MINIMAX-OPTIMAL MULTI-AGENT ROBUST REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent robust reinforcement learning, also known as multi-player robust Markov games (RMGs), is a crucial framework for modeling competitive interactions under environmental uncertainties, with wide applications in multi-agent systems. However, existing results on sample complexity typically suffer from the curse of multiple agents, where the complexity scales exponentially with the number of players. The Follow-the-Regularized-Leader (FTRL) framework has emerged as a powerful technique for addressing this issue in standard Markov games (MGs). However, its application to RMGs remains suboptimal, as the resulting complexity has significantly worse dependence on both the horizon H and the accuracy ε compared to prior works. To address these challenges, we study the problem in the finite-horizon setting, assuming an R-contamination uncertainty set and access to a generative model. We prove that applying FTRL in this setting achieves an ε-robust coarse correlated equilibrium (CCE) with a sample complexity (up to log factors) of $\widetilde{O}\left(H^3 S \sum_{i=1}^{m} A_i \min\{H, 1/R\} / \varepsilon^2\right)$, where S denotes the number of states, A_i is the number of actions of the i-th agent, and R is uncertainty level. Furthermore, we establish a matching information-theoretic lower bound under the same uncertainty model, thereby proving the minimax optimality of our result. To the best of our knowledge, this is the first work that achieves minimax-optimal sample complexity in the setting of RMGs.

1 INTRODUCTION

The rapidly evolving field of multi-agent reinforcement learning (MARL), also referred to as Markov games (MGs) (Littman, 1994; Shapley, 1953), explores how a group of agents interacts in a shared environment to maximize their individual expected cumulative rewards (Zhang et al., 2020a; Lanctot et al., 2019; Silver et al., 2017; Vinyals et al., 2019). This area has found wide applications in fields such as ecosystem management (Fang et al., 2015), autonomous driving (Zhou et al., 2020), and others (Silver et al., 2017; Qu et al., 2022). However, in real-world applications, environmental uncertainties — stemming from factors such as model misalignment — can significantly alter the qualitative outcomes of the game (Slumbers et al., 2023; Shi et al., 2024c; Balaji et al., 2019; Yeh et al., 2021; Zeng et al., 2022; Zhang et al., 2020b).

These challenges motivate the study of robust Markov games (RMGs), which assume that each agent aims to maximize its worst-case cumulative reward in an environment where the transition model is constrained by an uncertainty set centered around an unknown nominal model. Given the competitive nature of the game, the objective of RMGs is to reach an equilibrium where no agent has an incentive to unilaterally change its policy to increase its own payoff. A classical type of equilibrium is the robust Nash equilibrium (NE) (Blanchet et al., 2024; Nash Jr, 1950), where each agent's policy is independent, and no agent can improve its worst-case performance by deviating from its current strategy. Due to the high computational cost of solving robust NEs, this concept is often relaxed to the robust coarse correlated equilibrium (CCE), where agents' policies may be correlated (Moulin & Vial, 1978).

In the context of RMGs, achieving equilibrium with minimal samples is of particular interest, as data is often limited in practical applications. While there has been extensive research on the sample complexity of MGs, studies on RMGs remain limited (Kardeş et al., 2011), including the asymptotic convergence analysis (Zhang et al., 2020b) and the finite-sample analysis for RMGs (Blanchet

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2024; Ma et al., 2023; Shi et al., 2024b;a; Gai & Shi, 2026; Farhat et al., 2026) under various definitions of uncertainty sets. Specifically, Ma et al. (2023) showed that robust Correlated Equilibrium (CE) can be achieved with $\widetilde{O}\left(H^5 S \max_{1\le i\le m} A_i^2/\varepsilon^2\right)$ samples when the uncertainty level R is sufficiently small, i.e., $R \le \varepsilon/(SH^2)$, where S denotes the number of states, A_i denotes the number of actions for the i-th agent, H denotes the finite horizon length, and m denotes the number of agents. However, the applicability of this result is limited in scenarios where high robustness is required due to its restriction on a small R. Farhat et al. (2026) relaxed this constraint by establishing a sample complexity of $\widetilde{O}\left(H^3 S \prod_{i=1}^m A_i \min\{H, 1/R\}/\varepsilon^2\right)$ for arbitrary $R \in [0,1]$ under a failure states assumption. Blanchet et al. (2024) and Shi et al. (2024c) assumed access to an offline dataset or a generative model, and established that robust NEs can be achieved with sample complexities $\widetilde{O}\left(H^4 S^3 \prod_{i=1}^m A_i^2/\varepsilon^2\right)$ and $\widetilde{O}\left(H^3 S \prod_{i=1}^m A_i \min\{H, 1/R\}/\varepsilon^2\right)$, respectively. Despite this progress, a key limitation persists: these complexities scale sublinearly or even exponentially with the number of actions $\max_i A_i$, which is commonly referred to as the curse of multiple agents.

These complexities, which suffers from the curse of multiple agents, become prohibitive in environments involving many agents, such as multi-robot systems and market simulations (Casgrain & Jaimungal, 2020). However, under a policy-induced uncertainty set, Shi et al. (2024a) established a lower bound of $\widetilde{O}(H^3 \min\{H, 1/R\} S \max_{i=1} A_i/\varepsilon^2)$ for RMGs, which scales with $\max_{i=1} A_i$, indicating that the exponential cost may not be necessary. A powerful tool in addressing the curse of multiple agents is the Follow-the-Regularized-Leader (FTRL) framework, an online learning method for handling sequential decision-making problems. It has been effectively used in standard MGs to achieve minimax-optimal sample complexity (Li et al., 2022). However, its application in RMGs remains less developed. Shi et al. (2024a) leveraged FTRL to achieve a sample complexity of $\widetilde{O}\left(H^6 S \sum_{i=1}^m A_i \min\{H, 1/R\}/\varepsilon^4\right)$ for $\varepsilon \in (0, \sqrt{\min\{H, 1/R\}}]$. Gai & Shi (2026) extended this result to the linear function approximation setting and achieved a sample complexity of $H^9 d^3/\varepsilon^4$ under the same policy-induced uncertainty set. Specialized to the tabular setting, the sample complexity becomes $H^9 S^3 \max_i A_i^3/\varepsilon^4$. Though these approaches addressing the curse of multiple agents, they still exhibit several limitations when compared to the lower bound:

- *A restrictive range of accuracy ε.* This implies the existence of burn-in cost, which refers to the minimum number of samples required for the algorithm to be theoretically guaranteed to achieve minimax-optimal sample complexity. Results in (Shi et al., 2024a) only hold for $\varepsilon \in (0, \sqrt{\min\{H, 1/R\}}]$, which implies a burn-in cost of $\widetilde{O}\left(H^5 S \sum_{i=1}^m A_i \max\{1, HR\}\right)$. The algorithm's behavior before this threshold is met remains unknown. A large burn-in cost significantly reduces efficiency in applications where sample collection is expensive, making it difficult to meet the required threshold.
- *The barrier of long horizons H.* The dependency on H and target error ε is significantly larger than the information-theoretic lower bound, leading to inefficiencies in long-horizon or high-accuracy settings.

1.1 Our contributions

In this paper, we apply the FTRL framework to RMGs and, for the first time, establish minimax-optimal sample complexity up to the dependence on the number of agents in this setting. In particular, we focus on finite-horizon RMGs with a generative model, in which uncertainty is modeled via the R-contamination set (Huber, 1965) — a set previously studied in robust single-agent RL (Wang & Zou, 2021) and online robust MARL (Ma et al., 2023). Our contributions are as follows:

- Assuming access to a generative model, we demonstrate that applying the FTRL framework to RMGs can achieve an ε-robust CCE with the following sample complexity (up to logarithmic factors):

$$\widetilde{O}\left(\frac{H^3 S \sum_{i=1}^m A_i}{\varepsilon^2} \min\left\{H, \frac{1}{R}\right\}\right)$$

[2] In particular, Shi et al. (2024a) considered policy-dependent uncertainty sets; Shi et al. (2024c); Farhat et al. (2026); Gai & Shi (2026) considered uncertainty sets defined by $\mathsf{TV}(P, P') \le R$; Ma et al. (2023) studied general rectangular uncertainty sets and used those defined by $\mathsf{KL}(P, P') \le R$ and the R-contamination model; Blanchet et al. (2024) considered uncertainty sets defined by $\mathsf{TV}(P, P') \le R$ or $\mathsf{KL}(P, P') \le R$.

Algorithm	Equil.	setting	ε and R range	Sample compl.
Ma et al. (2023)	CE	online	$\varepsilon \in (0, H]$, $R \in \left(0, \frac{\varepsilon}{SH^2}\right]$	$\frac{H^6 S \max_{1\le i\le m} A_i^2}{\varepsilon^2}$
			$\varepsilon \in (0, H]$, $R \in \left(\frac{\varepsilon}{SH^2}, \frac{p_{\min}}{H}\right]$	$\frac{H^6 S \max_{1\le i\le m} A_i^2}{\varepsilon^2 p_{\min}^2}$
Farhat et al. (2026)	NE/CE/ CCE	online	$\varepsilon \in (0, H]$, $R \in [0, 1]$	$\frac{H^3 S \prod_{i=1}^m A_i}{\varepsilon^2} \min\{H, \frac{1}{R}\}$
Blanchet et al. (2024)	NE	offline	$\varepsilon \in (0, H]$, $R \in [0, 1]$	$\frac{H^4 S^3 \prod_{i=1}^m A_i^2}{\varepsilon^2}$
Gai & Shi (2026)	CCE	simulator	$\varepsilon \in (0, H]$, $R \in [0, 1]$	$\frac{H^9 d^3}{\varepsilon^4}$
Shi et al. (2024c)	NE/CE/ CCE	simulator	$\varepsilon \in \left(0, \sqrt{\min\{H, \frac{1}{R}\}}\right]$, $R \in [0, 1]$	$\frac{H^3 S \prod_{i=1}^m A_i}{\varepsilon^2} \min\{H, \frac{1}{R}\}$
Shi et al. (2024a)	CCE	simulator	$\varepsilon \in \left(0, \sqrt{\min\{H, \frac{1}{R}\}}\right]$, $R \in [0, 1]$	$\frac{H^6 S \sum_{i=1}^m A_i}{\varepsilon^4} \min\{H, \frac{1}{R}\}$
this work(Theorem 1)	CCE	simulator	$\varepsilon \in (0, H]$, $R \in [0, 1]$	$\frac{H^3 S \sum_{i=1}^m A_i}{\varepsilon^2} \min\{H, \frac{1}{R}\}$

Table 1: Our results and comparisons with prior art (up to log factors) for RMG algorithms in terms of equilibrium type, settings (data collection mechanism), range of ε and R, and the sample complexity. Here Gai & Shi (2026) studied the linear function approximation setting with dimension d, which becomes $d = S \max_i A_i$ when specialized to the tabular setting in this work. In addition, Farhat et al. (2026) required the failure states assumption. notation $p_{\min}$ denotes the minimal number of nominal transition probability. We note that the uncertainty sets may differ across literature [2].

for a full range of $\varepsilon \in (0, H]$ and uncertainty level $R \in [0, 1]$. Compared to previous results (Shi et al., 2024a;c; Ma et al., 2023; Blanchet et al., 2024; Gai & Shi, 2026; Farhat et al., 2026), our approach effectively addresses the challenges posed by multiple agents and long horizon lengths. Furthermore, our result holds for the entire range of ε and R, implying that the algorithm requires no burn-in cost, and allows for sufficient robustness. Detailed comparisons are in Table 1. More related works are compared in Appendix A.

- The information-theoretic lower bound for RMGs can vary depending on the structure of the uncertainty set. In this work, we derive a lower bound on the sample complexity under the R-contamination set for $R \in [0, c_R]$, where c_R is a constant smaller than 1. Up to logarithmic factors, this bound is given by:

$$\widetilde{O}\left(\frac{H^3 S \max_{1\le i\le m} A_i}{\varepsilon^2} \min\left\{H, \frac{1}{R}\right\}\right).$$

Combining with our algorithm's sample complexity, we conclude that the proposed algorithm achieves minimax-optimal sample complexity for $R \in [0, c_R]$ up to the dependence on the number of agents.

2 Background

2.1 Standard Markov games

A finite-horizon, non-stationary multi-player general-sum Markov game (MG) with m players competing against each other is represented by $\mathcal{MG} = \{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le m}, H, P, r\}$, where $\mathcal{S} = \{1, \cdots, S\}$ represents the state space of the shared environment, $\mathcal{A}_i = \{1, \cdots, A_i\}$ is the action space for the i-th player, H denotes the horizon of the Markov game, $P = \{P_h\}_{1\le h\le H}$ with $P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ describes the transition probability kernel for $\mathcal{MG}$, and $r = \{r_{i,h}\}_{1\le h\le H, 1\le i\le m}$ with $r_{i,h} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the reward received by the i-th player. Denote $\mathcal{A}$ as the Cartesian product $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$.

Markov policies. This paper focuses on Markov policies, which select actions based solely on the current state s and step h, independent of previous trajectories. Specifically, denote policy of the i-th player as $\pi_i = \{\pi_{i,h}\}_{1\le h\le H}$, where $\pi_{i,h}(\cdot|s) \in \Delta(\mathcal{A}_i)$ denotes the probability distribution over space $\mathcal{A}_i$ at step h. The joint Markov policy of all players is defined as $\pi = (\pi_1, \cdots, \pi_m) = \{\pi_h\}_{1\le h\le H} : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$, where $\pi_h(\cdot|s) = (\pi_{1,h}, \cdots, \pi_{m,h})(\cdot|s) \in \Delta(\mathcal{A})$ specifies the

joint action distribution for all players. For any joint policy π, we use π_{-i} to denote the policies of all players except the i-th player, and use $\pi_{-i,h}$ to represent the policies of all but the i-th player at step h.

Value functions and Q functions. For a given joint policy π and transition kernel P, the long-term cumulative reward of agent i in MG is quantified through the value function $V_{i,h}^{\pi,P} : \mathcal{S} \to \mathbb{R}$ and Q-function $Q_{i,h}^{\pi,P} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for the i-th agent, defined as follows:

$$V_{i,h}^{\pi,P}(s) := \mathbb{E}_{\pi,P}\left[\sum_{t=h}^{H} r_{i,t}(s_t, \boldsymbol{a}_t) \mid s_h = s\right], \quad \forall (i,h,s) \in [m] \times [H] \times \mathcal{S},$$

$$Q_{i,h}^{\pi,P}(s, \boldsymbol{a}) := \mathbb{E}_{\pi,P}\left[\sum_{t=h}^{H} r_{i,t}(s_t, \boldsymbol{a}_t) \mid s_h = s, \boldsymbol{a}_h = \boldsymbol{a}\right], \quad \forall (i,h,s,\boldsymbol{a}) \in [m] \times [H] \times \mathcal{S} \times \mathcal{A}.$$

where the expectation is taken over the distribution $\boldsymbol{a}_t \sim \pi_t(\cdot \mid s_t), s_{t+1} \sim P_t(\cdot \mid s_t, \boldsymbol{a}_t), t \geq h$ for $V_{i,h}^{\pi,P}(s)$, and $\boldsymbol{a}_{t+1} \sim \pi_{t+1}(\cdot \mid s_{t+1}), s_{t+1} \sim P_t(\cdot \mid s_t, \boldsymbol{a}_t), t \geq h$ for $Q_{i,h}^{\pi,P}(s, \boldsymbol{a})$.

2.2 Robust Markov games

A multi-agent general-sum robust Markov game (RMG) in the finite-horizon setting is represented by

$$\mathcal{MG}_{\mathsf{rob}} = \left\{\mathcal{S}, \{\mathcal{A}_i\}_{1 \leq i \leq m}, \mathcal{U}^R(P^0), r, H\right\},$$

where $\mathcal{S}$, $\{\mathcal{A}_i\}$, r, H and policies are defined as in standard MGs (cf. Section 2.1). The key distinction from standard MGs is that: instead of transitioning according to a fixed probability kernel, players in an RMG aim to maximize their worst-case cumulative reward, assuming the probability kernel is restricted within an uncertainty set $\mathcal{U}^R(P^0)$ centered around the nominal probability kernel P^0.

Uncertainty set with agent-wise (s,a)-rectangularity. In this work, the uncertainty set is constructed using the R-contamination model (Huber, 1965), which has been extensively used in the context of robust statistics (Huber, 1973; Huber & Ronchetti, 2011), robust reinforcement learning (Wang & Zou, 2021) and others (Nishimura & Ozaki, 2004; 2006; Duchi et al., 2019). Following Shi et al. (2024c), we consider a multi-agent variant of rectangularity in RMGs — agent-wise (s,a)-rectangularity, which enables the robust counterpart of Bellman equations and computational tractability of the problems (Iyengar, 2005; Shi et al., 2024b; Wiesemann et al., 2013; Zhou et al., 2021).

Specifically, denote nominal probability kernel $P^0 = \{P_h^0\}_{h=1}^H$ with $P_h^0 : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. The uncertainty set $\mathcal{U}^R(P^0)$ is mathematically defined as:

$$\mathcal{U}^R(P^0) = \otimes \mathcal{U}^R(P_{h,s,\boldsymbol{a}}^0), \quad \text{with} \quad \mathcal{U}^R(P_{h,s,\boldsymbol{a}}^0) = \left\{(1-R)P_{h,s,\boldsymbol{a}}^0 + RP' | P' \in \Delta(\mathcal{S})\right\}, \quad (1)$$

where $\otimes$ represents the Cartesian product, and $P_{h,s,\boldsymbol{a}}^0 := P_h^0(\cdot \mid s, \boldsymbol{a}) \in \mathbb{R}^S$. Here $R \in [0,1]$ is referred to as the uncertainty level.

Robust value functions and robust Bellman equations. For any given joint policy π, the worst-case performance of the i-th agent at time step h, over all possible transition kernels within uncertainty set $\mathcal{U}^R(P^0)$, is denoted by the robust value function $V_{i,h}^{\pi,R}$ and the robust Q-function $Q_{i,h}^{\pi,R}$:

$$V_{i,h}^{\pi,R}(s) = \inf_{P \in \mathcal{U}^R(P^0)} V_{i,h}^{\pi,P}(s) = \inf_{P \in \mathcal{U}^R(P^0)} \mathbb{E}_{\pi,P}\left[\sum_{t=h}^{H} r_{i,t}(s_t, \boldsymbol{a}_t) | s_h = s\right],$$

$$Q_{i,h}^{\pi,R}(s, \boldsymbol{a}) = \inf_{P \in \mathcal{U}^R(P^0)} Q_{i,h}^{\pi,P}(s, \boldsymbol{a}) = \inf_{P \in \mathcal{U}^R(P^0)} \mathbb{E}_{\pi,P}\left[\sum_{t=h}^{H} r_{i,t}(s_t, \boldsymbol{a}_t) | s_h = s, \boldsymbol{a}_h = \boldsymbol{a}\right].$$

In the context of RMGs, each agent aims to compete with others and maximize its worst-case performance. To characterize this, given the joint policy π_{-i} of all players except the i-th one, we use $V_{i,h}^{\star,\pi_{-i},R}$ to represent the optimal value function of agent i by playing an independent policy π_i':

$$V_{i,h}^{\star,\pi_{-i},R}(s) := \max_{\pi_i' : \mathcal{S} \times [H] \to \Delta(\mathcal{A}_i)} V_{i,h}^{\pi_i' \times \pi_{-i},R}(s) = \max_{\pi_i' : \mathcal{S} \times [H] \to \Delta(\mathcal{A}_i)} \inf_{P \in \mathcal{U}^R(P^0)} V_{i,h}^{\pi_i' \times \pi_{-i},P}(s).$$

As in standard MGs, it can be shown that there exists at least one policy — referred to as the robust best-response policy for the i-th agent—that achieves $V_{i,1}^{\star,\pi_{-i},R}(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$ (see Blanchet et al. (2024), Section A.2).

The robust value functions $\{V_{i,h}^{\pi,R}\}$ of RMGs with any joint policy π satisfy the following robust Bellman consistency equation: for all $(i,h,s) \in [m] \times [H] \times \mathcal{S}$,

$$\begin{aligned} V_{i,h}^{\pi,R}(s) &= \mathbb{E}_{\boldsymbol{a}\sim\pi_h(\cdot|s)}\left[r_{i,h}(s,\boldsymbol{a}) + \inf_{P_{h,s,\boldsymbol{a}}\in\mathcal{U}^R(P^0_{h,s,\boldsymbol{a}})} P_{h,s,\boldsymbol{a}} V_{i,h+1}^{\pi,R}\right] \\ &= \mathbb{E}_{\boldsymbol{a}\sim\pi_h(\cdot|s)}\left[r_{i,h}(s,\boldsymbol{a}) + (1-R)P^0_{h,s,\boldsymbol{a}} V_{i,h+1}^{\pi,R} + R\min_{s\in\mathcal{S}} V_{i,h+1}^{\pi,R}\right]. \end{aligned} \tag{2}$$

This robust Bellman equation is inherently tied to the agent-wise (s,a)-rectangularity condition and the R-contamination model (cf. (1)).

Equilibria in robust Markov games. Finding an equilibrium is central to the study of RMGs. In this paper, we consider robust Coarse Correlated Equilibrium (CCE). A (possibly correlated) joint policy $\pi \in \mathcal{S} \times [H] \to \Delta(\mathcal{A})$ is a robust CCE if

$$\forall (i,s) \in [m] \times \mathcal{S}: \quad V_{i,1}^{\pi,R}(s) \geq V_{i,1}^{\star,\pi_{-i},R}(s). \tag{3}$$

Robust CCE guarantees that no player can benefit from unilaterally deviating its current strategy. In particular, CCE assumes that each agent receives a recommended action drawn from a joint distribution (the CCE policy), and no agent can gain by rejecting the recommendation, provided others follow theirs.

Computing exact robust equilibria is a challenging task and may not always be necessary. A slight relaxation of the definition (3) introduces the concept of ε-robust CCE. A possibly correlated joint policy $\pi \in \mathcal{S} \times [H] \to \Delta(\mathcal{A})$ is said to be an ε-robust CCE if

$$\mathsf{gap}_{\mathsf{CCE}}(\pi) = \max_{s\in\mathcal{S}} \mathsf{gap}(\pi; s) \leq \varepsilon, \quad \text{with} \quad \mathsf{gap}(\pi; s) := \max_{1\leq i\leq m}\left\{V_{i,1}^{\star,\pi_{-i},R}(s) - V_{i,1}^{\pi,R}(s)\right\}. \tag{4}$$

The existence of robust CCEs has been proved under general conditions for the uncertainty set (Blanchet et al., 2024).

