Taming the Curses of Multiagency in Robust Markov Games with Large State Space through Linear Function Approximation

Jingchu Gai
Machine Learning Department
Carnegie Mellon University
jgai@andrew.cmu.edu

Laixi Shi
Department of Electrical and Computer Engineering
Johns Hopkins University
laixis@jhu.edu

Abstract

Multi-agent reinforcement learning (MARL) holds great potential but faces robustness challenges due to environmental uncertainty. To address this, distributionally robust Markov games (RMGs) optimize worst-case performance when the environment deviates from the nominal model within a uncertainty set. Beyond robustness, an equally urgent goal for MARL is data efficiency—sampling from vast state and action spaces that grow exponentially with the number of agents potentially leads to the curse of multiagency. However, current provably data-efficient algorithms for RMGs are limited to tabular settings with finite state and action spaces, which are only computationally manageable for small-scale problems, leaving RMGs with large-scale (or infinite) state spaces largely unexplored. The only existing work beyond tabular settings focuses on linear function approximation (LFA) for a restrictive class of RMGs using vanish minimal value assumption and still suffers from sample complexity with the curse of multiagency. In this work, we focuses on general RMGs with LFA. For uncertainty sets defined by total variation distance, we develop provably data-efficient algorithms that break the curse of multiagency in both the generative model setting and a newly proposed online interactive setting. To our knowledge, our results are the first to break the curse of multiagency of sample complexity for RMGs with large (possibly infinite) state spaces, regardless of the uncertainty set construction.

1 Introduction

Multi-agent reinforcement learning (MARL), formulated as Markov games (MGs) for strategic games, provides a powerful framework for sequential decision-making problems involving multiple agents interacting in a shared environment. MARL has demonstrated strong empirical success across a wide range of domains, including fine-tuning multi-LLM agent systems [56], autonomous driving [90], games [69], and ecosystem protection [23]. Due to the inherently conflicting objectives among agents, solution concepts for MGs are typically defined in terms of equilibrium notions: joint strategies under which no rational agent has an incentive to deviate unilaterally, assuming all other agents keep their strategies fixed. Widely studied examples include Nash equilibria (NE) [50, 63], correlated equilibria (CE), and coarse correlated equilibria (CCE) [4, 49].

Robustness becomes a central challenge in MARL, as learned solutions can be highly sensitive to subtle perturbations in human behavior and environmental dynamics, potentially leading to substantial degradation in both individual-agent performance and overall return [7, 83, 84, 86]. This motivates the development of algorithms that remain reliable under uncertainty—particularly in high-stakes or safety-critical applications—and has spurred growing interest in robust MARL through the lens of distributionally robust Markov games (RMGs) [39, 67, 86]. Unlike standard MGs, which

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Data Mechanism	Algorithm	Robust MG Model	Uncertainty Set	Break Curse	Sample Complexity/ Regret(T)
Generative Model	P^{2}MPO [12]	Tabular	$(s, \boldsymbol{a})$-rectangularity	No	$S^4 \left(\prod_{i=1}^n A_i\right)^3 \frac{H^4}{\varepsilon^2}$
	DR-NVI [67]	Tabular	$(s, \boldsymbol{a})$-rectangularity	No	$\frac{SH^3 \prod_{i=1}^n A_i}{\varepsilon^2} \min\left\{H, \frac{1}{\min_i \sigma_i}\right\}$
	Robust-Q-FTRL [66]	Tabular	fictitious (s, a_i)-rectangularity	Yes	$\frac{SH^6 \sum_{i=1}^n A_i}{\varepsilon^4} \min\left\{H, \frac{1}{\min_i \sigma_i}\right\}$
	This Work (Generative)	Linear, $d = S \max_{i\in[n]} A_i$	fictitious (s, a_i)-rectangularity	Yes	$\frac{H^9 d^2}{\varepsilon^4}$
Online Interaction	DRMG [73]	Tabular	$(s, \boldsymbol{a})$-rectangularity	No	$\tilde{\mathcal{O}}\left(\sqrt{\min\{\sigma_{\min}^{-1}, H\} HS(\prod_{i\in[n]} A_i)T}\right)$,
	DR-CCE-LSI [89]	Linear, $d_{\mathsf{curse}} = S \prod_{i\in[n]} A_i$	$(s, \boldsymbol{a})$-rectangularity	No	$d_{\mathsf{curse}} H \min\left\{H, \frac{1}{\min\{\sigma_i\}}\right\} \sqrt{T}$
	This Work (Online)	Linear, $d = S \max_{i\in[n]} A_i$	fictitious (s, a_i)-rectangularity	Yes	$d \max_{i\in[n]} A_i \, H^2 \sqrt{T}$

Table 1: The existing results of sample complexity in the generative model setting (or regret in the online setting) for finding an ε-approximate equilibrium in finite-horizon, multi-agent, general-sum robust Markov games (RMGs), omitting logarithmic factors. To our knowledge, this work provides the first algorithms that break the curse of multiagency for large (possibly infinite) state spaces using linear function approximation.

assume a fixed environment, RMGs explicitly pursue robustness by optimizing worst-case performance over environments within a prescribed uncertainty set.

Beyond robustness, an equally critical goal in MARL is data efficiency. In extensive real-world applications, agents must sample from environments with vast state and action spaces. This challenge is further amplified in MARL by the exponential growth of the joint action space as the number of agents increases, leading to potentially exploding sample complexity—a phenomenon known as the *curse of multiagency*. Consequently, substantial research effort has focused on designing provably sample-efficient algorithms not only for RMGs [12, 66] but also for broader MARL problems [37, 75].

However, current sample-efficient algorithms are often limited to tabular RMGs with finite state and action spaces. In practice, (multi-agent) RL problems inherently involve large-scale or even continuous and infinite-dimensional state spaces—e.g., transportation with autonomous vehicles [28], robotics and swarm manipulation [53], water and energy systems [32], and financial markets [48]. Scaling learning algorithms to such large-scale state spaces poses significant data and computational challenges. To address this, function approximation—such as deep neural networks—is widely employed in RL [15, 68, 74, 75]. Nevertheless, incorporating function approximation into provably sample-efficient solutions for RMGs remains largely unexplored. The only existing work [89] studies linear function approximation in an online setting, focusing on a restrictive class of RMGs [35, 75] and requiring an additional vanishing minimal value assumption to obtain meaningful guarantees; moreover, its sample complexity still suffers from the curse of multiagency. This raises an open question:

Can the curse of multiagency be tamed in robust Markov games with large-scale state spaces?

To address this gap, to the best of our knowledge, this work provides the first provably sample-efficient algorithm for RMGs with large-scale (or infinite) state spaces that avoids the curse of multiagency. In particular, we focus on linear function approximation. Let S and $A_1, \ldots, A_n$ denote the cardinality of the state space and the action spaces of n agents, respectively; let d denote the feature dimension of the linear function class for each state-action pair (see (2) for details). Let H denote the horizon (number of steps per episode), and let T denote the total number of episodes in the online learning process. The main contributions are summarized as below:

• We focus on *distributionally robust linear Markov games* (R-LMGs), i.e., RMGs with linear function approximation. Under the data collection mechanism with a generative model, we develop a provable algorithm that achieves an ε-approximate CCE for R-LMGs using total variation uncertainty set, with sample complexity at most $\tilde{\mathcal{O}}(H^9 d^2/\varepsilon^4)$ (cf. Theorem 2). To the best of our knowl-

edge, this is the first provable sample-complexity guarantee for R-LMGs under the generative model setting, irrespective of the uncertainty-set formulation; moreover, it effectively overcomes the curse of multi-agency. Technically, unlike tabular RMGs, large-scale (possibly infinite) state spaces make exhaustive sampling over all state–action pairs infeasible. We therefore employ a infinite-to-finite reduction to sample a carefully chosen finite subset and generalize to the full state space.

• Beyond the generative model setting, we consider a more practical data-collection regime—an online interactive protocol in which agents sample Markovian trajectories from transition kernels within the uncertainty set of R-LMGs, rather than solely from the nominal kernel. In this setting, we propose a sample-efficient algorithm that attains sublinear regret $\tilde{\mathcal{O}}\big(d\,\max_{i\in[n]} A_i\, H^2\sqrt{T}\big)$ (cf. Theorem 3), thereby breaking the curse of multiagency. Technically, since the true robust value function is unknown, we cannot sample directly from the worst-case transition kernel and must instead approximate the adversarial environment. To this end, beyond the optimistic value estimates commonly used in standard linear MGs for policy updates, we introduce a new sampling strategy that leverages an additional sequence of pessimistic robust value estimates to approximate the adversarial model, combined with a hybrid sampling scheme to establish our guarantees.

2 Problem Formulation

2.1 Background: standard linear Markov games

Finite-horizon multi-agent general-sum Markov games (MGs). We consider an episodic finite-horizon general-sum MG represented by the tuple $(H, \mathcal{S}, \{\mathcal{A}_i\}_{i\in[n]}, P, \{r_i\}_{i\in[n]})$, where H denotes the length of each episode and $\mathcal{S}$ represents the state space. The joint action space for all agents is defined as $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$, where, for each $i \in [n]$, $\mathcal{A}_i$ is a finite action set of agent i with cardinality A_i. We use $\mathbf{a} \in \mathcal{A}$ to represent the joint action of all agents, and $\mathbf{a}_{-i}$ to denote the joint action of all agents except for agent i. Additionally, $P = \{P_h\}_{h\in[H]}$ is a collection of transition kernels, where $P_h(\cdot \mid s, \mathbf{a})$ gives the distribution of the next state when agents take action $\mathbf{a}$ at state s and step h. For each agent i, $r_i = \{r_{i,h}\}_{h\in[H]}$ represents the reward functions, where agent i receives $r_{i,h}(s, \mathbf{a}) \in [0, 1]$ when all agents take action $\mathbf{a}$ at state s and step h.

To continue, we focus on Markov policies, where the action selection rule depends only on the current state, independent of the previous trajectory. Specifically, for each agent $i \in [n]$, actions are chosen according to the policy $\pi_i = \{\pi_{i,h} : \mathcal{S} \to \Delta(\mathcal{A}_i)\}_{1\le h\le H}$. Here, $\pi_{i,h}(a_i \mid s)$ denotes the probability of selecting action $a_i \in \mathcal{A}_i$ in state s at time step h.The joint policy for all agents is represented as $\pi = \pi_1 \times \pi_2 \times \cdots \times \pi_n : \mathcal{S} \times [H] \to \Delta(\mathcal{A})$. Additionally, we define $\pi_{-i} = \otimes_{j\neq i}\pi_j$ to denote the joint policy of all agents except for agent i. For any joint policy π and transition kernel P, the long-term cumulative reward for agent i is captured by the value function $V_{i,h}^{\pi,P} : \mathcal{S} \to \mathbb{R}$ and Q-function $Q_{i,h}^{\pi,P} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, defined for all $(h, s, \mathbf{a}) \in [H] \times \mathcal{S} \times \mathcal{A}$ by $Q_{i,h}^{\pi,P}(s, \mathbf{a}) := \mathbb{E}_{\pi,P}[\sum_{t=h}^{H} r_{i,t}(s_t, \mathbf{a}_t) \mid s_h = s, \mathbf{a}_h = \mathbf{a}]$ and $V_{i,h}^{\pi,P}(s) := \mathbb{E}_{\pi,P}[\sum_{t=h}^{H} r_{i,t}(s_t, \mathbf{a}_t) \mid s_h = s]$ (1), where the expectation is over the trajectory $\{(s_t, \mathbf{a}_t)\}_{h\le t\le H}$ generated by π under P.

Linear Markov games: MGs with per-agent independent linear function approximation. In this work, we focus on MGs with linear function approximation (LFA), abbreviated by linear MGs later, assuming that each agents' transitions and reward functions can be described by independent linear functions, as proposed by Cui et al. [15], Dai et al. [16], Wang et al. [75]. Specifically, in any linear MG with transition kernel $P = P^0$, for each agent $i \in [n]$, there exists a known d-dimensional feature mapping $\phi_i : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}^d$ such that for all joint policy π and $(i, h, s, a_i) \in [n] \times [H] \times \mathcal{S} \times \mathcal{A}_i$, $\mathbb{E}_{a_{-i}\sim\pi_{-i}(s)}[P_h^0(\cdot \mid s, (a_i, \mathbf{a}_{-i}))] = \phi_i(s, a_i)^\top \mu_{i,h}^{\pi_{-i}}$ and $\mathbb{E}_{a_{-i}\sim\pi_{-i}(s)}[r_{i,h}(s, (a_i, \mathbf{a}_{-i}))] = \phi_i(s, a_i)^\top \theta_{i,h}^{\pi_{-i}}$ (2), where $\mu_{i,h}^{\pi_{-i}} = (\mu_{i,h}^{\pi_{-i},(1)}, \ldots, \mu_{i,h}^{\pi_{-i},(d)})^\top$ has d rows, each row $\mu_{i,h}^{\pi_{-i},(j)}$ a signed measure over $\mathcal{S}$, and $\theta_{i,h}^{\pi_{-i}} \in \mathbb{R}^d$ is the unknown reward parameter. We adopt the standard normalization $\|\theta_{i,h}^{\pi_{-i}}\|_2 \le d$ and $\|\mu_{i,h}^{\pi_{-i}}(s)\|_2 \le \sqrt{d}$ for all (s, h).

Remark. The above assumption on the linear structure (cf. (2)) is strictly weaker than that in the closely related prior work [89, Assumption 3.1], which also studies robust linear MGs but adopts a centralized linear structure. Specifically, Zheng and Lin [89] requires the transition kernel $P_h(s' \mid s, \mathbf{a})$ for each state-joint action $(s, \mathbf{a})$ to admit a linear representation proportional to a feature vector $\phi(s, \mathbf{a})$.

2.2 Distributionally robust linear MGs

To consider robustness for linear MGs, we focus on a distributionally robust linear Markov game (R-LMG), characterized by the tuple $\{\mathcal{S}, \{\mathcal{A}_i\}_{i\in[n]}, \{\mathcal{U}_\rho^{\sigma_i}(P^0,\cdot)\}_{i\in[n]}, r, H\}$. The key difference between an R-LMG and a standard linear MG is that, in an R-LMG, the transition kernel for each agent i can vary within its own uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0,\cdot)$ around a nominal transition kernel P^0. Note that the nominal transition P^0 and r satisfy the linear function approximation described in (2), while other transitions within the uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0,\cdot)$ do not need to satisfy LFA or any other modeling assumptions. We shall specify the uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0,\cdot)$ with ρ and σ_i shortly.

Fictitious uncertainty set. Robust MGs (and also robust linear MGs) form a broad class due to the flexibility in specifying uncertainty sets that capture different notions of model mismatch and shifts. For a generic uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0,\cdot)$, ρ represents a divergence function that quantifies the discrepancy between transition kernels, and $\sigma_i \geq 0$ controls the uncertainty radius for agent i. Various choices of the divergence function ρ can be considered, including but not limited to f-divergence family (such as total variation and Kullback-Leibler (KL) divergence), Wasserstein distance Xu et al. [81], and learning-based approximation Ding et al. [20]. In this work, we adopt the fictitious uncertainty sets $\mathcal{U}_\rho^{\sigma_i}(P^0,\pi)$ from Shi et al. [66], which model each agent as addressing and responding to the aggregate uncertainty induced by all other agents, rather than decomposing uncertainty into agent-by-agent individual uncertainty. This aligns with realistic human decision-making under uncertainty, as observed in behavioral economics [27, 60].

To specify, we firstly introduce some useful notations. For all $(s,\mathbf{a},h) \in \mathcal{S}\times\mathcal{A}\times[H]$, let $P^0_{h,s,\mathbf{a}} \in \Delta(\mathcal{S})$ denote the nominal transition kernel with $P^0_{h,s,\mathbf{a}}(s') = P^0_h(s' \mid s,\mathbf{a})$ for all $s' \in \mathcal{S}$, and let $\mathcal{A}_{-i} := \prod_{j\neq i}\mathcal{A}_j$. We define the expected nominal transition kernel conditioned on agent i selecting action a_i while the other agents follow a conditional policy as $P^{\pi_{-i}}_{h,s,a_i}(s') := \sum_{\mathbf{a}_{-i}\in\mathcal{A}_{-i}} \frac{\pi_h(a_i,\mathbf{a}_{-i}|s)}{\pi_{i,h}(a_i|s)} P^0_{h,s,\mathbf{a}}(s')$ for all $s' \in \mathcal{S}$ (3). Armed with this notation, for any joint policy π and uncertainty levels $\{\sigma_i\}$, the uncertainty set is $\mathcal{U}_\rho^{\sigma_i}(P^0,\pi) = \otimes\mathcal{U}_\rho^{\sigma_i}(P^{\pi_{-i}}_{h,s,a_i})$, where $\mathcal{U}_\rho^{\sigma_i}(P^{\pi_{-i}}_{h,s,a_i}) := \{P\in\Delta(\mathcal{S}) : \rho(P, P^{\pi_{-i}}_{h,s,a_i}) \leq \sigma_i\}$ (4).

Robust value functions. In a robust linear MG, each agent optimizes its worst-case performance. For a joint policy π, agent i's robust value function and robust Q-function are defined for all (i,h,s,a_i) as $Q^{\pi,\sigma_i}_{i,h}(s,a_i) := \inf_{P\in\mathcal{U}_\rho^{\sigma_i}(P^0,\pi)} Q^{\pi,P}_{i,h}(s,a_i)$ and $V^{\pi,\sigma_i}_{i,h}(s) := \inf_{P\in\mathcal{U}_\rho^{\sigma_i}(P^0,\pi)} V^{\pi,P}_{i,h}(s)$ (5). Notably, unlike in (1), the Q-function for agent i is evaluated based solely on its individual action a_i rather than the joint action $\mathbf{a}$. The robust best-response benchmark is, for each agent i, given the policies of the other agents, $V^{\star,\pi_{-i},\sigma_i}_{i,h}(s) := \max_{\pi_i':\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} V^{\pi_i'\times\pi_{-i},\sigma_i}_{i,h}(s)$ (6), and any maximizer π_i' is a robust best response of agent i.

Solution concepts: robust equilibria. Analogous to standard MGs, we introduce two types of equilibria as the learning objectives in R-LMGs, the robust variants — robust Nash equilibrium (NE) and robust coarse correlated equilibrium (CCE):

• Robust Nash equilibrium (Robust NE): A product policy $\pi = \pi_1\times\cdots\times\pi_n$ is a robust NE if $V^{\pi,\sigma_i}_{i,1}(s) = V^{\star,\pi_{-i},\sigma_i}_{i,1}(s)$ for all $(i,s)\in[n]\times\mathcal{S}$ (7). In other words, fixing other agents' policy π_{-i}, agent i cannot improve its worst-case outcome by unilaterally changing its own policy.

• Robust coarse correlated equilibrium (Robust CCE): A distribution ξ over joint product policies is a robust CCE if, for all $(i,s)\in[n]\times\mathcal{S}$, $\mathbb{E}_{\pi\sim\xi}[V^{\pi,\sigma_i}_{i,1}(s)] \geq \max_{\pi_i'}\mathbb{E}_{\pi\sim\xi}[V^{\pi_i'\times\pi_{-i},\sigma_i}_{i,1}(s)]$. Intuitively, when $\pi\sim\xi$, no agent can increase its expected robust value by committing to an alternative policy.

Building on the definition of the fictitious uncertainty set, the following theorem establishes the existence of both a robust NE and a robust CCE in robust linear MGs (R-LMGs). The proof is deferred to Appendix D.

Theorem 1 (Existence of robust NE)**.** *For any robust linear MG $\{\mathcal{S}, \{\mathcal{A}_i\}_{1\leq i\leq n}, \{\mathcal{U}_\rho^{\sigma_i}(P^0,\cdot)\}_{1\leq i\leq n}, r, H\}$ with linear approximation (cf.* (2)*) and fictitious uncertainty sets, there exists at least one robust NE, which further implies the existence of at least one robust CCE.*

This theorem extends existing robust-NE existence results in [66] for robust MGs with finite state spaces to settings with possibly infinite state spaces, by combining Glicksberg's fixed-point theorem [26] with additional analytic arguments.

Specification of the uncertainty set: total variation (TV). In this work, we adopt the TV distance as the divergence function ρ, motivated by its simplicity, practical utility [41, 54, 71], and theoretical advantages [65, 67]: for any $P, P' \in \Delta(\mathcal{S})$, $\|P - P'\|_{d_{\mathsf{TV}}} = \frac{1}{2}\sum_{s\in\mathcal{S}} |P(s) - P'(s)|$. For simplicity, we denote $\mathcal{U}^{\sigma_i}(\cdot) \coloneqq \mathcal{U}^{\sigma_i}_{d_{\mathsf{TV}}}(\cdot)$ throughout the rest of the paper.

