, we have

$$\begin{aligned}I_1 &= \sum_{k=1}^K\sum_{h=1}^H \big[\mathbb{P}_h\overline{V}_{k,h+1}^2(s_h^k,a_h^k,b_h^k) - \mathbb{P}_h[V_{h+1}^{\mu^k}]^2(s_h^k,a_h^k,b_h^k)\big]\\
&\quad - \sum_{k=1}^K\sum_{h=1}^H \big[[\mathbb{P}_h\overline{V}_{k,h+1}]^2(s_h^k,a_h^k,b_h^k) - [\mathbb{P}_hV_{h+1}^{\mu^k}]^2(s_h^k,a_h^k,b_h^k)\big]\\
&\leq \sum_{k=1}^K\sum_{h=1}^H \mathbb{P}_h[(\overline{V}_{k,h+1} - V_{h+1}^{\mu^k})(\overline{V}_{k,h+1} + V_{h+1}^{\mu^k})](s_h^k,a_h^k,b_h^k)\\
&\quad - \sum_{k=1}^K\sum_{h=1}^H [(\mathbb{P}_h\overline{V}_{k,h+1} - \mathbb{P}_hV_{h+1}^{\mu^k})(\mathbb{P}_h\overline{V}_{k,h+1} + \mathbb{P}_hV_{h+1}^{\mu^k})](s_h^k,a_h^k,b_h^k)\\
&\leq 4H\sum_{k=1}^K\sum_{h=1}^H \mathbb{P}_h[\overline{V}_{k,h+1} - V_{h+1}^{\mu^k}](s_h^k,a_h^k,b_h^k)\end{aligned}$$

$$= 4H \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{P}_h[\overline{V}_{k,h+1} - V_{h+1}^{\mu^k}](s_h^k, a_h^k, b_h^k),$$

where the first inequality is by $|\overline{V}_{k,h+1}|, |V_{h+1}^{\mu^k}| \leq H$, and the second inequality is by $\overline{V}_{k,h+1} - V_{h+1}^{\mu^k} \geq 0$ due to Lemma C.3. To bound I_2, we have

$$\begin{aligned}
I_2 &\leq 2 \sum_{k=1}^{K} \sum_{h=1}^{H} \beta_k^{(1)} \min\Big\{1, \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k) \Big\|_2 \Big\} \\
&\quad + 4H \sum_{k=1}^{K} \sum_{h=1}^{H} \beta_k^{(2)} \bar{\sigma}_{k,h} \min\Big\{1, \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)/\bar{\sigma}_{k,h} \Big\|_2 \Big\} \\
&\leq 2\beta_K^{(1)} \sqrt{T} \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \min\Big\{1, \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k) \Big\|_2^2 \Big\}} \\
&\quad + 7\beta_K^{(1)} H^2 \sqrt{T} \sqrt{\sum_{k=1}^{K} \sum_{h=1}^{H} \min\Big\{1, \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k) \Big\|_2^2 / \bar{\sigma} \Big\}} \\
&\leq 2\beta_K^{(2)} \sqrt{T} \sqrt{2dH \log(1 + KH^4/(d\lambda))} + 7\beta_K^{(1)} H^2 \sqrt{T} \sqrt{2dH \log(1 + K/\lambda)},
\end{aligned}$$

where the first inequality holds due to $\beta_k^{(1)} \geq H^2$ and $\beta_k^{(2)} \bar{\sigma}_{k,h} \geq \sqrt{d} \cdot H/\sqrt{d} = H$, the second inequality holds due to Cauchy-Schwartz inequality, $\beta_k^{(1)} \leq \beta_K^{(1)}$, $\beta_k^{(2)} \leq \beta_K^{(2)}$,

$$\overline{\sigma}_{k,h}^2 = \max\{H^2/d, \mathbb{V}^{\text{est}} \overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) + \overline{E}_{k,h}\} \leq \max\{H^2/d, H^2 + 2H^2\} = 3H^2,$$

the third inequality holds due to Lemma C.4. Next we bound I_3, since event $\mathcal{E}_2$ holds , we have

$$I_3 \leq 3(HT + H^3 \log(1/\delta)).$$

Finally, due to on event $\mathcal{E}$, we have $I_4 \leq 0$. We finish the proof by substituting I_1, I_2, I_3, I_4 into (C.10). ■