Sampling mechanism: a generative model. In this paper, we adopt a commonly used sampling mechanism in standard MARL (Li et al., 2022; Zhang et al., 2020b). Specifically, we assume access to a generative model (simulator) (Kearns & Singh, 1998), which returns an independent sample of state s' based on the nominal transition kernel P^0, as follows:

$$s' \sim P_h^0(\cdot|s,\boldsymbol{a}),$$

for any pair $(s,\boldsymbol{a},h) \in \mathcal{S} \times \mathcal{A} \times [H]$ selected by the learners. The objective of this paper is to compute an ε-robust equilibrium for the game $\mathcal{MG}_{\mathsf{rob}}$ with as few samples as possible.

3 Sample-efficient robust Markov games

In this section, we apply the FTRL framework in a similar way to its use in standard MGs (Li et al., 2022), and show that it retains the minimax sample complexity in this setting.

3.1 Algorithm description

Overall, the algorithm proceeds backward from step $h = H$ to $h = 1$. At each step h, each agent i performs K iterations of the FTRL procedure: in each iteration, it draws SA samples from the nominal transition probability, estimates the worst-case Q-function under the current policy, and updates its policy $\widehat{\pi}_{i,h}^{k+1}$ via the FTRL rule. After K iterations, the resulting joint correlated policy approximates a robust CCE at step h. The algorithm then computes the value function under this optimized policy and moves to step $h-1$.

We now present the mathematical formulation. The algorithm initializes the value functions as $\widehat{V}_{i,H+1} = 0 \in \mathbb{R}^S$ for $1 \leq i \leq m$, then iterates over each step h, using estimates $\{\widehat{V}_{i,h+1}\}_{i=1}^m$ from the previous step. Each step h consists of two phases: policy updates and value function estimation.

Algorithm 1: Sample-efficient Robust Q-FTRL

Input: number of rounds K for each step, learning rates $\{\alpha_k\}$ and $\{\eta_{k+1}\}$.

Initialize: for any $(s, a_i, h, i) \in \mathcal{S} \times \mathcal{A}_i \times [H] \times [m]$, set $\widehat{V}_{i,H+1}(s) = Q^0_{i,h}(s, a_i) = 0$ and $\pi^1_{i,h}(a_i|s) = 1/A_i$.
for $h = H, \cdots, 1$ **do**
  for $k = 1, \cdots, K$ **do**
    for $i = 1, \cdots, m$ **do**
      Draw independent samples from generative model according to (6) for all $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$, and construct the model estimation $r^k_{i,h}, P^k_{i,h}$ using (7) and (8).
      Compute $Q^k_{i,h}$ according to (9).
      Update $\pi^{k+1}_{i,h}(\cdot|s)$ according to (11) for all $s \in \mathcal{S}$.
    end for
  end for
  Compute $\widehat{V}_{i,h}$ according to (13).
end for
Output: $\widehat{\pi} = \{\widehat{\pi}_h\}_{1 \le h \le H}$ where $\widehat{\pi}_h = \sum_{k=1}^K \alpha_k^K (\pi^k_{1,h} \times \cdots \times \pi^k_{m,h})$.

Sampling and policy updates. At each step, the algorithm draws K samples from the generative model for each $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$ and each agent $i \in [m]$, and uses these samples to update policies. In particular, in the k-th sample collection step, a tuple $\{\boldsymbol{a}^k_h(s, a_i) = \{a^k_{j,h}(s, a_i)\}_{j=1}^m, r^k_{i,h}(s, a_i), s^k_{i,h+1}(s, a_i)\}_{(s,a_i) \in \mathcal{S} \times \mathcal{A}_i}$ is drawn according to the following rules:

$$a^k_{j,h}(s, a_i) \sim \pi^k_{j,h}(\cdot|s), \ (j \neq i) \quad \text{and} \quad a^k_{i,h}(s, a_i) = a_i; \tag{5}$$

$$r^k_{i,h}(s, a_i) = r_{i,h}(s, \boldsymbol{a}^k_h(s, a_i)); \qquad s^k_{i,h+1}(s, a_i) \sim P^0_h(\cdot|s, \boldsymbol{a}^k_h(s, a_i)), \tag{6}$$

where $\pi^k_{j,h}(\cdot|s)$ is initialized as $\pi^1_{j,h}(a_j|s) = 1/A_j$ for all $(s, a_j, j) \in \mathcal{S} \times \mathcal{A}_j \times [m]$. For convenience, we collect samples for all state-action pairs into a reward vector $r^k_{i,h} \in \mathbb{R}^{SA_i}$ and an empirical transition probability matrix $P^k_{i,h} \in \mathbb{R}^{SA_i \times S}$, given by

$$r^k_{i,h} = [r^k_{i,h}(s, a_i)]_{(s,a_i) \in \mathcal{S} \times \mathcal{A}_i} \in \mathbb{R}^{SA_i}, \tag{7}$$

$$P^k_{i,h}(s'|s, a_i) = 1, \text{ for } s' = s^k_{h+1}(s, a_i), \quad \text{and} \quad P^k_{i,h}(s'|s, a_i) = 0, \text{ otherwise.} \tag{8}$$

Next, we update $Q^k_{i,h}$ following the framework of FTRL:

$$Q^k_{i,h} = (1 - \alpha_k) Q^{k-1}_{i,h} + \alpha_k q^k_{i,h}, \tag{9}$$

where $0 < \alpha_k < 1$ is the learning rate, $Q^k_{i,h} \in \mathbb{R}^{SA_i}$ is initialized as $Q^0_{i,h} = 0$, and $q^k_{i,h}$ is computed according to Bellman operator:

$$q^k_{i,h} = r^k_{i,h} + (1 - R) P^k_{i,h} \widehat{V}_{i,h+1} + R \min_{s \in \mathcal{S}} \widehat{V}_{i,h+1}. \tag{10}$$

After updating the Q-function estimates, the algorithm uses the FTRL strategy (Shalev-Shwartz, 2012) to update the policy of the i-th agent as follows:

$$\pi^{k+1}_{i,h}(\cdot|s) = \frac{\exp\left(\eta_{k+1} Q^k_{i,h}(s, \cdot)\right)}{\sum_{a \in \mathcal{A}_i} \exp\left(\eta_{k+1} Q^k_{i,h}(s, a)\right)}, \quad \forall s \in \mathcal{S}, \tag{11}$$

where $\eta_{k+1} > 0$ is another learning rate associated with the policy updates.

Value function estimation. After K rounds of the above procedure, we construct the final policy estimate $\widehat{\pi}_h : \mathcal{S} \to \Delta(\mathcal{A})$ as

$$\widehat{\pi}_h(\boldsymbol{a}|s) = \sum_{k=1}^K \alpha_k^K \prod_{i=1}^m \pi^k_{i,h}(a_i|s), \quad \forall (s, \boldsymbol{a}) \in \mathcal{S} \times \mathcal{A},$$

where $\alpha_k^K = \alpha_k \prod_{j=k+1}^K (1-\alpha_j)$ for $1 \le k \le K$. Moreover, we estimate the value function $\widehat{V}_{i,h} \in \mathbb{R}^S$ for step h under $\widehat{\pi}_h$. To obtain an optimistic estimation, we introduce a bonus term $\beta_{i,h} \in \mathbb{R}^S$ with the s-th entry given by

$$\beta_{i,h}(s) = c_{\mathsf{b}} \sqrt{\frac{\log^3 \frac{KS\sum_{i=1}^m A_i}{\delta}}{K \min\left\{H, \frac{1}{R}\right\}}} \sum_{k=1}^K \alpha_k^K \left\{ \mathsf{Var}_{\pi_{i,h,s}^k}(q_{i,h,s}^k) + \min\left\{H, \frac{1}{R}\right\} \right\}, \tag{12}$$

where $q_{i,h,s}^k = q_{i,h}^k(\cdot|s) \in \mathbb{R}^{A_i}$, and $\pi_{i,h,s}^k = \pi_{i,h}^k(\cdot|s) \in \Delta(\mathcal{A}_i)$. The value function $\widehat{V}_{i,h}$ is then estimated as

$$\widehat{V}_{i,h}(s) = \min\left\{ \sum_{k=1}^K \alpha_k^K \mathbb{E}_{a_i \sim \pi_{i,h,s}^k} q_{i,h,s}^k + \beta_{i,h}(s), H-h+1 \right\}, \quad \forall s \in \mathcal{S}. \tag{13}$$

Equipped with $\widehat{\pi}_h$ and $\widehat{V}_{i,h}$, the algorithm proceeds to step $h-1$ and repeats the above operations.

The entire procedure is summarized in Algorithm 1, with the total number of samples being $KHS\sum_{i=1}^m A_i$. The robust parameter R appears in the computation of the bonus term $\beta_{i,h}$ in (12) and $q_{i,h}^k$ in (10), and consequently influences the computation of $Q_{i,h}^k$ in (9).

3.2 Main results

The sample complexity of Algorithm 1 is formalized in the following theorem, the proof of which is postponed to Appendix C.

Theorem 1. *Consider an m-player robust general-sum Markov game with uncertainty set $\mathcal{U}^R$ defined in (1) and uncertainty level R. Suppose that the step-size*

$$\alpha_k = \frac{c_\alpha \log K}{k - 1 + c_\alpha \log K}, \quad \eta_k = \sqrt{\frac{\log K}{\alpha_k \min\left\{H, \frac{1}{R}\right\}}}, \tag{14}$$

where $c_\alpha > 0$ is a constant. For any $\varepsilon \in (0, H]$, $0 \le R \le 1$, and $0 < \delta < 1$, with probability at least $1-\delta$, the joint policy $\widehat{\pi}$ returned by Algorithm 1 achieves an ε-robust CCE (cf. (4))

$$\mathsf{gap}_{\mathsf{CCE}}(\widehat{\pi}) \le \varepsilon$$

as long as the total number of samples

$$KHS\sum_{i=1}^m A_i \ge \frac{CH^3 S\sum_{i=1}^m A_i}{\varepsilon^2} \min\left\{H, \frac{1}{R}\right\} \log^3\left(\frac{KS\sum_{i=1}^m A_i}{\delta}\right), \tag{15}$$

where $C > 0$ is a large enough universal constant.

The following theorem establishes an information-theoretic lower bound for learning RMGs with a generative model. Its proof is postponed to Appendix E. Although lower bounds exist for RMGs under various uncertainty sets, none applies to the set in (1). We therefore derive one, which coincides with the result for the total variation distance-based set in Shi et al. (2024a).

Theorem 2. *Consider any $m \ge 2$, $H \ge 16$, $0 < \varepsilon < c_0 H$, and any $R \in [0, c_R]$, where $c_0 > 0$ is a small enough constant and c_R is a constant within $(0,1)$. We can construct a collection of m-player general-sum robust Markov games — denoted by $\{\mathcal{MG}_\theta\}_{\theta\in\Theta}$, such that for any dataset comprising N independent samples generated from the nominal environment for each state-action pairs $(s, a_i) \in \mathcal{S}\times\mathcal{A}_i$, $1 \le i \le m$, one has*

$$\inf_{\widehat{\pi}} \max_{\theta\in\Theta} \mathbb{P}_\theta\left(\mathsf{gap}_{\mathsf{CCE}}(\widehat{\pi}) > \varepsilon\right) \ge \frac{1}{4},$$

provided that

$$NHS \max_{1\le i\le m} A_i \le \frac{C_1 H^3 S \max_{1\le i\le m} A_i}{\varepsilon^2} \min\left\{H, \frac{1}{R}\right\}.$$

Here, C_1 is some small enough constant, the infimum is taken over all estimators $\widehat{\pi}$, and $\mathbb{P}_\theta$ denotes the probability when the game is $\mathcal{MG}_\theta$ for all $\theta \in \Theta$.

According to Theorem 1, the sample complexity of Algorithm 1 is given by (up to logarithmic factors)

$$\widetilde{O}\left(\frac{H^3 S\sum_{i=1}^m A_i}{\varepsilon^2}\min\left\{H,\frac{1}{R}\right\}\right).$$

This result overcomes the curse of multiple agents due to the use FTRL. Furthermore, when compared to the information-theoretic lower bound established in Theorem 2, we observe that this complexity is minimax optimal (up to logarithmic terms) for $R\in[0,c_R]$, provided the number of players m remains fixed or grows polylogarithmically with respect to the problem parameters. To the best of our knowledge, this is the first result for RMGs to achieve minimax optimal sample complexity.

In addition, our sample complexity holds for any uncertainty level $R\in[0,1]$. This makes our theory applicable across a wide range of practical scenarios that may require various robustness levels. Furthermore, it also holds for the entire range of $\varepsilon\in(0,H]$. This is particularly favoring for data-limited applications, as it implies that the algorithm incurs no burn-in cost. A detailed explanation is provided in Section 1. To the best of our knowledge, this is the first work to simultaneously address all of the following three challenges within the RMG framework: the restrictive range of uncertainty level or accuracy, the curse of multiple agents, and the barrier of long horizons (cf. Section 1).

Finally, we would like to clarify why a larger R yields a smaller sample complexity. Although larger R corresponds to a larger uncertainty set, the robust Bellman operator has the form $(1-R)P^0_{h,s,\boldsymbol{a}}V+R\min_{s'}V(s')$ (cf. (2)). Hence, the contribution that requires statistical estimation is weighted by $1-R$, whereas the worst-case component can be evaluated directly. This explains why the statistical complexity can decrease as R increases.

4 Proof of Theorem 1

Step 1. introduce auxiliary sequences and policies. We begin by defining the notation $\overline{V}$, which denotes the value function estimate obtained by substituting every expectation operator with the corresponding empirical estimation based on $(r^k_{i,h},P^k_{i,h})$. Specifically, we set

$$\overline{V}^{\widehat{\pi},R}_{i,h}(s)=\sum_{k=1}^K\alpha_k^K\mathbb{E}_{a_i\sim\pi^k_{i,h,s}}\left[r^k_{i,h}(s,a_i)+(1-R)P^k_{i,h}(s,a_i)\overline{V}^{\widehat{\pi},R}_{i,h+1}+R\min_{s\in\mathcal{S}}\overline{V}^{\widehat{\pi},R}_{i,h+1}\right],$$

where $P^k_{i,h}(s,a_i):=P^k_{i,h}(\cdot|s,a_i)\in\mathbb{R}^S$ and $\overline{V}^{\widehat{\pi},R}_{i,H+1}=\overline{V}^{\star,\widehat{\pi}_{-i},R}_{i,H+1}=0$. Note that the expectation here is taken only with respect to the action a_i, while the randomness in the actions of the other players $\boldsymbol{a}_{-i}$ is implicitly captured in the samples $r^k_{i,h}$ and $P^k_{i,h}$. Moreover, define $\widehat{\pi}^\star_i$ as the optimal policy of the i-th agent when other agents adopt correlated policy $\widehat{\pi}_{-i}$:

$$\widehat{\pi}^\star_i=\{\widehat{\pi}^\star_{i,h}\}_{h=1}^H=\arg\max_{\pi_i:\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)}V^{\pi_i\times\widehat{\pi}_{-i},R}_{i,1}. \tag{16}$$

By Bellman operator, the corresponding value function estimation $\overline{V}^{\widehat{\pi}^\star_i\times\widehat{\pi}_{-i},R}_{i,h}$ under $\widehat{\pi}^\star_i\times\widehat{\pi}_{-i}$ is

$$\begin{aligned}\overline{V}^{\widehat{\pi}^\star_i\times\widehat{\pi}_{-i},R}_{i,h}(s)=\sum_{k=1}^K\alpha_k^K\mathbb{E}_{a_i\sim\widehat{\pi}^\star_{i,h,s}}\Big[&r^k_{i,h}(s,a_i)+(1-R)P^k_{i,h}(s,a_i)\overline{V}^{\widehat{\pi}^\star_i\times\widehat{\pi}_{-i},R}_{i,h+1}\\&+R\min_{s\in\mathcal{S}}\overline{V}^{\widehat{\pi}^\star_i\times\widehat{\pi}_{-i},R}_{i,h+1}\Big],\end{aligned} \tag{17}$$

where $\overline{V}^{\widehat{\pi}^\star_i,\widehat{\pi}_{-i},R}_{i,H+1}=0$, and $\widehat{\pi}^\star_{i,h,s}:=\widehat{\pi}^\star_{i,h}(\cdot|s)\in\Delta(\mathcal{A}_i)$. According to the definition of $\mathsf{gap}_{\mathsf{CCE}}(\widehat{\pi})$ in (4), it suffices to prove that for all $i\in[m]$, $V^{\star,\widehat{\pi}_{-i},R}_{i,1}(s)-V^{\widehat{\pi},R}_{i,1}(s)\le\varepsilon$, $\forall s\in\mathcal{S}$. Towards this, we decompose this gap into several errors term:

$$V^{\star,\widehat{\pi}_{-i},R}_{i,h}-V^{\widehat{\pi},R}_{i,h}=\underbrace{V^{\widehat{\pi}^\star_i\times\widehat{\pi}_{-i},R}_{i,h}-\overline{V}^{\widehat{\pi}^\star_i\times\widehat{\pi}_{-i},R}_{i,h}}_{\mathcal{E}_1}+\underbrace{\overline{V}^{\widehat{\pi}^\star_i\times\widehat{\pi}_{-i},R}_{i,h}-\widehat{V}_{i,h}}_{\mathcal{E}_2}+\underbrace{\widehat{V}_{i,h}-\overline{V}^{\widehat{\pi},R}_{i,h}}_{\mathcal{E}_3}+\underbrace{\overline{V}^{\widehat{\pi},R}_{i,h}-V^{\widehat{\pi},R}_{i,h}}_{\mathcal{E}_4},$$

where we remind the reader that $V^{\star,\widehat{\pi}_{-i},R}_{i,h}=V^{\widehat{\pi}^\star_i\times\widehat{\pi}_{-i},R}_{i,h}$ (cf. (16)). Among the four decomposed error terms $\mathcal{E}_1,\dots,\mathcal{E}_4$, terms $\mathcal{E}_1$ and $\mathcal{E}_4$ share a similar structure, both reflecting the estimation error caused by the sampling of $\boldsymbol{a}_{-i}$ and previous policies. In the remainder of the proof, we handle these two jointly, analyze the remaining two separately, and then combine the results to complete the proof.

Step 2. bounding $\mathcal{E}_1$ and $\mathcal{E}_4$. By the definition of $V_{i,h}^{\widehat{\pi},R}$ and the Bellman equation, this term can be expressed in terms of the expected reward $r_{i,h}^{\widehat{\pi}} = \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}} r_{i,h}(s,\boldsymbol{a})$ and the transition distribution $P_{h,s}^{\widehat{\pi}} = \mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}} P_h^0(\cdot|s,\boldsymbol{a})$ under policy $\widehat{\pi}$:

$$V_{i,h}^{\widehat{\pi},R}(s) = r_{i,h}^{\widehat{\pi}}(s) + P_{h,s}^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}} V_{i,h+1}^{\widehat{\pi},R},$$

Here, for any given function V, the worst-case transition distribution is defined as $P_{h,s}^{\widehat{\pi},V} = (1-R)P_{h,s}^{\widehat{\pi}} + Re_{\arg\min V}$, with e_i denotes the i-th natural basis vector. Notice that due to sampling randomness, $\overline{V}_{i,h}^{\widehat{\pi},R}$ does not follow the Bellman equation under $r_{i,h}^{\widehat{\pi}}$ and $P_{h,s}^{\widehat{\pi}}$, which leads to a discrepancy with $V_{i,h}^{\widehat{\pi},R}(s)$. By controlling this difference recursively, we yield (cf. (30) - (35b) and Lemma 2)

$$\mathcal{E}_1 + \mathcal{E}_4 \lesssim H\sqrt{\frac{\min\{H, \frac{1}{R}\}\log^3 \frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}. \tag{18}$$

Step 3. bounding $\mathcal{E}_3$. Comparing the definition of $\widehat{V}_{i,h}$ and $\overline{V}_{i,h}^{\widehat{\pi},R}$, we observe that their difference arises from the inclusion of upper confidence bound $\beta_{i,h}$, i.e., (cf. (27a), (27b), and (36))

$$\mathcal{E}_3 \le \sum_{h'=h}^{H}\prod_{j=h}^{h'-1} \widehat{P}_{i,j}^{\widehat{\pi},\overline{V}_{i,j+1}^{\widehat{\pi},R}}\beta_{i,h'}, \tag{19}$$

where $\widehat{P}_{i,h}^{\widehat{\pi},V} = (1-R)\widehat{P}_{i,h,s}^{\widehat{\pi}} + Re_{\arg\min V}$ with $\widehat{P}_{i,h,s}^{\widehat{\pi}} = \sum_{k=1}^K \alpha_k^K \mathbb{E}_{a_i\sim\pi_{i,h,s}^k} P_{i,h}^k(s,a_i)$. By analyzing $\beta_{i,h}$ (cf. Lemma 3), we could further obtain

$$\mathcal{E}_3 \lesssim H\sqrt{\frac{\min\left\{H, \frac{1}{R}\right\}\log^3 \frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}.$$

Step 4. bounding $\mathcal{E}_2$. Given the sampled reward $r_{i,h}^k$ and transition kernel $P_{i,h}^k$, the value function $\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}$ satisfies a Bellman equation in single-agent reinforcement learning (cf. (17)), with reward $\sum_{k=1}^K \alpha_k^K r_{i,h}^k$, transition kernel $\sum_{k=1}^K \alpha_k^K P_{i,h}^k$, and policy $\widehat{\pi}_i^\star$. We define $\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}$ as the optimal value function of the above robust MDP (cf. (39)). It is obvious that $\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R} \le \overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}$. Moreover, we reveal that $\widehat{V}_{i,h}$ is an optimism estimation of $\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}$ due to the inclusion of the upper confidence bound, i.e., $\widehat{V}_{i,h} \ge \overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}$ (cf. Lemma 4), which further gives $\mathcal{E}_2 \le \overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R} - \widehat{V}_{i,h} \le 0$. Combining these results completes the proof.

5 Discussion

This paper investigates multi-player robust Markov games in a finite-horizon setting, and for the first time achieves minimax sample complexity through establishing matching upper and lower bounds. Specifically, we consider an R-contamination set with uncertainty level R as uncertainty set for the transition probabilities. We apply the FTRL framework to this robust setting, and demonstrate that it achieves an ε-robust CCE with a sample complexity of $H^3S\sum_{i=1}^m A_i \min\{H, 1/R\}/\varepsilon^2$ (up to log factors), where S denotes the number of states, A_i is the number of actions for the i-th player, and H represents the horizon length. Moreover, we establish a matching information-theoretic lower bound under the same uncertainty set, thereby demonstrating the minimax optimality of our result. Compared to previous results, our result simultaneously addresses three challenges: the restrictive range of uncertainty level or accuracy, curse of multiple agents, and the barrier of long horizon.

