# Contrastive UCB: Provably Efficient Contrastive Self-Supervised Learning in Online Reinforcement Learning

**Shuang Qiu** [1] **Lingxiao Wang** [2] **Chenjia Bai** [3] **Zhuoran Yang** [4] **Zhaoran Wang** [2]

## Abstract

In view of its power in extracting feature representation, contrastive self-supervised learning has been successfully integrated into the practice of (deep) reinforcement learning (RL), leading to efficient policy learning on various applications. Despite its tremendous empirical successes, the understanding of contrastive learning for RL remains elusive. To narrow such a gap, we study contrastive-learning empowered RL for a class of Markov decision processes (MDPs) and Markov games (MGs) with low-rank transitions. For both models, we propose to extract the correct feature representations of the low-rank model by minimizing a contrastive loss. Moreover, under the online setting, we propose novel upper confidence bound (UCB)-type algorithms that incorporate such a contrastive loss with online RL algorithms for MDPs or MGs. We further theoretically prove that our algorithm recovers the true representations and simultaneously achieves sample efficiency in learning the optimal policy and Nash equilibrium in MDPs and MGs. We also provide empirical studies to demonstrate the efficacy of the UCB-based contrastive learning method for RL. To the best of our knowledge, we provide the first provably efficient online RL algorithm that incorporates contrastive learning for representation learning.

## 1. Introduction

Deep reinforcement learning (DRL) has achieved great empirical successes in various real-world decision-making problems (e.g., Mnih et al. (2015); Silver et al. (2016; 2017); Sallab et al. (2017); Sutton & Barto (2018); Silver et al. (2018); Vinyals et al. (2019)). A key to the success of DRL is the superior representation power of the neural networks, which extracts the effective information from raw input pixel states. Nevertheless, learning such effective representation of states typically demands millions of interactions with the environment, which limits the usefulness of RL algorithms in domains where the interaction with environments is expensive or prohibitive, such as healthcare (Yu et al., 2021) and autonomous driving (Kiran et al., 2021).

To improve the sample efficiency of RL algorithms, recent works propose to learn low-dimensional representations of the states via solving auxiliary problems (Jaderberg et al., 2016; Hafner et al., 2019a;b; Gelada et al., 2019; François-Lavet et al., 2019; Bellemare et al., 2019; Srinivas et al., 2020; Zhang et al., 2020; Liu et al., 2021; Yang & Nachum, 2021; Stooke et al., 2021). Among the recent breakthroughs in representation learning for RL, contrastive self-supervised learning gains popularity for its superior empirical performance (Oord et al., 2018b; Sermanet et al., 2018; Dwibedi et al., 2018; Anand et al., 2019; Schwarzer et al., 2020; Srinivas et al., 2020; Liu et al., 2021). A typical paradigm for such contrastive RL is to construct an auxiliary contrastive loss for representation learning, add it to the loss function in RL, and deploy an RL algorithm with the learned representation being the state and action input. However, the theoretical underpinnings of such an enterprise remain elusive. To summarize, we raise the following question:

*Can contrastive self-supervised learning provably improve the sample efficiency of RL via representation learning?*

To answer such a question, we face two challenges. Firstly, how to integrate contrastive self-supervised learning into the provably efficient online exploration strategies, such as exploration with the upper confidence bound (UCB), remains unknown. Secondly, how to analyze the sample complexity of such integration between self-supervised learning and RL remains unknown. This paper takes an initial step towards tackling such challenges based on an instantiation of the contrastive self-supervised learning empowered RL. Concretely, we first investigate the sample complexity of the online single-agent RL problem under the low-rank MDP

[1]University of Chicago. [2]Northwestern University. [3]Shanghai AI Laboratory. [4]Yale University. Correspondence to: Shuang Qiu <qiush@umich.edu>, Lingxiao Wang <lingxiaowang2022@u.northwestern.edu>, Chenjia Bai <baichenjia255@gmail.com>.

*Proceedings of the $39^{th}$ International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022. 

setting, where the representations are learned via the temporal contrastive self-supervised learning similar to Oord et al. (2018b); Sermanet et al. (2018). The algorithm we propose iteratively solves a temporal contrastive loss to obtain the state-action representations and then constructs a UCB bonus with such representations to conduct exploration in a provably efficient way. Our theory shows that the proposed algorithm can recover the true representations under the low-rank MDP setting. Moreover, our proposed algorithm achieves a $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity for attaining the $\varepsilon$-approximate optimal value function, where the notation $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors. Furthermore, we extend our proposed algorithm to the zero-sum MG under the low-rank setting, a multi-agent extension of MDPs in a competitive environment. We construct upper and lower confidence bounds (ULCB) for such a competitive RL setting and prove that the proposed approach achieves an $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ sample complexity to attain an $\varepsilon$-approximate Nash equilibrium. To the best of our knowledge, we propose the first provably efficient online RL algorithms that employ contrastive learning for representation learning. Our major contributions are summarized as follows:

**Contribution.** Our contributions are three-fold. First, We show that contrastive self-supervised learning recovers the underlying true transition dynamics, which reveals the benefit of incorporating representation learning into RL in a provable way. Second, we propose the first provably efficient exploration strategy incorporated with contrastive self-supervised learning. Our proposed UCB-based method is readily adapted to existing representation learning methods for RL, which then demonstrates improvements over the previous empirical results as shown in our experiments. Finally, we extend our findings to the zero-sum MG, which reveals a potential direction of utilizing the contrastive self-supervised learning for RL.

**Related Work.** Our work is closely related to the line of research on RL with low-rank transition kernels, which assumes that the transition dynamics take the form of an inner product of two unknown feature vectors for the current state-action pair and the next state (see Assumption 2.1 for details) (Jiang et al., 2017; Agarwal et al., 2020; Uehara et al., 2021). In contrast, as a special case of the low-rank model, linear MDPs have a similar form of structures but with an extra assumption that the linear representation is known a priori (Du et al., 2019b; Yang & Wang, 2019; Jin et al., 2020; Xie et al., 2020; Ayoub et al., 2020; Cai et al., 2020; Yang & Wang, 2020; Chen et al.; Zhou et al., 2021a;b). Our work focuses on the more challenging low-rank setting and aims to recover the unknown state-action representation via contrastive self-supervised learning. Our theory is motivated by the recent progress in low-rank MDPs (Agarwal et al., 2020; Uehara et al., 2021), which show that the transition dynamics can be effectively recovered via maximum likelihood estimation (MLE). In contrast, our work recovers the representation via contrastive self-supervised learning. Upon acceptance of our work, we notice a concurrent work (Zhang et al., 2022) studies contrastive learning in RL on linear MDPs.

There is a large amount of literature studying contrastive learning in RL empirically. To improve the sample efficiency of RL, previous empirical works leverages different types of information for representation learning, e.g., temporal information (Sermanet et al., 2018; Dwibedi et al., 2018; Oord et al., 2018b; Anand et al., 2019; Schwarzer et al., 2020), local spatial structure(Anand et al., 2019), image augmentation(Srinivas et al., 2020), and return feedback(Liu et al., 2021). Our work follows the utilization of contrastive learning for RL to extract temporal information. Similar to our work, recent work by Misra et al. (2020) shows that contrastive learning provably recovers the latent embedding under the restrictive Block MDP setting (Du et al., 2019a). In contrast, our work analyzes contrastive learning in RL under the more general low-rank setting, which includes Block MDP as a special case (Agarwal et al., 2020) for both MDPs and MGs.

## 2. Preliminaries

In this section, we introduce the backgrounds of single-agent MDPs, zero-sum MGs, and the low-rank assumption.

**Single-Agent MDP.** An episodic single-agent MDP is defined by $(\mathcal{S}, \mathcal{A}, H, r, \mathbb{P})$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $H$ is the length of an episode, $r = \{r_h\}_{h=1}^H$ is the reward function with $r_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$, and $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ denotes the transition model with $\mathbb{P}_h(s'|s, a)$ being the probability density of an agent transitioning to $s' \in \mathcal{S}$ from state $s \in \mathcal{S}$ after taking action $a \in \mathcal{A}$ at the step $h$. Specifically, $\mathcal{S}$ can be an infinite state space[1] and the action space $\mathcal{A}$ is assumed to be finite with the size of $|\mathcal{A}|$. A deterministic policy is denoted as $\pi = \{\pi_h\}_{h=1}^H$ where $\pi_h : \mathcal{S} \mapsto \mathcal{A}$ is the map from the agent's state $s$ to an action $a$ at the $h$-th step. We further denote the policy learned at the $k$-th episode by $\pi^k = \{\pi_h^k\}_{h=1}^H$. For simplicity, assume the initial state is fixed as $s_1^k = s_1$ for any episode $k$.

For the single-agent MDP, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define the associated Q-function and value function as $Q_h^\pi(s, a) = \mathbb{E}[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \,|\, s_h = s, a_h = a, \pi, \mathbb{P}]$ and $V_h^\pi(s) = \mathbb{E}[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \,|\, s_h = s, \pi, \mathbb{P}]$. Then, we further have the Bellman equation as $Q_h^\pi(s, a) = r_h(s, a) + \mathbb{P}_h V_{h+1}^\pi(s, a)$ and $V_h^\pi(s) = Q_h^\pi(s, \pi_h(s))$ where, for the ease of notation, we denote $\mathbb{P}_h V(s, a) = \int_{s'} \mathbb{P}_h(s'|s, a) V(s') \mathrm{d}s'$ for any value function $V$. Moreover,

[1] We assume that the volume (Lebesgue measure) of the infinite state space $\mathcal{S}$ satisfies $\mathrm{Vol}(\mathcal{S}) \leq c$, where $\mathrm{Vol}(\cdot)$ denotes the volume of a space. WOLG, we let $c = 1$ for simplicity.

we define the *optimal policy* as $\pi^* := \operatorname{argmax}_\pi V_1^\pi(s_1)$. We say a policy $\pi$ is an $\varepsilon$-*suboptimal policy* if

$$V_1^{\pi^*}(s_1) - V_1^\pi(s_1) \le \varepsilon.$$

**Zero-Sum Markov Game.** Our work further studies the zero-sum two-player Markov game that can be defined by $(\mathcal{S}, \mathcal{A}, \mathcal{B}, H, r, \mathbb{P})$, where $\mathcal{S}$ is the infinite state space with $\mathrm{Vol}(\mathcal{S}) \le 1$, $\mathcal{A}$ and $\mathcal{B}$ are the finite action spaces for two players with the sizes of $|\mathcal{A}|$ and $|\mathcal{B}|$, $H$ is the length of an episode, $r = \{r_h\}_{h=1}^H$ is the reward function with $r_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \mapsto [-1, 1]$, and $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ denotes the transition model with $\mathbb{P}_h(s'|s, a, b)$ being the probability density of the two players transitioning to $s' \in \mathcal{S}$ from state $s \in \mathcal{S}$ after taking action $a \in \mathcal{A}$ and $b \in \mathcal{B}$ at step $h$. The policies of the two players are denoted as $\pi = \{\pi_h\}_{h=1}^H$ and $\nu = \{\nu_h\}_{h=1}^H$, where $\pi_h(a|s)$ and $\nu_h(b|s)$ are the probabilities of taking actions $a \in \mathcal{A}$ or $b \in \mathcal{B}$ at the state $s \in \mathcal{S}$. Moreover, we denote $\sigma = \{\sigma_h\}_{h=1}^H$ as a joint policy, where $\sigma_h(a, b|s)$ is the probability of taking actions $a \in \mathcal{A}$ and $b \in \mathcal{B}$ at the state $s \in \mathcal{S}$. Note that the actions $a$ and $b$ are not necessarily mutually independent conditioned on state $s$. One special case of a joint policy is the product of a policy pair $\pi \times \nu$. Here we also assume the initial state is fixed as $s_1^k = s_1$ for any episode $k$. The Markov game is a multi-agent extension of the MDP model under a competitive environment.

For any $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ and joint policy $\sigma$, we define the Q-function and value function as $Q_h^\sigma(s, a, b) = \mathbb{E}[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}, b_{h'}) \,|\, s_h = s, a_h = a, b_h = b, \sigma, \mathbb{P}]$ and $V_h^\sigma(s) = \mathbb{E}[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}, b_{h'}) \,|\, s_h = s, \sigma, \mathbb{P}]$. We have the Bellman equation as $Q_h^\sigma(s, a, b) = r_h(s, a, b) + \mathbb{P}_h V_{h+1}^\sigma(s, a, b)$ and $V_h^\sigma(s) = \langle \sigma_h(\cdot, \cdot|s), Q_h^\sigma(s, \cdot, \cdot)\rangle$. We denote $\mathbb{P}_h V(s, a, b) = \int_{s'} \mathbb{P}_h(s'|s, a, b) V(s') \mathrm{d}s'$ for any value function $V$. We say $(\pi^\dagger, \nu^\dagger)$ is a *Nash equilibrium (NE)* if it is a solution to the max-min optimization problem $\max_\pi \min_\nu V_1^{\pi,\nu}(s_1)$. Then, $(\pi, \nu)$ is an $\varepsilon$-*approximate NE* if it satisfies

$$\max_{\pi'} V_1^{\pi',\nu}(s_1) - \min_{\nu'} V_1^{\pi,\nu'}(s_1) \le \varepsilon.$$

In addition, we denote $\mathrm{br}(\cdot)$ as the best response, which is defined as $\mathrm{br}(\nu) = \operatorname{argmax}_\pi V_1^{\pi,\nu}(s_1)$ and $\mathrm{br}(\pi) = \operatorname{argmin}_\nu V_1^{\pi,\nu}(s_1)$.

**Low-Rank Transition Kernel.** In this paper, we consider the low-rank structures with the dimension $d$ (Jiang et al., 2017; Agarwal et al., 2020; Uehara et al., 2021) for both single-agent MDPs and Markov games, in which the transition model admits the structure in the following assumption. To unify both settings, with a slight abuse of notation, we let $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$ for single-agent MDPs and $\mathcal{Z} := \mathcal{S} \times \mathcal{A} \times \mathcal{B}$ for Markov games.

**Assumption 2.1** (Low-Rank Transition Kernel)**.** *Assuming there exist two unknown maps $\psi^* : \mathcal{S} \mapsto \mathbb{R}^d$ and $\phi^* : \mathcal{Z} \mapsto \mathbb{R}^d$, the true transition kernel admits the following low-rank decomposition for all $h \in [H]$, $(z, s') \in \mathcal{Z} \times \mathcal{S}$,*

$$\mathbb{P}_h(s'|z) = \psi_h^*(s')^\top \phi_h^*(z),$$

*where $\|\phi_h^*(z)\|_2 \le 1$ and $\|\psi_h^*(s')\|_2 \le \sqrt{d}$.*

**Remark 2.2.** *In contrast to linear MDPs (Jin et al., 2020) or linear Markov games (Xie et al., 2020) where $\phi_h^*$ is known* a priori*, we adopt the more challenging setting that both $\psi_h^*$ and $\phi_h^*$ are unknown and hence should be identified via contrastive learning. Moreover, our work also extends the scenario of the low-rank transition model from single-agent RL (Jiang et al., 2017; Agarwal et al., 2020; Uehara et al., 2021) to the multi-agent competitive RL.*

**Algorithm 1** Online Contrastive RL for Single-Agent MDPs

**Initialize:** $\pi_h^0(a|s) = 1/|\mathcal{A}|, \forall (s, a) \in \mathcal{S} \times \mathcal{A}$. $\mathcal{D}_h^0 = \varnothing, \forall h \in [H]$. $\delta > 0$, $\beta > 0$, and $\varepsilon > 0$.
**for** episode $k = 1, \ldots, K$ **do**
  Let $V_{H+1}^k(\cdot) = \mathbf{0}$ and $Q_{H+1}^k(\cdot, \cdot) = \mathbf{0}$
  Collect bonus data $\{\widetilde{\mathcal{D}}_h^k = \{(\widetilde{s}_h^\tau, \widetilde{a}_h^\tau)\}_{\tau=1}^k\}_{h=1}^H$ and contrastive training data $\{\mathcal{D}_h^k\}_{h=1}^H$ by Alg. 3.
  **for** step $h = H, H-1, \ldots, 1$ **do**
    Obtain $\widetilde{\phi}_h^k$ and $\widetilde{\psi}_h^k$ by solving (3) with $\mathcal{D}_h^k$.
    Normalize $\widetilde{\phi}_h^k$ and $\widetilde{\psi}_h^k$ by (1) to obtain $\widehat{\phi}_h^k$ and $\widehat{\psi}_h^k$.
    Estimate $\mathbb{P}_h$ by $\widehat{\mathbb{P}}_h^k(\cdot|\cdot, \cdot) = \widehat{\psi}_h^k(\cdot)^\top \widehat{\phi}_h^k(\cdot, \cdot)$.
    $\widehat{\Sigma}_h^k = \frac{1}{k}\sum_{\tau=1}^k \widehat{\phi}_h^k(\widetilde{s}_h^\tau, \widetilde{a}_h^\tau)\widehat{\phi}_h^k(\widetilde{s}_h^\tau, \widetilde{a}_h^\tau)^\top + \lambda_k I$.
    Bonus $\beta_h^k(\cdot, \cdot) = \min\{\gamma_k \|\widehat{\phi}_h^k(\cdot, \cdot)\|_{(\widehat{\Sigma}_h^k)^{-1}}, 2H\}$.
    $\overline{Q}_h^k(\cdot, \cdot) = (r_h + \beta_h^k + \widehat{\mathbb{P}}_h^k \overline{V}_{h+1}^k)(\cdot, \cdot)$.
    $\overline{V}_h^k(\cdot) = \max_{a \in \mathcal{A}} \overline{Q}_h^k(\cdot, a)$.
    $\pi_h^k(\cdot) = \operatorname{argmax}_{a \in \mathcal{A}} \overline{Q}_h^k(\cdot, a)$.
  **end for**
**end for**

## 3. Contrastive Learning for Single-Agent MDP

### 3.1. Algorithm

**Algorithmic Framework.** We propose an online UCB-type contrastive RL algorithm, Contrastive UCB, for MDPs in Algorithm 1. At the $k$-th episode, we execute the learned policy from the last round to collect the datasets $\{\widetilde{\mathcal{D}}_h^k\}_{h=1}^H$ and $\{\mathcal{D}_h^k\}_{h=1}^H$ as bonus construction data and the contrastive learning data according to the sampling strategy in Algorithm 3. Specifically, the contrastive learning sample is composed of positive and negative data points. At a state-action pair $(s_h, a_h)$ that is sampled independently following a certain distribution formed by the current policy and the true transition, with probability $1/2$, we collect the positive transition data point as $(s_h, a_h, s_{h+1}, 1)$ with $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)$ and a label $y = 1$. On the other hand, with probability $1/2$, we generate the negative tran-

sition data point as $(s_h, a_h, s_{h+1}^-, 0)$ with $s_{h+1}^- \sim \mathcal{P}_{\mathcal{S}}^-(\cdot)$ and a label $y = 0$, where $\mathcal{P}_{\mathcal{S}}^-(\cdot)$ is a designed negative sampling distribution. Given the data sample for contrastive learning $\{\mathcal{D}_h^k\}_{h=1}^H$, we propose to solve the minimization problem (3) at each step $h$ with $\mathcal{L}_h(\psi, \phi; \mathcal{D}_h^k)$ denoting the contrastive loss defined in (2) to learn the low-rank representation $\widetilde{\phi}_h^k$ and $\widetilde{\psi}_h^k$. More detailed implementation of data sampling and the contrastive loss will be elaborated below. According to our analysis in Section 5.1, the true transition kernel $\mathbb{P}_h(s'|s,a)$ can be well approximated by the learned representation $\widetilde{\phi}_h^k(s')^\top \widetilde{\psi}_h^k(s,a)\mathcal{P}_{\mathcal{S}}^-(s')$. However, such learned features are not guaranteed to satisfy the relation $\int_{s'\in\mathcal{S}} \widetilde{\phi}_h^k(s')^\top \widetilde{\psi}_h^k(s,a)\mathcal{P}_{\mathcal{S}}^-(s')\mathrm{d}s' = 1$ or $\widetilde{\phi}_h^k(\cdot)^\top \widetilde{\psi}_h^k(s,a)\mathcal{P}_{\mathcal{S}}^-(\cdot)$ may not be a distribution over $\mathcal{S}$. Thus, we further normalize learned representations by

$$\begin{aligned} &\widehat{\psi}_h^k(s') := \mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\psi}_h^k(s'), \\ &\widehat{\phi}_h^k(z) := \widetilde{\phi}_h^k(z) \big/ \textstyle\int_{s'\in\mathcal{S}} \mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(z)^\top \widetilde{\psi}_h^k(s')\mathrm{d}s', \end{aligned} \tag{1}$$

where $z = (s,a)$. Then, we obtain an approximated transition kernel $\widehat{\mathbb{P}}_h^k(\cdot|s,a) := \widehat{\psi}_h^k(\cdot)^\top \widehat{\phi}_h^k(s,a)$. Our analysis in Section 5.1 shows that $\widehat{\mathbb{P}}_h^k(\cdot|s,a)$ lies in a probability simplex and can well approximate the true transition $\mathbb{P}_h(\cdot|s,a)$.

Simultaneously, we construct the UCB bonus term $\beta_h^k$ with the learned representation $\widehat{\phi}_h^k$ and the empirical covariance matrix $\widehat{\Sigma}_h^k$ using the bonus construction data sampled online via Algorithm 3. Then, with the estimated transition $\widehat{\mathbb{P}}_h^k$ and the UCB bonus term $\beta_h^k$, we obtain a UCB estimation of the Q-function and value function in Line 11 and Line 12. The policy $\pi_h^k$ is then the greedy policy corresponding to the estimated Q-function $\overline{Q}_h^k$.

**Remark 3.1.** *To focus our analysis on the contrastive learning for the transition dynamics, we only consider the setting where the reward function $r_h(\cdot,\cdot)$ is known. One might further modify the proposed algorithm to the unknown reward setting under the linear reward function assumption by considering to minimize a square loss with observed rewards as the regression target to learn the parameters. The corresponding analysis would then take the statistical error of such a procedure into consideration.*

**Dataset for Contrastive Learning.** For our algorithm, we make the following assumption for the negative sampling distribution $\mathcal{P}_{\mathcal{S}}^-(\cdot)$.

**Assumption 3.2** (Negative Sampling Distribution)**.** *Let $\mathcal{P}_{\mathcal{S}}^-(\cdot)$ be a distribution over $\mathcal{S}$. The distribution $\mathcal{P}_{\mathcal{S}}^-(\cdot)$ satisfies $\inf_{s\in\mathcal{S}} \mathcal{P}_{\mathcal{S}}^-(s) \geq C_{\mathcal{S}}^- > 0$ for a constant $C_{\mathcal{S}}^-$.*

The detailed sampling scheme for the contrastive learning dataset is presented in Algorithm 3 in Appendix. Here we provide a brief idea of this algorithm. Letting $d_h^\pi(\cdot)$ be the state distribution at step $h$ under the true transition $\mathbb{P}$ and a policy $\pi$, we define two state-action distributions induced by $\pi$ and $\mathbb{P}$ at step $h$ as $\widetilde{d}_h^\pi(s,a) = d_h^\pi(s)\mathrm{Unif}(a)$ and $\breve{d}_h^\pi(s,a) = \widetilde{d}_{h-1}^\pi(s',a')\mathbb{P}_{h-1}(s|s',a')\mathrm{Unif}(a)$, where $\mathrm{Unif}(a) = 1/|\mathcal{A}|$. Then, at each round $k$, we sample the temporal data as follows:

- Sample $(\widetilde{s}_h^k, \widetilde{a}_h^k) \sim \widetilde{d}_h^{\pi^{k-1}}(\cdot,\cdot)$ for all $h \in [H]$ and $(\breve{s}_h^k, \breve{a}_h^k) \sim \breve{d}_h^{\pi^{k-1}}(\cdot,\cdot)$ for all $h \geq 2$.
- For each $(\widetilde{s}_h^k, \widetilde{a}_h^k)$ or $(\breve{s}_h^k, \breve{a}_h^k)$, generate a label $y \in \{0,1\}$ from a Bernoulli distribution $\mathrm{Ber}(1/2)$ independently.
- Sample the next state from the true transition as $\widetilde{s}_{h+1}^k \sim \mathbb{P}_h(\cdot|\widetilde{s}_h^k, \widetilde{a}_h^k)$ or $\breve{s}_{h+1}^k \sim \mathbb{P}_h(\cdot|\breve{s}_h^k, \breve{a}_h^k)$ when the associated labels are 1 and sample negative transition data points by $\widetilde{s}_{h+1}^{k,-} \sim \mathcal{P}_{\mathcal{S}}^-(\cdot)$ or $\breve{s}_{h+1}^{k,-} \sim \mathcal{P}_{\mathcal{S}}^-(\cdot)$ if labels are 0.
- Given the dataset $\mathcal{D}_h^{k-1}$ from the last round, add the new transition data with labels, i.e., $(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{s}_{h+1}^k, 1)$ or $(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{s}_{h+1}^{k,-}, 0)$ and $(\breve{s}_h^k, \breve{a}_h^k, \breve{s}_{h+1}^k, 1)$ or $(\breve{s}_h^k, \breve{a}_h^k, \breve{s}_{h+1}^{k,-}, 0)$, into it to compose a new set $\mathcal{D}_h^k$.

In addition, we also acquire a dataset $\widetilde{\mathcal{D}}_h^k$ via Algorithm 3 for the construction of the UCB bonus term in Algorithm 1, where $\widetilde{\mathcal{D}}_h^k$ is composed of the present and historical state-action pairs sampled from $\widetilde{d}_h^{\pi^{k'}}(\cdot,\cdot)$ for all $k' \in [0, k-1]$. Algorithm 3 illustrates how to sample the above data by interacting with the environment in an online manner, which can also guarantee the data points are mutually independent within $\mathcal{D}_h^k$ and $\widetilde{\mathcal{D}}_h^k$.

**Contrastive Loss.** Given the dataset $\{\mathcal{D}_h^k\}_{h=1}^H$ for contrastive learning, we further define the following contrastive loss for each step $h \in [H]$

$$\begin{aligned} \mathcal{L}_h(\psi, \phi; \mathcal{D}_h^k) := \mathbb{E}_{\mathcal{D}_h^k}\big[&y \log(1 + 1/\psi(s')^\top \phi(z)) \\ &+ (1-y)\log(1 + \psi(s')^\top \phi(z))\big], \end{aligned} \tag{2}$$

where $z = (s,a)$ and $\mathbb{E}_{\mathcal{D}_h^k}$ indicates taking average over all $(s,a,s',y)$ in the collected contrastive training dataset $\mathcal{D}_h^k$. Here $\phi$ and $\psi$ are two functions lying in the function classes $\Phi$ and $\Phi$ as defined below. Letting $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$, we define:

**Definition 3.3** (Function Class)**.** *Let $\mathcal{F} := \{\psi(\cdot)^\top \phi(\cdot,\cdot) : \psi \in \Psi, \phi \in \Phi\}$ be a function class where $\Psi := \{\phi : \mathcal{S} \mapsto \mathbb{R}^d\}$ and $\Phi := \{\psi : \mathcal{Z} \mapsto \mathbb{R}^d\}$ are two finite function classes. For any $\psi \in \Psi$, $\sup_{s\in\mathcal{S}} \|\psi(s)\|_2 \leq \sqrt{d}/C_{\mathcal{S}}^-$. And for any $\phi \in \Phi$, $\sup_{s\in\mathcal{S}} \|\phi(z)\|_2 \leq 1$. The cardinality of $\mathcal{F}$ is $|\mathcal{F}| = |\Psi| \cdot |\Phi|$.*

The fundamental idea of designing (2) is to consider a negative log-likelihood loss for the probability $\mathrm{Pr}_h(y|s,a,s') := \left(\frac{f_h(s,a,s')}{1+f_h(s,a,s')}\right)^y \left(\frac{1}{1+f_h(s,a,s')}\right)^{1-y}$ where $f_h(s,a,s') = \psi(s')^\top \phi(s,a)$ and $\mathrm{Pr}_h$ denote the associated probability at step $h$. Then (2) is equivalent to $\mathcal{L}_h(\psi, \phi; \mathcal{D}_h^k) = -\mathbb{E}_{\mathcal{D}_h^k}[\log \mathrm{Pr}_h(y|s,a,s')]$. Thus, to learn the contrastive

feature representation, we seek to solve the following problem of contrastive loss minimization

$$(\widetilde{\psi}_h^k, \widetilde{\phi}_h^k) = \operatorname*{argmin}_{\psi\in\Psi,\phi\in\Phi} \mathcal{L}_h(\psi, \phi; \mathcal{D}_h^k). \tag{3}$$

According to Lemma C.1 in Appendix, letting $z = (s, a)$, the learning target of the above minimization problem is

$$f_h^*(z, s') = \mathbb{P}_h(s'|z)/\mathcal{P}_{\mathcal{S}}^-(s'). \tag{4}$$

Since $\mathbb{P}_h(s'|z) = \psi_h^*(s')^\top \phi_h^*(z)$ with $\|\phi_h^*(z)\|_2 \le 1$ and $\|\psi_h^*(s')\|_2 \le \sqrt{d}$ as in Assumption 2.1, by Definition 3.3, we know $f_h^* \in \mathcal{F}$, i.e., $\psi_h^*(\cdot)/\mathcal{P}_{\mathcal{S}}^-(\cdot) \in \Psi$ and $\phi_h^*(\cdot) \in \Phi$.

**Remark 3.4.** *The parameter $C_{\mathcal{S}}^-$ in Assumption 3.2 captures the fundamental difficulty of contrastive learning in RL by characterizing how large the function class (Definition 3.3) should be to include the underlying true density ratio in (4). Technically, it also guarantees that the problem is mathematically well-defined. In particular, the true density ratio (4) has non-zero denominator $\mathcal{P}_{\mathcal{S}}^-(s), \forall s \in \mathcal{S}$ if the parameter $C_{\mathcal{S}}^-$ is positive.*

**Remark 3.5.** *One can further extend the setting of the finite function class to the infinite function class setting by utilizing the covering argument as in Van de Geer (2000); Uehara & Sun (2021) such that the terms depending on the cardinality of $\mathcal{F}$ would be replaced by terms related to the covering number of $\mathcal{F}$. We leave such an analysis under the online setting as our future work.*

### 3.2. Main Result for Single-Agent MDP Setting

**Theorem 3.6** (Sample Complexity)**.** *Letting $\lambda_k = c_0 d \log(H|\mathcal{F}|k/\delta)$ for a sufficiently large constant $c_0 > 0$ and $\gamma_k = 4H\big(12\sqrt{|\mathcal{A}|d} + \sqrt{c_0 d}\big)/C_{\mathcal{S}}^- \cdot \sqrt{\log(2Hk|\mathcal{F}|/\delta)}$, with probability at least $1 - 3\delta$, we have*

$$1/K \cdot \textstyle\sum_{k=1}^K \left[V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1)\right] \lesssim \sqrt{C \log(H|\mathcal{F}|K/\delta) \log(c_0' K)/K},$$

*where $C = H^4 d^4 |\mathcal{A}|/(C_{\mathcal{S}}^-)^2 + H^4 d^3 |\mathcal{A}|^2/(C_{\mathcal{S}}^-)^2 + H^6 d^2 |\mathcal{A}|/(C_{\mathcal{S}}^-)^2 + H^6 d^3$ and $c_0'$ is an absolute constant.*

Letting $\widehat{\pi}$ be a policy uniformly sampled from $\{\pi^k\}_{k=1}^K$ generated by Algorithm 1, the above theorem indicates $\widehat{\pi}$ is an $\varepsilon$-suboptimial policy with probability at least $1 - 3\delta$ after executing Algorithm 1 for $K \ge \widetilde{\mathcal{O}}(1/\varepsilon^2)$ episodes. Here $\widetilde{\mathcal{O}}$ hides logarithmic dependence on $|\mathcal{F}|, H, K, 1/\delta$, and $1/\varepsilon$.