Appendix D. Full Version of Algorithm 1

In this section, we present the full version of Algorithm 1 in Algorithm 2.
Update of optimistic action-value function:

$$\begin{aligned}
\overline{Q}_{k,h}(\cdot,\cdot,\cdot) &\leftarrow \Big[r_h(\cdot,\cdot,\cdot) + \langle \overline{\boldsymbol{\theta}}_{k,h}^{(0)}, \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(\cdot,\cdot,\cdot) \rangle + \beta_k^{(0)} \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(\cdot,\cdot,\cdot) \Big\|_2 \Big]_{[-H,H]} \\
\underline{Q}_{k,h}(\cdot,\cdot,\cdot) &\leftarrow \Big[r_h(\cdot,\cdot,\cdot) + \langle \underline{\boldsymbol{\theta}}_{k,h}^{(0)}, \boldsymbol{\phi}_{\underline{V}_{k,h+1}}(\cdot,\cdot,\cdot) \rangle - \beta_k^{(0)} \Big\| \big[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}}(\cdot,\cdot,\cdot) \Big\|_2 \Big]_{[-H,H]}.
\end{aligned} \tag{D.1}$$

Update of variance estimation:

$$\mathbb{V}^{\text{est}} \overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) \leftarrow \big[\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(1)} \rangle \big]_{[0,H^2]} - \big[\langle \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(0)} \rangle \big]_{[-H,H]}^2,$$

Algorithm 2 Nash-UCRL

Input: Regularization parameter λ, Number of episode K, number of horizon H.
For any h, $\overline{\mathbf{\Sigma}}_{1,h}^{(i)} \leftarrow \underline{\mathbf{\Sigma}}_{1,h}^{(i)} \leftarrow \lambda\mathbf{I}$; $\overline{\mathbf{b}}_{1,h}^{(i)} \leftarrow \underline{\mathbf{b}}_{1,h}^{(i)} \leftarrow \mathbf{0}$; $\overline{\boldsymbol{\theta}}_{1,h}^{(i)} \leftarrow \underline{\boldsymbol{\theta}}_{1,h}^{(i)} \leftarrow \mathbf{0}$, for $i \in \{0,1\}$.
for $k = 1, \ldots, K$ **do**
  $\overline{V}_{k,H+1}(\cdot) \leftarrow 0$, $\underline{V}_{k,H+1}(\cdot) \leftarrow 0$
  for $h = H, \ldots, 1$ **do**
    Set $\overline{Q}_{k,h}(\cdot,\cdot,\cdot)$ and $\underline{Q}_{k,h}(\cdot,\cdot,\cdot)$ as in (D.1).
    for $s \in \mathcal{S}$ **do**
      Let $\mu_h^k(\cdot,\cdot|s) = \epsilon\text{-CCE}(\overline{Q}_{k,h}(s,\cdot,\cdot), \underline{Q}_{k,h}(s,\cdot,\cdot))$.
      $\overline{V}_{k,h}(s) = \mathbb{E}_{(a,b)\sim\mu_h^k(\cdot,\cdot|s)}\overline{Q}_{k,h}(s,a,b)$, $\underline{V}_{k,h}(s) = \mathbb{E}_{(a,b)\sim\mu_h^k(\cdot,\cdot|s)}\underline{Q}_{k,h}(s,a,b)$
      $\pi_h^k(\cdot|s) = \mathcal{P}_{\max}\mu_h^k(\cdot,\cdot|s)$, $\nu_h^k(\cdot|s) = \mathcal{P}_{\min}\mu_h^k(\cdot,\cdot|s)$
    end for
  end for
  receives s_1^k
  for $h = 1, \ldots, H$ **do**
    Take action $a_h^k \sim \pi_h^k(s_h^k)$ and $b_h^k \sim \nu_h^k(s_h^k)$ and receives $s_{h+1}^k \sim \mathbb{P}(\cdot|s_h^k, a_h^k, b_h^k)$.
    Set $\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)$ and $\mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k)$ as in (D.2).
    Set $\overline{E}_{k,h}, \underline{E}_{k,h}, \overline{\boldsymbol{\sigma}}_{k,h}, \underline{\boldsymbol{\sigma}}_{k,h}, \overline{\mathbf{\Sigma}}_{k+1,h}^{(0)}, \underline{\mathbf{\Sigma}}_{k+1,h}^{(0)}, \overline{\mathbf{b}}_{k+1,h}^{(0)}, \underline{\mathbf{b}}_{k+1,h}^{(0)}, \overline{\mathbf{\Sigma}}_{k+1,h}^{(1)}, \underline{\mathbf{\Sigma}}_{k+1,h}^{(1)}, \overline{\mathbf{b}}_{k+1,h}^{(1)}, \underline{\mathbf{b}}_{k+1,h}^{(1)}$ as defined in (D.3).
    Set $\overline{\boldsymbol{\theta}}_{k+1,h}^{(i)} \leftarrow \big[\overline{\mathbf{\Sigma}}_{k+1,h}^{(i)}\big]^{-1}\overline{\mathbf{b}}_{k+1,h}^{(i)}$, $\underline{\boldsymbol{\theta}}_{k+1,h}^{(i)} \leftarrow \big[\underline{\mathbf{\Sigma}}_{k+1,h}^{(i)}\big]^{-1}\underline{\mathbf{b}}_{k+1,h}^{(i)}$, $i = 0, 1$
  end for
end for