Future work will focus on extending this algorithm to handle other types of uncertainty sets, such as those defined by total variation distance or Kullback-Leibler divergence. In addition, we will investigate the possibility of developing algorithms that can achieve minimax-optimal sample complexity without relying on generative models — for example, in the online setting, which is more challenging and may require exploration mechanisms such as UCB techniques. Finally, it remains an open problem whether the dependence on the action space can be further improved from $\sum_{i=1}^m A_i$ to $\max_i A_i$, which we leave for future work.

AI USE STATEMENT

In preparing this paper, large language models (LLMs) were employed solely as assistive tools for minor language refinement. All technical contributions, results, and conclusions are entirely the work of the authors.

REFERENCES

Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning*, pp. 511–520. PMLR, 2021.

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

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in neural information processing systems*, 33:2159–2170, 2020.

Bharathan Balaji, Sunil Mallya, Sahika Genc, Saurabh Gupta, Leo Dirac, Vineet Khare, Gourav Roy, Tao Sun, Yunzhe Tao, Brian Townsend, et al. Deepracer: Educational autonomous racing platform for experimentation with sim2real reinforcement learning. *arXiv preprint arXiv:1911.01562*, 2019.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *Advances in Neural Information Processing Systems*, 36, 2024.

Philippe Casgrain and Sebastian Jaimungal. Mean-field games with differing beliefs for algorithmic trading. *Mathematical Finance*, 30(3):995–1034, 2020.

Zhi Chen, Melvyn Sim, and Huan Xu. Distributionally robust optimization with infinitely constrained ambiguity sets. *Operations Research*, 67(5):1328–1344, 2019.

Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In *International Conference on Algorithmic Learning Theory*, pp. 227–261. PMLR, 2022.

Qiwen Cui and Simon S Du. When are offline two-player zero-sum markov games solvable? *Advances in Neural Information Processing Systems*, 35:25779–25791, 2022.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of markov equilibrium in stochastic games. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4180–4234. PMLR, 2023.

Zehao Dou, Zhuoran Yang, Zhaoran Wang, and Simon Du. Gap-dependent bounds for two-player markov games. In *International Conference on Artificial Intelligence and Statistics*, pp. 432–455. PMLR, 2022.

John C Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses against mixture covariate shifts. *Under review*, 2(1), 2019.

Fei Fang, Peter Stone, and Milind Tambe. When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *IJCAI*, volume 15, pp. 2589–2595, 2015.

Zain Ulabedeen Farhat, Debamita Ghosh, George Atia, and Yue Wang. Sample-efficient distributionally robust multi-agent reinforcement learning via online interaction. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 151781–151838, 2026.

Jingchu Gai and Laixi Shi. Taming the curses of multiagency in robust markov games with large state space through linear function approximation. *arXiv preprint arXiv:2605.03125*, 2026.

E. N. Gilbert. A comparison of signalling alphabets. *The Bell System Technical Journal*, 31(3):504–522, 1952. doi: https://doi.org/10.1002/j.1538-7305.1952.tb01393.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1952.tb01393.x.

Vineet Goyal and Julien Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

Songyang Han, Sanbao Su, Sihong He, Shuo Han, Haizhao Yang, Shaofeng Zou, and Fei Miao. What is the solution for state-adversarial multi-agent reinforcement learning? *arXiv preprint arXiv:2212.02705*, 2022.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Fast bellman updates for robust mdps. In *International Conference on Machine Learning*, pp. 1979–1988. PMLR, 2018.

Chin Pang Ho, Marek Petrik, and Wolfram Wiesemann. Partial policy iteration for l1-robust markov decision processes. *Journal of Machine Learning Research*, 22(275):1–46, 2021.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pp. 1753–1758, 1965.

Peter J Huber. The use of choquet capacities in statistics. *Bulletin of the International Statistical Institute*, 45(4):181–188, 1973.

Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. John Wiley & Sons, 2011.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

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

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

Shyam Sundar Kannan, Vishnunandan LN Venkatesh, and Byung-Cheol Min. Smart-llm: Smart multi-agent robot task planning using large language models. *arXiv preprint arXiv:2309.10062*, 2023.

Erim Kardeş, Fernando Ordóñez, and Randolph W Hall. Discounted robust stochastic games and an application to queueing control. *Operations research*, 59(2):365–382, 2011.

Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent rl in markov games with a generative model. *arXiv preprint arXiv:2208.10458*, 2022.

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

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

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

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. Decentralized robust v-learning for solving markov games with model uncertainty. *Journal of Machine Learning Research*, 24(371):1–40, 2023.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, 13(1):165–186, 2023.

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

John F Nash Jr. Equilibrium points in n-person games. *Proceedings of the national academy of sciences*, 36(1):48–49, 1950.

Kiyohiko G Nishimura and Hiroyuki Ozaki. Search and knightian uncertainty. *Journal of Economic Theory*, 119(2):299–333, 2004.

Kiyohiko G Nishimura and Hiroyuki Ozaki. An axiomatic approach to ϵ-contamination. *Economic Theory*, 27(2):333–340, 2006.

Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning for multiagent networked systems. *Operations Research*, 70(6):3601–3628, 2022.

Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. *Advances in neural information processing systems*, 30, 2017.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning,*, 4(2):107–194, 2012.

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

Laixi Shi, Jingchu Gai, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Breaking the curse of multiagency in robust multi-agent reinforcement learning. *arXiv preprint arXiv:2409.20067*, 2024a.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 36, 2024b.

Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. *arXiv preprint arXiv:2404.18909*, 2024c.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Oliver Slumbers, David Henry Mguni, Stefano B Blumberg, Stephen Marcus Mcaleer, Yaodong Yang, and Jun Wang. A game-theoretic framework for managing risk in multi-agent systems. In *International Conference on Machine Learning*, pp. 32059–32087. PMLR, 2023.

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

Aviv Tamar, Shie Mannor, and Huan Xu. Scaling up robust mdps using function approximation. In *International conference on machine learning*, pp. 181–189. PMLR, 2014.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. In *International conference on machine learning*, pp. 10279–10288. PMLR, 2021.

A. Tsybakov. Introduction to nonparametric estimation. In *Springer*, 2009.

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

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3370–3398. PMLR, 2023.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust q-learning. *Journal of Machine Learning Research*, 25(341):1–77, 2024.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30, 2017.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Huan Xu and Shie Mannor. Distributionally robust markov decision processes. *Advances in Neural Information Processing Systems*, 23, 2010.

Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. Model-based reinforcement learning is minimax-optimal for offline zero-sum markov games. *arXiv preprint arXiv:2206.04044*, 5:21, 2022.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *The Annals of Statistics*, 50(6): 3223–3248, 2022.

Wenhao Yang, Han Wang, Tadashi Kozuno, Scott M Jordan, and Zhihua Zhang. Robust markov decision processes without model estimation. *arXiv preprint arXiv:2302.01248*, 2023.

Yuepeng Yang and Cong Ma. $o(t^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum markov games. *arXiv preprint arXiv:2209.12430*, 2022.

Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B Lobell, and Stefano Ermon. Sustainbench: Benchmarks for monitoring the sustainable development goals with machine learning. *arXiv preprint arXiv:2111.04724*, 2021.

Lanting Zeng, Dawei Qiu, and Mingyang Sun. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. *Applied Energy*, 324:119688, 2022.

Kaiqing Zhang, Sham Kakade, Tamer Basar, and Lin Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33:1166–1178, 2020a.

Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33:10571–10583, 2020b.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pp. 27117–27142. PMLR, 2022.

Ming Zhou, Jun Luo, Julian Villella, Yaodong Yang, David Rusu, Jiayu Miao, Weinan Zhang, Montgomery Alban, Iman Fadakar, Zheng Chen, et al. Smarts: Scalable multi-agent reinforcement learning training school for autonomous driving, 2020. *arXiv preprint arXiv:2010.09776*, 2020.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

A Related works

In this section, we shall briefly review some works related to robust Markov games.

Standard Markov games. Early research on Markov games, also known as Multi-agent reinforcement learning (MARL) (Shapley, 1953; Oroojlooy & Hajinezhad, 2023; Zhang et al., 2021), primarily focused on asymptotic analysis (Littman & Szepesvári, 1996; Littman et al., 2001), but recent theoretical efforts have increasingly concentrated on finite-sample analysis. A significant portion of the literature has focused on the special case of two-player zero-sum Markov games, examples include Bai et al. (2020); Bai & Jin (2020); Zhong et al. (2022); Chen et al. (2022); Cui & Du (2022); Chen et al. (2019); Dou et al. (2022); Jia et al. (2019); Mao & Başar (2023); Tian et al. (2021); Wei et al. (2017); Yan et al. (2022); Yang & Ma (2022). Other studies have aimed to develop computationally efficient algorithms for multi-agent general-sum MGs, including Daskalakis et al. (2023); Jin et al. (2021); Liu et al. (2021); Mao & Başar (2023); Song et al. (2021). The minimax-optimal sample complexity for finite-horizon settings has been established as $\widetilde{O}\left(\frac{H^4 S \sum_{i=1}^m A_i}{\varepsilon^2}\right)$, and has been achieved by Li et al. (2022) under the same generative model assumption employed in this work, which corresponds to the case of $R = 0$ in our work.

Robust reinforcement learning. In the context of single-agent reinforcement learning (RL), the study of model robustness through distributionally robust dynamic programming and robust Markov Decision Processes (MDPs) has garnered significant attention, both empirically and theoretically (Badrinath & Kalathil, 2021; Goyal & Grand-Clement, 2023; Ho et al., 2018; 2021; Iyengar, 2005; Roy et al., 2017; Tamar et al., 2014; Xu & Mannor, 2010). A growing body of work has focused on the finite-sample analysis for robust RL, where the agent aims to optimize its worst-case performance under a transition model constrained by an uncertainty set (Panaganti & Kalathil, 2022; Yang et al., 2022; Zhou et al., 2021; Wang et al., 2023; 2024; Yang et al., 2023). Among these studies, Wang & Zou (2021) explored the robust variant of Q-learning algorithm under the R-contamination set and derived a sample complexity of $\widetilde{O}\left(\frac{SA}{(1-\gamma)^5\varepsilon^2}\right)$ in the discounted infinite-horizon setting, where γ represents the discount factor. However, this work does not explicitly provide the impact of robustness on the sample complexity.

Robust Markov games. Research on RMGs has explored robustness across various dimensions, including states (Han et al., 2022), environment (probabilistic transition kernels) (Shi et al., 2024c;a), and policies (Kannan et al., 2023). This work focuses on the robustness in environment.

The most relevant works to this study include Kardeş et al. (2011); Zhang et al. (2020b); Blanchet et al. (2024); Ma et al. (2023); Shi et al. (2024c;a); Gai & Shi (2026); Farhat et al. (2026). In particular, Kardeş et al. (2011) established the existence of NE in RMGs under mild conditions on the uncertainty set. Zhang et al. (2020a) demonstrated the asymptotic convergence of a Q-learning type algorithm under certain assumptions. Ma et al. (2023) proposed an algorithm for RMGs with a sample complexity of $\widetilde{O}\left(H^6 S \max_{1\le i\le m} A_i^2/\varepsilon^2\right)$ when the uncertainty level $R \le \varepsilon/(SH^2)$. This result implies that very little robustness can be tolerated when either H or S is large. In cases of $\varepsilon/(SH^2) < R \le p_{\min}/H$, where $p_{\min}$ represents the minimal nominal transition probability, the sample complexity is proved to be $\widetilde{O}\left(H^6 S \max_{1\le i\le m} A_i^2/(\varepsilon^2 p_{\min}^2)\right)$.

The algorithm proposed in Blanchet et al. (2024) can reach robust NEs with sample complexity $\widetilde{O}\left(H^4S^3\prod_{i=1}^m A_i^2/\varepsilon^2\right)$. Later, Shi et al. (2024c) achieved a better sample complexity of $\widetilde{O}\left(H^3S\prod_{i=1}^m A_i \min\left\{H,1/R\right\}/\varepsilon^2\right)$ for reaching a robust NE when $\varepsilon \in \left(0, \sqrt{\min\left\{H,1/R\right\}}\right]$, which implies a $\widetilde{O}(H^3S\prod_{i=1}^m A_i)$ burn-in cost. Shi et al. (2024a) achieved a sample complexity of $\widetilde{O}(H^6S\sum_{i=1}^m A_i \min\{H,1/R\}/\varepsilon^4)$ for any $\varepsilon \in \left(0, \sqrt{\min\left\{H,1/R\right\}}\right]$, which implies a $\widetilde{O}(H^5S\sum_{i=1}^m A_i \max\{1,HR\})$ burn-in cost. Recent work by Farhat et al. (2026) established a sample complexity of $\widetilde{O}\left(H^3S\prod_{i=1}^m A_i \min\{H,1/R\}/\varepsilon^2\right)$ in the online setting for arbitrary $R \in [0,1]$ under a failure states assumption, which requires the existence of an absorbing set of zero-reward states whose support is not escaped by the nominal transition kernel. Gai & Shi (2026) considered the linear function approximation setting and achieved a sample complexity of H^9d^3/ε^4 under a policy-induced uncertainty set, where d denotes the feature dimension. Specialized to the tabular setting, where $d = S\max_i A_i$, the sample complexity becomes $H^9S^3\max_i A_i^3/\varepsilon^4$.

B Notations

The notation $x = [x(s,a_i)]_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \in \mathbb{R}^{SA_i}$ (resp. $x = [x(s)]_{s\in\mathcal{S}} \in \mathbb{R}^S$) denotes a vector assigning values to each state-action pair for the i-th agent (resp. each state). The Hadamard product of two vectors x and y is denoted as $x\circ y = [x(s)\cdot y(s)]_{s\in\mathcal{S}}$ or as $x\circ y = [x(s,a_i)\cdot y(s,a_i)]_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}$. Notation $\mathsf{Var}_P(V) = P(V\circ V) - (PV)\circ(PV)$ denotes the variance of V following the distribution of P. Notation $\Delta(\mathcal{S})$ represents the probability simplex over the set $\mathcal{S}$. Moreover, with $\mathcal{X} := \{S, \{A_i\}_{i=1}^m, H, R, 1/\delta, 1/\varepsilon\}$. we also use $f(\mathcal{X}) = O(g(\mathcal{X}))$ or $f(\mathcal{X}) \lesssim g(\mathcal{X})$ to indicate that there exists a universal constant $C > 0$ such that $f(\mathcal{X}) \le Cg(\mathcal{X})$. Similarly $f(\mathcal{X}) \gtrsim g(\mathcal{X})$ indicates $g(\mathcal{X}) = O(f(\mathcal{X}))$. Additionally, the notation $\widetilde{O}(\cdot)$ is defined in the same way as $O(\cdot)$ except that it hides logarithmic factors.

Furthermore, for each $i \in [m]$, we denote $\mathcal{A}_{-i} := \prod_{j\neq i}\mathcal{A}_j$, which represents the action space of all players except the i-th one. We use notation $\boldsymbol{a} \in \mathcal{A}$ and $\boldsymbol{a}_{-i} \in \mathcal{A}_{-i}$ to denote a joint action of all players (or a joint action excluding the i-th player's action).

C Analysis

C.1 Notations and technical lemmas

For convenience, we restate some definitions from Section 4 and introduce additional notations. Define $\overline{V}$ as the estimation for V by replacing all expectation operation with the empirical estimation calculated according to $(r_{i,h}^k, P_{i,h}^k)$. Specifically, define

$$\overline{V}_{i,h}^{\widehat{\pi},R}(s) = \sum_{k=1}^K \alpha_k^K \mathbb{E}_{a_i\sim\pi_{i,h,s}^k}\left[r_{i,h}^k(s,a_i) + (1-R)P_{i,h}^k(s,a_i)\overline{V}_{i,h+1}^{\widehat{\pi},R} + R\min_{s\in\mathcal{S}}\overline{V}_{i,h+1}^{\widehat{\pi},R}\right], \quad (20)$$

where $P_{i,h}^k(s,a_i) \coloneqq P_{i,h}^k(\cdot|s,a_i) \in \mathbb{R}^S$ and $\overline{V}_{i,H+1}^{\widehat{\pi},R} = \overline{V}_{i,H+1}^{\star,\widehat{\pi}_{-i},R} = 0$. Note that the expectation here is taken only with respect to the action a_i, while the randomness in the actions of the other players $\boldsymbol{a}_{-i}$ is implicitly captured in the samples $r_{i,h}^k$ and $P_{i,h}^k$. Moreover, define $\widehat{\pi}_i^\star$ as the optimal policy (in the sense of achieving optimal worst-case performance) of the i-th agent when other agents adopt correlated policy $\widehat{\pi}_{-i}$:

$$\widehat{\pi}_i^\star = \{\widehat{\pi}_{i,h}^\star\}_{h=1}^H = \arg\max_{\pi_i:\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} V_{i,1}^{\pi_i\times\widehat{\pi}_{-i},R}. \quad (21)$$

By using Bellman operator, the corresponding value function estimation $\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}$ under policy $\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}$ is

$$\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}(s) = \sum_{k=1}^K \alpha_k^K \mathbb{E}_{a_i\sim\widehat{\pi}_{i,h,s}^\star}\left[r_{i,h}^k(s,a_i) + (1-R)P_{i,h}^k(s,a_i)\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R} + R\min_{s\in\mathcal{S}}\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right], \quad (22)$$

where with $\overline{V}_{i,H+1}^{\widehat{\pi}_i^\star,\widehat{\pi}_{-i},R}=0$, and $\widehat{\pi}_{i,h,s}^\star \coloneqq \widehat{\pi}_{i,h}^\star(\cdot|s)\in\Delta(\mathcal{A}_i)$. Define the expected reward and the nominal transition probability matrix under policy $\widehat{\pi}$ as

$$r_{i,h}^{\widehat{\pi}}(s)=\mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}}r_{i,h}(s,\boldsymbol{a})=\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_j\sim\pi_{j,h,s}^k}r_{i,h}(s,\boldsymbol{a}) \tag{23a}$$

$$=\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}\left[\mathbb{E}_{a_j\sim\pi_{j,h,s}^k,j\neq i}[r_{i,h}(s,\boldsymbol{a})|a_i]\right], \tag{23b}$$

$$P_{h,s}^{\widehat{\pi}}=\mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}}P_{h,s,\boldsymbol{a}}^0=\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_j\sim\pi_{j,h,s}^k}P_{h,s,\boldsymbol{a}}^0 \tag{23c}$$

$$=\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}\left[\mathbb{E}_{a_j\sim\pi_{j,h,s}^k,j\neq i}\left[P_{h,s,\boldsymbol{a}}^0|a_i\right]\right]. \tag{23d}$$

Different with definitions (20), here the expectation is taken with respect to all actions $\boldsymbol{a}$. Similarly, the expected reward and the nominal transition probability matrix under policy $\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}$ is denoted as

$$r_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}}(s)=\mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}}r_{i,h}(s,\boldsymbol{a}), \tag{24a}$$

$$P_h^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}}(s)=\mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}}P_{h,s,\boldsymbol{a}}^0. \tag{24b}$$

In addition, we define an empirical probability transition matrix $\widehat{P}_{i,h}^{\widehat{\pi}}\in\mathbb{R}^{S\times S}$ which is the weighted sum of sample result. Its s-th row is

$$\widehat{P}_{i,h,s}^{\widehat{\pi}}=\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}P_{i,h}^k(s,a_i). \tag{25}$$

Finally, define $P_h^{\widehat{\pi},V}\in\mathbb{R}^{S\times S}$(resp. $P_h^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},V}$, $\widehat{P}_{i,h}^{\widehat{\pi},V}$) as the transition matrix that achieves the worst-case performance when the value function at the $(h+1)$-stage is V. Particularly, its s-th row is defined as $P_{h,s}^{\widehat{\pi},V}\coloneqq\mathbb{E}_{\boldsymbol{a}\sim\widehat{\pi}_h(\cdot|s)}P_h^V(\cdot|s,\boldsymbol{a})$ with $P_h^V(\cdot|s,\boldsymbol{a})\coloneqq\arg\min_{P\in\mathcal{U}^R(P_{h,s,\boldsymbol{a}}^0)}PV$. One can easily verify that it satisfies

$$P_{h,s}^{\widehat{\pi},V}=(1-R)P_{h,s}^{\widehat{\pi}}+Re_{\arg\min V},\quad P_{h,s}^{\widehat{\pi}^\star\times\widehat{\pi}_{-i},V}=(1-R)P_{h,s}^{\widehat{\pi}^\star\times\widehat{\pi}_{-i}}+Re_{\arg\min V}, \tag{26a}$$

$$\widehat{P}_{i,h,s}^{\widehat{\pi},V}=(1-R)\widehat{P}_{i,h,s}^{\widehat{\pi}}+Re_{\arg\min V}, \tag{26b}$$

where $e_i\in\mathbb{R}^S$ denotes the i-th standard basis vector. Based on this notation, the estimation $\widehat{V}_{i,h}$ in (13) and the $\overline{V}_{i,h}^{\widehat{\pi},R}$ in (20) can be rewritten as

$$\widehat{V}_{i,h}(s)=\min\left\{\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}r_{i,h}^k(s,a_i)+\widehat{P}_{i,h,s}^{\widehat{\pi},\widehat{V}_{i,h+1}}\widehat{V}_{i,h+1}+\beta_{i,h}(s),H-h+1\right\}, \tag{27a}$$

$$\overline{V}_{i,h}^{\widehat{\pi},R}(s)=\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}r_{i,h}^k(s,a_i)+\widehat{P}_{i,h,s}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}}\overline{V}_{i,h+1}^{\widehat{\pi},R}. \tag{27b}$$

Before diving into the proof details, we first present a technical lemma that states some properties about $\widehat{V}_{i,h}$. Notice that $\widehat{V}_{i,h}$ depends on $\beta_{i,h}$, this lemma and the subsequent Lemma 3 are proved jointly via mathematical induction, which is deferred to Appendix D.1 in Supplementary Materials.