3 Robust Linear MGs with a Generative Model

In both single-agent and multi-agent RL, three data-collection mechanisms are widely used: the generative model setting [42, 88], the offline setting [11], and the online setting [34]. In this section, we begin with the fundamental generative model setting, in which we assume access to a generative model or simulator that provides samples $s' \sim P^0_h(\cdot \mid s, \mathbf{a})$ from the nominal environment P^0 for any given state, joint action, and time-step tuple $(s, \mathbf{a}, h) \in \mathcal{S} \times \mathcal{A} \times [H]$.

To learn equilibria in MARL, we focus on finding an *ε-approximate CCE* using as few samples as possible. This is motivated by the fact that computing a Nash equilibrium is generally intractable [17], while finding an exact robust equilibrium is computationally hard and often impractical [1, 76]. Mathematically, our goal is to find a distribution $\xi = \{\xi_h\}_{h\in[H]} : [H] \to \Delta(\mathcal{S} \to \prod_{i\in[n]} \Delta(\mathcal{A}_i))$ over policies such that

$$\max_{s\in\mathcal{S}, i\in[n]} \left\{ \mathbb{E}_{\pi\sim\xi}\left[V^{\star,\pi_{-i},\sigma_i}_{i,1}(s)\right] - \mathbb{E}_{\pi\sim\xi}\left[V^{\pi,\sigma_i}_{i,1}(s)\right] \right\} \le \varepsilon. \tag{8}$$

3.1 Algorithm Design

To solve robust linear MGs (R-LMGs), we propose L-Robust-Q-FTRL (outlined in Algorithm 2). Motivated by Robust-Q-FTRL algorithm developed for robust MGs in tabular setting [66], it incorporates designs tailored to linear function approximation (LFA) for large-scale possibly infinite state space, through an infinite-to-finite sampling design and leveraging ridge regression to estimate the transition probabilities and reward functions. Due to the space limit, we will introduce this key module of infinite-to-finite sampling for nominal model estimation, which differs fundamentally from the tabular robust MG setting. The details of the entire algorithm are postponed to the Appendix B.

Nominal model estimation: constructing finite set reflecting (infinite) state space. Unlike the tabular case, where the state space is finite and relatively small, sampling costs become prohibitive when enumerating all state-action pairs in a large-scale (or infinite) state space. To address this challenge, we employ the following lemma, a widely-used tool for linear Markov games [75]:

Lemma 1. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact, full-dimensional set [1] There exists a distribution ρ supported on at most $d(d+1)/2$ points in $\mathcal{X}$ such that, defining $\Sigma = \mathbb{E}_{x\sim\rho}[xx^\top]$, we have Σ is invertible and $\|x\|^2_{\Sigma^{-1}} \le d$ for all $x \in \mathcal{X}$.*

The proof of Lemma 1 can be found in Lattimore and Szepesvári [40]. Without loss of generality, we assume that the feature set of each agent i, denoted by $\mathcal{X}_i \coloneqq \{\phi_i(s, a_i)\}_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}$, is full-dimensional. If this is not the case, the model parameters can be represented using a lower-dimensional linear parameterization (i.e., a smaller d). Applying Lemma 1, we can construct a support set of state-action pairs $\mathcal{Y}_i \subset \mathcal{S} \times \mathcal{A}_i$ with $|\mathcal{Y}_i| \le d(d+1)/2$, and a distribution ρ_i over $\mathcal{Y}_i$ such that $\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \|\phi_i(s, a_i)\|^2_{\Sigma_i^{-1}} \le d$, where $\Sigma_i = \mathbb{E}_{(s,a_i)\sim\rho_i}[\phi_i(s, a_i)\phi_i(s, a_i)^\top]$. Given the finite set $\mathcal{Y}_i$, we query the generative model $\lceil N\rho_i(s, a_i)\rceil$ times for each $(s, a_i) \in \mathcal{Y}_i$. This yields a total sample size of at most $N + d(d+1)/2$. Crucially, this sampling design relies only on the finite set $\mathcal{Y}_i$, making the process feasible even when the original state-action space $\mathcal{X}_i$ is infinite.

Armed with the constructed finite set $\mathcal{Y}_i$ for each agent i, we estimate nominal transitions and rewards via ridge regression. L-Robust-Q-FTRL then follows an adversarial online learning procedure similar to Robust-Q-FTRL [42], using backward recursion from $h = H$ to $h = 1$. At each

[1] Here, full-dimensional implies that the linear span of $\mathcal{X}$ covers the entire space, i.e., span$(\mathcal{X}) = \mathbb{R}^d$. Namely, there exists a distribution ρ over $\mathcal{X}$ so that the second moment matrix $\Sigma = \mathbb{E}_{x\sim\rho}[xx^\top]$ is positive definite and invertible. See Bertsekas [8].

step h, the algorithm executes a K-iteration loop that samples around size N based on $\mathcal{Y}_i$, estimates the nominal model, and updates the robust Q-function via the dual form of robust Bellman equation [30] (cf. (12) in Appendix B). After completing the K iterations, it constructs an optimistic robust value function by adding a bonus term $\beta_{i,h}$. See Appendix B for more details.

3.2 Theoretical Guarantees

We provide the theoretical guarantee for the sample complexity of L-Robust-Q-FTRL in Algorithm 2.

Theorem 2. *Consider any robust linear MG $\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le n}, \{\mathcal{U}^{\sigma_i}(P^0,\cdot)\}_{1\le i\le n}, r, H\}$ with uncertainty level $\{\sigma_i\}_{i\in[n]}$ and some $\delta \in (0,1)$. Algorithm 2 outputs an ε-approximate CCE ξ, i.e., $\max_{s\in\mathcal{S},\, i\in[n]}\{\mathbb{E}_{\pi\sim\xi}[V_{i,1}^{\star,\pi_{-i},\sigma_i}(s)] - \mathbb{E}_{\pi\sim\xi}[V_{i,1}^{\pi,\sigma_i}(s)]\} \le \varepsilon$ with probability at least $1-\delta$, provided that $N \ge CH^4d^2\ln(KNnH/\delta)/\varepsilon^2$ and $K \ge CH^4\ln(\max_{i\in[n]} A_i)/\varepsilon^2$ for some universal constant C. Consequently, the total number of samples drawn during learning satisfies*

$$N_{\mathsf{all}} \le HK(N + d(d+1)/2) = \tilde{\mathcal{O}}\big(H^9d^2/\varepsilon^4\big).$$

The implications of the above theorem are as below:

Breaking the curse of multi-agency in large state spaces. To the best of our knowledge, this provides the first provable sample complexity guarantees for robust linear MGs using a generative model, irrespective of the uncertainty set formulation. To illustrate the implications of our results, we consider the reduction to the tabular setting, which corresponds to $d = S\max_{i\in[n]} A_i$. Substituting this into Theorem 2 yields a complexity of $\tilde{\mathcal{O}}\big(H^9S^2\max_{i\in[n]}(A_i)^2/\varepsilon^4\big)$, which depends polynomially on all salient parameters. This demonstrates that L-Robust-Q-FTRL is a sample-efficient algorithm for robust linear MGs, effectively overcoming the curse of multi-agency in the generative model setting. Comparisons with existing results in the tabular setting are summarized in Table 1 (in Appendix C's related-work section).

Technical insights. In contrast to robust Markov games in the tabular setting, estimating the transition model and reward function in large-scale (possibly infinite) state spaces presents a challenge, since sampling every state-action pair—as done in tabular robust Markov games—is infeasible. To address this, we construct a specific finite subset of the state-action space to sample from; this enables us to obtain sufficient model information for the entire space, followed by ridge regression to estimate the parameters. On the other hand, compared to standard linear MGs, the robust value function in R-LMGs is highly nonlinear with respect to the nominal transition kernel. As a result, the statistical error to be controlled is no longer a linear aggregation of the estimation errors across K iterations, which typically cancel out in standard linear MGs. We address this challenge by drawing N samples in each iteration (compared to a single sample in standard linear MGs) and employing a tailored decomposition of the robust value function to control the auxiliary statistical error, ultimately achieving a sharp sample complexity guarantee.

4 Robust Linear MGs with Online Adversarial Environment

In this section, we move beyond the generative model setting and focus on learning R-LMGs in a more practical *online interactive setting*, where agents interact with the environment and collect Markovian trajectories, reflecting many real-world applications [47, 74]. Notably, the online setting in RMGs can be substantially more challenging than the generative model setting, due to the need to balance exploration and exploitation and to cope with statistical dependence along sampled trajectories.

In contrast to the generative model setting discussed in Section 3, the online setting concerns not only the final output policy but also the performance throughout the entire exploration process, relative to the best possible performance in hindsight. Specifically, over a T-round online process, the goal is to minimize regret incurred along the way. Let $\xi^t : [H] \to \Delta\big(\mathcal{S} \to \Delta(\prod_{i\in[n]}\mathcal{A}_i)\big)$ denote the distribution over product policies executed by the algorithm at round $t \in [T]$. The regret is then

defined as

$$\text{Regret}(T) \coloneqq \max_{i\in[n]} \sum_{t=1}^{T} \left\{ \mathbb{E}_{\pi\sim\xi^t}\left[V_{i,1}^{\star,\pi_{-i},\sigma_i}(s_1)\right] - \mathbb{E}_{\pi\sim\xi^t}\left[V_{i,1}^{\pi,\sigma_i}(s_1)\right] \right\}. \tag{9}$$

4.1 Interactive Data Collection within the Uncertainty Set

Many online interaction paradigms in practical studies allow agents to interact not only with the nominal environment but also with perturbed environments for adversarial training aimed at safety and sim-to-real transfer [2, 19]. For instance, to address the sim-to-real gap in robotics [2] and safety in autonomous driving [19], the training environment is explicitly generated using approximate worst-case parameters or scenarios, actively configured to expose the agent's vulnerabilities. This interaction paradigm is also reflected in recent theoretical works: Jackson et al. [31] and Zhang et al. [87] optimize policies against estimated adversarial environments, while Ren and Majumdar [57] approximate worst-case models within a Wasserstein ball centered on nominal estimates.

Motivated by such widely used online interaction paradigms, we propose an practically meaningful online protocol in which agents may sample from transition kernels within the prescribed uncertainty set, rather than only from the nominal kernel. Specifically, we can sample from the nominal environment and also utilize the following sampling mechanism interacting with an "adversarial environment" inside the uncertainty set, defined as:

Definition 1 (**Interaction with an Adversarial Environment**). *For any agent $i \in [n]$, given a collection of vectors $\{V_h : \mathcal{S} \to [0,H]\}_{h\in[H]}$ (which need not be the robust value functions) and a joint policy π, we are allowed to sample trajectories $(s_1, \mathbf{a}_1, r_1(s_1,\mathbf{a}_1), \ldots, \mathbf{a}_{H-1}, r_{H-1}(s_1,\mathbf{a}_1), s_H)$, where the transitions following $s_{h+1} \sim P_{i,h,s_h,a_i}^{\pi_{-i},V_{h+1}}(\cdot)$ with $P_{i,h,s_h,a_i}^{\pi_{-i},V_{h+1}} \coloneqq \arg\min_{P\in\mathcal{U}^{\sigma_i}(P_{h,s_h,a_i}^{\pi_{-i}})} PV_{h+1}$.*

Namely, we are allowed to interact with and sample from the *still unknown* approximate adversarial environment induced by the input approximate robust value function vectors $\{V_h\}_{h\in[H]}$. This environment lies within the fictitious uncertainty set $\mathcal{U}^{\sigma_i}(P_{h,s_h,a_i}^{\pi_{-i}})$ (see (4)).

Comparison to prior online interaction mechanisms. Prior works on robust RL with online data collection [44, 89] consider agents to interact with *only the fixed nominal environment*. However, this mechanism is inherently ill-posed: even for a special subclass of tabular robust Markov games (single-agent robust Markov decision processes), achieving sublinear regret is generally impossible in the worst case, as noted by Lu et al. [44, Theorem 3.2]. Consequently, analyses in this setting are typically established under additional assumptions on the RMGs considered, namely, that the robust value function satisfies a vanishing minimality condition [44, 89].

Challenges. We emphasize that the proposed online interaction setting introduces *unique* challenges beyond those in standard linear MGs and in existing online interaction settings considered for R-LMGs.

1. No degeneration to standard linear Markov games. Even if we can query a worst-case transition kernel with respect to a given value function, without access to the ground-truth robust value function, it is impossible to identify and sample from the exact worst-case transition kernel within the uncertainty set; therefore, the problem does not degenerate to solving a standard linear MG. The key challenge is therefore to approximate the worst-case environment using progressively more accurate estimates of the robust value function.

2. The sampled adversarial environment can be non-linear. Prior work [89] focuses on sampling only in the nominal transition environment satisfying linear function approximation, where the samples from the nominal model can be directly leveraged for estimating a linear model. While in our problems, we sample from both the nominal kernel and online adversarial environments lacking the linearity assumption. Consequently, the collected trajectories may be generated by a potentially non-linear model, which requires extra care in model estimation and in the design of subsequent procedures based on these samples.

4.2 Algorithm Design

To solve R-LMGs in the proposed online interactive setting, we develop Online-L-Robust-Q-FTRL (summarized in Algorithm 3), which achieves an ε-approximate CCE for general R-LMGs (cf. (8))

and simultaneously minimizes regret. Compared to the generative model setting in Section 3, the main challenge is to design a sampling procedure that mimics drawing Markovian trajectories under the exact worst-case environment using only an estimated robust value function. Before presenting the full algorithmic pipeline, we first describe one key component: the sampling strategy employed by Online-L-Robust-Q-FTRL. This strategy differs substantially from those used in the generative model setting and from those for standard linear Markov games [37, 75], which serve as inspiration for our approach.

Algorithm 1: Hybrid-Sampling

1: **Input:** Agent i, step h, episode t, distribution ξ_h^l, values $\{V_j^l\}_{(j,l)\in[H]\times[t-1]}$.
2: Repeat for each $l \in [t-1]$, iterating $\lceil N/t \rceil$ times:
3: *// Sample from the approximate adversarial environment for the first $h-1$ steps*
4: **for** $j \in [h-1]$ **do**
5: *// Random policy draw from historical FTRL iterates at step j.*
6: Select $\pi \sim \mathrm{Unif}\big(\{\otimes_{i'\in[n]}\pi_{i',j}^{l,k}\}_{k\in[K]}\big)$, take action $\mathbf{a}_j \sim \pi(\cdot \mid s_j)$
7: Collect state sample $s_{j+1} \sim P_{i,j,s_j,a_{i,j}}^{\pi_{-i},V_{j+1}^l}(\cdot)$ (see Definition 1))
8: *// At step h, explore a_i uniformly; opponents from past policies.*
9: Select $\pi_{-i} \sim \mathrm{Unif}(\{\otimes_{i'\neq i}\pi_{i',h}^{l,k}\}_{k\in[K]})$, $a_{-i,h} \sim \pi_{-i}(s_h)$, $a_i \sim \mathrm{Unif}(\mathcal{A}_i)$.
10: Collect $s_{h+1} \sim P_{h,s_h,a_{i,h}}^{\pi_{-i}}(\cdot)$ (defined in (3)).
11: **Output:** Samples are collected in
$\mathcal{D}_i = \{(s_1^{(m)}, \mathbf{a}_1^{(m)}, \mathbf{r}_1^{(m)}, \ldots, \mathbf{a}_h^{(m)}, \mathbf{r}_h^{(m)}, s_{h+1}^{(m)})\}_{m=1}^{|\mathcal{D}_i|}$.

Collecting online trajectories in the approximate adversarial environment. At the t-th online sampling round, for $t \in [T]$, we run a K-iteration procedure at each time step $h \in [H]$. For a fixed round $t \in [T]$, step $h \in [H]$, and iteration k, let $\pi_h^{t,k}$ denote the current learned policy, and apply the hybrid sampling scheme in Algorithm 1. Concretely, for each agent i, we collect N samples by interacting with an approximate adversarial environment for the first $h-1$ steps, and then sampling from the nominal environment only at the final step h. The approximate adversarial environment in Algorithm 1 is specified by querying an additional collection of pessimistic robust value estimates $\{\underline{V}_j^l\}_{(j,l)\in[H]\times[t-1]}$. These auxiliary pessimistic estimates are designed to handle robust linear Markov games and are crucial for sampling under adversarial dynamics, a difficulty that does not arise in standard linear Markov games [37, 75].

Specifically, for a given agent i, each $l \in [t-1]$, and each step $j \in [h-1]$, the agents first select a joint policy π uniformly from the set $\{\otimes_{i'\in[n]}\pi_{i',j}^{l,k}\}_{k\in[K]}$ of K policies from round l. For the rest of the paper, we denote $a_{i,h}$ as the action of the i-th agent at time step h for all $(i,h) \in [n]\times[H]$. They then sample an action $\mathbf{a}_j \sim \pi(\cdot \mid s_j)$ (Line 6) and observe the next state s_{j+1} from the adversarial kernel $P_{i,j,s_j,a_{i,j}}^{\pi_{-i},\underline{V}_{i,j+1}^l}(\cdot)$, as in Definition 1. At step h, we switch strategies: the opponents' joint policy π_{-i} is selected uniformly from $\{\otimes_{i'\neq i}\pi_{i',h}^{l,k}\}_{k\in[K]}$ and their actions are sampled as $\mathbf{a}_{-i,h} \sim \pi_{-i}(\cdot \mid s_h)$, while agent i samples $a_{i,h}$ uniformly from its action space (Line 9). Finally, we draw s_{h+1} from the nominal kernel $P_{h,s_h,a_{i,h}}^{\pi_{-i}}(\cdot)$ (Line 10). We repeat this procedure $\lceil N/t \rceil$ times for each $l \in [t-1]$, yielding a dataset $\mathcal{D}_i$ containing approximately N samples, denoted by $\{(s_1^{(m)}, \mathbf{a}_1^{(m)}, \mathbf{r}_1^{(m)}, \ldots, \mathbf{a}_h^{(m)}, \mathbf{r}_h^{(m)}, s_{h+1}^{(m)})\}_{m=1}^{|\mathcal{D}_i|}$.

Overall pipeline of Online-L-Robust-Q-FTRL**.** With the hybrid sampling design in hand, the central mechanism of Online-L-Robust-Q-FTRL in Algorithm 3 is to construct both optimistic and pessimistic estimates of the robust value function for subsequent procedures. Using the pessimistic estimate, the algorithm collects online trajectory samples from an adversarial environment, while using the optimistic estimate to update the policy. Online-L-Robust-Q-FTRL follows an adversarial online learning procedure using backward recursion from $h = H$ to $h = 1$. At each round $t \in [T]$, for each step h, the algorithm executes a K-iteration loop that samples around size N using estimates the nominal model, and updates both the pessimistic and optimistic robust Q-function via the dual form of robust Bellman equation. See Appendix C for more details.

4.3 Theoretical Guarantee

Theorem 3. *Consider any R-LMG* $\{\mathcal{S}, \{\mathcal{A}_i\}_{1\le i\le n}, \{\mathcal{U}^{\sigma_i}(P^0,\cdot)\}_{1\le i\le n}, r, H\}$ *with a fixed initial state* $s_1 \in \mathcal{S}$*, uncertainty levels* $\{\sigma_i\}_{i\in[n]}$*, and fix* $\delta \in (0,1)$*. Let* $N = K = T$*. Then, with probability at least* $1-\delta$*, the regret of Algorithm 3 satisfies*

$$\textit{Regret}(T) \le \tilde{\mathcal{O}}\big(H^2 d \max_{i\in[n]} A_i \sqrt{T}\big).$$

And if the online iteration time T *satisfies* $T > H^4 d^2 \max_{i\in[n]}(A_i)^2/\epsilon^2$*, then Algorithm 3 outputs an* ε*-CCE.*

Breaking the curse of multiagency and comparison to prior work. To illustrate the implications of our results, we consider the specialization to the tabular setting, which corresponds to setting the feature dimension as $d = S\max_{i\in[n]} A_i$. In this case, Theorem 3 yields the regret bound $\tilde{\mathcal{O}}\Big(H^2 S \max_{i\in[n]}(A_i)^2\sqrt{T}\Big)$, which is polynomial in all salient parameters. This shows that Linear-Robust-Q-FTRL breaks the curse of multiagency in the proposed online interactive setting.