$$\mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) \leftarrow \big[\langle \boldsymbol{\phi}_{\underline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k), \underline{\boldsymbol{\theta}}_{k,h}^{(1)}\rangle\big]_{[0,H^2]} - \big[\langle \boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \underline{\boldsymbol{\theta}}_{k,h}^{(0)}\rangle\big]_{[-H,H]}^2. \tag{D.2}$$

Update of other parameters:

$$\begin{aligned}
\overline{E}_{k,h} &= \min\big\{H^2, \beta_k^{(1)}\Big\|\big[\overline{\mathbf{\Sigma}}_{k,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k)\Big\|_2\big\} \\
&\quad + \min\big\{H^2, 2H\beta_k^{(2)}\Big\|\overline{\mathbf{\Sigma}}_{k,h}^{(0)-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\Big\|_2\big\} \\
\underline{E}_{k,h} &= \min\big\{H^2, \beta_k^{(1)}\Big\|\big[\underline{\mathbf{\Sigma}}_{k,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\underline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k)\Big\|_2\big\} \\
&\quad + \min\big\{H^2, 2H\beta_k^{(2)}\Big\|\underline{\mathbf{\Sigma}}_{k,h}^{(0)-1/2}\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\Big\|_2\big\}, \\
\overline{\boldsymbol{\sigma}}_{k,h} &= \sqrt{\max\{H^2/4d, \mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) + \overline{E}_{k,h}\}}, \\
\underline{\boldsymbol{\sigma}}_{k,h} &= \sqrt{\max\{H^2/4d, \mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) + \underline{E}_{k,h}\}}, \\
\overline{\mathbf{\Sigma}}_{k+1,h}^{(0)} &\leftarrow \overline{\mathbf{\Sigma}}_{k,h}^{(0)} + \overline{\boldsymbol{\sigma}}_{k,h}^{-2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)^\top, \\
\underline{\mathbf{\Sigma}}_{k+1,h}^{(0)} &\leftarrow \underline{\mathbf{\Sigma}}_{k,h}^{(0)} + \underline{\boldsymbol{\sigma}}_{k,h}^{-2}\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)^\top, \\
\overline{\mathbf{b}}_{k+1,h}^{(0)} &= \overline{\mathbf{b}}_{k,h}^{(0)} + \overline{\boldsymbol{\sigma}}_{k,h}^{-2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\overline{V}_{k,h+1}(s_{h+1}^k),
\end{aligned}$$

$$\begin{aligned}
\underline{\mathbf{b}}^{(0)}_{k+1,h} &= \underline{\mathbf{b}}^{(0)}_{k,h} + \underline{\boldsymbol{\sigma}}^{-2}_{k,h}\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\underline{V}_{k,h+1}(s_{h+1}^k),\\
\overline{\boldsymbol{\Sigma}}^{(1)}_{k+1,h} &\leftarrow \overline{\boldsymbol{\Sigma}}^{(1)}_{k,h} + \boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)^\top,\\
\underline{\boldsymbol{\Sigma}}^{(1)}_{k+1,h} &\leftarrow \underline{\boldsymbol{\Sigma}}^{(1)}_{k,h} + \boldsymbol{\phi}_{\underline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\boldsymbol{\phi}_{\underline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)^\top,\\
\overline{\mathbf{b}}^{(1)}_{k+1,h} &= \overline{\mathbf{b}}^{(1)}_{k,h} + \boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\overline{V}^2_{k,h+1}(s_{h+1}^k),\\
\underline{\mathbf{b}}^{(1)}_{k+1,h} &= \underline{\mathbf{b}}^{(1)}_{k,h} + \boldsymbol{\phi}_{\underline{V}^2_{k,h+1}}(s_h^k, a_h^k, b_h^k)\underline{V}^2_{k,h+1}(s_{h+1}^k). \qquad \text{(D.3)}
\end{aligned}$$