Lemma 1. *For any $i\in[m]$, $1\le h\le H$, $s\in\mathcal{S}$ and $1\le k\le K$, assume that K satisfies (15). We have*

$$\widehat{V}_{i,h}-\min_{s\in\mathcal{S}}\widehat{V}_{i,h}\le 3\sum_{h'=h}^{H}(1-R)^{h'-h}\le 3\min\left\{H,\frac{1}{R}\right\}, \tag{28a}$$

$$\mathsf{Var}_{\pi_{i,h,s}^k}(q_{i,h,s}^k)\lesssim\left(\min\left\{H,\frac{1}{R}\right\}\right)^2,\quad q_{i,h,s}^k-\min q_{i,h,s}^k\lesssim\min\left\{H,\frac{1}{R}\right\}. \tag{28b}$$

C.2 Proof of Theorem 1

According to the definition of $\mathsf{gap}_{\mathsf{CCE}}(\widehat{\pi})$ in (4), it suffices to prove that for all $i \in [m]$, the following inequality holds.

$$V_{i,1}^{\star,\widehat{\pi}_{-i},R}(s) - V_{i,1}^{\widehat{\pi},R}(s) \le \varepsilon, \quad \forall s \in \mathcal{S}.$$

Towards this, we decompose the above gap into several estimation errors, and the difference between the estimate and the variable $\widehat{V}_{i,h}$ in the algorithm:

$$\begin{aligned} & V_{i,h}^{\star,\widehat{\pi}_{-i},R} - V_{i,h}^{\widehat{\pi},R} \\ &= V_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R} - \overline{V}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R} + \overline{V}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R} - \widehat{V}_{i,h} + \widehat{V}_{i,h} - \overline{V}_{i,h}^{\widehat{\pi},R} + \overline{V}_{i,h}^{\widehat{\pi},R} - V_{i,h}^{\widehat{\pi},R}, \end{aligned} \tag{29}$$

where we remind the reader that $V_{i,h}^{\star,\widehat{\pi}_{-i},R} = V_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R}$ (cf. (21)). In the right-hand-side of (29), there are four error terms. Among them, the estimation errors $V_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R} - \overline{V}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R}$ and $\overline{V}_{i,h}^{\widehat{\pi},R} - V_{i,h}^{\widehat{\pi},R}$ share a similar structure, both reflecting the estimation error caused by the sampling of $\boldsymbol{a}_{-i}$ and previous policies. In the remainder of the proof, we analyze these two terms jointly, followed by a separate analysis of the remaining two error terms. We then combine the results to complete the analysis. Accordingly, the proof is organized into the following four steps.

Step 1. bounding $V_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R} - \overline{V}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R}$ and $\overline{V}_{i,h}^{\widehat{\pi},R} - V_{i,h}^{\widehat{\pi},R}$. We begin by analyzing the difference term $\overline{V}_{i,h}^{\widehat{\pi},R} - V_{i,h}^{\widehat{\pi},R}$. By the definition of $V_{i,h}^{\widehat{\pi},R}$ and the Bellman equation, this term can be expressed in terms of the expected reward $r_{i,h}^{\widehat{\pi}}$ (cf. (23a)) and the transition distribution $P_{h,s}^{\widehat{\pi}}$ under policy $\widehat{\pi}$ (cf. (23c)):

$$\begin{aligned} V_{i,h}^{\widehat{\pi},R}(s) &= \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{a_j \sim \pi_{j,h,s}^k} \left[r_{i,h}(s, \boldsymbol{a}) + (1-R) P_{h,s,\boldsymbol{a}}^0 V_{i,h+1}^{\widehat{\pi},R} + R \min V_{i,h+1}^{\widehat{\pi},R} \right] \\ &= r_{i,h}^{\widehat{\pi}}(s) + (1-R) P_{h,s}^{\widehat{\pi}} V_{i,h+1}^{\widehat{\pi},R} + R \min_{s \in \mathcal{S}} V_{i,h+1}^{\widehat{\pi},R} = r_{i,h}^{\widehat{\pi}}(s) + P_{h,s}^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}} V_{i,h+1}^{\widehat{\pi},R}, \end{aligned}$$

where the worst-case transition distribution $P_{h,s}^{\widehat{\pi},V}$ under policy $\widehat{\pi}$ is defined in (26a). Notice that due to sampling randomness, $\overline{V}_{i,h}^{\widehat{\pi},R}$ does not follow the Bellman equation under $r_{i,h}^{\widehat{\pi}}$ and $P_{h,s}^{\widehat{\pi}}$, which leads to a discrepancy with $V_{i,h}^{\widehat{\pi},R}(s)$. To control this difference recursively, we introduce the error between $\overline{V}_{i,h}^{\widehat{\pi},R}$ and the result of applying the Bellman operator to $\overline{V}_{i,h+1}^{\widehat{\pi},R}$:

$$\zeta_{i,h}^{\widehat{\pi},R} = \overline{V}_{i,h}^{\widehat{\pi},R} - r_{i,h}^{\widehat{\pi}} - P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \overline{V}_{i,h+1}^{\widehat{\pi},R}. \tag{30}$$

Thus we could derive that the difference between $\overline{V}_{i,h}^{\widehat{\pi},R}$ and $V_{i,h}^{\widehat{\pi},R}$ is the cumulative sum of $\zeta_{i,h}^{\widehat{\pi},R}$:

$$\begin{aligned} \overline{V}_{i,h}^{\widehat{\pi},R} - V_{i,h}^{\widehat{\pi},R} &= \zeta_{i,h}^{\widehat{\pi},R} + r_{i,h}^{\widehat{\pi}} + P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \overline{V}_{i,h+1}^{\widehat{\pi},R} - V_{i,h}^{\widehat{\pi},R} \\ &= \zeta_{i,h}^{\widehat{\pi},R} + P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \overline{V}_{i,h+1}^{\widehat{\pi},R} - P_h^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}} V_{i,h+1}^{\widehat{\pi},R} \\ &\overset{(i)}{\le} \zeta_{i,h}^{\widehat{\pi},R} + P_h^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}} \left(\overline{V}_{i,h+1}^{\widehat{\pi},R} - V_{i,h+1}^{\widehat{\pi},R} \right) \le \sum_{h'=h}^{H} \prod_{j=h}^{h'-1} P_j^{\widehat{\pi},V_{i,j+1}^{\widehat{\pi},R}} \zeta_{i,h'}^{\widehat{\pi},R}, \end{aligned} \tag{31}$$

where (i) uses the fact that $P_h^{\widehat{\pi},V} V \le P_h^{\widehat{\pi},V'} V$ for any pair of value functions V and V'. Similarly, the term $V_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R} - \overline{V}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R}$ can be decomposed by introducing the error term

$$\zeta_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R} = \overline{V}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R} - r_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}} - P_h^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R}} \overline{V}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i},R}, \tag{32}$$

where $r_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}}$ and $P_h^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, V}$ are defined in (24a) and (24b), respectively, and $P_h^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, V}$ is defined in (26a). Then the term $V_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} - \overline{V}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}$ is decomposed as

$$
\begin{aligned}
V_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} - \overline{V}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} &= -\zeta_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} + V_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} - r_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}} - P_h^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, \overline{V}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}} \overline{V}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} \\
&= -\zeta_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} + P_h^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, V_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}} V_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} \\
&\quad - P_h^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, \overline{V}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}} \overline{V}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} \\
&\le -\zeta_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} + P_h^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, \overline{V}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}} \left(V_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} - \overline{V}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} \right) \\
&= -\sum_{h'=h}^{H} \prod_{j=h}^{h'-1} P_j^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, \overline{V}_{i,j+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}} \zeta_{i,h'}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}.
\end{aligned}
$$

The following lemma provides upper bounds for the single error terms $\zeta_{i,h}^{\widehat{\pi}, R}$ and $\zeta_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}$. Its proof is deferred to Appendix D.2 in Supplementary Materials.

Lemma 2. *Assume that K satisfies (15). Consider error terms $\zeta_{i,h}^{\widehat{\pi}, R}$ and $\zeta_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}$ defined in (30) and (32), respectively. With probability at least $1-\delta$, the following inequalities show that the terms are controlled by $(P_h^{\widehat{\pi}, V} - I)(V \circ V)$, which has a favorable structure for summarization.*

$$
\begin{aligned}
\left| \zeta_{i,h}^{\widehat{\pi}, R} \right| &\le 8 c_\alpha \sqrt{\frac{\log^3 \frac{KS \sum_{i=1}^m A_i}{\delta}}{K \min\{H, \frac{1}{R}\}}} \left(P_h^{\widehat{\pi}, V_{i,h+1}^{\widehat{\pi}, R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi}, R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi}, R} \right) - \widetilde{\overline{V}}_{i,h}^{\widehat{\pi}, R} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi}, R} \right) \\
&\quad + 2c' \sqrt{\frac{\min\{H, \frac{1}{R}\} \log^3 \frac{KS \sum_{i=1}^m A_i}{\delta}}{K}}, \qquad (33a) \\
\left| \zeta_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} \right| &\le 8 c_\alpha \sqrt{\frac{\log^3 \frac{KS \sum_{i=1}^m A_i}{\delta}}{K \min\{H, \frac{1}{R}\}}} \left(P_h^{\widehat{\pi}, V_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} \right) \right. \\
&\quad \left. - \widetilde{\overline{V}}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi}_i^\star \times \widehat{\pi}_{-i}, R} \right) + 2c' \sqrt{\frac{\min\{H, \frac{1}{R}\} \log^3 \frac{KS \sum_{i=1}^m A_i}{\delta}}{K}}. \qquad (33b)
\end{aligned}
$$

Inequality (33a) implies that the summarization in the right-hand-side of (31) is bounded by

$$
\begin{aligned}
&\left\| \sum_{h=1}^{H} \prod_{j=1}^{h-1} P_j^{\widehat{\pi}, V_{i,j+1}^{\widehat{\pi}, R}} \zeta_{i,h}^{\widehat{\pi}, R} \right\|_\infty \le \sum_{h=1}^{H} \prod_{j=1}^{h-1} P_j^{\widehat{\pi}, V_{i,j+1}^{\widehat{\pi}, R}} \left| \zeta_{i,h}^{\widehat{\pi}, R} \right| \\
&\lesssim \sqrt{\frac{\log^3 \frac{KS \sum_{i=1}^m A_i}{\delta}}{K \min\{H, \frac{1}{R}\}}} \sum_{h=1}^{H} \prod_{j=1}^{h-1} P_j^{\widehat{\pi}, V_{i,j+1}^{\widehat{\pi}, R}} \left(P_h^{\widehat{\pi}, V_{i,h+1}^{\widehat{\pi}, R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi}, R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi}, R} \right) - \widetilde{\overline{V}}_{i,h}^{\widehat{\pi}, R} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi}, R} \right) \\
&\quad + H \sqrt{\frac{\min\{H, \frac{1}{R}\} \log^3 \frac{KS \sum_{i=1}^m A_i}{\delta}}{K}} \overset{\text{(i)}}{\lesssim} H \sqrt{\frac{\min\{H, \frac{1}{R}\} \log^3 \frac{KS \sum_{i=1}^m A_i}{\delta}}{K}},
\end{aligned}
$$

where (i) calculates the summarization by canceling the successive term as:

$$\begin{aligned}
&\sum_{h=1}^{H}\prod_{j=1}^{h-1} P_j^{\widehat{\pi},V_{i,j+1}^{\widehat{\pi},R}}\left(P_h^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}}\left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\circ\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right)-\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}\circ\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}\right)\\
&=\sum_{h=1}^{H}\prod_{j=1}^{h} P_j^{\widehat{\pi},V_{i,j+1}^{\widehat{\pi},R}}\left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\circ\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right)-\sum_{h=1}^{H}\prod_{j=1}^{h-1} P_j^{\widehat{\pi},V_{i,j+1}^{\widehat{\pi},R}}\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}\circ\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}\\
&=\sum_{h=2}^{H+1}\prod_{j=1}^{h-1} P_j^{\widehat{\pi},V_{i,j+1}^{\widehat{\pi},R}}\left(\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}\circ\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}\right)-\sum_{h=1}^{H}\prod_{j=1}^{h-1} P_j^{\widehat{\pi},V_{i,j+1}^{\widehat{\pi},R}}\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}\circ\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}=-\widetilde{\overline{V}}_{i,1}^{\widehat{\pi},R}\circ\widetilde{\overline{V}}_{i,1}^{\widehat{\pi},R}\le 0.
\end{aligned} \tag{34}$$

Substituting this result into (31) yields

$$\overline{V}_{i,1}^{\widehat{\pi},R}-V_{i,1}^{\widehat{\pi},R}\lesssim H\sqrt{\frac{\min\{H,\frac{1}{R}\}\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}. \tag{35a}$$

A similar inequality for $Vi, 1^{\widehat{\pi}i^\star\times\widehat{\pi}-i,R}$ can be derived similarly by applying (33b):

$$V_{i,1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}-\overline{V}_{i,1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\lesssim H\sqrt{\frac{\min\{H,\frac{1}{R}\}\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}. \tag{35b}$$

Step 2. bounding $\widehat{V}_{i,h}-\overline{V}_{i,h}^{\widehat{\pi},R}$. Comparing the definition of $\widehat{V}_{i,h}$ and $\overline{V}_{i,h}^{\widehat{\pi},R}$ (cf. (27a) and (27b), respectively), we observe that their difference arises from the inclusion of upper confidence bound $\beta_{i,h}$. Consequently, it is natural to express their difference as a cumulative sum of $\beta_{i,h}$.

$$\begin{aligned}
\widehat{V}_{i,h}-\overline{V}_{i,h}^{\widehat{\pi},R}&\le\widehat{P}_{i,h}^{\widehat{\pi},\widehat{V}_{i,h+1}}\widehat{V}_{i,h+1}-\widehat{P}_{i,h}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}}\overline{V}_{i,h+1}^{\widehat{\pi},R}+\beta_{i,h}\\
&\le\widehat{P}_{i,h}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}}\left(\widehat{V}_{i,h+1}-\overline{V}_{i,h+1}^{\widehat{\pi},R}\right)+\beta_{i,h}\le\sum_{h'=h}^{H}\prod_{j=h}^{h'-1}\widehat{P}_{i,j}^{\widehat{\pi},\overline{V}_{i,j+1}^{\widehat{\pi},R}}\beta_{i,h'},
\end{aligned} \tag{36}$$

where $\widehat{P}_{i,h}^{\widehat{\pi},V}$ is defined in (26b). The following lemma shows how to decompose $\beta_{i,h}$. Its proof is deferred to Section D.3 in Supplementary Materials.

Lemma 3. *Define $\widetilde{\widehat{V}}_{i,h+1}=\widehat{V}_{i,h+1}-\min_{s\in\mathcal{S}}\widehat{V}_{i,h+1}$. For $\beta_{i,h}$ defined in (12), we have*

$$\beta_{i,h}\le 2c_{\mathsf{b}}\sqrt{\frac{\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\left\{H,\frac{1}{R}\right\}}}\left(\widehat{P}_{i,h}^{\widehat{\pi},\widehat{V}_{i,h+1}}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)+\frac{3}{2}\min\left\{H,\frac{1}{R}\right\}\right). \tag{37a}$$

Moreover, if $\widetilde{\widehat{V}}_{i,h+1}\le 3\min\{H,\frac{1}{R}\}$, then we have

$$\begin{aligned}
\beta_{i,h}\le{}&4c_{\mathsf{b}}\sqrt{\frac{\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\left\{H,\frac{1}{R}\right\}}}\left((1-R)\widehat{P}_{i,h}^{\widehat{\pi}}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)-\widetilde{\widehat{V}}_{i,h}\circ\widetilde{\widehat{V}}_{i,h}\right)\\
&+14c_{\mathsf{b}}\sqrt{\frac{\min\left\{H,\frac{1}{R}\right\}\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}.
\end{aligned} \tag{37b}$$

Lemma 1 implies that (37b) holds. Recalling the definition of $\widehat{P}_{i,h}^{\widehat{\pi},V}$, we have $\widehat{P}_{i,h}^{\widehat{\pi},V}V\le\widehat{P}_{i,h}^{\widehat{\pi},V'}V$ for any $V,V'\in\mathbb{R}^S$. Thus we have the following upper bound of $\beta_{i,h}$.

$$\beta_{i,h}\le 4c_{\mathsf{b}}\sqrt{\frac{\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\left\{H,\frac{1}{R}\right\}}}\left[\left(\widehat{P}_{i,h}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)-\widetilde{\widehat{V}}_{i,h}\circ\widetilde{\widehat{V}}_{i,h}\right)+\frac{7}{2}\min\left\{H,\frac{1}{R}\right\}\right].$$

By substituting it into (36), the difference $\widehat{V}_{i,h} - \overline{V}_{i,h}^{\widehat{\pi},R}$ can be bounded using a similar summation as in (31):

$$\begin{aligned}\widehat{V}_{i,h} - \overline{V}_{i,h}^{\widehat{\pi},R} \le 4c_{\mathrm{b}}\sqrt{\frac{\log^3 \frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\left\{H, \frac{1}{R}\right\}}} \sum_{h'=h}^{H}\prod_{j=h}^{h'-1} \widehat{P}_{i,j}^{\widehat{\pi},\overline{V}_{i,j+1}^{\widehat{\pi},R}}\left(\widehat{P}_{i,h'}^{\widehat{\pi},\overline{V}_{i,h'+1}^{\widehat{\pi},R}}\left(\widetilde{\widehat{V}}_{i,h'+1}\circ\widetilde{\widehat{V}}_{i,h'+1}\right)\right. \\ \left. - \widetilde{\widehat{V}}_{i,h'}\circ\widetilde{\widehat{V}}_{i,h'}\right) + 14c_{\mathrm{b}}H\sqrt{\frac{\min\left\{H,\frac{1}{R}\right\}\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}} \\ \overset{(i)}{\lesssim} H\sqrt{\frac{\min\left\{H,\frac{1}{R}\right\}\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}, \end{aligned} \tag{38}$$

where (i) uses the cancellation between successive terms, similar to (34):

$$\begin{aligned}&\sum_{h'=h}^{H}\prod_{j=h}^{h'-1} \widehat{P}_{i,j}^{\widehat{\pi},\overline{V}_{i,j+1}^{\widehat{\pi},R}}\left(\widehat{P}_{i,h'}^{\widehat{\pi},\overline{V}_{i,h'+1}^{\widehat{\pi},R}}\left(\widetilde{\widehat{V}}_{i,h'+1}\circ\widetilde{\widehat{V}}_{i,h'+1}\right) - \widetilde{\widehat{V}}_{i,h'}\circ\widetilde{\widehat{V}}_{i,h'}\right) \\ &= \sum_{h'=h}^{H}\prod_{j=h}^{h'} \widehat{P}_{i,j}^{\widehat{\pi},\overline{V}_{i,j+1}^{\widehat{\pi},R}}\left(\widetilde{\widehat{V}}_{i,h'+1}\circ\widetilde{\widehat{V}}_{i,h'+1}\right) - \sum_{h'=h}^{H}\prod_{j=h}^{h'-1} \widehat{P}_{i,j}^{\widehat{\pi},\overline{V}_{i,j+1}^{\widehat{\pi},R}}\widetilde{\widehat{V}}_{i,h'}\circ\widetilde{\widehat{V}}_{i,h'} \\ &= -\widetilde{\widehat{V}}_{i,h}\circ\widetilde{\widehat{V}}_{i,h} \le 0.\end{aligned}$$

Step 3. bounding $\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R} - \widehat{V}_{i,h}$. Given the sampled reward $r_{i,h}^k$ and transition kernel $P_{i,h}^k$, the value function $\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}$ satisfies a Bellman equation in single-agent reinforcement learning (cf. (22)), with reward $r_{i,h}^k$, transition kernel $P_{i,h}^k$, and policy $\widehat{\pi}_i$. We introduce the following value function, which corresponds to the optimal value function in the above robust Markov Decision Process:

$$\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}(s) = \max_{a_i\in\mathcal{A}_i}\sum_{k=1}^{K}\alpha_k^K\left[r_{i,h}^k(s,a_i) + (1-R)P_{i,h}^k(s,a_i)\overline{V}_{i,h+1}^{\star,\widehat{\pi}_{-i},R} + R\min_{s\in\mathcal{S}}\overline{V}_{i,h+1}^{\star,\widehat{\pi}_{-i},R}\right]. \tag{39}$$

The following lemma reveals that $\widehat{V}_{i,h}$ is an optimism estimation of $\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}$ due to the inclusion of the upper confidence bound. Its proof is deferred to Section D.4 in the Supplementary Materials.

Lemma 4. *For $\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}$ defined in (39), and K satisfying (15), with probability at least $1-\delta$, we have for all $i\in[m]$ and $1\le h\le H$,*

$$\widehat{V}_{i,h} \ge \overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}. \tag{40}$$

Recalling that $\overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}$ is the optimal value function, we obviously have $\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R} \le \overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R}$. Then Lemma 4 gives that

$$\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R} - \widehat{V}_{i,h} \le \overline{V}_{i,h}^{\star,\widehat{\pi}_{-i},R} - \widehat{V}_{i,h} \le 0. \tag{41}$$

Step 4. combining the above results. Inserting (35a), (35b), (38), and (41) into (29), we have

$$V_{i,h}^{\star,\widehat{\pi}_{-i},R} - V_{i,h}^{\widehat{\pi},R} \le CH\sqrt{\frac{\min\left\{H,\frac{1}{R}\right\}\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}} \le \varepsilon, \tag{42}$$

as long as K satisfying (15), where C is a sufficiently large constant. Thus we complete the proof.

D PROOF OF TECHNICAL LEMMAS

Before diving into the proof details, we first present the following lemma, which states some properties about step-size α_k.