A closely related work is Zheng and Lin [89], which also studies robust linear MGs, but under a different online interactive setting: agents interact only with the fixed nominal transition kernel. Zheng and Lin [89] focuses on a restrictive problem class in which the robust value function is required to satisfy a *vanishing minimal value* condition. Moreover, under some additional assumptions in [89, Corollary 5.3], Zheng and Lin [89] obtains a sublinear regret bound $\mathrm{Regret}(T) \le \tilde{\mathcal{O}}\Big(H\min\big\{H, \frac{1}{\min\{\sigma_i\}}\big\} d_{\mathsf{curse}}\sqrt{T}\Big)$, where, in the tabular reduction, the feature dimension becomes $d_{\mathsf{curse}} = S\prod_{i=1}^n A_i$. Consequently, the regret bound in [89, Corollary 5.3] scales as $\tilde{\mathcal{O}}\big(H\min\big\{H, \frac{1}{\min\{\sigma_i\}}\big\} S\prod_{i=1}^n A_i \sqrt{T}\big)$, which grows exponentially in the number of agents through the joint action space $\prod_{i=1}^n A_i$. In contrast, Linear-Robust-Q-FTRL guarantees $\tilde{\mathcal{O}}\Big(H^2 S \max_{i\in[n]}(A_i)^2 \sqrt{T}\Big)$, thereby avoiding exponential dependence on the number of agents—a benefit that is retained even in the tabular setting.

Technical Contributions

- **Constructing an adversarial environment using a pessimistic value function.** As discussed above, we cannot sample directly from the true worst-case adversarial transition dynamics because the ground-truth robust value function is unknown. To approximate that adversarial environment, in addition to the optimistic robust value estimates typically used to guide policy updates in standard (linear) MGs [37, 75], we introduce a sequence of *pessimistic* robust value estimates. These pessimistic estimates are fed into the hybrid sampling module (cf. Algorithm 1) to obtain a tight lower bound on the robust value function under the true adversarial environment.
- **Maintaining nominal-model estimation via hybrid-environment sampling.** A key technical challenge is that the linear function approximation assumption applies only to the nominal kernel; kernels in the uncertainty set need not preserve this structure, making direct estimation of the worst-case kernel intractable. To address this, we continue to estimate the nominal model while preserving the state–action occupancy distribution induced by the worst-case environment. Specifically, we use a hybrid sampling mechanism that rolls out Markovian trajectories under the approximate worst-case environment for the first $h-1$ steps and queries the nominal kernel only at the final step h. This design allows us to estimate the nominal kernel while preserving the state-action occupancy distribution induced by the worst-case environment up to step $h-1$, which play a crucial role in subsequent estimation of the value function corresponding to the worst-case environment.

5 Conclusion

In this work, we develop provably data-efficient algorithms for robust linear Markov games (R-LMGs) for large-scale, possibly infinite state spaces under linear function approximation. Focusing on a fictitious uncertainty set defined via total variation, we study both the generative-model setting and a more practical online interactive data-collection setting. In both settings, to the best of our

knowledge, we provide the first provably sample-efficient algorithms for R-LMGs that overcome the curse of multi-agency, regardless of the uncertainty set formulation. Notably, the proposed online interactive setting is practically meaningful while also preserving well-posedness for achieving sublinear regret. On the technical side, we introduce a new hybrid sampling mechanism to address the challenge of sampling from an approximate adversarial environment in the online setting, which may be of independent interest for future research.

Acknowledgments and Disclosure of Funding

The authors thank Prof. Yuejie Chi for her guidance throughout the process and Prof. Eric Mazumdar for very helpful discussions. Laixi Shi acknowledges funding support from Mitsubishi Electric Research Laboratories.

References

[1] Michele Aghassi and Dimitris Bertsimas. Robust game theory. *Mathematical programming*, 107(1):231–273, 2006.

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

[3] Pragnya Alatur, Anas Barakat, and Niao He. Independent policy mirror descent for markov potential games: Scaling to large number of players, 2024.

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

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

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

[7] 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.

[8] Dimitri P Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific Belmont, 2009.

[9] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical programming*, 167(2):235–292, 2018.

[10] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[11] 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*, 36:66845–66859, 2023.

[12] 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*, 36:66845–66859, 2023.

[13] Fan Chen, Song Mei, and Yu Bai. Unified algorithms for rl with decision-estimation coefficients: Pac, reward-free, preference-based learning and beyond, 2025.

[14] 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*, pages 227–261. PMLR, 2022.

[15] Qiwen Cui, Kaiqing Zhang, and Simon S Du. Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. 2023.

[16] Yan Dai, Qiwen Cui, and Simon S Du. Refined sample complexity for markov games with independent linear function approximation, 2024.

[17] Constantinos Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013.

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

[19] Wenhao Ding, Baiming Chen, Minjun Xu, and Ding Zhao. Learning to collide: An adaptive safety-critical scenarios generating method. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2243–2250. IEEE, 2020.

[20] Wenhao Ding, Laixi Shi, Yuejie Chi, and Ding Zhao. Seeing is not believing: Robust reinforcement learning against spurious correlation. *Advances in Neural Information Processing Systems*, 36:66328–66363, 2023.

[21] Jing Dong, Baoxiang Wang, and Yaoliang Yu. Convergence to nash equilibrium and no-regret guarantee in (markov) potential games, 2024.

[22] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[23] 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, pages 2589–2595, 2015.

[24] Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306, 2023.

[25] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

[26] Irving L Glicksberg. A further generalization of the kakutani fixed theorem, with application to nash equilibrium points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

[27] Jacob K Goeree, Charles A Holt, and Thomas R Palfrey. Regular quantal response equilibrium. *Experimental economics*, 8(4):347–367, 2005.

[28] Min Hua, Xinda Qi, Dong Chen, Kun Jiang, Zemin Eitan Liu, Hongyu Sun, Quan Zhou, and Hongming Xu. Multi-agent reinforcement learning for connected and automated vehicles control: Recent advancements and future prospects. *IEEE Transactions on Automation Science and Engineering*, 22:16266–16286, 2025.

[29] Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. 2021.

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

[31] Matthew T Jackson, Minqi Jiang, Jack Parker-Holder, Risto Vuorio, Chris Lu, Greg Farquhar, Shimon Whiteson, and Jakob Foerster. Discovering general reinforcement learning algorithms with adversarial environment design. *Advances in Neural Information Processing Systems*, 36: 79980–79998, 2023.

[32] Lei Jia and Yan Pei. Recent advances in multi-agent reinforcement learning for intelligent automation and control of water environment systems. *Machines*, 13(6):503, 2025.

[33] Yuchen Jiao and Gen Li. Minimax-optimal multi-agent robust reinforcement learning. *arXiv preprint arXiv:2412.19873*, 2024.

[34] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael Jordan. Is q-learning provably efficient? volume 31, 2018.

[35] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation, 2020.

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

[37] 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.

[38] Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279. PMLR, 2022.

[39] 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.

[40] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*, volume 355. 2019.

[41] Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. Optidice: Offline policy optimization via stationary distribution correction estimation. In *International Conference on Machine Learning*, pages 6120–6130. PMLR, 2021.

[42] Gen Li, Yuejie Chi, Yuting Wei, and Yuxin Chen. Minimax-optimal multi-agent rl in markov games with a generative model. *Advances in Neural Information Processing Systems*, 35: 15353–15367, 2022.

[43] Fengming Lin, Xiaolei Fang, and Zheming Gao. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization*, 12(1):159–212, 2022.

[44] Miao Lu, Han Zhong, Tong Zhang, and Jose Blanchet. Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithms. *Advances in Neural Information Processing Systems*, 37:12528–12580, 2024.

[45] 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.

[46] A Rupam Mahmood, Dmytro Korenkevych, Gautham Vasan, William Ma, and James Bergstra. Benchmarking reinforcement learning algorithms on real-world robots. In *Conference on robot learning*, pages 561–591. PMLR, 2018.

[47] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[48] Valentin Mohl, Sascha Frey, Reuben Leyland, Kang Li, George Nigmatulin, Mihai Cucuringu, Stefan Zohren, Jakob Foerster, and Anisoara Calinescu. Jaxmarl-hft: Gpu-accelerated large-scale multi-agent reinforcement learning for high-frequency trading. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pages 18–26, 2025.

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

[50] John F Nash. Non-cooperative games. pages 329–340, 2024.

[51] Chengzhuo Ni, Yuda Song, Xuezhou Zhang, Chi Jin, and Mengdi Wang. Representation learning for general-sum low-rank markov games. *arXiv preprint arXiv:2210.16976*, 2022.

[52] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[53] James Orr and Ayan Dutta. Multi-agent deep reinforcement learning for multi-robot applications: A survey. *Sensors*, 23(7):3625, 2023.

[54] Yuxin Pan, Yize Chen, and Fangzhen Lin. Adjustable robust reinforcement learning for online 3d bin packing. *Advances in Neural Information Processing Systems*, 36:51926–51954, 2023.

[55] Omiros Papaspiliopoulos. *High-dimensional probability: An introduction with applications in data science*. Taylor & Francis, 2020.

[56] Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30215–30248, 2025.

[57] Allen Z Ren and Anirudha Majumdar. Distributionally robust policy learning via adversarial environment generation. *IEEE Robotics and Automation Letters*, 7(2):1379–1386, 2022.

[58] Alex P Robertson and Wendy Robertson. Topological vector spaces. volume 53. CUP Archive, 1980.

[59] Aviad Rubinstein. Settling the complexity of computing approximate two-player nash equilibria. *ACM SIGecom Exchanges*, 15(2):45–49, 2017.

[60] Fedor Sandomirskiy, Po Hyun Sung, Omer Tamuz, and Ben Wincelberg. Narrow bracketing and risk in games. *arXiv preprint arXiv:2502.11243*, 2025.

[61] Shai Shalev-Shwartz. Online learning: Theory, algorithms, and applications. 2007.

[62] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2025.

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

[64] Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *Journal of Machine Learning Research*, 25(200):1–91, 2024.

[65] 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*, 36:79903–79917, 2023.

[66] Laixi Shi, Jingchu Gai, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Can we break the curse of multiagency in robust multi-agent reinforcement learning?, 2024.

[67] 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*, 2024.

[68] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529 (7587):484–489, 2016.

[69] 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.

[70] 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.

[71] István Szita, Bálint Takács, and András Lörincz. ε-mdps: Learning in varying environments. *Journal of Machine Learning Research*, 3(Aug):145–174, 2002.

[72] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

[73] Zain Ulabedeen Farhat, Debamita Ghosh, George K Atia, and Yue Wang. Online robust multi-agent reinforcement learning under model uncertainties. *arXiv e-prints*, pages arXiv–2508, 2025.

[74] 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.

[75] Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2793–2848. PMLR, 2023.

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

[77] Stephen Willard. *General topology*. Courier Corporation, 2012.

[78] David G Wright. Tychonoff's theorem. *Proceedings of the American Mathematical Society*, 120(3):985–987, 1994.

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

[80] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, and Tong Zhang. A self-play posterior sampling algorithm for zero-sum markov games. In *International Conference on Machine Learning*, pages 24496–24523. PMLR, 2022.

[81] Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. pages 9728–9754, 2023.

[82] Lin Yang and Mengdi Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[83] 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.

[84] 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.

[85] Wenhao Zhan, Jason D Lee, and Zhuoran Yang. Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in markov games. 2022.

[86] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in neural information processing systems*, 33:21024–21037, 2020.

[87] Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*, 2021.

[88] Kaiqing Zhang, Sham M Kakade, Tamer Basar, and Lin F Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Journal of Machine Learning Research*, 24(175):1–53, 2023.

[89] Zewu Zheng and Yuanyuan Lin. Distributionally robust online markov game with linear function approximation. 40(34):28892–28900, 2026.

[90] 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. *arXiv preprint arXiv:2010.09776*, 2020.

[91] 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*, pages 3331–3339. PMLR, 2021.

A Related work

Distributionally Robust RL. Standard reinforcement learning (RL) methods can be highly sensitive to perturbations of environment dynmaics (transition kernels and reward function) [46, 86]. To enhance robustness, distributionally robust RL (robust RL) seeks policies that optimize worst-case performance over a prescribed uncertainty set by incorporating distributionally robust optimization (DRO) into sequential decision making. DRO has also been extensively studied in supervised learning Bertsimas et al. [9], Blanchet and Murthy [10], Duchi and Namkoong [22], Gao [24], Gao and Kleywegt [25], Lin et al. [43].

Distributionally robust RL was first studied in the single-agent setting [5, 30, 52, 64, 65, 91]. More recently, DRO formulations have been incorporated to multi-agent RL and formulates robust Markov games (RMGs) Blanchet et al. [11], Kardeş et al. [39], Ma et al. [45], Shi et al. [67], Zhang et al. [86]. Existing sample-complexity guarantees for algorithms of RMGs largely focus on tabular settings with finite state and action spaces Blanchet et al. [12], Ma et al. [45], Shi et al. [66, 67], while comparatively few results address large-scale or infinite state spaces. To our knowledge, the only existing work [89] studied RMGs under linear function approximation. [89] focuses on a restrictive class of RMGs [44, 89] that requires an additional vanishing minimal value assumption, which is a more restrictive subclass of RMGs than the setting considered in this work; moreover, its sample complexity still suffers from the curse of multiagency when transferred to the tabular setting. In this work, we focus on general RMGs under independent per-agent linear function approximation modeling with no further assumption on the structure of the value function, and achieve sample complexity that breaks the curse of multi-agency. See the discussion after Theorem 3 for a detailed comparison.

Multi-Agent RL with Function Approximation. Linear function approximation (LFA) was first studied in the single-agent setting [36, 82]. Building on these techniques, Xie et al. [79] and Chen et al. [13] investigated Markov games under centralized LFA, where a central function class models value functions across the joint actions of all agents. This setting makes circumventing the curse of multi-agency challenging. To address this issue, Cui et al. [15], Wang et al. [75] introduced independent per-agent linear function classes and correspondingly developed provably efficient algorithms that break the curse of multi-agency; further, Dai et al. [16] obtained optimal dependence on the action-set size via sharper analysis. Beyond LFA, several works have studied Markov games with general function approximation in both centralized settings [14, 29, 38, 51, 80, 85] and per-agent settings [75]. In this paper, we extend independent per-agent LFA to a popular robust counterpart of standard MGs—robust linear MGs (R-LMGs). Compared to standard MGs, in both generative and online settings, we must handle the additional statistical error induced by the high nonlinearity of R-LMGs with respect to the nominal transition kernel. In the online interactive setting in particular, additional challenges arise from the need to sample from the approximate adversarial environment and from the fact that the LFA assumption applies solely to the nominal kernel; kernels within the uncertainty set do not necessarily preserve this linear structure. To address this, we introduce a *hybrid sampling strategy* that uses a sequence of pessimistic robust value estimates, allowing us to estimate the nominal transition kernel satisfying LFA through ridge regression and then apply the pessimistic estimates to approximate the adversarial model for subsequent robust objective estimation.

Breaking the Curse of Multiagency in Multi-Agent RL. In multi-agent reinforcement learning (MARL), the joint action space grows exponentially with the number of agents, making it crucial to develop algorithms whose sample complexity does not scale exponentially in the number of agents—often referred to as breaking the curse of multi-agency. This challenge is ubiquitous in MARL and has attracted substantial interest. Bai and Jin [6], Rubinstein [59], Song et al. [70] show that learning a Nash equilibrium in general Markov games can require sample complexity exponential in the number of agents. This has motivated work on alternative solution concepts, such as correlated equilibrium (CE) and coarse correlated equilibrium (CCE), for which non-exponential guarantees are possible. For standard MGs, Daskalakis et al. [18] provide a complexity bound of $\widetilde{O}(H^{11}S^3 \max_{i\in[n]} A_i/\epsilon^3)$ for learning a CCE, and Jin et al. [37], Li et al. [42], Song et al. [70] also propose algorithms with non-exponential sample complexity, both for tabular MGs. Beyond the tabular setting, Cui et al. [15] and Wang et al. [75] develop algorithms that break the curse of multi-agency under independent per-agent LFA. Beyond standard Markov games, Alatur et al.

[3], Dong et al. [21], Song et al. [70] study Markov potential games and give algorithms that learn Nash equilibria with sample complexity that does not scale exponentially with the number of agents.

For robust Markov games, breaking the curse remains largely underexplored. The only existing results that breaks the curse of multi-agency Jiao and Li [33], Shi et al. [66] are restricted to tabular RMGs with R-contamination or TV distance. To the best of our knowledge, this is the first work to break the curse of multi-agency for large-scale (continuous) state space, in both generative model and online data collection setting.

Algorithm 2: L-Robust-Q-FTRL

1: **Input:** The support set $\{\mathcal{Y}_i\}_{i\in[n]}$ and the sampling distribution $\{\rho_i\}_{i\in[n]}$; number of iterations K.
2: **Initialization:** $\widehat{V}_{i,H+1}(s) = Q^0_{i,h}(s,a_i) = 0$, $\pi^1_{i,h}(a_i \mid s) = 1/A_i$ for all $(i,h,s,a_i) \in [n]\times[H]\times\mathcal{S}\times\mathcal{A}_i$.
3: **for** $h = H, H-1, \ldots, 1$, and $k \in [K]$ **do**
4: **for** $i = 1, 2, \ldots, n$ **do**
5: *// Sampling from the nominal model based on the finite set $\mathcal{Y}_i$ and probability distribution ρ_i.*
6: **for** $(s,a_i) \in \mathcal{Y}_i$ and $u = 1, 2, \cdots \lceil N\rho_i(s,a_i)\rceil$ **do**
7: Sample $\mathbf{a}^{(u)}(s,a_i) = [a_j(s,a_i)]_{j\in[n]}$ with $a_j(s,a_i) \overset{\text{ind.}}{\sim} \pi_{j,h}(\cdot \mid s)$ $(j \neq i)$, $a_i(s,a_i) = a_i$.
8: Sample from the generative model:

$$r^{(u)}_{i,h} = r_{i,h}(s, \mathbf{a}^{(u)}(s,a_i)), \quad s^{(u)}_{h+1} \sim P_h(\cdot \mid s, \mathbf{a}^{(u)}(s,a_i)).$$

9: *// Observe reward and next state from the nominal model P_h.*
10: Insert the data tuple $\{s^{(m)}_h = s, a^{(m)}_{i,h} = a_i, r^{(m)}_{i,h} = r^{(u)}_{i,h}, s^{(m)}_{h+1} = s^{(u)}_{h+1}\}$ into $\mathcal{D}_i$.
11: Estimate model parameters $\{r^k_{i,h}, P^k_{i,h}\}$ with $\mathcal{D}_i$ via ridge regression using (11).
12: *// Ridge regression for estimating nominal reward and transition in the feature space.*
13: Compute q-function $q^k_{i,h}(s,a_i)$ with (12).
14: *// Follow the Regularized Leader (FTRL) algorithm to update the policy.*
15: Update policy for all $(s,a_i) \in \mathcal{S}\times\mathcal{A}_i$ with

$$\pi^{k+1}_{i,h}(a_i \mid s) = \frac{\exp(\eta_{k+1}\sum_{l=1}^k q^l_{i,h}(s,a_i))}{\sum_{a_i\in\mathcal{A}_i}\exp(\eta_{k+1}\sum_{l=1}^k q^l_{i,h}(s,a_i))} \tag{10}$$

16: *// Update optimistic robust values with bonus $\beta_{i,h}$ after K rounds at step h.*
17: Construct an upper bound $\{\widehat{V}_{i,h}\}$ for value functions using (14).
18: **Output:** Product policies $\{\pi^k_h = (\pi^k_{1,h}\times\cdots\times\pi^k_{n,h})\}_{h\in[H],k\in[K]}$ and a probability distribution $\xi = \{\xi_h\}_{h\in[H]}$ over product policies so that $\xi_h(\pi^k_h) = 1/K$ for all $h \in [H]$.

B Algorithm Design with a Generative Model

Due to the space limit, we postpone the entire pipleline of Algorithm 2 in this section as follows.