In line 8 of Algorithm 2, we need to call the ϵ-CCE subroutine. In detail, for any fixed state s, and two matrices $\overline{Q}_{k,h}(s,\cdot,\cdot)$ and $\underline{Q}_{k,h}(s,\cdot,\cdot) \in [0,1]^{|\mathcal{A}_{\max}|\times|\mathcal{A}_{\min}|}$, the subroutine ϵ-CCE$(\cdot,\cdot)$ returns a distribution $\sigma \in \Delta_{|\mathcal{A}_{\max}|\times|\mathcal{A}_{\min}|}$ that satisfies

$$\begin{aligned}
\mathbb{E}_{(a,b)\sim\sigma}\overline{Q}_{k,h}(s,a,b) &\geq \max_{a'\in\mathcal{A}_{\max}} \mathbb{E}_{(a,b)\sim\sigma}\overline{Q}_{k,h}(s,a',b) - \epsilon,\\
\mathbb{E}_{(a,b)\sim\sigma}\underline{Q}_{k,h}(s,a,b) &\leq \min_{b'\in\mathcal{A}_{\min}} \mathbb{E}_{(a,b)\sim\sigma}\underline{Q}_{k,h}(s,a,b') + \epsilon. \qquad \text{(D.4)}
\end{aligned}$$

(D.4) is a feasibility problem, where the constraints can be rewritten as $|\mathcal{A}_{\max}| + |\mathcal{A}_{\min}|$ linear constraints on $\sigma \in \Delta_{|\mathcal{A}_{\max}|\times|\mathcal{A}_{\min}|}$. Thus it can be efficiently resolved by any linear programming algorithms. See also Appendix B in Liu et al. (2020) and Xie et al. (2020) for more detailed discussions.

Appendix E. Extensions to Turn-based Games

In this section, we extend our algorithm and results to turn-based Markov games.

Turn-based MGs A two-player zero-sum turn-based episodic MG is denoted by a tuple $M(\mathcal{S},\mathcal{A},H,\{r_h\}_{h=1}^H,\{\mathbb{P}_h\}_{h=1}^H)$, where $\mathcal{S} = \mathcal{S}_{\max}\cup\mathcal{S}_{\min}$, $\mathcal{S}_{\max}$ ($\mathcal{S}_{\min}$) are the states where the max (min)-player plays, $\mathcal{S}_{\max}\cap\mathcal{S}_{\min} = \emptyset$. Note that the partition of state space suggests that at each step, only one player can play. $\mathcal{A}$ is the action space, H is the length of game/episode, $r_h : \mathcal{S}\times\mathcal{A}\to[-1,1]$ is the reward function, $\mathbb{P}_h(s'|s,a)$ denotes the transition probability for the max (min)-player ($s \in \mathcal{S}_{\max}$ or $\mathcal{S}_{\max}$) to take action a and transit to next state s'. Similar to the linear mixture MGs, we can define linear mixture turn-based MGs as follows.

Definition E.1 *$M(\mathcal{S},\mathcal{A},H,\{r_h\}_{h=1}^H,\{\mathbb{P}_h\}_{h=1}^H)$ is called a time-inhomogeneous, episodic B-bounded linear mixture turn-based Markov game if there exist $\{\boldsymbol{\theta}_h\}_{h=1}^H \subset \mathbb{R}^d$ and $\widetilde{\boldsymbol{\phi}}(s'|s,a)\in\mathbb{R}^d$ satisfying*

$$\|\boldsymbol{\theta}_h\|_2 \leq B, \qquad \forall V:\mathcal{S}\to[-1,1], \left\|\sum_{s'\in\mathcal{S}}\widetilde{\boldsymbol{\phi}}(s'|s,a)V(s')\right\|_2 \leq 1,$$

such that $\mathbb{P}_h(s'|s,a) = \langle\boldsymbol{\phi}(s'|s,a),\boldsymbol{\theta}_h\rangle$ for any state-action-state triplet (s,a,s') and any step h.