## 4. Contrastive Learning for Markov Game

### 4.1. Algorithm

**Algorithmic Framework.** We propose an online algorithm, Contrastive ULCB, for contrastive learning on Markov games in Algorithm 2. At the $k$-th round, we execute the learned joint policy $\sigma^{k-1}$ from the last round to collect the bonus construction data $\{\widetilde{\mathcal{D}}_h^k\}_{h=1}^H$ and the contrastive learning data $\{\mathcal{D}_h^k\}_{h=1}^H$ via the sampling algorithm in Algorithm 4. At a state-action pair $(s_h, a_h, b_h)$ sampled at the $h$-th step, with probability $1/2$ respectively, we collect the positive transition data point $(s_h, a_h, b_h, s_{h+1}, 1)$ with $s_{h+1} \sim \mathbb{P}(\cdot|s_h, a_h, b_h)$ and the negative transition data point $(s_h, a_h, s_{h+1}^-, 0)$ with $s_{h+1}^- \sim \mathcal{P}_{\mathcal{S}}^-(\cdot)$, where $\mathcal{P}_{\mathcal{S}}^-(\cdot)$ is the negative sampling distribution. Given the dataset $\{\mathcal{D}_h^k\}_{h=1}^H$ for contrastive learning, we define the contrastive loss $\mathcal{L}_h(\psi, \phi; \mathcal{D}_h^k)$ as in (2) with setting $z = (s, a, b)$. The function class $\mathcal{F}$ is then defined the same as in Definition 3.3 by setting $z = (s, a, b)$. We solve the contrastive loss minimization problem as (3) at each step $h$ to learn the representation $\widetilde{\phi}_h^k$ and $\widetilde{\psi}_h^k$. Since it is not guaranteed that $\widetilde{\phi}_h^k(\cdot)^\top \widetilde{\psi}_h^k(s, a, b)\mathcal{P}_{\mathcal{S}}^-(\cdot)$ is a distribution over $\mathcal{S}$, we normalize $\widetilde{\phi}_h^k$ and $\widetilde{\psi}_h^k$ as (1) where $z = (s, a, b)$. Then we obtain an approximated transition kernel $\widehat{\mathbb{P}}_h^k(\cdot|s, a, b) := \widehat{\psi}_h^k(\cdot)^\top \widehat{\phi}_h^k(s, a, b)$. Furthermore, we use the bonus dataset to construct the empirical covariance matrix $\widehat{\Sigma}_h^k$ and then the bonus term $\beta_h^k$. The major differences between algorithms for single-agent MDPs and Markov games lie in the following two steps: **(1)** In Lines 11 and 12, we have two types of Q-functions with both addition and subtraction of bonus terms such that Algorithm 2 is an upper and lower confidence bound (ULCB)-type algorithm. **(2)** We update policies of two players by first finding an $\iota_k$-coarse corre-

**Algorithm 2** Online Contrastive RL for Markov Games

1: **Initialize:** $\sigma_h^0(a, b|s) = 1/(|\mathcal{A}||\mathcal{B}|), \forall (s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$. $\mathcal{D}_h^0 = \varnothing, \forall h \in [H]$. $\delta > 0$, $\beta > 0$, and $\varepsilon > 0$.
2: **for** episode $k = 1, \ldots, K$ **do**
3: Let $V_{H+1}^k(\cdot) = \mathbf{0}$ and $Q_{H+1}^k(\cdot, \cdot, \cdot) = \mathbf{0}$
4: Collect bonus data $\{\widetilde{\mathcal{D}}_h^k = \{(\widetilde{s}_h^\tau, \widetilde{a}_h^\tau, \widetilde{b}_h^\tau)\}_{\tau=1}^k\}_{h=1}^H$ and contrastive training data $\{\mathcal{D}_h^k\}_{h=1}^H$ by Alg. 4.
5: **for** step $h = H, H-1, \ldots, 1$ **do**
6: Obtain $\widetilde{\phi}_h^k$ and $\widetilde{\psi}_h^k$ by solving (3) with $\mathcal{D}_h$.
7: Normalize $\widetilde{\phi}_h^k$ and $\widetilde{\psi}_h^k$ by (1) to obtain $\widehat{\phi}_h^k$ and $\widehat{\psi}_h^k$.
8: Estimate $\mathbb{P}_h$ by $\widehat{\mathbb{P}}_h^k(\cdot|\cdot, \cdot, \cdot) = \widehat{\psi}_h^k(\cdot)^\top \widehat{\phi}_h^k(\cdot, \cdot, \cdot)$.
9: $\widehat{\Sigma}_h^k = \frac{1}{k}\sum_{\tau=1}^k \widehat{\phi}_h^k(\widetilde{s}_h^\tau, \widetilde{a}_h^\tau, \widetilde{b}_h^\tau)\widehat{\phi}_h^k(\widetilde{s}_h^\tau, \widetilde{a}_h^\tau, \widetilde{b}_h^\tau)^\top + \lambda_k I$
10: $\beta_h^k(\cdot, \cdot, \cdot) = \min\{\gamma_k \|\widehat{\phi}_h^k(\cdot, \cdot, \cdot)\|_{(\widehat{\Sigma}_h^k)^{-1}}, 2H\}$.
11: $\overline{Q}_h^k(\cdot, \cdot, \cdot) = (r_h + \widehat{\mathbb{P}}_h^k \overline{V}_{h+1}^k + \beta_h^k)(\cdot, \cdot, \cdot)$.
12: $\underline{Q}_h^k(\cdot, \cdot, \cdot) = (r_h + \widehat{\mathbb{P}}_h^k \underline{V}_{h+1}^k - \beta_h^k)(\cdot, \cdot, \cdot)$.
13: $\overline{V}_h^k(\cdot) = \langle \sigma_h^k(\cdot, \cdot|\cdot), \overline{Q}_h^k(\cdot, \cdot, \cdot)\rangle$.
14: $\underline{V}_h^k(\cdot) = \langle \sigma_h^k(\cdot, \cdot|\cdot), \underline{Q}_h^k(\cdot, \cdot, \cdot)\rangle$.
15: $\sigma_h^k(\cdot, \cdot|s) = \iota_k$-CCE$(\overline{Q}_h^k(s, \cdot, \cdot), \underline{Q}_h^k(s, \cdot, \cdot)), \forall s$.
16: $\pi_h^k = \mathcal{P}_1 \sigma_h^k$ and $\nu_h^k = \mathcal{P}_2 \sigma_h^k$.
17: **end for**
18: **end for**

lated equilibrium (CCE) with the two Q-functions as a joint policy $\{\sigma_h^k\}_{h=1}^H$ in Line 15 and then applying marginalization to obtain the policies as in Line 16, where $\mathcal{P}_1$ and $\mathcal{P}_2$ denote getting marginal distributions over $\mathcal{A}$ and $\mathcal{B}$ respectively. In particular, the notion of an $\iota$-CCE (Moulin & Vial, 1978; Aumann, 1987) is defined as follows:

**Definition 4.1** ($\iota$-CCE). *For two payoff matrices $\overline{Q}, \underline{Q} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$, a distribution $\mu$ over $\mathcal{A}\times\mathcal{B}$ is $\iota$-CCE if it satisfies*

$$\mathbb{E}_{(a,b)\sim\mu}[\overline{Q}(a,b)] \geq \mathbb{E}_{b\sim\mathcal{P}_2\mu}[\overline{Q}(a',b)] - \iota, \forall a' \in \mathcal{A},$$
$$\mathbb{E}_{(a,b)\sim\mu}[\underline{Q}(a,b)] \leq \mathbb{E}_{a\sim\mathcal{P}_1\mu}[\underline{Q}(a,b')] + \iota, \forall b' \in \mathcal{B}.$$

An $\iota$-CCE may not have mutually independent marginals since the two players take actions in a correlated way. The $\iota$-CCE can be found *efficiently* by the method developed in Xie et al. (2020) for arbitrary $\iota > 0$.

**Dataset for Contrastive Learning.** Summarized in Algorithm 4 in Appendix, the sampling algorithm for Markov games follows a similar sampling strategy to Algorithm 3 with extending the action space from $\mathcal{A}$ to $\mathcal{A}\times\mathcal{B}$. Letting $d_h^\sigma(s)$ be a state probability at step $h$ under $\mathbb{P}$ and a joint policy $\sigma$, we define $\widetilde{d}_h^\sigma(s,a,b) = d_h^\sigma(s)\mathrm{Unif}(a)\mathrm{Unif}(b)$ and $\breve{d}_h^\sigma(s,a,b) = \widetilde{d}_{h-1}^\sigma(s',a',b')\mathbb{P}_{h-1}(s|s',a',b')\mathrm{Unif}(a)\mathrm{Unif}(b)$, where we define $\mathrm{Unif}(a) = 1/|\mathcal{A}|$ and $\mathrm{Unif}(b) = 1/|\mathcal{B}|$. Analogously, at round $k$, we sample state-action pairs following $\widetilde{d}_h^{\sigma^{k-1}}(\cdot,\cdot,\cdot)$ for all $h\in[H]$ and $\breve{d}_h^{\sigma^{k-1}}(\cdot,\cdot,\cdot)$ for all $h\geq 2$ and then sample the next state from $\mathbb{P}_h$ or negative sampling distribution $\mathcal{P}_{\mathcal{S}}^-$ with probability $1/2$. We also acquire a dataset for the construction of the bonus term in Algorithm 2 by sampling from $\widetilde{d}_h^{\sigma^{k'}}(\cdot,\cdot,\cdot)$ for all $k'\in[0,k-1]$.

### 4.2. Main Result for Markov Game Setting

**Theorem 4.2** (Sample Complexity). *Letting $\lambda_k = c_0 d\log(H|\mathcal{F}|k/\delta)$ for a sufficiently large constant $c_0 > 0$, $\gamma_k = 4H\big(12\sqrt{|\mathcal{A}||\mathcal{B}|d} + \sqrt{c_0 d}\big)/C_{\mathcal{S}}^- \cdot \sqrt{\log(2Hk|\mathcal{F}|/\delta)}$, and $\iota_k \leq \mathcal{O}(\sqrt{1/k})$, with probability at least $1-3\delta$, we have*

$$1/K\cdot\textstyle\sum_{k=1}^K \big[V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)\big] \lesssim \sqrt{C\log(H|\mathcal{F}|K/\delta)\log(c_0' K)/K},$$

*where $C = H^4d^4|\mathcal{A}||\mathcal{B}|/(C_{\mathcal{S}}^-)^2 + H^4d^3|\mathcal{A}|^2|\mathcal{B}|^2/(C_{\mathcal{S}}^-)^2 + H^6d^2|\mathcal{A}||\mathcal{B}|/(C_{\mathcal{S}}^-)^2 + H^6d^3$ and $c_0'$ is an absolute constant.*

This theorem further implies a PAC bound for learning an approximate NE (Xie et al., 2020). Specifically, Theorem 4.2 implies that there exists $k_0\in[K]$ such that $(\pi^{k_0},\nu^{k_0})$ is an $\varepsilon$-approximate NE with probability at least $1-3\delta$ after executing Algorithm 2 for $K\geq\widetilde{\mathcal{O}}(1/\varepsilon^2)$ episodes, i.e., letting $k_0 := \min_{k\in[K]}[V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)]$, we then have

$$V_1^{\mathrm{br}(\nu^{k_0}),\nu^{k_0}}(s_1) - V_1^{\pi^{k_0},\mathrm{br}(\pi^{k_0})}(s_1) \leq 1/K\cdot\textstyle\sum_{k=1}^K[V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)] \leq \varepsilon$$

with probability at least $1-3\delta$.

## 5. Theoretical Analysis

This section provides the analysis of the transition kernel recovery via contrastive learning and the proofs of the main results for single-agent MDPs and zero-sum MGs. Our theoretical analysis integrates contrastive self-supervised learning for transition recovery and low-rank MDPs in a unified manner. Part of our analysis is motivated by the recent work (Uehara et al., 2021) for learning the low-rank MDPs. In contrast to this work, our paper analyzes the representation recovery via contrastive learning under the online setting. In addition, we consider an episodic setting distinct from the infinite-horizon setting in the aforementioned work. On the other hand, the existing work on low-rank MDPs only focuses on a single-agent setting. Our analysis further considers a Markov game setting where a natural challenge of non-stationarity arises due to competitive policies of multiple players. We develop the first representation learning analysis for Markov games based on the proposed ULCB algorithm.

We first define several notations for our analysis. Recall that we have defined $d_h^\pi$, $\widetilde{d}_h^\pi$, and $\breve{d}_h^\pi$ as in Section 3.1. Then, we subsequently define $\rho_h^k(s,a) := 1/k\cdot\sum_{k'=0}^{k-1} d_h^{\pi^{k'}}(s,a)$, $\widetilde{\rho}_h^k(s,a) := 1/k\cdot\sum_{k'=0}^{k-1}\widetilde{d}_h^{\pi^{k'}}(s,a)$, and $\breve{\rho}_h^k(s,a) := 1/k\cdot\sum_{k'=0}^{k-1}\breve{d}_h^{\pi^{k'}}(s,a)$, which are the averaged distributions across $k$ episodes for the corresponding state-action distributions. In addition, for any $\rho$ and $\phi$, we define the associated covariance matrix $\Sigma_{\rho,\phi} := k\cdot\mathbb{E}_{(s,a)\sim\rho_h^k(\cdot,\cdot)}\big[\phi(s,a)\phi(s,a)^\top\big] + \lambda_k I$. On the other hand, for zero-sum MGs, in Section 4.1, we have defined $d_h^\sigma$, $\widetilde{d}_h^\sigma$, and $\breve{d}_h^\sigma$ for any joint policy $\sigma$. Then, we can analogously define $\rho_h^k$, $\widetilde{\rho}_h^k$, $\breve{\rho}_h^k$, and $\Sigma_{\rho,\phi}$ for MGs by extending action spaces from $\mathcal{A}$ to $\mathcal{A}\times\mathcal{B}$. We summarize these notations in a table in Section B. Moreover, for abbreviation, letting $z=(s,a)$ for MDPs and $z=(s,a,b)$ for MGs and $\widetilde{\rho}_h^k$, $\breve{\rho}_h^k$ be corresponding distributions, we define

$$\zeta_h^k := \mathbb{E}_{z\sim\widetilde{\rho}_h^k}[\|\mathbb{P}_1(\cdot|z) - \widehat{\mathbb{P}}_1^k(\cdot|z)\|_1^2],$$
$$\xi_h^k := \mathbb{E}_{z\sim\breve{\rho}_h^k}[\|\mathbb{P}_h(\cdot|z) - \widehat{\mathbb{P}}_h^k(\cdot|z)\|_1^2]. \tag{5}$$

### 5.1. Analysis for Single-Agent MDP

Based on the above definitions and notations, we have the following lemma to show the transition recovery via contrastive learning.

**Lemma 5.1** (Transition Recovery). *After executing Algorithm 1 for $k$ rounds, with probability at least $1-2\delta$,*

$$\zeta_h^k \le 32d/(C_{\mathcal{S}}^-)^2 \cdot \log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \ge 1,$$
$$\xi_h^k \le 32d/(C_{\mathcal{S}}^-)^2 \cdot \log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \ge 2,$$

*where $\zeta_h^k$ and $\xi_h^k$ are defined as (5).*

This lemma indicates that via the contrastive learning step in Algorithm 1, we can successfully learn a correct representation and recover the transition model. Next, we give the proof sketch of this lemma.

**Proof Sketch of Lemma 5.1**. Letting $\Pr_h^f(y|s,a,s')$ be defined as in Section 3.1, we have $\Pr_h^f(y,s'|s,a) = \Pr_h^f(y|s,a,s')\Pr_h(s'|s,a)$ with defining $f_h(s,a,s') := \psi(s')^\top\phi(s,a)$. Furthermore, we can calculate that $\Pr_h(s'|s,a) = \frac{1}{2}[\mathbb{P}_h(s'|s,a) + \mathcal{P}_{\mathcal{S}}^-(s')] \ge \frac{1}{2}C_{\mathcal{S}}^- > 0$ by Assumption 3.2. Thus, the contrastive loss minimization (3) is equivalent to $\max_{\phi_h,\psi_h}\mathbb{E}_{\mathcal{D}_h^k}\log\Pr_h^f(y|s,a,s')$, which further equals $\max_{\phi_h,\psi_h}\mathbb{E}_{\mathcal{D}_h^k}\log\Pr_h^f(y,s'|s,a)$, since $\Pr_h(s'|s,a)$ is only determined by $\mathbb{P}_h(s'|s,a)$ and $\mathcal{P}_{\mathcal{S}}^-(s')$ and is independent of $f_h$. Denoting the solution as $\widehat{f}_h^k(s,a,s') = \widetilde{\psi}_h^k(s')^\top\widetilde{\phi}_h^k(s,a)$. With Algorithm 3, further by the MLE guarantee in Lemma E.2, we can show with high probability, $\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\Pr_h^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr_h^{f^*}(\cdot,\cdot|s,a)\|_{\mathrm{TV}}^2 \le \epsilon_k$ and $\mathbb{E}_{(s,a)\sim\check{\rho}_h^k(\cdot,\cdot)}\|\Pr_h^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr_h^{f^*}(\cdot,\cdot|s,a)\|_{\mathrm{TV}}^2 \le \epsilon_k$, where $f_h^*$ is defined in (4) and $\epsilon_k := 2\log(2kH|\mathcal{F}|/\delta)/k$.

Next, we show the recovery error bound of the transition model based on $\widehat{f}_h^k$. By expanding $\Pr_h^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr_h^{f^*}(\cdot,\cdot|s,a)$, we further obtain $\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\mathbb{P}_h(\cdot|s,a) - \mathcal{P}_{\mathcal{S}}^-(\cdot)\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(\cdot)\|_{\mathrm{TV}}^2 \le 4d\epsilon_k/(C_{\mathcal{S}}^-)^2$ as well as $\mathbb{E}_{(s,a)\sim\check{\rho}_h^k(\cdot,\cdot)}\|\mathbb{P}_h(\cdot|s,a) - \mathcal{P}_{\mathcal{S}}^-(\cdot)\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(\cdot)\|_{\mathrm{TV}}^2 \le 4d\epsilon_k/(C_{\mathcal{S}}^-)^2$.

Now we define $\widehat{g}_h^k(s,a,s') := \mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(s')$. Since that $\int_{s'\in\mathcal{S}}\widehat{g}_h^k(s,a,s')\mathrm{d}s'$ may not be guaranteed to be 1 though $\widehat{g}_h^k(s,a,\cdot)$ is close to the true transition model $\mathbb{P}_h(\cdot|s,a)$. Therefore, to obtain a distribution approximator of the transition model $\mathbb{P}_h$, we further normalize $\widehat{g}_h^k(s,a,s')$ and define $\widehat{\mathbb{P}}_h^k(s'|s,a) := \widehat{g}_h^k(s,a,s')/\|\widehat{g}_h^k(s,a,\cdot)\|_1 = \widehat{\psi}_h^k(s')^\top\widehat{\phi}_h^k(s,a)$ which is equivalent to (1). According to the definition of the approximation error $\zeta_h^k := \mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a) - \mathbb{P}_h(\cdot|s,a)\|_{\mathrm{TV}}^2$, we can further prove in our formal proof that $\zeta_h^k \le 4\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\mathbb{P}_h(\cdot|s,a) - \mathcal{P}_{\mathcal{S}}^-(\cdot)\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(\cdot)\|_{\mathrm{TV}}^2 \le 16d\epsilon_k/(C_{\mathcal{S}}^-)^2$. We similarly can have $\xi_h^k \le 16d\epsilon_k/(C_{\mathcal{S}}^-)^2$. Plugging in $\epsilon_k = 2\log(2kH|\mathcal{F}|/\delta)/k$ gives the desired results. Please see Section C.2 for a detailed proof.

Based on Lemma 5.1, we give the analysis of Theorem 3.6.

**Proof Sketch of Theorem 3.6**. We first define that $\overline{V}_{k,h}^{\pi}$ is the value function on an auxiliary MDP defined by $\widehat{\mathbb{P}}^k$ and $r+\beta^k$. Then we can decompose $V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1)$ as

$$\begin{aligned} V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1) &= V_1^{\pi^*}(s_1) - \overline{V}_{k,1}^{\pi^*}(s_1) \\ &\quad + \overline{V}_{k,1}^{\pi^*}(s_1) - V_1^k(s_1) + V_1^k(s_1) - V_1^{\pi^k}(s_1) \\ &\le \underbrace{V_1^{\pi^*}(s_1) - \overline{V}_{k,1}^{\pi^*}(s_1)}_{(i)} + \underbrace{\overline{V}_{k,1}^{\pi^k}(s_1) - V_1^{\pi^k}(s_1)}_{(ii)}, \end{aligned} \tag{6}$$

where the first inequality is by Lemma C.6 that $\overline{V}_{k,1}^{\pi^*}(s_1) \le V_1^k(s_1)$ due to the value iteration step in Algorithm 1. Moreover, by the definition of $\overline{V}_h^k$ above, we known $\overline{V}_h^k = \overline{V}_{k,h}^{\pi^k}$ for any $h \in [H]$. Thus, we need to bound $(i)$ and $(ii)$.

To bound term $(i)$, by Lemma C.2 and Lemma C.4, we have

$$(i) = V_1^{\pi^*}(s_1) - V_1^{\pi^*}(s_1) \le \sqrt{|\mathcal{A}|\zeta_1^k},$$

which indicates a near-optimism (Uehara et al., 2021) with a bias $\sqrt{|\mathcal{A}|\zeta_1^k} \le \widetilde{\mathcal{O}}(\sqrt{1/k})$ according to Lemma 5.1. This is guaranteed by adding a UCB bonus to the Q-function.

Term $(ii)$ reflects the model difference between the defined auxiliary MDP and the true MDP under the learned policy $\pi^k$. By Lemma C.3 and Lemma C.5, we have that $(ii) \le [\sqrt{3d|\mathcal{A}|\gamma_k^2/k} + 3H^2\sqrt{|\mathcal{A}|\zeta_1^k}] + \sum_{h=1}^{H-1}[\sqrt{3d|\mathcal{A}|\gamma_k^2 + 4H^2\lambda_k d} + 3H^2\sqrt{k|\mathcal{A}|\zeta_{h+1}^k + 4\lambda_k d}] \cdot \mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}}\|\phi_h^*(s,a)\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}}$. In fact, we can bound the term $\sum_{k=1}^K \mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}}\|\phi_h^*(s,a)\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}} \le \widetilde{O}(\sqrt{dK})$ by Lemma E.3. According to Lemma 5.1, with high probability, we can bound $\zeta_h^k$ and $\xi_h^k$. Then, $\frac{1}{K}\sum_{k=1}^K(ii) \le \widetilde{O}(1/\sqrt{K})$ with polynomial dependence on $|\mathcal{A}|, H, d$ by setting parameters as in Theorem 3.6.

By (6), we have $\frac{1}{K}[V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1)] \le \frac{1}{K}\sum_{k=1}^K[(i)+(ii)]$. Then, plugging in the upper bounds for terms $(i)$ and $(ii)$, setting the parameters $\gamma_k$ and $\lambda_k$ as in Theorem 3.6, we obtain the desired bound. Please see Section C.3 in Appendix for a detailed proof.

### 5.2. Analysis for Markov Game

We further have a transition recovery lemma for Algorithm 2 similar to Lemma 5.1.

**Lemma 5.2** (Transition Recovery). *After executing Algorithm 2 for $k$ rounds, with probability at least $1-2\delta$,*

$$\zeta_h^k \le 32d/(C_{\mathcal{S}}^-)^2 \cdot \log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \ge 1,$$
$$\xi_h^k \le 32d/(C_{\mathcal{S}}^-)^2 \cdot \log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \ge 2,$$

*where $\zeta_h^k$ and $\xi_h^k$ are defined as (5).*

The proof idea for Lemma 5.2 is nearly identical to the one for Lemma 5.1 with extending the action space from $\mathcal{A}$ to $\mathcal{A} \times \mathcal{B}$. We defer the proof to Section D.2 in Appendix. Based on Lemma 5.2, we further give the analysis of Theorem 4.2.

**Proof Sketch of Theorem 4.2.** We define two auxiliary MGs respectively by reward function $r + \beta^k$ and transition model $\widehat{\mathbb{P}}^k$, and $r - \beta^k$, $\widehat{\mathbb{P}}^k$. Then, for any joint policy $\sigma$, let $\overline{V}^{\sigma}_{k,h}$ and $\underline{V}^{\sigma}_{k,h}$ be the associated value functions on the two auxiliary MGs respectively. Recall that $\overline{V}^k_h$ and $\underline{V}^k_h$ are generated by Algorithm 2. We then decompose $V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)$ as follows

$$
\begin{aligned}
V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1) &= \underbrace{V_1^{\sigma^k_\nu}(s_1) - \overline{V}^{\sigma^k_\nu}_{k,1}(s_1)}_{(i)} \\
&+ \underbrace{\overline{V}^{\sigma^k_\nu}_{k,1}(s_1) - \overline{V}^k_1(s_1)}_{(ii)} + \underbrace{\overline{V}^k_1(s_1) - \underline{V}^k_1(s_1)}_{(iii)} \qquad (7) \\
&+ \underbrace{\underline{V}^k_1(s_1) - \underline{V}^{\sigma^k_\pi}_{k,1}(s_1)}_{(iv)} + \underbrace{\underline{V}^{\sigma^k_\pi}_{k,1}(s_1) - V_1^{\sigma^k_\pi}(s_1)}_{(v)}.
\end{aligned}
$$

Here we let $\sigma^k_\nu := (\mathrm{br}(\nu^k), \nu^k)$ and $\sigma^k_\pi := (\pi^k, \mathrm{br}(\pi^k))$ for abbreviation. Terms $(ii)$ and $(iv)$ depict the planning error on the two auxiliary MGs, which is guaranteed to be small by finding $\iota_k$-CCE in Algorithm 2. Thus, by Lemma D.6, we have

$$
(ii) \le H\iota_k, \quad (iv) \le H\iota_k,
$$

which can be controlled by setting a proper value to $\iota_k$ as in Theorem 4.2.

Moreover, by Lemma D.2 and Lemma D.4, we obtain

$$
(i) \le \sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k}, \quad (v) \le \sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k},
$$

which is guaranteed by the design of ULCB-type Q-functions with the bonus term in our algorithm. Thus we obtain the near-optimism and near-pessimism properties for terms $(i)$ and $(v)$ respectively.

Term $(iii)$ is the model difference between the two auxiliary MGs under the learned joint policy $\sigma^k$. By Lemma D.3 and Lemma D.5, we have that $(iii) \le [2\sqrt{3d|\mathcal{A}|\gamma_k^2/k} + 6H^2\sqrt{|\mathcal{A}|\zeta_1^k}] + \sum_{h=1}^{H-1}[2\sqrt{3d|\mathcal{A}|\gamma_k^2 + 4H^2\lambda_k d} + 6H^2\sqrt{k|\mathcal{A}|\zeta_{h+1}^k + 4\lambda_k d}] \cdot \mathbb{E}_{d_h^{\sigma^k,\mathbb{P}}} \|\phi_h^*\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}}$. Furthermore, we obtain that $\sum_{k=1}^K \mathbb{E}_{d_h^{\sigma^k,\mathbb{P}}} \|\phi_h^*\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}} \le \widetilde{O}(\sqrt{dK})$ by Lemma E.3. According to Lemma 5.2 for the contrastive learning, with high probability, we can bound $\frac{1}{K}\sum_{k=1}^K (iii) \le \widetilde{O}(1/\sqrt{K})$ under the same conditions in Theorem 3.6.

According to (7), we have $\frac{1}{K}\sum_{k=1}^K [V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)] \le \frac{1}{K}\sum_{k=1}^K [(i) + (ii) + (iii) + (iv) + (v)]$. Thus, plugging in the above upper bounds for terms $(i), (ii), (iii), (iv)$, and $(v)$, setting the parameters $\iota_k$, $\gamma_k$ and, $\lambda_k$ as in Theorem 3.6, we get the desired result. Please see Section D.3 in Appendix for a detailed proof.

## 6. Proof of Concept Experiments

In this section, we present the experimental justification of the UCB-based exploration in practice inspired by our theory. The codes are available at `https://github.com/Baichenjia/Contrastive-UCB`.

### 6.1. Implementation of Bonus

**Representation Learning with SPR.** Our goal is to examine whether the proposed UCB bonus practically enhances the exploration of deep RL algorithms with contrastive learning. To this end, we adopt the SPR method (Schwarzer et al., 2021), the state-of-the-art RL approach with contrastive learning on the benchmark Atari 100K (Kaiser et al., 2020). SPR utilizes the temporal information and learns the representation via maximizing the similarity between the future state representations and the corresponding predicted next state representations based on the observed state and action sequences. The representation learning under the framework of SPR is different from the proposed representation learning from the following aspects: **(1)** SPR considers multi-step consistency in addition to the one-step prediction of our proposed contrastive objective, namely, SPR incorporates the information of multiple steps ahead of $(s_h, a_h)$ in the representation $\widehat{\phi}(s_h, a_h)$. Although representation learning with one-step prediction is sufficient according to our theory, such a multi-step approach further enhances the temporal consistency of the learned representation empirically. Similar techniques also arise in various empirical studies (Oord et al., 2018a; Guo et al., 2018). **(2)** SPR utilizes the cosine similarity to maximize the similarity of state-action representations and the embeddings of the corresponding next states. We remark that we adopt the architecture of SPR as an empirical simplification to our proposed contrastive objective, which does not require explicit negative sampling and the corresponding parameter tuning (Schwarzer et al., 2021). This leads to better computational efficiency and avoidance of defining an improper negative sampling distribution. In addition, we remark that the representations obtained from SPR contain sufficient temporal information of the transition dynamics required for exploration, as shown in our experiments.