For initialization, L-Robust-Q-FTRL set the robust value function, robust Q-function $\widehat{V}_{i,H+1}(s) = Q^0_{i,h}(s,a_i) = 0$, and the policy $\pi^1_{i,h}(a_i \mid s) = 1/A_i$ for all $(i,s) \in [n]\times\mathcal{S}$. Then L-Robust-Q-FTRL performs a backward recursive learning path, learning recursively from the final time step $h = H$ to $h = 1$. At each time step $h \in [H]$, a K-iterations online learning process is applied. At each interaction iteration $k \in [K]$, parallelly for each agent $i \in [n]$, we execute the following steps:

- **Sampling $\mathcal{D}_i$ from the nominal model.** For each $(s,a_i) \in \mathcal{Y}_i$, we collect $\lceil N\rho_i(s,a_i)\rceil$ samples from the generative model (cf. line 7 to 10) and yield the dataset $\mathcal{D}_i = \{s^{(m)}_h, a^{(m)}_{i,h}, r^{(m)}_{i,h}, s^{(m)}_{h+1}\}_{m\in\mathcal{D}_i}$

- **Estimating the nominal model.** Armed with the dataset $\mathcal{D}_i$, we estimate the model parameters $\{r_{i,h}^k, P_{i,h}^k\}$ via ridge regression as below:

$$\Lambda_{i,h}^k = \sum_{m\in\mathcal{D}_i} \phi_i(s_h^{(m)}, a_{i,h}^{(m)})\phi_i(s_h^{(m)}, a_{i,h}^{(m)})^\top + \lambda I, \tag{11a}$$

$$\widehat{\theta}_{i,h}^k = (\Lambda_{i,h}^k)^{-1} \sum_{m\in\mathcal{D}_i} \phi_i(s_h^{(m)}, a_{i,h}^{(m)}) r_{i,h}^{(m)}, \tag{11b}$$

$$\widehat{\mu}_{i,h}^k = (\Lambda_{i,h}^k)^{-1} \sum_{m\in\mathcal{D}_i} \phi_i(s_h^{(m)}, a_{i,h}^{(m)}) \delta_{s_{h+1}^{(m)}}, \tag{11c}$$

where $\Lambda_{i,h}^k \in \mathbb{R}^{d\times d}, \phi_i(\cdot,\cdot) \in \mathbb{R}^d, \widehat{\mu}_{i,h}^k \in \mathbb{R}^d \to \Delta(\mathcal{S})$. Here, we denote δ_s as a measure over state space $\mathcal{S}$ so that the measure on state s is 1, and the measure on other states is 0, for all $s \in \mathcal{S}$. λ is the regularization parameter in ridge regression satisfying $\lambda \geq 1$. For all $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$, we compute the estimated reward function and transition probabilities as:

$$r_{i,h}^k(s, a_i) = \phi_i(s, a_i)^\top \widehat{\theta}_{i,h}^k, \quad P_{i,h}^k(\cdot \mid s, a_i) = \phi_i(s, a_i)^\top \widehat{\mu}_{i,h}^k. \tag{11d}$$

- **Updating the robust-Q-function and policy.** Denoting the current joint policy by π_h^k, with the estimated model $\{r_{i,h}^k, P_{i,h}^k\}_{i\in[n]}$, we compute a robust Q-function at each iteration. Specifically, for all $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$, we compute $q_{i,h}^k(s, a_i)$ as follows:

$$\begin{aligned} q_{i,h}^k(s, a_i) = r_{i,h}^k(s, a_i) + \max_{\alpha\in\left[\min_s \widehat{V}_{i,h+1}(s), \max_s \widehat{V}_{i,h+1}(s)\right]} \big\{ & P_{i,h}^k(\cdot \mid s, a_i)\big[\widehat{V}_{i,h+1}\big]_\alpha \\ & - \sigma_i\big(\alpha - \min_{s'} \big[\widehat{V}_{i,h+1}\big]_\alpha (s')\big)\big\}, \end{aligned} \tag{12}$$

which is the dual form of the robust Bellman equation [65, Lemma 1]

$$q_{i,h}^k(s, a_i) = r_{i,h}^k(s, a_i) + \inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}(P_{i,h}^k(\cdot \mid s,a_i))} \mathcal{P}\widehat{V}_{i,h+1} \tag{13}$$

and is computationally tractable [30]. Note that even $\mathcal{S}$ may be infinite, the model estimate $P_{i,h}^k(\cdot \mid s, a_i)$ obtained from ridge regression in (11) is supported on the finite number of observed next states $\{s_{h+1}^{(m)}\}_{m\in\mathcal{D}_i}$ in dataset $\mathcal{D}_i$. Consequently, $P_{i,h}^k(\cdot \mid s, a_i)[\widehat{V}_{i,h+1}]_\alpha$ is a finite weighted sum, and $\widehat{V}_{i,h+1}$ need only be evaluated at finitely many points, as in the lazy-evaluation implementation in standard linear MDP/MG algorithms [35, 75]. After calculating the q-function, we employ the Follow-the-Regularized-Leader (FTRL) strategy [42, 62] to update the learner's policy towards $\pi_{i,h}^{k+1}$ via (10).

- **Estimate robust value function.** Subsequently, after completing the K iterations at step h, we compute an optimistic estimate of the value function as follows: for all $(i, s) \in [n] \times \mathcal{S}$,

$$\widehat{V}_{i,h}(s) = \min\big\{\textstyle\sum_{k=1}^K \frac{1}{K}\big\langle \pi_{i,h}^k(\cdot \mid s), q_{i,h}^k(s, \cdot)\big\rangle + \beta_{i,h}(s), H - h + 1\big\}, \tag{14}$$

Here, $\beta_{i,h}(s)$ is a bonus term specified as follows:

$$\beta_{i,h}(s) = 8\sqrt{\tfrac{d}{N}}\big(2H\sqrt{d\ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big) + 2H\sqrt{\tfrac{\ln A_i}{K}}. \tag{15}$$

After completing the learning for all $h \in [H]$, each with K iterations at every step, we output the joint correlated policy $\big\{\frac{1}{K}\sum_{k=1}^K \pi_h^k = \frac{1}{K}\sum_{k=1}^K \{(\pi_{1,h}^k \times \cdots \times \pi_{n,h}^k)\}_{h\in[H]}\big\}$ as the learned solution, which can be shown later as an ε-CCE with high probability.

C Algorithm Design in Online Interactive Setting

This section describe the algorithm in detail, along with the full pseudocode for the algorithms used in the main body in Algorithm 3.

At each online round $t \in [T]$, the algorithm learns a distribution ξ^t over policies by running K iterations at each time step $h \in [H]$. For any fixed round t and step $h \in [H]$, the k-th iteration proceeds as follows:

- **Estimate nominal model.** In the generative-model setting of Section 3.1, one actively queries a pre-designed per-agent state-action set $\mathcal{Y}_i$ from Lemma 1. In contrast, the online setting's visited state-action pairs are determined entirely by the sampled trajectories in $\mathcal{D}_i$ and need not—nor can they—be chosen as *a prior*. We therefore directly apply ridge regression to the trajectory samples in $\mathcal{D}_i$ collected by Hybrid-Sampling to estimate the transition probabilities and the reward function:

$$\Lambda_{i,h}^{k,t} = \textstyle\sum_{m=1}^{|\mathcal{D}_i|} \phi_i(s_h^{(m)}, a_{i,h}^{(m)}) \phi_i(s_h^{(m)}, a_{i,h}^{(m)})^\top + \lambda I, \tag{16a}$$

$$\widehat{\theta}_{i,h}^{k,t} = (\Lambda_{i,h}^{k,t})^{-1} \textstyle\sum_{m=1}^{|\mathcal{D}_i|} \phi_i(s_h^{(m)}, a_{i,h}^{(m)}) r_{i,h}^{(m)}, \tag{16b}$$

$$\widehat{\mu}_{i,h}^{k,t} = (\Lambda_{i,h}^{k,t})^{-1} \textstyle\sum_{m=1}^{|\mathcal{D}_i|} \phi_i(s_h^{(m)}, a_{i,h}^{(m)}) \delta_{s_{h+1}^{(m)}}, \tag{16c}$$

where $\Lambda_{i,h}^{k,t} \in \mathbb{R}^{d\times d}, \phi_i(\cdot,\cdot) \in \mathbb{R}^d, \widehat{\mu}_{i,h}^{k,t} \in \mathbb{R}^d \to \Delta(\mathcal{S})$. Here, we denote δ_s as a measure over state space $\mathcal{S}$ so that the measure on state s is 1, and the measure on other states is 0, for all $s \in \mathcal{S}$. λ is the regularization parameter in ridge regression satisfying $\lambda \geq 1$. For all $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$, we compute the estimated reward function and transition probabilities as:

$$r_{i,h}^{k,t}(s, a_i) = \phi_i(s, a_i)^\top \widehat{\theta}_{i,h}^{k,t}, \quad P_{i,h}^{k,t}(\cdot \mid s, a_i) = \phi_i(s, a_i)^\top \widehat{\mu}_{i,h}^{k,t}. \tag{16d}$$

- **Update robust-Q-function and policy.** Different from generative model setting, we now compute both the optimistic estimation and pessimistic estimation of the current-step Q-function. Specifically, for all $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$, we compute $\overline{q}_{i,h}^{k,t}(s, a_i)$ and $\underline{q}_{i,h}^{k,t}(s, a_i)$ as follows:

$$\begin{aligned}\overline{q}_{i,h}^{k,t}(s, a_i) = r_{i,h}^{k,t}(s, a_i) + \max_{\alpha \in \left[\overline{V}_{\min}, \overline{V}_{\max}\right]} \big\{ & P_{i,h}^{k,t}(\cdot \mid s, a_i) \big[\overline{V}_{i,h+1}^{t}\big]_\alpha \\ & - \sigma_i\big(\alpha - \min_{s'} \big[\overline{V}_{i,h+1}^{t}\big]_\alpha (s')\big)\big\},\end{aligned} \tag{17a}$$

$$\begin{aligned}\underline{q}_{i,h}^{k,t}(s, a_i) = r_{i,h}^{k,t}(s, a_i) + \max_{\alpha \in \left[\underline{V}_{\min}, \underline{V}_{\max}\right]} \big\{ & P_{i,h}^{k,t}(\cdot \mid s, a_i) \big[\underline{V}_{i,h+1}^{t}\big]_\alpha \\ & - \sigma_i\big(\alpha - \min_{s'} \big[\underline{V}_{i,h+1}^{t}\big]_\alpha (s')\big)\big\}.\end{aligned} \tag{17b}$$

where we define $\overline{V}_{\min} = \min_s \overline{V}_{i,h+1}^{t}(s)$, $\overline{V}_{\max} = \max_s \overline{V}_{i,h+1}^{t}(s)$, $\underline{V}_{\min} = \min_s \underline{V}_{i,h+1}^{t}(s)$ and $\underline{V}_{\max} = \max_s \underline{V}_{i,h+1}^{t}(s)$. After calculating the q-function, we employ the Follow-the-Regularized-Leader (FTRL) strategy to update the learner's policy, i,e.,

$$\forall (i, s, a_i) \in [n] \times \mathcal{S} \times \mathcal{A}_i : \pi_{i,h}^{k+1,t}(a_i \mid s) = \frac{\exp(\eta_{k+1} \sum_{l=1}^{k} \overline{q}_{i,h}^{l,t}(s, a_i))}{\sum_{a_i \in \mathcal{A}_i} \exp(\eta_{k+1} \sum_{l=1}^{k} \overline{q}_{i,h}^{l,t}(s, a_i))}.$$

- **Estimate robust value function.** Subsequently, we compute the optimistic and pessimistic estimates of the value function as follows:

$$\begin{aligned}\overline{V}_{i,h}^{t}(s) &= \min\big\{ \textstyle\sum_{k=1}^{K} \frac{1}{K} \big\langle \pi_{i,h}^{k,t}(\cdot \mid s), \overline{q}_{i,h}^{k,t}(s, \cdot) \big\rangle + \beta_{i,h,1}^{t}(s),\ H - h + 1 \big\} \\ \underline{V}_{i,h}^{t}(s) &= \max\big\{ \textstyle\sum_{k=1}^{K} \frac{1}{K} \big\langle \pi_{i,h}^{k,t}(\cdot \mid s), \underline{q}_{i,h}^{k,t}(s, \cdot) \big\rangle - \beta_{i,h,2}^{t}(s), 0 \big\}.\end{aligned} \tag{18}$$

Here, for all $(i, h, t, s) \in [n] \times [H] \times [T] \times \mathcal{S}$, the bonus terms are defined as

$$\begin{aligned}\beta_{i,h,1}^{t}(s) = & \max_{a_i \in \mathcal{A}_i} \textstyle\sum_{k=1}^{K} \frac{1}{K} \big\|\phi_i(s, a_i)\big\|_{\left(\Lambda_{i,h}^{k,t}\right)^{-1}} \\ & \cdot \big(4H\sqrt{d \ln(NH+1) + \ln\big(\tfrac{3TNHnK}{\delta}\big)} + 2H\sqrt{d}\big) + \tfrac{1}{N} + 2H\sqrt{\tfrac{\ln A_i}{K}},\end{aligned} \tag{19}$$

$$\begin{aligned}\beta_{i,h,2}^{t}(s) = & \textstyle\sum_{k=1}^{K} \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big\|\phi_i(s, a_i)\big\|_{\left(\Lambda_{i,h}^{k,t}\right)^{-1}} \\ & \cdot \big(4H\sqrt{d \ln(NH+1) + \ln\big(\tfrac{3TNHnK}{\delta}\big)} + 2H\sqrt{d}\big) + \tfrac{1}{N}.\end{aligned} \tag{20}$$

Algorithm pseudocode. The algorithm in online interactive setting is summarized in Algorithm 3.

Algorithm 3: Online-Robust-Q-Linear-FTRL

Input: Iterations T, learning rate $\{\eta_{t+1}\}$.
for $t = 1, 2, \ldots, T$ **do**
  Initialization: $\overline{V}^t_{i,H+1} = \underline{V}^t_{i,H+1} = q^t_{i,H+1} = 0$, $\pi^1_{i,h}(a_i|s) = 1/A_i$ for all i, h, s, a_i.
  for $h = H, H-1, \ldots, 1$ and $k = 1, \ldots, K$ **do**
    // Backward procedure over $[H]$ with K inner FTRL steps;
    for $i = 1, 2, \ldots, n$ **do**
      // Collect samples from the approximate adversarial environment
      Collect N samples with Hybrid-Sampling$(i, h, t, \{\xi^l\}_{l<t}, \{\underline{V}^l_{i,j}\}_{(j,l)\in[H]\times[t-1]})$ in Algo 1.
      // Estimate the transition probabilities, reward function, and current step Q-function
      Estimate $r^{k,t}_{i,h}, P^{k,t}_{i,h}$ via (16) and q-function $\overline{q}^{k,t}_{i,h}$ and $\underline{q}^{k,t}_{i,h}$ via (17).
      // Update policy using FTRL with optimistic $\overline{q}$ *only.*
      Update policy for all (s, a_i): $\pi^{k+1,t}_{i,h}(a_i|s) = \frac{\exp(\eta_{k+1}\sum_{l=1}^k \overline{q}^{l,t}_{i,h}(s,a_i))}{\sum_{a'_i\in\mathcal{A}_i}\exp(\eta_{k+1}\sum_{l=1}^k \overline{q}^{l,t}_{i,h}(s,a'_i))}$
  Update $\overline{V}^t_{i,h}(s), \underline{V}^t_{i,h}(s)$ via (18).
  Let ξ^t be the uniform distribution over product policies $\{\pi^{k,t}_h\}_{(h,k)\in[H]\times[K]}$.
Output: All distributions $\{\xi^t\}_{t\in[T]}$ over product policies.

D Proof of Theorem 1

The proof of equilibrium existence often relies on fixed-point theorems. In this paper, to establish equilibrium existence for large-scale (possibly infinite) state spaces—including those that may be infinite or continuous—we appeal to the following fixed-point theorem.

Definition 2 (Upper semi-continuous)**.** *A point-to-set mapping $x \in X \mapsto \phi(x) \in Y$ is upper semi-continuous if $\lim_{n\to\infty} x^n = x_0, y^n \in \phi(x^n), \lim_{n\to\infty} y^n = y_0$ imply that $y_0 \in \phi(x_0)$.*

Theorem 4 (Glicksberg's fixed point theorem [26])**.** *If X is a compact and convex subset in a Hausdorff locally convex topological vector space, and ϕ is an upper semi-continuous correspondence mapping X into the family of all convex subsets of X, then there exists $x \in X$ so that $x \in \phi(x)$.*

It is worth noting that for cases with finite state spaces, the existence of Nash equilibria for RMGs—a broader class of games that includes R-LMGs—has been established in Shi et al. [66, Theorem1], where another theorem — Kakutani's fixed-point theorem is used. Theorem 4 relaxes the requirement in Kakutani's fixed-point theorem, where X must belong to a finite-dimensional Euclidean space, extending it to a broader class of Hausdorff linear topological spaces that is crucial for us to handle infinite state spaces.

Proof of Theorem 1. With Theorem 4 in hand, we prove Theorem 1 by following the proof outline of Shi et al. [66, Theorem 1], with modifications to accommodate a large-scale (possibly infinite) state space. In the large-state setting, we work directly with general robust MGs and do not exploit the game's linear structure to narrow down to the subclass of robust MGs (robust linear MGs). Following Shi et al. [66, Section B.1], we first establish the existence of a Nash equilibrium in an auxiliary one-step game; this result is then used recursively to verify equilibrium existence for the full sequential RMG.

Fix a step $h \in [H]$ and consider the associated auxiliary one-step game. Let $V_{i,h+1} \in \mathbb{R}^{\mathcal{S}}$ be a fixed "value function" for agent i, with $0 \le V_{i,h+1}(s) \le H$ for all $s \in \mathcal{S}$. Given any joint product policy $\pi : \mathcal{S} \to \prod_{i\in[n]} \Delta(\mathcal{A}_i)$ at step h, define agent i's payoff at state s by

$$\forall s \in \mathcal{S}: \quad f_{i,s}\big(\pi_i(s), \pi_{-i}(s); V_{i,h+1}\big) = \mathbb{E}_{\mathbf{a}\sim\pi(s)}\big[r_{i,h}(s,\mathbf{a})\big] + \mathbb{E}_{a_i\sim\pi_i(s)}\big[\inf_{\mathcal{U}^{\sigma_i}\left(P^{\pi_{-i}}_{h,s,a_i}\right)} PV_{i,h+1}\big]. \tag{21}$$

Based on these payoffs, we can introduce the best-response correspondence mapping ϕ as follows: for any $\pi : \mathcal{S} \to \prod_{i\in[n]} \Delta(\mathcal{A}_i)$,

$$\phi(\pi) \coloneqq \big\{u : \mathcal{S} \mapsto \textstyle\prod_{i\in[n]} \Delta(\mathcal{A}_i) \,|\, u_i(s) \in \operatorname{argmax}_{\pi_i'(s)\in\Delta(\mathcal{A}_i)} f_{i,s}(\pi_i'(s), \pi_{-i}(s); V_{i,h+1}), \\ \forall (i,s) \in [n] \times \mathcal{S}\big\}. \tag{22}$$

For this one-step game, the (joint) product policy space is

$$X \coloneqq \Big\{\pi : \mathcal{S} \to \textstyle\prod_{i\in[n]} \Delta(\mathcal{A}_i)\Big\}.$$

Since fixed points of ϕ correspond to Nash equilibria (NE), Theorem 4 yields the existence of NE once its conditions are verified. To show the three conditions, we begin by showing that X is a compact and convex subset of a convex Hausdorff linear topological space. First, $\{\pi : \mathcal{S} \mapsto \prod_{i\in[n]} \Delta(\mathcal{A}_i)\}$ is convex since the finite product of simplex $X' \coloneqq \prod_{i\in[n]} \Delta(\mathcal{A}_i)$ is convex and $\{\pi : \mathcal{S} \mapsto \prod_{i\in[n]} \Delta(\mathcal{A}_i)\}$ can be seen as a infinite product of a convex set, whose convexity is easily verified since product operations preserve convexity. Next, the compactness of $\{\pi : \mathcal{S} \mapsto \prod_{i\in[n]} \Delta(\mathcal{A}_i)\}$ follows from the compactness of $\prod_{i\in[n]} \Delta(\mathcal{A}_i)$ and Tychonoff's theorem [78], which guarantees that an arbitrary product of compact sets remains compact. Therefore, we verified that X is a compact and convex set.

Next, define

$$Y' \coloneqq \textstyle\prod_{i\in[n]} \mathbb{R}^{|\mathcal{A}_i|}, \qquad Y \coloneqq (Y')^{\mathcal{S}} = \Big\{\pi : \mathcal{S} \to Y'\Big\},$$

Since Y' is a finite-dimensional Euclidean space, it is a locally convex Hausdorff topological vector space. And arbitrary products of Hausdorff (resp. locally convex) topological vector spaces are again Hausdorff [77, Theorem 5] (resp. locally convex [58, Page 207]). It follows that the finite product Y' is a locally convex Hausdorff topological vector space, and then the arbitrary product $Y = (Y')^{\mathcal{S}}$ is a locally convex Hausdorff topological vector space. Moreover, because $X' \subset Y'$, we have $X \subset Y$. Consequently, X is a compact and convex subset of the locally convex Hausdorff topological vector space Y.

The remaining conditions of Theorem 4, as well as the subsequent recursive argument, can be verified similarly as in Shi et al. [66, Theorem 1] and are omitted for brevity.