Lemma 5. *(Li et al., 2022, Lemma 1) Assume that $c_\alpha \geq 24$, and $K \geq c_\alpha \log K + 1$. For any $k \geq 1$, we have*

$$\sum_{k=1}^{K} \alpha_k^K = 1, \qquad \max_{1\leq k\leq K} \alpha_k^K \leq \frac{2c_\alpha \log K}{K}, \qquad \max_{1\leq k\leq K/2} \alpha_k^K \leq \frac{1}{K^6}.$$

D.1 Proof of Lemma 1

Proof of (28a). The key of this proof is to establish (28a) by mathematical induction. When $h = H + 1$, (28a) holds obviously. Assume that (28a) holds for all $h' \geq h + 1$. Now we intend to prove that (28a) holds for h. To this end, we make the following observation.

$$\begin{aligned}
\widetilde{\widehat{V}}_{i,h} &= \widehat{V}_{i,h} - \min \widehat{V}_{i,h} \leq 1 + (1-R)\widehat{P}_{i,h}^{\widehat{\pi}} \widehat{V}_{i,h+1} + R \min \widehat{V}_{i,h+1} - \min \widehat{V}_{i,h} + \beta_{i,h} \\
&\overset{(i)}{\leq} 1 + (1-R)\widehat{P}_{i,h}^{\widehat{\pi}} \widetilde{\widehat{V}}_{i,h+1} + \beta_{i,h} \\
&\leq \sum_{h'=h}^{H} (1-R)^{h'-h} \prod_{j=h}^{h'-1} \widehat{P}_{i,j}^{\widehat{\pi}} \left(1 + \beta_{i,h'}\right) \\
&= \sum_{h'=h}^{H} (1-R)^{h'-h} + \sum_{h'=h}^{H} (1-R)^{h'-h} \prod_{j=h}^{h'-1} \widehat{P}_{i,j}^{\widehat{\pi}} \beta_{i,h'},
\end{aligned} \tag{43}$$

where (i) uses the fact that the minimal value of $\widehat{V}_{i,h}$ increases as h decrease, because a new positive reward is added:

$$\begin{aligned}
\min \widehat{V}_{i,h} &\geq \min \left\{ (1-R) \min \widehat{V}_{i,h+1} + R \min \widehat{V}_{i,h+1} + \beta_{i,h}, H - h + 1 \right\} \\
&\geq \min \left\{ \min \widehat{V}_{i,h+1} + \beta_{i,h}, H - h + 1 \right\} \geq \min \widehat{V}_{i,h+1}.
\end{aligned}$$

For ease of notations, we define the variable

$$\rho_0 = c_{\mathsf{b}} \sqrt{\frac{\log^3 \frac{KS\sum_{i=1}^m A_i}{\delta}}{K \min\left\{H, \frac{1}{R}\right\}}}.$$

With this definition, the upper bound on $\beta_{i,h}$ in (37a) of Lemma 3 simplifies to

$$\begin{aligned}
\beta_{i,h} &\leq 2\rho_0 \widehat{P}_{i,h}^{\widehat{\pi}, \widehat{V}_{i,h+1}} \left(\widetilde{\widehat{V}}_{i,h+1} \circ \widetilde{\widehat{V}}_{i,h+1} \right) + 3\rho_0 \min\left\{H, \frac{1}{R}\right\} \\
&= 2\rho_0 (1-R) \widehat{P}_{i,h}^{\widehat{\pi}} \left(\widetilde{\widehat{V}}_{i,h+1} \circ \widetilde{\widehat{V}}_{i,h+1} \right) + 3\rho_0 \min\left\{H, \frac{1}{R}\right\},
\end{aligned}$$

Recalling our assumption that $\widetilde{\widehat{V}}_{i,h'} \leq 3\min\{H, \frac{1}{R}\}$ for all $h' \geq h + 1$, Lemma 3 implies that a stronger conclusion, given by (37b), holds in this case. This result can be further simplified using the newly introduced notation ρ_0, leading to the following inequality:

$$\beta_{i,h'} \leq 4\rho_0 \left((1-R) \widehat{P}_{i,h'}^{\widehat{\pi}} \left(\widetilde{\widehat{V}}_{i,h'+1} \circ \widetilde{\widehat{V}}_{i,h'+1} \right) - \widetilde{\widehat{V}}_{i,h'} \circ \widetilde{\widehat{V}}_{i,h'} \right) + 14\rho_0 \min\left\{H, \frac{1}{R}\right\}, \quad h' \geq h+1,$$

We decompose the sum $\sum_{h'=h}^{H} (1-R)^{h'-h} \prod_{j=h}^{h'-1} \widehat{P}_{i,j}^{\widehat{\pi}} \beta_{i,h'}$ into two parts: the term corresponding to $h' = h$ and the remaining terms with $h' > h$. We then apply the two bounds for $\beta_{i,h}$ separately,

obtaining:

$$
\begin{aligned}
&\sum_{h'=h}^{H}(1-R)^{h'-h}\prod_{j=h}^{h'-1}\widehat{P}^{\widehat{\pi}}_{i,j}\beta_{i,h'} \\
&\le 4\rho_0\sum_{h'=h+1}^{H}(1-R)^{h'-h}\prod_{j=h}^{h'-1}\widehat{P}^{\widehat{\pi}}_{i,j}\left((1-R)\widehat{P}^{\widehat{\pi}}_{i,h'}\left(\widetilde{\widehat{V}}_{i,h'+1}\circ\widetilde{\widehat{V}}_{i,h'+1}\right)-\widetilde{\widehat{V}}_{i,h'}\circ\widetilde{\widehat{V}}_{i,h'}\right) \\
&\quad+2\rho_0(1-R)\widehat{P}^{\widehat{\pi}}_{i,h}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)+14\rho_0\min\left\{H,\frac{1}{R}\right\}\sum_{h'=h+1}^{H}(1-R)^{h'-h}+3\rho_0\min\left\{H,\frac{1}{R}\right\} \\
&\overset{(i)}{\le}-4\rho_0(1-R)\widehat{P}^{\widehat{\pi}}_{i,h}\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}+2\rho_0(1-R)\widehat{P}^{\widehat{\pi}}_{i,h}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right) \\
&\quad+14\rho_0\min\left\{H,\frac{1}{R}\right\}\sum_{h'=h}^{H}(1-R)^{h'-h}\le 14\rho_0\min\left\{H,\frac{1}{R}\right\}\sum_{h'=h}^{H}(1-R)^{h'-h},
\end{aligned}
\tag{44}
$$

where (i) absorbs the term $3\rho_0\min\left\{H,\frac{1}{R}\right\}$ into the sum of $14\rho_0\min\left\{H,\frac{1}{R}\right\}\sum_{h'=h}^{H}(1-R)^{h'-h}$, and it uses the fact that the following sum is negative due to cancellation between successive terms:

$$
\begin{aligned}
&\sum_{h'=h+1}^{H}(1-R)^{h'-h}\prod_{j=h}^{h'-1}\widehat{P}^{\widehat{\pi}}_{i,j}\left((1-R)\widehat{P}^{\widehat{\pi}}_{i,h'}\left(\widetilde{\widehat{V}}_{i,h'+1}\circ\widetilde{\widehat{V}}_{i,h'+1}\right)-\widetilde{\widehat{V}}_{i,h'}\circ\widetilde{\widehat{V}}_{i,h'}\right) \\
&=\sum_{h'=h+1}^{H}(1-R)^{h'-h+1}\prod_{j=h}^{h'}\widehat{P}^{\widehat{\pi}}_{i,j}\widetilde{\widehat{V}}_{i,h'+1}\circ\widetilde{\widehat{V}}_{i,h'+1}-\sum_{h'=h+1}^{H}(1-R)^{h'-h}\prod_{j=h}^{h'-1}\widehat{P}^{\widehat{\pi}}_{i,j}\widetilde{\widehat{V}}_{i,h'}\circ\widetilde{\widehat{V}}_{i,h'} \\
&=\sum_{h'=h+2}^{H+1}(1-R)^{h'-h}\prod_{j=h}^{h'-1}\widehat{P}^{\widehat{\pi}}_{i,j}\widetilde{\widehat{V}}_{i,h'}\circ\widetilde{\widehat{V}}_{i,h'}-\sum_{h'=h+1}^{H}(1-R)^{h'-h}\prod_{j=h}^{h'-1}\widehat{P}^{\widehat{\pi}}_{i,j}\widetilde{\widehat{V}}_{i,h'}\circ\widetilde{\widehat{V}}_{i,h'} \\
&=-(1-R)\widehat{P}^{\widehat{\pi}}_{i,h+1}\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}.
\end{aligned}
$$

Inserting (44) into (43), the range of $\widehat{V}_{i,h}$ is bounded by

$$
\widetilde{\widehat{V}}_{i,h}\le\sum_{h'=h}^{H}(1-R)^{h'-h}+14\rho_0\min\left\{H,\frac{1}{R}\right\}\sum_{h'=h}^{H}(1-R)^{h'-h}\le 3\sum_{h'=h}^{H}(1-R)^{h'-h},
$$

as long as $14\rho_0\min\left\{H,\frac{1}{R}\right\}\le 2$ for K in (15). Thus we complete the proof of (28a).

Proof of (28b). With (28a) established, now we are ready to prove (28b). According to the definition of $q^k_{i,h,s}$, a trivial bound for $q^k_{i,h,s}$ is that

$$
\min q^k_{i,h,s}\ge\min q^k_{i,h}\ge(1-R)P^k_{i,h}\widehat{V}_{i,h+1}+R\min\widehat{V}_{i,h+1}\ge\min\widehat{V}_{i,h+1}.
$$

Based on the computation of $q^k_{i,h,s}$, we use the range of $\widehat{V}_{i,h+1}$ to bound its range as

$$
\begin{aligned}
q^k_{i,h,s}-\min q^k_{i,h,s}&\le q^k_{i,h,s}-\min q^k_{i,h}\le 1+(1-R)P^k_{i,h}\left(\widehat{V}_{i,h+1}-\min\widehat{V}_{i,h+1}\right) \\
&=1+(1-R)P^k_{i,h}\widetilde{\widehat{V}}_{i,h+1}\le 1+3(1-R)\sum_{h'=h+1}^{H}(1-R)^{h'-h-1} \\
&\le 3\sum_{h'=h}^{H}(1-R)^{h'-h}\le 3\min\left\{H,\frac{1}{R}\right\}.
\end{aligned}
$$

Based on this result, it is obviously that $\mathsf{Var}_{\pi^k_{i,h,s}}(q^k_{i,h,s})\le 9\min\left\{H,\frac{1}{R}\right\}^2$. Thus we complete the proof.

D.2 Proof of Lemma 2

Before refined analysis, we first demonstrate that due to the introduction of robustness, the estimate $\overline{V}_{i,h+1}^{\widehat{\pi},R}$ and $\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}$ have an upper bound related to R. Notice that according to the definition, the range of $\overline{V}_{i,h+1}^{\widehat{\pi},R}$ bounded by

$$\begin{aligned}\overline{V}_{i,h}^{\widehat{\pi},R}(s)-\min\overline{V}_{i,h}^{\widehat{\pi},R}&\le 1+(1-R)\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}P_{i,h}^k(s,a_i)\overline{V}_{i,h+1}^{\widehat{\pi},R}+R\min\overline{V}_{i,h+1}^{\widehat{\pi},R}-\overline{V}_{i,h}^{\widehat{\pi},R}\\&\le 1+(1-R)\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}P_{i,h}^k(s,a_i)\left(\overline{V}_{i,h+1}^{\widehat{\pi},R}-\min\overline{V}_{i,h+1}^{\widehat{\pi},R}\right),\end{aligned}\tag{45}$$

where the last inequality holds because the minimal value increases as h decrease due to the addition of a new positive reward:

$$\min\overline{V}_{i,h}^{\widehat{\pi},R}\ge\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}(1-R)P_{i,h}^k(s,a_i)\overline{V}_{i,h+1}^{\widehat{\pi},R}+R\min\overline{V}_{i,h+1}^{\widehat{\pi},R}\ge\min\overline{V}_{i,h+1}^{\widehat{\pi},R}.\tag{46}$$

Recalling the definition of $\widehat{P}_{i,h}^{\widehat{\pi}}$ in (25), inequality (45) can be rewritten in terms of $\widehat{P}_{i,h}^{\widehat{\pi}}$. By recursion, this yields a bound that depends on H and R:

$$\overline{V}_{i,h}^{\widehat{\pi},R}-\min\overline{V}_{i,h}^{\widehat{\pi},R}\le 1+(1-R)\widehat{P}_{i,h}^{\widehat{\pi}}\left(\overline{V}_{i,h+1}^{\widehat{\pi},R}-\min\overline{V}_{i,h+1}^{\widehat{\pi},R}\right)\le\sum_{h'=h}^{H}(1-R)^{h'-h}.\tag{47}$$

Similarly, by replacing $\widehat{\pi}$ with $\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}$, we have bound the range of $\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}$ as

$$\begin{aligned}\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}-\min\overline{V}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}&\le 1+(1-R)\widehat{P}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}}\left(\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}-\min\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)\\&\le\sum_{h'=h}^{H}(1-R)^{h'-h},\end{aligned}\tag{48}$$

where the s-th row of matrix $\widehat{P}_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}}\in\mathbb{R}^{S\times S}$ is defined in a similar way as $\widehat{P}_{i,h}^{\widehat{\pi}}$, i.e.,

$$\widehat{P}_{i,h,s}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}}=\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^\star}P_{i,h}^k(s,a_i).$$

Next, we shall focus on the proof for (33a). Recalling that $\overline{V}_{i,h}^{\widehat{\pi},R}$ satisfies the Bellman equation for a single-agent RL problem with empirical reward $r_{i,h}^k$ and transition kernel $P_{i,h}^k$ under the policy $\widehat{\pi}_{i,h}^k$, and that $\zeta_{i,h}^{\widehat{\pi},R}$ characterizes its deviation from the Bellman equation in RMGs under policy $\widehat{\pi}_h$:

$$\begin{aligned}\zeta_{i,h}^{\widehat{\pi},R}(s)&=\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}\left[r_{i,h}^k(s,a_i)+P_{i,h}^{k,\overline{V}_{i,h+1}^{\widehat{\pi},R}}(s,a_i)\overline{V}_{i,h+1}^{\widehat{\pi},R}\right]-r_{i,h}^{\widehat{\pi}}-P_{h,s}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}}\overline{V}_{i,h+1}^{\widehat{\pi},R}\\&\overset{(i)}{=}\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{a_i\sim\pi_{i,h,s}^k}\left[r_{i,h}^k(s,a_i)-\mathbb{E}_{k-1}r_{i,h}^k(s,a_i)+(1-R)\left(P_{i,h}^k-\mathbb{E}_{k-1}P_{i,h}^k\right)(s,a_i)\overline{V}_{i,h+1}^{\widehat{\pi},R}\right]\end{aligned}\tag{49}$$

where $P_{i,h}^{k,V}(s,a_i)\in\mathbb{R}^S$ is defined as

$$P_{i,h}^{k,V}(s,a_i)=(1-R)P_{i,h}^k(s,a_i)+Re_{\arg\min V}.$$

Here, equation (i) uses the definition of $r_{i,h}^{\widehat{\pi}}$ and $P_h^{\widehat{\pi}}$ in (23b) and (23d), respectively, and $\mathbb{E}_{k-1}[\cdot]$ denotes the expectation with history policies $\pi_{i,h}^j$ for $j\le k$ conditioned. Next we intend to use Freedman's inequality to bound this error term, which is stated as below.

Lemma 6. *(Li et al., 2022, Theorem 5) Suppose that $Y_n = \sum_{k=1}^n X_k \in \mathbb{R}$, where $\{X_k\}$ is a real-valued scalar sequence obeying*

$$|X_k| \le R, \quad \text{and} \quad \mathbb{E}[X_k|\{X_j\}_{j\le k}] = 0 \quad \forall k \ge 1$$

for some quantity $R > 0$. Define

$$W_n := \sum_{k=1}^n \mathbb{E}_{k-1}[X_k^2],$$

where $\mathbb{E}_{k-1}[\cdot]$ stands for the expectation conditional on $\{X_j\}_{j<k}$. Consider any arbitrary quantity $\kappa > 0$. With probability at least $1-\delta$, one has

$$|Y_n| \le \sqrt{8W_n \log\frac{3n}{\delta}} + 5R\log\frac{3n}{\delta} \le \kappa W_n + \left(\frac{2}{\kappa} + 5R\right)\log\frac{3n}{\delta}.$$

To apply Freedman's inequality, we define the random variable

$$\begin{aligned} X_k =& \alpha_k^K \mathbb{E}_{a_i\sim\pi_{i,h,s}^k}\left[r_{i,h}^k(s,a_i) - \mathbb{E}_{k-1}r_{i,h}^k(s,a_i) + (1-R)\left(P_{i,h}^k - \mathbb{E}_{k-1}P_{i,h}^k\right)(s,a_i)\overline{V}_{i,h+1}^{\widehat{\pi},R}\right] \\ \overset{(i)}{=}& \alpha_k^K \mathbb{E}_{a_i\sim\pi_{i,h,s}^k}\left[r_{i,h}^k(s,a_i) - \mathbb{E}_{k-1}r_{i,h}^k(s,a_i) + (1-R)\left(P_{i,h}^k - \mathbb{E}_{k-1}P_{i,h}^k\right)(s,a_i)\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right], \end{aligned}$$

where $\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} = \overline{V}_{i,h+1}^{\widehat{\pi},R} - \min\overline{V}_{i,h+1}^{\widehat{\pi},R}$, and (i) holds because equation $\left(P_{i,h}^k - \mathbb{E}_{k-1}P_{i,h}^k\right)\min\overline{V}_{i,h+1}^{\widehat{\pi},R} = 0$, as both $P_{i,h}^k$ and its expectation are transition kernels. It is obvious that X_k has zero mean, i.e., $\mathbb{E}_{k-1}[X_k] = 0$. A trivial bound for $|X_k|$ is given by

$$\begin{aligned} |X_k| &\le \alpha_k^K\left(1 + (1-R)\left(\overline{V}_{i,h+1}^{\widehat{\pi},R} - \min\overline{V}_{i,h+1}^{\widehat{\pi},R}\right)\right) \\ &\le \max_k \alpha_k^K \min\left\{H, \frac{1}{R}\right\} \overset{(i)}{\le} \frac{2c_\alpha\log K}{K}\min\left\{H,\frac{1}{R}\right\}, \end{aligned}$$

where (i) uses the property of α_k^K in Lemma 5. Moreover, the second order momentum is given by

$$\begin{aligned} W_K &= \sum_{k=1}^K \mathbb{E}_{k-1}X_k^2 = \sum_{k=1}^K (\alpha_k^K)^2 \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k) \\ &\le \max_k \alpha_k^K \sum_{k=1}^K \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k) \le \frac{2c_\alpha\log K}{K}\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k), \end{aligned}$$

where $\mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k)$ denotes the variance of random variable $\widetilde{q}_{i,h,s}^k$ conditioned on $\pi_{i,h}^j$ with $j \le k$, and

$$\widetilde{q}_{i,h,s}^k := \mathbb{E}_{a_i\sim\pi_{i,h,s}^k}\left[r_{i,h}^k(s,a_i) + (1-R)P_{i,h}^k(s,a_i)\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right]. \tag{50}$$

Now we are ready to apply Freedman's inequality (c.f. Lemma 6). By taking $\kappa = \sqrt{\frac{K\log\frac{KS\sum_{i=1}^m A_i}{\delta}}{\min\{H,\frac{1}{R}\}}}$, we have

$$\begin{aligned} \left|\zeta_{i,h}^{\widehat{\pi},R}\right| &\le 2c_\alpha\sqrt{\frac{\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\{H,\frac{1}{R}\}}\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k)} \\ &\quad + \left(2\sqrt{\frac{\min\{H,\frac{1}{R}\}}{K\log\frac{KS\sum_{i=1}^m A_i}{\delta}}} + \frac{10c_\alpha\min\{H,\frac{1}{R}\}\log K}{K}\right)\log\frac{3KS\sum_{i=1}^m A_i}{\delta} \\ &\le 2c_\alpha\sqrt{\frac{\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\{H,\frac{1}{R}\}}\sum_{k=1}^K \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k)} + c\sqrt{\frac{\min\{H,\frac{1}{R}\}\log\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}, \end{aligned} \tag{51}$$

where c is a constant. Now we focus on the term $\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k)$. Recalling that the variance arises from both the reward $r_{i,h}^k$ and the transition kernel $P_{i,h}^k$ (cf. (50)), we handle these two sources separately by introducing the following decomposition:

$$\begin{aligned}\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k) &\le 2\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{k-1}(r_{i,h,s}^k) + 2\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k - r_{i,h,s}^k) \\ &\overset{(i)}{\le} 2 + 2\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k - r_{i,h,s}^k), \end{aligned} \tag{52}$$

where $r_{i,h,s}^k := \mathbb{E}_{a_i \sim \pi_{i,h,s}^k} r_{i,h}^k(s, a_i)$, and (i) holds because $r_{i,h,s}^k$ is bounded by $|r_{i,h,s}^k| \le 1$. Furthermore, the variance of $\widetilde{q}_{i,h,s}^k - r_{i,h,s}^k$ is decomposed as:

$$\begin{aligned}&\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k - r_{i,h,s}^k) \\ &= \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{k-1}(\widetilde{q}_{i,h,s}^k - r_{i,h,s}^k)^2 - \sum_{k=1}^{K} \alpha_k^K \left(\mathbb{E}_{k-1}(\widetilde{q}_{i,h,s}^k - r_{i,h,s}^k)\right)^2. \end{aligned} \tag{53}$$

Below we shall handle these two terms separately.

- For the first term, which is the expectation of square, we use the Cauchy-Schwarz inequality and get

$$\begin{aligned}\sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{k-1}(\widetilde{q}_{i,h,s}^k - r_{i,h,s}^k)^2 &= \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{k-1}\left[\mathbb{E}_{a_i \sim \pi_{i,h,s}^k}\left[(1-R)P_{i,h}^k(s,a_i)\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right]\right]^2 \\ &\le \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{k-1, a_i \sim \pi_{i,h,s}^k}\left[(1-R)P_{i,h}^k(s,a_i)\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right]^2 \\ &\overset{(i)}{\le} (1-R)^2 P_{h,s}^{\widehat{\pi}}\left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) \\ &\le P_{h,s}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}}\left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right), \end{aligned} \tag{54a}$$

  where (i) uses the fact that each row of $P_{i,h}^k(s,a_i)$ has only a non-zero value, and the vector $P_{h,s}^{\widehat{\pi}} \in \mathbb{R}^S$ is defined in (23c).