**Architecture and UCB Bonus.** In our experiments, we adopt the same architecture as SPR. We further construct the UCB bonus based on SPR and propose the SPR-UCB method. In particular, we adopt the same hyper-parameters as that of SPR (Schwarzer et al., 2021). Meanwhile, we adopt the last layer of the Q-network as our learned rep-

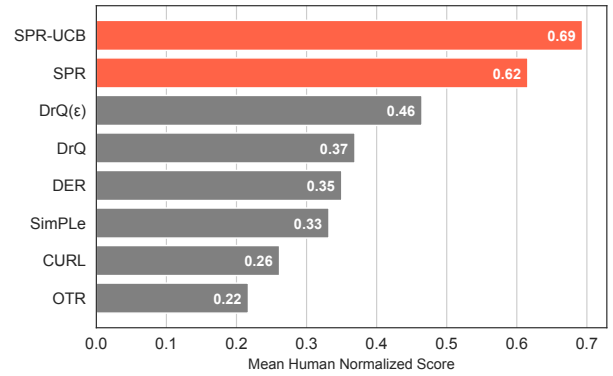


*Figure 1.* Mean human-normalized score in Atari-100K benchmark. The results of baseline algorithms are adopted from Agarwal et al. (2021). We observe that SPR-UCB outperforms SPR and other baseline algorithms.

resentation $\widehat{\phi}$ which is linear in the estimated Q-function. In the training stage, we update the empirical covariance matrix $\widehat{\Sigma}_h^k \in \mathbb{R}^{d\times d}$ by adding the feature covariance $\widehat{\phi}(s_h^k, a_h^k)\widehat{\phi}(s_h^k, a_h^k)^\top$ over the sampled transition tuples $\{(s_h^k, a_h^k, s_h^{k+1})\}_{h\in[H]}$ from the replay buffer, where $\widehat{\phi} \in \mathbb{R}^{d\times 1}$ is the learned representation from the Q-network of SPR. The transition data is sampled from the interaction history. The bonus for the state-action pair $(s, a)$ is calculated by $\beta^k(s, a) = \gamma_k \cdot [\widehat{\phi}(s, a)^\top (\widehat{\Sigma}_h^k)^{-1} \widehat{\phi}(s, a)]^{\frac{1}{2}}$, where we set the hyperparameter $\gamma_k = 1$ for all iterations $k \in [K]$. Upon computing the bonus for each state-action pair of the sampled transition tuples from the replay buffer, we follow our proposed update in Algorithm 1 and add the bonus on the target of Q-functions in fitting the Q-network.

### 6.2. Environments and Baselines

In our experiments, we use Atari 100K (Kaiser et al., 2020) benchmark for evaluation, which contains 26 Atari games from various domains. The benchmark Atari 100K only allows the agent to interact with the environment for 100K steps. Such a setup aims to test the sample efficiency of RL algorithms.

We compare the SPR-UCB method with several baselines in Atari 100K benchmark, including **(1)** SimPLe (Kaiser et al., 2020), which learns an environment model based on the video prediction task and trains a policy under the learned model; **(2)** DER (van Hasselt et al., 2019) and **(3)** OTR (Kielak, 2020), which improve Rainbow (van Hasselt et al., 2019) to perform sample-efficient model-free RL; **(4)** CURL (Laskin et al., 2020), which incorporates contrastive learning based on data augmentation; **(5)** DrQ (Yarats et al., 2021), which directly utilizes data augmentation based on the image observations; and **(6)** SPR (Schwarzer et al., 2021), which learns temporal consistent representation for model-free RL. For all methods, we calculate the human normalized score by $\frac{\text{agent score}-\text{random score}}{\text{human score}-\text{random score}}$. In our experiments, we run the proposed SPR-UCB over 10 different random seeds.

### 6.3. Result Comparison

We illustrate the aggregated mean of human normalized scores among all tasks in Figure 1. We report the score for each task in Appendix F. In our experiments, we observe that **(1)** Both SPR and SPR-UCB outperform baselines that do not learn temporal consistent representations significantly, including DER, OTR, SimPLe, CURL, and DrQ. **(2)** By incorporating the UCB bonus, SPR-UCB outperforms SPR. In addition, we remark that SPR-UCB outperforms SPR significantly in challenging environments including *Boxing*, *Freeway*, *Frostbite*, *KungfuMaster*, *PrivateEye*, and *RoadRunner*. Please see Appendix F for the details.

## 7. Conclusion

We study contrastive-learning empowered RL for MDPs and MGs with low-rank transitions. We propose novel online RL algorithms that incorporate such a contrastive loss with temporal information for MDPs or MGs. We further theoretically prove that our algorithms recover the true representations and simultaneously achieve sample efficiency in learning the optimal policy and Nash equilibrium in MDPs and MGs respectively. We also provide empirical studies to demonstrate the efficacy of the UCB-based contrastive learning method for RL. To the best of our knowledge, we provide the first provably efficient online RL algorithm that incorporates contrastive learning for representation learning.

## Acknowledgements

The authors would like to thank all reviewers for their valuable comments. The authors would also like to thank Sirui Zheng for helpful discussions. The contribution from Chenjia Bai was made during his time as a visiting student at the University of Toronto (Vector Institute for Artificial Intelligence), working with Animesh Garg. The theory, methods, and codes developed in this paper are shared publicly without any proprietary or other restrictions.

## References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *arXiv preprint arXiv:2006.10814*, 2020.

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M.-A., and Hjelm, R. D. Unsupervised state representation learning in atari. *arXiv preprint arXiv:1906.08226*, 2019.

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. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020.

Bellemare, M., Dabney, W., Dadashi, R., Ali Taiga, A., Castro, P. S., Le Roux, N., Schuurmans, D., Lattimore, T., and Lyle, C. A geometric perspective on optimal representations for reinforcement learning. *Advances in neural information processing systems*, 32:4358–4369, 2019.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Chen, Z., Zhou, D., and Gu, Q. Almost optimal algorithms for two-player zero-sum markov games with linear function approximation.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019a.

Du, S. S., Kakade, S. M., Wang, R., and Yang, L. F. Is a good representation sufficient for sample efficient reinforcement learning? *arXiv preprint arXiv:1910.03016*, 2019b.

Dwibedi, D., Tompson, J., Lynch, C., and Sermanet, P. Learning actionable representations from visual observations. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1577–1584. IEEE, 2018.

François-Lavet, V., Bengio, Y., Precup, D., and Pineau, J. Combined reinforcement learning via abstract representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3582–3589, 2019.

Gelada, C., Kumar, S., Buckman, J., Nachum, O., and Bellemare, M. G. Deepmdp: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, pp. 2170–2179. PMLR, 2019.

Guo, Z. D., Azar, M. G., Piot, B., Pires, B. A., and Munos, R. Neural predictive belief representations. *arXiv preprint arXiv:1811.06407*, 2018.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pp. 2555–2565. PMLR, 2019b.

Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv:1611.05397*, 2016.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Kaiser, Ł., Babaeizadeh, M., Miłos, P., Osiński, B., Campbell, R. H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., et al. Model based reinforcement learning for atari. In *International Conference on Learning Representations*, 2020.

Kielak, K. P. Do recent advancements in model-based deep reinforcement learning really improve data efficiency?, 2020. URL https://openreview.net/forum?id=Bke9u1HFwB.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Laskin, M., Srinivas, A., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. In *International Conference on Machine Learning*, pp. 5639–5650. PMLR, 2020.

Liu, G., Zhang, C., Zhao, L., Qin, T., Zhu, J., Li, J., Yu, N., and Liu, T.-Y. Return-based contrastive representation learning for reinforcement learning. *arXiv preprint arXiv:2102.10960*, 2021.

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pp. 6961–6971. PMLR, 2020.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

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.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018a.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018b.

Sallab, A. E., Abdou, M., Perot, E., and Yogamani, S. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.

Schwarzer, M., Anand, A., Goel, R., Hjelm, R. D., Courville, A., and Bachman, P. Data-efficient reinforcement learning with self-predictive representations. *arXiv preprint arXiv:2007.05929*, 2020.

Schwarzer, M., Anand, A., Goel, R., Hjelm, R. D., Courville, A., and Bachman, P. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021.

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., and Brain, G. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 1134–1141. IEEE, 2018.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Srinivas, A., Laskin, M., and Abbeel, P. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

Stooke, A., Lee, K., Abbeel, P., and Laskin, M. Decoupling representation learning from reinforcement learning. In *International Conference on Machine Learning*, pp. 9870–9879. PMLR, 2021.

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

Taiga, A. A., Fedus, W., Machado, M. C., Courville, A., and Bellemare, M. G. On bonus based exploration methods in the arcade learning environment. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJewlyStDr.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021.

Van de Geer, S. A. *Applications of empirical process theory*, volume 91. Cambridge University Press Cambridge, 2000.

van Hasselt, H. P., Hessel, M., and Aslanides, J. When to use parametric models in reinforcement learning? *Advances in Neural Information Processing Systems*, 32:14322–14333, 2019.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

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

Yang, L. and Wang, M. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pp. 6995–7004. PMLR, 2019.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Yang, M. and Nachum, O. Representation matters: Offline pretraining for sequential decision making. *arXiv preprint arXiv:2102.05815*, 2021.

Yarats, D., Kostrikov, I., and Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021.

Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. *arXiv preprint arXiv:2103.12923*, 2021.

Zhang, A., McAllister, R., Calandra, R., Gal, Y., and Levine, S. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.

Zhang, T., Ren, T., Yang, M., Gonzalez, J. E., Schuurmans, D., and Dai, B. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*. PMLR, 2022.

Zhou, D., Gu, Q., and Szepesvari, C. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021b.

# A. Omitted Sampling Algorithm

**Algorithm 3** Contrastive Data Sampling for Single-Agent MDP

**for** step $h = 1, \ldots, H-1$ **do**
  Sample $(\widetilde{s}_h^k, \widetilde{a}_h^k) \sim \widetilde{d}_h^{\pi^{k-1}}(\cdot,\cdot)$, $\widetilde{s}_{h+1}^k \sim \mathbb{P}_h(\cdot|\widetilde{s}_h^k, \widetilde{a}_h^k)$
  Let $\breve{s}_{h+1}^k = \widetilde{s}_{h+1}^k$. Sample $\breve{a}_{h+1}^k \sim \mathrm{Unif}(\mathcal{A})$, $\breve{s}_{h+2}^k = \mathbb{P}_{h+1}(\cdot|\breve{s}_{h+1}^k, \breve{a}_{h+1}^k)$, and $y_h^k \sim \mathrm{Ber}(1/2)$
  $\widetilde{\mathcal{D}}_h^k = \widetilde{\mathcal{D}}_h^{k-1} \cup \{(\widetilde{s}_h^k, \widetilde{a}_h^k)\}$.
  **if** $y_h^k = 1$ **then**
    $\mathcal{D}_h^k = \mathcal{D}_h^{k-1} \cup \{(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{s}_{h+1}^k, 1)\}$.
    $\mathcal{D}_{h+1}^k = \mathcal{D}_{h+1}^{k-1} \cup \{(\breve{s}_{h+1}^k, \breve{a}_{h+1}^k, \breve{s}_{h+2}^k, 1)\}$.
  **else if** $y_h^k = 0$ **then**
    Sample negative transition $\widetilde{s}_{h+1}^{k,-}, \breve{s}_{h+2}^{k,-} \sim \mathcal{P}_{\mathcal{S}}^-(\cdot)$.
    $\mathcal{D}_h^k = \mathcal{D}_h^{k-1} \cup \{(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{s}_{h+1}^{k,-}, 0)\}$.
    $\mathcal{D}_{h+1}^k = \mathcal{D}_{h+1}^{k-1} \cup \{(\breve{s}_{h+1}^k, \breve{a}_{h+1}^k, \breve{s}_{h+2}^{k,-}, 0)\}$.
  **end if**
**end for**
$(\widetilde{s}_H^k, \widetilde{a}_H^k) \sim \widetilde{d}_H^{\pi^{k-1}}(\cdot,\cdot)$, $\widetilde{s}_{H+1}^k \sim \mathbb{P}_h(\cdot|\widetilde{s}_H^k, \widetilde{a}_H^k)$, and $y_H^k \sim \mathrm{Ber}(1/2)$
$\widetilde{\mathcal{D}}_H^k = \widetilde{\mathcal{D}}_H^{k-1} \cup \{(\widetilde{s}_H^k, \widetilde{a}_H^k)\}$.
**if** $y_H^k = 1$ **then**
  $\mathcal{D}_H^k = \mathcal{D}_H^{k-1} \cup \{(\widetilde{s}_H^k, \widetilde{a}_H^k, \widetilde{s}_{H+1}^k, 1)\}$.
**else if** $y_H^k = 0$ **then**
  Sample negative transition $\widetilde{s}_{H+1}^{k,-} \sim \mathcal{P}_{\mathcal{S}}^-(\cdot)$.
  $\mathcal{D}_H^k = \mathcal{D}_H^{k-1} \cup \{(\widetilde{s}_H^k, \widetilde{a}_H^k, \widetilde{s}_{H+1}^{k,-}, 0)\}$.
**end if**
**return** $\{\mathcal{D}_h^k\}_{h=1}^H$ and $\{\widetilde{\mathcal{D}}_h^k\}_{h=1}^H$.

**Algorithm 4** Contrastive Data Sampling for Markov Game

**for** step $h = 1, \ldots, H-1$ **do**
  Sample $(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{b}_h^k) \sim \widetilde{d}_h^{\pi^{k-1}}(\cdot,\cdot,\cdot)$, $\widetilde{s}_{h+1}^k \sim \mathbb{P}_h(\cdot|\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{b}_h^k)$
  Let $\breve{s}_{h+1}^k = \widetilde{s}_{h+1}^k$. Sample $\breve{a}_{h+1}^k \sim \mathrm{Unif}(\mathcal{A})$, $\breve{b}_{h+1}^k \sim \mathrm{Unif}(\mathcal{B})$, $\breve{s}_{h+2}^k = \mathbb{P}_{h+1}(\cdot|\breve{s}_{h+1}^k, \breve{a}_{h+1}^k, \breve{b}_{h+1}^k)$, $y_h^k \sim \mathrm{Ber}(1/2)$
  $\widetilde{\mathcal{D}}_h^k = \widetilde{\mathcal{D}}_h^{k-1} \cup \{(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{b}_h^k)\}$.
  **if** $y_h^k = 1$ **then**
    $\mathcal{D}_h^k = \mathcal{D}_h^{k-1} \cup \{(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{b}_h^k, \widetilde{s}_{h+1}^k, 1)\}$.
    $\mathcal{D}_{h+1}^k = \mathcal{D}_{h+1}^{k-1} \cup \{(\breve{s}_{h+1}^k, \breve{a}_{h+1}^k, \breve{b}_{h+1}^k, \breve{s}_{h+2}^k, 1)\}$.
  **else if** $y_h^k = 0$ **then**
    Sample negative transition $\widetilde{s}_{h+1}^{k,-}, \breve{s}_{h+2}^{k,-} \sim \mathcal{P}_{\mathcal{S}}^-(\cdot)$.
    $\mathcal{D}_h^k = \mathcal{D}_h^{k-1} \cup \{(\widetilde{s}_h^k, \widetilde{a}_h^k, \widetilde{b}_h^k, \widetilde{s}_{h+1}^{k,-}, 0)\}$.
    $\mathcal{D}_{h+1}^k = \mathcal{D}_{h+1}^{k-1} \cup \{(\breve{s}_{h+1}^k, \breve{a}_{h+1}^k, \breve{b}_{h+1}^k, \breve{s}_{h+2}^{k,-}, 0)\}$.
  **end if**
**end for**
Sample $(\widetilde{s}_H^k, \widetilde{a}_H^k, \widetilde{b}_H^k) \sim \widetilde{d}_H^{\pi^{k-1}}(\cdot,\cdot,\cdot)$, $\widetilde{s}_{H+1}^k \sim \mathbb{P}_h(\cdot|\widetilde{s}_H^k, \widetilde{a}_H^k, \widetilde{b}_H^k)$, and $y_H^k \sim \mathrm{Ber}(1/2)$
$\widetilde{\mathcal{D}}_H^k = \widetilde{\mathcal{D}}_H^{k-1} \cup \{(\widetilde{s}_H^k, \widetilde{a}_H^k, \widetilde{b}_H^k)\}$.
**if** $y_H^k = 1$ **then**
  $\mathcal{D}_H^k = \mathcal{D}_H^{k-1} \cup \{(\widetilde{s}_H^k, \widetilde{a}_H^k, \widetilde{b}_H^k, \widetilde{s}_{H+1}^k, 1)\}$.
**else if** $y_H^k = 0$ **then**
  Sample negative transition $\widetilde{s}_{H+1}^{k,-} \sim \mathcal{P}_{\mathcal{S}}^-(\cdot)$.
  $\mathcal{D}_H^k = \mathcal{D}_H^{k-1} \cup \{(\widetilde{s}_H^k, \widetilde{a}_H^k, \widetilde{b}_H^k, \widetilde{s}_{H+1}^{k,-}, 0)\}$.
**end if**
**return** $\{\mathcal{D}_h^k\}_{h=1}^H$ and $\{\widetilde{\mathcal{D}}_h^k\}_{h=1}^H$.

## B. Table of Notation

We present the following table of notations. We denote by $\sigma$ an arbitrary joint policy. If the joint policy $\sigma$ is equivalent to a product of two separate policies for each player, i.e., $\sigma(a,b|s)=\pi(a|s)\times\nu(b|s)$, then we can replace $\sigma$ by $\pi,\nu$.

*Table 1.* Table of Notation

| Notation | Meaning |
|---|---|
| $d_h^\pi(s)$ | state probability at step $h$ under the true transition $\mathbb{P}$ and a policy $\pi$ |
| $d_h^\pi(s,a)$ | state-action probability at step $h$ under the true transition $\mathbb{P}$ and a policy $\pi$ |
| $\widetilde{d}_h^\pi(s,a)$ | $\widetilde{d}_h^\pi(s,a):=d_h^\pi(s)\mathrm{Unif}(a)$ |
| $\breve{d}_h^\pi(s,a)$ | $\breve{d}_h^\pi(s,a):=\widetilde{d}_{h-1}^\pi(s',a')\mathbb{P}_{h-1}(s\|s',a')\mathrm{Unif}(a)$ |
| $\rho_h^k(s,a)$ | $\rho_h^k(s,a):=1/k\cdot\sum_{k'=0}^{k-1}d_h^{\pi^{k'}}(s,a)$ |
| $\widetilde{\rho}_h^k(s,a)$ | $\widetilde{\rho}_h^k(s,a):=1/k\cdot\sum_{k'=0}^{k-1}\widetilde{d}_h^{\pi^{k'}}(s,a)$ |
| $\breve{\rho}_h^k(s,a)$ | $\breve{\rho}_h^k(s,a):=1/k\cdot\sum_{k'=0}^{k-1}\breve{d}_h^{\pi^{k'}}(s,a)$ |
| $\Sigma_{\rho,\phi}$ | covariance matrix defined as $k\cdot\mathbb{E}_{(s,a)\sim\rho_h^k(\cdot,\cdot)}\left[\phi(s,a)\phi(s,a)^\top\right]+\lambda_k I$ for any $\rho$ and $\phi$ |
| $d_h^\sigma(s)$ | state probability at step $h$ under the true transition $\mathbb{P}$ and a joint policy $\sigma$ |
| $d_h^\sigma(s,a,b)$ | state-action probability at step $h$ under the true transition $\mathbb{P}$ and a joint policy $\sigma$ |
| $\widetilde{d}_h^\sigma(s,a,b)$ | $\widetilde{d}_h^\sigma(s,a,b):=d_h^\sigma(s)\mathrm{Unif}(a)\mathrm{Unif}(b)$ |
| $\breve{d}_h^\sigma(s,a,b)$ | $\breve{d}_h^\sigma(s,a,b):=\widetilde{d}_{h-1}^\sigma(s',a',b')\mathbb{P}_{h-1}(s\|s',a',b')\mathrm{Unif}(a)\mathrm{Unif}(b)$ |
| $\rho_h^k(s,a,b)$ | $\rho_h^k(s,a,b):=1/k\cdot\sum_{k'=0}^{k-1}d_h^{\sigma^{k'}}(s,a,b)$ |
| $\widetilde{\rho}_h^k(s,a,b)$ | $\widetilde{\rho}_h^k(s,a,b):=1/k\cdot\sum_{k'=0}^{k-1}\widetilde{d}_h^{\sigma^{k'}}(s,a,b)$ |
| $\breve{\rho}_h^k(s,a,b)$ | $\breve{\rho}_h^k(s,a,b):=1/k\cdot\sum_{k'=0}^{k-1}\breve{d}_h^{\sigma^{k'}}(s,a,b)$ |
| $\Sigma_{\rho,\phi}$ | covariance matrix defined as $k\cdot\mathbb{E}_{(s,a,b)\sim\rho(\cdot,\cdot,\cdot)}\left[\phi(s,a,b)\phi(s,a,b)^\top\right]+\lambda_k I$ |
| $V_h^\pi,Q_h^\pi$ | value and Q-functions at step $h$ under the policy $\pi$ and the true transition and reward $\mathbb{P},r$ |
| $\overline{V}_h^k,\overline{Q}_h^k$ | value and Q-functions generated in Lines 11 and 12 of Algorithm 1 |
| $\overline{V}_{k,h}^\pi,\overline{Q}_{k,h}^\pi$ | value and Q-functions at step $h$ on the auxiliary MDP defined by $r+\beta^k$ and $\widehat{\mathbb{P}}^k$ |
| $V_h^\sigma,Q_h^\sigma$ | value and Q-functions at step $h$ under the joint policy $\sigma$ and the true transition and reward $\mathbb{P},r$ |
| $\overline{V}_h^k,\overline{Q}_h^k$ | value and Q-functions generated in Lines 11 and 13 of Algorithm 2 |
| $\underline{V}_h^k,\underline{Q}_h^k$ | value and Q-functions generated in Lines 12 and 14 of Algorithm 2 |
| $\overline{V}_{k,h}^\sigma,\overline{Q}_{k,h}^\sigma$ | value and Q-functions at step $h$ on the auxiliary MG defined by $r+\beta^k$ and $\widehat{\mathbb{P}}^k$ |
| $\underline{V}_{k,h}^\sigma,\underline{Q}_{k,h}^\sigma$ | value and Q-functions at step $h$ on the auxiliary MG defined by $r-\beta^k$ and $\widehat{\mathbb{P}}^k$ |
| $\mathrm{Unif}(\mathcal{A}),\mathrm{Unif}(\mathcal{B})$ | uniform distribution over spaces $\mathcal{A}$ or $\mathcal{B}$ |
| $\mathrm{Unif}(a),\mathrm{Unif}(b)$ | probabilities for the above distributions: $\mathrm{Unif}(a)=1/\|\mathcal{A}\|$ and $\mathrm{Unif}(b)=1/\|\mathcal{B}\|$ |
| $\|\cdot\|_{\mathrm{TV}}$ | total variation distance |
| $\|\cdot\|_1$ | define $\|f\|_1:=\int_x\|f(x)\|\mathrm{d}x$ |

## C. Theoretical Analysis for Single-Agent MDP

### C.1. Lemmas

**Lemma C.1** (Learning Target of Contrastive Loss)**.** *For any $(s,a)\in\mathcal{S}\times\mathcal{A}$ that is reachable under certain sampling strategy, the learning target of the contrastive loss in (2) is*

$$f_h^*(s,a,s')=\frac{\mathbb{P}_h(s'|s,a)}{\mathcal{P}_{\mathcal{S}}^-(s')}.$$

*Proof.* For any $h\in[H]$, we let $\Pr_h$ denote the probability for some event at the $h$-th step of an MDP. Our contrastive loss in (2) implicitly assumes

$$\Pr{}_h(y|s,a,s')=\left(\frac{f_h^*(s,a,s')}{1+f_h^*(s,a,s')}\right)^y\left(\frac{1}{1+f_h^*(s,a,s')}\right)^{1-y}.$$

On the other hand, by Bayes' rule, we know $\Pr_h(y|s,a,s')$ can be rewritten as

$$\Pr_h(y|s,a,s') = \frac{\Pr_h(s,a,s'|y)\Pr_h(y)}{\sum_{y\in\{0,1\}}\Pr_h(s,a,s'|y)\Pr_h(y)} = \frac{\Pr_h(s,a,s'|y)}{\Pr_h(s,a)\mathbb{P}_h(s'|s,a)+\Pr_h(s,a)\mathcal{P}_{\mathcal{S}}^-(s')},$$

where the last equation uses the fact that $\Pr_h(y) = 1/2$ for any $y \in \{0,1\}$ at the $h$-th step according to our sampling algorithm. In the last equality, we also have

$$\Pr_h(s,a,s'|y=1) = \Pr_h(s,a|y=1)\Pr_h(s'|y=1,s,a) = \Pr_h(s,a)\mathbb{P}_h(s'|s,a),$$
$$\Pr_h(s,a,s'|y=0) = \Pr_h(s,a|y=0)\Pr_h(s'|y=0,s,a) = \Pr_h(s,a)\mathcal{P}_{\mathcal{S}}^-(s'),$$

where we use $\Pr_h(s,a|y=1) = \Pr_h(s,a|y=0) = \Pr_h(s,a)$ since $(s,a)$ and $y$ are independent at each step, and also $\Pr_h(s'|y=1,s,a) = \mathbb{P}_h(s'|s,a)$ as well as $\Pr_h(s'|y=0,s,a) = \mathcal{P}_{\mathcal{S}}^-(s')$.

Therefore, combining the above results, when $y=1$ at the $h$-th step, we obtain

$$\frac{f_h^*(s,a,s')}{1+f_h^*(s,a,s')} = \frac{\Pr_h(s,a)\mathbb{P}_h(s'|s,a)}{\Pr_h(s,a)\mathbb{P}_h(s'|s,a)+\Pr_h(s,a)\mathcal{P}_{\mathcal{S}}^-(s')},$$

which further gives

$$f_h^*(s,a,s') = \frac{\mathbb{P}_h(s'|s,a)}{\mathcal{P}_{\mathcal{S}}^-(s')},$$

since$(s,a)$ is reachable under the sampling algorithm, namely $\Pr_h(s,a) > 0$. Equivalently, when $y=0$, we get the same result. This completes the proof. □

**Lemma C.2.** *Let $\pi^* := \operatorname{argmax}_\pi V_1^\pi(s_1)$ be the optimal policy and $\overline{V}_{k,1}^\pi$ be the value function under any policy $\pi$ associated with an MDP defined by the reward function $r+\beta^k$ and the estimated transition $\widehat{\mathbb{P}}^k$ with $\beta^k$ and $\widehat{\mathbb{P}}^k$ obtained at episode $k$ of Algorithm 1. We have the decomposition of the difference between the following two value functions as*

$$V_1^{\pi^*}(s_1) - \overline{V}_{k,1}^{\pi^*}(s_1) = \mathbb{E}\left[\sum_{h=1}^H \left(-\beta_h^k(s_h,a_h) + (\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)V_{h+1}^{\pi^*}(s_h,a_h)\right) \,\middle|\, \pi^*, \widehat{\mathbb{P}}^k\right].$$

*Proof.* We consider two MDPs defined by $(\mathcal{S},\mathcal{A},H,r,\mathbb{P})$ and $(\mathcal{S},\mathcal{A},H,r+\beta,\mathbb{P}')$ where $\mathbb{P}$ and $\mathbb{P}'$ are any transition models and $r$ and $\beta$ are arbitrary reward function and bonus term. Then, for any deterministic policy $\pi$, we let $Q_h^\pi$ and $V_h^\pi$ be the associated Q-function and value function at the $h$-th step for the MDP defined by $(\mathcal{S},\mathcal{A},H,r,\mathbb{P})$, and $\widetilde{Q}_h^\pi$ and $\widetilde{V}_h^\pi$ be the associated Q-function and value function at the $h$-th step for the MDP defined by $(\mathcal{S},\mathcal{A},H,r+\beta,\mathbb{P}')$. Then, we have for any $(s_h,a_h) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}
&Q_h^\pi(s_h,a_h) - \widetilde{Q}_h^\pi(s_h,a_h)\\
&\quad= -\beta_h(s_h,a_h) + \mathbb{P}_h V_{h+1}^\pi(s_h,a_h) - \mathbb{P}_h'\widetilde{V}_{h+1}^\pi(s_h,a_h)\\
&\quad= -\beta_h(s_h,a_h) + \mathbb{P}_h V_{h+1}^\pi(s_h,a_h) - \mathbb{P}_h' V_{h+1}^\pi(s_h,a_h) + \mathbb{P}_h' V_{h+1}^\pi(s_h,a_h) - \mathbb{P}_h'\widetilde{V}_{h+1}^\pi(s_h,a_h)\\
&\quad= -\beta_h(s_h,a_h) + (\mathbb{P}_h - \mathbb{P}_h')V_{h+1}^\pi(s_h,a_h) + \mathbb{P}_h'[V_{h+1}^\pi(s_h,a_h) - \widetilde{V}_{h+1}^\pi(s_h,a_h)],
\end{aligned}$$

where we use the Bellman equation for the above equalities. Thus, further by the Bellman equation and the above result, we have

$$\begin{aligned}
&V_h^\pi(s_h) - \widetilde{V}_h^\pi(s_h)\\
&\quad= Q_h^\pi(s_h,\pi_h(s_h)) - \widetilde{Q}_h^\pi(s_h,\pi_h(s_h))\\
&\quad= -b_h(s_h,\pi_h(s_h)) + (\mathbb{P}_h - \mathbb{P}_h')V_{h+1}^\pi(s_h,\pi_h(s_h)) + \mathbb{P}_h'[V_{h+1}^\pi(s_h,\pi_h(s_h)) - \widetilde{V}_{h+1}^\pi(s_h,\pi_h(s_h))].
\end{aligned}$$

By the fact that $V_{H+1}^\pi(s) = \widetilde{V}_{H+1}^\pi(s) = 0$ for any $s\in\mathcal{S}$ and $\pi$, recursively applying the above relation, we have

$$V_1^\pi(s_1) - \widetilde{V}_1^\pi(s_1) = \mathbb{E}\left[\sum_{h=1}^H \left(-\beta_h(s_h,a_h) + (\mathbb{P}_h - \mathbb{P}_h')V_{h+1}^\pi(s_h,a_h)\right) \,\middle|\, \pi, \mathbb{P}'\right].$$

Note that the above results can be straightforwardly extended to any randomized policy $\pi = \{\pi_h\}_{h=1}^H$ with $\pi_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$.

