Almost Optimal Algorithms for Two-player Zero-Sum Markov Games with Linear Function Approximation

Zixiang Chen [1] **Dongruo Zhou** [1] **Quanquan Gu** [1]

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 very efficient. Specifically, we show that Nash-UCRL can provably achieve an $\widetilde{O}(\sqrt{d^2H^2T})$ 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 $\widetilde{\Omega}(\sqrt{d^2H^2T})$ lower bound on the regret. Our upper bound matches the lower bound up to logarithmic factors, which suggests the optimality of our algorithm.

1. Introduction

Multi-agent reinforcement learning (MARL) has achieved tremendous practical success across a wide range of machine learning tasks. For Markov games with large state and action spaces, it is natural to use linear function approximation. In particular, Xie et al. (2020) proposed an 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, with d being the dimension of the linear structure. However, as we will show in this paper, there is still a gap between the upper and lower bounds of existing algorithms for Markov games with linear structures.

*Equal contribution [1]Department of Computer Science, University of California, Los Angeles, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

In this paper, we consider Markov games with a linear mixture structure and design a minimax optimal algorithm for learning zero-sum Markov games 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 (which can be specialized to turn-based games) that can provably achieve an $\widetilde{O}(\sqrt{d^2H^2T})$ 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.
- To access the optimality of our algorithm Nash-UCRL, we also prove a regret lower bound $\Omega(\sqrt{d^2H^2T})$. Our upper bound matches the lower bound up to logarithmic factors, which suggests the optimality of our algorithm.

Notations 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 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 $\exists C > 0, N > 0, \forall n > N, a_n \le Cb_n$; $a_n = \Omega(b_n)$ if $a_n \ge Cb_n$. The notation $\widetilde{O}$ is used to hide logarithmic factors.

2. Preliminaries

Simultaneous-move MG. Two-player zero-sum Markov game (MG) (Shapley, 1953; Littman, 1994) is a generalization of the standard Markov decision process (MDP). 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.

We now define the stochastic policies, which give distributions over the actions. A policy $\pi = \{\pi_h : \mathcal{S} \rightarrow \Delta_{\mathcal{A}_{\max}}\}_{h=1}^H$ 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 ν for the min-player. 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 use the notation $Q_h^{\pi,\nu} : \mathcal{S} \times \mathcal{A}_{\max} \times \mathcal{A}_{\min} \rightarrow \mathbb{R}$ to present the action-value function (a.k.a., Q function), and the notation $V_h^{\pi,\nu} : \mathcal{S} \rightarrow \mathbb{R}$ to present the value function..

Learning Objective. 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-mean 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:

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

By symmetry, we can also define $Q_h^{\pi,*}(s,a,b), V_h^{\pi,*}(s)$. A Nash Equilibrium (NE) of the game is a pair of policies π^*, ν^* such that

$$V_1^{\pi^*,\nu^*}(s) = V_1^{\pi^*,*}(s) = V_1^{*,\nu^*}(s), \text{for all } s \in \mathcal{S}. \quad (2.1)$$

For most applications, they are the ultimate solutions we want to pursue. We can measure the suboptimality of learned policies $\{\pi^k, \nu^k\}_k$ by the gap between their performance and the performance of the optimal strategy (i.e., Nash equilibrium) when playing against the best responses respectively:

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

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. This regret has been widely used in previous work that studies the offline learning of two-player game (Jin et al., 2018; Xie et al., 2020; Liu et al., 2020).

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 $(s,a,b,s') \in \mathcal{S} \times \mathcal{A}_{\max} \times \mathcal{A}_{\max} \times \mathcal{S}$. 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 $V : \mathcal{S} \rightarrow [-1,1]$ and any tuple $(s,a,b) \in \mathcal{S} \times \mathcal{A}_{\max} \times \mathcal{A}_{\min}$, we have

$$\|\boldsymbol{\phi}_V(s,a,b)\|_2 \leq 1, \quad (2.2)$$

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 2.1. $M(\mathcal{S}, \mathcal{A}_{\max}, \mathcal{A}_{\min}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$ is called a time inhomogeneous, episodic B-bounded linear mixture MG if there exist H vectors $\boldsymbol{\theta}_h \in \mathbb{R}^d$ satisfying for any $h \in [H]$, $\|\boldsymbol{\theta}_h\|_2 \leq B$, and feature mapping $\boldsymbol{\phi}$ satisfying (2.2), 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 $M_{\boldsymbol{\theta}^*}$.

3. 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. For completeness, we will present the full algorithm as Algorithm 2 in Appendix C, and a turn-based version as Algorithm 3 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. For any function $V : \mathcal{S} \rightarrow \mathbb{R}$ we introduce the shorthands:

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

where V^2 stands for the function whose value at s is $V^2(s)$.