- For the second term, which is the square of the expectation, we use the Jensen's inequality to compute a lower bound as follows:

$$\begin{aligned}\sum_{k=1}^{K} \alpha_k^K \left(\mathbb{E}_{k-1}(\widetilde{q}_{i,h,s}^k - r_{i,h,s}^k)\right)^2 &\ge (1-R)^2 \sum_{k=1}^{K} \alpha_k^K \left(\mathbb{E}_{k-1, a_i \sim \pi_{i,h,s}^k} P_{i,h}^k(s,a_i)\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right)^2 \\ &\overset{(i)}{\ge} (1-R)^2 \left(\sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{k-1, a_i \sim \pi_{i,h,s}^k} P_{i,h}^k(s,a_i)\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right)^2 \\ &= (1-R)^2 \left(P_{h,s}^{\widehat{\pi}} \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right)^2 = \left(P_{h,s}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right)^2. \end{aligned} \tag{54b}$$

Combining (54a), (54b), and (53), we have

$$\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k - r_{i,h,s}^k) \le P_{h,s}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}}\left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) - \left(P_{h,s}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right)^2. \tag{55}$$

Notice that this result is not exactly what we aim for, as it involves the transition kerne $P_{h,s}^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}}$ rather than $P_{h,s}^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}}$ as required in Lemma 2. Fortunately, $\overline{V}_{i,h+1}^{\widehat{\pi},R}$ is a good approximation of $V_{i,h+1}^{\widehat{\pi},R}$, so the associated transition kernels should be close. To formalize this intuition, we define the error term:

$$\vartheta_{i,h}^{\widehat{\pi}} := \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} - P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}. \tag{56}$$

According to its definition, this error term is upper bounded by

$$\begin{aligned}
\vartheta_{i,h}^{\widehat{\pi}} &= \overline{V}_{i,h}^{\widehat{\pi},R} - P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \overline{V}_{i,h+1}^{\widehat{\pi},R} - \min \overline{V}_{i,h}^{\widehat{\pi},R} + \min \overline{V}_{i,h+1}^{\widehat{\pi},R} \\
&\overset{(i)}{\leq} \overline{V}_{i,h}^{\widehat{\pi},R} - P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \overline{V}_{i,h+1}^{\widehat{\pi},R} \\
&\overset{(ii)}{\leq} r_{i,h}^{\widehat{\pi}} + \zeta_{i,h}^{\widehat{\pi},R} \leq 1 + \left|\zeta_{i,h}^{\widehat{\pi},R}\right|.
\end{aligned}$$

where (i) arises from the fact that $\min \overline{V}_{i,h}^{\widehat{\pi},R} \geq \min \overline{V}_{i,h+1}^{\widehat{\pi},R}$ as shown in (46), and (ii) arises from the definition of $\zeta_{i,h}^{\widehat{\pi},R}$ in (30). By applying the definition (56), we have

$$\begin{aligned}
&P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) - \left(P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) \circ \left(P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) \\
&= P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) - \left(\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} - \vartheta_{i,h}^{\widehat{\pi}}\right) \circ \left(\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} - \vartheta_{i,h}^{\widehat{\pi}}\right) \\
&\leq P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) - \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} + 2\vartheta_{i,h}^{\widehat{\pi}} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} \\
&\leq P_h^{\widehat{\pi},\overline{V}_{i,h+1}^{\widehat{\pi},R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) - \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} + 2\min\left\{H, \frac{1}{R}\right\}\left(1 + \left|\zeta_{i,h}^{\widehat{\pi},R}\right|\right) \\
&\leq P_h^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) - \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} + 2\min\left\{H, \frac{1}{R}\right\}\left(1 + \left|\zeta_{i,h}^{\widehat{\pi},R}\right|\right).
\end{aligned} \tag{57}$$

Now this bound depends on the transition kernel $P_h^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}}$. Inserting (57) into (55) and (52), we have

$$\begin{aligned}
\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s}^k) \leq 2P_h^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) - 2\widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} \\
+ 6\min\left\{H, \frac{1}{R}\right\}\left(1 + \frac{2}{3}\left|\zeta_{i,h}^{\widehat{\pi},R}\right|\right).
\end{aligned}$$

Combining with (51), we have

$$\begin{aligned}
\left|\zeta_{i,h}^{\widehat{\pi},R}\right| &\leq 4c_\alpha \sqrt{\frac{\log^3 \frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\{H, \frac{1}{R}\}}} \left(P_h^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) - \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}\right) \\
&\quad + 4c_\alpha \sqrt{\frac{\log^3 \frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\{H, \frac{1}{R}\}}} \left(\left(3 + \frac{c}{4c_\alpha}\right)\min\left\{H, \frac{1}{R}\right\} + 2\left|\zeta_{i,h}^{\widehat{\pi},R}\right|\min\left\{H, \frac{1}{R}\right\}\right) \\
&\leq 4c_\alpha \sqrt{\frac{\log^3 \frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\{H, \frac{1}{R}\}}} \left(P_h^{\widehat{\pi},V_{i,h+1}^{\widehat{\pi},R}} \left(\widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h+1}^{\widehat{\pi},R}\right) - \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R} \circ \widetilde{\overline{V}}_{i,h}^{\widehat{\pi},R}\right) \\
&\quad + c' \sqrt{\frac{\min\{H, \frac{1}{R}\}\log^3 \frac{KS\sum_{i=1}^m A_i}{\delta}}{K}} + \frac{1}{2}\left|\zeta_{i,h}^{\widehat{\pi},R}\right|,
\end{aligned} \tag{58}$$

where $c' = 12c_\alpha + c$, and the last inequality holds because $4c_\alpha\sqrt{\frac{\min\{H,\frac{1}{R}\}\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}} \le \frac{1}{4}$ as long as K satisfies (15). Finally, by solving $\left|\zeta_{i,h}^{\widehat{\pi},R}\right|$ from this inequality, we establish (33a).

The proof of (33b) follows similarly by replacing $\pi_{i,h}^k$ with $\widehat{\pi}_i^\star$, but requires additional clarification regarding the statistical dependency between the learned policy $\widehat{\pi}_i^\star$ and the samples used to construct $P_{i,h}^k$. We therefore provide a simplified proof for clarity.

Recall the (32) for $\zeta_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}$, which gives

$$\zeta_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}(s) = \mathbb{E}_{a_i\sim\widehat{\pi}_{i,h,s}^\star}\left[\varphi_{i,h}^{\widehat{\pi}_{-i},R}(s,a_i)\right].$$

Here, we define

$$\varphi_{i,h}^{\widehat{\pi}_{-i},R}(s,a_i) \coloneqq \sum_{k=1}^K \alpha_k^K\Big[r_{i,h}^k(s,a_i) - \mathbb{E}_{k-1}r_{i,h}^k(s,a_i) + (1-R)\left(P_{i,h}^k - \mathbb{E}_{k-1}P_{i,h}^k\right)(s,a_i)\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\Big],$$

with

$$\mathbb{E}_{k-1}r_{i,h}^k(s,a_i) = \mathbb{E}_{a_j\sim\pi_{j,h,s}^k, j\neq i}[r_{i,h}(s,\mathbf{a})], \qquad \mathbb{E}_{k-1}P_{i,h}^k(s,a_i) = \mathbb{E}_{a_j\sim\pi_{j,h,s}^k, j\neq i}[P_{h,s,\mathbf{a}}^0].$$

Then, notice the relation

$$|\zeta_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}(s)| \le \mathbb{E}_{a_i\sim\widehat{\pi}_{i,h,s}^\star}\left[|\varphi_{i,h}^{\widehat{\pi}_{-i},R}(s,a_i)|\right]. \tag{59}$$

This motivates us to control $\zeta_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}(s)$ through $\varphi_{i,h}^{\widehat{\pi}_{-i},R}(s,a_i)$ for every action $a_i\in\mathcal{A}_i$, which has no dependency issue. Now the remaining task lies in (i) controlling $\varphi_{i,h}^{\widehat{\pi}_{-i},R}$; and (ii) making use of the bound of $\varphi_{i,h}^{\widehat{\pi}_{-i},R}$ to establish (33b).

Controlling $\varphi_{i,h}^{\widehat{\pi}_{-i},R}$. Since that $\varphi_{i,h}^{\widehat{\pi}_{-i},R}(s,a_i)$ does not suffer from the dependency issue, we are able to apply Lemma 6 and follow the same argument as in (49) - (51) to obtain

$$\left|\varphi_{i,h}^{\widehat{\pi}_{-i},R}(s,a_i)\right| \le 2c_\alpha\sqrt{\frac{\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\{H,\frac{1}{R}\}}\sum_{k=1}^K\alpha_k^K\mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s,a_i}^k)} + c\sqrt{\frac{\min\{H,\frac{1}{R}\}\log\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}, \tag{60}$$

where

$$\widetilde{q}_{i,h,s,a_i}^k = r_{i,h}^k(s,a_i) + (1-R)P_{i,h}^k(s,a_i)\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R},$$

and the Var_{k-1} is with respect to distribution $a_j\sim\pi_{j,h,s}^k$, $j\neq i$ and $s_{i,h+1}^k(s,a_i)$. Following the same variance decomposition as in (52) - (53), we have

$$\begin{aligned}\sum_{k=1}^K\alpha_k^K\mathsf{Var}_{k-1}(\widetilde{q}_{i,h,s,a_i}^k) \le 1 &+ 2P_{h,s,a_i}^{\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\left(\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\circ\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)\\ &- 2\left(P_{h,s,a_i}^{\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)^2,\end{aligned} \tag{61}$$

where

$$\begin{aligned}P_{h,s,a_i}^{\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}} &= (1-R)\mathbb{E}_{\mathbf{a}_{-i}\sim\widehat{\pi}_{-i}}P_{h,s,\mathbf{a}}^0 + Re_{\arg\min\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}}}\\ &= (1-R)\sum_{k=1}^K\alpha_k^K\mathbb{E}_{k-1}P_{i,h}^k(s,a_i) + Re_{\arg\min\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i}}}.\end{aligned}$$

Invoking (61) into (60) yields:

$$\left|\varphi_{i,h}^{\widehat{\pi}_{-i},R}(s,a_i)\right| \le 4c_\alpha\sqrt{\frac{\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\{H,\frac{1}{R}\}}}\left(P_{h,s,a_i}^{\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\left(\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\circ\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)\right.$$
$$\left.-\left(P_{h,s,a_i}^{\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)^2\right)+c'\sqrt{\frac{\min\{H,\frac{1}{R}\}\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}. \tag{62}$$

Establishing (33b) from the bound of $\varphi_{i,h}^{\widehat{\pi}_{-i},R}$. Notice that

$$\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h,s}^\star}\left[P_{h,s,a_i}^{\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\left(\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\circ\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)\right]=P_{h,s,a_i}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\left(\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\circ\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right),$$

$$\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h,s}^\star}\left(P_{h,s,a_i}^{\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)^2\ge\left(\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h,s}^\star}\left[P_{h,s,a_i}^{\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right]\right)^2$$
$$=\left(P_{h,s,a_i}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)^2.$$

Combining (59), (62), and the above relations yields

$$|\zeta_{i,h}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}(s)|\le 4c_\alpha\sqrt{\frac{\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\{H,\frac{1}{R}\}}}\left(P_{h,s,a_i}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\left(\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\circ\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)\right.$$
$$\left.-\left(P_{h,s,a_i}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},\overline{V}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}}\overline{\widetilde{V}}_{i,h+1}^{\widehat{\pi}_i^\star\times\widehat{\pi}_{-i},R}\right)^2\right)+c'\sqrt{\frac{\min\{H,\frac{1}{R}\}\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K}}.$$

Finally, following the same argument as in (56) - (58) gives (33b).

D.3 Proof of Lemma 3

In the definition of $\beta_{i,h}$ (c.f. (12)), for convenience, we denote

$$\rho_0:=c_{\mathsf{b}}\sqrt{\frac{\log^3\frac{KS\sum_{i=1}^m A_i}{\delta}}{K\min\left\{H,\frac{1}{R}\right\}}},\qquad \rho_1=\rho_0\min\left\{H,\frac{1}{R}\right\}.$$

By using these new notations, we could rewrite the definition of $\beta_{i,h}$ as

$$\beta_{i,h}=\rho_0\sum_{k=1}^K\alpha_k^K\mathsf{Var}_{\pi_{i,h,s}^k}(q_{i,h,s}^k)+\rho_1$$
$$\overset{(i)}{\le}2\rho_0\sum_{k=1}^K\alpha_k^K\mathsf{Var}_{\pi_{i,h,s}^k}(q_{i,h,s}^k-r_{i,h,s}^k)+2\rho_0+\rho_1, \tag{63}$$

where (i) uses the fact that

$\mathsf{Var}_{\pi_{i,h,s}^k}(q_{i,h,s}^k)\le 2\mathsf{Var}_{\pi_{i,h,s}^k}(r_{i,h,s}^k)+2\mathsf{Var}_{\pi_{i,h,s}^k}(q_{i,h,s}^k-r_{i,h,s}^k)\le 2+2\mathsf{Var}_{\pi_{i,h,s}^k}(q_{i,h,s}^k-r_{i,h,s}^k)$,

and $\sum_{k=1}^K\alpha_k^K=1$. The variance is further decomposed as

$$\sum_{k=1}^K\alpha_k^K\mathsf{Var}_{\pi_{i,h,s}^k}\left(q_{i,h,s}^k-r_{i,h,s}^k\right)$$
$$=\sum_{k=1}^K\alpha_k^K\mathbb{E}_{\pi_{i,h,s}^k}\left(q_{i,h,s}^k-r_{i,h,s}^k\right)^2-\sum_{k=1}^K\alpha_k^K\left(\mathbb{E}_{\pi_{i,h,s}^k}\left(q_{i,h,s}^k-r_{i,h,s}^k\right)\right)^2. \tag{64}$$

The expectation of the square is bounded by

$$\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{\pi^k_{i,h,s}}(q^k_{i,h,s}-r^k_{i,h,s})^2\le\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{\pi^k_{i,h,s}}\left((1-R)P^k_{i,h}\widehat{V}_{i,h+1}+R\min_s\widehat{V}_{i,h+1}\right)^2$$
$$=\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\left(\widehat{V}_{i,h+1}\circ\widehat{V}_{i,h+1}\right),\tag{65}$$

and the square of the expectation is bounded by

$$\sum_{k=1}^{K}\alpha_k^K\left(\mathbb{E}_{\pi^k_{i,h,s}}\left(q^k_{i,h,s}-r^k_{i,h,s}\right)\right)^2\overset{(i)}{\ge}\left(\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{\pi^k_{i,h,s}}\left(q^k_{i,h,s}-r^k_{i,h,s}\right)\right)^2$$
$$\ge\left(\sum_{k=1}^{K}\alpha_k^K\mathbb{E}_{\pi^k_{i,h,s}}\left((1-R)P^k_{i,h}\widehat{V}_{i,h+1}+R\min_s\widehat{V}_{i,h+1}\right)\right)^2$$
$$=\left(\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widehat{V}_{i,h+1}\right)\circ\left(\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widehat{V}_{i,h+1}\right),\tag{66}$$

where (i) uses the Jensen's inequality. Inserting (64) and (65) into (66), we have

$$\sum_{k=1}^{K}\alpha_k^K\mathsf{Var}_{\pi^k_{i,h,s}}\left(q^k_{i,h,s}-r^k_{i,h,s}\right)$$
$$\le\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\left(\widehat{V}_{i,h+1}\circ\widehat{V}_{i,h+1}\right)-\left(\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widehat{V}_{i,h+1}\right)\circ\left(\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widehat{V}_{i,h+1}\right)$$
$$=\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)-\left(\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widetilde{\widehat{V}}_{i,h+1}\right)\circ\left(\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widetilde{\widehat{V}}_{i,h+1}\right),\tag{67}$$

where

$$\widetilde{\widehat{V}}_{i,h}:=\widehat{V}_{i,h}-\min\widehat{V}_{i,h}.$$

Inserting it into (63), we have

$$\beta_{i,h}\le2\rho_0\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)+2\rho_0+\rho_1$$
$$\le2\rho_0\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)+3\rho_1.$$

Recalling the definition of ρ_0 and ρ_1, we could complete the proof of (37a).

Next, we shall prove (37b). Define

$$\widehat{\vartheta}_{i,h}:=\widetilde{\widehat{V}}_{i,h}-\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widetilde{\widehat{V}}_{i,h+1}.$$

Noticing that

$$\min\widehat{V}_{i,h}\ge\min\left\{(1-R)\min\widehat{V}_{i,h+1}+R\min\widehat{V}_{i,h+1}+\beta_{i,h},H-h+1\right\}$$
$$\ge\min\left\{\min\widehat{V}_{i,h+1}+\beta_{i,h},H-h+1\right\}\overset{(i)}{\ge}\min\widehat{V}_{i,h+1},$$

where (i) uses $\widehat{V}_{i,h+1}\le H-h\le H-h+1$, we have

$$\widehat{\vartheta}_{i,h}=\widehat{V}_{i,h}-\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widehat{V}_{i,h+1}-\min\widehat{V}_{i,h}+\min\widehat{V}_{i,h+1}\le\widehat{V}_{i,h}-\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widehat{V}_{i,h+1}\le1+\beta_{i,h}.\tag{68}$$

By using this, we have

$$\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)-\left(\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widetilde{\widehat{V}}_{i,h+1}\right)\circ\left(\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\widetilde{\widehat{V}}_{i,h+1}\right)$$
$$=\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)-\left(\widetilde{\widehat{V}}_{i,h}-\widehat{\vartheta}_{i,h}\right)\circ\left(\widetilde{\widehat{V}}_{i,h}-\widehat{\vartheta}_{i,h}\right)$$
$$\le\widehat{P}^{\widehat{\pi},\widehat{V}_{i,h+1}}_{i,h}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)-\widetilde{\widehat{V}}_{i,h}\circ\widetilde{\widehat{V}}_{i,h}+2\widehat{\vartheta}_{i,h}\circ\widetilde{\widehat{V}}_{i,h}$$
$$\overset{(i)}{\le}(1-R)\widehat{P}^{\widehat{\pi},\widehat{\pi}}_{i,h}\left(\widetilde{\widehat{V}}_{i,h+1}\circ\widetilde{\widehat{V}}_{i,h+1}\right)-\widetilde{\widehat{V}}_{i,h}\circ\widetilde{\widehat{V}}_{i,h}+2(1+\beta_{i,h})\circ\widetilde{\widehat{V}}_{i,h},$$

where (i) uses the definition of $\widehat{P}^{\widehat{\pi}}_{i,h}$ and (68). Inserting it into (67), we have

$$\begin{aligned}
&\sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{\pi^k_{i,h,s}} \left(q^k_{i,h,s} - r^k_{i,h,s}\right) \\
&\leq (1-R)\widehat{P}^{\widehat{\pi}}_{i,h} \left(\widetilde{\widehat{V}}_{i,h+1} \circ \widetilde{\widehat{V}}_{i,h+1}\right) - \widetilde{\widehat{V}}_{i,h} \circ \widetilde{\widehat{V}}_{i,h} + 2(1+\beta_{i,h}) \circ \widetilde{\widehat{V}}_{i,h}. \qquad (69)
\end{aligned}$$

Inserting (69) into (63), and noticing that $\widetilde{\widehat{V}}_{i,h} \leq 3\min\{H, \frac{1}{R}\}$, we have

$$\begin{aligned}
\beta_{i,h} &\leq 2\rho_0 \left((1-R)\widehat{P}^{\widehat{\pi}}_{i,h} \left(\widetilde{\widehat{V}}_{i,h+1} \circ \widetilde{\widehat{V}}_{i,h+1}\right) - \widetilde{\widehat{V}}_{i,h} \circ \widetilde{\widehat{V}}_{i,h} + 2\right) \\
&\quad + 6\rho_0 \beta_{i,h} \min\left\{H, \frac{1}{R}\right\} + 2\rho_0 + \rho_1.
\end{aligned}$$

Recalling that for K satisfying (15), we have $6\rho_0 \min\left\{H, \frac{1}{R}\right\} \leq \frac{1}{2}$. Thus we have

$$\begin{aligned}
\beta_{i,h} &\leq 4\rho_0 \left((1-R)\widehat{P}^{\widehat{\pi}}_{i,h} \left(\widetilde{\widehat{V}}_{i,h+1} \circ \widetilde{\widehat{V}}_{i,h+1}\right) - \widetilde{\widehat{V}}_{i,h} \circ \widetilde{\widehat{V}}_{i,h}\right) + 12\rho_0 + 2\rho_1 \\
&\leq 4\rho_0 \left((1-R)\widehat{P}^{\widehat{\pi}}_{i,h} \left(\widetilde{\widehat{V}}_{i,h+1} \circ \widetilde{\widehat{V}}_{i,h+1}\right) - \widetilde{\widehat{V}}_{i,h} \circ \widetilde{\widehat{V}}_{i,h}\right) + 14\rho_1.
\end{aligned}$$

Then we complete the proof of (37b).

D.4 Proof of Lemma 4

We shall prove Lemma 4 by mathematical induction. For $h = H+1$, it is obviously that $\widehat{V}_{i,H+1} = \overline{V}^{\star,\widehat{\pi}_{-i},R}_{i,H+1} = 0$. Assume that (40) holds for $h+1$, that is, $\widehat{V}_{i,h+1} \geq \overline{V}^{\star,\widehat{\pi}_{-i},R}_{i,h+1}$. It suffices to prove that (40) holds for h.

According to the definition of $\overline{V}^{\star,\widehat{\pi}_{-i},R}_{i,h}$, we have

$$\begin{aligned}
\overline{V}^{\star,\widehat{\pi}_{-i},R}_{i,h}(s) &= \max_{a_i \in \mathcal{A}_i} \sum_{k=1}^{K} \alpha_k^K \left[r^k_{i,h}(s,a_i) + (1-R)P^k_{i,h}(s,a_i)\overline{V}^{\star,\widehat{\pi}_{-i},R}_{i,h+1} + R\min_{s\in\mathcal{S}} \overline{V}^{\star,\widehat{\pi}_{-i},R}_{i,h+1}\right] \\
&\leq \max_{a_i \in \mathcal{A}_i} \sum_{k=1}^{K} \alpha_k^K \left[r^k_{i,h}(s,a_i) + (1-R)P^k_{i,h}(s,a_i)\widehat{V}_{i,h+1} + R\min_{s\in\mathcal{S}} \widehat{V}_{i,h+1}\right] \\
&= \max_{a_i \in \mathcal{A}_i} \sum_{k=1}^{K} \alpha_k^K q^k_{i,h,s} = \max_{a_i \in \mathcal{A}_i} Q^K_{i,h,s}, \qquad (70)
\end{aligned}$$

where $Q^K_{i,h,s} = Q^K_{i,h}(s,\cdot) \in \mathbb{R}^{A_i}$, and the last equality uses the fact that $Q^K_{i,h,s} = \sum_{k=1}^K \alpha_k^K q^k_{i,h,s}$ because of $Q^0_{i,h,s} = 0$ for all i,h,s.

Below we intend to find the upper bound of $Q^k_{i,h,s}$. Towards this, we introduce the following auxiliary lemma.

Lemma 7. *(Li et al., 2022, Theorem 3) Suppose that $0 < \alpha_1 \leq 1$ and $\eta_1 = \eta_2(1-\alpha_1)$. Also assume that $0 < \alpha_k < 1$ and $0 < \eta_{k+1}(1-\alpha_k) \leq \eta_k$ for all $k \geq 2$. In addition, define $\widehat{\eta}_1 = \eta_2$ and $\widehat{\eta}_k = \frac{\eta_k}{1-\alpha_k}$ for $k \geq 2$. Then we have*

$$\begin{aligned}
Q^K_{i,h,s} &\leq \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{\pi^k_{i,h,s}} q^k_{i,h,s} + \frac{5}{3}\sum_{k=1}^{K} \alpha_k^K \widehat{\eta}_k \alpha_k \mathsf{Var}_{\pi^k_{i,h,s}}(q^k_{i,h,s}) \\
&\quad + \frac{\log A_i}{\eta_{K+1}} + 3\sum_{k=1}^{K} \alpha_k^K \widehat{\eta}_k^2 \alpha_k^2 \left\|q^k_{i,h,s} - \min q^k_{i,h,s}\right\|_\infty^3 \mathbb{1}\left(\widehat{\eta}_k \alpha_k \left\|q^k_{i,h,s} - \min q^k_{i,h,s}\right\|_\infty > \frac{1}{3}\right).
\end{aligned}$$

Careful readers may find that this lemma is slightly different from Theorem 3 in Li et al. (2022) by replacing $\left\|q^k_{i,h,s}\right\|_\infty$ with $\left\|q^k_{i,h,s} - \min q^k_{i,h,s}\right\|_\infty$. This can be easily verified by replacing all ℓ_k with $\ell_k - \min \ell_k$ when bounding $\left\langle \pi_k - \pi^-_{k+1}, \ell_k \right\rangle$ in the proof of Li et al. (2022).