For any episode $k$, setting $\mathbb{P}', \pi, \beta$ to be $\widehat{\mathbb{P}}^k, \pi^*, \beta^k$ defined in Algorithm 1 and $\mathbb{P}, r$ to be the true transition model and reward function, by the above equation and the definition of $V_h^\pi$ and $\overline{V}_h^\pi$, we obtain

$$V_1^{\pi^*}(s_1) - \overline{V}_{k,1}^{\pi^*}(s_1) = \mathbb{E}\left[\sum_{h=1}^H \left(-\beta_h^k(s_h, a_h) + (\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)V_{h+1}^{\pi^*}(s_h, a_h)\right) \,\middle|\, \pi^*, \widehat{\mathbb{P}}^k\right].$$

This completes the proof. □

**Lemma C.3.** *Let $\pi^k$ be the learned policy at episode $k$ of Algorithm 1 and $\overline{V}_{k,1}^\pi$ be the value function under any policy $\pi$ associated with an MDP defined by the reward function $r + \beta^k$ and the estimated transition $\widehat{\mathbb{P}}^k$ with $\beta^k$ and $\widehat{\mathbb{P}}^k$ obtained at episode $k$ of Algorithm 1. We have the decomposition of the difference between the following two value functions as*

$$V_1^{\pi^k}(s_1) - \overline{V}_{k,1}^{\pi^k}(s_1) = \mathbb{E}\left[\sum_{h=1}^H \left(-\beta_h^k(s_h, a_h) + (\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)\overline{V}_{h+1}^{\pi^k}(s_h, a_h)\right) \,\middle|\, \pi^k, \mathbb{P}\right].$$

*Proof.* Similar to Proof of Lemma C.2, we consider two arbitrary MDPs defined by $(\mathcal{S}, \mathcal{A}, H, r, \mathbb{P})$ and $(\mathcal{S}, \mathcal{A}, H, r + \beta, \mathbb{P}')$. For any deterministic policy $\pi$, let $Q_h^\pi$ and $V_h^\pi$ be the associated Q-function and value function at the $h$-th step for the MDP defined by $(\mathcal{S}, \mathcal{A}, H, r, \mathbb{P})$, and $\widetilde{Q}_h^\pi$ and $\widetilde{V}_h^\pi$ be the associated Q-function and value function at the $h$-th step for the MDP defined by $(\mathcal{S}, \mathcal{A}, H, r + \beta, \mathbb{P}')$. For any $(s_h, a_h) \in \mathcal{S} \times \mathcal{A}$, by Bellman equation, we have

$$\begin{aligned}
&Q_h^\pi(s_h, a_h) - \widetilde{Q}_h^\pi(s_h, a_h) \\
&\quad = -\beta_h(s_h, a_h) + \mathbb{P}_h V_{h+1}^\pi(s_h, a_h) - \mathbb{P}'_h \widetilde{V}_{h+1}^\pi(s_h, a_h) \\
&\quad = -\beta_h(s_h, a_h) + \mathbb{P}_h V_{h+1}^\pi(s_h, a_h) - \mathbb{P}_h \widetilde{V}_{h+1}^\pi(s_h, a_h) + \mathbb{P}_h \widetilde{V}_{h+1}^\pi(s_h, a_h) - \mathbb{P}'_h \widetilde{V}_{h+1}^\pi(s_h, a_h) \\
&\quad = -\beta_h(s_h, a_h) + \mathbb{P}_h[V_{h+1}^\pi(s_h, a_h) - \widetilde{V}_{h+1}^\pi(s_h, a_h)] + (\mathbb{P}_h - \mathbb{P}'_h)\widetilde{V}_{h+1}^\pi(s_h, a_h).
\end{aligned}$$

Then, further by the Bellman equation and the above result, we have

$$\begin{aligned}
&V_h^\pi(s_h) - \widetilde{V}_h^\pi(s_h) \\
&\quad = Q_h^\pi(s_h, \pi_h(s_h)) - \widetilde{Q}_h^\pi(s_h, \pi_h(s_h)) \\
&\quad = -\beta_h(s_h, \pi_h(s_h)) + \mathbb{P}_h[V_{h+1}^\pi(s_h, \pi_h(s_h)) - \widetilde{V}_{h+1}^\pi(s_h, \pi_h(s_h))] + (\mathbb{P}_h - \mathbb{P}'_h)\widetilde{V}_{h+1}^\pi(s_h, \pi_h(s_h))].
\end{aligned}$$

By the fact that $V_{H+1}^\pi(s) = \widetilde{V}_{H+1}^\pi(s) = 0$ for any $s \in \mathcal{S}$ and $\pi$, recursively applying the above relation, we have

$$V_1^\pi(s_1) - \widetilde{V}_1^\pi(s_1) = \mathbb{E}\left[\sum_{h=1}^H \left(-\beta_h(s_h, a_h) + (\mathbb{P}_h - \mathbb{P}'_h)\widetilde{V}_{h+1}^\pi(s_h, a_h)\right) \,\middle|\, \pi, \mathbb{P}\right].$$

The above results can be straightforwardly extended to any randomized policy $\pi = \{\pi_h\}_{h=1}^H$ with $\pi_h : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$.

For any episode $k$, setting $\mathbb{P}', \pi, \beta$ to be $\widehat{\mathbb{P}}^k, \pi^k, \beta^k$ defined in Algorithm 1 and $\mathbb{P}, r$ to be the true transition model and reward function, by the above equation and the definition of $V_h^\pi$ and $\overline{V}_h^\pi$, we obtain

$$V_1^{\pi^k}(s_1) - \overline{V}_{k,1}^{\pi^k}(s_1) = \mathbb{E}\left[\sum_{h=1}^H \left(-\beta_h^k(s_h, a_h) + (\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)\overline{V}_{h+1}^{\pi^k}(s_h, a_h)\right) \,\middle|\, \pi^k, \mathbb{P}\right].$$

This completes the proof. □

**Lemma C.4.** *Let $\widehat{\mathbb{P}}^k$ be the estimated transition obtained at episode $k$ of Algorithm 1. Define $\zeta_{h-1}^k := \mathbb{E}_{(s'',a'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot)} \|\widehat{\mathbb{P}}_{h-1}^k(\cdot|s'', a'') - \mathbb{P}_{h-1}(\cdot|s'', a'')\|_1^2$ for all $h \geq 2$, $\widetilde{\rho}_h^k(\cdot, \cdot) := \frac{1}{k}\sum_{k'=0}^{k-1} \widetilde{d}_h^{\pi^{k'}}(\cdot, \cdot)$ for all $h \geq 1$ with $\widetilde{\rho}_1^k(s_1, a) = \mathrm{Unif}(a)$,*

*and* $\breve{\rho}_h^k(\cdot,\cdot) := \frac{1}{k}\sum_{k'=0}^{k-1} \breve{d}_h^{\pi^{k'}}(\cdot,\cdot)$ *for all* $h \geq 2$*. Then for any function* $g : \mathcal{S}\times\mathcal{A} \mapsto [0,B]$ *and policy* $\pi$*, we have for any* $h \geq 2$*, the following inequality holds*

$$\begin{aligned}&\left|\mathbb{E}_{(s,a)\sim d_h^{\pi,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}[g(s,a)]\right| \\ &\quad\leq \sqrt{2kB^2\zeta_{h-1}^k + 2k|\mathcal{A}|\cdot\mathbb{E}_{(s,a)\sim\breve{\rho}_h^k(\cdot,\cdot)}[g(s,a)^2] + \lambda_k B^2 d/(C_{\mathcal{S}}^-)^2}\cdot\mathbb{E}_{(s',a')\sim d_{h-1}^{\pi,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}\left\|\widehat{\phi}_{h-1}^k(s',a')\right\|_{\Sigma^{-1}_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}}.\end{aligned}$$

*Moreover, for* $h = 1$*, we have*

$$\left|\mathbb{E}_{(s,a)\sim d_1^{\pi,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}[g(s,a)]\right| = \sqrt{g(s_1,\pi_1(s_1))^2} \leq \sqrt{|\mathcal{A}|\mathbb{E}_{a\sim\widetilde{\rho}_1^k(s_1,\cdot)}[g(s_1,a)^2]},$$

*where* $\widetilde{\rho}_1^k(s_1,a) = \mathrm{Unif}(a)$.

*Proof.* For any function $g : \mathcal{S}\times\mathcal{A}\mapsto[0,B]$ and any deterministic policy $\pi$, under the estimated transition model $\widehat{\mathbb{P}}^k$ at the episode $k$, for any $h \geq 2$, we have

$$\begin{aligned}&\left|\mathbb{E}_{(s,a)\sim d_h^{\pi,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}[g(s,a)]\right| \\ &\quad= \left|\mathbb{E}_{(s',a')\sim d_{h-1}^{\pi,\widehat{\mathbb{P}}^k}(\cdot,\cdot),s\sim\widehat{\mathbb{P}}_{h-1}^k(\cdot|s',a')}[g(s,\pi_h(s))]\right| \\ &\quad= \left|\mathbb{E}_{(s',a')\sim d_{h-1}^{\pi,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}\left[\widehat{\phi}_{h-1}^k(s',a')^\top\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right]\right| \\ &\quad\leq \mathbb{E}_{(s',a')\sim d_{h-1}^{\pi,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}\left\|\widehat{\phi}_{h-1}^k(s',a')\right\|_{\Sigma^{-1}_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}}\left\|\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right\|_{\Sigma_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}},\end{aligned} \tag{8}$$

where the inequality is due to the Cauchy-Schwarz inequality. Hereafter, we define the covariance matrix $\Sigma_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k} := k\mathbb{E}_{(s,a)\sim\widetilde{\rho}_{h-1}^k}[\widehat{\phi}_{h-1}^k(s,a)\widehat{\phi}_{h-1}^k(s,a)^\top] + \lambda_k I$ with $\widetilde{\rho}_{h-1}^k(s,a) = \frac{1}{k}\sum_{k'=0}^{k-1}\widetilde{d}_{h-1}^{\pi^{k'}}(s,a)$.

Next, we can bound

$$\begin{aligned}&\left\|\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right\|^2_{\Sigma_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}} \\ &\quad= k\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right)^\top\mathbb{E}_{\widetilde{\rho}_{h-1}^k}\left[\widehat{\phi}_{h-1}^k(\widehat{\phi}_{h-1}^k)^\top\right]\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right) \\ &\qquad+\lambda_k\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right) \\ &\quad= k\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot)}\left[\int_{\mathcal{S}}\widehat{\phi}_{h-1}^k(s'',a'')^\top\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right] \\ &\qquad+\lambda_k\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right) \\ &\quad\leq k\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot)}\left[\int_{\mathcal{S}}\widehat{\phi}_{h-1}^k(s'',a'')^\top\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right]^2+\lambda_k B^2 d/(C_{\mathcal{S}}^-)^2,\end{aligned} \tag{9}$$

where the last inequality is due to

$$\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)g(s,\pi_h(s))\mathrm{d}s\right) \leq B^2\left|\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\mathrm{d}s\right|_2^2 \leq B^2 d/(C_{\mathcal{S}}^-)^2,$$

since $\|\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\mathrm{d}s\|_2^2 := \|\int_{\mathcal{S}}\mathcal{P}_{\mathcal{S}}^-(s)\widetilde{\psi}_{h-1}^k(s)\mathrm{d}s\|_2^2 \leq \|\int_{\mathcal{S}}\widetilde{\psi}_{h-1}^k(s)\mathrm{d}s\|_2^2 \leq (\int_{\mathcal{S}}\|\widetilde{\psi}_{h-1}^k(s)\|_2\mathrm{d}s)^2 \leq d/(C_{\mathcal{S}}^-)^2$ according to the definition of the function class in Definition 3.3 and the assumption that all states are normalized such that $\mathrm{Vol}(\mathcal{S}) \leq 1$.

Moreover, we have

$$
\begin{aligned}
&k\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}^k_{h-1}(\cdot,\cdot)}\left[\int_{\mathcal{S}}\widehat{\phi}^k_{h-1}(s'',a'')^\top\widehat{\psi}^k_{h-1}(s)g(s,\pi_h(s))\mathrm{d}s\right]^2\\
&\quad\le 2k\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}^k_{h-1}(\cdot,\cdot)}\left[\int_{\mathcal{S}}\left(\widehat{\mathbb{P}}^k_{h-1}(s|s'',a'')-\mathbb{P}_{h-1}(s|s'',a'')\right)g(s,\pi_h(s))\mathrm{d}s\right]^2\\
&\qquad+2k\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}^k_{h-1}(\cdot,\cdot)}\left[\int_{\mathcal{S}}\mathbb{P}_{h-1}(s|s'',a'')g(s,\pi_h(s))\mathrm{d}s\right]^2\\
&\quad\le 2kB^2\zeta^k_{h-1}+2k\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}^k_{h-1}(\cdot,\cdot)}\left[\int_{\mathcal{S}}\mathbb{P}_{h-1}(s|s'',a'')g(s,\pi_h(s))\mathrm{d}s\right]^2\\
&\quad\le 2kB^2\zeta^k_{h-1}+2k\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}^k_{h-1}(\cdot,\cdot),s\sim\mathbb{P}_{h-1}(\cdot|s'',a'')}[g(s,\pi_h(s))^2]\\
&\quad\le 2kB^2\zeta^k_{h-1}+2k\frac{1}{\mathrm{Unif}(a)}\mathbb{E}_{(s,a)\sim\breve{\rho}^k_h(\cdot,\cdot)}[g(s,a)^2]\\
&\quad= 2kB^2\zeta^k_{h-1}+2k|\mathcal{A}|\cdot\mathbb{E}_{(s,a)\sim\breve{\rho}^k_h(\cdot,\cdot)}[g(s,a)^2],
\end{aligned}
\tag{10}
$$

where the first inequality is due to $(x+y)^2\le 2x^2+2y^2$, the second inequality is due to $\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}^k_{h-1}(\cdot,\cdot)}[\int_{\mathcal{S}}(\widehat{\mathbb{P}}^k_{h-1}(s|s'',a'')-\mathbb{P}_{h-1}(s|s'',a''))g(s,\pi_h(s))\mathrm{d}s]^2\le B^2\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}^k_{h-1}(\cdot,\cdot)}\|\widehat{\mathbb{P}}^k_{h-1}(\cdot|s'',a'')-\mathbb{P}_{h-1}(\cdot|s'',a'')\|_1^2=B^2\zeta^k_{h-1}$ with $\zeta^k_{h-1}:=\mathbb{E}_{(s'',a'')\sim\widetilde{\rho}^k_{h-1}(\cdot,\cdot)}\|\widehat{\mathbb{P}}^k_{h-1}(\cdot|s'',a'')-\mathbb{P}_{h-1}(\cdot|s'',a'')\|_1^2$, the third inequality is by Jensen's inequality, and the fourth inequality is due to $g(s,\pi_h(s))^2\le\sum_{a\in\mathcal{A}}g(s,a)^2=1/\mathrm{Unif}(a)\cdot\mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})}[g(s,a)^2]$ and $\breve{\rho}^k_h(s,a):=\widetilde{\rho}^k_{h-1}(s',a')\mathbb{P}_{h-1}(s|s',a')\mathrm{Unif}(a)$ for all $h\ge 2$.

Combining (8),(9), and (10), we have for any $h\ge 2$,

$$
\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim d^{\pi,\widehat{\mathbb{P}}^k}_h(\cdot,\cdot)}[g(s,a)]\right|\\
&\quad\le\sqrt{2kB^2\zeta^k_{h-1}+2k|\mathcal{A}|\cdot\mathbb{E}_{(s,a)\sim\breve{\rho}^k_h(\cdot,\cdot)}[g(s,a)^2]+\lambda_kB^2d/(C^-_{\mathcal{S}})^2}\cdot\mathbb{E}_{(s',a')\sim d^{\pi,\widehat{\mathbb{P}}^k}_{h-1}(\cdot,\cdot)}\left\|\widehat{\phi}^k_{h-1}(s',a')\right\|_{\Sigma^{-1}_{\widetilde{\rho}^k_{h-1},\widehat{\phi}^k_{h-1}}}.
\end{aligned}
$$

For $h=1$, we have

$$
\left|\mathbb{E}_{(s,a)\sim d^{\pi,\widehat{\mathbb{P}}^k}_1(\cdot,\cdot)}[g(s,a)]\right|=\sqrt{g(s_1,\pi_1(s_1))^2}\le\sqrt{|\mathcal{A}|\mathbb{E}_{a\sim\widetilde{\rho}^k_1(s_1,\cdot)}[g(s_1,a)^2]},
$$

where we let $\widetilde{\rho}^k_1(s_1,a)=\mathrm{Unif}(a)$. Note that the above derivations also hold for any randomized policy $\pi$. The proof is completed. □

**Lemma C.5.** *Define $\widetilde{\rho}^k_h(\cdot,\cdot):=\frac{1}{k}\sum_{k'=0}^{k-1}\widetilde{d}^{\pi^{k'}}_h(\cdot,\cdot)$ for all $h\ge 1$ with $\widetilde{\rho}^k_1(s_1,a)=\mathrm{Unif}(a)$ and $\rho^k_h(\cdot,\cdot):=\frac{1}{k}\sum_{k'=0}^{k-1}d^{\pi^{k'}}_h(\cdot,\cdot)$ for all $h\ge 2$. Then for any function $g:\mathcal{S}\times\mathcal{A}\mapsto[0,B]$ and policy $\pi$, we have for any $h\ge 2$, the following inequality holds*

$$
\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim d^{\pi,\mathbb{P}}_h(\cdot,\cdot)}[g(s,a)]\right|\\
&\quad\le\sqrt{k|\mathcal{A}|\cdot\mathbb{E}_{(s,a)\sim\widetilde{\rho}^k_h(\cdot,\cdot)}[g(s,a)^2]+\lambda_kB^2d}\cdot\mathbb{E}_{(s',a')\sim d^{\pi,\mathbb{P}}_{h-1}(\cdot,\cdot)}\left\|\phi^*_{h-1}(s',a')\right\|_{\Sigma^{-1}_{\rho^k_{h-1},\phi^*_{h-1}}}.
\end{aligned}
$$

*Moreover, for $h=1$, we have*

$$
\left|\mathbb{E}_{(s,a)\sim d^{\pi,\mathbb{P}}_1(\cdot,\cdot)}[g(s,a)]\right|\le\sqrt{g(s_1,\pi_1(s_1))^2}\le\sqrt{|\mathcal{A}|\mathbb{E}_{a\sim\widetilde{\rho}^k_1(s_1,\cdot)}[g(1s,a)^2]}.
$$

*Proof.* For any function $g:\mathcal{S}\times\mathcal{A}\mapsto[0,B]$ and any deterministic policy $\pi$, under the true transition model $\mathbb{P}$, for any

$h \geq 2$, we have

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim d_h^{\pi,\mathbb{P}}(\cdot,\cdot)}[g(s,a)]\right| \\
&\quad= \left|\mathbb{E}_{(s',a')\sim d_{h-1}^{\pi,\mathbb{P}}(\cdot,\cdot),s\sim\mathbb{P}_{h-1}(\cdot|s',a')}[g(s,\pi_h(s))]\right| \\
&\quad= \left|\mathbb{E}_{(s',a')\sim d_{h-1}^{\pi,\mathbb{P}}(\cdot,\cdot)}\left[\phi_{h-1}^*(s',a')^\top \int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right]\right| \\
&\quad\leq \mathbb{E}_{(s',a')\sim d_{h-1}^{\pi,\mathbb{P}}(\cdot,\cdot)} \left\|\phi_{h-1}^*(s',a')\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}} \left\|\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right\|_{\Sigma_{\rho_{h-1}^k,\phi_{h-1}^*}},
\end{aligned} \tag{11}$$

where the inequality is due to the Cauchy-Schwarz inequality. Here, we define the covariance matrix $\Sigma_{\rho_{h-1}^k,\phi_{h-1}^*} := k\mathbb{E}_{(s,a)\sim\rho_{h-1}^k}[\phi_{h-1}^*(s,a)\phi_{h-1}^*(s,a)^\top] + \lambda_k I$ with $\rho_{h-1}^k(s,a) = \frac{1}{k}\sum_{k'=0}^{k-1} d_{h-1}^{\pi^{k'}}(s,a)$.

Next, we have

$$\begin{aligned}
&\left\|\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right\|^2_{\Sigma_{\rho_{h-1}^k,\phi_{h-1}^*}} \\
&\quad= k\left(\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right)^\top \mathbb{E}_{\rho_{h-1}^k}\left[\phi_{h-1}^*(\phi_{h-1}^*)^\top\right]\left(\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right) \\
&\qquad+ \lambda_k\left(\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right) \\
&\quad= k\mathbb{E}_{(s'',a'')\sim\rho_{h-1}^k(\cdot,\cdot)}\left[\int_{\mathcal{S}} \phi_{h-1}^*(s'',a'')^\top \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right] \\
&\qquad+ \lambda_k\left(\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right) \\
&\quad\leq k\mathbb{E}_{(s'',a'')\sim\rho_{h-1}^k(\cdot,\cdot)}\left[\int_{\mathcal{S}} \phi_{h-1}^*(s'',a'')^\top \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right]^2 + \lambda_k B^2 d,
\end{aligned} \tag{12}$$

where, by Assumption 2.1, the last inequality is due to

$$\left(\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}} \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right) \leq B^2\left|\int_{\mathcal{S}} \psi_{h-1}^*(s)\mathrm{d}s\right|_2^2 \leq B^2 d.$$

Furthermore, we have

$$\begin{aligned}
&k\mathbb{E}_{(s'',a'')\sim\rho_{h-1}^k(\cdot,\cdot)}\left[\int_{\mathcal{S}} \phi_{h-1}^*(s'',a'')^\top \psi_{h-1}^*(s) g(s,\pi_h(s))\mathrm{d}s\right]^2 \\
&\quad= k\mathbb{E}_{(s'',a'')\sim\rho_{h-1}^k(\cdot,\cdot)}\left[\int_{\mathcal{S}} \mathbb{P}_{h-1}(s|s'',a'') g(s,\pi_h(s))\mathrm{d}s\right]^2 \\
&\quad\leq k\mathbb{E}_{(s'',a'')\sim\rho_{h-1}^k(\cdot,\cdot),s\sim\mathbb{P}_{h-1}(\cdot|s'',a'')}[g(s,\pi_h(s))^2] \\
&\quad\leq k\frac{1}{\mathrm{Unif}(a)}\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}[g(s,a)^2] \\
&\quad= k|\mathcal{A}|\cdot\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}[g(s,a)^2],
\end{aligned} \tag{13}$$

where the first inequality is due to Jensen's inequality and the second inequality is by $g(s,\pi_h(s))^2 \leq \sum_{a\in\mathcal{A}} g(s,a)^2 = 1/\mathrm{Unif}(a)\cdot\mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})}[g(s,a)^2]$ and $\widetilde{\rho}_h^k(s,a) := \rho_{h-1}^k(s',a')\mathbb{P}_{h-1}(s|s',a')\mathrm{Unif}(a)$ for all $h \geq 2$.

Combining (11), (12), and (13), we have for any $h \geq 2$,

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a)\sim d_h^{\pi,\mathbb{P}}(\cdot,\cdot)}[g(s,a)]\right| \\
&\quad\leq \sqrt{k|\mathcal{A}|\cdot\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}[g(s,a)^2] + \lambda_k B^2 d}\cdot\mathbb{E}_{(s',a')\sim d_{h-1}^{\pi,\mathbb{P}}(\cdot,\cdot)}\left\|\phi_{h-1}^*(s',a')\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}.
\end{aligned}$$

For $h = 1$, we have

$$\left|\mathbb{E}_{(s,a)\sim d_1^{\pi,\mathbb{P}}(\cdot,\cdot)}[g(s,a)]\right| \leq \sqrt{g(s_1,\pi_1(s_1))^2} \leq \sqrt{|\mathcal{A}|\mathbb{E}_{a\sim\widetilde{\rho}_1^k(s_1,\cdot)}[g(_1s,a)^2]},$$

where we define $\widetilde{\rho}_1^k(s_1,a) = \mathrm{Unif}(a)$. The above derivations also hold for any randomized policy $\pi$. The proof is completed. □

**Lemma C.6.** *Let $\pi^* := \mathrm{argmax}_\pi V_1^\pi(s_1)$, $\overline{V}_1^k(s_1)$ be the value function updated in Algorithm 1, and $\overline{V}_{k,1}^\pi$ be the value function under any policy $\pi$ associated with an MDP defined by the reward function $r+\beta^k$ and the estimated transition $\widehat{\mathbb{P}}^k$ with $\beta^k$ and $\widehat{\mathbb{P}}^k$ obtained at episode $k$ of Algorithm 1. Then we have*

$$\overline{V}_1^k(s_1) \geq \overline{V}_{k,1}^{\pi^*}(s_1).$$

*Proof.* We prove this lemma by induction. First, we have $\overline{V}_{H+1}^k(s) = \overline{V}_{k,H+1}^\pi(s) = 0$ for any $s\in\mathcal{S}$ and any (randomized) policy $\pi$ such that the Bellman equation is written as $Q_h^\pi(s,a) = r_h(s,a)+\mathbb{P}_h V_{h+1}^\pi(s,a)$ and $V_h^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_h^\pi(s,a)]$. Here, we aim to prove this lemma holds for any policy $\pi$, we slightly abuse the notation $\pi$ and let $\pi_h(a|s)$ be the probability of taking action $a$ under the state $s$. Next, we assume the following inequality holds

$$\overline{V}_{h+1}^k(s) \geq \overline{V}_{k,h+1}^\pi(s).$$

Then, with the above inequality, by the Bellman equation, we have

$$\begin{aligned}
&\overline{Q}_h^k(s,a) - \overline{Q}_{k,h}^\pi(s,a) \\
&\quad= r_h(s,a) + \beta_h^k(s,a) + \widehat{\mathbb{P}}_h^k\overline{V}_{h+1}^k(s,a) - r_h(s,a) - \beta_h^k(s,a) - \widehat{\mathbb{P}}_h^k\overline{V}_{k,h+1}^\pi(s,a) \\
&\quad= \widehat{\mathbb{P}}_h^k\overline{V}_{h+1}^k(s) - \widehat{\mathbb{P}}_h^k\overline{V}_{k,h+1}^\pi(s) \geq 0.
\end{aligned} \tag{14}$$

Then, we have

$$\begin{aligned}
\overline{V}_h^k(s) &= \max_{a\in\mathcal{A}}\overline{Q}_h^k(s,a) \\
&\geq \max_{a\in\mathcal{A}}\overline{Q}_{k,h}^\pi(s,a) \\
&\geq \mathbb{E}_{a\sim\pi_h(\cdot|s)}[\overline{Q}_{k,h}^\pi(s,a)] = \overline{V}_{k,h}^\pi(s),
\end{aligned}$$

where the first inequality is by (14) and the second inequality is due to the fact that $\max_i \mathbf{v}_i \geq \langle \mathbf{v}, \mathbf{d}\rangle$ when $\mathbf{v}$ is any vector and $\mathbf{d}$ is a vector in a probability simplex satisfying $\sum_i \mathbf{d}_i = 1$ and $\mathbf{d}_i \geq 0$. Thus, we obtain for any policy $\pi$,

$$\overline{V}_1^k(s_1) \geq \overline{V}_{k,1}^\pi(s_1),$$

which further implies

$$\overline{V}_1^k(s_1) \geq \overline{V}_{k,1}^{\pi^*}(s_1).$$

This completes the proof. □

### C.2. Proof of Lemma 5.1

*Proof.* For any function $f_h \in \mathcal{F}$, we let $\Pr_h^f(y|s,a,s')$ denote the conditional probability characterized by the function $f_h$ at the step $h$, which is

$$\Pr{}_h^f(y|s,a,s') = \left(\frac{f_h(s,a,s')}{1+f_h(s,a,s')}\right)^y \left(\frac{1}{1+f_h(s,a,s')}\right)^{1-y}.$$

Furthermore, we have

$$\Pr{}_h^f(y,s'|s,a) = \Pr{}_h^f(y|s,a,s')\Pr{}_h(s'|s,a) = \left(\frac{f_h(s,a,s')\Pr{}_h(s'|s,a)}{1+f_h(s,a,s')}\right)^y \left(\frac{\Pr{}_h(s'|s,a)}{1+f_h(s,a,s')}\right)^{1-y},$$

where we have

$$\begin{aligned}\Pr{}_h(s'|s,a) &= \Pr{}_h(y=1|s,a)\Pr{}_h(s'|y=1,s,a) + \Pr{}_h(y=0|s,a)\Pr{}_h(s'|y=0,s,a)\\ &= \Pr{}_h(y=1)\Pr{}_h(s'|y=1,s,a) + \Pr{}_h(y=0)\Pr{}_h(s'|y=0,s,a)\\ &= \frac{1}{2}[\mathbb{P}_h(s'|s,a) + \mathcal{P}_{\mathcal{S}}^-(s')] \geq \frac{1}{2}C_{\mathcal{S}}^- > 0,\end{aligned} \tag{15}$$

since we assume $P_{\mathcal{S}}^-(s') \geq C_{\mathcal{S}}^-$.