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

Algorithm 1 Nash-UCRL

1: **Input:** Regularization parameter λ, number of episode K, number of horizon H.
2: For any h, $\overline{\boldsymbol{\Sigma}}^{(i)}_{1,h} \leftarrow \underline{\boldsymbol{\Sigma}}^{(i)}_{1,h} \leftarrow \lambda\mathbf{I}$; $\overline{\mathbf{b}}^{(i)}_{1,h} \leftarrow \underline{\mathbf{b}}^{(i)}_{1,h} \leftarrow \mathbf{0}$; $\overline{\boldsymbol{\theta}}^{(i)}_{1,h} \leftarrow \underline{\boldsymbol{\theta}}^{(i)}_{1,h} \leftarrow \mathbf{0}$, for $i \in \{0,1\}$.
3: **for** $k = 1, \ldots, K$ **do**
4: $\quad\overline{V}_{k,H+1}(\cdot) \leftarrow 0$, $\underline{V}_{k,H+1}(\cdot) \leftarrow 0$
5: $\quad$**for** $h = H, \ldots, 1$ **do**
6: $\quad\quad$Set $\overline{Q}_{k,h}(\cdot,\cdot,\cdot) \leftarrow \Big[r_h(\cdot,\cdot,\cdot) + \langle \overline{\boldsymbol{\theta}}^{(0)}_{k,h}, \boldsymbol{\phi}_{\overline{V}_{k,h+1}}(\cdot,\cdot,\cdot)\rangle + \beta^{(0)}_k \big\|\big[\overline{\boldsymbol{\Sigma}}^{(0)}_{k,h}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(\cdot,\cdot,\cdot)\big\|_2\Big]_{[-H,H]}$.
7: $\quad\quad$Set $\underline{Q}_{k,h}(\cdot,\cdot,\cdot) \leftarrow \Big[r_h(\cdot,\cdot,\cdot) + \langle \underline{\boldsymbol{\theta}}^{(0)}_{k,h}, \boldsymbol{\phi}_{\underline{V}_{k,h+1}}(\cdot,\cdot,\cdot)\rangle - \beta^{(0)}_k \big\|\big[\underline{\boldsymbol{\Sigma}}^{(0)}_{k,h}\big]^{-1/2}\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(\cdot,\cdot,\cdot)\big\|_2\Big]_{[-H,H]}$.
8: $\quad\quad$**for** $s \in \mathcal{S}$ **do**
9: $\quad\quad\quad$Let $\mu^k_h(\cdot,\cdot|s) = \epsilon$-CCE$(\overline{Q}_{k,h}(s,\cdot,\cdot), \underline{Q}_{k,h}(s,\cdot,\cdot))$.
10: $\quad\quad\quad\overline{V}_{k,h}(s) = \mathbb{E}_{(a,b)\sim\mu^k_h(\cdot,\cdot|s)}\overline{Q}_{k,h}(s,a,b)$, $\underline{V}_{k,h}(s) = \mathbb{E}_{(a,b)\sim\mu^k_h(\cdot,\cdot|s)}\underline{Q}_{k,h}(s,a,b)$
11: $\quad\quad\quad\pi^k_h(\cdot|s) = \mathcal{P}_{\max}\mu^k_h(\cdot,\cdot|s)$, $\nu^k_h(\cdot|s) = \mathcal{P}_{\min}\mu^k_h(\cdot,\cdot|s)$
12: $\quad\quad$**end for**
13: $\quad$**end for**
14: $\quad$Receives s^k_1
15: $\quad$**for** $h = 1, \ldots, H$ **do**
16: $\quad\quad$Take action $(a^k_h, b^k_h) \sim \mu^k_h(\cdot,\cdot|s^k_h)$ and central controller receives $s^k_{h+1} \sim \mathbb{P}(\cdot|s^k_h, a^k_h, b^k_h)$.
17: $\quad\quad$Set $\overline{E}_{k,h}$ as in (3.3), $\overline{\sigma}_{k,h}$ as in (3.2), and $\underline{E}_{k,h}, \underline{\sigma}_{k,h}$ in similar ways.
18: $\quad\quad$Update $\overline{\boldsymbol{\Sigma}}^{(i)}_{k+1,h}, \overline{\mathbf{b}}^{(i)}_{k+1,h}, \underline{\boldsymbol{\Sigma}}^{(i)}_{k+1,h}, \underline{\mathbf{b}}^{(i)}_{k+1,h}$, $i = 0, 1$. (See Algorithm 2)
19: $\quad\quad$Set $\overline{\boldsymbol{\theta}}^{(i)}_{k+1,h} \leftarrow \big[\overline{\boldsymbol{\Sigma}}^{(i)}_{k+1,h}\big]^{-1}\overline{\mathbf{b}}^{(i)}_{k+1,h}$, $\underline{\boldsymbol{\theta}}^{(i)}_{k+1,h} \leftarrow \big[\underline{\boldsymbol{\Sigma}}^{(i)}_{k+1,h}\big]^{-1}\underline{\mathbf{b}}^{(i)}_{k+1,h}$, $i = 0, 1$
20: $\quad$**end for**
21: **end for**

good accuracy. Inspired by the UCRL with "value-targeted regression" (UCRL-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. To encourage the agent to explore, Nash-UCRL constructs an optimistic action-value function $\overline{Q}_{k,h}$ following the "optimism-in-the-face-of-uncertainty" principle as showed in Line 6 of Algorithm 1. Finally, Nash-UCRL constructs the optimistic value functions $\overline{V}_{k,h}, \underline{V}_{k,h}$ and the policy μ^k_h 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). 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:

Definition 3.1 (Moulin & Vial 1978; Aumann 1987)**.** Given two payoff matrices $Q_{\text{max}}, Q_{\text{min}} \in \mathbb{R}^{|\mathcal{A}_{\text{max}}|\cdot|\mathcal{A}_{\text{min}}|}$, we denote the ϵ-Coarse Correlated Equilibrium (ϵ-CCE) as a joint distribution σ over $\mathcal{A}_{\text{max}}$ and $\mathcal{A}_{\text{min}}$ satisfying that

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

Nash-UCRL computes the distribution $\mu^k_h(\cdot,\cdot|s)$, a ϵ-CCE of $\overline{Q}_{k,h}, \underline{Q}_{k,h}$ for each state s in Line 9. 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^k_h as in Line 10 of Algorithm 1. After obtaining μ^k, Nash-UCRL sets $\pi^k_h(\cdot|s)$ as the marginal distribution for the max-player $\mathcal{P}_{\text{max}}\mu^k_h(\cdot,\cdot|s)$, and sets $\nu^k_h(\cdot|s)$ as the marginal distribution for the min-player $\mathcal{P}_{\text{min}}\mu^k_h(\cdot,\cdot|s)$.