To verify that the condition in Lemma 7 holds true, it suffices to prove that

$$\frac{\eta_k}{\eta_{k+1}} = \sqrt{\frac{\alpha_{k+1}}{\alpha_k}} = \sqrt{\frac{k-1+c_\alpha \log K}{k + c_\alpha \log K}} = \sqrt{1 - \frac{1}{k + c_\alpha \log K}} \overset{(i)}{\geq} \sqrt{1-\alpha_k} \geq 1 - \alpha_k,$$

where (i) holds as long as $c_\alpha \geq 1$ and $K \geq 3$. Then applying Lemma 7, we have

$$Q^K_{i,h,s} \leq \sum_{k=1}^K \alpha^K_k \mathbb{E}_{\pi^k_{i,h,s}} q^k_{i,h,s} + \xi_1 + \xi_2 + \xi_3,$$

where

$$\xi_1 = \frac{5}{3}\sum_{k=1}^K \alpha^K_k \widehat{\eta}_k \alpha_k \mathsf{Var}_{\pi^k_{i,h,s}}(q^k_{i,h,s}), \quad \xi_2 = \frac{\log A_i}{\eta_{K+1}},$$
$$\xi_3 = 3\sum_{k=1}^K \alpha^K_k \widehat{\eta}^2_k \alpha^2_k \left\|q^k_{i,h,s} - \min q^k_{i,h,s}\right\|^3_\infty \mathbb{1}\left(\widehat{\eta}_k \alpha_k \left\|q^k_{i,h,s} - \min q^k_{i,h,s}\right\|_\infty > \frac{1}{3}\right).$$

Below we shall bound ξ_1, ξ_2, and ξ_3 separately.

- For ξ_1, we have

$$\begin{aligned}
\xi_1 &= \frac{5}{3}\alpha^K_1 \alpha_1 \eta_2 \mathsf{Var}_{\pi^1_{i,h,s}}(q^1_{i,h,s}) + \frac{5}{3}\sum_{k=2}^K \alpha^K_k \frac{\alpha_k \eta_k}{1-\alpha_k} \mathsf{Var}_{\pi^k_{i,h,s}}(q^k_{i,h,s}) \\
&\overset{(i)}{\lesssim} \frac{\min\left\{H, \frac{1}{R}\right\}^2}{K^6} + \sum_{k=2}^{K/2} \frac{(\min\{H, \frac{1}{R}\}\log K)^{\frac{3}{2}}}{K^6} \\
&\quad + \sum_{k=K/2+1}^K \sqrt{\frac{\log^2 K}{K \min\{H, \frac{1}{R}\}}} \alpha^K_k \mathsf{Var}_{\pi^k_{i,h,s}}(q^k_{i,h,s}) \\
&\overset{(ii)}{\lesssim} \sqrt{\frac{\log^2 K}{K \min\{H, \frac{1}{R}\}}} \sum_{k=1}^K \alpha^K_k \mathsf{Var}_{\pi^k_{i,h,s}}(q^k_{i,h,s}) + \frac{1}{K^3},
\end{aligned} \tag{71}$$

  where (i) holds since $\alpha^K_k \leq \frac{1}{K^6}$ according to Lemma 5, and $\mathsf{Var}_{\pi^k_{i,h,s}}(q^k_{i,h,s}) \lesssim \min\{H, \frac{1}{R}\}$, and

$$\begin{aligned}
&\frac{1}{1-\alpha_k} \leq \frac{1}{1-\alpha_2} = 1 + c_\alpha \log K \lesssim \log K, \quad k \geq 2, \\
&\frac{1}{1-\alpha_k} \leq \frac{1}{1-\alpha_{K/2}} \leq 2, \quad k \geq \frac{K}{2}, \\
&\frac{\eta_k \alpha_k}{1-\alpha_k} = \frac{\sqrt{\alpha_k}}{1-\alpha_k}\sqrt{\frac{\log K}{\min\{H, \frac{1}{R}\}}} \leq 2\sqrt{\frac{\log^2 K}{k \min\{H, \frac{1}{R}\}}}.
\end{aligned} \tag{72}$$

  as long as $K \geq 2c_\alpha \log K + 2$, and (ii) uses the fact that $K \gtrsim \log K \min\{H, \frac{1}{R}\}$ for any $\varepsilon \in (0, H]$.

- For ξ_2, we have

$$\xi_2 = \frac{\log A_i}{\eta_{K+1}} = \log A_i \sqrt{\frac{\min\{H, \frac{1}{R}\}}{\log K}} \sqrt{\frac{c_\alpha \log K}{K + c_\alpha \log K}} \leq \log A_i \sqrt{\frac{c_\alpha \min\{H, \frac{1}{R}\}}{K}}. \tag{73}$$

- For term ξ_3, we have

$$\xi_3 \le 3\sum_{k=2}^{K} \alpha_k^K \widehat{\eta}_k^2 \alpha_k^2 \left\|q_{i,h,s}^k - \min q_{i,h,s}^k\right\|_\infty^3 \mathbb{1}\left(\widehat{\eta}_k \alpha_k \left\|q_{i,h,s}^k - \min q_{i,h,s}^k\right\|_\infty > \frac{1}{3}\right) + 3\alpha_1^K \eta_2^2 \alpha_1^2 \left\|q_{i,h,s}^1 - \min q_{i,h,s}^1\right\|_\infty^3. \tag{74}$$

Here by using Lemma 1, the second term in the right-hand-side of (74) satisfy

$$3\alpha_1^K \eta_2^2 \alpha_1^2 \left\|q_{i,h,s}^1 - \min q_{i,h,s}^1\right\|_\infty^3 \lesssim \frac{\min\left\{H, \frac{1}{R}\right\}^3}{K^6} \le \frac{1}{K^3}. \tag{75}$$

To bound the first term, notice (72), we have

$$\widehat{\eta}_k \alpha_k \left\|q_{i,h,s}^k - \min q_{i,h,s}^k\right\|_\infty = \frac{\eta_k \alpha_k}{1-\alpha_k} \left\|q_{i,h,s}^k - \min q_{i,h,s}^k\right\|_\infty \le c \log K \sqrt{\frac{\min\{H, \frac{1}{R}\}}{k}} \le \frac{1}{3}, \quad k \ge 9c^2 \log^2 K \min\{H, \frac{1}{R}\},$$

where c is a constant. Then the first term is bound by

$$\begin{aligned} &3\sum_{k=2}^{K} \alpha_k^K \widehat{\eta}_k^2 \alpha_k^2 \left\|q_{i,h,s}^k - \min q_{i,h,s}^k\right\|_\infty^3 \mathbb{1}\left(\widehat{\eta}_k \alpha_k \left\|q_{i,h,s}^k - \min q_{i,h,s}^k\right\|_\infty > \frac{1}{3}\right) \\ &\le 3 \sum_{k=2}^{9c^2 \log^2 K \min\{H, \frac{1}{R}\}} \alpha_k^K \widehat{\eta}_k^2 \alpha_k^2 \left\|q_{i,h,s}^k - \min q_{i,h,s}^k\right\|_\infty^3 \\ &\overset{(i)}{\lesssim} \frac{\log^2 K \min\{H, \frac{1}{R}\}^4}{K^6} \overset{(i)}{\lesssim} \frac{1}{K^2}, \end{aligned} \tag{76}$$

where (i) holds because $\alpha_k^K \le \frac{1}{K^6}$ for $K/2 \ge 9c^2 \log^2 K \min\{H, \frac{1}{R}\}$, and (ii) holds because $K \gtrsim \log K \min\{H, \frac{1}{R}\}$ for any $\varepsilon \in (0, H]$. Inserting (75) and (76) into (74), we have

$$\xi_3 \lesssim \frac{1}{K^2}. \tag{77}$$

Combining (71), (73), and (77), and notice that $K \gtrsim \min\{H, \frac{1}{R}\}$, we have

$$\begin{aligned} Q_{i,h,s}^K &\le \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{\pi_{i,h,s}^k} q_{i,h,s}^k + c\sqrt{\frac{\log^2 K}{K \min\{H, \frac{1}{R}\}}} \sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{\pi_{i,h,s}^k}(q_{i,h,s}^k) \\ &\quad + \log A_i \sqrt{\frac{(c_\alpha + 1)\min\{H, \frac{1}{R}\}}{K}}. \end{aligned}$$

According to the definition of $\widehat{V}_{i,h}$ in (13), if $\widehat{V}_{i,h} = H - h + 1$, then (40) holds obviously. Otherwise, we have

$$\begin{aligned} \widehat{V}_{i,h}(s) &= \sum_{k=1}^{K} \alpha_k^K \mathbb{E}_{a_i \sim \pi_{i,h,s}^k} q_{i,h,s} + \beta_{i,h}(s) \\ &\ge \max Q_{i,h,s}^K - c\sqrt{\frac{\log^2 K}{K \min\{H, \frac{1}{R}\}}} \sum_{k=1}^{K} \alpha_k^K \mathsf{Var}_{\pi_{i,h,s}^k}(q_{i,h,s}^k) \\ &\quad - \log A_i \sqrt{\frac{(c_\alpha + 1)\min\{H, \frac{1}{R}\}}{K}} + \beta_{i,h}(s) \\ &\overset{(i)}{\ge} \max Q_{i,h,s}^K \overset{(ii)}{\ge} \overline{V}_{i,h}^{\star, \widehat{\pi}_{-i}, R}, \end{aligned}$$

where (i) holds as long as $c_{\mathsf{b}} \ge c + \sqrt{c_\alpha + 1}$, and (ii) arises from $\overline{V}_{i,h}^{\star, \widehat{\pi}_{-i}, R} \le \max_{a_i} Q_{i,h,s}^k$ as shown in (70). Thus we complete the proof.

E PROOF OF THEOREM 2

According to the proof of Theorem 2 in Shi et al. (2024c), we could construct a collection of robust Markov games, such that finding a robust NE/CE/CCE of these RMGs degrades to finding the optimal policy of the first agent over a corresponding robust Markov Decision Processes. Thus the proof of Theorem 2 degrades to prove that: one can construct a collection of robust Markov Decision Processes $\{\mathcal{M}_\theta|\theta \in \Theta\}$ with S states, A actions, horizon $H \geq 16$, and the uncertanty level $R \in [0, c_R]$, such that

$$\inf_{\pi} \max_{\theta\in\Theta} \mathbb{P}_\theta \left(\max_s \left[V_1^{\star,R}(s) - V_1^{\widehat{\pi},R}(s) \right] \geq \varepsilon \right) \geq \frac{1}{4},$$

provided that the sample size for each state-action pair over the nominal transition kernel obeys

$$N \leq \frac{cH^3}{\varepsilon^2} \min\left\{ H, \frac{1}{R} \right\}$$

for some sufficiently small constant $c > 0$. Here, the infimum is over all policy estimators $\widehat{\pi}$, and $\mathbb{P}_\theta$ denotes the probability when the RMDP is $\mathcal{M}_\theta$. Below we shall construct a hard instance.

E.1 CONSTRUCTION OF THE HARD PROBLEM INSTANCES

Construction of the hard MDPs. To begin with, define the state space and the action space as $\mathcal{S} = \{0, 1, \cdots, S-1\}$ and $\mathcal{A} = \{0, 1, \cdots, A-1\}$, respectively. To construct the nominal transition probability matrix, for any integer $H \geq 16$, let us consider a set $\Theta \subset \{0,1\}^H$ of H-dimensional vectors, which we shall construct shortly. We then generate a collection of MDPs

$$\mathsf{MDP}(\Theta) = \left\{ \mathcal{M}_\theta = \left(\mathcal{S}, \mathcal{A}, P^\theta = \{P_h^\theta\}_{h=1}^H, \{r_h\}_{h=1}^H, H\right) | \theta = [\theta_h]_{1\leq h\leq H} \in \Theta \right\}.$$

Given any state $s \in \{2, \cdots, S-1\}$, the corresponding action space is $\mathcal{A} = \{0, \cdots, A-1\}$. While for state $s = 0$, the action space is $\mathcal{A}' = \{0, 1\}$. For any $\theta = \{\theta_h\}_{h=1}^H \in \{0,1\}^H$, the nominal transition kernel P^θ of the constructed robust MDP $\mathcal{M}_\phi$ is defined as

$$P_h^\theta(s_{h+1}|s_h, a_h) = \begin{cases} p\mathbb{1}(s_{h+1}=1) + (1-p)\mathbb{1}(s_{h+1}=0), & \text{if } (s_h, a_h) = (0, \theta_h), \\ q\mathbb{1}(s_{h+1}=1) + (1-q)\mathbb{1}(s_{h+1}=0), & \text{if } (s_h, a_h) = (0, 1-\theta_h), \\ \mathbb{1}(s_{h+1}=1), & \text{if } s_h \geq 1, \end{cases}$$

where $0 < q < p < \frac{3}{4}$ are defined as

$$p = c\max\left\{\frac{1}{H}, R\right\}, \quad q = p - \Delta,$$

where

$$\Delta := \frac{c_1\varepsilon}{H\min\left\{H, \frac{1}{R}\right\}}, \tag{78}$$

and $c \leq 3/4$. Moreover, we assume that $\varepsilon \leq c_0 H$. Here c_1 is a sufficiently large constant, c_0 is a sufficiently small constant, and meanwhile $c_0 c_1 \leq c/2$. With this definition, one can easily check that $q \geq p - \frac{c_1 c_0}{c} \geq \frac{p}{2}$ for any $\varepsilon \in (0, H]$.

Then, we define the reward function as

$$r_h(s,a) = \begin{cases} 1, & \text{if } s = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, let us choose the set $\Theta \subset \{0,1\}^H$. By virtue of the Gilbert-Varshamov lemma (Gilbert, 1952), one can construct $\Theta \subset \{0,1\}^H$ in a way that

$$|\Theta| \geq \mathrm{e}^{H/8} \quad \text{and} \quad \left\|\theta - \tilde{\theta}\right\|_1 \geq \frac{H}{8} \quad \text{for any } \theta, \tilde{\theta} \in \Theta \text{ obeying } \theta \neq \tilde{\theta}. \tag{79}$$

Uncertainty set of the transition kernels. Recalling the uncertainty set assumed is

$$\mathcal{U}^R(P_h^\theta) = \otimes \mathcal{U}^R(P_h^\theta(\cdot|s,a)), \quad \mathcal{U}^R(P_h^\theta(\cdot|s,a)) := \left\{(1-R)P_h^\theta(\cdot|s,a) + RP' | P' \in \Delta(\mathcal{S})\right\}.$$

For any policy $\pi = \{\pi_h\}_{h=1}^H$, its corresponding transition probability matrix is

$$P_h^{\theta,\pi}(s_{h+1}|s_h) = \begin{cases} z_h^\pi \mathbb{1}(s_{h+1}=1) + (1-z_h^\pi)\,\mathbb{1}(s_{h+1}=0), & \text{if } s_h = 0, \\ \mathbb{1}(s_{h+1}=1), & \text{if } s_h \geq 1, \end{cases}$$

where

$$z_h^\pi = \pi_h(\theta_h|0)p + \pi_h(1-\theta_h|0)q.$$

We pause to show by mathematical induction that for any policy π, we always have

$$V_h^{\pi,R}(0) = \min_s V_h^{\pi,R}(s), \quad \forall 1 \leq h \leq H. \tag{80}$$

It is obvious that $V_{H+1}^{\pi,R}(0) = \min_s V_{H+1}^{\pi,R}(s) = 0$. Suppose that $V_{h+1}^{\pi,R}(0) = \min_s V_{h+1}^{\pi,R}(s)$. Then we have

$$\begin{aligned} V_h^{\pi,R}(0) &= (1-R)z_h^\pi V_{h+1}^{\pi,R}(1) + (1-R)(1-z_h^\pi)V_{h+1}^{\pi,R}(0) + RV_{h+1}^{\pi,R}(0) \\ &\leq (1-R)V_{h+1}^{\pi,R}(1) + RV_{h+1}^{\pi,R}(0) \\ &\leq (1-R)V_{h+1}^{\pi,R}(1) + RV_{h+1}^{\pi,R}(0) = \min_{s\geq 1} V_h^{\pi,R}(s) \\ &\leq 1 + (1-R)V_{h+1}^{\pi,R}(1) + RV_{h+1}^{\pi,R}(0) = V_h^{\pi,R}(1). \end{aligned}$$

Thus we complete the proof.

Therefore, for any policy π, the perturbed transition kernel that minimizes the expected reward is

$$\begin{aligned} P_h^{\theta,\pi,V_{h+1}^{\pi,R}}(\cdot|s_h) &= \arg\min_{P\in\mathcal{U}_h^R(P_h^{\theta,\pi}(\cdot|s_h))} PV_{h+1}^{\pi,R} \\ &= \begin{cases} (1-R)z_h^\pi \mathbb{1}(s_{h+1}=1) + (1-(1-R)z_h^\pi)\,\mathbb{1}(s_{h+1}=0), & \text{if } s_h = 0, \\ (1-R)\mathbb{1}(s_{h+1}=1) + R\mathbb{1}(s_{h+1}=0), & \text{if } s_h \geq 1. \end{cases} \end{aligned}$$

Optimal robust policy and value function. To proceed, we are ready to identify the the optimal robust policies, and derive the corresponding robust value functions. According to (80), the optimal robust policy $\pi^{\star,\theta} = \{\pi_h^{\star,\theta}\}_{h=1}^H$ for $\mathcal{M}_\theta$ is

$$\pi_h^{\star,\theta}(\theta_h|0) = 1, \quad \pi_h^{\star,\theta}(1-\theta_h|0) = 0. \tag{81}$$

According to robust Bellman equation, the corresponding value function is

$$\begin{aligned} V_h^{\star,R}(0) &= \frac{\widetilde{p}}{R+\widetilde{p}}\left(H-h+1-\frac{1-(1-R-\widetilde{p})^{H-h+1}}{R+\widetilde{p}}\right) \\ V_h^{\star,R}(1) &= \frac{1}{R+\widetilde{p}}\left(\widetilde{p}(H-h+1) + R\frac{1-(1-R-\widetilde{p})^{H-h+1}}{R+\widetilde{p}}\right), \end{aligned} \tag{82}$$

where

$$\widetilde{p} = (1-R)p. \tag{83}$$

The derivation is deferred to the end of this section. Here we remark that for all θ, RMPDs $\mathcal{M}_\theta$ share the same optimal value function. Thus we omit the superscript θ in $V_h^{\star,R}$.

E.2 Establishing the minimax lower bound

We are now positioned to establish our sample complexity lower bounds. It suffices to prove that

$$\inf_{\widehat{\pi}} \max_{\theta\in\Theta} \mathbb{P}_\theta\left(V_1^{\star,R}(0) - V_1^{\widehat{\pi},R}(0) > \varepsilon\right) \geq \frac{1}{4}, \tag{84}$$

where $\widehat{\pi}$ is any policy estimator computed based on the independent samples.

Step 1: converting the goal to estimate θ. For any policy estimator $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^H$, We can construct a corresponding estimator $\widehat{\theta} = \{\widehat{\theta}_h\}_{h=1}^H$ for θ as

$$\widehat{\theta} := \arg\min_{\bar{\theta}\in\Theta} \sum_{h=1}^{H} \left\| \widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\bar{\theta}}(\cdot|0) \right\|_1, \tag{85}$$

where $\pi_h^{\star,\bar{\theta}}$ is defined in (81). According to the construction of Θ in (79), a good policy estimator $\widehat{\pi}$ implies a correct estimator $\widehat{\theta}$. Mathematically, with samples from $\mathcal{M}_\theta$ for $\theta \in \Theta$, we have

$$\mathbb{P}_\theta \left(V_1^{\star,R}(0) - V_1^{\widehat{\pi},R}(0) < \varepsilon \right) \leq \mathbb{P}_\theta \left(\widehat{\theta} = \theta \right), \tag{86}$$

where $\mathbb{P}_\theta$ denotes the probability distribution when the MDP is $\mathcal{M}_\theta$. The proof of this claim is postponed to the end of this section.

Step 2: probability of error in testing multiple hypotheses. Next, we turn attention to the hypothesis testing problem on Θ. Specifically, we focus on the minimax probability of error defined as follows:

$$p_e := \inf_{\psi} \max_{\theta\in\Theta} \mathbb{P}_\theta(\psi \neq \theta), \tag{87}$$

where the infimum is taken over all possible tests ψ constructed based on the samples from $\mathcal{M}_\theta$.