Based on above definition, we show that any turn-based linear mixture MG can be regarded as a special case of linear mixture simultaneous-move MG. In fact, for any turn-based linear mixture MG with feature mapping $\widetilde{\boldsymbol{\phi}}(\cdot|\cdot,\cdot)$ and reward $\widetilde{r}_h(\cdot,\cdot)$, we can define the corresponding linear mixture simultaneous-move MG with feature mapping $\boldsymbol{\phi}(\cdot|\cdot,\cdot,\cdot)$ and reward $r_h(\cdot,\cdot,\cdot)$ as follows: for each $s\in\mathcal{S}_{\max}$,

$$\boldsymbol{\phi}(s'|s,a,b) = \widetilde{\boldsymbol{\phi}}_h(s'|s,a),\ r_h(s'|s,a,b) = \widetilde{r}_h(s'|s,a),$$

and for each $s \in \mathcal{S}_{\min}$,

$$\boldsymbol{\phi}(s'|s,a,b) = \widetilde{\boldsymbol{\phi}}_h(s'|s,b),\ r_h(s'|s,a,b) = \widetilde{r}_h(s'|s,b).$$

Therefore, we can still use Algorithm 1 to find the Nash equilibrium. Notice that for the turn-based game, at each step only one player can take action. Thus, the ϵ-CCE routine in Line 8 of Algorithm 1 needs be replaced by two separate subroutines: taking π_h^k and ν_h^k as greedy policies w.r.t. $\overline{Q}_{k,h}$ and $\underline{Q}_{k,h}$. For completeness, we present the turn-based version of Algorithm 1 as Algorithm 3.

Algorithm 3 Turn-based Nash-UCRL

For any h, $\overline{\boldsymbol{\Sigma}}_{1,h}^{(i)} \leftarrow \underline{\boldsymbol{\Sigma}}_{1,h}^{(i)} \leftarrow \lambda\mathbf{I}$; $\overline{\mathbf{b}}_{1,h}^{(i)} \leftarrow \underline{\mathbf{b}}_{1,h}^{(i)} \leftarrow \mathbf{0}$; $\overline{\boldsymbol{\theta}}_{1,h}^{(i)} \leftarrow \underline{\boldsymbol{\theta}}_{1,h}^{(i)} \leftarrow \mathbf{0}$, for $i \in \{0,1\}$.
for $k = 1, \ldots, K$ **do**
  $\overline{V}_{k,H+1}(\cdot) \leftarrow 0$, $\underline{V}_{k,H+1}(\cdot) \leftarrow 0$
  for $h = H, \ldots, 1$ **do**
    Set $\overline{Q}_{k,h}(\cdot,\cdot)$ and $\underline{Q}_{k,h}(\cdot,\cdot)$ as in (E.1).
    for $s \in \mathcal{S}_{\max}$ **do**
      $\pi_h^k(\cdot|s) = \max_{a\in\mathcal{A}} \overline{Q}_{k,h}(s,a)$, $\overline{V}_{k,h}(s) = \mathbb{E}_{a\sim\pi_h^k(\cdot|s)}\overline{Q}_{k,h}(s,a)$.
    end for
    for $s \in \mathcal{S}_{\min}$ **do**
      $\nu_h^k(\cdot|s) = \min_{b\in\mathcal{A}} \underline{Q}_{k,h}(s,b)$, $\underline{V}_{k,h}(s) = \mathbb{E}_{b\sim\nu_h^k(\cdot|s)}\underline{Q}_{k,h}(s,b)$.
    end for
  end for
  receives s_1^k
  for $h = 1, \ldots, H$ **do**
    if $s_h^k \in \mathcal{S}_{\max}$ **then**
      Take action $a_h^k \sim \pi_h^k(\cdot|s_h^k)$ and receives $s_{h+1}^k \sim \mathbb{P}(\cdot|s_h^k, a_h^k)$.
    else
      Take action $a_h^k \sim \nu_h^k(\cdot|s_h^k)$ and receives $s_{h+1}^k \sim \mathbb{P}(\cdot|s_h^k, a_h^k)$.
    end if
    Set $\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k)$ and $\mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s_h^k, a_h^k)$ as in (E.2).
    Set $\overline{E}_{k,h}, \underline{E}_{k,h}, \overline{\boldsymbol{\sigma}}_{k,h}, \underline{\boldsymbol{\sigma}}_{k,h}$, $\overline{\boldsymbol{\Sigma}}_{k+1,h}^{(0)}$, $\underline{\boldsymbol{\Sigma}}_{k+1,h}^{(0)}$, $\overline{\mathbf{b}}_{k+1,h}^{(0)}$, $\underline{\mathbf{b}}_{k+1,h}^{(0)}$, $\overline{\boldsymbol{\Sigma}}_{k+1,h}^{(1)}$, $\underline{\boldsymbol{\Sigma}}_{k+1,h}^{(1)}$, $\overline{\mathbf{b}}_{k+1,h}^{(1)}$, $\underline{\mathbf{b}}_{k+1,h}^{(1)}$ as defined in (E.3).
    Set $\overline{\boldsymbol{\theta}}_{k+1,h}^{(i)} \leftarrow \big[\overline{\boldsymbol{\Sigma}}_{k+1,h}^{(i)}\big]^{-1}\overline{\mathbf{b}}_{k+1,h}^{(i)}$, $\underline{\boldsymbol{\theta}}_{k+1,h}^{(i)} \leftarrow \big[\underline{\boldsymbol{\Sigma}}_{k+1,h}^{(i)}\big]^{-1}\underline{\mathbf{b}}_{k+1,h}^{(i)}$, $i = 0, 1$
  end for
end for