E Proof of Theorem 2

E.1 Preliminaries

To begin with, for any joint policy π, the robust value function satisfies the robust Bellman equation $V_{i,h}^{\pi,\sigma_i}(s) = \mathbb{E}_{a_i\sim\pi_{i,h}(s)}[r_{i,h}^{\pi_{-i}}(s,a_i) + \inf_{\mathcal{U}_\rho^{\sigma_i}(P_{h,s,a_i}^{\pi_{-i}})} PV_{i,h+1}^{\pi,\sigma_i}]$ (23), where $r_{i,h}^{\pi_{-i}}(s,a_i) = \mathbb{E}_{a_{-i}\sim\pi_{-i,h}(s)}[r_{i,h}(s,\mathbf{a})]$ for all (s,a_i).

Then we introduce a standard bound on the ridge-regression estimation error under linear function approximation, adapted from [35, Theorem 3.1], which plays a key role in the proof of both the generative case and the online interactive case.

Theorem 5 (Upper bound for ridge-regression estimation 35, Theorem 3.1)**.** *For any fixed agent $i \in [n]$, step $h \in [H]$, and iteration $k \in [K]$, let $\widehat{\pi}^k$ be the product policy at current iteration k, and $\{(s^m, a_i^m, s_{+1}^m)\}_{m=1}^N$ be N samples collected independenly in this iteration, with $s_{+1}^m \sim P_{h,s^m,a_i^m}^{\widehat{\pi}_{-i}^k}(\cdot)$. Let $\phi_i^m \coloneqq \phi_i(s^m, a_i^m) \in \mathbb{R}^d$ satisfy $\|\phi_i^m\|_2 \le 1$. Fix the regularization parameter $\lambda \ge 1$ and define the Gram matrix*

$$\Lambda_{i,h}^k \coloneqq \textstyle\sum_{m=1}^N \phi_i^m (\phi_i^m)^\top + \lambda I_d.$$

Let $\mu_{i,h}^{\widehat{\pi}_{-i}^k} : \mathbb{R}^d \to \Delta(\mathcal{S})$ be the linear model parameter of the true transition kernel: $P_{h,s,a_i}^{\widehat{\pi}_{-i}^k}(\cdot) = \phi_i(s,a_i)^\top \mu_{i,h}^{\widehat{\pi}_{-i}^k}(\cdot)$ for all $(s,a_i) \in \mathcal{S} \times \mathcal{A}_i$, and assume the standard normalization $\|\mu_{i,h}^{\widehat{\pi}_{-i}^k}(\mathcal{S})\|_2 \le \sqrt{d}$. For any fixed value function $f : \mathcal{S} \to [0,H]$, define

$$\mu_{i,h}^{\widehat{\pi}_{-i}^k} f \coloneqq \textstyle\int_{\mathcal{S}} f(s')\, \mu_{i,h}^{\widehat{\pi}_{-i}^k}(ds') \in \mathbb{R}^d \qquad \text{and} \qquad \widehat{\mu}_{i,h}^k f \coloneqq (\Lambda_{i,h}^k)^{-1} \textstyle\sum_{m=1}^N \phi_i^m\, f(s_{+1}^m) \in \mathbb{R}^d.$$

Then, with probability at least $1-\delta$, it holds that

$$\big\|\widehat{\mu}_{i,h}^k f - \mu_{i,h}^{\widehat{\pi}_{-i}^k} f\big\|_{\Lambda_{i,h}^k} \;\le\; 2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3NH}{\delta}\big)} + H\sqrt{d\lambda},$$

and by Cauchy–Schwarz, for all $(s,a_i)\in\mathcal{S}\times\mathcal{A}_i$,

$$\big|\phi_i(s,a_i)^\top\big(\widehat{\mu}_{i,h}^k f - \mu_{i,h}^{\widehat{\pi}_{-i}^k} f\big)\big| \;\le\; \big\|\phi_i(s,a_i)\big\|_{(\Lambda_{i,h}^k)^{-1}}\Big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3NH}{\delta}\big)} + H\sqrt{d\lambda}\Big).$$

Proof. We adapt the argument of [35, Theorem 3.1] to our multi-agent linear-function-approximation setting.

Step 1: error decomposition. Recalling the definitions of $\Lambda_{i,h}^k$ and $\widehat{\mu}_{i,h}^k f$, combined with the fact that $(\phi_i^m)^\top \mu_{i,h}^{\widehat{\pi}_{-i}^k} f = \mathbb{E}_{s'\sim P_{h,s^m,a_i^m}^{\widehat{\pi}_{-i}^k}}[f(s')]$, we have

$$\begin{aligned}\widehat{\mu}_{i,h}^k f - \mu_{i,h}^{\widehat{\pi}_{-i}^k} f &= (\Lambda_{i,h}^k)^{-1}\Big[\textstyle\sum_{m=1}^N \phi_i^m\, f(s_{+1}^m) - \Lambda_{i,h}^k\, \mu_{i,h}^{\widehat{\pi}_{-i}^k} f\Big] \\ &= (\Lambda_{i,h}^k)^{-1}\underbrace{\textstyle\sum_{m=1}^N \phi_i^m\big(f(s_{+1}^m) - (\phi_i^m)^\top \mu_{i,h}^{\widehat{\pi}_{-i}^k} f\big)}_{:=\,\xi_{i,h}^k\ \text{(stochastic term)}} - \lambda\,(\Lambda_{i,h}^k)^{-1}\mu_{i,h}^{\widehat{\pi}_{-i}^k} f.\end{aligned}$$

Taking the $\Lambda_{i,h}^k$-norm and applying the triangle inequality together with the identity $\|(\Lambda_{i,h}^k)^{-1}v\|_{\Lambda_{i,h}^k} = \|v\|_{(\Lambda_{i,h}^k)^{-1}}$,

$$\big\|\widehat{\mu}_{i,h}^k f - \mu_{i,h}^{\widehat{\pi}_{-i}^k} f\big\|_{\Lambda_{i,h}^k} \;\le\; \underbrace{\big\|\xi_{i,h}^k\big\|_{(\Lambda_{i,h}^k)^{-1}}}_{\text{(I) stochastic}} + \underbrace{\lambda\big\|\mu_{i,h}^{\widehat{\pi}_{-i}^k} f\big\|_{(\Lambda_{i,h}^k)^{-1}}}_{\text{(II) bias}}. \tag{24}$$

Step 2: bounding the stochastic term (I). We first recall the lemma in [35, Lemma D.4] (renaming their time index k to τ to avoid clash with our iteration index $k\in[K]$): *let $\{x_\tau\}_{\tau\ge1}$ be a stochastic process on state space $\mathcal{S}$ with corresponding filtration $\{\mathcal{F}_\tau\}_{\tau\ge0}$. Let $\{\widetilde{\phi}_\tau\}_{\tau\ge1}$ be an $\mathbb{R}^d$-valued stochastic process where $\widetilde{\phi}_\tau\in\mathcal{F}_{\tau-1}$, and $\|\widetilde{\phi}_\tau\|_2\le1$. Let $\widetilde{\Lambda}_n := \lambda I_d + \sum_{\tau=1}^n \widetilde{\phi}_\tau\widetilde{\phi}_\tau^\top$. Then for any $\delta>0$, with probability at least $1-\delta$, for all $n\ge0$, and any $V\in\mathcal{V}$ so that $\sup_x|V(x)|\le H$, we have:*

$$\Big\|\textstyle\sum_{\tau=1}^n \widetilde{\phi}_\tau\big\{V(x_\tau) - \mathbb{E}[V(x_\tau)\mid\mathcal{F}_{\tau-1}]\big\}\Big\|_{\widetilde{\Lambda}_n^{-1}}^2 \le 4H^2\Big[\tfrac{d}{2}\log\tfrac{n+\lambda}{\lambda} + \log\tfrac{\mathcal{N}_\varepsilon}{\delta}\Big] + \tfrac{8n^2\varepsilon^2}{\lambda}, \tag{25}$$

where $\mathcal{N}_\varepsilon = \mathcal{N}_\varepsilon(\mathcal{V})$ is the ε-covering number of $\mathcal{V}$ with respect to the distance $\operatorname{dist}(V,V') = \sup_x|V(x)-V'(x)|$.

To apply the lemma into our setting, the sample index $m\in[N]$ plays the role of τ, so we apply (25) with $n=N$ and let

$\widetilde{\phi}_\tau = \phi_i^m = \phi_i(s^m,a_i^m)$,	(predictor)
$x_\tau = s_{+1}^m$,	(next-state observation)
$\mathcal{F}_m = \sigma\big(\{(s^j,a_i^j,s_{+1}^j)\}_{j\le m}\big)$,	(natural filtration)
$\widetilde{\Lambda}_n = \Lambda_{i,h}^k$,	(Gram matrix at $n=N$, fixed iteration k)
$V = f:\mathcal{S}\to[0,H]$.	(value function)

Therefore, $(s^m,a_i^m)\in\mathcal{F}_{m-1}$, $\phi_i^m\in\mathcal{F}_{m-1}$ as required, and $s_{+1}^m\sim P_{h,s^m,a_i^m}^{\widehat{\pi}_{-i}^k}(\cdot)$, which gives

$$\mathbb{E}[V(x_\tau)\mid\mathcal{F}_{\tau-1}] = \mathbb{E}_{s'\sim P_{h,s^m,a_i^m}^{\widehat{\pi}_{-i}^k}}[f(s')] = (\phi_i^m)^\top\mu_{i,h}^{\widehat{\pi}_{-i}^k} f.$$

Hence the LHS of (25) is exactly $\big\|\xi_{i,h}^k\big\|_{(\Lambda_{i,h}^k)^{-1}}^2$.

To proceed, since f is data-independent, we apply (25) to the singleton class $\mathcal{V} = \{f\}$, for which $\mathcal{N}_\varepsilon(\{f\}) = 1$ for every $\varepsilon \geq 0$. Letting $\varepsilon \to 0$ then drives $8N^2\varepsilon^2/\lambda \to 0$, so that with probability at least $1-\delta$,

$$\left\|\xi_{i,h}^k\right\|_{(\Lambda_{i,h}^k)^{-1}}^2 \leq 4H^2\Big[\tfrac{d}{2}\log\tfrac{N+\lambda}{\lambda} + \log\tfrac{1}{\delta}\Big] \overset{\lambda\geq 1}{\leq} 2H^2\, d\log(N+1) + 4H^2\log\tfrac{1}{\delta},$$

and consequently

$$\left\|\xi_{i,h}^k\right\|_{(\Lambda_{i,h}^k)^{-1}} \leq 2H\sqrt{d\log(NH+1) + \log\big(\tfrac{3NH}{\delta}\big)}. \tag{26}$$

Step 3: bounding the bias term (II). Since $\Lambda_{i,h}^k \succeq \lambda I_d$, we have $(\Lambda_{i,h}^k)^{-1} \preceq \lambda^{-1} I_d$, and thus

$$\left\|\mu_{i,h}^{\widehat{\pi}_{-i}^k} f\right\|_{(\Lambda_{i,h}^k)^{-1}} \leq \lambda^{-1/2}\left\|\mu_{i,h}^{\widehat{\pi}_{-i}^k} f\right\|_2.$$

Invoking the standard normalization $\|\mu_{i,h}^{\widehat{\pi}_{-i}^k}(\mathcal{S})\|_2 \leq \sqrt{d}$ together with $f(s') \in [0,H]$, one has

$$\left\|\mu_{i,h}^{\widehat{\pi}_{-i}^k} f\right\|_2 = \left\|\int f(s')\,\mu_{i,h}^{\widehat{\pi}_{-i}^k}(ds')\right\|_2 \leq \|f\|_\infty \cdot \left\|\mu_{i,h}^{\widehat{\pi}_{-i}^k}(\mathcal{S})\right\|_2 \leq H\sqrt{d}.$$

Therefore the bias term is bounded by

$$\lambda\left\|\mu_{i,h}^{\widehat{\pi}_{-i}^k} f\right\|_{(\Lambda_{i,h}^k)^{-1}} \leq \sqrt{\lambda}\cdot H\sqrt{d} \;=\; H\sqrt{d\lambda}. \tag{27}$$

The bias bound $\sqrt{\lambda}\,H\sqrt{d} = H\sqrt{d\lambda}$ requires smaller λ for a tighter bound, while the stochastic bound (I) deteriorates as $\lambda \to 0$; we therefore restrict to $\lambda \geq 1$. Note that the generative-model case takes $\lambda = 1$, recovering the standard $H\sqrt{d}$ bias (cf. 35); the online algorithm later sets $\lambda \geq c_0\log(2dnKHT/\delta) = \widetilde{\mathcal{O}}(1)$ (see the matrix-Bernstein step around (60)), so the $\sqrt{\lambda}$ factor in the bias is also $\widetilde{\mathcal{O}}(1)$ and the bias term becomes $H\sqrt{d\lambda} = \widetilde{\mathcal{O}}(H\sqrt{d})$.

Step 4: combining and pointwise version. Plugging the stochastic bound (26) and the bias bound (27) into (24), we obtain the first inequality of the theorem:

$$\left\|\widehat{\mu}_{i,h}^k f - \mu_{i,h}^{\widehat{\pi}_{-i}^k} f\right\|_{\Lambda_{i,h}^k} \leq 2H\sqrt{d\log(NH+1) + \log\big(\tfrac{3NH}{\delta}\big)} + H\sqrt{d\lambda}$$

holds with probability at least $1-\delta$. Then it follows by Cauchy–Schwarz inequality that: for all $(s,a_i) \in \mathcal{S}\times\mathcal{A}_i$,

$$\left|\phi_i(s,a_i)^\top\big(\widehat{\mu}_{i,h}^k f - \mu_{i,h}^{\widehat{\pi}_{-i}^k} f\big)\right| \leq \|\phi_i(s,a_i)\|_{(\Lambda_{i,h}^k)^{-1}}\left\|\widehat{\mu}_{i,h}^k f - \mu_{i,h}^{\widehat{\pi}_{-i}^k} f\right\|_{\Lambda_{i,h}^k}.$$

□

E.2 Proof pipeline of Theorem 2

Before presenting the proof for Algorithm 2, we introduce some general notations and auxiliary robust value functions for clarity.

Step 1: Notation conventions and worst-case transition operators.

- **Notation abbreviations**: Recall the definition of the value function:

$$V_{i,h}^{\pi,\sigma_i}(s) = \mathbb{E}_{a_i\sim\pi_{i,h}(s)}\Big[r_{i,h}^{\pi_{-i}}(s,a_i) + \inf_{\mathcal{U}^{\sigma_i}(P_{h,s,a_i}^{\pi_{-i}})} PV_{i,h+1}^{\pi,\sigma_i}\Big].$$

  Throughout the remainder of the proof, we suppress the superscript σ_i and write

$$\forall(i,h)\in[n]\times[H]: \quad V_{i,h}^{\pi} \coloneqq V_{i,h}^{\pi,\sigma_i}. \tag{28}$$

  And we denote the estimated transitions in (11) as

$$\forall(h,k)\in[H]\times[K]: P_{i,h,s,a_i}^k \coloneqq P_{i,h}^k(\cdot\mid s,a_i). \tag{29}$$

  In addition, for any distribution ζ over the space of product policies, for all $(s,i)\in\mathcal{S}\times[n]$ and any policy π_i' for the agent i, we define

$$V_{i,1}^{\star,\zeta}(s) \coloneqq \mathbb{E}_{\pi\sim\zeta}\big[V_{i,1}^{\star,\pi_{-i}}(s)\big], \quad V_{i,1}^{\zeta}(s) \coloneqq \mathbb{E}_{\pi\sim\zeta}\big[V_{i,1}^{\pi}(s)\big],$$
$$V_{i,1}^{\pi_i',\zeta}(s) \coloneqq \mathbb{E}_{\pi_{-i}\sim\zeta}\left[V_{i,1}^{\pi_i',\pi_{-i}}(s)\right]. \tag{30}$$

- **Notations for worst-case transitions**: Consider any set of policies $\{\widehat{\pi}_h^k\}_{(h,k)\in[H]\times[K]}$ with $\widehat{\pi}_h^k \in [H]\times\mathcal{S}\mapsto\Delta(\mathcal{A})$. For any value vector $V:\mathcal{S}\to[0,H]$, for any $(h,i,k)\in[H]\times[n]\times[K]$, we define the transition operators $P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V}V$ and $\widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V}V$ as:

$$P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V}V := \inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P_{h,s,a_i}^{\widehat{\pi}_{-i}^k}\right)}\mathcal{P}V, \quad \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V}V := \inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P_{i,h,s,a_i}^k\right)}\mathcal{P}V. \tag{31}$$

Step 2: Auxiliary aggregated value functions. Let $\zeta := \{\zeta_h\}_{h\in[H]}$ be a distribution over the product policy space, namely $\zeta_h:\mathcal{S}\mapsto\Delta(\prod_{i\in[n]}\Delta(\mathcal{A}_i))$. For all $(i,h,s)\in[n]\times[H]\times\mathcal{S}$, we can define the aggregated value functions $\overline{V}_{i,h}^{\zeta}(s)$ and $\overline{V}_{i,h}^{\star,\zeta}(s)$ as follows:

$$\begin{aligned}
\overline{V}_{i,h}^{\zeta}(s) &:= \tfrac{1}{K}\textstyle\sum_{k=1}^K \mathbb{E}_{a_i\sim\pi_{i,h}^k(s)}\Big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\overline{V}_{i,h+1}^{\zeta}}\overline{V}_{i,h+1}^{\zeta}\Big],\\
\overline{V}_{i,h}^{\star,\zeta}(s) &:= \max_{a_i\in\mathcal{A}_i}\tfrac{1}{K}\textstyle\sum_{k=1}^K \Big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\overline{V}_{i,h+1}^{\star,\zeta}}\overline{V}_{i,h+1}^{\star,\zeta}\Big].
\end{aligned} \tag{32}$$

Finally, for any policy π_i of agent i, we define the value function $\overline{V}_{i,h}^{\pi_i,\zeta}(s)$:

$$\overline{V}_{i,h}^{\pi_i,\zeta}(s) := \tfrac{1}{K}\textstyle\sum_{k=1}^K \mathbb{E}_{a_i\sim\widetilde{\pi}_{i,h}^{\star}(s)}\Big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\overline{V}_{i,h+1}^{\pi_i,\zeta}}\overline{V}_{i,h+1}^{\pi_i,\zeta}\Big]. \tag{33}$$

These objects will serve as the estimate to the true robust value functions by aggregating the polices output by Algorithm 2.

With the above setup in place, we are ready to prove the main theorem of Algorithm 2, starting from an error decomposition:

Step 3: Error decomposition. Our goal is to show that for all $i\in[n]$ and $s\in\mathcal{S}$, the distribution ξ over the product policies, produced by Algorithm 2, satisfies

$$\max_{i\in[n],s\in\mathcal{S}}\Big(V_{i,1}^{\star,\xi}(s) - V_{i,1}^{\xi}(s)\Big) = \max_{i\in[n],s\in\mathcal{S}}\Big(\mathbb{E}_{\pi\sim\xi}\big[V_{i,1}^{\star,\pi_{-i}}(s)\big] - \mathbb{E}_{\pi\sim\xi}\big[V_{i,1}^{\pi}(s)\big]\Big) \le \varepsilon.$$

Towards this, before proceeding, we introduce several auxiliary definitions that will be used throughout the proof. We define the best response policy as follows:

$$\forall i\in[n]: \quad \widetilde{\pi}_i^{\star} = \arg\max_{\pi_i':\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} V_{i,1}^{\pi_i',\xi}.$$

Accordingly, we have the fundamental fact that for all $s\in\mathcal{S}$

$$\overline{V}_{i,H+1}^{\xi}(s) = \overline{V}_{i,H+1}^{\widetilde{\pi}_i^{\star},\xi}(s) = \overline{V}_{i,H+1}^{\star,\xi}(s). \tag{34}$$

We then decompose the error of the value function as follows: for any agent $i\in[n]$,

$$\begin{aligned}
V_{i,1}^{\star,\xi}(s) - V_{i,1}^{\xi}(s) &= \Big(V_{i,1}^{\star,\xi}(s) - \overline{V}_{i,1}^{\widehat{\pi}_i^{\star},\xi}(s)\Big) + \Big(\overline{V}_{i,1}^{\widehat{\pi}_i^{\star},\xi}(s) - \overline{V}_{i,1}^{\xi}(s)\Big) + \Big(\overline{V}_{i,1}^{\xi}(s) - V_{i,1}^{\xi}(s)\Big)\\
&\le \underbrace{\Big(V_{i,1}^{\star,\xi}(s) - \overline{V}_{i,1}^{\widehat{\pi}_i^{\star},\xi}(s)\Big)}_{\text{term}(A)} + \underbrace{\Big(\overline{V}_{i,1}^{\star,\xi}(s) - \overline{V}_{i,1}^{\xi}(s)\Big)}_{\text{term}(B)} + \underbrace{\Big(\overline{V}_{i,1}^{\xi}(s) - V_{i,1}^{\xi}(s)\Big)}_{\text{term}(C)},
\end{aligned}$$

where the last line is due to the elementary fact that

$$\overline{V}_{i,1}^{\star,\xi}(s) = \max_{\pi_i':\mathcal{S}\times[H]\to\Delta(\mathcal{A}_i)} V_{i,1}^{\pi_i',\xi}(s) \ge \overline{V}_{i,1}^{\widehat{\pi}_i^{\star},\xi}(s).$$

We now control term A, term B and term C separately.