Let μ_h^θ represent the distribution of a sample tuple (s_h, a_h, s_{h+1}) for step h under the nominal transition kernel P_h^θ associated with $\mathcal{M}_\theta$ for $1 \leq h \leq H$. Assume that we collect N independent samples for each state-action pair and each time step H. Combined with Fano's inequality from Tsybakov (2009) (c.f.Theorem 2.2) and the additivity of the KL divergence (cf. Tsybakov (2009), Page 85), we have

$$\begin{aligned} p_e &\geq 1 - \frac{N \max_{\theta,\tilde{\theta}\in\Theta} \sum_{h=1}^H \mathsf{KL}(\mu_h^\theta || \mu_h^{\tilde{\theta}}) + \log 2}{\log|\Theta|} \\ &\geq \frac{1}{2} - \frac{8N \max_{\theta,\tilde{\theta}\in\Theta} \sum_{h=1}^H \mathsf{KL}(\mu_h^\theta || \mu_h^{\tilde{\theta}})}{H}, \end{aligned} \tag{88}$$

where we use the assumption that $H \geq 16\log 2$ and $|\Theta| \geq \mathrm{e}^{H/8}$. Notice that

$$\begin{aligned} \mathsf{KL}(\mu_h^\theta || \mu_h^{\tilde{\theta}}) &= \sum_{a=\{0,1\}} \mathsf{KL}\left(P_h^\theta(\cdot|0,a) || P_h^{\tilde{\theta}}(\cdot|0,a) \right) \\ &= \left(p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} + q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p} \right) \mathbb{1}(\theta_h \neq \tilde{\theta}_h) \\ &\leq \left(\frac{p^2}{q} - p + \frac{(1-p)^2}{1-q} - 1 + p + \frac{q^2}{p} - 1 + \frac{(1-q)^2}{1-p} - 1 + q \right) \mathbb{1}(\theta_h \neq \tilde{\theta}_h) \\ &= \left(\frac{(p-q)^2}{p(1-p)} + \frac{(p-q)^2}{q(1-q)} \right) \mathbb{1}(\theta_h \neq \tilde{\theta}_h) \\ &\overset{(i)}{\leq} \frac{8(p-q)^2}{q} \mathbb{1}(\theta_h \neq \tilde{\theta}_h), \end{aligned}$$

where (i) uses the fact that $p(1-p) \geq \frac{p}{4}$ and $q(1-q) \geq \frac{q}{4}$ for $q < p < \frac{3}{4}$. Inserting into (88), we have

$$\begin{aligned} p_e &\geq \frac{1}{2} - \frac{8N \max_{\theta,\tilde{\theta}\in\Theta} \left\| \theta - \tilde{\theta} \right\|_1}{H} \frac{8(p-q)^2}{q} \\ &\overset{(i)}{\geq} \frac{1}{2} - \frac{64N(p-q)^2}{q} \overset{(ii)}{\geq} \frac{1}{2} - \frac{128Nc_1^2\varepsilon^2}{cH^2 \min\left\{H, \frac{1}{R}\right\}}, \end{aligned}$$

where (i) uses the value of $p-q$ in (78), and the fact that $\max_{\theta,\tilde{\theta}\in\Theta}\left\|\theta-\tilde{\theta}\right\|_1 \leq H$, (ii) uses the fact that $q \geq \frac{p}{2}$. Provided that the total number of samples

$$NHSA \leq \frac{cH^3SA}{512c_1^2\varepsilon^2}\min\left\{H, \frac{1}{R}\right\},$$

we have

$$p_e \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \tag{89}$$

Step 3: combining the above results. Combining (87) and (89), we have that for any estimator ψ for θ, the error probability

$$\max_{\theta\in\Theta}\mathbb{P}_\theta(\psi \neq \theta) \geq \frac{1}{4}.$$

Then for the estimator $\widehat{\theta}$ constructed in (85), we have

$$\max_{\theta\in\Theta}\mathbb{P}_\theta(\widehat{\theta} \neq \theta) \geq \frac{1}{4}.$$

According to (86), we have for any policy estimator $\widehat{\pi}$,

$$\max_{\theta\in\Theta}\mathbb{P}_\theta\left(V_1^{\star,R}(0) - V_1^{\widehat{\pi},R}(0) \geq \varepsilon\right) \geq \max_{\theta\in\Theta}\mathbb{P}_\theta\left(\widehat{\theta} \neq \theta\right) \geq \frac{1}{4}.$$

Considering that this holds for arbitrary policy estimator $\widehat{\pi}$, we could prove that (84) holds true, and complete the proof.

Calculation of (82). Considering that we only consider about the value of $V_h^{\star,R}(0)$ and $V_h^{\star,R}(1)$, we rewrite vector $V_h^{\star,R} = [V_h^{\star,R}(0), V_h^{\star,R}(1)]^\top$ for convenience. According to robust Bellman's equation, we have

$$V_h^{\star,R} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 1-\widetilde{p} & \widetilde{p} \\ R & 1-R \end{bmatrix} V_{h+1}^{\star,R} =: r + P^{\star,R}V_{h+1}^{\star,R}, \tag{90}$$

with $V_{H+1}^{\star,R} = 0$, where we redefine

$$P^{\star,R} = P_h^{\star,V_{h+1}^{\star,R}}, \quad r = r_h.$$

By eigenvalue decomposition, $P^{\star,R}$ can be decomposed as

$$\begin{aligned}\begin{bmatrix} 1-\widetilde{p} & \widetilde{p} \\ R & 1-R \end{bmatrix} &= \begin{bmatrix} 1 & \widetilde{p} \\ 1 & -R \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1-R-\widetilde{p} \end{bmatrix}\begin{bmatrix} 1 & \widetilde{p} \\ 1 & -R \end{bmatrix}^{-1} \\ &= \frac{1}{R+\widetilde{p}}\begin{bmatrix} 1 & \widetilde{p} \\ 1 & -R \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1-R-\widetilde{p} \end{bmatrix}\begin{bmatrix} R & \widetilde{p} \\ 1 & -1 \end{bmatrix}.\end{aligned} \tag{91}$$

Thus we have

$$\begin{aligned}V_h^{\star,R} &= \sum_{h'=h}^{H}(P^{\star,R})^{h'-h}r = \frac{1}{R+\widetilde{p}}\begin{bmatrix} 1 & \widetilde{p} \\ 1 & -R \end{bmatrix}\begin{bmatrix} \sum_{h'=h}^{H}1 & 0 \\ 0 & \sum_{h'=h}^{H}(1-R-\widetilde{p})^{h'-h} \end{bmatrix}\begin{bmatrix} R & \widetilde{p} \\ 1 & -1 \end{bmatrix}r \\ &= \frac{1}{R+\widetilde{p}}\begin{bmatrix} \widetilde{p}\left(H-h+1-\frac{1-(1-R-\widetilde{p})^{H-h+1}}{R+\widetilde{p}}\right) \\ \widetilde{p}(H-h+1)+R\frac{1-(1-R-\widetilde{p})^{H-h+1}}{R+\widetilde{p}} \end{bmatrix}.\end{aligned}$$

Proof of (86). It suffices to prove that for policy estimation $\widehat{\pi}$ obtained from samples of $\mathcal{M}_\theta$, if $V_1^{\star,R}(0) - V_1^{\widehat{\pi},R}(0) \le \varepsilon$, then we have $\widehat{\theta} = \theta$, where $\widehat{\theta}$ is defined by (85).

Towards this, we assume that $\widehat{\theta} \neq \theta$. Then according to the definition of Θ in (79), and recalling $\widehat{\theta} \in \Theta$, we have $\left\|\theta - \widehat{\theta}\right\|_1 \ge H/8$, and

$$\begin{aligned}\sum_{h=1}^{H}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0)\right\|_1 &\ge \sum_{h=1}^{H}\left\|\pi_h^{\star,\theta}(\cdot|0) - \pi_h^{\star,\widehat{\theta}}(\cdot|0)\right\|_1 - \sum_{h=1}^{H}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\widehat{\theta}}(\cdot|0)\right\|_1 \\ &= 2\left\|\theta - \widehat{\theta}\right\|_1 - \sum_{h=1}^{H}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\widehat{\theta}}(\cdot|0)\right\|_1 \\ &\ge \frac{H}{4} - \sum_{h=1}^{H}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\widehat{\theta}}(\cdot|0)\right\|_1. \end{aligned} \tag{92}$$

To proceed, we claim that $V_1^{\star,R}(0) - V_1^{\widehat{\pi},R}(0) < \varepsilon$ implies

$$\sum_{h=1}^{H}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0)\right\|_1 < \frac{H}{8}. \tag{93}$$

Then according to (92), we have

$$\sum_{h=1}^{H}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0)\right\|_1 > \frac{H}{4} - \frac{H}{8} = \frac{H}{8},$$

which contradicts (93). Thus we prove that $\widehat{\theta} = \theta$.

Now it suffices to prove that the claim (93) holds true. Specifically, we intend to prove that if

$$\sum_{h=1}^{H}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0)\right\|_1 \ge \frac{H}{8},$$

then one has

$$V_1^{\star,R}(0) - V_1^{\widehat{\pi},R}(0) \ge \varepsilon.$$

Towards this, for any policy $\widehat{\pi}$, notice that $V_1^{\widehat{\pi},R} = [V_1^{\widehat{\pi},R}(0), V_1^{\widehat{\pi},R}(1)]^\top$ is computed as

$$V_1^{\widehat{\pi},R} = \sum_{h=1}^{H}\prod_{j=1}^{h-1} A_j^{\widehat{\pi}} r,$$

where $A_j^{\widehat{\pi}}$ is defined as

$$A_j^{\widehat{\pi}} = \begin{bmatrix} 1 - \tilde{z}_j^{\widehat{\pi}} & \tilde{z}_j^{\widehat{\pi}} \\ R & 1 - R \end{bmatrix}, \quad r = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

with

$$\tilde{z}_j^{\widehat{\pi}} = (1-R)\left(\widehat{\pi}_j(\theta_j|0)p + \widehat{\pi}_j(1-\theta_j|0)q\right). \tag{94}$$

It is obvious that $(1-R)q \le \tilde{z}_j^{\widehat{\pi}} \le (1-R)p$. Recall that the optimal value function $V_1^{\star,R} = [V_1^{\star,R}(0), V_1^{\star,R}(1)]^\top$ is calculated as (cf. (90))

$$V_1^{\star,R} = \sum_{h=1}^{H}\prod_{j=1}^{h-1} A_j^{\star} r,$$

where

$$A_j^{\star} = \begin{bmatrix} 1 - \widetilde{p} & \widetilde{p} \\ R & 1 - R \end{bmatrix},$$

and $\widetilde{p}$ is defined in (83). To decompose the difference between $V_1^{\star,R}$ and $V_1^{\widehat{\pi},R}$, we introduce a sequence of vectors $\widetilde{V}_0, \widetilde{V}_1, \cdots, \widetilde{V}_H$ defined as

$$\widetilde{V}_h = \sum_{h'=1}^{h+1} \prod_{j=1}^{h'-1} A_j^\star r + \sum_{h'=h+2}^{H} \prod_{j=1}^{h} A_j^\star \prod_{j=h+1}^{h'-1} A_j^{\widehat{\pi}} r,$$

We remark that $\widetilde{V}_h$ is actually the value function at step 1 for the policy $\widetilde{\pi} = \{\widetilde{\pi}_h\}_{h=1}^H$ satisfying $\widetilde{\pi}_j = \pi_j^{\star,\theta}$ for $j \le h$ and $\widetilde{\pi}_j = \widehat{\pi}_j$ for $j > h$. Moreover, we have $\widetilde{V}_H = V_1^{\star,R}$ and $\widetilde{V}_0 = V_1^{\widehat{\pi},R}$. Now we can decompose the difference between $V_1^{\star,R}$ and $V_1^{\widehat{\pi},R}$ as

$$V_1^{\star,R} - V_1^{\widehat{\pi},R} = \sum_{h=1}^{H} \left(\widetilde{V}_h - \widetilde{V}_{h-1} \right). \tag{95}$$

For the h-th term, we have

$$\begin{aligned}
\widetilde{V}_h - \widetilde{V}_{h-1} &= \sum_{h'=1}^{h+1} \prod_{j=1}^{h'-1} A_j^\star r + \sum_{h'=h+2}^{H} \prod_{j=1}^{h} A_j^\star \prod_{j=h+1}^{h'-1} A_j^{\widehat{\pi}} r - \sum_{h'=1}^{h} \prod_{j=1}^{h'-1} A_j^\star r - \sum_{h'=h+1}^{H} \prod_{j=1}^{h-1} A_j^\star \prod_{j=h}^{h'-1} A_j^{\widehat{\pi}} r \\
&= \prod_{j=1}^{h-1} A_j^\star \left(A_h^\star - A_h^{\widehat{\pi}} \right) \left(\sum_{h'=h+1}^{H} \prod_{j=h+1}^{h'-1} A_j^{\widehat{\pi}} r \right),
\end{aligned}$$

where $\prod_{j=h+1}^{h} A_j^{\widehat{\pi}} r$ is defined as r. We claim that

$$\begin{aligned}
\widetilde{V}_h - \widetilde{V}_{h-1} &\ge \prod_{j=1}^{h-1} A_j^\star \left(A_h^\star - A_h^{\widehat{\pi}} \right) \left(\sum_{h'=h+1}^{H} \prod_{j=h+1}^{h'-1} A_j^\star r \right) \\
&\overset{(i)}{=} (A^\star)^{h-1} \left(A_h^\star - A_h^{\widehat{\pi}} \right) \sum_{h'=0}^{H-h-1} (A^\star)^{h'} r,
\end{aligned} \tag{96}$$

where in (i) we rewrite $A_j^\star$ as $A^\star$ since it does not vary with j. The proof of (96) is postponed to the end of this section. According to the decomposition in (91), we have

$$\begin{aligned}
\sum_{h'=0}^{H-h-1} (A^\star)^{h'} r &= \frac{1}{R+\widetilde{p}} \begin{bmatrix} \widetilde{p}\left(H - h - \frac{1-(1-R-\widetilde{p})^{H-h}}{R+\widetilde{p}} \right) \\ \widetilde{p}(H-h) + R\frac{1-(1-R-\widetilde{p})^{H-h}}{R+\widetilde{p}} \end{bmatrix}, \\
(A^\star)^{h-1} &= \frac{1}{R+\widetilde{p}} \begin{bmatrix} 1 & \widetilde{p} \\ 1 & -R \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & (1-R-\widetilde{p})^{h-1} \end{bmatrix} \begin{bmatrix} R & \widetilde{p} \\ 1 & -1 \end{bmatrix}.
\end{aligned}$$

by inserting them, we have

$$\begin{aligned}
\widetilde{V}_h(0) - \widetilde{V}_{h-1}(0) &= \frac{\widetilde{p} - \widetilde{z}_h^{\widehat{\pi}}}{(R+\widetilde{p})^2} \left(R + \widetilde{p}(1-R-\widetilde{p})^{h-1} \right) \left(1 - (1-R-\widetilde{p})^{H-h} \right) \\
&\overset{(i)}{=} \frac{(1-R)(p-q)}{2(R+\widetilde{p})^2} \left(R + \widetilde{p}(1-R-\widetilde{p})^{h-1} \right) \\
&\quad \cdot \left(1 - (1-R-\widetilde{p})^{H-h} \right) \left\| \widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0) \right\|_1 \\
&\ge 0,
\end{aligned} \tag{97}$$

where (i) uses the definition of $\widetilde{p}$ and $\widetilde{z}_h^{\widehat{\pi}}$ in (83) and (94). Below we consider the case of $h \le 31H/32$, which gives

$$(1-R-\widetilde{p})^{H-h} \le (1-R-\widetilde{p})^{\frac{H}{32}} \le (1-p)^{\frac{H}{32}} \overset{(i)}{\le} 1 - \frac{c}{32},$$

where (i) holds because $p \ge \frac{c}{H} \ge 1 - (1-\frac{c}{32})^{\frac{32}{H}}$ for $c \le 3/4$. Thus we have

$$\widetilde{V}_h(0) - \widetilde{V}_{h-1}(0) \ge \frac{c(1-R)(p-q)}{64(R+\widetilde{p})^2} \left(R + \widetilde{p}(1-R-\widetilde{p})^{h-1} \right) \left\| \widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0) \right\|_1. \tag{98}$$

We shall proceed with a case-by-case analysis as below.

- Consider the case $R+\widetilde{p} \geq \frac{7}{8(h-1)}$. Notice that if $R \leq \frac{1}{H}$, then we have

$$R+\widetilde{p} = R + c(1-R)\max\left\{\frac{1}{H}, R\right\} = R + \frac{c(1-R)}{H} \overset{(i)}{\leq} R + \frac{3}{4H},$$

where (i) uses the fact that $c \leq 3/4$. This can be rewritten as

$$R \geq R+\widetilde{p} - \frac{3}{4H} \geq \frac{7}{8H} - \frac{3}{4H} = \frac{1}{8H}. \tag{99}$$

While if $R > \frac{1}{H}$, then $R \geq \frac{1}{8H}$ obviously. Inserting into the definition of p, we have

$$p = c\max\left\{\frac{1}{H}, R\right\} \leq c\max\{8R, R\} = 8cR.$$

This upper bound on p implies an upper bound on the following term.

$$\begin{aligned}
&\frac{(1-R)(p-q)}{(R+\widetilde{p})^2}\left(R+\widetilde{p}(1-R-\widetilde{p})^{h-1}\right) \\
&\geq \frac{(1-R)(p-q)R}{(R+\widetilde{p})^2} \geq \frac{(1-R)(p-q)R}{(R+8c(1-R)R)^2} \\
&= \frac{(1-R)(p-q)}{(1+8c(1-R))^2 R} \geq \frac{(1-R)(p-q)}{49R} \geq \frac{(1-R)(p-q)}{49\max\{\frac{1}{H}, R\}}.
\end{aligned}$$

- Consider the case $R+\widetilde{p} < \frac{7}{8(h-1)}$. In this case, we have

$$(1-R-\widetilde{p})^{h-1} \geq 1-(R+\widetilde{p})(h-1) \geq 1-\frac{7}{8} = \frac{1}{8}.$$

We further have

$$\begin{aligned}
&\frac{(1-R)(p-q)}{(R+\widetilde{p})^2}\left(R+\widetilde{p}(1-R-\widetilde{p})^{h-1}\right) \\
&\geq \frac{(1-R)(p-q)}{(R+\widetilde{p})^2}\left(R+\frac{\widetilde{p}}{8}\right) \geq \frac{(1-R)(p-q)(R+\widetilde{p})}{8(R+\widetilde{p})^2} \\
&= \frac{(1-R)(p-q)}{8(R+\widetilde{p})} \geq \frac{(1-R)(p-q)}{8(1+c)\max\{\frac{1}{H}, R\}} \geq \frac{(1-R)(p-q)}{14\max\{\frac{1}{H}, R\}}.
\end{aligned}$$

Summarizing the two cases above, we obtain the following inequality:

$$\frac{(1-R)(p-q)}{(R+\widetilde{p})^2}\left(R+\widetilde{p}(1-R-\widetilde{p})^{h-1}\right) \geq \frac{(1-R)(p-q)}{50}\min\left\{H, \frac{1}{R}\right\}.$$

Inserting into (98), we have

$$\widetilde{V}_h(0) - \widetilde{V}_{h-1}(0) \geq \frac{c(1-R)(p-q)}{3200}\min\left\{H, \frac{1}{R}\right\}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0)\right\|_1, \quad h \leq \frac{31H}{32}.$$

Combining with (95), we have

$$\begin{aligned}
V_1^{\star,R}(0) - V_1^{\widehat{\pi},R}(0) &\overset{\text{(i)}}{\geq} \sum_{h=1}^{\frac{31H}{32}}\left(\widetilde{V}_h - \widetilde{V}_{h-1}\right) \\
&\geq \frac{c(1-R)(p-q)}{3200}\min\left\{H, \frac{1}{R}\right\}\sum_{h=1}^{\frac{31H}{32}}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0)\right\|_1 \\
&\overset{\text{(ii)}}{\geq} \frac{c(1-R)(p-q)H}{51200}\min\left\{H, \frac{1}{R}\right\} \overset{\text{(iii)}}{\geq} \varepsilon.
\end{aligned} \tag{100}$$

where (i) holds because $\widetilde{V}_h - \widetilde{V}_{h-1} \geq 0$ for any h (c.f. (97)), (ii) holds because $\sum_{h=1}^{\frac{31H}{32}}\left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0)\right\|_1 \geq \left\|\widehat{\pi}_h(\cdot|0) - \pi_h^{\star,\theta}(\cdot|0)\right\|_1 - \frac{H}{16} \geq \frac{H}{16}$, and (iii) holds as long as $c_1 \geq \frac{51200}{c(1-c_R)}$.

Proof of (96). It suffices to prove that

$$\prod_{j=1}^{h-1} A_j^\star \left(A_h^\star - A_h^{\widehat{\pi}}\right) \left(\sum_{h'=h+1}^{H} \prod_{j=h+1}^{h'-1} A_j^{\widehat{\pi}} r\right) \geq \prod_{j=1}^{h-1} A_j^\star \left(A_h^\star - A_h^{\widehat{\pi}}\right) \left(\sum_{h'=h+1}^{H} \prod_{j=h+1}^{h'-1} A_j^\star r\right).$$

We remind that all entries in matrix $A_j^\star$ is positive. Thus it suffices to prove that

$$\left(A_h^\star - A_h^{\widehat{\pi}}\right) \left(\sum_{h'=h+1}^{H} \prod_{j=h+1}^{h'-1} A_j^{\widehat{\pi}} r\right) \geq \left(A_h^\star - A_h^{\widehat{\pi}}\right) \left(\sum_{h'=h+1}^{H} \prod_{j=h+1}^{h'-1} A_j^\star r\right). \tag{101}$$

To proceed, we remind that

$$A_h^\star - A_h^{\widehat{\pi}} = (\widetilde{p} - \widetilde{z}_h^{\widehat{\pi}}) \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}. \tag{102}$$

For convenience, we introduce two new notations:

$$x_k = \sum_{h'=k}^{H} \prod_{j=k}^{h'-1} A_j^\star r, \quad \widehat{x}_k = \sum_{h'=k}^{H} \prod_{j=k}^{h'-1} A_j^{\widehat{\pi}} r, \quad h+1 \leq k \leq H.$$

One can easily check that x_k and $\widehat{x}_k$ satisfy the following recursive formula:

$$x_k = r + A_k^\star x_{k+1}, \quad \widehat{x}_k = r + A_k^{\widehat{\pi}} \widehat{x}_{k+1}, \quad x_H = r, \quad \widehat{x}_H = r. \tag{103}$$

Combining with (102), proving (101) is equivalent to showing that

$$\widehat{x}_k(1) - \widehat{x}_k(0) \geq x_k(1) - x_k(0) \geq 0, \quad \forall h+1 \leq k \leq H. \tag{104}$$

We shall establish this by mathematical induction. Notice that (104) holds for $k = H$ due to (103). Assume that $\widehat{x}_{k+1}(1) - \widehat{x}_{k+1}(0) \geq x_{k+1}(1) - x_{k+1}(0)$. According to (103), we have

$$\begin{aligned} \widehat{x}_k(1) - \widehat{x}_k(0) &= 1 + R\widehat{x}_{k+1}(0) + (1-R)\widehat{x}_{k+1}(1) - (1 - \widetilde{z}_k^{\widehat{\pi}})\widehat{x}_{k+1}(0) - \widetilde{z}_k^{\widehat{\pi}} \widehat{x}_{k+1}(1) \\ &= 1 + (1 - R - \widetilde{z}_k^{\widehat{\pi}})(\widehat{x}_{k+1}(1) - \widehat{x}_{k+1}(0)) \\ &\geq 1 + (1 - R - \widetilde{p})(x_{k+1}(1) - x_{k+1}(0)). \end{aligned}$$

Moreover, by using (104), we have

$$\begin{aligned} x_k(1) - x_k(0) &= 1 + Rx_{k+1}(0) + (1-R)x_{k+1}(1) - (1-\widetilde{p})x_{k+1}(0) - \widetilde{p}x_{k+1}(1) \\ &= 1 + (1 - R - \widetilde{p})(x_{k+1}(1) - x_{k+1}(0)) \geq 0. \end{aligned}$$

In comparison, it is obvious that $\widehat{x}_k(1) - \widehat{x}_k(0) \geq x_k(1) - x_k(0)$. Thus we complete the proof.