Weighted linear regression for value function estimation. Now we specify how to construct the estimators $\overline{\boldsymbol{\theta}}^{(0)}_{k,h}, \underline{\boldsymbol{\theta}}^{(0)}_{k,h}$. We set the estimator $\overline{\boldsymbol{\theta}}_{k,h}$ as the minimizer to a *weighted ridge regression* problem with square loss over context-target pairs $(\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s^k_h, a^k_h, b^k_h), \overline{V}_{k,h+1}(s^k_{h+1}))$ as follows:

$$\overline{\boldsymbol{\theta}}^{(0)}_{k,h} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}}\, \lambda\|\boldsymbol{\theta}\|_2^2 + \sum_{j=1}^{k-1}\big[\big\langle \boldsymbol{\phi}_{\overline{V}_{j,h+1}}(s^j_h, a^j_h, b^j_h), \boldsymbol{\theta}\big\rangle - \overline{V}_{j,h+1}(s^j_{h+1})\big]^2/\overline{\sigma}^2_{j,h}, \quad (3.1)$$

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

$$\overline{\sigma}_{k,h} = \sqrt{\max\{H^2/d, \mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) + \overline{E}_{k,h}\}}, \tag{3.2}$$

where $[\mathbb{V}_{k,h}^{\text{est}}\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)$ 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 (3.1) has a closed-form solution $\overline{\boldsymbol{\theta}}_{k,h}^{(0)} = \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1}\overline{\mathbf{b}}_{k,h}^{(0)}$ based on the covariance matrix $\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}$ and correlation vector $\overline{\mathbf{b}}_{k,h}^{(0)}$. $\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}$ and $\overline{\mathbf{b}}_{k,h}^{(0)}$ can be updated in the online fashion, and their update rules are deferred to Appendix C. We deferred the construction of the empirical variance term $[\mathbb{V}_{k,h}^{\text{est}}\overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k)$ to Appendix C.

Finally, by the standard self-normalized concentration inequality for vector-valued martingales of 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

$$\overline{E}_{k,h} = \min\big\{H^2, \beta_k^{(1)}\|\big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}^2}(s_h^k, a_h^k, b_h^k)\|_2\big\} + \min\big\{H^2, 2H\beta_k^{(2)}\|\big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\|_2\big\}. \tag{3.3}$$

Here $\beta_k^{(1)}, \beta_k^{(2)}$ together with $\beta_k^{(0)}$ in Line 6 are all constants setting as follows

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

4. Main Results

In this section, we present the main theoretical results. We first present the regret of Nash-UCRL.

Theorem 4.1. 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

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

where $T = KH$.

Theorem 4.1 suggests that when $d \geq H$ and $T \geq d^4H^2$, the regret of Nash-UCRL is bounded by $\widetilde{O}(dH\sqrt{T})$.

Remark 4.2. Our Nash-UCRL also enjoys a finite sample complexity. By the standard online-to-batch conversion (Xie et al., 2020), we can show that Nash-UCRL is guaranteed to find an ϵ-approximate NE, i.e., (π, ν) satisfying $V_1^{*,\nu} - V_1^{\pi,*} \leq \epsilon$, within $\widetilde{O}((d^2H^3 + dH^4)/\epsilon^2)$ episodes.

Here, we present a lower bound for linear mixture MGs. It has been shown in Zhou et al. (2020) 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 4.3 (Regret lower bound, Theorem 5.6 in Zhou et al. 2020)**.** Let $B > 1$ and $K \geq \max\{(d-1)^2H/2, (d-1)/(32H(B-1))\}$, $d \geq 4$, $H \geq 3$. 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:

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

where $T = KH$.

Remark 4.4. When $d \geq H$ and $T \geq d^4H^2$, the regret of Nash-UCRL matches the lower bound up to logarithmic factors. Therefore, Nash-UCRL is nearly minimax optimal.

Remark 4.5. Based on a similar argument made in Zhou et al. (2020), we can show that the same lower bound holds for the Markov games with linear structures studied in Xie et al. (2020). 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 (Xie et al., 2020).

5. Conclusions

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 $\widetilde{O}(\sqrt{d^2H^2T})$ 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 $\widetilde{\Omega}(\sqrt{d^2H^2T})$ lower bound on the regret. Our upper bound matches the lower bound up to logarithmic factors, which suggests the optimality of our algorithm.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

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

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. F. Model-based reinforcement learning with value-targeted regression. *arXiv preprint arXiv:2006.01107*, 2020.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272. JMLR. org, 2017.

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

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

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

Jia, Z., Yang, L., Szepesvari, C., and Wang, M. Model-based reinforcement learning with value-targeted regression. In *L4DC*, 2020.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4868–4878, 2018.

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

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. *arXiv preprint arXiv:2010.01604*, 2020.

Modi, A., Jiang, N., Tewari, A., and Singh, S. Sample complexity of reinforcement learning using linearly combined model ensembles. In *International Conference on Artificial Intelligence and Statistics*, pp. 2010–2020. PMLR, 2020.

Moulin, H. and Vial, J.-P. 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.

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

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. *arXiv preprint arXiv:2002.07066*, 2020.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. *arXiv preprint arXiv:2012.08507*, 2020.

A. Proof of Results in Section 4

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

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

A.1. Proof of Theorem 4.1

Nash-UCRL constructs $\overline{\boldsymbol{\theta}}_{k,h}^{(0)}$ as the estimator of $\boldsymbol{\theta}_h^*$ based on linear regression on $(\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k), \overline{V}_{k,h+1}(s_{h+1}^k))$ (as mentioned in section 3). 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:

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

Here $\left[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\right]^{1/2}$ 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. 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 A.1. Setting $\beta_k^{(0)}$ in (A.1) and $\beta_k^{(1)}, \beta_k^{(2)}$ in (3.3) to

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

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

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

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

Lemma A.2. Denote events $\mathcal{E}_1$ and $\mathcal{E}_2$ as follows

$$
\begin{aligned}
\mathcal{E}_1 = \Big\{\forall h' \in [H], \sum_{k=1}^K \sum_{h=h'}^H \Big[&[\mathbb{P}_h \overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) \\
&- \overline{V}_{k,h+1}(s_{h+1}^k) + \underline{V}_{k,h+1}(s_{h+1}^k)\Big] \le 8H\sqrt{2T \log(H/\delta)}\Big\} \\
\mathcal{E}_2 = \Big\{ \sum_{k=1}^K \sum_{h=1}^H \mathbb{V}_h V_{h+1}^{\mu^k}&(s_h^k, a_h^k, b_h^k) \le 3(HT + H^3 \log(1/\delta))\Big\}.
\end{aligned}
$$

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 A.3. 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 A.4. 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 A.5. 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 4.1.