Update of optimistic action-value function:

$$\overline{Q}_{k,h}(\cdot,\cdot) \leftarrow \min\{H, \widetilde{r}_h(\cdot,\cdot) + \langle \overline{\boldsymbol{\theta}}_{k,h}^{(0)}, \widetilde{\boldsymbol{\phi}}_{\overline{V}_{k,h+1}}(\cdot,\cdot)\rangle + \beta_k^{(0)} \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \widetilde{\boldsymbol{\phi}}_{\overline{V}_{k,h+1}}(\cdot,\cdot)\Big\|_2\}$$
$$\underline{Q}_{k,h}(\cdot,\cdot) \leftarrow \max\{-H, \widetilde{r}_h(\cdot,\cdot) + \langle \underline{\boldsymbol{\theta}}_{k,h}^{(0)}, \widetilde{\boldsymbol{\phi}}_{\underline{V}_{k,h+1}}(\cdot,\cdot)\rangle - \beta_k^{(0)} \Big\| \big[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \widetilde{\boldsymbol{\phi}}_{\underline{V}_{k,h+1}}(\cdot,\cdot)\Big\|_2\}. \quad \text{(E.1)}$$

Update of variance estimation:

$$\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k) \leftarrow \big[\langle \widetilde{\boldsymbol{\phi}}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(1)}\rangle\big]_{[0,H^2]} - \big[\langle \widetilde{\boldsymbol{\phi}}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k), \overline{\boldsymbol{\theta}}_{k,h}^{(0)}\rangle\big]_{[-H,H]}^2,$$

$$\mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s_h^k, a_h^k) \leftarrow \big[\langle \widetilde{\boldsymbol{\phi}}_{\underline{V}^2_{k,h+1}}(s_h^k, a_h^k), \underline{\boldsymbol{\theta}}^{(1)}_{k,h}\rangle\big]_{[0,H^2]} - \big[\langle \widetilde{\boldsymbol{\phi}}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k), \underline{\boldsymbol{\theta}}^{(0)}_{k,h}\rangle\big]^2_{[-H,H]}. \quad \text{(E.2)}$$

Update of other parameters:

$$\begin{aligned}
\overline{E}_{k,h} &= \min\big\{H^2, \beta_k^{(1)}\Big\|\big[\overline{\boldsymbol{\Sigma}}^{(1)}_{k,h}\big]^{-1/2}\widetilde{\boldsymbol{\phi}}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k)\Big\|_2\big\} \\
&\quad + \min\big\{H^2, 2H\beta_k^{(2)}\Big\|\overline{\boldsymbol{\Sigma}}^{(0)-1/2}_{k,h}\widetilde{\boldsymbol{\phi}}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k)\Big\|_2\big\}, \\
\underline{E}_{k,h} &= \min\big\{H^2, \beta_k^{(1)}\Big\|\big[\underline{\boldsymbol{\Sigma}}^{(1)}_{k,h}\big]^{-1/2}\widetilde{\boldsymbol{\phi}}_{\underline{V}^2_{k,h+1}}(s_h^k, a_h^k)\Big\|_2\big\} \\
&\quad + \min\big\{H^2, 2H\beta_k^{(2)}\Big\|\underline{\boldsymbol{\Sigma}}^{(0)-1/2}_{k,h}\widetilde{\boldsymbol{\phi}}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k)\Big\|_2\big\}, \\
\overline{\boldsymbol{\sigma}}_{k,h} &= \sqrt{\max\{H^2/d, \mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k) + \overline{E}_{k,h}\}}, \\
\underline{\boldsymbol{\sigma}}_{k,h} &= \sqrt{\max\{H^2/d, \mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s_h^k, a_h^k) + \underline{E}_{k,h}\}}, \\
\overline{\boldsymbol{\Sigma}}^{(0)}_{k+1,h} &\leftarrow \overline{\boldsymbol{\Sigma}}^{(0)}_{k,h} + \overline{\boldsymbol{\sigma}}^{-2}_{k,h}\widetilde{\boldsymbol{\phi}}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k)\widetilde{\boldsymbol{\phi}}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k)^\top \\
\underline{\boldsymbol{\Sigma}}^{(0)}_{k+1,h} &\leftarrow \underline{\boldsymbol{\Sigma}}^{(0)}_{k,h} + \underline{\boldsymbol{\sigma}}^{-2}_{k,h}\widetilde{\boldsymbol{\phi}}_{\underline{V}_{k,h+1}}(s_h^k, b_h^k)\widetilde{\boldsymbol{\phi}}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k)^\top \\
\overline{\mathbf{b}}^{(0)}_{k+1,h} &= \overline{\mathbf{b}}^{(0)}_{k,h} + \overline{\boldsymbol{\sigma}}^{-2}_{k,h}\widetilde{\boldsymbol{\phi}}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k)\overline{V}_{k,h+1}(s_{k,h+1}) \\
\underline{\mathbf{b}}^{(0)}_{k+1,h} &= \underline{\mathbf{b}}^{(0)}_{k,h} + \underline{\boldsymbol{\sigma}}^{-2}_{k,h}\widetilde{\boldsymbol{\phi}}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k)\underline{V}_{k,h+1}(s_{k,h+1}) \\
\overline{\boldsymbol{\Sigma}}^{(1)}_{k+1,h} &\leftarrow \overline{\boldsymbol{\Sigma}}^{(1)}_{k,h} + \widetilde{\boldsymbol{\phi}}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k)\widetilde{\boldsymbol{\phi}}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k)^\top \\
\underline{\boldsymbol{\Sigma}}^{(1)}_{k+1,h} &\leftarrow \underline{\boldsymbol{\Sigma}}^{(1)}_{k,h} + \widetilde{\boldsymbol{\phi}}_{\underline{V}^2_{k,h+1}}(s_h^k, a_h^k)\widetilde{\boldsymbol{\phi}}_{\underline{V}^2_{k,h+1}}(s_h^k, a_h^k)^\top \\
\overline{\mathbf{b}}^{(1)}_{k+1,h} &= \overline{\mathbf{b}}^{(1)}_{k,h} + \widetilde{\boldsymbol{\phi}}_{\overline{V}^2_{k,h+1}}(s_h^k, a_h^k)\overline{V}^2_{k,h+1}(s_{h+1}^k), \\
\underline{\mathbf{b}}^{(1)}_{k+1,h} &= \underline{\mathbf{b}}^{(1)}_{k,h} + \widetilde{\boldsymbol{\phi}}_{\underline{V}^2_{k,h+1}}(s_h^k, a_h^k)\underline{V}^2_{k,h+1}(s_{h+1}^k). \qquad \text{(E.3)}
\end{aligned}$$

By Theorem 5.2, we immediately have that the regret of our turn-based algorithm in Algorithm 3 is also bounded by

$$\widetilde{\mathcal{O}}(\sqrt{d^2H^2 + dH^3}\sqrt{T} + d^2H^3 + d^3H^2),$$

where $T = KH$. Similarly, we can show that if $d \geq H$ and $T \geq d^4H^2$, our turn-based algorithm is nearly minimax optimal.