Step 4: Controlling Term A and C. Terms (A) and (C) both characterize the difference between a true robust value function and an estimated one, so we derive one template bound and then specialize it to these two cases. Towards this, we consider a more general case involving a given set of policies $\{\widehat{\pi}_h^k\}_{(h,k)\in[H]\times[K]}$, where either $\widehat{\pi}_h^k = \pi_h^k$, or $\widehat{\pi}_h^k = \widetilde{\pi}_i^\star \times \pi_{-i,h}^k$ for all $(h,k) \in [H]\times[K]$. Additionally, we define a uniform distribution over the aforementioned set of policies $\zeta := \{\zeta_h\}_{h\in[H]}$ so that $\zeta_h(\widehat{\pi}_h^k) = 1/K$ for all $h \in [H]$. Our objective is to derive an upper bound on $\big|V_{i,h}^\zeta(s) - \overline{V}_{i,h}^\zeta(s)\big|$ for all $i \in [n]$ over any such policy set $\{\widehat{\pi}_h^k\}_{(h,k)\in[H]\times[K]}$, which directly yields bounds for terms A and C.

To proceed, we start with the following lemma in terms of estimation errors of transition model and reward function, where the proof is postponed to Appendix E.3.1.

Lemma 2. *Let $\delta \in (0,1)$. With probability at least $1-\delta$, for all $(h,i,k) \in [H]\times[n]\times[K]$, all $(s,a_i) \in \mathcal{S}\times\mathcal{A}_i$, and any fixing value function $V : \mathcal{S} \to [0,H]$, we have*

$$\big|P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V} V - \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V} V\big| \le 2\sqrt{\tfrac{d}{N}}\big(2H\sqrt{d\ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big), \tag{35}$$

$$\big|\mathbb{E}_{a_{-i}\sim\widehat{\pi}_{-i}^k(s)}\big[r_{i,h}(s,\mathbf{a})\big] - r_{i,h}^k(s,a_i)\big| \le \sqrt{\tfrac{d}{N}}\big(2H\sqrt{d\ln(NH+1) + \ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big). \tag{36}$$

To continue, for the clarity of notations, in this subsection, we use

$$\widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}} := \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}_{i,h+1}^\zeta}, \quad \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V} := \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V_{i,h+1}^\zeta}. \tag{37}$$

We begin by expressing out the gap $\big|V_{i,h}^\zeta(s) - \overline{V}_{i,h}^\zeta(s)\big|$ under controlling using the Bellman structure and then splitting the expression into two one-sided terms that can be controlled separately. Initially, recalling the definitions in (32), we have

$$\begin{aligned}
&\big|V_{i,h}^\zeta(s) - \overline{V}_{i,h}^\zeta(s)\big| \\
=&\big|V_{i,h}^\zeta(s) - \textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i)\big] - \sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[\widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}}\,\overline{V}_{i,h+1}^\zeta\big]\big| \\
\le&\max\Big\{\underbrace{V_{i,h}^\zeta(s) - \textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i)\big] - \sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[\widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}}\,\overline{V}_{i,h+1}^\zeta\big]}_{(a)}, \\
&\underbrace{\textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i)\big] + \sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[\widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}}\,\overline{V}_{i,h+1}^\zeta\big] - V_{i,h}^\zeta(s)}_{(b)}\Big\}.
\end{aligned}$$

Now we proceed to control term (a) and term (b) separately, starting with term (a). At any time step $h \in [H]$, we have

$$\begin{aligned}
&V_{i,h}^\zeta(s) - \textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i)\big] - \sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[\widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}}\,\overline{V}_{i,h+1}^\zeta\big] \\
\overset{\text{(i)}}{=}&\textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big|\mathbb{E}_{a_{-i}\sim\widehat{\pi}_{-i}^k(s)}\big[r_{i,h}(s,\mathbf{a})\big] - r_{i,h}^k(s,a_i)\big| + \sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V} V_{i,h+1}^\zeta\big] \\
&- \textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[\widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}}\,\overline{V}_{i,h+1}^\zeta\big] \\
\le&\textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big|\mathbb{E}_{a_{-i}\sim\widehat{\pi}_{-i}^k(s)}\big[r_{i,h}(s,\mathbf{a})\big] - r_{i,h}^k(s,a_i)\big| + \sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V} V_{i,h+1}^\zeta\big] \\
&- \textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}}\,\overline{V}_{i,h+1}^\zeta\big] + \sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big|P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}}\,\overline{V}_{i,h+1}^\zeta - \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}}\,\overline{V}_{i,h+1}^\zeta\big|,
\end{aligned}$$

where (i) follows from the robust Bellman equation (23). Then applying Lemma 2 to control the reward and the transition estimation errors, we arrive at

$$\begin{aligned}
&V_{i,h}^\zeta(s) - \overline{V}_{i,h}^\zeta(s) \\
=&V_{i,h}^\zeta(s) - \textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i)\big] - \sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)}\big[\widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,\overline{V}}\,\overline{V}_{i,h+1}^\zeta\big]
\end{aligned} \tag{38}$$

$$\leq \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big(P^{\widehat{\pi}^k_{-i}, V}_{i,h,s,a_i} V^{\zeta}_{i,h+1} - P^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \big)$$
$$+ 3\sqrt{\tfrac{d}{N}} \big(2H \sqrt{d \ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d} \big) \tag{39}$$
$$\leq \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big(P^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} V^{\zeta}_{i,h+1} - P^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \big)$$
$$+ 3\sqrt{\tfrac{d}{N}} \big(2H \sqrt{d \ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d} \big) \tag{40}$$

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

By symmetry of the argument, we derive an analogous upper bound for term (b).

$$\overline{V}^{\zeta}_{i,h}(s) - V^{\zeta}_{i,h}(s) \tag{41}$$
$$= \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} [r^k_{i,h}(s,a_i)] + \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big[\widehat{P}^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \big] - V^{\zeta}_{i,h}(s)$$
$$= \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} [r^k_{i,h}(s,a_i)] + \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big[\widehat{P}^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \big]$$
$$- \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{\mathbf{a} \sim \widehat{\pi}^k_h(s)} \big[r_{i,h}(s,\mathbf{a}) \big] - \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} P^{\widehat{\pi}^k_{-i}, V}_{i,h,s,a_i} V^{\zeta}_{i,h+1}$$
$$\leq \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big[P^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \big] - \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \Big[P^{\widehat{\pi}^k_{-i}, V}_{i,h,s,a_i} V^{\zeta}_{i,h+1} \Big]$$
$$+ \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big| r^k_{i,h}(s,a_i) - \mathbb{E}_{a_{-i} \sim \widehat{\pi}^k_{-i}(s)} \big[r_{i,h}(s,\mathbf{a}) \big] \big|$$
$$+ \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big| P^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} - \widehat{P}^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \big|$$
$$\overset{\text{(i)}}{\leq} \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big[P^{\widehat{\pi}^k_{-i}, V}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \big] - \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \Big[P^{\widehat{\pi}^k_{-i}, V}_{i,h,s,a_i} V^{\zeta}_{i,h+1} \Big]$$
$$+ \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big| r^k_{i,h}(s,a_i) - \mathbb{E}_{a_{-i} \sim \widehat{\pi}^k_{-i}(s)} \big[r_{i,h}(s,\mathbf{a}) \big] \big|$$
$$+ \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big| P^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} - \widehat{P}^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \big|$$
$$\leq \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \big[P^{\widehat{\pi}^k_{-i}, V}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \big] - \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} \Big[P^{\widehat{\pi}^k_{-i}, V}_{i,h,s,a_i} V^{\zeta}_{i,h+1} \Big]$$
$$+ 3\sqrt{\tfrac{d}{N}} \big(2H \sqrt{d \ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d} \big). \tag{42}$$

where the first equality is obtained by the robust Bellman equation (23), (i) holds due to the elementary fact that $P^{\widehat{\pi}^k_{-i}, V}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1} \geq P^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \overline{V}^{\zeta}_{i,h+1}$, and the final inequality holds with probability at least $1-\delta$ for all $(i,h,s) \in [n] \times [H] \times \mathcal{S}$, by applying Lemma 2.

Now we have controlled the two one-sided differences $V^{\zeta}_{i,h}(s) - \overline{V}^{\zeta}_{i,h}(s)$ and $\overline{V}^{\zeta}_{i,h}(s) - V^{\zeta}_{i,h}(s)$ *separately* in (40) and (42).

Denoting

$$C_0 \coloneqq \tfrac{3\sqrt{d}}{\sqrt{N}} \Big(2H \sqrt{d \ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d} \Big)$$

for brevity, we rewrite (40) as below:

$$V^{\zeta}_{i,h}(s) - \overline{V}^{\zeta}_{i,h}(s) \leq \sum_{k=1}^{K} \tfrac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} P^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i} \big(V^{\zeta}_{i,h+1} - \overline{V}^{\zeta}_{i,h+1} \big) + C_0. \tag{43}$$

Note that the operator $\mathcal{Q}^{\overline{V}}_h \coloneqq \sum_{k=1}^{K} \frac{1}{K} \mathbb{E}_{a_i \sim \widehat{\pi}^k_{i,h}(s)} P^{\widehat{\pi}^k_{-i}, \overline{V}}_{i,h,s,a_i}$. Iterating (43) from h to H and using the terminal condition $V^{\zeta}_{i,H+1} = \overline{V}^{\zeta}_{i,H+1} = 0$ yields

$$V^{\zeta}_{i,h}(s) - \overline{V}^{\zeta}_{i,h}(s) \leq \mathcal{Q}^{\overline{V}}_h \mathcal{Q}^{\overline{V}}_{h+1} \cdots \mathcal{Q}^{\overline{V}}_H \big(V^{\zeta}_{i,H+1} - \overline{V}^{\zeta}_{i,H+1} \big) + (H+1-h)\, C_0$$
$$= 0 + (H+1-h)\, C_0 \;\leq\; H\, C_0 \qquad \text{for all } s \in \mathcal{S}.$$

Applying the identical argument to (42) with the probability operator $\mathcal{Q}_h^V := \sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\widehat{\pi}_{i,h}^k(s)} P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V}$ gives

$$\overline{V}_{i,h}^{\zeta}(s) - V_{i,h}^{\zeta}(s) \le H\, C_0 \qquad \text{for all } s \in \mathcal{S}.$$

Combining the two one-sided bounds via $|x| = \max\{x, -x\}$, we obtain that for all $(i,h,s) \in [n]\times[H]\times\mathcal{S}$, with probability at least $1-\delta$:

$$\left|V_{i,h}^{\zeta}(s) - \overline{V}_{i,h}^{\zeta}(s)\right| \le H\, C_0 = \frac{3\sqrt{d}\,H}{\sqrt{N}}\Big(2H\sqrt{d\ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\Big). \quad (44)$$

Finally, we instantiate (44) for the two cases $\widehat{\pi}_h^k = \pi_h^k$ and $\widehat{\pi}_h^k = \widetilde{\pi}_i^\star \times \pi_{-i,h}^k$, which correspond to term (C) and term (A) respectively. With probability at least $1-\delta$, there exists an absolute constant c such that

$$\begin{aligned}
\text{term}(A) &= V_{i,1}^{\star,\xi}(s) - \overline{V}_{i,1}^{\widetilde{\pi}_i^\star,\xi}(s) \le \frac{c\sqrt{d}H^2}{\sqrt{N}}\sqrt{d\ln\big(\tfrac{NHnK}{\delta}\big)} \\
\text{term}(C) &= \overline{V}_{i,1}^{\xi}(s) - V_{i,1}^{\xi}(s) \le \frac{c\sqrt{d}H^2}{\sqrt{N}}\sqrt{d\ln\big(\tfrac{NHnK}{\delta}\big)}.
\end{aligned} \quad (45)$$

Step 5: Controlling Term B. Recall that $\text{term}(B) = \overline{V}_{i,1}^{\star,\xi}(s) - \overline{V}_{i,1}^{\xi}(s)$, which measures the gap between the best-response aggregated value $\overline{V}_{i,1}^{\star,\xi}$ and the own-policy aggregated value $\overline{V}_{i,1}^{\xi}$ at $h=1$. This is an intrinsic optimization error arising from the FTRL-based policy updates under the coupled, non-stationary multi-agent dynamics. We control it by first replacing $\overline{V}_{i,h}^{\star,\xi}$ with an optimistic surrogate $\widehat{V}_{i,h}$ satisfying $\widehat{V}_{i,h} \ge \overline{V}_{i,h}^{\star,\xi}$ (Lemma 3), then bounding the gap $\widehat{V}_{i,h}(s) - \overline{V}_{i,h}^{\xi}(s)$. For brevity, for all $(h,k,i,s,a_i) \in [H]\times[K]\times[n]\times\mathcal{S}\times\mathcal{A}_i$, we abbreviate the definitions in (31) in this subsection as below:

$$\begin{aligned}
\widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\widehat{V}} &:= \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\widehat{V}_{i,h+1}} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P_{i,h,s,a_i}^k\right)} \mathcal{P}\widehat{V}_{i,h+1}, \\
P_{i,h,s,a_i}^{\pi_{-i}^k,\overline{V}} &:= P_{i,h,s,a_i}^{\pi_{-i}^k,\overline{V}_{i,h+1}^{\xi}} = \operatorname{argmin}_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P_{h,s,a_i}^{\pi_{-i}^k}\right)} \mathcal{P}\overline{V}_{i,h+1}^{\xi}.
\end{aligned} \quad (46)$$

To continue, we first introduce the following lemma, which establishes that the value $\widehat{V}_{i,h}$ is an optimistic version of the best-response surrogate $\overline{V}_{i,h}^{\star,\xi}$; its proof is deferred to Appendix E.3.2:

Lemma 3. *With probability at least $1-\delta$, for all $(i,h,s) \in [n]\times[H]\times\mathcal{S}$, with the bonus factors $\beta_{i,h}(s)$ defined in* (15)*, it holds that*

$$\widehat{V}_{i,h}(s) \ge \overline{V}_{i,h}^{\star,\xi}.$$

Applying Lemma 3, we first replace $\overline{V}_{i,h}^{\star,\xi}$ by its optimistic quantity $\widehat{V}_{i,h}$ and choose to control the corresponding gap as below:

$$\overline{V}_{i,h}^{\star,\xi}(s) - \overline{V}_{i,h}^{\xi}(s) \le \widehat{V}_{i,h}(s) - \overline{V}_{i,h}^{\xi}(s). \quad (47)$$

Using techniques and arguments analogous to those in the proof of Lemma 3, we obtain the following lemma and corrolary:

Lemma 4. *With probability at least $1-\delta$, it holds that*

$$\widehat{V}_{i,h}(s) \ge \overline{V}_{i,h}^{\xi}(s), \quad \textit{for all } (i,h,s) \in [n]\times[H]\times\mathcal{S}.$$

Corollary 5.1. *With probability at least $1-\delta$, it holds that for all $(i,h,s) \in [n]\times[H]\times\mathcal{S}$,*

$$\begin{aligned}
&\textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\widehat{V}}\widehat{V}_{i,h+1}\big] + \beta_{i,h}(s) \\
\ge &\textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\overline{V}}\overline{V}_{i,h+1}^{\xi}\big]
\end{aligned}$$

Combining Lemma 4 and Corollary 5.1,

recalling the definition of $\widehat{V}_{i,h}$ in (14) gives that

$$\begin{aligned}\widehat{V}_{i,h}(s) &= \min\Big\{\textstyle\sum_{k=1}^{K}\frac{1}{K}\big\langle\pi_{i,h}^{k}(\cdot\mid s), q_{i,h}^{k}(s,\cdot)\big\rangle + \beta_{i,h}(s), H-h+1\Big\},\\ &= \min\big\{\textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[r_{i,h}^{k}(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k},\widehat{V}}\widehat{V}_{i,h+1}\big] + \beta_{i,h}(s), H-h+1\big\}\end{aligned}$$

Therefore, term (B) can be controlled by

$$\begin{aligned}&\overline{V}_{i,h}^{\star,\xi}(s) - \overline{V}_{i,h}^{\xi}(s) \le \widehat{V}_{i,h}(s) - \overline{V}_{i,h}^{\xi}(s)\\ &\overset{\text{(i)}}{\le} \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[r_{i,h}^{k}(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k},\widehat{V}}\widehat{V}_{i,h+1}\big] + \beta_{i,h}(s)\\ &\qquad - \textstyle\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\Big[r_{i,h}^{k}(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}^{\zeta}_{i,h+1}}\overline{V}^{\zeta}_{i,h+1}\Big]\\ &= \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[\widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k},\widehat{V}}\widehat{V}_{i,h+1}\big] + \beta_{i,h}(s)\\ &\qquad - \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[\widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}}\overline{V}^{\xi}_{i,h+1}\big]\\ &\le \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[P_{i,h,s,a_i}^{\pi_{-i}^{k},\widehat{V}}\widehat{V}_{i,h+1}\big] + \beta_{i,h}(s)\\ &\qquad - \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[P_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}}\overline{V}^{\xi}_{i,h+1}\big]\\ &\quad + \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\Big|\widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k},\widehat{V}}\widehat{V}_{i,h+1} - P_{i,h,s,a_i}^{\pi_{-i}^{k},\widehat{V}}\widehat{V}_{i,h+1}\Big|\\ &\quad + \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\Big|P_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}}\overline{V}^{\xi}_{i,h+1} - \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}}\overline{V}^{\xi}_{i,h+1}\Big|. \end{aligned}\tag{48}$$

Here, (i) holds by recalling the definition of $\widehat{V}_{i,h}$ in (14) and the definition of $\overline{V}_{i,h}^{\xi}(s)$ in (32). Furthermore, we apply Lemma 2 and arrive at: with probability at least $1-\delta$,

$$\begin{aligned}&\overline{V}_{i,h}^{\star,\xi}(s) - \overline{V}_{i,h}^{\xi}(s) \le \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[P_{i,h,s,a_i}^{\pi_{-i}^{k},\widehat{V}}\widehat{V}_{i,h+1}\big] + \beta_{i,h}(s)\\ &\quad - \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[P_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}}\overline{V}^{\xi}_{i,h+1}\big] + \frac{4\sqrt{d}}{\sqrt{N}}\big(2H\sqrt{d\ln(NH+1)+2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big)\\ &\le \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[P_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}}\widehat{V}_{i,h+1}\big] + \beta_{i,h}(s)\\ &\quad - \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\big[P_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}}\overline{V}^{\xi}_{i,h+1}\big] + \frac{4\sqrt{d}}{\sqrt{N}}\big(2H\sqrt{d\ln(NH+1)+2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big)\\ &= \textstyle\sum_{k=1}^{K}\frac{1}{K}\mathbb{E}_{a_i\sim\pi_{i,h}^{k}(s)}\Big[P_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}}\Big(\widehat{V}_{i,h+1} - \overline{V}^{\xi}_{i,h+1}\Big)\Big]\\ &\quad + \beta_{i,h}(s) + \frac{4\sqrt{d}}{\sqrt{N}}\big(2H\sqrt{d\ln(NH+1)+2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big)\end{aligned}\tag{49}$$

where the penultimate inequality holds due to the elementary fact that $P_{i,h,s,a_i}^{\pi_{-i}^{k},\widehat{V}}\widehat{V}_{i,h+1} \le P_{i,h,s,a_i}^{\pi_{-i}^{k},\overline{V}}\widehat{V}_{i,h+1}$. Finally, applying the recursive bound in (49) recursively for $h=1,2,\ldots,H$ yields

$$\begin{aligned}\text{term}(B) =& \overline{V}_{1,h}^{\star,\xi}(s) - \overline{V}_{1,h}^{\xi}(s)\\ \le& \textstyle\sum_{i=1}^{H}\beta_{i,h}(s) + H\cdot\frac{4\sqrt{d}}{\sqrt{N}}\big(2H\sqrt{d\ln(NH+1)+2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big)\\ \le& \textstyle\frac{12\sqrt{d}H}{\sqrt{N}}\big(2H\sqrt{d\ln(NH+1)+2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big) + 2H^2\sqrt{\frac{\ln A_i}{K}}.\end{aligned}\tag{50}$$

where the first inequality holds by $\widehat{V}_{i,H+1}(s)=0$ and the fundamental fact $\overline{V}_{i,H+1}^{\xi}(s)=0$ in (34) for all $(i,h)\in[n]\times[H]$.