Thus, we have the equivalency of solving the following two problems with $f_h(s,a,s') = \phi_h(s,a)^\top \psi_h(s')$, which is

$$\max_{\phi_h\in\Phi,\psi_h\in\Psi} \sum_{(s,a,s',y)\in\mathcal{D}_h^k} \log \Pr{}_h^f(y|s,a,s') = \max_{\phi_h,\psi_h} \sum_{(s,a,s',y)\in\mathcal{D}_h^k} \log \Pr{}_h^f(y,s'|s,a), \tag{16}$$

since the conditional probability $\Pr_h(s'|s,a)$ is only determined by $\mathbb{P}_h(s'|s,a)$ and $\mathcal{P}_{\mathcal{S}}^-(s')$ and is independent of $f_h$ as shown in (15). We denote the solution of (16) as $\widetilde{\phi}_h^k$ and $\widetilde{\psi}_h^k$ such that

$$\widehat{f}_h^k(s,a,s') = \widetilde{\psi}_h^k(s')^\top \widetilde{\phi}_h^k(s,a).$$

According to Algorithm 3, we know that for each $h \geq 2$, at each episode $k' \in [k]$, the data $(s,a)$ is sampled from both $\widetilde{d}_h^{\pi^{k'}}(\cdot,\cdot)$ and $\breve{d}_h^{\pi^{k'}}(\cdot,\cdot)$. Therefore, further with Lemma E.2, by solving the contrastive loss in (2) or equivalently as in (16), with probability at least $1-\delta$, for all $h \geq 2$, we have

$$\begin{aligned}\sum_{k'=1}^{k}\Bigg[&\mathbb{E}_{(s,a)\sim\widetilde{d}_h^{\pi^{k'}}(\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2\\ &+ \mathbb{E}_{(s,a)\sim\breve{d}_h^{\pi^{k'}}(\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2\Bigg] \leq 2\log(2kH|\mathcal{F}|/\delta),\end{aligned}$$

where the factor $2H$ inside $\log$ is due to the data being sampled from two distributions and applying union bound for all $h \geq 2$. The above inequality is equivalent to

$$\begin{aligned}&\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2\\ &\quad+ \mathbb{E}_{(s,a)\sim\breve{\rho}_h^k(\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2 \leq 2\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \geq 2,\end{aligned} \tag{17}$$

where we use the fact that $\widetilde{\rho}_h^k(s,a) = \frac{1}{k}\sum_{k'=0}^{k-1}\widetilde{d}_h^{\pi^{k'}}(s,a)$ and $\breve{\rho}_h^k(s,a) = \frac{1}{k}\sum_{k'=0}^{k-1}\breve{d}_h^{\pi^{k'}}(s,a)$. On the other hand, for $h=1$, the data is only sampled from $\widetilde{d}_1^{\pi^{k'}}(\cdot,\cdot)$ for any $k' \in [k]$. Therefore, we have

$$\sum_{k'=1}^{k}\left[\mathbb{E}_{(s,a)\sim\widetilde{d}_1^{\pi^{k'}}(\cdot,\cdot)}\left\|\Pr{}_1^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr{}_1^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2\right] \leq 2\log(2k|\mathcal{F}|/\delta),$$

which, analogously, gives

$$\mathbb{E}_{(s,a)\sim\widetilde{\rho}_1^k(\cdot,\cdot)}\left\|\Pr{}_1^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr{}_1^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2 \leq 2\log(2k|\mathcal{F}|/\delta)/k. \tag{18}$$

Thus, by (17) and (18), with probability at least $1-2\delta$, we have

$$\begin{aligned}&\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2 \leq 2\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \geq 1,\\ &\mathbb{E}_{(s,a)\sim\breve{\rho}_h^k(\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2 \leq 2\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \geq 2,\end{aligned} \tag{19}$$

Next, we show the recovery error bound of the transition model based on (19). We have

$$\begin{aligned}
&\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a)-\Pr{}_h^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2\\
&\quad=\left(\left\|\Pr{}_h^{\widehat{f}^k}(y=0,\cdot|s,a)-\Pr{}_h^{f^*}(y=0,\cdot|s,a)\right\|_{\mathrm{TV}}+\left\|\Pr{}_h^{\widehat{f}^k}(y=1,\cdot|s,a)-\Pr{}_h^{f^*}(y=1,\cdot|s,a)\right\|_{\mathrm{TV}}\right)^2\\
&\quad=4\left\|\frac{\Pr{}_h(\cdot|s,a)}{1+\widehat{f}_h^k(s,a,\cdot)}-\frac{\Pr{}_h(\cdot|s,a)}{1+f_h^*(s,a,\cdot)}\right\|_{\mathrm{TV}}^2\\
&\quad=2\left[\int_{s'\in\mathcal{S}}\frac{\Pr{}_h(s'|s,a)\cdot|f_h^*(s,a,s')-\widehat{f}_h^k(s,a,s')|}{[1+\widehat{f}_h^k(s,a,s')]\cdot[1+f_h^*(s,a,s')]}\mathrm{d}s'\right]^2,
\end{aligned}$$

where $f^*(s,a,s')=\frac{\mathbb{P}(s'|s,a)}{\mathcal{P}_{\mathcal{S}}^-(s')}$ with $\mathcal{P}_{\mathcal{S}}^-(s')\geq C_{\mathcal{S}}^-,\ \forall s'\in\mathcal{S}$ and the second equation is due to $\|\Pr{}_h^{\widehat{f}^k}(y=0,\cdot|s,a)-\Pr{}_h^{f^*}(y=0,\cdot|s,a)\|_{\mathrm{TV}}=\|\Pr{}_h^{\widehat{f}^k}(y=1,\cdot|s,a)-\Pr{}_h^{f^*}(y=1,\cdot|s,a)\|_{\mathrm{TV}}=\left\|\frac{\Pr{}_h(\cdot|s,a)}{1+\widehat{f}_h^k(s,a,\cdot)}-\frac{\Pr{}_h(\cdot|s,a)}{1+f_h^*(s,a,\cdot)}\right\|_{\mathrm{TV}}$. Moreover, according to Lemma C.1 and (15), we have

$$\begin{aligned}
&\frac{\Pr{}_h(s'|s,a)\cdot|f_h^*(s,a,s')-\widehat{f}_h^k(s,a,s')|}{[1+\widehat{f}_h^k(s,a,s')]\cdot[1+f_h^*(s,a,s')]}\\
&\quad=\frac{1/2\cdot[\mathbb{P}_h(s'|s,a)+\mathcal{P}_{\mathcal{S}}^-(s')]\cdot|\mathbb{P}_h(s'|s,a)/\mathcal{P}_{\mathcal{S}}^-(s')-\widehat{f}_h^k(s,a,s')|}{[1+\widehat{f}_h^k(s,a,s')]\cdot[1+\mathbb{P}_h(s'|s,a)/\mathcal{P}_{\mathcal{S}}^-(s')]}\\
&\quad=\frac{1/2\cdot|\mathbb{P}_h(s'|s,a)-\mathcal{P}_{\mathcal{S}}^-(s')\widehat{f}_h^k(s,a,s')|}{1+\widehat{f}_h^k(s,a,s')}\geq\frac{|\mathbb{P}_h(s'|s,a)-\mathcal{P}_{\mathcal{S}}^-(s')\widehat{f}_h^k(s,a,s')|}{4\sqrt{d}/C_{\mathcal{S}}^-},
\end{aligned}$$

where the inequality is due to $[1+\widehat{f}_h^k(s,a,s')]\leq(1+\sqrt{d}/C_{\mathcal{S}}^-)\leq 2\sqrt{d}/C_{\mathcal{S}}^-$ since $\widehat{f}_h^k(s,a,s')\leq\sqrt{d}/C_{\mathcal{S}}^-$ with $d\geq 1$ and $0<C_{\mathcal{S}}^-\leq 1$. Thus, the above results further give

$$\frac{(C_{\mathcal{S}}^-)^2}{8d}\left[\int_{s'\in\mathcal{S}}\left|\mathbb{P}_h(s'|s,a)-\mathcal{P}_{\mathcal{S}}^-(s')\widehat{f}_h^k(s,a,s')\right|\mathrm{d}s'\right]^2\leq\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a)-\Pr{}_h^{f^*}(\cdot,\cdot|s,a)\right\|_{\mathrm{TV}}^2.$$

Therefore, combining this inequality with (19), we obtain

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\left\|\mathbb{P}_h(\cdot|s,a)-\mathcal{P}_{\mathcal{S}}^-(\cdot)\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(\cdot)\right\|_{\mathrm{TV}}^2 \qquad (20)\\
&=1/2\cdot\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\left[\int_{s'\in\mathcal{S}}\left|\mathbb{P}_h(s'|s,a)-\mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(s')\right|\mathrm{d}s'\right]^2\leq 8d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k,\quad\forall h\geq 2.
\end{aligned}$$

Similarly, we can obtain

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\widetilde{\rho}_1^k(\cdot,\cdot)}\left\|\mathbb{P}_h(\cdot|s,a)-\mathcal{P}_{\mathcal{S}}^-(\cdot)\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(\cdot)\right\|_{\mathrm{TV}}^2\leq 8d/(C_{\mathcal{S}}^-)^2\cdot\log(2k|\mathcal{F}|/\delta)/k,\\
&\mathbb{E}_{(s,a)\sim\check{\rho}_h^k(\cdot,\cdot)}\left\|\mathbb{P}_h(\cdot|s,a)-\mathcal{P}_{\mathcal{S}}^-(\cdot)\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(\cdot)\right\|_{\mathrm{TV}}^2\leq 8d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k,\quad\forall h\geq 2.
\end{aligned}\qquad(21)$$

Now we define

$$\widehat{g}_h^k(s,a,s'):=\mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(s').$$

Note that $\int_{s'\in\mathcal{S}}\widehat{g}_h^k(s,a,s')\mathrm{d}s'$ may not be guaranteed to be 1 though $\widehat{g}_h^k(s,a,\cdot)$ is close to the true transition model $\mathbb{P}_h(\cdot|s,a)$ according to (20) and (21). Therefore, to obtain an approximator of the transition model $\mathbb{P}_h$ lying on a probability simplex, we should further normalize $\widehat{g}_h^k(s,a,s')$. Thus, we define for all $(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}$,

$$\widehat{\mathbb{P}}_h^k(s'|s,a):=\frac{\widehat{g}_h^k(s,a,s')}{\|\widehat{g}_h^k(s,a,\cdot)\|_1}=\frac{\widehat{g}_h^k(s,a,s')}{\int_{s'\in\mathcal{S}}\widehat{g}_h^k(s,a,s')\mathrm{d}s'}=\frac{\mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(s')}{\int_{s'\in\mathcal{S}}\mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a)^\top\widetilde{\psi}_h^k(s')\mathrm{d}s'}.$$

We further let

$$\widehat{\phi}_h^k(s,a) := \widetilde{\phi}_h^k(s,a) / \int_{s'\in\mathcal{S}} \mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a)^\top \widetilde{\psi}_h^k(s')\mathrm{d}s', \qquad \widehat{\psi}_h^k(s') := \mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\psi}_h^k(s'),$$

such that

$$\widehat{\mathbb{P}}_h^k(s'|s,a) = \widehat{\psi}_h^k(s')^\top \widehat{\phi}_h^k(s,a).$$

Next, based on the above definitions and results, we will give the upper bound of the approximation error $\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a) - \mathbb{P}_h(\cdot|s,a)\|_{\mathrm{TV}}^2$. We have

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a) - \mathbb{P}_h(\cdot|s,a)\|_{\mathrm{TV}}^2 \\
&\quad\le 2\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a) - \widehat{g}_h^k(s,a,\cdot)\|_{\mathrm{TV}}^2 + 2\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{g}_h^k(s,a,\cdot) - \mathbb{P}_h(\cdot|s,a)\|_{\mathrm{TV}}^2 \\
&\quad\le 2\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a) - \widehat{g}_h^k(s,a,\cdot)\|_{\mathrm{TV}}^2 + 16d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k,
\end{aligned} \tag{22}$$

where the first inequality is by $(x+y)^2 \le 2x^2+2y^2$ and the last inequality is by (20). Moreover, we have

$$\begin{aligned}
&\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a) - \widehat{g}_h^k(s,a,\cdot)\|_{\mathrm{TV}}^2 \\
&\quad= \mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\left\|\frac{\widehat{g}_h^k(s,a,s')}{\|\widehat{g}_h^k(s,a,\cdot)\|_1} - \widehat{g}_h^k(s,a,\cdot)\right\|_{\mathrm{TV}}^2 \\
&\quad= \frac{1}{4}\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\left(\|\widehat{g}_h^k(s,a,\cdot)\|_1 - 1\right)^2 \\
&\quad\le \frac{1}{4}\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\left(\|\widehat{g}_h^k(s,a,\cdot) - \mathbb{P}_h(\cdot|s,a)\|_1 + \|\mathbb{P}_h(\cdot|s,a)\|_1 - 1\right)^2 \\
&\quad\le \frac{1}{4}\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{g}_h^k(s,a,\cdot) - \mathbb{P}_h(\cdot|s,a)\|_1^2 \\
&\quad= \mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{g}_h^k(s,a,\cdot) - \mathbb{P}_h(\cdot|s,a)\|_{\mathrm{TV}}^2 \le 8d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k.
\end{aligned}$$

Combining the above inequality with (22), we eventually obtain

$$\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a) - \mathbb{P}_h(\cdot|s,a)\|_{\mathrm{TV}}^2 \le 32d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h\ge 1.$$

Thus, we similarly have

$$\mathbb{E}_{(s,a)\sim\check{\rho}_h^k(\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a) - \mathbb{P}_h(\cdot|s,a)\|_{\mathrm{TV}}^2 \le 32d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h\ge 2.$$

The above three inequalities hold with probability at least $1-2\delta$. This completes the proof. □

### C.3. Proof of Theorem 3.6

*Proof.* We first decompose the term $V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1)$ as follows

$$\begin{aligned}
&V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1) \\
&\quad= V_1^{\pi^*}(s_1) - \overline{V}_{k,1}^{\pi^*}(s_1) + \overline{V}_{k,1}^{\pi^*}(s_1) - V_1^k(s_1) + V_1^k(s_1) - V_1^{\pi^k}(s_1) \\
&\quad\le V_1^{\pi^*}(s_1) - \overline{V}_{k,1}^{\pi^*}(s_1) + \overline{V}_1^k(s_1) - V_1^{\pi^k}(s_1) \\
&\quad= \underbrace{V_1^{\pi^*}(s_1) - \overline{V}_{k,1}^{\pi^*}(s_1)}_{(i)} + \underbrace{\overline{V}_{k,1}^{\pi^k}(s_1) - V_1^{\pi^k}(s_1)}_{(ii)},
\end{aligned} \tag{23}$$

where the first inequality is by the result of Lemma C.6 that $\overline{V}_{k,1}^{\pi^*}(s_1) \le V_1^k(s_1)$ and the second equation is by the definition of $\overline{V}_h^k$ as in Algorithm 1 such that $\overline{V}_h^k = \overline{V}_{k,h}^{\pi^k}$ for any $h\in[H]$. Thus, to bound the term $V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1)$, we only need to bound the terms $(i)$ and $(ii)$ as in (23).

To bound term $(i)$, by Lemma C.2, we have

$$
\begin{aligned}
(i) = V_1^{\pi^*}(s_1) - \overline{V}_{k,1}^{\pi^*}(s_1) &= \mathbb{E}\left[\sum_{h=1}^{H}\left(-\beta_h^k(s_h,a_h) + (\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)V_{h+1}^{\pi^*}(s_h,a_h)\right)\,\middle|\,\pi^*,\widehat{\mathbb{P}}^k\right] \\
&\le \mathbb{E}\left[\sum_{h=1}^{H}\left(-\beta_h^k(s_h,a_h) + H\|\mathbb{P}_h(\cdot|s_h,a_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h)\|_1\right)\,\middle|\,\pi^*,\widehat{\mathbb{P}}^k\right],
\end{aligned} \tag{24}
$$

where the first inequality is by the fact $\sup_{s\in\mathcal{S}} V_{h+1}^{\pi}(s) \le H$. Next, we bound the term $\mathbb{E}[\sum_{h=1}^{H} H\|\mathbb{P}_h(\cdot|s_h,a_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h)\|_1 \,|\, \pi^*,\widehat{\mathbb{P}}^k]$. Note that for the term $\|\mathbb{P}_h(\cdot|s_h,a_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h)\|_1$, we first have a trivial bound that $\|\mathbb{P}_h(\cdot|s_h,a_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h)\|_1 \le \|\mathbb{P}_h(\cdot|s_h,a_h)\|_1 + \|\widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h)\|_1 = 2$. Moreover, according to Lemma C.4, we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}\|\mathbb{P}_h(\cdot|s_h,a_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h)\|_1\,\middle|\,\pi^*,\widehat{\mathbb{P}}^k\right] \\
&= \sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim d_h^{\pi^*,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}[\|\mathbb{P}_h(\cdot|s_h,a_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h)\|_1] \\
&= \sum_{h=2}^{H}\sqrt{8k\zeta_{h-1}^k + 2k|\mathcal{A}|\mathbb{E}_{(s,a)\sim\breve{\rho}_h^k(\cdot,\cdot)}[\|\mathbb{P}_h(\cdot|s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a)\|_1^2] + 4\lambda_k d/(C_{\mathcal{S}}^-)^2}\cdot\mathbb{E}_{(s,a)\sim d_{h-1}^{\pi^*,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}\left\|\widehat{\phi}_{h-1}^k(s,a)\right\|_{\Sigma^{-1}_{\breve{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}} \\
&\quad + \sqrt{|\mathcal{A}|\mathbb{E}_{a\sim\widetilde{\rho}_1^k(s_1,\cdot)}[\|\mathbb{P}_1(\cdot|s_1,a) - \widehat{\mathbb{P}}_1^k(\cdot|s_1,a)\|_1^2]} \\
&= \sum_{h=2}^{H}\sqrt{8k\zeta_{h-1}^k + 2k|\mathcal{A}|\xi_h^k + 4\lambda_k d/(C_{\mathcal{S}}^-)^2}\cdot\mathbb{E}_{(s,a)\sim d_{h-1}^{\pi^*,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}\left\|\widehat{\phi}_{h-1}^k(s,a)\right\|_{\Sigma^{-1}_{\breve{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}} + \sqrt{|\mathcal{A}|\zeta_1^k},
\end{aligned}
$$

where the last equation is by the below definitions for all $(h,k)\in[H]\times[K]$,

$$
\begin{aligned}
\zeta_h^k &:= \mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}[\|\mathbb{P}_1(\cdot|s,a) - \widehat{\mathbb{P}}_1^k(\cdot|s,a)\|_1^2], \\
\xi_h^k &:= \mathbb{E}_{(s,a)\sim\breve{\rho}_h^k(\cdot,\cdot)}[\|\mathbb{P}_h(\cdot|s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a)\|_1^2],
\end{aligned} \tag{25}
$$

whose upper bound will be characterized later. Therefore, the above results imply that

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^{H}H\|\mathbb{P}_h(\cdot|s_h,a_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h)\|_1\,\middle|\,\pi^*,\widehat{\mathbb{P}}^k\right] \\
&\quad\le \min\left\{H\sqrt{|\mathcal{A}|\zeta_1^k} + \sum_{h=2}^{H}H\sqrt{8k\zeta_{h-1}^k + 2k|\mathcal{A}|\xi_h^k + 4\lambda_k d/(C_{\mathcal{S}}^-)^2}\cdot\mathbb{E}_{(s,a)\sim d_{h-1}^{\pi^*,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}\left\|\widehat{\phi}_{h-1}^k(s,a)\right\|_{\Sigma^{-1}_{\breve{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}},\ 2H^2\right\}.
\end{aligned}
$$

On the other hand, we further bound the term $\mathbb{E}[\sum_{h=1}^H -\beta_h^k(s_h,a_h)\,|\,\pi^*,\widehat{\mathbb{P}}^k]$ in (24). We have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^H -\beta_h^k(s_h,a_h)\,\middle|\,\pi^*,\widehat{\mathbb{P}}^k\right]\\
&\quad=\mathbb{E}\left[\sum_{h=1}^H -\min\{\gamma_k\|\widehat{\phi}_h^k(s_h,a_h)\|_{(\widehat{\Sigma}_h^k)^{-1}},2H\}\,\middle|\,\pi^*,\widehat{\mathbb{P}}^k\right]\\
&\quad\le\mathbb{E}\left[\sum_{h=1}^H -\min\left\{\frac{3}{5}\gamma_k\|\widehat{\phi}_h^k(s_h,a_h)\|_{\Sigma^{-1}_{\widehat{\rho}_h^k,\widehat{\phi}_h^k}},2H\right\}\,\middle|\,\pi^*,\widehat{\mathbb{P}}^k\right]\\
&\quad=-\min\left\{\frac{3}{5}\gamma_k\sum_{h=1}^H\mathbb{E}_{(s,a)\sim d_h^{\pi^*,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}\|\widehat{\phi}_h^k(s,a)\|_{\Sigma^{-1}_{\widehat{\rho}_h^k,\widehat{\phi}_h^k}},2H^2\right\}\\
&\quad\le-\min\left\{\frac{3}{5}\gamma_k\sum_{h=1}^{H-1}\mathbb{E}_{(s,a)\sim d_h^{\pi^*,\widehat{\mathbb{P}}^k}(\cdot,\cdot)}\|\widehat{\phi}_h^k(s,a)\|_{\Sigma^{-1}_{\widehat{\rho}_h^k,\widehat{\phi}_h^k}},2H^2\right\},
\end{aligned}$$

when $\lambda_k\ge c_0 d\log(H|\Phi|k/\delta)$ with probability at least $1-\delta$. The first inequality is obtained by applying Lemma E.1 for all $h\in[H]$. Thus, plugging in the above results into (24), for a sufficient large $c_0$, setting

$$\lambda_k=c_0 d\log(H|\Phi|k/\delta),\qquad \gamma_k=\frac{5}{3}H\sqrt{8k\zeta_{h-1}^k+2k|\mathcal{A}|\xi_h^k+4\lambda_k d/(C_{\mathcal{S}}^-)^2},\tag{26}$$

we have that

$$(i)=V_1^{\pi^*}(s_1)-V_1^{\pi^*}(s_1)\le\sqrt{|\mathcal{A}|\zeta_1^k},\tag{27}$$

where the inequality is due to $\min\{x+y,2H^2\}-\min\{y,2H^2\}\le x,\forall x,y\ge 0$. The above inequality (27) looks similar to the optimism in linear MDPs (Jin et al., 2020) but has an additional positive bias $\sqrt{|\mathcal{A}|\zeta_1^k}$ which depends on $\sqrt{1/k}$. Uehara et al. (2021) refers to such a biased optimism as *near-optimism*. In our work, we prove that our algorithm for the low-rank MDP in an episodic setting also leads to near-optimism.

Next, we show the upper bound of the term $(ii)$ in (23). By Lemma C.3, we have

$$\begin{aligned}
(ii)=\overline{V}_{k,1}^{\pi^k}(s_1)-V_1^{\pi^k}(s_1)&=\mathbb{E}\left[\sum_{h=1}^H\left(\beta_h^k(s_h,a_h)-(\mathbb{P}_h-\widehat{\mathbb{P}}_h^k)\overline{V}_{h+1}^{\pi^k}(s_h,a_h)\right)\,\middle|\,\pi^k,\mathbb{P}\right]\\
&\le\mathbb{E}\left[\sum_{h=1}^H\left(\beta_h^k(s_h,a_h)+3H^2\|\mathbb{P}_h(\cdot|s_h,a_h)-\widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h)\|_1\right)\,\middle|\,\pi^k,\mathbb{P}\right]\\
&=\sum_{h=1}^H\mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}(\cdot,\cdot)}\left(\beta_h^k(s,a)+3H^2\|\mathbb{P}_h(\cdot|s,a)-\widehat{\mathbb{P}}_h^k(\cdot|s,a)\|_1\right).
\end{aligned}\tag{28}$$

where the first inequality is due to $\sup_{s\in\mathcal{S},a\in\mathcal{A}}(r_h+\beta_h^k)(s,a)\le 1+2H\le 3H$ such that $\sup_{s\in\mathcal{S}}\overline{V}_h^{\pi}(s)\le 3H^2,\forall h\in[H]$ and the last equation is by the definition of $d_h^{\pi^k,\mathbb{P}}$. Then, we need to separately bound the two terms in the last equation above. By Lemma C.5, since $\sup_{s\in\mathcal{S},a\in\mathcal{A}}\beta_h^k(s,a)\le 2H$ according to the definition of $\beta_h^k$ in Algorithm 1, we have

$$\begin{aligned}
&\sum_{h=1}^H\mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}(\cdot,\cdot)}[\beta_h^k(s,a)]\\
&\le\sqrt{|\mathcal{A}|\mathbb{E}_{a\sim\widetilde{\rho}_1^k(s_1,\cdot)}[\beta_1^k(s_1,a)^2]}+\sum_{h=2}^H\sqrt{k|\mathcal{A}|\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}[\beta_h^k(s,a)^2]+4H^2\lambda_k d}\,\mathbb{E}_{(s,a)\sim d_{h-1}^{\pi^k,\mathbb{P}}(\cdot,\cdot)}\left\|\phi_{h-1}^*(s,a)\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}\\
&\le\sqrt{|\mathcal{A}|\gamma_k^2\mathbb{E}_{a\sim\widetilde{\rho}_1^k(s_1,\cdot)}\|\widehat{\phi}_1^k(s_1,a)\|^2_{(\widehat{\Sigma}_1^k)^{-1}}}\\
&\quad+\sum_{h=2}^H\sqrt{k|\mathcal{A}|\gamma_k^2\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}\|\widehat{\phi}_h^k(s,a)\|^2_{(\widehat{\Sigma}_h^k)^{-1}}+4H^2\lambda_k d}\,\mathbb{E}_{(s,a)\sim d_{h-1}^{\pi^k,\mathbb{P}}(\cdot,\cdot)}\left\|\phi_{h-1}^*(s,a)\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}},
\end{aligned}$$

where the second inequality is due to $\beta_h^k(s,a) \le \|\widehat{\phi}_h^k(s,a)\|_{(\widehat{\Sigma}_h^k)^{-1}}$. Furthermore, we have that with $\lambda_k \ge c_0 d \log(H|\Phi|k/\delta)$, with probability at least $1-\delta$, for all $h \in [H]$,

$$\begin{aligned}
\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)} \|\widehat{\phi}_h^k(s,a)\|^2_{(\widehat{\Sigma}_h^k)^{-1}} &\le 3\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)} \|\widehat{\phi}_h^k(s,a)\|^2_{\Sigma^{-1}_{\widetilde{\rho}_h^k,\widehat{\phi}_h^k}} \\
&= 3\mathbb{E}_{\widetilde{\rho}_h^k}\left[\widehat{\phi}_h^{k\top}\left(k\mathbb{E}_{\widetilde{\rho}_h^k}[\widehat{\phi}_h^k(\widehat{\phi}_h^k)^\top] + \lambda_k I\right)^{-1}\widehat{\phi}_h^k\right] \\
&= \frac{3}{k}\operatorname{tr}\left\{k\mathbb{E}_{\widetilde{\rho}_h^k}[\widehat{\phi}_h^k\widehat{\phi}_h^{k\top}]\left(k\mathbb{E}_{\widetilde{\rho}_h^k}[\widehat{\phi}_h^k(\widehat{\phi}_h^k)^\top] + \lambda_k I\right)^{-1}\right\} \\
&\le \frac{3}{k}\operatorname{tr}(I) = \frac{3d}{k},
\end{aligned}$$

where the first inequality is by Lemma E.1 and $\mathbb{E}_{\widetilde{\rho}_h^k}$ is short for $\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}$. Thus, combining the above results, we have the following inequality holds with probability at least $1-\delta$,

$$\begin{aligned}
&\sum_{h=1}^H \mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}(\cdot,\cdot)}[\beta_h^k(s,a)] \\
&\quad\le \sqrt{3d|\mathcal{A}|\gamma_k^2/k} + \sum_{h=2}^H \sqrt{3d|\mathcal{A}|\gamma_k^2 + 4H^2\lambda_k d}\,\mathbb{E}_{(s,a)\sim d_{h-1}^{\pi^k,\mathbb{P}}(\cdot,\cdot)} \left\|\phi_{h-1}^*(s,a)\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}.
\end{aligned} \tag{29}$$

In addition, further by Lemma C.5, due to $\|\mathbb{P}_h(\cdot|s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a)\|_1 \le \|\mathbb{P}_h(\cdot|s,a)\|_1 + \|\widehat{\mathbb{P}}_h^k(\cdot|s,a)\|_1 \le 2$, we have

$$\begin{aligned}
&\sum_{h=1}^H \mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}(\cdot,\cdot)}[\|\mathbb{P}_h(\cdot|s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a)\|_1] \\
&\quad\le \sqrt{|\mathcal{A}|\mathbb{E}_{a\sim\widetilde{\rho}_1^k(s_1,\cdot)}[\|\mathbb{P}_h(\cdot|s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a)\|_1^2]} \\
&\qquad + \sum_{h=2}^H \sqrt{k|\mathcal{A}|\mathbb{E}_{(s,a)\sim\widetilde{\rho}_h^k(\cdot,\cdot)}[\|\mathbb{P}_h(\cdot|s,a) - \widehat{\mathbb{P}}_h^k(\cdot|s,a)\|_1^2] + 4\lambda_k d}\,\mathbb{E}_{(s,a)\sim d_{h-1}^{\pi^k,\mathbb{P}}(\cdot,\cdot)} \left\|\phi_{h-1}^*(s,a)\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}} \\
&\quad= \sqrt{|\mathcal{A}|\zeta_1^k} + \sum_{h=2}^H \sqrt{k|\mathcal{A}|\zeta_h^k + 4\lambda_k d}\,\mathbb{E}_{(s,a)\sim d_{h-1}^{\pi^k,\mathbb{P}}(\cdot,\cdot)} \left\|\phi_{h-1}^*(s,a)\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}.
\end{aligned} \tag{30}$$

Therefore, combining (28), (29), and (30), we obtain

$$\begin{aligned}
(ii) \le &\left[\sqrt{3d|\mathcal{A}|\gamma_k^2/k} + 3H^2\sqrt{|\mathcal{A}|\zeta_1^k}\right] \\
&+ \sum_{h=2}^H \left[\sqrt{3d|\mathcal{A}|\gamma_k^2 + 4H^2\lambda_k d} + 3H^2\sqrt{k|\mathcal{A}|\zeta_h^k + 4\lambda_k d}\right] \mathbb{E}_{(s,a)\sim d_{h-1}^{\pi^k,\mathbb{P}}(\cdot,\cdot)} \left\|\phi_{h-1}^*(s,a)\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}.
\end{aligned} \tag{31}$$

Now we characterize the upper bound of $\zeta_h^k$ and $\xi_h^k$ as defined in (25). According to Lemma 5.1, we have with probability at least $1-2\delta$,

$$\begin{aligned}
\zeta_h^k &\le 32d\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \ge 1, \\
\xi_h^k &\le 32d\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \ge 2,
\end{aligned} \tag{32}$$

Plugging (32) and (26) into (27) and (31), we obtain

$$(i) = V_1^{\pi^*}(s_1) - V_1^{\pi^*}(s_1) \lesssim \sqrt{d|\mathcal{A}|\log(KH|\mathcal{F}|/\delta)/k},$$

$$(ii) = \overline{V}_{k,1}^{\pi^k}(s_1) - V_1^{\pi^k}(s_1) \lesssim \sqrt{C_1\log(H|\mathcal{F}|K/\delta)/k} + \sqrt{(C_1+C_2)\log(H|\mathcal{F}|K/\delta)}\sum_{h=1}^{H-1}\mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}(\cdot,\cdot)}\|\phi_h^*(s,a)\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}}.$$

where we let $C_1 = H^2d^3|\mathcal{A}|/(C_{\mathcal{S}}^-)^2 + H^2d^2|\mathcal{A}|^2/(C_{\mathcal{S}}^-)^2 + H^4d|\mathcal{A}|/(C_{\mathcal{S}}^-)^2$ and $C_2 = H^4d^2$. Further by (23), we have

$$\frac{1}{K}\sum_{k=1}^{K}\Big[V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1)\Big] \lesssim \sqrt{(C_1+C_2)\log(H|\mathcal{F}|K/\delta)}/K \cdot \sum_{h=1}^{H-1}\sum_{k=1}^{K}\mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}(\cdot,\cdot)}\|\phi_h^*(s,a)\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}} + \sqrt{C_1\log(H|\mathcal{F}|K/\delta)/K}.$$

Moreover, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}(\cdot,\cdot)}\|\phi_h^*(s,a)\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}} \\
&\le \sqrt{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}(\cdot,\cdot)}\|\phi_h^*(s,a)\|^2_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}}} \\
&= \sqrt{\frac{1}{K}\sum_{k=1}^{K}\mathrm{tr}\Big(\mathbb{E}_{(s,a)\sim d_h^{\pi^k,\mathbb{P}}(\cdot,\cdot)}(\phi_h^*(s,a)\phi_h^*(s,a)^\top)\,\Sigma^{-1}_{\rho_h^k,\phi_h^*}\Big)} \\
&\le \sqrt{d\log(1+kd/\lambda_k)/K} \le \sqrt{d\log(1+c_1K)/K}.
\end{aligned}$$

where the first inequality is by Jensen's inequality and the second inequality is by Lemma E.3 with $c_1$ being some absolute constant. Thus, we have

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\Big[V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1)\Big] &\lesssim \sqrt{(C_1+C_2)\log(H|\mathcal{F}|K/\delta)H^2d\log(1+c_1K)/K} + \sqrt{C_1\log(H|\mathcal{F}|K/\delta)/K} \\
&\lesssim \sqrt{H^2d(C_1+C_2)\log(H|\mathcal{F}|K/\delta)\log(1+c_1K)/K}.
\end{aligned}$$

Taking union bound for all events in this proof, due to $|\mathcal{F}| \ge |\Phi|$, setting

$$\lambda_k = c_0 d\log(H|\mathcal{F}|k/\delta), \quad \gamma_k = 4H\big(12\sqrt{|\mathcal{A}|d} + \sqrt{c_0}d\big)/C_{\mathcal{S}}^- \cdot \sqrt{\log(2Hk|\mathcal{F}|/\delta)},$$

we obtain with probability at least $1-3\delta$,

$$\frac{1}{K}\sum_{k=1}^{K}\Big[V_1^{\pi^*}(s_1) - V_1^{\pi^k}(s_1)\Big] \lesssim \sqrt{C\log(H|\mathcal{F}|K/\delta)\log(c_0'K)/K},$$

where $C = H^4d^4|\mathcal{A}|/(C_{\mathcal{S}}^-)^2 + H^4d^3|\mathcal{A}|^2/(C_{\mathcal{S}}^-)^2 + H^6d^2|\mathcal{A}|/(C_{\mathcal{S}}^-)^2 + H^6d^3$ and $c_0, c_0'$ are absolute constants. This completes the proof. □

# D. Theoretical Analysis for Markov Game

## D.1. Lemmas

**Lemma D.1** (Learning Target of Contrastive Loss). *For any $(s,a,b) \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}$ that is reachable under certain sampling strategy, the learning target of the contrastive loss in (2) with setting $z = (s,a,b)$ is*

$$f_h^*(s,a,b,s') = \frac{\mathbb{P}_h(s'|s,a,b)}{\mathcal{P}_{\mathcal{S}}^-(s')}.$$

*Proof.* For any $h \in [H]$, we let $\Pr_h$ denote the probability for some event at the $h$-th step of a Markov game. The contrastive loss in (2) with setting $z = (s,a,b)$ implicitly assumes

$$\Pr_h(y|s,a,b,s') = \left(\frac{f_h^*(s,a,b,s')}{1+f_h^*(s,a,b,s')}\right)^y \left(\frac{1}{1+f_h^*(s,a,b,s')}\right)^{1-y}.$$

In addition, by Bayes' rule, we also have

$$\Pr_h(y|s,a,b,s') = \frac{\Pr_h(s,a,b,s'|y)\Pr_h(y)}{\sum_{y\in\{0,1\}}\Pr_h(s,a,b,s'|y)\Pr_h(y)} = \frac{\Pr_h(s,a,b,s'|y)}{\Pr_h(s,a,b)\mathbb{P}_h(s'|s,a,b)+\Pr_h(s,a,b)\mathcal{P}_{\mathcal{S}}^-(s')},$$

where we use $\Pr_h(y) = 1/2$ for any $y \in \{0,1\}$ according to Algorithm 4. In the last equality, we also have

$$\Pr_h(s,a,b,s'|y=1) = \Pr_h(s,a,b|y=1)\Pr_h(s'|y=1,s,a,b) = \Pr_h(s,a,b)\mathbb{P}_h(s'|s,a,b),$$
$$\Pr_h(s,a,b,s'|y=0) = \Pr_h(s,a,b|y=0)\Pr_h(s'|y=0,s,a,b) = \Pr_h(s,a,b)\mathcal{P}_{\mathcal{S}}^-(s'),$$

where we use $\Pr_h(s,a,b|y=1) = \Pr_h(s,a,b|y=0) = \Pr_h(s,a,b)$ and also $\Pr_h(s'|y=1,s,a,b) = \mathbb{P}_h(s'|s,a,b)$, $\Pr_h(s'|y=0,s,a,b) = \mathcal{P}_{\mathcal{S}}^-(s')$.