Proof of Theorem 4.1. By definition of Regret we have that

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

where the first inequality is by Lemma A.3, the second inequality is by the bound of accumulated difference between the UCB and LCB in Lemma A.4, 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 A.1.

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

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

where the first inequality is by Lemma A.5, the second inequality is by Lemma A.4 and the last inequality is due to the choice of $\beta_K^{(0)} = \widetilde{O}(\sqrt{d})$ in Lemma A.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 $x \le a\sqrt{x} + b \Rightarrow x \le 2a^2 + b$, (A.3) suggests that

$$\begin{aligned}
\sum_{k=1}^K \sum_{h=1}^H \overline{\sigma}_{k,h}^2 + \underline{\sigma}_{k,h}^2 &= \widetilde{O}(d^2H^5 + H^2T/d + TH + \sqrt{T}d^{1.5}H^{2.5} + H^3\sqrt{T})\\
&= \widetilde{O}(d^2H^5 + d^4H^3 + TH + H^2T/d), && \text{(A.5)}
\end{aligned}$$

where the inequality holds by $\sqrt{T}d^{1.5}H^{2.5} \le (TH^2/4d + d^4H^3)/2$ and $H^3\sqrt{T} \le (d^2H^5 + H^2T/d)/2$. Plugging (A.5) into (A.2) 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 finishing the proof. □

A.2. Proof of Theorem 4.3

Proof of Theorem 4.3. 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,

$$\text{Regret}(M_{\boldsymbol{\theta}^*}, K) = \sum_{k=1}^K V_1^{*,\nu^k}(s_1^k) - \sum_{k=1}^K V_1^{\pi^k,*}(s_1^k)$$

$$= \sum_{k=1}^{K} \widetilde{V}_1^*(s_1^k) - \sum_{k=1}^{K} \widetilde{V}_1^{\pi^k}(s_1^k).$$

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., 2020)], 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 (2.2) because by the definition of linear mixture MDP in (Zhou et al., 2020), 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 $\widetilde{\boldsymbol{\phi}}_V(s,a) = \sum_{s' \in \mathcal{S}} \widetilde{\boldsymbol{\phi}}(s'|s,a)V(s')$.

□

B. Proof of Lemmas in Appendix A

B.1. Proof of Lemma A.1

For simplicity we denote the following confident sets:

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

By the selection $\beta_k^{(0)} < \beta_k^{(2)}$ in Lemma A.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 (Bernstein inequality in (Zhou et al., 2020))**.** Let $\{\mathcal{G}_t\}_{t=1}^\infty$ be a filtration, $\{\mathbf{x}_t, \eta_t\}_{t\geq 1}$ 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 $t \geq 1$ let $y_t = \langle \boldsymbol{\mu}^*, \mathbf{x}_t \rangle + \eta_t$ and suppose that $\eta_t, \mathbf{x}_t$ also satisfy

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

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

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

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

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

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

Proof. For simplicity, we only prove the results for the max-player.

By the triangle inequality we have that

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

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

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

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

$$
I_1 \le \min\Big\{H^2, \Big\|\big[\overline{\mathbf{\Sigma}}_{k,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k,a_h^k,b_h^k)\Big\|_2\Big\|\big[\overline{\mathbf{\Sigma}}_{k,h}^{(1)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(1)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\}, \qquad \text{(B.3)}
$$

For the term I_2,

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

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 $I_2 \le H^2$, so we have that

$$
I_2 \le \min\Big\{H^2, 2H\Big\|\mathbf{\Sigma}_{k,h}^{(0)-1/2}\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s_h^k,a_h^k,b_h^k)\Big\|_2\Big\|\big[\overline{\mathbf{\Sigma}}_{k,h}^{(0)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\}. \qquad \text{(B.4)}
$$

Plugging (B.4) and (B.3) into (B.2) gets

$$
|\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s_h^k,a_h^k,b_h^k) - \mathbb{V}\overline{V}_{k,h+1}(s_h^k,a_h^k,b_h^k)|
$$

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

□

Now we present the Proof of Lemma A.1.

Proof of Lemma A.1. For simplicity, we only prove the results for the max-player. Fix $h \in [H]$.

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

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

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 $k \le K$, $\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*)\Big\|_2$ is bounded by

$$16d\sqrt{\log(1+k/\lambda)\log(8k^2H/\delta)} + 8\sqrt{d}\log(8k^2H/\delta) + \sqrt{\lambda}B = \beta_k^{(2)}, \tag{B.5}$$

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

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 $\Big\|\big[\boldsymbol{\Sigma}_{k,h}^{(1)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(1)} - \boldsymbol{\theta}_h^*)\Big\|_2$ bounded by

$$16\sqrt{dH^4\log(1+kH^4/(d\lambda))\log(8k^2H/\delta)} + 8H^2\log(8k^2H/\delta) + \sqrt{\lambda}B = \beta_k^{(1)}, \tag{B.6}$$

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

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

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

$\mathcal{G}_i = \mathcal{F}_{i,h}$, $\boldsymbol{\mu}^* = \boldsymbol{\theta}_h^*$, $y_i = \langle\boldsymbol{\mu}^*, \mathbf{x}_i\rangle + \eta_i$, $\mathbf{Z}_i = \lambda\mathbf{I} + \sum_{i'=1}^i \mathbf{x}_{i'}\mathbf{x}_{i'}^\top$, $\mathbf{b}_i = \sum_{i'=1}^i \mathbf{x}_{i'}y_{i'}$ and $\boldsymbol{\mu}_i = \mathbf{Z}_i^{-1}\mathbf{b}_i$. Still we have that $\|\mathbf{x}_i\|_2 \le \overline{\sigma}_{i,h}^{-1}H \le \sqrt{d}$.Because $\mathbb{1}\{\boldsymbol{\theta}_h^* \in \overline{\mathcal{C}}_{i,h}^{(1)} \cap \overline{\mathcal{C}}_{i,h}^{(2)}\}$ is $\mathcal{G}_i$-measurable, we have $\mathbb{E}[\eta_i|\mathcal{G}_i] = 0$. We also have $|\eta_i| \le \overline{\sigma}_{i,h}^{-1}2H \le 2\sqrt{d}$ since $|\overline{V}_{i,h+1}(\cdot)| \le H$ and $\overline{\sigma}_{i,h} \ge H/\sqrt{d}$. To get better bound $\beta_k^{(0)}$ rather than $\beta_k^{(2)}$ in (B.5), we need more careful computation of $\mathbb{E}[\eta_i^2|\mathcal{G}_i]$ as follows,