Step 6: Summing up the results. To complete the proof, combining the bounds for terms (A) and (C) in (45) with that for term (B) in (50), we obtain

$$V_{i,1}^{\star,\xi}(s) - V_{i,1}^{\xi}(s) \lesssim \frac{dH^2}{\sqrt{N}}\sqrt{\ln\left(\frac{KNnH}{\delta}\right)} + H^2\sqrt{\frac{\ln A_i}{K}}.$$

Consequently, Algorithm 2 outputs an ε-approximate CCE ξ, i.e.,

$$\max_{i\in[n],s\in\mathcal{S}}\left(\mathbb{E}_{\pi\sim\xi}\left[V_{i,1}^{\star,\pi_{-i}}(s)\right] - \mathbb{E}_{\pi\sim\xi}\left[V_{i,1}^{\pi}(s)\right]\right) \le \varepsilon$$

with probability at least $1-\delta$, provided that $N \ge CH^4d^2\ln(KNnH/\delta)/\varepsilon^2$ and $K \ge CH^4\ln(\max_{i\in[n]}A_i)/\varepsilon^2$ for some universal constant C, since then each of the two terms above is at most $\varepsilon/2$. Consequently, the total number of samples drawn during learning satisfies

$$N_{\mathsf{all}} \le HK(N + d(d+1)/2) = \tilde{\mathcal{O}}\big(H^9d^2/\varepsilon^4\big).$$

Here the inequality holds because each step h and iteration k draws $\sum_{(s,a_i)\in\mathcal{Y}_i}\lceil N\rho_i(s,a_i)\rceil \le N + |\mathcal{Y}_i| \le N + d(d+1)/2$ samples, and the last step uses $d(d+1)/2 \le d^2 \le N$. Thus, we finish the proof of Theorem 2.

E.3 Proof of auxiliary results

E.3.1 Proof of Lemma 2

Step 1: Proof of the transition estimation gap. First, let's consider any fixed triple $(h,i,k) \in [H]\times[n]\times[K]$. Applying strong duality property [30], we can rewrite the transition estimation error as

$$\begin{aligned}
&\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}\left|P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V}V - \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^k,V}V\right| \\
\overset{\text{(i)}}{=}&\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}\Bigg|\max_{\alpha\in\left[\min_s V(s),\max_s V(s)\right]}\left[P_{h,s,a_i}^{\widehat{\pi}_{-i}^k}[V]_\alpha - \sigma_i\big(\alpha - \min_{s'}[V]_\alpha(s')\big)\right] \\
&- \max_{\alpha\in\left[\min_s V(s),\max_s V(s)\right]}\left[P_{i,h,s,a_i}^{k}[V]_\alpha - \sigma_i\big(\alpha - \min_{s'}[V]_\alpha(s')\big)\right]\Bigg| \\
\le&\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}\max_{\alpha\in\left[\min_s V(s),\max_s V(s)\right]}\left|P_{h,s,a_i}^{\widehat{\pi}_{-i}^k}[V]_\alpha - P_{i,h,s,a_i}^{k}[V]_\alpha\right|,
\end{aligned}$$

where (i) holds by refering to the definitions in (31) and then apply the dual form of robust Bellman operator [65, Lemma 1] due to strong duality property [30]. See the definiton of P_{i,h,s,a_i}^k in (29). For a fixed α, we have

$$\begin{aligned}
&\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}\left|P_{h,s,a_i}^{\widehat{\pi}_{-i}^k}[V]_\alpha - P_{i,h,s,a_i}^{k}[V]_\alpha\right| \\
\overset{\text{(i)}}{=}&\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}\left|\phi_i(s,a_i)^\top\mu_{i,h}^{\widehat{\pi}_{-i}^k}[V]_\alpha - \phi_i(s,a_i)^\top\widehat{\mu}_{i,h}^{k}[V]_\alpha\right| \\
\overset{\text{(ii)}}{\le}&\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}\left\|\phi_i(s,a_i)\right\|_{(\Lambda_{i,h}^k)^{-1}}\Big(2H\sqrt{d\ln(NH+1)+\ln\left(\tfrac{3NH}{\delta}\right)} + H\sqrt{d}\Big),
\end{aligned}$$

where (i) holds due to the linear assumption of Markov games in (2) and the definition of the estimated transition kernel in (11), and (ii) follows by applying Theorem 5 with $f = [V]_\alpha$ (which is a fixed function with range $[0,H]$ for the given α) and $\lambda = 1$, which holds with probability at least $1-\delta$.

Furthermore, we apply Lemma 1 with $\Sigma = \frac{\Lambda_{i,h}^k}{|\mathcal{D}_i|}$ and $x = \phi_i(s,a_i)$ for all (s,a_i) and arrive at

$$\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}\left|P_{h,s,a_i}^{\widehat{\pi}_{-i}^k}[V]_\alpha - P_{i,h,s,a_i}^{k}[V]_\alpha\right| \le \sqrt{\tfrac{d}{N}}\Big(2H\sqrt{d\ln(NH+1)+\ln\left(\tfrac{3NH}{\delta}\right)} + H\sqrt{d}\Big), \tag{51}$$

with probability at least $1-\delta$. To extend the pointwise bound towards a union bound for all $\alpha \in [0,H]$, we define a ε_1-net denoted as N_{ε_1} for α. To cover $\alpha \in [0,H]$, the net size bounded by $|N_{\varepsilon_1}| \le 3H/\varepsilon_1$ [55] allows us to apply the union bound:

$$\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i}\max_{\alpha\in\left[\min_s V(s),\max_s V(s)\right]}\left|P_{h,s,a_i}^{\widehat{\pi}_{-i}^k}[V]_\alpha - P_{i,h,s,a_i}^{k}[V]_\alpha\right|$$

$$
\begin{aligned}
&\leq \max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \max_{\alpha\in N_{\varepsilon_1}} \left|P^{\widehat{\pi}^k_{-i}}_{h,s,a_i}[V]_\alpha - P^k_{i,h,s,a_i}[V]_\alpha\right| + \varepsilon_1 \\
&\leq \sqrt{\tfrac{d}{N}}\big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3NH|N_{\varepsilon_1}|}{\delta}\big)} + H\sqrt{d}\big) + \varepsilon_1 \\
&\leq 2\sqrt{\tfrac{d}{N}}\big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3NH}{\delta}\big)} + H\sqrt{d}\big),
\end{aligned} \tag{52}
$$

where the first inequality follows because the chosen α lies within an ε-ball centered at some point in N_{ε_1}, and the function $\left|P^{\widehat{\pi}^k_{-i}}_{h,s,a_i}[V]_\alpha - P^k_{i,h,s,a_i}[V]_\alpha\right|$ is 1-Lipschitz in α; the penultimate line holds by appying (51); and the last line is induced by letting $\varepsilon_1 = 1/N$ and $|N_{\varepsilon_1}| \leq 3HN$.

Note that (52) is established for a *fixed* triple $(h,i,k) \in [H]\times[n]\times[K]$, with probability at least $1-\delta$. To upgrade this pointwise bound to a union bound over all $(h,i,k) \in [H]\times[n]\times[K]$, we apply for all nKH triples and arrive at: with probability at least $1-\delta$, for all $(h,i,k) \in [H]\times[n]\times[K]$, and any fixed value function $V : \mathcal{S} \to [0,H]$, it obeys

$$
\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \left|P^{\widehat{\pi}^k_{-i},V}_{i,h,s,a_i} V - \widehat{P}^{\widehat{\pi}^k_{-i},V}_{i,h,s,a_i} V\right| \leq 2\sqrt{\tfrac{d}{N}}\big(2H\sqrt{d\ln(NH+1)+2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big).
$$

Proof of the reward estimation gap. The reward bound follows the same proof structure as that used for the transition kernel, but with a simpler one-dimensional parameter vector. Based on the linear assumption of the reward function $r_{i,h}$ in (2) and the construction of the estimated reward function $r^k_{i,h}$ in (11), we derive that

$$
\begin{aligned}
&\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \left|\mathbb{E}_{a_{-i}\sim\widehat{\pi}^k_{-i}(s)} r_{i,h}(s,\mathbf{a}) - r^k_{i,h}(s,a_i)\right| \\
&\quad = \max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \left|\phi_i(s,a_i)^\top \theta^{\widehat{\pi}^k_{-i}}_{i,h} - \phi_i(s,a_i)^\top \widehat{\theta}^k_{i,h}\right|.
\end{aligned}
$$

Applying Theorem 5 with the scalar target $f = \mathbb{E}_{a_{-i}\sim\widehat{\pi}^k_{-i}(s)}[r_{i,h}(s,\mathbf{a})] \in [0,H]$ and $\lambda = 1$, we instantiate the ridge-regression concentration bound for the one-dimensional reward parameter and obtain that with probability at least $1-\delta$, for a fixed (h,i,k):

$$
\begin{aligned}
&\max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \left|\mathbb{E}_{a_{-i}\sim\widehat{\pi}^k_{-i}(s)} r_{i,h}(s,\mathbf{a}) - r^k_{i,h}(s,a_i)\right| \\
&\leq \max_{(s,a_i)\in\mathcal{S}\times\mathcal{A}_i} \left\|\phi_i(s,a_i)\right\|_{\left(\Lambda^k_{i,h}\right)^{-1}} \cdot \big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{1}{\delta}\big)} + H\sqrt{d}\big) \\
&\leq \frac{\sqrt{d}}{\sqrt{N}} \cdot \big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{1}{\delta}\big)} + H\sqrt{d}\big),
\end{aligned} \tag{53}
$$

where the last inequality holds by applying Lemma 1 with $\Sigma = \frac{\Lambda^k_{i,h}}{|\mathcal{D}_i|}$ and $x = \phi_i(s,a_i)$ for all (s,a_i).

As in the transition case, we will transfer this pointwise bound (53) the high-probability event underlying the ridge-regression bound above is defined relative to the samples used to estimate $\widehat{\theta}^k_{i,h}$ for a *fixed* triple $(h,i,k) \in [H]\times[n]\times[K]$ for any fixed(h,i,k) to a union bound over the nKH triples of the space $[H]\times[n]\times[K]$. Consequently, it obeys that with probability at least $1-\delta$, for all $(h,i,k) \in [H]\times[n]\times[K]$,

$$
\left|\mathbb{E}_{a_{-i}\sim\widehat{\pi}^k_{-i}(s)} r_{i,h}(s,\mathbf{a}) - r^k_{i,h}(s,a_i)\right| \leq \frac{\sqrt{d}}{\sqrt{N}}\big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3NHnK}{\delta}\big)} + H\sqrt{d}\big).
$$

E.3.2 Proof of Lemma 3

Before proceeding, we first introduce an exsiting result for Follow-the-Regularized-Leader (FTRL) (see Shalev-Shwartz [61, Theorem 6] and also Li et al. [42, Theorem 3]) that we will resort to: for all $(i,h) \in [n]\times[H]$:

$$
\begin{aligned}
&\max_{a_i\in\mathcal{A}_i} \textstyle\sum_{k=1}^K \frac{1}{K}\big[r^k_{i,h}(s,a_i) + \widehat{P}^{\pi^k_{-i},\widehat{V}}_{i,h,s,a_i} \widehat{V}_{i,h+1}\big] \\
&\leq \textstyle\sum_{k=1}^K \frac{1}{K}\mathbb{E}_{a_i\sim\pi^k_{i,h}(s)}\big[r^k_{i,h}(s,a_i) + \widehat{P}^{\pi^k_{-i},\widehat{V}}_{i,h,s,a_i} \widehat{V}_{i,h+1}\big] + 2H\sqrt{\tfrac{\ln A_i}{K}},
\end{aligned} \tag{54}
$$

which can be obtained byc instantiating the standard FTRL regret bound in our setting as follows. For any fixed time step h, agent i, and state $s \in \mathcal{S}$, view the K iterations as an

online-learning process for the agent. At each iteration k, the per-round reward is defnied as $\ell_{i,h}^k(a_i) \coloneqq r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\widehat{V}} \widehat{V}_{i,h+1}$ for all $a_i \in \mathcal{A}_i$ and $\widehat{V}_{i,h+1} \in [0,H]$. During the prpocess, the agent i plays the mixed strategy $\pi_{i,h}^k(\cdot \mid s) \in \Delta(\mathcal{A}_i)$ produced by FTRL with entropy regularizer. Applying Shalev-Shwartz [61, Theorem 6] (equivalently Li et al. [42, Theorem 3]) with per-round reward range size H and action-space size $|\mathcal{A}_i| = A_i$ yields the regret bound $\max_{a_i} \sum_{k=1}^K \ell_{i,h}^k(a_i) - \sum_{k=1}^K \mathbb{E}_{a_i \sim \pi_{i,h}^k(s)}\big[\ell_{i,h}^k(a_i)\big] \le 2H\sqrt{K \ln A_i}$; dividing by K gives (54).x

Armed with the key results, we shall prove Lemma 3 using induction arguments. Initially, for the base case at the final time step $H+1$, we directly have

$$\text{for all } (i,s) \in [n] \times \mathcal{S}: \quad \widehat{V}_{i,H+1}(s) = 0 = \overline{V}_{i,H+1}^{\star,\xi}(s),$$

due to the definition of $\widehat{V}_{i,H+1}$ and the fundamental fact in (34). We proceed by backward induction from the last step $h = H+1$ to $h = 1$. Assume that $\widehat{V}_{i,h+1}(s) \ge \overline{V}_{i,h+1}^{\star,\xi}(s)$ holds for all $(i,s) \in [n] \times \mathcal{S}$. Our goal is to show that $\widehat{V}_{i,h}(s) \ge \overline{V}_{i,h}^{\star,\xi}(s)$ for all $(i,s) \in [n] \times \mathcal{S}$.

Before proceeding, to invoke the FTRL regret bound in (54), we must bound the range of the per-round gain $\ell_{i,h}^k(a_i) \coloneqq r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi - i^k,\widehat{V}} \widehat{V}_{i,h+1}$. By the definition of $\widehat{V}_{i,h}$ in (14), we have $\widehat{V}_{i,h}(s) = \min\big\{ \sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\widehat{V}} \widehat{V}_{i,h+1}\big] + \beta_{i,h}(s),\, H - h + 1\big\}$. It immediately yields $\widehat{V}_{i,h}(s) \le H - h + 1 \le H$. For the lower bound, we proceed by backward induction on h: with the base case $\widehat{V}_{i,H+1} = 0$, and given $\widehat{V}_{i,h+1} \ge 0$, each summand of the first argument is non-negative—$r_{i,h}^k(s,a_i) \ge 0$ by assumption, $\widehat{P}_{i,h,s,a_i}^{\pi - i^k,\widehat{V}} \widehat{V}_{i,h+1} \ge 0$ as a probability kernel applied to a non-negative function, and $\beta_{i,h}(s) \ge 0$. Hence $\widehat{V}_{i,h}(s) \in [0,H]$ for all $(i,h) \in [n] \times [H]$. Now we are ready to verify the induction hypothesis for step h. Towards this, recalling the definition of $\overline{V}_{i,h}^{\star,\xi}(s)$ (see (32)) gives

$$\begin{aligned}
&\overline{V}_{i,h}^{\star,\xi}(s) = \max_{a_i \in \mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\overline{V}_{i,h+1}^{\star,\xi}} \overline{V}_{i,h+1}^{\star,\xi}\big] \\
&\overset{\text{(i)}}{\le} \max_{a_i \in \mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\big[r_{i,h}^k(s,a_i) + P_{i,h,s,a_i}^{\pi_{-i}^k,\overline{V}_{i,h+1}^{\star,\xi}} \overline{V}_{i,h+1}^{\star,\xi}\big] \\
&\quad + \tfrac{2\sqrt{d}}{\sqrt{N}}\big(2H\sqrt{d\ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big) \\
&\overset{\text{(ii)}}{\le} \max_{a_i \in \mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\big[r_{i,h}^k(s,a_i) + P_{i,h,s,a_i}^{\pi_{-i}^k,\widehat{V}} \widehat{V}_{i,h+1}\big] \\
&\quad + \tfrac{2\sqrt{d}}{\sqrt{N}}\big(2H\sqrt{d\ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big) \\
&\overset{\text{(iii)}}{\le} \max_{a_i \in \mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\widehat{V}} \widehat{V}_{i,h+1}\big] \\
&\quad + \tfrac{4\sqrt{d}}{\sqrt{N}}\big(2H\sqrt{d\ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big) \\
&\overset{\text{(iv)}}{\le} \sum_{k=1}^K \tfrac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\widehat{V}} \widehat{V}_{i,h+1}\big] + \beta_{i,h}(s),
\end{aligned} \tag{55}$$

Here, (i) and (iii) follows from applying Lemma 2, (ii) follows from the induction assumption, and (iv) is based on the results of FTRL in (54) and recalling $\beta_{i,h}(s)$ in (15) for all $(i,s) \in [n] \times \mathcal{S}$:

$$\beta_{i,h}(s) = 8\sqrt{\tfrac{d}{N}}\big(2H\sqrt{d\ln(NH+1) + 2\ln\big(\tfrac{3KNHn}{\delta}\big)} + H\sqrt{d}\big) + 2H\sqrt{\tfrac{\ln A_i}{K}}.$$

On the other hand, recalling the definition of $\widehat{V}_{i,h}$ in (14) gives that

$$\begin{aligned}
\widehat{V}_{i,h}(s) &= \min\Big\{\textstyle\sum_{k=1}^K \frac{1}{K}\big\langle \pi_{i,h}^k(\cdot \mid s), q_{i,h}^k(s,\cdot)\big\rangle + \beta_{i,h}(s), H - h + 1\Big\}, \\
&= \min\big\{ \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^k(s)}\big[r_{i,h}^k(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^k,\widehat{V}} \widehat{V}_{i,h+1}\big] + \beta_{i,h}(s), H - h + 1\big\}
\end{aligned}$$

where the last equality holds by the definitions of $q_{i,h}^k$ in (12). Therefore, with the fact that $\overline{V}_{i,h}^{\star,\xi}(s) \leq H-h+1$ for all $(i,s) \in [n] \times \mathcal{S}$, we obtain that

$$\forall (i,s) \in [n] \times \mathcal{S}: \quad \overline{V}_{i,h}^{\star,\xi}(s) \leq \widehat{V}_{i,h}(s).$$

Therefore, we have finished the proof of the lemma by induction arguments.

F Proof of Theorem 3

We now turn to the proof for the online interactive setting. The overall structure parallels the generative case: we begin with estimation lemmas, then construct optimistic and pessimistic surrogates, and finally combine them to derive a recursive relationship, which leads to the regret bound.

Step 1: error decomposition. First, recall that the objective in the online interactive setting is to upper bound the following regret term induced by the process of Algoarithm 3:

$$\max_{i\in[n]} \sum_{t=1}^T \mathbb{E}_{\pi\sim\xi^t}\left[V_{i,1}^{\star,\pi_{-i},\sigma_i}(s)\right] - \mathbb{E}_{\pi\sim\xi^t}\left[V_{i,1}^{\pi,\sigma_i}(s)\right] = \max_{i\in[n]} \sum_{t=1}^T V_{i,1}^{\star,\xi^t}(s) - V_{i,1}^{\xi^t}(s).$$

Before continuning, for any fixed set of policies $\left\{\widehat{\pi}^{k.t}\right\}_{(k,t)\times[K]\times[T]}$ with $\widehat{\pi}^{k.t} \in [H]\times\mathcal{S} \mapsto \prod_{i=1}^n \Delta(\mathcal{A}_i)$, we define the following useful auxiliary notations of transition kernels:

$$P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^{k,t},V} V \coloneqq \inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P_{h,s,a_i}^{\widehat{\pi}_{-i}^{k,t}}\right)} \mathcal{P}V, \qquad \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^{k,t},V} V \coloneqq \inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P_{i,h,s,a_i}^{k,t}\right)} \mathcal{P}V,$$

where $P_{i,h,s,a_i}^{k,t}$ is the estimated transition kernel output by Algoarithm 3.

Armed with the notations, we first introduce the following lemma, which bounds the estimation errors of the transition probability and the reward function. The proof, which is analogous to that of Lemma 2 with the added union bound over $t \in [T]$, is provided in Appendix F.1.1.