Combining the above results, when $y=1$ at the $h$-th step, we obtain

$$\frac{f_h^*(s,a,b,s')}{1+f_h^*(s,a,b,s')} = \frac{\Pr_h(s,a,b)\mathbb{P}_h(s'|s,a,b)}{\Pr_h(s,a,b)\mathbb{P}_h(s'|s,a,b)+\Pr_h(s,a,b)\mathcal{P}_{\mathcal{S}}^-(s')},$$

which gives

$$f_h^*(s,a,b,s') = \frac{\mathbb{P}_h(s'|s,a,b)}{\mathcal{P}_{\mathcal{S}}^-(s')}.$$

Equivalently, when $y=0$, we get the same result. This completes the proof. □

**Lemma D.2.** *Suppose the policies $\pi^k$, $\nu^k$, the estimated transition $\widehat{\mathbb{P}}^k$, and the bonus $\beta^k$ are obtained at episode $k$ of Algorithm 2. Let $\mathrm{br}(\cdot)$ denote the best response policy given the opponent's policy. Moreover, $\underline{V}_{k,1}^{\sigma}(s_1)$ denotes the value function under any joint policy $\sigma$ for the zero-sum Markov game defined by the reward function $r-\beta^k$ and $\widehat{\mathbb{P}}^k$ while $\overline{V}_{k,1}^{\sigma}(s_1)$ denotes the value function for the zero-sum Markov game defined by $r+\beta^k$ and $\widehat{\mathbb{P}}^k$. Then, we have the following value function differences decomposed as*

$$V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - \overline{V}_{k,1}^{\mathrm{br}(\nu^k),\nu^k}(s_1) = \mathbb{E}\left[\sum_{h=1}^{H}\left(-\beta_h^k(s_h,a_h,b_h) + (\mathbb{P}_h-\widehat{\mathbb{P}}_h^k)V_{h+1}^{\mathrm{br}(\nu^k),\nu^k}(s_h,a_h,b_h)\right)\,\middle|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k\right],$$
$$\underline{V}_{k,1}^{\pi^k,\mathrm{br}(\pi^k)}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1) = \mathbb{E}\left[\sum_{h=1}^{H}\left(-\beta_h^k(s_h,a_h,b_h) - (\mathbb{P}_h-\widehat{\mathbb{P}}_h^k)V_{h+1}^{\pi^k,\mathrm{br}(\pi^k)}(s_h,a_h,b_h)\right)\,\middle|\,\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k\right].$$

*Proof.* Consider two zero-sum Markov games defined by $(\mathcal{S},\mathcal{A},\mathcal{B},H,r,\mathbb{P})$ and $(\mathcal{S},\mathcal{A},\mathcal{B},H,r+\beta,\mathbb{P}')$ where $\mathbb{P}$ and $\mathbb{P}'$ are any transition models and $r$ and $\beta$ are arbitrary reward function and bonus term. Then, for any joint policy $\sigma$, we let $Q_h^\sigma$ and $V_h^\sigma$ be the associated Q-function and value function at the $h$-th step for the Markov game defined by $(\mathcal{S},\mathcal{A},\mathcal{B},H,r,\mathbb{P})$, and $\widetilde{Q}_h^\sigma$ and $\widetilde{V}_h^\sigma$ be the associated Q-function and value function at the $h$-th step for the Markov game defined by $(\mathcal{S},\mathcal{A},\mathcal{B},H,r+\beta,\mathbb{P}')$. Then, by Bellman equation, we have for any $(s_h,a_h,b_h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}$,

$$\begin{aligned}
&Q_h^\sigma(s_h,a_h,b_h) - \widetilde{Q}_h^\sigma(s_h,a_h,b_h)\\
&\quad= -\beta_h(s_h,a_h,b_h) + \mathbb{P}_hV_{h+1}^\sigma(s_h,a_h,b_h) - \mathbb{P}'_h\widetilde{V}_{h+1}^\sigma(s_h,a_h,b_h)\\
&\quad= -\beta_h(s_h,a_h,b_h) + \mathbb{P}_hV_{h+1}^\sigma(s_h,a_h,b_h) - \mathbb{P}'_h\widetilde{V}_{h+1}^\sigma(s_h,a_h,b_h)\\
&\quad= -\beta_h(s_h,a_h,b_h) + (\mathbb{P}_h-\mathbb{P}'_h)V_{h+1}^\sigma(s_h,a_h,b_h) + \mathbb{P}'_h[V_{h+1}^\sigma(s_h,a_h,b_h) - \widetilde{V}_{h+1}^\sigma(s_h,a_h,b_h)].
\end{aligned}$$

Further by the Bellman equation and the above result, we have

$$\begin{aligned}
&V_h^\sigma(s_h) - \widetilde{V}_h^\sigma(s_h)\\
&\quad= \langle\sigma_h(\cdot,\cdot|s_h), Q_h^\sigma(s_h,\cdot,\cdot) - \widetilde{Q}_h^\sigma(s_h,\cdot,\cdot)\rangle\\
&\quad= \langle\sigma_h(\cdot,\cdot|s_h), -\beta_h(s_h,\cdot,\cdot) + (\mathbb{P}_h-\mathbb{P}'_h)V_{h+1}^\sigma(s_h,\cdot,\cdot) + \mathbb{P}'_h[V_{h+1}^\sigma(s_h,\cdot,\cdot) - \widetilde{V}_{h+1}^\sigma(s_h,\cdot,\cdot)]\rangle.
\end{aligned}$$

Since $V_{H+1}^{\sigma}(s) = \widetilde{V}_{H+1}^{\sigma}(s) = 0$ for any $s \in \mathcal{S}$ and $\sigma$, recursively applying the above relation, we have

$$V_1^{\sigma}(s_1) - \widetilde{V}_1^{\sigma}(s_1) = \mathbb{E}\left[\sum_{h=1}^{H}\left(-\beta_h(s_h,a_h,b_h) + (\mathbb{P}_h - \mathbb{P}'_h)V_{h+1}^{\sigma}(s_h,a_h,b_h)\right) \,\middle|\, \sigma, \mathbb{P}'\right].$$

For any episode $k$, setting $\mathbb{P}', \sigma, \beta$ to be $\widehat{\mathbb{P}}^k, (\mathrm{br}(\nu^k), \nu^k), \beta^k$ defined in Algorithm 2 and $\mathbb{P}, r$ to be the true transition model and reward function, by the above equality, according to the definition of $V_h^{\sigma}$ and $\overline{V}_{k,h}^{\sigma}$, we obtain

$$V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - \overline{V}_{k,1}^{\mathrm{br}(\nu^k),\nu^k}(s_1) = \mathbb{E}\left[\sum_{h=1}^{H}\left(-\beta_h^k(s_h,a_h,b_h) + (\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)V_{h+1}^{\mathrm{br}(\nu^k),\nu^k}(s_h,a_h,b_h)\right) \,\middle|\, \mathrm{br}(\nu^k), \nu^k, \widehat{\mathbb{P}}^k\right].$$

Moreover, setting $\mathbb{P}', \sigma, \beta$ to be $\widehat{\mathbb{P}}^k, (\pi^k, \mathrm{br}(\pi^k)), -\beta^k$ defined in Algorithm 2 and $\mathbb{P}, r$ to be the true transition model and reward function, by the definition of $V_h^{\sigma}$ and $\underline{V}_h^{\sigma}$, we obtain

$$V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1) - \underline{V}_{k,1}^{\pi^k,\mathrm{br}(\pi^k)}(s_1) = \mathbb{E}\left[\sum_{h=1}^{H}\left(\beta_h^k(s_h,a_h,b_h) + (\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)V_{h+1}^{\pi^k,\mathrm{br}(\pi^k)}(s_h,a_h,b_h)\right) \,\middle|\, \pi^k, \mathrm{br}(\pi^k), \widehat{\mathbb{P}}^k\right],$$

which leads to

$$\underline{V}_{k,1}^{\pi^k,\mathrm{br}(\pi^k)}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1) = \mathbb{E}\left[\sum_{h=1}^{H}\left(-\beta_h^k(s_h,a_h,b_h) - (\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)V_{h+1}^{\pi^k,\mathrm{br}(\pi^k)}(s_h,a_h,b_h)\right) \,\middle|\, \pi^k, \mathrm{br}(\pi^k), \widehat{\mathbb{P}}^k\right].$$

This completes the proof. □

**Lemma D.3.** *Suppose the joint policy $\sigma^k$, the estimated transition $\widehat{\mathbb{P}}^k$, and the bonus $\beta^k$ are obtained at episode $k$ of Algorithm 2. Moreover, $\overline{V}_1^k(s_1)$ and $\underline{V}_1^k(s_1)$ are the estimated value functions based on UCB and LCB obtained at episode $k$ of Algorithm 2. Then, their difference can be decomposed as*

$$\overline{V}_1^k(s_1) - \underline{V}_1^k(s_1) = \mathbb{E}\left[\sum_{h=1}^{H} 2\beta_h^k(s_h,a_h,b_h) + (\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h,a_h,b_h) \,\middle|\, \sigma^k, \mathbb{P}\right].$$

*Proof.* For the episode $k$, we consider two Markov games defined by $(\mathcal{S}, \mathcal{A}, \mathcal{B}, H, r + \beta^k, \widehat{\mathbb{P}}^k)$ and $(\mathcal{S}, \mathcal{A}, \mathcal{B}, H, r - \beta^k, \widehat{\mathbb{P}}^k)$. Then, for the joint policy $\sigma^k$, by Algorithm 2, we have for any $(s_h, a_h, b_h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$,

$$\begin{aligned}
&\overline{Q}_h^k(s_h,a_h,b_h) - \underline{Q}_h^k(s_h,a_h,b_h) \\
&\quad = 2\beta_h^k(s_h,a_h,b_h) + \widehat{\mathbb{P}}_h^k(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h,a_h,b_h) \\
&\quad = 2\beta_h^k(s_h,a_h,b_h) + (\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h,a_h,b_h) + \mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h,a_h,b_h).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\overline{V}_h^k(s_h) - \underline{V}_h^k(s_h) \\
&\quad = \langle \sigma_h^k(\cdot,\cdot|s_h), \overline{Q}_h^k(s_h,\cdot,\cdot) - \underline{Q}_h^k(s_h,\cdot,\cdot)\rangle \\
&\quad = 2\left\langle \sigma_h^k(\cdot,\cdot|s_h), \beta_h^k(s_h,a_h,b_h)\right\rangle + \left\langle \sigma_h^k(\cdot,\cdot|s_h), (\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h,\cdot,\cdot)\right\rangle \\
&\qquad + \left\langle \sigma_h^k(\cdot,\cdot|s_h), \mathbb{P}_h(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h,\cdot,\cdot)\right\rangle.
\end{aligned}$$

By the fact that $\overline{V}_{H+1}^k(s) = \underline{V}_{H+1}^k(s) = 0$ for any $s \in \mathcal{S}$, recursively applying the above relation, we have

$$\overline{V}_1^k(s_1) - \underline{V}_1^k(s_1) = \mathbb{E}\left[\sum_{h=1}^{H} 2\beta_h^k(s_h,a_h,b_h) + (\widehat{\mathbb{P}}_h^k - \mathbb{P}_h)(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k)(s_h,a_h,b_h) \,\middle|\, \sigma^k, \mathbb{P}\right].$$

This completes the proof. □

**Lemma D.4.** *Suppose that $\widehat{\mathbb{P}}^k$ is the estimated transition obtained at episode $k$ of Algorithm 2. We define $\zeta_{h-1}^k := \mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)}\|\widehat{\mathbb{P}}_{h-1}^k(\cdot|s'',a'',b'') - \mathbb{P}_{h-1}(\cdot|s'',a'',b'')\|_1^2$ for all $h \geq 2$, $\widetilde{\rho}_h^k(\cdot,\cdot,\cdot) := \frac{1}{k}\sum_{k'=0}^{k-1}\widetilde{d}_h^{\sigma^{k'}}(\cdot,\cdot,\cdot)$ for all $h \geq 1$ with $\widetilde{\rho}_1^k(s_1,a,b) = \mathrm{Unif}(a)\mathrm{Unif}(b)$, and $\breve{\rho}_h^k(\cdot,\cdot,\cdot) := \frac{1}{k}\sum_{k'=0}^{k-1}\breve{d}_h^{\sigma^{k'}}(\cdot,\cdot,\cdot)$ for all $h \geq 2$. Then for any function $g : \mathcal{S}\times\mathcal{A}\times\mathcal{B}\mapsto[0,B]$ and joint policy $\sigma$, we have for any $h \geq 2$, the following inequality holds*

$$\begin{aligned}&\left|\mathbb{E}_{(s,a,b)\sim d_h^{\sigma,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}[g(s,a,b)]\right| \\ &\leq \sqrt{2kB^2\zeta_{h-1}^k + 2k|\mathcal{A}||\mathcal{B}|\cdot\mathbb{E}_{(s,a,b)\sim\breve{\rho}_h^k(\cdot,\cdot,\cdot)}[g(s,a,b)^2] + \lambda_k B^2 d/(C_{\mathcal{S}}^-)^2}\cdot\mathbb{E}_{(s',a',b')\sim d_{h-1}^{\sigma,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}\left\|\widehat{\phi}_{h-1}^k(s',a',b')\right\|_{\Sigma_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}^{-1}}.\end{aligned}$$

*In addition, for $h = 1$, we have*

$$\left|\mathbb{E}_{(s,a,b)\sim d_1^{\sigma,\mathbb{P}}(\cdot,\cdot,\cdot)}[g(s,a,b)]\right| = \sqrt{\mathbb{E}_{(a,b)\sim\sigma_1(\cdot,\cdot|s_1)}[g(s_1,a,b)^2]} \leq \sqrt{|\mathcal{A}||\mathcal{B}|\mathbb{E}_{(a,b)\sim\widetilde{\rho}_1^k(s_1,\cdot,\cdot)}[g(s_1,a,b)^2]}.$$

*Proof.* For any function $g : \mathcal{S}\times\mathcal{A}\times\mathcal{B}\mapsto[0,B]$ and any joint policy $\sigma$, under the estimated transition model $\widehat{\mathbb{P}}^k$ at the $k$-th episode, for any $h \geq 2$, we have

$$\begin{aligned}&\left|\mathbb{E}_{(s,a,b)\sim d_h^{\sigma,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}[g(s,a,b)]\right| \\ &= \left|\mathbb{E}_{(s',a',b')\sim d_{h-1}^{\sigma,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot),s\sim\widehat{\mathbb{P}}_{h-1}^k(\cdot|s',a',b'),(a,b)\sim\sigma_h(\cdot,\cdot|s)}[g(s,a,b)]\right| \\ &= \left|\mathbb{E}_{(s',a',b')\sim d_{h-1}^{\sigma,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}\left[\widehat{\phi}_{h-1}^k(s',a',b')^\top\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right]\right| \\ &\leq \mathbb{E}_{(s',a',b')\sim d_{h-1}^{\sigma,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}\left\|\widehat{\phi}_{h-1}^k(s',a',b')\right\|_{\Sigma_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}^{-1}}\left\|\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right\|_{\Sigma_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}},\end{aligned} \tag{33}$$

where the inequality is due to the Cauchy-Schwarz inequality. We define the covariance matrix as $\Sigma_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k} := k\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)}[\widehat{\phi}_{h-1}^k(s,a,b)\widehat{\phi}_{h-1}^k(s,a,b)^\top] + \lambda_k I$ with $\widetilde{\rho}_{h-1}^k(s,a,b) = \frac{1}{k}\sum_{k'=0}^{k-1}\widetilde{d}_{h-1}^{\sigma^{k'}}(s,a,b)$.

Moreover, we have

$$\begin{aligned}&\left\|\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right\|_{\Sigma_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}}^2 \\ &= k\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right)^\top\mathbb{E}_{\widetilde{\rho}_{h-1}^k}\left[\widehat{\phi}_{h-1}^k(\widehat{\phi}_{h-1}^k)^\top\right]\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right) \\ &\quad+\lambda_k\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right) \\ &= k\mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)}\left[\int_{\mathcal{S}}\widehat{\phi}_{h-1}^k(s'',a'',b'')^\top\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right] \\ &\quad+\lambda_k\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}}\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right) \\ &\leq k\mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)}\left[\int_{\mathcal{S}}\widehat{\phi}_{h-1}^k(s'',a'',b'')^\top\widehat{\psi}_{h-1}^k(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right]^2+\lambda_k B^2 d/(C_{\mathcal{S}}^-)^2,\end{aligned} \tag{34}$$

where the last inequality is by

$$\left(\int_{\mathcal{S}} \widehat{\psi}_{h-1}^k(s) \sum_{a\in\mathcal{A},b\in\mathcal{B}} \sigma_h(a,b|s) g(s,a,b) \mathrm{d}s\right)^\top \left(\int_{\mathcal{S}} \widehat{\psi}_{h-1}^k(s) \sum_{a\in\mathcal{A},b\in\mathcal{B}} \sigma_h(a,b|s) g(s,a,b) \mathrm{d}s\right) \le B^2 d/(C_{\mathcal{S}}^-)^2,$$

since $0 \le g(s,a,b) \le B$ and $\|\int_{\mathcal{S}} \widehat{\psi}_{h-1}^k(s)\mathrm{d}s\|_2^2 := \|\int_{\mathcal{S}} \mathcal{P}_{\mathcal{S}}^-(s)\widetilde{\psi}_{h-1}^k(s)\mathrm{d}s\|_2^2 \le \|\int_{\mathcal{S}} \widetilde{\psi}_{h-1}^k(s)\mathrm{d}s\|_2^2 \le (\int_{\mathcal{S}} \|\widetilde{\psi}_{h-1}^k(s)\|_2 \mathrm{d}s)^2 \le d/(C_{\mathcal{S}}^-)^2$ according to the definition of the function class in Definition 3.3 and the assumption that all states are normalized such that $\mathrm{Vol}(\mathcal{S}) \le 1$. In addition, we have

$$\begin{aligned}
&k\mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)} \left[\int_{\mathcal{S}} \widehat{\phi}_{h-1}^k(s'',a'',b'')^\top \widehat{\psi}_{h-1}^k(s) \sum_{a\in\mathcal{A},b\in\mathcal{B}} \sigma_h(a,b|s) g(s,a,b)\mathrm{d}s\right]^2 \\
&\le 2k\mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)} \left[\int_{\mathcal{S}} \left(\widehat{\mathbb{P}}_{h-1}^k(s|s'',a'',b'') - \mathbb{P}_{h-1}(s|s'',a'',b'')\right) \sum_{a\in\mathcal{A},b\in\mathcal{B}} \sigma_h(a,b|s) g(s,a,b)\mathrm{d}s\right]^2 \\
&\quad + 2k\mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)} \left[\int_{\mathcal{S}} \mathbb{P}_{h-1}(s|s'',a'',b'') \sum_{a\in\mathcal{A},b\in\mathcal{B}} \sigma_h(a,b|s) g(s,a,b)\mathrm{d}s\right]^2 \\
&\le 2kB^2\zeta_{h-1}^k + 2k\mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)} \left[\int_{\mathcal{S}} \mathbb{P}_{h-1}(s|s'',a'',b'') \sum_{a\in\mathcal{A},b\in\mathcal{B}} \sigma_h(a,b|s) g(s,a,b)\mathrm{d}s\right]^2 \\
&\le 2kB^2\zeta_{h-1}^k + 2k\mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot),s\sim\mathbb{P}_{h-1}(\cdot|s'',a'',b''),(a,b)\sim\sigma_h(\cdot,\cdot|s)}[g(s,a,b)^2] \\
&\le 2kB^2\zeta_{h-1}^k + 2k \sup_{a\in\mathcal{A},b\in\mathcal{B},s\in\mathcal{S}} \frac{\sigma_h(a,b|s)}{\mathrm{Unif}(a)\mathrm{Unif}(b)} \mathbb{E}_{(s,a,b)\sim\breve{\rho}_h^k(\cdot,\cdot,\cdot)}[g(s,a,b)^2] \\
&= 2kB^2\zeta_{h-1}^k + 2k|\mathcal{A}||\mathcal{B}| \cdot \mathbb{E}_{(s,a,b)\sim\breve{\rho}_h^k(\cdot,\cdot,\cdot)}[g(s,a,b)^2],
\end{aligned} \tag{35}$$

where the first inequality is by $(a+b)^2 \le 2a^2 + 2b^2$, the second inequality is by $\mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)} [\int_{\mathcal{S}} (\widehat{\mathbb{P}}_{h-1}^k(s|s'',a'',b'') - \mathbb{P}_{h-1}(s|s'',a'',b'')) \sum_{a\in\mathcal{A},b\in\mathcal{B}} \sigma_h(a,b|s) g(s,a,b)\mathrm{d}s]^2 \le B^2 \mathbb{E}_{(s'',a'',b'')\sim\widetilde{\rho}_{h-1}^k(\cdot,\cdot,\cdot)} \|\widehat{\mathbb{P}}_{h-1}^k(\cdot|s'',a'',b'') - \mathbb{P}_{h-1}(\cdot|s'',a'',b'')\|_1^2 \le B^2\zeta_{h-1}^k$, the third inequality is by Jensen's inequality, and the fourth inequality is by substituting the joint policy $\sigma$ with the uniform distribution.

Combining (33),(34), and (35), we have for any $h \ge 2$,

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a,b)\sim d_h^{\sigma,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}[g(s,a,b)]\right| \\
&\le \sqrt{2kB^2\zeta_{h-1}^k + 2k|\mathcal{A}||\mathcal{B}| \cdot \mathbb{E}_{(s,a,b)\sim\breve{\rho}_h^k(\cdot,\cdot,\cdot)}[g(s,a,b)^2] + \lambda_k B^2 d/(C_{\mathcal{S}}^-)^2} \cdot \mathbb{E}_{(s',a',b')\sim d_{h-1}^{\sigma,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)} \left\|\widehat{\phi}_{h-1}^k(s',a',b')\right\|_{\Sigma_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}^{-1}}.
\end{aligned}$$

On the other hand, for $h = 1$, we have

$$\left|\mathbb{E}_{(s,a,b)\sim d_1^{\sigma,\mathbb{P}}(\cdot,\cdot,\cdot)}[g(s,a,b)]\right| = \sqrt{\mathbb{E}_{(a,b)\sim\sigma_1(\cdot,\cdot|s_1)}[g(s_1,a,b)^2]} \le \sqrt{|\mathcal{A}||\mathcal{B}|\mathbb{E}_{(a,b)\sim\widetilde{\rho}_1^k(s_1,\cdot,\cdot)}[g(s_1,a,b)^2]},$$

where we let $\widetilde{\rho}_1^k(s_1,a,b) = \mathrm{Unif}(a)\mathrm{Unif}(b)$ and the last inequality is by $\mathbb{E}_{(a,b)\sim\sigma_1(\cdot,\cdot|s_1)}[g(s_1,a,b)^2] \le \max_{a,b} \frac{\sigma_1(a,b|s_1)}{\mathrm{Unif}(a)\mathrm{Unif}(b)} \mathbb{E}_{(a,b)\sim\widetilde{\rho}_1^k(s_1,\cdot,\cdot)}[g(s_1,a,b)^2]$. The proof is completed. □

**Lemma D.5.** *Suppose that $\widehat{\mathbb{P}}^k$ is the estimated transition obtained at episode $k$ of Algorithm 2. We define $\widetilde{\rho}_h^k(\cdot,\cdot,\cdot) := \frac{1}{k}\sum_{k'=0}^{k-1} \widetilde{d}_h^{\sigma^{k'}}(\cdot,\cdot,\cdot)$ for all $h \ge 1$ with $\widetilde{\rho}_1^k(s_1,a,b) = \mathrm{Unif}(a)\mathrm{Unif}(b)$ and $\rho_h^k(\cdot,\cdot,\cdot) := \frac{1}{k}\sum_{k'=0}^{k-1} d_h^{\sigma^{k'}}(\cdot,\cdot,\cdot)$ for all $h \ge 2$. Then for any function $g : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \mapsto [0,B]$ and joint policy $\sigma$, we have for any $h \ge 2$, the following inequality holds*

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a,b)\sim d_h^{\sigma,\mathbb{P}}(\cdot,\cdot,\cdot)}[g(s,a,b)]\right| \\
&\le \sqrt{k|\mathcal{A}||\mathcal{B}| \cdot \mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}[g(s,a,b)^2] + \lambda_k B^2 d} \cdot \mathbb{E}_{(s',a',b')\sim d_{h-1}^{\sigma,\mathbb{P}}(\cdot,\cdot,\cdot)} \left\|\phi_{h-1}^*(s',a',b')\right\|_{\Sigma_{\rho_{h-1}^k,\phi_{h-1}^*}^{-1}}.
\end{aligned}$$

*In addition, for $h = 1$, we have*

$$\left|\mathbb{E}_{(s,a,b)\sim d_1^{\sigma,\mathbb{P}}(\cdot,\cdot,\cdot)}[g(s,a,b)]\right| \le \sqrt{\mathbb{E}_{(a,b)\sim\sigma_1(\cdot,\cdot|s_1)}[g(s_1,a,b)^2]} \le \sqrt{|\mathcal{A}||\mathcal{B}|\mathbb{E}_{(a,b)\sim\widetilde{\rho}_1^k(s_1,\cdot,\cdot)}[g(s_1,a,b)^2]}.$$

*Proof.* For any function $g : \mathcal{S}\times\mathcal{A}\times\mathcal{B}\mapsto\mathbb{R}$ and any joint policy $\sigma$, under the true transition model $\mathbb{P}$, for any $h \ge 2$, we have

$$\begin{aligned}
&\left|\mathbb{E}_{(s,a,b)\sim d_h^{\sigma,\mathbb{P}}(\cdot,\cdot,\cdot)}[g(s,a,b)]\right| \\
&= \left|\mathbb{E}_{(s',a',b')\sim d_{h-1}^{\sigma,\mathbb{P}}(\cdot,\cdot,\cdot),s\sim\mathbb{P}_{h-1}(\cdot|s',a'),(a,b)\sim\sigma_h(\cdot,\cdot|s)}[g(s,a,b)]\right| \\
&= \left|\mathbb{E}_{(s',a',b')\sim d_{h-1}^{\sigma,\mathbb{P}}(\cdot,\cdot,\cdot)}\left[\phi_{h-1}^*(s',a',b')^\top\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right]\right| \qquad (36)\\
&\le \mathbb{E}_{(s',a',b')\sim d_{h-1}^{\sigma,\mathbb{P}}(\cdot,\cdot,\cdot)}\left\|\phi_{h-1}^*(s',a',b')\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}\left\|\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right\|_{\Sigma_{\rho_{h-1}^k,\phi_{h-1}^*}},
\end{aligned}$$

where the inequality is by Cauchy-Schwarz inequality. We define the covariance matrix $\Sigma_{\rho_{h-1}^k,\phi_{h-1}^*} := k\mathbb{E}_{(s,a,b)\sim\rho_{h-1}^k}[\phi_{h-1}^*(s,a,b)\phi_{h-1}^*(s,a,b)^\top] + \lambda_k I$ with $\rho_{h-1}^k(s,a,b) = \frac{1}{k}\sum_{k'=0}^{k-1} d_{h-1}^{\pi^{k'}}(s,a,b)$.