$$\mathbb{E}[\eta_i^2|\mathcal{G}_i] = \overline{\sigma}_{i,h}^{-2}\,\mathbb{1}\{\boldsymbol{\theta}_h^* \in \overline{\mathcal{C}}_{i,h}^{(1)} \cap \overline{\mathcal{C}}_{i,h}^{(2)}\}[\mathbb{V}_h\overline{V}_{i,h+1}](s_h^i, a_h^i, b_h^i)$$
$$\le \overline{\sigma}_{i,h}^{-2}\,\mathbb{1}\{\boldsymbol{\theta}_h^* \in \overline{\mathcal{C}}_{i,h}^{(1)} \cap \overline{\mathcal{C}}_{i,h}^{(2)}\}\Bigg[[\mathbb{V}_{i,h}^{\text{est}}\overline{V}_{i,h+1}](s_h^i, a_h^i, b_h^i)$$
$$+ \min\Big\{H^2, \Big\|\big[\boldsymbol{\Sigma}_{i,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{i,h+1}^2}(s_h^i, a_h^i, b_h^i)\Big\|_2\Big\|\big[\boldsymbol{\Sigma}_{i,h}^{(1)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{i,h}^{(1)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\}$$
$$+ \min\Big\{H^2, 2H\Big\|\big[\boldsymbol{\Sigma}_{i,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{i,h+1}}(s_h^i, a_h^i, b_h^i)\Big\|_2\Big\|\big[\boldsymbol{\Sigma}_{i,h}^{(0)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{i,h}^{(0)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\}\Bigg]$$
$$\le \overline{\sigma}_{i,h}^{-2}\Bigg[[\mathbb{V}_{i,h}^{\text{est}}\overline{V}_{i,h+1}](s_h^i, a_h^i, b_h^i) + \min\Big\{H^2, \beta_i^{(1)}\Big\|\big[\boldsymbol{\Sigma}_{i,h}^{(1)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{i,h+1}^2}(s_h^i, a_h^i, b_h^i)\Big\|_2\Big\}$$

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

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 $1-\delta/(2H)$, $\forall k \le K$,

$$
\|\boldsymbol{\mu}_k - \boldsymbol{\mu}^*\|_{\mathbf{Z}_i} \le 16\sqrt{d\log(1+k/\lambda)\log(8k^2H/\delta)} + 8\sqrt{d}\log(8k^2H/\delta) + \sqrt{\lambda}B = \beta_k^{(0)}, \tag{B.7}
$$

where the equality uses the definition of $\beta_k^{(0)}$. Let $\mathcal{E}'$ be the event when $\boldsymbol{\theta}_h^* \in \cap_{k\le K}\overline{\mathcal{C}}_{k,h}^{(1)} \cap \overline{\mathcal{C}}_{k,h}^{(2)}$ and (B.7) hold. By the union bound, $\mathbb{P}(\mathcal{E}') \ge 1 - 3\delta/(2H)$.

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 $k \le K$, $\boldsymbol{\mu}_k = \overline{\boldsymbol{\theta}}_{k+1,h}^{(0)}$ and for any $i \le K$,

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

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

$$
\boldsymbol{\theta}_h^* \in \cap_{k\le K}\overline{\mathcal{C}}_{k,h}^{(1)} \cap \overline{\mathcal{C}}_{k,h}^{(2)} \tag{B.8}
$$

To finish our proof, it is thus sufficient to show that on the event when (B.8) holds, it also holds that

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

However, by the definition of $\overline{E}_{k,h}$, this is immediate from substituting (B.5), (B.6) into Lemma B.2. □

B.2. Proof of Lemma A.2

We first present the Azuma-Hoeffding inequality:

Lemma B.3 (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 of Lemma A.2. To prove $\mathbb{P}(\mathcal{E}_1) \ge 1-\delta$, we apply the Azuma-Hoeffding inequality (Lemma B.3). 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

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

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). □

B.3. Proof of Lemma A.3

Following Lemma directly from the definition of ϵ-CCE,

Lemma B.4. For each (k,h,s), $\mu_h^k(\cdot,\cdot|s), \pi_h^k(\cdot|s), \nu_h^k(\cdot|s)$ satisfy that

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

Proof of Lemma A.3. For simplicity, we only prove the following UCB by induction,

$$Q_h^{*,\nu^k}(s,a,b) \leq \overline{Q}_{k,h}(s,a,b) + (H-h+1)\epsilon, V_h^{*,\nu^k}(s) \leq \overline{V}_{k,h}(s) + (H-h+2)\epsilon. \tag{B.9}$$

The base case $h = H+1$ holds trivially since the terminal cost is zero. Now we assume that the bounds (B.9) holds for step $h+1$. That is,

$$Q_{h+1}^{*,\nu^k}(s,a,b) \leq \overline{Q}_{k,h+1}(s,a,b) + (H-h)\epsilon, V_{h+1}^{*,\nu^k}(s) \leq \overline{V}_{k,h+1}(s) + (H-h+1)\epsilon. \tag{B.10}$$

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

$$\begin{aligned}
&\overline{Q}_{k,h}(s,a,b) - Q_h^{*,\nu^k}(s,a,b) \\
&= \langle \overline{\boldsymbol{\theta}}_{k,h}^{(0)}, \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \rangle + \beta_k^{(0)} \Big\| \big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 - \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \rangle \\
&\quad + \mathbb{P}_h \overline{V}_{k,h+1}(s,a,b) - \mathbb{P}_h V_{h+1}^{*,\nu^k}(s) \\
&\geq \beta_k^{(0)} \Big\| \big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 - \Big\| \big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{1/2} (\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*) \Big\|_2 \Big\| \big[\boldsymbol{\Sigma}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} \Big\|_2 \\
&\quad + \mathbb{P}_h \overline{V}_{k,h+1}(s) - \mathbb{P}_h V_{h+1}^{*,\nu^k}(s) \\
&\geq \mathbb{P}_h \overline{V}_{k,h+1}(s) - \mathbb{P}_h V_{h+1}^{*,\nu^k}(s) \\
&\geq -(H-h+1)\epsilon,
\end{aligned} \tag{B.11}$$