Next, we have

$$\begin{aligned}
&\left\|\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right\|^2_{\Sigma_{\rho_{h-1}^k,\phi_{h-1}^*}} \\
&= k\left(\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right)^\top\mathbb{E}_{\rho_{h-1}^k}\left[\phi_{h-1}^*(\phi_{h-1}^*)^\top\right]\left(\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right) \\
&\quad+\lambda_k\left(\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right) \\
&= k\mathbb{E}_{(s'',a'',b'')\sim\rho_{h-1}^k(\cdot,\cdot,\cdot)}\left[\int_{\mathcal{S}}\phi_{h-1}^*(s'',a'',b'')^\top\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right] \\
&\quad+\lambda_k\left(\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right) \\
&\le k\mathbb{E}_{(s'',a'',b'')\sim\rho_{h-1}^k(\cdot,\cdot,\cdot)}\left[\int_{\mathcal{S}}\phi_{h-1}^*(s'',a'',b'')^\top\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right]^2+\lambda_k B^2 d, \qquad (37)
\end{aligned}$$

where, by Assumption 2.1, the last inequality is due to

$$\begin{aligned}
&\left(\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right)^\top\left(\int_{\mathcal{S}}\psi_{h-1}^*(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right) \\
&\qquad\le B^2\left|\int_{\mathcal{S}}\psi_{h-1}^*(s)\mathrm{d}s\right|_2^2\le B^2 d.
\end{aligned}$$

Moreover, we have

$$
\begin{aligned}
&k\mathbb{E}_{(s'',a'',b'')\sim\rho^k_{h-1}(\cdot,\cdot,\cdot)}\left[\int_{\mathcal{S}}\phi^*_{h-1}(s'',a'',b'')^\top\psi^*_{h-1}(s)\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right]^2\\
&\quad= k\mathbb{E}_{(s'',a'',b'')\sim\rho^k_{h-1}(\cdot,\cdot,\cdot)}\left[\int_{\mathcal{S}}\mathbb{P}_{h-1}(s|s'',a'',b'')\sum_{a\in\mathcal{A},b\in\mathcal{B}}\sigma_h(a,b|s)g(s,a,b)\mathrm{d}s\right]^2\\
&\quad\le k\mathbb{E}_{(s'',a'',b'')\sim\rho^k_{h-1}(\cdot,\cdot,\cdot),s\sim\mathbb{P}_{h-1}(\cdot|s'',a'',b''),(a,b)\sim\sigma_h(\cdot,\cdot|s)}[g(s,a,b)^2]\\
&\quad\le k\sup_{a\in\mathcal{A},b\in\mathcal{A},s\in\mathcal{S}}\frac{\sigma_h(a,b|s)}{\mathrm{Unif}(a)\mathrm{Unif}(b)}\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}^k_h(\cdot,\cdot,\cdot)}[g(s,a,b)^2]\\
&\quad= k|\mathcal{A}||\mathcal{B}|\cdot\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}^k_h(\cdot,\cdot,\cdot)}[g(s,a,b)^2],
\end{aligned}
\tag{38}
$$

where the first inequality is due to Jensen's inequality and the second inequality is by substituting the joint policy $\sigma$ with the uniform distribution and $\widetilde{\rho}^k_h(s,a,b) := \rho^k_{h-1}(s',a',b')\mathbb{P}_{h-1}(s|s',a',b')\mathrm{Unif}(a)\mathrm{Unif}(b)$ for all $h\ge 2$. Combining (36),(37), and (38), we have for any $h\ge 2$,

$$
\begin{aligned}
&\left|\mathbb{E}_{(s,a,b)\sim d^{\sigma,\mathbb{P}}_h(\cdot,\cdot,\cdot)}[g(s,a,b)]\right|\\
&\quad\le\sqrt{k|\mathcal{A}||\mathcal{B}|\cdot\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}^k_h(\cdot,\cdot,\cdot)}[g(s,a,b)^2]+\lambda_k B^2 d}\cdot\mathbb{E}_{(s',a',b')\sim d^{\sigma,\mathbb{P}}_{h-1}(\cdot,\cdot,\cdot)}\left\|\phi^*_{h-1}(s',a',b')\right\|_{\Sigma^{-1}_{\rho^k_{h-1},\phi^*_{h-1}}}.
\end{aligned}
$$

For $h=1$, we have

$$
\left|\mathbb{E}_{(s,a,b)\sim d^{\sigma,\mathbb{P}}_1(\cdot,\cdot,\cdot)}[g(s,a,b)]\right|\le\sqrt{\mathbb{E}_{(a,b)\sim\sigma_1(\cdot,\cdot|s_1)}[g(s_1,a,b)^2]}\le\sqrt{|\mathcal{A}||\mathcal{B}|\mathbb{E}_{(a,b)\sim\widetilde{\rho}^k_1(s_1,\cdot,\cdot)}[g(s_1,a,b)^2]},
$$

where we let $\widetilde{\rho}^k_1(s_1,a,b) = \mathrm{Unif}(a)\mathrm{Unif}(b)$ and the last inequality is by $\mathbb{E}_{(a,b)\sim\sigma_1(\cdot,\cdot|s_1)}[g(s_1,a,b)^2]\le\max_{a,b}\frac{\sigma_1(a,b|s_1)}{\mathrm{Unif}(a)\mathrm{Unif}(b)}\mathbb{E}_{(a,b)\sim\widetilde{\rho}^k_1(s_1,\cdot,\cdot)}[g(s_1,a,b)^2]$. The proof is completed. □

**Lemma D.6.** *Suppose at the $k$-th episode of Algorithm 2, $\pi^k,\nu^k$ are learned policies , $\iota_k$ is the CCE learning accuracy, and $\overline{V}^k_1(s_1)$ and $\underline{V}^k_1(s_1)$ are the value functions updated as in the algorithm. Moreover, for any joint policy $\sigma$, $\overline{V}^\sigma_{k,1}(s_1)$ is the value function associated with the Markov game defined by the reward function $r+\beta^k$ and the estimated transition $\widehat{\mathbb{P}}^k$ while $\underline{V}^\sigma_{k,1}(s_1)$ is the value function associated with the Markov game defined by the reward function $r-\beta^k$ and the estimated transition $\widehat{\mathbb{P}}^k$. Then we have*

$$
\begin{aligned}
\overline{V}^{\mathrm{br}(\nu^k),\nu^k}_{k,1}(s_1)&\le\overline{V}^k_1(s_1)+H\iota_k,\\
\underline{V}^{\pi^k,\mathrm{br}(\pi^k)}_{k,1}(s_1)&\ge\underline{V}^k_1(s_1)-H\iota_k.
\end{aligned}
$$

*Proof.* We prove this lemma by induction. For the first inequality in this lemma, we have $\overline{V}^{\mathrm{br}(\nu^k),\nu^k}_{k,H+1}(s)=\overline{V}^k_{H+1}(s)=0$ for any $s\in\mathcal{S}$. Next, we assume the following inequality holds

$$
\overline{V}^{\mathrm{br}(\nu^k),\nu^k}_{k,h+1}(s)\le\overline{V}^k_{h+1}(s)+(H-h)\iota_k.
$$

Then, with the above inequality, by the Bellman equation, we have

$$
\begin{aligned}
&\overline{Q}^{\mathrm{br}(\nu^k),\nu^k}_{k,h}(s,a,b)-\overline{Q}^k_h(s,a,b)\\
&\quad= r_h(s,a,b)+\beta^k_h(s,a,b)+\mathbb{P}_h\overline{V}^{\mathrm{br}(\nu^k),\nu^k}_{k,h+1}(s,a,b)-r_h(s,a,b)-\beta^k_h(s,a,b)-\mathbb{P}_h\overline{V}^k_{h+1}(s,a,b)\\
&\quad=\mathbb{P}_h\overline{V}^{\mathrm{br}(\nu^k),\nu^k}_{k,h+1}(s,a,b)-\mathbb{P}_h\overline{V}^k_{h+1}(s,a,b)\le(H-h)\iota_k.
\end{aligned}
\tag{39}
$$

Then, we have

$$\begin{aligned}\overline{V}_{k,h}^{\mathrm{br}(\nu^k),\nu^k}(s) &= \mathbb{E}_{a\sim\mathrm{br}(\nu^k)_h, b\sim\nu_h^k}\left[\overline{Q}_{k,h}^{\mathrm{br}(\nu^k),\nu^k}(s,a,b)\right] \\ &\le \mathbb{E}_{a\sim\mathrm{br}(\nu^k)_h, b\sim\nu_h^k}\left[\overline{Q}_h^k(s,a,b)\right] + (H-h)\iota_k \\ &\le \mathbb{E}_{(a,b)\sim\sigma_h^k}\left[\overline{Q}_h^k(s,a,b)\right] + (H+1-h)\iota_k \\ &= \overline{V}_h^k(s) + (H+1-h)\iota_k,\end{aligned}$$

where the first inequality is by (39) and the second inequality is by the definition of $\iota_k$-CCE as in Definition 4.1. Thus, we obtain

$$\overline{V}_{k,1}^{\mathrm{br}(\nu^k),\nu^k}(s_1) \le \overline{V}_1^k(s_1) + H\iota_k.$$

For the second inequality in this lemma, we have $\underline{V}_{k,H+1}^{\pi^k,\mathrm{br}(\pi^k)}(s) = \underline{V}_{H+1}^k(s) = 0$. Then, we assume that

$$\underline{V}_{k,h+1}^{\pi^k,\mathrm{br}(\pi^k)}(s) \ge \underline{V}_{h+1}^k(s) - (H-h)\iota_k.$$

Then, by the Bellman equation, we have

$$\begin{aligned}&\underline{Q}_{k,h}^{\pi^k,\mathrm{br}(\pi^k)}(s,a,b) - \underline{Q}_h^k(s,a,b) \\ &\quad= r_h(s,a,b) + \beta_h^k(s,a,b) + \mathbb{P}_h\underline{V}_{k,h+1}^{\pi^k,\mathrm{br}(\pi^k)}(s,a,b) - r_h(s,a,b) - \beta_h^k(s,a,b) - \mathbb{P}_h\underline{V}_{h+1}^k(s,a,b) \\ &\quad= \mathbb{P}_h\underline{V}_{k,h+1}^{\pi^k,\mathrm{br}(\pi^k)}(s,a,b) - \mathbb{P}_h\underline{V}_{h+1}^k(s,a,b) \ge -(H-h)\iota_k.\end{aligned} \tag{40}$$

Then, we have

$$\begin{aligned}\underline{V}_{k,h}^{\pi^k,\mathrm{br}(\pi^k)}(s) &= \mathbb{E}_{a\sim\pi_h^k, b\sim\mathrm{br}(\pi^k)_h}\left[\underline{Q}_{k,h}^{\pi^k,\mathrm{br}(\pi^k)}(s,a,b)\right] \\ &\ge \mathbb{E}_{a\sim\pi_h^k, b\sim\mathrm{br}(\pi^k)_h}\left[\underline{Q}_h^k(s,a,b)\right] - (H-h)\iota_k \\ &\ge \mathbb{E}_{(a,b)\sim\sigma_h^k}\left[\underline{Q}_h^k(s,a,b)\right] - (H+1-h)\iota_k \\ &= \underline{V}_h^k(s) - (H+1-h)\iota_k,\end{aligned}$$

where the first inequality is by (40) and the second inequality is by the definition of $\iota_k$-CCE as in Definition 4.1. Thus, we obtain

$$\underline{V}_{k,1}^{\pi^k,\mathrm{br}(\pi^k)}(s_1) \ge \underline{V}_1^k(s_1) - H\iota_k.$$

This completes the proof. □

### D.2. Proof of Lemma 5.2

The proof of this lemma follows from Proof of Lemma 5.1 by expanding the action space from $\mathcal{A}$ to $\mathcal{A}\times\mathcal{B}$. In this subsection, we briefly present the major steps of the proof.

*Proof.* For any function $f_h \in \mathcal{F}$, we let $\Pr_h^f(y|s,a,b,s')$ denote the conditional probability characterized by the function $f_h$ at the step $h$, which is

$$\Pr{}_h^f(y|s,a,b,s') = \left(\frac{f_h(s,a,b,s')}{1+f_h(s,a,b,s')}\right)^y \left(\frac{1}{1+f_h(s,a,b,s')}\right)^{1-y}.$$

Moreover, there is

$$\Pr{}_h^f(y,s'|s,a,b) = \Pr{}_h^f(y|s,a,b,s')\Pr{}_h(s'|s,a,b) = \left(\frac{f_h(s,a,b,s')\Pr{}_h(s'|s,a,b)}{1+f_h(s,a,b,s')}\right)^y \left(\frac{\Pr{}_h(s'|s,a,b)}{1+f_h(s,a,b,s')}\right)^{1-y},$$

where we have

$$\begin{aligned}\Pr{}_h(s'|s,a,b) &= \Pr{}_h(y=1|s,a,b)\Pr{}_h(s'|y=1,s,a,b) + \Pr{}_h(y=0|s,a,b)\Pr{}_h(s'|y=0,s,a,b)\\ &= \Pr{}_h(y=1)\Pr{}_h(s'|y=1,s,a,b) + \Pr{}_h(y=0)\Pr{}_h(s'|y=0,s,a,b)\\ &= \frac{1}{2}[\mathbb{P}_h(s'|s,a,b) + \mathcal{P}_{\mathcal{S}}^-(s')] \geq \frac{1}{2}C_{\mathcal{S}}^- > 0.\end{aligned} \tag{41}$$

Thus, we have the equivalency of solving the following two problems with $f_h(s,a,b,s') = \phi_h(s,a,b)^\top \psi_h(s')$, which is

$$\max_{\phi_h\in\Phi,\psi_h\in\Psi} \sum_{(s,a,s',y)\in\mathcal{D}_h^k} \log \Pr{}_h^f(y|s,a,b,s') = \max_{\phi_h,\psi_h} \sum_{(s,a,s',y)\in\mathcal{D}_h^k} \log \Pr{}_h^f(y,s'|s,a,b). \tag{42}$$

We denote the solution of (42) as $\widetilde{\phi}_h^k$ and $\widetilde{\psi}_h^k$ such that

$$\widehat{f}_h^k(s,a,b,s') = \widetilde{\psi}_h^k(s')^\top \widetilde{\phi}_h^k(s,a,b).$$

According to Algorithm 4, for any $h \geq 2$ and $k' \in [k]$, the data $(s,a,b)$ is sampled from both $\widetilde{d}_h^{\sigma^{k'}}(\cdot,\cdot,\cdot)$ and $\breve{d}_h^{\sigma^{k'}}(\cdot,\cdot,\cdot)$. Then, by Lemma E.2, solving the contrastive loss in (2) with letting $z=(s,a,b)$ gives, with probability at least $1-\delta$, for all $h \geq 2$,

$$\begin{aligned}\sum_{k'=1}^{k}\Bigg[&\mathbb{E}_{(s,a,b)\sim\widetilde{d}_h^{\sigma^{k'}}(\cdot,\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a,b) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a,b)\right\|_{\text{TV}}^2\\ &+ \mathbb{E}_{(s,a,b)\sim\breve{d}_h^{\sigma^{k'}}(\cdot,\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a,b) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a,b)\right\|_{\text{TV}}^2\Bigg] \leq 2\log(2kH|\mathcal{F}|/\delta),\end{aligned}$$

which is equivalent to

$$\begin{aligned}&\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a,b) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a,b)\right\|_{\text{TV}}^2\\ &+ \mathbb{E}_{(s,a,b)\sim\breve{\rho}_h^k(\cdot,\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a,b) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a,b)\right\|_{\text{TV}}^2 \leq 2\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \geq 2,\end{aligned} \tag{43}$$

where $\widetilde{\rho}_h^k(s,a,b) = \frac{1}{k}\sum_{k'=0}^{k-1}\widetilde{d}_h^{\pi^{k'}}(s,a,b)$ and $\breve{\rho}_h^k(s,a,b) = \frac{1}{k}\sum_{k'=0}^{k-1}\breve{d}_h^{\pi^{k'}}(s,a,b)$. Moreover, since for $h=1$, the data is only sampled from $\widetilde{d}_1^{\pi^{k'}}(\cdot,\cdot,\cdot)$ for any $k' \in [k]$, then we analogously have

$$\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_1^k(\cdot,\cdot,\cdot)}\left\|\Pr{}_1^{\widehat{f}^k}(\cdot,\cdot|s,a,b) - \Pr{}_1^{f^*}(\cdot,\cdot|s,a,b)\right\|_{\text{TV}}^2 \leq 2\log(2k|\mathcal{F}|/\delta)/k. \tag{44}$$

Thus, combining (43) and (44), with probability at least $1-2\delta$, we have

$$\begin{aligned}&\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a,b) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a,b)\right\|_{\text{TV}}^2 \leq 2\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \geq 1,\\ &\mathbb{E}_{(s,a,b)\sim\breve{\rho}_h^k(\cdot,\cdot,\cdot)}\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a,b) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a,b)\right\|_{\text{TV}}^2 \leq 2\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \geq 2,\end{aligned} \tag{45}$$

Next, we show the recovery error bound of the transition model based on (45). We have

$$\begin{aligned}&\left\|\Pr{}_h^{\widehat{f}^k}(\cdot,\cdot|s,a,b) - \Pr{}_h^{f^*}(\cdot,\cdot|s,a,b)\right\|_{\text{TV}}^2\\ &= \left(\left\|\Pr{}_h^{\widehat{f}^k}(y=0,\cdot|s,a,b) - \Pr{}_h^{f^*}(y=0,\cdot|s,a,b)\right\|_{\text{TV}} + \left\|\Pr{}_h^{\widehat{f}^k}(y=1,\cdot|s,a,b) - \Pr{}_h^{f^*}(y=1,\cdot|s,a,b)\right\|_{\text{TV}}\right)^2\\ &= 4\left\|\frac{\Pr{}_h(\cdot|s,a,b)}{1+\widehat{f}_h^k(s,a,b,\cdot)} - \frac{\Pr{}_h(\cdot|s,a,b)}{1+f_h^*(s,a,b,\cdot)}\right\|_{\text{TV}}^2\\ &= 2\left[\int_{s'\in\mathcal{S}}\frac{\Pr{}_h(s'|s,a,b)\cdot|f_h^*(s,a,b,s') - \widehat{f}_h^k(s,a,b,s')|}{[1+\widehat{f}_h^k(s,a,b,s')]\cdot[1+f_h^*(s,a,b,s')]}\mathrm{d}s'\right]^2,\end{aligned}$$

where $f^*(s,a,b,s') = \frac{\mathbb{P}(s'|s,a,b)}{\mathcal{P}_{\mathcal{S}}^-(s')}$ with $\mathcal{P}_{\mathcal{S}}^-(s') \geq C_{\mathcal{S}}^-,\ \forall s' \in \mathcal{S}$ and the second equation is due to $\big\| \Pr_h^{\widehat{f}^k}(y=0,\cdot|s,a,b) - \Pr_h^{f^*}(y=0,\cdot|s,a,b)\big\|_{\mathrm{TV}} = \big\| \Pr_h^{\widehat{f}^k}(y=1,\cdot|s,a,b) - \Pr_h^{f^*}(y=1,\cdot|s,a,b)\big\|_{\mathrm{TV}} = \Big\| \frac{\Pr_h(\cdot|s,a,b)}{1+\widehat{f}_h^k(s,a,b,\cdot)} - \frac{\Pr_h(\cdot|s,a,b)}{1+f_h^*(s,a,b,\cdot)} \Big\|_{\mathrm{TV}}$. Moreover, according to Lemma C.1 and (15), we have

$$
\begin{aligned}
&\frac{\Pr_h(s'|s,a,b)\cdot|f_h^*(s,a,b,s') - \widehat{f}_h^k(s,a,b,s')|}{[1+\widehat{f}_h^k(s,a,b,s')]\cdot[1+f_h^*(s,a,b,s')]} \\
&\quad= \frac{1/2\cdot[\mathbb{P}_h(s'|s,a,b)+\mathcal{P}_{\mathcal{S}}^-(s')]\cdot|\mathbb{P}_h(s'|s,a,b)/\mathcal{P}_{\mathcal{S}}^-(s') - \widehat{f}_h^k(s,a,b,s')|}{[1+\widehat{f}_h^k(s,a,b,s')]\cdot[1+\mathbb{P}_h(s'|s,a,b)/\mathcal{P}_{\mathcal{S}}^-(s')]} \\
&\quad= \frac{1/2\cdot|\mathbb{P}_h(s'|s,a,b) - \mathcal{P}_{\mathcal{S}}^-(s')\widehat{f}_h^k(s,a,b,s')|}{1+\widehat{f}_h^k(s,a,b,s')} \geq \frac{|\mathbb{P}_h(s'|s,a,b) - \mathcal{P}_{\mathcal{S}}^-(s')\widehat{f}_h^k(s,a,b,s')|}{4\sqrt{d}/C_{\mathcal{S}}^-},
\end{aligned}
$$

where the inequality is due to $[1+\widehat{f}_h^k(s,a,b,s')] \leq (1+\sqrt{d}) \leq 2\sqrt{d}/C_{\mathcal{S}}^-$ since $\widehat{f}_h^k(s,a,b,s') \leq \sqrt{d}/C_{\mathcal{S}}^-$ with $d \geq 1$ and $0 < C_{\mathcal{S}}^- \leq 1$. Thus, combining this inequality with (45), we further have, $\forall h \geq 2$,

$$
\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)} \Big\|\mathbb{P}_h(\cdot|s,a,b) - \mathcal{P}_{\mathcal{S}}^-(\cdot)\widetilde{\phi}_h^k(s,a,b)^\top\widetilde{\psi}_h^k(\cdot)\Big\|_{\mathrm{TV}}^2 \leq 8d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k. \tag{46}
$$

Similarly, we can obtain

$$
\begin{aligned}
&\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_1^k(\cdot,\cdot,\cdot)} \Big\|\mathbb{P}_h(\cdot|s,a,b) - \mathcal{P}_{\mathcal{S}}^-(\cdot)\widetilde{\phi}_h^k(s,a,b)^\top\widetilde{\psi}_h^k(\cdot)\Big\|_{\mathrm{TV}}^2 \leq 8d/(C_{\mathcal{S}}^-)^2\cdot\log(2k|\mathcal{F}|/\delta)/k, \\
&\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)} \Big\|\mathbb{P}_h(\cdot|s,a,b) - \mathcal{P}_{\mathcal{S}}^-(\cdot)\widetilde{\phi}_h^k(s,a,b)^\top\widetilde{\psi}_h^k(\cdot)\Big\|_{\mathrm{TV}}^2 \leq 8d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \geq 2.
\end{aligned} \tag{47}
$$

Now we define

$$
\widehat{g}_h^k(s,a,b,s') := \mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a,b)^\top\widetilde{\psi}_h^k(s').
$$

Since $\int_{s'\in\mathcal{S}}\widehat{g}_h^k(s,a,b,s')\mathrm{d}s'$ may not be guaranteed to be 1, to obtain an approximator of the transition model $\mathbb{P}_h$ lying on a probability simplex, we should further normalize $\widehat{g}_h^k(s,a,b,s')$. Thus, we define for all $(s,a,b,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{B}\times\mathcal{S}$,

$$
\widehat{\mathbb{P}}_h^k(s'|s,a,b) := \frac{\widehat{g}_h^k(s,a,b,s')}{\|\widehat{g}_h^k(s,a,b,\cdot)\|_1} = \frac{\widehat{g}_h^k(s,a,b,s')}{\int_{s'\in\mathcal{S}}\widehat{g}_h^k(s,a,b,s')\mathrm{d}s'} = \frac{\mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a,b)^\top\widetilde{\psi}_h^k(s')}{\int_{s'\in\mathcal{S}}\mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a,b)^\top\widetilde{\psi}_h^k(s')\mathrm{d}s'}.
$$

We further let

$$
\widehat{\phi}_h^k(s,a,b) := \widetilde{\phi}_h^k(s,a,b)\big/\textstyle\int_{s'\in\mathcal{S}}\mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\phi}_h^k(s,a,b)^\top\widetilde{\psi}_h^k(s')\mathrm{d}s', \qquad \widehat{\psi}_h^k(s') := \mathcal{P}_{\mathcal{S}}^-(s')\widetilde{\psi}_h^k(s'),
$$

such that

$$
\widehat{\mathbb{P}}_h^k(s'|s,a,b) = \widehat{\psi}_h^k(s')^\top\widehat{\phi}_h^k(s,a,b).
$$

Next, we give the upper bound of the approximation error $\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a,b) - \mathbb{P}_h(\cdot|s,a,b)\|_{\mathrm{TV}}^2$. We have

$$
\begin{aligned}
&\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a,b) - \mathbb{P}_h(\cdot|s,a,b)\|_{\mathrm{TV}}^2 \\
&\quad\leq 2\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a,b) - \widehat{g}_h^k(s,a,b,\cdot)\|_{\mathrm{TV}}^2 + 2\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{g}_h^k(s,a,b,\cdot) - \mathbb{P}_h(\cdot|s,a,b)\|_{\mathrm{TV}}^2 \\
&\quad\leq 2\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a,b) - \widehat{g}_h^k(s,a,b,\cdot)\|_{\mathrm{TV}}^2 + 16d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k,
\end{aligned} \tag{48}
$$

where the first inequality is by $(x+y)^2 \le 2x^2+2y^2$ and the last inequality is by (46). Moreover, we have

$$\begin{aligned}
&\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a,b)-\widehat{g}_h^k(s,a,b,\cdot)\|_{\mathrm{TV}}^2\\
&\quad=\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\left\|\frac{\widehat{g}_h^k(s,a,b,s')}{\|\widehat{g}_h^k(s,a,b,\cdot)\|_1}-\widehat{g}_h^k(s,a,b,\cdot)\right\|_{\mathrm{TV}}^2\\
&\quad=\frac{1}{4}\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\left(\|\widehat{g}_h^k(s,a,b,\cdot)\|_1-1\right)^2\\
&\quad\le\frac{1}{4}\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\left(\|\widehat{g}_h^k(s,a,b,\cdot)-\mathbb{P}_h(\cdot|s,a,b)\|_1+\|\mathbb{P}_h(\cdot|s,a,b)\|_1-1\right)^2\\
&\quad\le\frac{1}{4}\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{g}_h^k(s,a,b,\cdot)-\mathbb{P}_h(\cdot|s,a,b)\|_1^2\\
&\quad=\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{g}_h^k(s,a,b,\cdot)-\mathbb{P}_h(\cdot|s,a,b)\|_{\mathrm{TV}}^2\le 8d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k.
\end{aligned}$$

Combining the above inequality with (22), we eventually obtain

$$\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a,b)-\mathbb{P}_h(\cdot|s,a,b)\|_{\mathrm{TV}}^2\le 32d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k,\quad\forall h\ge 1.$$

Thus, we similarly have

$$\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\mathbb{P}}_h^k(\cdot|s,a,b)-\mathbb{P}_h(\cdot|s,a,b)\|_{\mathrm{TV}}^2\le 32d/(C_{\mathcal{S}}^-)^2\cdot\log(2kH|\mathcal{F}|/\delta)/k,\quad\forall h\ge 2.$$

The above three inequalities hold with probability at least $1-2\delta$. This completes the proof. □

### D.3. Proof of Theorem 4.2

*Proof.* We define two auxiliary MGs respectively by reward function $r+\beta^k$ and transition model $\widehat{\mathbb{P}}^k$, and $r-\beta^k$, $\widehat{\mathbb{P}}^k$. Then for any joint policy $\sigma$, let $\overline{V}_{k,h}^{\sigma}$ and $\underline{V}_{k,h}^{\sigma}$ be the associated value functions on the two auxiliary MGs respectively. We first decompose the instantaneous regret term $V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)-V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)$ as follows

$$\begin{aligned}
&V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)-V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)\\
&\quad=\underbrace{V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)-\overline{V}_{k,1}^{\mathrm{br}(\nu^k),\nu^k}(s_1)}_{(i)}+\underbrace{\overline{V}_{k,1}^{\mathrm{br}(\nu^k),\nu^k}(s_1)-\overline{V}_1^k(s_1)}_{(ii)}+\underbrace{\overline{V}_1^k(s_1)-\underline{V}_1^k(s_1)}_{(iii)}\\
&\quad\quad+\underbrace{\underline{V}_1^k(s_1)-\underline{V}_{k,1}^{\pi^k,\mathrm{br}(\pi^k)}(s_1)}_{(iv)}+\underbrace{\underline{V}_{k,1}^{\pi^k,\mathrm{br}(\pi^k)}(s_1)-V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)}_{(v)}.
\end{aligned}\tag{49}$$

Terms $(ii)$ and $(iv)$ depict the planning error on two auxiliary Markov games. According to Lemma D.6, we have

$$\begin{aligned}
\overline{V}_{k,1}^{\mathrm{br}(\nu^k),\nu^k}(s_1)&\le\overline{V}_1^k(s_1)+H\iota_k,\\
\underline{V}_{k,1}^{\pi^k,\mathrm{br}(\pi^k)}(s_1)&\ge\underline{V}_1^k(s_1)-H\iota_k,
\end{aligned}$$

where $\iota_k$ is the learning accuracy of CCE. Thus, together with (49), we have

$$\begin{aligned}
&V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)-V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)\\
&\quad=\underbrace{V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)-\overline{V}_{k,1}^{\mathrm{br}(\nu^k),\nu^k}(s_1)}_{(i)}+\underbrace{\overline{V}_1^k(s_1)-\underline{V}_1^k(s_1)}_{(iii)}+\underbrace{\underline{V}_{k,1}^{\pi^k,\mathrm{br}(\pi^k)}(s_1)-V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)}_{(v)}+2H\iota_k.
\end{aligned}\tag{50}$$

Thus, to bound the term $V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)-V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)$, we only need to bound the terms $(i)$, $(iii)$, and $(v)$ as in (50).

To bound term $(i)$, by Lemma D.2, we have

$$\begin{aligned}(i) &= V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - \overline{V}_{k,1}^{\mathrm{br}(\nu^k),\nu^k}(s_1) \\ &= \mathbb{E}\left[\sum_{h=1}^{H}\left(-\beta_h^k(s_h,a_h,b_h) + (\mathbb{P}_h - \widehat{\mathbb{P}}_h^k)V_{h+1}^{\mathrm{br}(\nu^k),\nu^k}(s_h,a_h,b_h)\right)\,\middle|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k\right] \\ &\le \mathbb{E}\left[\sum_{h=1}^{H}\left(-\beta_h^k(s_h,a_h,b_h) + H\|\mathbb{P}_h(\cdot|s_h,a_h,b_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1\right)\,\middle|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k\right],\end{aligned} \tag{51}$$

where the first inequality is by the fact $\sup_{s\in\mathcal{S}}\left|V_{h+1}^{\mathrm{br}(\nu^k),\nu^k}(s)\right| \le H$. Next, we bound $\mathbb{E}\big[\sum_{h=1}^{H}\|\mathbb{P}_h(\cdot|s_h,a_h,b_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1\,\big|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k\big]$. Note that for $\|\mathbb{P}_h(\cdot|s_h,a_h,b_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1$, we have a trivial bound $\|\mathbb{P}_h(\cdot|s_h,a_h,b_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1 \le 2$. Furthermore, by Lemma D.4, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{h=1}^{H}\|\mathbb{P}_h(\cdot|s_h,a_h,b_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1\,\middle|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k\right] \\ &= \sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h,b_h)\sim d_h^{\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}[\|\mathbb{P}_h(\cdot|s_h,a_h,b_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1] \\ &= \sum_{h=2}^{H}\sqrt{8k\zeta_{h-1}^k + 2k|\mathcal{A}||\mathcal{B}|\mathbb{E}_{(s,a,b)\sim\breve{\rho}_h^k(\cdot,\cdot,\cdot)}[\|\mathbb{P}_h(\cdot|s,a,b) - \widehat{\mathbb{P}}_h^k(\cdot|s,a,b)\|_1^2] + 4\lambda_k d/(C_{\mathcal{S}}^-)^2}\cdot\mathbb{E}_{d_{h-1}^{\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k}}\left\|\widehat{\phi}_{h-1}^k\right\|_{\Sigma^{-1}_{\tilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}} \\ &\quad + \sqrt{|\mathcal{A}||\mathcal{B}|\mathbb{E}_{(a,b)\sim\widetilde{\rho}_1^k(s_1,\cdot,\cdot)}[\|\mathbb{P}_1(\cdot|s_1,a,b) - \widehat{\mathbb{P}}_1^k(\cdot|s_1,a,b)\|_1^2]} \\ &= \sum_{h=2}^{H}\sqrt{8k\zeta_{h-1}^k + 2k|\mathcal{A}||\mathcal{B}|\xi_h^k + 4\lambda_k d/(C_{\mathcal{S}}^-)^2}\cdot\mathbb{E}_{(s,a,b)\sim d_{h-1}^{\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}\left\|\widehat{\phi}_{h-1}^k(s,a,b)\right\|_{\Sigma^{-1}_{\tilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}} + \sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k},\end{aligned}$$

where the last equation is by the below definitions for all $(h,k)\in[H]\times[K]$,

$$\begin{aligned}\zeta_h^k &:= \mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}[\|\mathbb{P}_1(\cdot|s,a,b) - \widehat{\mathbb{P}}_1^k(\cdot|s,a,b)\|_1^2], \\ \xi_h^k &:= \mathbb{E}_{(s,a,b)\sim\breve{\rho}_h^k(\cdot,\cdot,\cdot)}[\|\mathbb{P}_h(\cdot|s,a,b) - \widehat{\mathbb{P}}_h^k(\cdot|s,a,b)\|_1^2],\end{aligned} \tag{52}$$

whose upper bound will be characterized later. Thus, the above results imply that

$$\begin{aligned}&\mathbb{E}\left[\sum_{h=1}^{H}H\|\mathbb{P}_h(\cdot|s_h,a_h,b_h) - \widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1\,\middle|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k\right] \\ &\le \min\left\{H\sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k} + \sum_{h=2}^{H}H\sqrt{8k\zeta_{h-1}^k + 2k|\mathcal{A}||\mathcal{B}|\xi_h^k + 4\lambda_k d/(C_{\mathcal{S}}^-)^2}\cdot\mathbb{E}_{d_{h-1}^{\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k}}\left\|\widehat{\phi}_{h-1}^k\right\|_{\Sigma^{-1}_{\tilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}},\ 2H^2\right\}.\end{aligned}$$

On the other hand, we bound $\mathbb{E}[\sum_{h=1}^H -\beta_h^k(s_h,a_h,b_h)\,|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k]$ in (51). We have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^H -\beta_h^k(s_h,a_h,b_h)\,\middle|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k\right]\\
&\quad=\mathbb{E}\left[\sum_{h=1}^H -\min\{\gamma_k\|\widehat{\phi}_h^k(s_h,a_h,b_h)\|_{(\widehat{\Sigma}_h^k)^{-1}},2H\}\,\middle|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k\right]\\
&\quad\le\mathbb{E}\left[\sum_{h=1}^H -\min\left\{\frac{3}{5}\gamma_k\|\widehat{\phi}_h^k(s_h,a_h,b_h)\|_{\Sigma^{-1}_{\widetilde{\rho}_h^k,\widehat{\phi}_h^k}},2H\right\}\,\middle|\,\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k\right]\\
&\quad=-\min\left\{\frac{3}{5}\gamma_k\sum_{h=1}^H\mathbb{E}_{(s,a,b)\sim d_h^{\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}\|\widehat{\phi}_h^k(s,a,b)\|_{\Sigma^{-1}_{\widetilde{\rho}_h^k,\widehat{\phi}_h^k}},2H^2\right\}\\
&\quad\le-\min\left\{\frac{3}{5}\gamma_k\sum_{h=1}^{H-1}\mathbb{E}_{(s,a,b)\sim d_h^{\mathrm{br}(\nu^k),\nu^k,\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}\|\widehat{\phi}_h^k(s,a,b)\|_{\Sigma^{-1}_{\widetilde{\rho}_h^k,\widehat{\phi}_h^k}},2H^2\right\},
\end{aligned}\tag{53}$$

when $\lambda_k\ge c_0 d\log(H|\Phi|k/\delta)$ with probability at least $1-\delta$. The first inequality is by Lemma E.1 for all $h\in[H]$. Thus, plugging in the above results into (51), for a sufficient large $c_0$, setting

$$\lambda_k\ge c_0 d\log(H|\Phi|k/\delta),\qquad \gamma_k\ge\frac{5}{3}H\sqrt{8k\zeta_{h-1}^k+2k|\mathcal{A}||\mathcal{B}|\xi_h^k+4\lambda_k d/(C_{\mathcal{S}}^-)^2},\tag{54}$$

we have that

$$(i)=V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)-\overline{V}_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)\le\sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k},\tag{55}$$

where the inequality is due to $\min\{x+y,2H^2\}-\min\{y,2H^2\}\le x,\forall x,y\ge 0$.