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 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

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

Then we have that

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

where the the first equality is by the property of ϵ-CCE in Lemma B.4, the second inequality is by (B.11), the last inequality is due to the Bellman equation.Therefore, our proof ends. □

B.4. Proof of Lemma A.4

Proof of Lemma A.4.

$$\overline{V}_{k,h}(s_h^k) - \underline{V}_{k,h}(s_h^k)$$

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

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

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

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,

$$
\begin{aligned}
&\overline{V}_{k,h}(s_h^k) - \underline{V}_{k,h}(s_h^k) - [\overline{V}_{k,h+1}(s_h^k) - \underline{V}_{k,h+1}(s_h^k)] \\
&\le 2\beta_k^{(0)} \overline{\sigma}_{k,h} \min\Big\{1, \Big\| \big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\overline{V}_{k,h+1}} / \overline{\sigma}_{k,h} \Big\|_2 \Big\} \\
&\quad + 2\beta_k^{(0)} \underline{\sigma}_{k,h} \min\Big\{1, \Big\| \big[\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2} \boldsymbol{\phi}_{\underline{V}_{k,h+1}} / \underline{\sigma}_{k,h} \Big\|_2 \Big\} \\
&\quad + [\mathbb{P}_h \overline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) - [\mathbb{P}_h \underline{V}_{k,h+1}](s_h^k, a_h^k, b_h^k) \\
&\quad - [\overline{V}_{k,h+1}(s_{h+1}^k) - \underline{V}_{k,h+1}(s_{h+1}^k)],
\end{aligned}
\tag{B.12}
$$

Taking summation of (B.12) from $k = 1 \dots K$ and $h = h' \dots H$, we have following inequality holds

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

$$
\begin{aligned}
&+ 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\} \\
&+ \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] \\
\leq\ & 2\beta_k^{(0)} \sum_{k=1}^{K} \sum_{h=1}^{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\} \\
&+ 2\beta_k^{(0)} \sum_{k=1}^{K} \sum_{h=1}^{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\} + 8H\sqrt{2T\log(H/\delta)} \\
\leq\ & 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{\boldsymbol{\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{\boldsymbol{\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)} \\
\leq\ & 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)} \\
\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{2Hd\log(1+K/\lambda)} + 8H\sqrt{2T\log(H/\delta)}, \qquad \text{(B.13)}
\end{aligned}
$$

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 $\big\|\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)/\overline{\sigma}_{k,h}\big\|_2 \leq \big\|\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\big\|_2 \cdot \sqrt{d}/H \leq \sqrt{d}$, $\big\|\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)/\underline{\sigma}_{k,h}\big\|_2 \leq \big\|\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s_h^k, a_h^k, b_h^k)\big\|_2 \cdot \sqrt{d}/H \leq \sqrt{d}$, the last inequality is by the fact that $\sqrt{a} + \sqrt{b} \leq 2\sqrt{a+b}$. (B.13) holds for any h', then we have following inequality holds

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

where the inequality holds due to (B.13) and on event $\mathcal{E}_1$. □

B.5. Proof of Lemma A.5

To estimate the variance in weighted ridge regression we need the following lemma, which is similar to the Lemma A.3 but without tolerant error ϵ.

Lemma B.5. Suppose the event $\mathcal{E}$ hold. Then we have for any s,a,b,k,h following inequalities hold,

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

and

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

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

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

The base case $h = H+1$ holds trivially since the terminal cost is zero. Now we assume that the bounds (B.9) holds for step $h+1$. That is,

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

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

$$\begin{aligned}
&\overline{Q}_{k,h}(s,a,b) - Q_h^{\mu^k}(s,a,b)\\
&= \langle \overline{\boldsymbol{\theta}}_{k,h}^{(0)}, \boldsymbol{\phi}_{\overline{V}_{k,h+1}}\rangle + \beta_k^{(0)}\Big\|\big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}\Big\|_2 - \langle \boldsymbol{\theta}_h^*, \boldsymbol{\phi}_{\overline{V}_{k,h+1}}\rangle\\
&\qquad + \mathbb{P}_h\overline{V}_{k,h+1}(s,a,b) - \mathbb{P}_h V_{h+1}^{*,\nu^k}(s)\\
&\ge \beta_k^{(0)}\Big\|\big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}\Big\|_2 - \Big\|\big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{1/2}(\overline{\boldsymbol{\theta}}_{k,h}^{(0)} - \boldsymbol{\theta}_h^*)\Big\|_2\Big\|\big[\overline{\boldsymbol{\Sigma}}_{k,h}^{(0)}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}\Big\|_2\\
&\qquad + \mathbb{P}_h\overline{V}_{k,h+1}(s) - \mathbb{P}_h V_{h+1}^{\mu^k}(s)\\
&\ge \mathbb{P}_h\overline{V}_{k,h+1}(s) - \mathbb{P}_h V_{h+1}^{\mu^k}(s)\\
&\ge 0,
\end{aligned} \tag{B.14}$$

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

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

where the inequality is by (B.14), the last inequality is due to the Bellman equation. Therefore, our proof ends. □

Lemma B.6. (Lemma 11 in (Abbasi-Yadkori et al., 2011)). For any $\{\mathbf{x}_t\}_{t=1}^T \subset \mathbb{R}^d$ satisfying that $\|\mathbf{x}_t\|_2 \le L$, 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