On the other hand, we prove the upper bound for term $(v)$. Specifically, by Lemma D.2, we have

$$\begin{aligned}
(v)&=\underline{V}_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)-V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)\\
&=\mathbb{E}\left[\sum_{h=1}^H\left(-\beta_h^k(s_h,a_h,b_h)-(\mathbb{P}_h-\widehat{\mathbb{P}}_h^k)V_{h+1}^{\pi^k,\mathrm{br}(\pi^k)}(s_h,a_h,b_h)\right)\,\middle|\,\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k\right]\\
&\le\mathbb{E}\left[\sum_{h=1}^H\left(-\beta_h^k(s_h,a_h,b_h)+H\|\mathbb{P}_h(\cdot|s_h,a_h,b_h)-\widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1\right)\,\middle|\,\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k\right],
\end{aligned}\tag{56}$$

where the first inequality is by the fact $\sup_{s\in\mathcal{S}}\left|V_{h+1}^{\pi^k,\mathrm{br}(\pi^k)}(s)\right|\le H$. Next, for $\|\mathbb{P}_h(\cdot|s_h,a_h,b_h)-\widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1$, we have a trivial bound $\|\mathbb{P}_h(\cdot|s_h,a_h,b_h)-\widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1\le 2$. In addition, by Lemma D.4, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^H\|\mathbb{P}_h(\cdot|s_h,a_h,b_h)-\widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1\,\middle|\,\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k\right]\\
&=\sum_{h=1}^H\mathbb{E}_{(s_h,a_h,b_h)\sim d_h^{\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}[\|\mathbb{P}_h(\cdot|s_h,a_h,b_h)-\widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1]\\
&=\sum_{h=2}^H\sqrt{8k\zeta_{h-1}^k+2k|\mathcal{A}||\mathcal{B}|\xi_h^k+4\lambda_k d}\cdot\mathbb{E}_{(s,a,b)\sim d_{h-1}^{\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}\left\|\widehat{\phi}_{h-1}^k(s,a,b)\right\|_{\Sigma^{-1}_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}}+\sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k},
\end{aligned}$$

where the last equation is by the definitions of $\zeta_h^k$ and $\xi_h^k$ in (52). Thus, the above results imply that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{h=1}^H H\|\mathbb{P}_h(\cdot|s_h,a_h,b_h)-\widehat{\mathbb{P}}_h^k(\cdot|s_h,a_h,b_h)\|_1\,\middle|\,\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k\right]\\
&\le\min\left\{H\sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k}+\sum_{h=2}^H H\sqrt{8k\zeta_{h-1}^k+2k|\mathcal{A}||\mathcal{B}|\xi_h^k+4\lambda_k d}\cdot\mathbb{E}_{d_{h-1}^{\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k}}\left\|\widehat{\phi}_{h-1}^k\right\|_{\Sigma^{-1}_{\widetilde{\rho}_{h-1}^k,\widehat{\phi}_{h-1}^k}},\ 2H^2\right\}.
\end{aligned}$$

On the other hand, we bound $\mathbb{E}[\sum_{h=1}^H -\beta_h^k(s_h,a_h,b_h)\,|\,\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k]$ in (56). Analogous to (53), we can obtain

$$\mathbb{E}\left[\sum_{h=1}^H -\beta_h^k(s_h,a_h,b_h)\,\middle|\,\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k\right] \le -\min\left\{\frac{3}{5}\gamma_k\sum_{h=1}^{H-1}\mathbb{E}_{(s,a,b)\sim d_h^{\pi^k,\mathrm{br}(\pi^k),\widehat{\mathbb{P}}^k}(\cdot,\cdot,\cdot)}\|\widehat{\phi}_h^k(s,a,b)\|_{\Sigma^{-1}_{\widetilde{\rho}_h^k,\widehat{\phi}_h^k}},2H^2\right\},$$

when $\lambda_k \ge c_0 d\log(H|\Phi|k/\delta)$ with probability at least $1-\delta$. Thus, plugging in the above results into (56), setting $\lambda_k$ and $\gamma_k$ as in (54), we have

$$(v) = \underline{V}_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1) \le \sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k}, \tag{57}$$

where the inequality is due to $\min\{x+y,2H^2\}-\min\{y,2H^2\}\le x, \forall x,y\ge 0$.

Now we have proved the near-optimism and near-pessimism in (55) and (57) respectively, which extends the related result for single-agent MDPs.

Next, we show the upper bound of the term $(iii)$ in (49). By Lemma D.3, we have

$$\begin{aligned}(iii) = \overline{V}_1^k(s_1) - \underline{V}_1^k(s_1) &= \mathbb{E}\left[\sum_{h=1}^H 2\beta_h^k(s_h,a_h,b_h) + (\widehat{\mathbb{P}}_h^k-\mathbb{P}_h)(\overline{V}_{h+1}^k-\underline{V}_{h+1}^k)(s_h,a_h,b_h)\,\middle|\,\sigma^k,\mathbb{P}\right]\\ &\le 2\sum_{h=1}^H \mathbb{E}_{(s,a,b)\sim d_h^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)}[\beta_h^k(s,a,b)] + 6H^2\sum_{h=1}^H \mathbb{E}_{(s,a,b)\sim d_h^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)}[\|\mathbb{P}_h^k(\cdot|s,a,b)-\mathbb{P}_h(\cdot|s,a,b)\|_1]\end{aligned} \tag{58}$$

where the above inequality is due to $\sup_{s\in\mathcal{S}}|\overline{V}_h^k(s)|\le H(1+2H)\le 3H^2$ and $\sup_{s\in\mathcal{S}}|\underline{V}_h^k(s)|\le H(1+2H)\le 3H^2$. By Lemma D.5, since $\sup_{s\in\mathcal{S},a\in\mathcal{A},b\in\mathcal{B}}\beta_h^k(s,a,b)\le 2H$ according to the definition of $\beta_h^k$ in Algorithm 2, we have

$$\begin{aligned}&\sum_{h=1}^H \mathbb{E}_{(s,a,b)\sim d_h^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)}[\beta_h^k(s,a,b)]\\ &\le \sqrt{|\mathcal{A}||\mathcal{B}|\mathbb{E}_{(a,b)\sim\widetilde{\rho}_1^k(s_1,\cdot,\cdot)}[\beta_1^k(s_1,a,b)^2]} + \sum_{h=2}^H\sqrt{k|\mathcal{A}||\mathcal{B}|\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}[\beta_h^k(s,a,b)^2]+4H^2\lambda_k d}\,\mathbb{E}_{d_{h-1}^{\sigma^k,\mathbb{P}}}\left\|\phi_{h-1}^*\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}\\ &\le \sqrt{|\mathcal{A}||\mathcal{B}|\gamma_k^2\mathbb{E}_{(a,b)\sim\widetilde{\rho}_1^k(s_1,\cdot,\cdot)}\|\widehat{\phi}_1^k(s_1,a,b)\|^2_{(\widehat{\Sigma}_1^k)^{-1}}}\\ &\quad+\sum_{h=2}^H\sqrt{k|\mathcal{A}||\mathcal{B}|\gamma_k^2\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\phi}_h^k(s,a,b)\|^2_{(\widehat{\Sigma}_h^k)^{-1}}+4H^2\lambda_k d}\,\mathbb{E}_{d_{h-1}^{\sigma^k,\mathbb{P}}}\left\|\phi_{h-1}^*\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}},\end{aligned}$$

where the second inequality is due to $\beta_h^k(s,a,b)\le\|\widehat{\phi}_h^k(s,a,b)\|_{(\widehat{\Sigma}_h^k)^{-1}}$. Furthermore, we have that with $\lambda_k\ge c_0 d\log(H|\Phi|k/\delta)$, with probability at least $1-\delta$, for all $h\in[H]$,

$$\begin{aligned}\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\phi}_h^k(s,a,b)\|^2_{(\widehat{\Sigma}_h^k)^{-1}} &\le 3\mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)}\|\widehat{\phi}_h^k(s,a,b)\|^2_{\Sigma^{-1}_{\widetilde{\rho}_h^k,\widehat{\phi}_h^k}}\\ &= 3\mathbb{E}_{\widetilde{\rho}_h^k}\left[\widehat{\phi}_h^{k\top}\left(k\mathbb{E}_{\widetilde{\rho}_h^k}[\widehat{\phi}_h^k(\widehat{\phi}_h^k)^\top]+\lambda_k I\right)^{-1}\widehat{\phi}_h^k\right]\\ &= \frac{3}{k}\,\mathrm{tr}\left\{k\mathbb{E}_{\widetilde{\rho}_h^k}[\widehat{\phi}_h^k\widehat{\phi}_h^{k\top}]\left(k\mathbb{E}_{\widetilde{\rho}_h^k}[\widehat{\phi}_h^k(\widehat{\phi}_h^k)^\top]+\lambda_k I\right)^{-1}\right\} \le \frac{3}{k}\,\mathrm{tr}(I) = \frac{3d}{k},\end{aligned}$$

where the first inequality is by Lemma E.1. Combining the above results, we have the following inequality holds with probability at least $1-\delta$,

$$\begin{aligned}&\sum_{h=1}^H \mathbb{E}_{(s,a,b)\sim d_h^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)}[\beta_h^k(s,a,b)]\\ &\le \sqrt{3d|\mathcal{A}||\mathcal{B}|\gamma_k^2/k}+\sum_{h=2}^H\sqrt{3d|\mathcal{A}||\mathcal{B}|\gamma_k^2+4H^2\lambda_k d}\,\mathbb{E}_{(s,a,b)\sim d_{h-1}^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)}\left\|\phi_{h-1}^*(s,a,b)\right\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}.\end{aligned} \tag{59}$$

Further by Lemma D.5, due to $\|\mathbb{P}_h(\cdot|s,a,b) - \widehat{\mathbb{P}}_h^k(\cdot|s,a,b)\|_1 \le 2$, we have

$$\begin{aligned}
&\sum_{h=1}^{H} \mathbb{E}_{(s,a,b)\sim d_h^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)} [\|\mathbb{P}_h(\cdot|s,a,b) - \widehat{\mathbb{P}}_h^k(\cdot|s,a,b)\|_1] \\
&\quad \le \sqrt{|\mathcal{A}||\mathcal{B}| \mathbb{E}_{a\sim\widetilde{\rho}_1^k(s_1,\cdot)} [\|\mathbb{P}_h(\cdot|s,a,b) - \widehat{\mathbb{P}}_h^k(\cdot|s,a,b)\|_1^2]} \\
&\qquad + \sum_{h=2}^{H} \sqrt{k|\mathcal{A}||\mathcal{B}| \mathbb{E}_{(s,a,b)\sim\widetilde{\rho}_h^k(\cdot,\cdot,\cdot)} [\|\mathbb{P}_h(\cdot|s,a,b) - \widehat{\mathbb{P}}_h^k(\cdot|s,a,b)\|_1^2] + 4\lambda_k d}\, \mathbb{E}_{(s,a,b)\sim d_{h-1}^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)} \big\|\phi_{h-1}^*(s,a,b)\big\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}} \\
&\quad = \sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k} + \sum_{h=2}^{H} \sqrt{k|\mathcal{A}||\mathcal{B}|\zeta_h^k + 4\lambda_k d}\, \mathbb{E}_{(s,a,b)\sim d_{h-1}^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)} \big\|\phi_{h-1}^*(s,a,b)\big\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}. \qquad (60)
\end{aligned}$$

Therefore, combining (58), (59), and (60), we obtain

$$\begin{aligned}
(iii) &\le \left(2\sqrt{3d|\mathcal{A}||\mathcal{B}|\gamma_k^2/k} + 6H^2\sqrt{|\mathcal{A}||\mathcal{B}|\zeta_1^k}\right) \qquad (61) \\
&\quad + \sum_{h=2}^{H} \left(2\sqrt{3d|\mathcal{A}||\mathcal{B}|\gamma_k^2 + 4H^2\lambda_k d} + 6H^2\sqrt{k|\mathcal{A}||\mathcal{B}|\zeta_h^k + 4\lambda_k d}\right) \mathbb{E}_{(s,a,b)\sim d_{h-1}^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)} \big\|\phi_{h-1}^*(s,a,b)\big\|_{\Sigma^{-1}_{\rho_{h-1}^k,\phi_{h-1}^*}}.
\end{aligned}$$

We characterize the upper bound of $\zeta_h^k$ and $\xi_h^k$ as defined in (52). According to Lemma 5.2, we have with probability at least $1-2\delta$,

$$\begin{aligned}
\zeta_h^k &\le 32d/(C_{\mathcal{S}}^-)^2 \cdot \log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \ge 1, \\
\xi_h^k &\le 32d/(C_{\mathcal{S}}^-)^2 \cdot \log(2kH|\mathcal{F}|/\delta)/k, \quad \forall h \ge 2,
\end{aligned} \qquad (62)$$

Plugging (62) and (54) into (55),(57), and (61), we obtain

$$\begin{aligned}
&(i) = V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - \overline{V}_{k,1}^{\mathrm{br}(\nu^k),\nu^k}(s_1) \lesssim \sqrt{d|\mathcal{A}||\mathcal{B}|/(C_{\mathcal{S}}^-)^2 \cdot \log(KH|\mathcal{F}|/\delta)/k}, \\
&(v) = \underline{V}_{k,1}^{\pi^k,\mathrm{br}(\pi^k)}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1) \lesssim \sqrt{d|\mathcal{A}||\mathcal{B}|/(C_{\mathcal{S}}^-)^2 \cdot \log(KH|\mathcal{F}|/\delta)/k}, \\
&(iii) = \overline{V}_1^k(s_1) - \underline{V}_1^k(s_1) \lesssim \sqrt{C_1 \log(H|\mathcal{F}|K/\delta)/k} + \sqrt{(C_1+C_2)\log(H|\mathcal{F}|K/\delta)} \sum_{h=1}^{H-1} \mathbb{E}_{d_h^{\sigma^k,\mathbb{P}}} \|\phi_h^*\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}},
\end{aligned}$$

where we let $C_1 = H^2d^3|\mathcal{A}||\mathcal{B}|/(C_{\mathcal{S}}^-)^2 + H^2d^2|\mathcal{A}|^2|\mathcal{B}|^2/(C_{\mathcal{S}}^-)^2 + H^4 d|\mathcal{A}||\mathcal{B}|/(C_{\mathcal{S}}^-)^2$ and $C_2 = H^4d^2$. Further by (50), we have

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K} \left[V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1) - V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)\right] &\lesssim \sqrt{(C_1+C_2)\log(H|\mathcal{F}|K/\delta)}/K \cdot \sum_{h=1}^{H-1}\sum_{k=1}^{K} \mathbb{E}_{d_h^{\sigma^k,\mathbb{P}}} \|\phi_h^*\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}} \\
&\quad + \sqrt{C_1 \log(H|\mathcal{F}|K/\delta)/K} + \frac{H}{K}\sum_{k=1}^{K} \iota_k.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{(s,a,b)\sim d_h^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)} \|\phi_h^*(s,a,b)\|_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}} \\
&\quad \le \sqrt{\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{(s,a,b)\sim d_h^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)} \|\phi_h^*(s,a,b)\|^2_{\Sigma^{-1}_{\rho_h^k,\phi_h^*}}} \\
&\quad = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \mathrm{tr}\left(\mathbb{E}_{(s,a,b)\sim d_h^{\sigma^k,\mathbb{P}}(\cdot,\cdot,\cdot)} (\phi_h^*(s,a,b)\phi_h^*(s,a,b)^\top)\, \Sigma^{-1}_{\rho_h^k,\phi_h^*}\right)} \\
&\quad \le \sqrt{d\log(1+kd/\lambda_k)/K} \le \sqrt{d\log(1+c_1K)/K}.
\end{aligned}$$

where the first inequality is by Jensen's inequality and the second inequality is by Lemma E.3 with $c_1$ being some absolute constant. Thus, we have

$$\frac{1}{K}\sum_{k=1}^{K}\Big[V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)-V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)\Big]\lesssim\sqrt{H^2/K}+\sqrt{H^2d(C_1+C_2)\log(H|\mathcal{F}|K/\delta)\log(1+c_1K)/K}.$$

Taking union bound for all events in this proof, due to $|\mathcal{F}|\geq|\Phi|$, letting

$$\lambda_k=c_0d\log(H|\mathcal{F}|k/\delta),\quad \gamma_k=4H\big(12\sqrt{|\mathcal{A}||\mathcal{B}|d}+\sqrt{c_0}d\big)/C_{\mathcal{S}}^{-}\cdot\sqrt{\log(2Hk|\mathcal{F}|/\delta)},\quad \iota_k\leq\mathcal{O}(\sqrt{1/k}),$$

we have with probability at least $1-3\delta$,

$$\frac{1}{K}\sum_{k=1}^{K}\Big[V_1^{\mathrm{br}(\nu^k),\nu^k}(s_1)-V_1^{\pi^k,\mathrm{br}(\pi^k)}(s_1)\Big]\lesssim\sqrt{C\log(H|\mathcal{F}|K/\delta)\log(c_0'K)/K},$$

where $C=H^4d^4|\mathcal{A}||\mathcal{B}|/(C_{\mathcal{S}}^{-})^2+H^4d^3|\mathcal{A}|^2|\mathcal{B}|^2/(C_{\mathcal{S}}^{-})^2+H^6d^2|\mathcal{A}||\mathcal{B}|/(C_{\mathcal{S}}^{-})^2+H^6d^3$ and $c_0,c_0'$ are absolute constants. This completes the proof. □

## E. Other Supporting Lemmas

**Lemma E.1** (Concentration of Inverse Covariances (Zanette et al., 2021)). *Let $\mu_i$ be the conditional distribution of $\phi$ given the sampled $\phi_1,\cdots,\phi_{i-1}$ with $\|\phi_i\|_2\leq 1$ holding for $\phi_i$ as the realization of $\phi$. Let $\Lambda=\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{\phi\sim\mu_i}[\phi\phi^\top]$. Then there exists an absolute constant $c_0>0$. If $\lambda\geq c_0d\log(|\Phi|k/\delta)$, we have with probability at least $1-\delta$, for all $k\geq 1$,*

$$\frac{3}{5}(k\Lambda+\lambda I)^{-1}\preceq\left(\sum_{i=1}^{k}\phi_i\phi_i^\top+\lambda I\right)^{-1}\preceq 3(k\Lambda+\lambda I)^{-1}.$$

*Proof.* The proof of this lemma is adapted from Lemma 39 in Zanette et al. (2021). Further applying Lemma 39 of Zanette et al. (2021) to all the elements in the function class $\Phi$, we obtain Lemma E.1. This completes the proof. □

**Lemma E.2** (Agarwal et al. (2020)). *Let $\mathcal{F}$ be a function class with $|\mathcal{F}|<\infty$ and $f^*\in\mathcal{F}$ where*

$$f^*(x,z)=P^*(z|x)$$

*is some conditional distribution. Given a dataset $\mathcal{D}:=\{(x_i,z_i)\}_{i=0}^{k-1}$, let $\mathcal{T}_i$ be some distribution that is dependent on $\{(x_{i'},z_{i'})\}_{i'=0}^{i-1}$ for all $i\leq k$. Suppose $x_i\sim\mathcal{T}_i$ and $z_i\sim P^*(\cdot|x)=f^*(x,\cdot)$ for all $i\leq k$. Then, we have with probability at least $1-\delta$,*

$$\sum_{i=0}^{k-1}\mathbb{E}_{x\sim\mathcal{T}_i}\|\widehat{f}(x,\cdot)-f^*(x,\cdot)\|_{TV}^2\leq 2\log(k|\mathcal{F}|/\delta),$$

*where*

$$\widehat{f}:=\operatorname*{argmax}_{f\in\mathcal{F}}\sum_{(x,z)\in\mathcal{D}}\log f(x,z).$$

**Lemma E.3** (Uehara et al. (2021); Jin et al. (2020)). *For $i=1,\ldots,k$, $\Sigma_i:=\Sigma_{i-1}+G_i$ where $\Sigma_0=\lambda I$ with $\lambda>0$ and $G_i\in\mathbb{R}^{d\times d}$ is a positive semidefinite matrix with eigenvalues upper bounded by 1 and $\mathrm{tr}(G_i)\leq C^2$ for some $C>0$. Then, we have the following inequality*

$$\sum_{i=1}^{k}\mathrm{tr}(G_i\Sigma_{i-1}^{-1})\leq 2\log\det(\Sigma_k)-2\log\det(\lambda I)\leq d\log(1+kC^2d/\lambda).$$

## F. Additional Experimental Results

In this section, we present the additional experimental results. In Table 2, we report the human normalized scores for all the algorithms under all the tasks of Atari 100K. In Figure 2, we follow Agarwal et al. (2021) and report the stratified bootstrap of experiments, which consists of the 95% confidence intervals (CIs) of median, interquartile mean (IQM), mean, and optimality gap, over the 26 Atari 100K tasks. Here IQM is the 25% trimmed mean obtained by discarding the top and bottom 25% score and calculating the mean. See Agarwal et al. (2021) for details. According to Figure 2, our proposed SPR-UCB performs similarly to SPR on average, without the top 5% scores. Nevertheless, we remark that SPR-UCB outperforms SPR significantly on some hard exploration tasks (Taiga et al., 2020), including *PrivateEye*, *Frostbite*, and *Freeway*, as shown in Table 2.

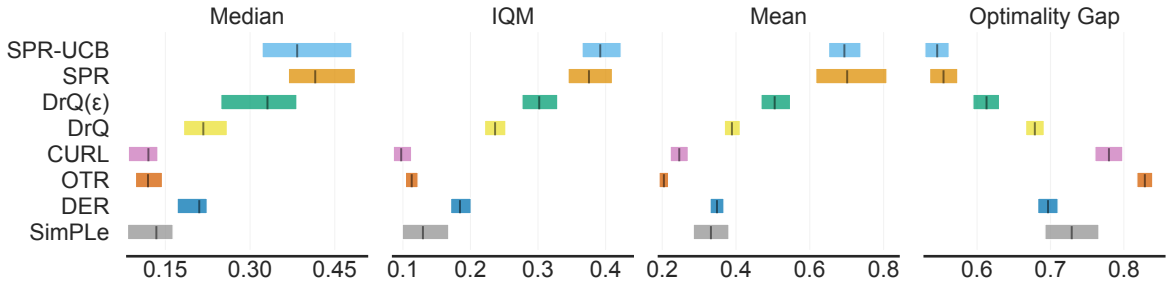


*Figure 2.* Stratified Bootstrap (Agarwal et al., 2021) of experiments, with 95% confidence intervals (CIs) based on 26 Atari 100K tasks. Higher mean, median, interquartile mean (IQM), and lower optimality gap a better. See Agarwal et al. (2021) for details. The results for baseline algorithms are collected from the report by Agarwal et al. (2021). The results for SPR-UCB are based on 10 runs per game.

*Table 2.* Table of the comparison of human normalized scores over tasks of Atari 100K. The scores of baselines are adopted from Agarwal et al. (2021), which runs each method over 100 seeds. We follow Agarwal et al. (2021) and evaluate the scores of SPR-UCB by evaluating the final policy obtained by SPR-UCB over 100 episodes. Highlighted scores are the highest and second highest among all algorithms.

| | CURL | OTR | DER | SIMPLE | DRQ | DRQ($\epsilon$) | SPR | **SPR-UCB** |
|---|---|---|---|---|---|---|---|---|
| ALIEN | 0.0700 | 0.0497 | 0.0833 | 0.0564 | 0.0734 | **0.0924** | 0.0890±0.03 | **0.0997**±0.02 |
| AMIDAR | 0.0630 | 0.0420 | 0.0701 | 0.0399 | 0.0516 | 0.0770 | **0.1015**±0.02 | **0.0973**±0.02 |
| ASSAULT | 0.5360 | 0.2088 | 0.6525 | 0.5866 | 0.4949 | **0.6875** | 0.6605±0.11 | **0.6729**±0.07 |
| ASTERIX | 0.0431 | 0.0150 | 0.0392 | **0.1107** | 0.0393 | 0.0668 | 0.0907±0.02 | **0.0965**±0.01 |
| BANKHEIST | 0.0692 | 0.0552 | 0.2318 | 0.0271 | 0.1884 | 0.2960 | **0.4483**±0.29 | **0.3011**±0.36 |
| BATTLEZONE | 0.1906 | 0.0798 | 0.1900 | 0.0480 | 0.2355 | 0.2241 | **0.3582**±0.14 | **0.3663**±0.09 |
| BOXING | 0.0708 | 0.1284 | -0.0340 | 0.6375 | 0.5443 | 0.7452 | **2.9667**±1.19 | **3.4332**±0.94 |
| BREAKOUT | 0.0297 | 0.2216 | 0.2609 | 0.5099 | 0.4759 | **0.6272** | 0.6208±0.46 | **0.7245**±0.47 |
| CHOPPERCOMMAND | -0.0042 | 0.0003 | 0.0175 | **0.0256** | -0.0028 | 0.0051 | **0.0206**±0.04 | 0.0041±0.04 |
| CRAZYCLIMBER | -0.0649 | 0.1684 | 0.9473 | **2.0681** | 0.4476 | 0.4295 | 1.0348±0.48 | **1.2936**±0.62 |
| DEMONATTACK | 0.2718 | 0.2911 | 0.2614 | 0.0308 | **0.5445** | **0.6429** | 0.2010±0.07 | 0.2214±0.10 |
| FREEWAY | **0.9550** | 0.3877 | 0.7046 | 0.5637 | 0.6006 | 0.6843 | 0.6512±0.47 | **0.9592**±0.11 |
| FROSTBITE | **0.2720** | 0.0374 | 0.1887 | 0.0402 | 0.1037 | 0.2223 | 0.2589±0.26 | **0.5591**±0.15 |
| GOPHER | 0.0665 | 0.1308 | 0.0972 | 0.1574 | 0.1673 | **0.1689** | **0.1870**±0.11 | 0.1666±0.05 |
| HERO | 0.1329 | 0.1654 | **0.1745** | 0.0547 | 0.0905 | 0.1054 | 0.1621±0.07 | **0.2096**±0.09 |
| JAMESBOND | 1.1032 | 0.2156 | 0.9009 | 0.2610 | 0.8136 | 1.1691 | **1.2326**±0.23 | **1.2124**±0.20 |
| KANGAROO | 0.2307 | 0.0994 | 0.1776 | -0.0003 | 0.3092 | 0.3474 | **1.1952**±1.08 | **1.0553**±0.96 |
| KRULL | 1.3595 | 1.9278 | 1.5540 | 0.5685 | 2.3732 | **2.6268** | 1.9519±0.43 | **2.4225**±0.23 |
| KUNGFUMASTER | 0.3513 | 0.2848 | 0.2812 | **0.6497** | 0.3068 | 0.4987 | 0.6462±0.32 | **0.8126**±0.27 |
| MSPACMAN | 0.1139 | 0.0904 | 0.1325 | **0.1765** | 0.1047 | 0.1371 | 0.1522±0.05 | **0.1557**±0.06 |
| PONG | 0.0627 | **0.5168** | 0.3113 | **0.9495** | 0.1827 | 0.3298 | 0.4331±0.30 | 0.4007±0.23 |
| PRIVATEEYE | 0.0008 | 0.0005 | 0.0007 | 0.0001 | 0.0000 | -0.0003 | **0.0009**±0.00 | **0.0011**±0.00 |
| QBERT | 0.0424 | 0.0292 | **0.1211** | 0.0846 | 0.0580 | **0.1239** | 0.0528±0.04 | 0.0606±0.04 |
| ROADRUNNER | 0.6376 | 0.3313 | 1.5104 | 0.7186 | 1.1123 | 1.4297 | **1.5576**±0.64 | **2.0051**±0.53 |
| SEAQUEST | 0.0059 | 0.0049 | 0.0056 | **0.0146** | 0.0058 | 0.0068 | 0.0117±0.00 | **0.0134**±0.00 |
| UPNDOWN | 0.1893 | 0.1611 | 0.2277 | 0.2524 | 0.2765 | 0.3397 | **0.9253**±1.44 | **0.6941**±0.59 |
| **AVERAGE** | 0.2615 | 0.2171 | 0.3503 | 0.3320 | 0.3691 | 0.4647 | **0.6158**±0.32 | **0.6938**±0.24 |