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

Proof of Lemma A.5. Suppose the event in Lemma 9.1 holds, we have the following results:

$$\sum_{k=1}^K\sum_{h=1}^H \overline{\sigma}_{k,h}^2 = \sum_{k=1}^K\sum_{h=1}^H \big[H^2/d + \mathbb{V}_{k,h}^{\text{est}}\overline{V}_{k,h+1}(s_h^k, a_h^k, b_h^k) + \overline{E}_{k,h}\big]$$

$$
\begin{aligned}
&= H^2T/d + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\big[\mathbb{V}_h\overline{V}_{k,h+1}(s_h^k,a_h^k,b_h^k) - \mathbb{V}_h V_{h+1}^{\mu^k}(s_h^k,a_h^k,b_h^k)\big]}_{I_1} \\
&\quad + \underbrace{2\sum_{k=1}^{K}\sum_{h=1}^{H}\overline{E}_{k,h}}_{I_2} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}_h V_{h+1}^{\mu^k}(s_h^k,a_h^k,b_h^k)}_{I_3} \\
&\quad + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\big[\mathbb{V}_h^{\text{est}}\overline{V}_{k,h+1}(s_h^k,a_h^k,b_h^k) - \mathbb{V}_h\overline{V}_{k,h+1}(s_h^k,a_h^k,b_h^k) - \overline{E}_{k,h}\big]}_{I_4}, \qquad \text{(B.15)}
\end{aligned}
$$

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

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

where the first inequality is by $|\overline{V}_{k,h+1}|, |V_{h+1}^{\mu^k}| \le H$, and the second inequality is by $\overline{V}_{k,h+1} - V_{h+1}^{\mu^k} \ge 0$ due to Lemma B.5. To bound I_2, we have

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

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

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

the third inequality holds due to Lemma B.6. Next we bound I_3, since event $\mathcal{E}_2$ holds , we have

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

Finally, due to on event $\mathcal{E}$, we have $I_4 \le 0$. We finish the proof by substituting I_1, I_2, I_3, I_4 into (B.15). □

C. Full Version of Algorithm 1

In this section, we present the full version of Algorithm 1 in Algorithm 2.

Algorithm 2 Nash-UCRL

Input: Regularization parameter λ, Number of episode K, number of horizon H.
For any h, $\overline{\boldsymbol{\Sigma}}^{(i)}_{1,h} \leftarrow \underline{\boldsymbol{\Sigma}}^{(i)}_{1,h} \leftarrow \lambda\mathbf{I}$; $\overline{\mathbf{b}}^{(i)}_{1,h} \leftarrow \underline{\mathbf{b}}^{(i)}_{1,h} \leftarrow \mathbf{0}$; $\overline{\boldsymbol{\theta}}^{(i)}_{1,h} \leftarrow \underline{\boldsymbol{\theta}}^{(i)}_{1,h} \leftarrow \mathbf{0}$, 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 (C.1).
    for $s \in \mathcal{S}$ **do**
      Let $\mu^k_h(\cdot,\cdot|s) = \epsilon\text{-CCE}(\overline{Q}_{k,h}(s,\cdot,\cdot), \underline{Q}_{k,h}(s,\cdot,\cdot))$.
      $\overline{V}_{k,h}(s) = \mathbb{E}_{(a,b)\sim\mu^k_h(\cdot,\cdot|s)}\overline{Q}_{k,h}(s,a,b)$, $\underline{V}_{k,h}(s) = \mathbb{E}_{(a,b)\sim\mu^k_h(\cdot,\cdot|s)}\underline{Q}_{k,h}(s,a,b)$
      $\pi^k_h(\cdot|s) = \mathcal{P}_{\max}\mu^k_h(\cdot,\cdot|s)$, $\nu^k_h(\cdot|s) = \mathcal{P}_{\min}\mu^k_h(\cdot,\cdot|s)$
    end for
  end for
  receives s^k_1
  for $h = 1, \ldots, H$ **do**
    Take action $a^k_h \sim \pi^k_h(s^k_h)$ and $b^k_h \sim \nu^k_h(s^k_h)$ and receives $s^k_{h+1} \sim \mathbb{P}(\cdot|s^k_h, a^k_h, b^k_h)$.
    Set $\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s^k_h, a^k_h, b^k_h)$ and $\mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s^k_h, a^k_h, b^k_h)$ as in (C.2).
    Set $\overline{E}_{k,h}, \underline{E}_{k,h}, \overline{\boldsymbol{\sigma}}_{k,h}, \underline{\boldsymbol{\sigma}}_{k,h}, \overline{\boldsymbol{\Sigma}}^{(0)}_{k+1,h}, \underline{\boldsymbol{\Sigma}}^{(0)}_{k+1,h}, \overline{\mathbf{b}}^{(0)}_{k+1,h}, \underline{\mathbf{b}}^{(0)}_{k+1,h}, \overline{\boldsymbol{\Sigma}}^{(1)}_{k+1,h}, \underline{\boldsymbol{\Sigma}}^{(1)}_{k+1,h}, \overline{\mathbf{b}}^{(1)}_{k+1,h}, \underline{\mathbf{b}}^{(1)}_{k+1,h}$ as defined in (C.3).
    Set $\overline{\boldsymbol{\theta}}^{(i)}_{k+1,h} \leftarrow \big[\overline{\boldsymbol{\Sigma}}^{(i)}_{k+1,h}\big]^{-1}\overline{\mathbf{b}}^{(i)}_{k+1,h}$, $\underline{\boldsymbol{\theta}}^{(i)}_{k+1,h} \leftarrow \big[\underline{\boldsymbol{\Sigma}}^{(i)}_{k+1,h}\big]^{-1}\underline{\mathbf{b}}^{(i)}_{k+1,h}$, $i = 0, 1$
  end for
end for

Update of optimistic action-value function:

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

Update of variance estimation:

$$\begin{aligned}
\mathbb{V}^{\text{est}}\overline{V}_{k,h+1}(s^k_h, a^k_h, b^k_h) &\leftarrow \big[\langle\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s^k_h, a^k_h, b^k_h), \overline{\boldsymbol{\theta}}^{(1)}_{k,h}\rangle\big]_{[0,H^2]} - \big[\langle\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s^k_h, a^k_h, b^k_h), \overline{\boldsymbol{\theta}}^{(0)}_{k,h}\rangle\big]^2_{[-H,H]}, \\
\mathbb{V}^{\text{est}}\underline{V}_{k,h+1}(s^k_h, a^k_h, b^k_h) &\leftarrow \big[\langle\boldsymbol{\phi}_{\underline{V}^2_{k,h+1}}(s^k_h, a^k_h, b^k_h), \underline{\boldsymbol{\theta}}^{(1)}_{k,h}\rangle\big]_{[0,H^2]} - \big[\langle\boldsymbol{\phi}_{\underline{V}_{k,h+1}}(s^k_h, a^k_h, b^k_h), \underline{\boldsymbol{\theta}}^{(0)}_{k,h}\rangle\big]^2_{[-H,H]}.
\end{aligned} \tag{C.2}$$

Update of other parameters:

$$\begin{aligned}
\overline{E}_{k,h} &= \min\big\{H^2, \beta^{(1)}_k\Big\|\big[\overline{\boldsymbol{\Sigma}}^{(1)}_{k,h}\big]^{-1/2}\boldsymbol{\phi}_{\overline{V}^2_{k,h+1}}(s^k_h, a^k_h, b^k_h)\Big\|_2\big\} \\
&\quad + \min\big\{H^2, 2H\beta^{(2)}_k\Big\|\overline{\boldsymbol{\Sigma}}^{(0)-1/2}_{k,h}\boldsymbol{\phi}_{\overline{V}_{k,h+1}}(s^k_h, a^k_h, b^k_h)\Big\|_2\big\} \\
\underline{E}_{k,h} &= \min\big\{H^2, \beta^{(1)}_k\Big\|\big[\underline{\boldsymbol{\Sigma}}^{(1)}_{k,h}\big]^{-1/2}\boldsymbol{\phi}_{\underline{V}^2_{k,h+1}}(s^k_h, a^k_h, b^k_h)\Big\|_2\big\}
\end{aligned}$$

$$
\begin{aligned}
&+ \min\big\{H^2, 2H\beta_k^{(2)}\Big\|\underline{\boldsymbol{\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{\boldsymbol{\Sigma}}_{k+1,h}^{(0)} &\leftarrow \overline{\boldsymbol{\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{\boldsymbol{\Sigma}}_{k+1,h}^{(0)} &\leftarrow \underline{\boldsymbol{\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),\\
\underline{\mathbf{b}}_{k+1,h}^{(0)} &= \underline{\mathbf{b}}_{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)\underline{V}_{k,h+1}(s_{h+1}^k),\\
\overline{\boldsymbol{\Sigma}}_{k+1,h}^{(1)} &\leftarrow \overline{\boldsymbol{\Sigma}}_{k,h}^{(1)} + \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}}_{k+1,h}^{(1)} &\leftarrow \underline{\boldsymbol{\Sigma}}_{k,h}^{(1)} + \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}}_{k+1,h}^{(1)} &= \overline{\mathbf{b}}_{k,h}^{(1)} + \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}}_{k+1,h}^{(1)} &= \underline{\mathbf{b}}_{k,h}^{(1)} + \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{(C.3)}
\end{aligned}
$$

D. 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 $\mathcal{S}=\mathcal{S}_{\text{max}}\cup\mathcal{S}_{\text{min}}$, $\mathcal{S}_{\text{max}}$ ($\mathcal{S}_{\text{min}}$) are the states where the max (min)-player plays, $\mathcal{S}_{\text{max}}\cap\mathcal{S}_{\text{min}}=\emptyset$. 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 ($s\in\mathcal{S}_{\text{max}}$ or $\mathcal{S}_{\text{max}}$) 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 D.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

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

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 9 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.

By Theorem 4.1, we immediately have that the regret of our turn-based algorithm 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.

Instead of dividing by state (Cui & Yang, 2020; Xie et al., 2020), some papers (Bai & Jin, 2020) consider dividing the steps $[H]$ into $\mathcal{H}_{\max}$ and $\mathcal{H}_{\min}$. This setting looks different at first glance but can actually be written as the turn-based MGs in this section by augmenting the state s by a latent variable h, i.e., $\widetilde{s} = (s, h)$, to indicate the max-player's step and the min-player's step.

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 (D.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 $s_{k,1}$
  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 (D.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 (D.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:

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

Update of variance estimation:

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

Update of other parameters:

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

$$
\begin{aligned}
&\quad + \min\big\{H^2, 2H\beta_k^{(2)}\Big\|\underline{\boldsymbol{\Sigma}}_{k,h}^{(0)-1/2}\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}^{\mathrm{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}^{\mathrm{est}}\underline{V}_{k,h+1}(s_h^k, a_h^k) + \underline{E}_{k,h}\}},\\
\overline{\boldsymbol{\Sigma}}_{k+1,h}^{(0)} &\leftarrow \overline{\boldsymbol{\Sigma}}_{k,h}^{(0)} + \overline{\boldsymbol{\sigma}}_{k,h}^{-2}\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}}_{k+1,h}^{(0)} &\leftarrow \underline{\boldsymbol{\Sigma}}_{k,h}^{(0)} + \underline{\boldsymbol{\sigma}}_{k,h}^{-2}\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}}_{k+1,h}^{(0)} &= \overline{\mathbf{b}}_{k,h}^{(0)} + \overline{\boldsymbol{\sigma}}_{k,h}^{-2}\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}}_{k+1,h}^{(0)} &= \underline{\mathbf{b}}_{k,h}^{(0)} + \underline{\boldsymbol{\sigma}}_{k,h}^{-2}\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}}_{k+1,h}^{(1)} &\leftarrow \overline{\boldsymbol{\Sigma}}_{k,h}^{(1)} + \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}}_{k+1,h}^{(1)} &\leftarrow \underline{\boldsymbol{\Sigma}}_{k,h}^{(1)} + \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}}_{k+1,h}^{(1)} &= \overline{\mathbf{b}}_{k,h}^{(1)} + \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}}_{k+1,h}^{(1)} &= \underline{\mathbf{b}}_{k,h}^{(1)} + \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).
\end{aligned} \tag{D.3}
$$