Lemma 5. *Consider any fixed $(h,i,k,t) \in [H]\times[n]\times[K]\times[T]$, fixed value function $V: \mathcal{S}\to[0,H]$ and $\delta \in (0,1)$. With probability at least $1-2\delta$, the following hold simultaneously for all $(s,a_i) \in \mathcal{S}\times\mathcal{A}_i$:*

$$\begin{aligned}
&\left|P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^{k,t},V} V - \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^{k,t},V} V\right| \\
&\quad \leq \left\|\phi(s,a_i)\right\|_{\left(\Lambda_{i,h}^{k,t}\right)^{-1}} \left(2H\sqrt{d\ln(NH+1) + 2\ln\left(\tfrac{3TKNHn}{\delta}\right)} + H\sqrt{d\lambda}\right) + \tfrac{1}{N}, \\
&\left|\mathbb{E}_{a_{-i}\sim\widehat{\pi}_{-i}^{k,t}(s)}\left[r_{i,h}(s,\mathbf{a})\right] - r_{i,h}^{k,t}(s,a_i)\right| \\
&\quad \leq \left\|\phi(s,a_i)\right\|_{\left(\Lambda_{i,h}^{k,t}\right)^{-1}} \left(2H\sqrt{d\ln(NH+1) + \ln\left(\tfrac{3NTHnK}{\delta}\right)} + H\sqrt{d\lambda}\right),
\end{aligned} \tag{56}$$

where $\Lambda_{i,h}^{k,t}$ was defined in (16) *with regularization $\lambda \geq 1$.*

To continue, recalling the definition of the estimated value function $\overline{V}_{i,h}^t(s)$ in (18), analogous to the generative case, we will show that $\overline{V}_{i,h}^t(s)$ is an optimistic estimation of the best-response value function $V_{i,h}^{\star,\xi^t}(s)$ simultaneously for all $h \in [H]$.

Lemma 6. *For all $(t,i,h,s) \in [T]\times[n]\times[H]\times\mathcal{S}$, invoking the optimistic bonus term defined in* (19)*, with probability at least $1-\delta$, the following inequality holds true:*

$$\forall s \in \mathcal{S}: \quad \overline{V}_{i,h}^t(s) \geq V_{i,h}^{\star,\xi^t}(s).$$

Unlike the generative case, we also maintain a collection of pessimistic estimates of the value function, and will show that $\underline{V}_{i,h}^t(s)$ serves as a lower bound for $V_{i,h}^{\xi^t}(s)$.

Lemma 7. *Invoking the pessimistic bonus in* (20)*, with probability at least $1-\delta$,*

$$\underline{V}_{i,h}^t(s) \leq V_{i,h}^{\xi^t}(s),$$

holds true for all $(i,h,t,s) \in [n]\times[H]\times[T]\times\mathcal{S}$.

The proof of Lemma 6 and Lemma 7 are postponed to Appendix F.1. Now we are now ready to prove Theorem 3. Applying Lemma 6 and Lemma 7, we have

$$
\begin{aligned}
&V_{i,h}^{\star,\xi^t}(s) - V_{i,h}^{\xi^t}(s) \le \overline{V}_{i,h}^t(s) - \underline{V}_{i,h}^t(s) \\
&\overset{\text{(i)}}{\le} \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big[r_{i,h}^{k,t}(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \overline{V}} \overline{V}_{i,h+1}^t \big] + \beta_{i,h,1}^t(s) + \beta_{i,h,2}^t(s) \\
&\quad - \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big[r_{i,h}^{k,t}(s,a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \underline{V}} \underline{V}_{i,h+1}^t \big] \\
&\le \beta_{i,h,1}^t(s) + \beta_{i,h,2}^t(s) + \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big[P_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \overline{V}} \overline{V}_{i,h+1}^t \big] \\
&\quad + \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big| \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \overline{V}} \overline{V}_{i,h+1}^t - P_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \overline{V}} \overline{V}_{i,h+1}^t \big| - \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big[P_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \underline{V}} \underline{V}_{i,h+1}^t \big] \\
&\quad + \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big| P_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \underline{V}} \underline{V}_{i,h+1}^t - \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \underline{V}} \underline{V}_{i,h+1}^t \big| \\
&\overset{\text{(ii)}}{\le} \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big[P_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \overline{V}} \overline{V}_{i,h+1}^t - P_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \underline{V}} \underline{V}_{i,h+1}^t \big] + \beta_{i,h,1}^t(s) + \beta_{i,h,2}^t(s) \\
&\quad + \textstyle\sum_{k=1}^K \frac{1}{K} \max_{a_i \in \mathcal{A}_i} \big\| \phi_i(s,a_i) \big\|_{\left(\Lambda_{i,h}^{k,t}\right)^{-1}} \\
&\qquad \cdot \Big(4H \sqrt{d \ln(NH+1) + \ln\big(\tfrac{3NTHnK}{\delta}\big)} + 2H\sqrt{d\lambda} \Big) + \tfrac{1}{N} \\
&\le \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big[P_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \underline{V}} \big(\overline{V}_{i,h+1}^t - \underline{V}_{i,h+1}^t \big) \big] + \beta_{i,h,1}^t(s) + \beta_{i,h,2}^t(s) \\
&\quad + \textstyle\sum_{k=1}^K \frac{1}{K} \max_{a_i \in \mathcal{A}_i} \big\| \phi_i(s,a_i) \big\|_{\left(\Lambda_{i,h}^{k,t}\right)^{-1}} \\
&\qquad \cdot \Big(4H \sqrt{d \ln(NH+1) + \ln\big(\tfrac{3NTHnK}{\delta}\big)} + 2H\sqrt{d\lambda} \Big) + \tfrac{1}{N}. \qquad (57)
\end{aligned}
$$

where (i) is obtained by recalling the definition of $\overline{V}_{i,h}^t(s)$ and $\underline{V}_{i,h}^t(s)$ in (18), (ii) is obtained by applying Lemma 5 to replace the empirical estimate $\widehat{P}$ by their ground truth, and the last inequality holds with probability at least $1-\delta$.

Step 2: controlling the second term in (57)**.** To continue, we focus on the second term in (57). We first show that, for $\lambda \ge c_0 \log(2dnKHT/\delta) = \widetilde{\mathcal{O}}(1)$, the empirical Gram matrix $\Lambda_{i,h}^{k,t}$ concentrates multiplicatively around its expectation.

Recall the empirical Gram matrix from (16):

$$
\Lambda_{i,h}^{k,t} = \textstyle\sum_{m=1}^{|\mathcal{D}_i|} \phi_i^{(m)} (\phi_i^{(m)})^\top + \lambda I_d, \qquad \text{where} \qquad \phi_i^{(m)} \coloneqq \phi_i\big(s_h^{(m)}, a_{i,h}^{(m)}\big).
$$

which is a positive definite matrix satisfying $\Lambda_{i,h}^{k,t} \succeq \lambda I_d$. By invoking the Hybrid-Sampling procedure in Algorithm 1, the dataset $\mathcal{D}_i$ at round t contains $|\mathcal{D}_i| \approx N$ samples. Specifically, for each $l \in [t-1]$, we independently collect $\lceil N/t \rceil$ samples conditioned on $\{\pi^{k,l}\}_{k\in[K]}$, which together form $\{(s_h^{(m)}, a_{i,h}^{(m)})\}_{m=1}^{|\mathcal{D}_i|}$. Recall the hybrid sampling procedure in Algorithm 1 for any round $t \in [T]$, time step $h \in [H]$, and agent i: for each $j \in [h-1]$,

$$
s_{j+1} \sim \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_{i,j} \sim \pi_{i,j}^{k,t}(s_j)} P_{i,j,s_j,a_{i,j}}^{\pi_{-i,j}^{k,t}, \underline{V}}(\cdot), \qquad (58)
$$

where $P_{i,j,s_j,a_{i,j}}^{\pi_{-i,j}^{k,t}, \underline{V}}$ denotes the pessimistic worst-case transition kernel defined in (74). Accordingly, we denote the resulting distribution of s_h by

$$
s_h \sim \left[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,1}^{\pi_1^{k,t}, \underline{V}} \right] \cdots \left[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,h-1}^{\pi_{h-1}^{k,t}, \underline{V}} \right].
$$

For all $(t,h) \in [T] \times [H]$, we define

$$
X_{t,h} = \mathbb{E}_{s_h, a_i} \big[\phi_i(s_h,a_i) \phi_i(s_h,a_i)^\top \,\big|\, s_h \sim \left[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,1}^{\pi_1^{k,t}, \underline{V}} \right] \cdots
$$

$$\cdots \Big[\sum_{k=1}^K \tfrac{1}{K} P_{i,h-1}^{\pi_{h-1}^{k,t},\underline{V}}\Big],\ a_i \sim \mathrm{Unif}(\mathcal{A}_i)\big].$$

Therefore, denoting the round index of any m-th sample as $\tau(m) \in [t-1]$, one has

$$\mathbb{E}\big[\phi_i^{(m)}(\phi_i^{(m)})^\top\big] = X_{\tau(m),h}$$

so that

$$\bar{\Lambda} := \mathbb{E}[\Lambda_{i,h}^{k,t}] = \textstyle\sum_{m=1}^{|\mathcal{D}_i|} X_{\tau(m),h} + \lambda I_d \qquad \text{and} \qquad \bar{\Lambda} \succeq \lambda I_d. \tag{59}$$

Then we introduce the following lemma whose proof is postponed to Appendix F.1.4.

Lemma 8. *Let $\lambda \geq c_0 \log(2dnKHT/\delta)$ for some large enough constant c_0. With probability at least $1-\delta$, for all $(k,t,h,i) \in [K]\times[T]\times[H]\times[n]$, one has*

$$\tfrac{1}{2}\,\bar{\Lambda} \ \preceq\ \Lambda_{i,h}^{k,t} \ \preceq\ \tfrac{3}{2}\,\bar{\Lambda}. \tag{60}$$

With the above lemma in hand, we introduce an auxiliary term

$$S_{t,h} = \textstyle\sum_{\tau=1}^{t-1} X_{\tau,h} + \lambda I_d$$

As each round $l \in [t-1]$ contributes at least one sample to $\mathcal{D}_i$, we have $\sum_{m=1}^{|\mathcal{D}_i|} X_{\tau(m),h} \succeq \sum_{l=1}^{t-1} X_{l,h}$, since each population covariance $X_{l,h}$ is positive semi-definite. Therefore, one has

$$\bar{\Lambda} \succeq S_{t,h} \succeq \lambda I_d, \tag{61}$$

and consequently

$$\forall \phi \in \mathbb{R}^d: \quad \bar{\Lambda}^{-1} \preceq S_{t,h}^{-1}, \qquad \big\|\phi\big\|_{\bar{\Lambda}^{-1}} \leq \big\|\phi\big\|_{S_{t,h}^{-1}}. \tag{62}$$

Inverting (60) in Lemma 8 yields the operator inequality $(\Lambda_{i,h}^{k,t})^{-1} \preceq 2\bar{\Lambda}^{-1}$. Combining with (62) gives that with probability at least $1-\delta$, simultaneously for all $(s,k,t,h,i) \in \mathcal{S}\times[K]\times[T]\times[H]\times[n]$,

$$\max_{a_i\in\mathcal{A}_i} \big\|\phi_i(s,a_i)\big\|_{\left(\Lambda_{i,h}^{k,t}\right)^{-1}} \lesssim \max_{a_i\in\mathcal{A}_i} \big\|\phi_i(s,a_i)\big\|_{\left(\mathbb{E}\Lambda_{i,h}^{k,t}\right)^{-1}} \leq \max_{a_i\in\mathcal{A}_i} \big\|\phi_i(s,a_i)\big\|_{S_{t,h}^{-1}}. \tag{63}$$

Inserting (63) back into (57) gives

$$\begin{aligned}
&V_{i,h}^{\star,\xi^t}(s) - V_{i,h}^{\xi^t}(s) \leq \overline{V}_{i,h}^t(s) - \underline{V}_{i,h}^t(s) \\
\leq& \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i\sim\pi_{i,h}^{k,t}(s)}\big[P_{i,h,s,a_i}^{\pi_{-i}^{k,t},\underline{V}}\big(\overline{V}_{i,h+1}^t - \underline{V}_{i,h+1}^t\big)\big] + \beta_{i,h,1}^t(s) + \beta_{i,h,2}^t(s) \\
&+ \textstyle\sum_{k=1}^K \frac{1}{K} \max_{a_i\in\mathcal{A}_i} \big\|\phi_i(s,a_i)\big\|_{S_{t,h}^{-1}} \big(4H\sqrt{d\ln(NH+1) + \ln\big(\frac{3NTHnK}{\delta}\big)} + 2H\sqrt{d\lambda}\big) + \frac{1}{N} \\
\leq& \tfrac{1}{K}\textstyle\sum_{k=1}^K \mathbb{E}_{a_i\sim\pi_{i,h}^{k,t}(s)} P_{i,h,s,a_i}^{\pi_{-i}^{k,t},\underline{V}}\big(\overline{V}_{i,h+1}^t - \underline{V}_{i,h+1}^t\big) + cH\sqrt{\tfrac{\ln A_i}{K}} + \tfrac{c}{N} \\
&\quad + c\textstyle\sum_{k=1}^K \frac{1}{K} \max_{a_i\in\mathcal{A}_i} \big\|\phi_i(s,a_i)\big\|_{S_{t,h}^{-1}} \cdot H\big(\sqrt{2d\ln(NH+1) + \ln(3TNHnK/\delta)} + \sqrt{d\lambda}\big),
\end{aligned} \tag{64}$$

which holds with probability at least $1-\delta$. Here the last inequality holds by invoking the definitions of $\beta_{i,h,1}^t(s), \beta_{i,h,2}^t(s)$ in (19) and (20).

Step 3: recursion arguments. Consequently, applying (64) recursively across layers $h = 1,\ldots,H$ arrives at

$$\begin{aligned}
&V_{i,1}^{\star,\xi^t}(s_1) - V_{i,1}^{\xi^t}(s_1) \leq \overline{V}_{i,1}^t(s_1) - \underline{V}_{i,1}^t(s_1) \\
\leq& \textstyle\sum_{h=1}^H c\Big[H\sqrt{\frac{\ln A_i}{K}} + \frac{1}{N} \\
&\quad + \mathbb{E}\big[\|\phi_i(s_h,a_i)\|_{S_{t,h}^{-1}} \,\big|\, s_h \sim \big[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,1}^{\pi_1^{k,t},\underline{V}}\big]\cdots\big[\sum_{k=1}^K \frac{1}{K} P_{i,h-1}^{\pi_{h-1}^{k,t},\underline{V}}\big],\ a_i\sim\mathrm{Unif}(\mathcal{A}_i)\big] \\
&\quad \cdot A_i H\big(\sqrt{2d\ln(NH+1) + \ln\big(\tfrac{3TNHnK}{\delta}\big)} + \sqrt{d\lambda}\big)\Big]
\end{aligned} \tag{65}$$

with probability at least $1-\delta$. Therefore, we can control the regret by applying (65) for all $t \in [T]$: For brevity, denote the conditional expectation

$$\mathcal{E}_{t,h} \coloneqq \mathbb{E}\big[\|\phi_i(s_h,a_i)\|_{S_{t,h}^{-1}} \,\big|\, s_h \sim \big[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,1}^{\pi_1^{k,t},\underline{V}}\big] \cdots \\ \cdots \big[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,h-1}^{\pi_{h-1}^{k,t},\underline{V}}\big],\ a_i \sim \mathrm{Unif}(\mathcal{A}_i)\big].$$

Then

$$\begin{aligned}
&\textstyle\sum_{t=1}^T V_{i,1}^{\star,\xi^t}(s_1) - V_{i,1}^{\xi^t}(s_1) \\
&\le \textstyle\sum_{h=1}^H c\Big[\sum_{t=1}^T \mathcal{E}_{t,h} \cdot A_i H\big(\sqrt{2d\ln(NH+1) + \ln\big(\frac{3TNHnK}{\delta}\big)} + \sqrt{d\lambda}\big)\Big] \\
&\qquad + \textstyle\sum_{h=1}^H \sum_{t=1}^T c\big(\sqrt{\frac{\ln A_i}{K}} + \frac{1}{N}\big) \\
&\le \textstyle\sum_{h=1}^H c\sqrt{T\sum_{t=1}^T \mathcal{E}_{t,h}^2} \cdot A_i H\big(\sqrt{2d\ln(NH+1) + \ln\big(\frac{3TNHnK}{\delta}\big)} + \sqrt{d\lambda}\big) \\
&\qquad + \textstyle\sum_{h=1}^H \sum_{t=1}^T c\big(\sqrt{\frac{\ln A_i}{K}} + \frac{1}{N}\big) \\
&\le \textstyle\sum_{h=1}^H c\sqrt{T\sum_{t=1}^T \mathsf{tr}\big(S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}\big)} \cdot A_i H\big(\sqrt{2d\ln(NH+1) + \ln\big(\frac{3TNHnK}{\delta}\big)} + \sqrt{d\lambda}\big) \\
&\qquad + \textstyle\sum_{h=1}^H \sum_{t=1}^T c\big(\sqrt{\frac{\ln A_i}{K}} + \frac{1}{N}\big) \qquad (66)
\end{aligned}$$

where the second inequality holds by the Cauchy–Schwarz inequality applied to the sum over t (i.e., $\sum_{t=1}^T x_t \le \sqrt{T\sum_{t=1}^T x_t^2}$ with $x_t = \mathbb{E}[\|\phi_i(s,a_i)\|_{S_{t,h}^{-1}} \mid \cdots]$, after also using Jensen's inequality, which yields $(\mathbb{E}[\|\phi_i(s,a_i)\|_{S_{t,h}^{-1}}])^2 \le \mathbb{E}[\|\phi_i(s,a_i)\|_{S_{t,h}^{-1}}^2]$).

The final inequality holds by the cyclic property of the trace: for any vector $\phi \in \mathbb{R}^d$ and any positive definite matrix $M \succ 0$,

$$\|\phi\|_{M^{-1}}^2 = \phi^\top M^{-1}\phi = \mathsf{tr}\big(M^{-1}\phi\phi^\top\big).$$

Let $M = S_{t,h}$ and noting that $S_{t,h} = \sum_{\tau=1}^{t-1} X_{\tau,h} + \lambda I_d$ depends only on iterations $\tau < t$ and is therefore measurable with respect to (and constant given) the conditioning event $\{s_h \sim [\sum_k \frac{1}{K} P_{i,1}^{\pi_1^{k,t},\underline{V}}] \cdots [\sum_k \frac{1}{K} P_{i,h-1}^{\pi_{h-1}^{k,t},\underline{V}}],\ a_i \sim \mathrm{Unif}(\mathcal{A}_i)\}$ used at iteration t, we may interchange the conditional expectation with the trace by linearity:

$$\begin{aligned}
&\mathbb{E}\big[\|\phi_i(s,a_i)\|_{S_{t,h}^{-1}}^2 \,\big|\, s_h \sim \big[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,1}^{\pi_1^{k,t},\underline{V}}\big] \cdots \big[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,h-1}^{\pi_{h-1}^{k,t},\underline{V}}\big],\ a_i \sim \mathrm{Unif}(\mathcal{A}_i)\big] \\
&= \mathsf{tr}\Big(S_{t,h}^{-1}\,\mathbb{E}\big[\phi_i(s_h,a_i)\phi_i(s_h,a_i)^\top \,\big|\, s_h \sim \big[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,1}^{\pi_1^{k,t},\underline{V}}\big] \cdots \\
&\qquad \cdots \big[\textstyle\sum_{k=1}^K \frac{1}{K} P_{i,h-1}^{\pi_{h-1}^{k,t},\underline{V}}\big],\ a_i \sim \mathrm{Unif}(\mathcal{A}_i)\big]\Big) \\
&= \mathsf{tr}\big(S_{t,h}^{-1} X_{t,h}\big) = \mathsf{tr}\big(S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}\big)
\end{aligned}$$

where the last line holds by invoking the definition of $X_{t,h}$ and the cyclic invariance of the trace.

We introduce the following lemme that controls $\sum_{t=1}^T \mathsf{tr}\big(S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}\big)$. The proof is postponed to Appendix F.1.5.

Lemma 9.

$$\textstyle\sum_{t=1}^T \mathsf{tr}\Big(S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}\Big) = 2d\log\big(1+\frac{T}{d\lambda}\big) = \widetilde{O}(d). \qquad (67)$$

Substituting this lemma into (66) gives

$$\begin{aligned}
&\textstyle\sum_{t=1}^T V_{i,1}^{\star,\xi^t}(s_1) - V_{i,1}^{\xi^t}(s_1) \\
&\le \textstyle\sum_{h=1}^H c\sqrt{T\sum_{t=1}^T \mathsf{tr}\Big(S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}\Big)} \cdot A_i H \cdot \Big(\sqrt{2d\ln(NH+1) + \ln(\tfrac{3TNHnK}{\delta})} + \sqrt{d\lambda}\Big)
\end{aligned}$$

$$
\begin{aligned}
&\quad + \textstyle\sum_{h=1}^{H}\sum_{t=1}^{T} c\big(\sqrt{\frac{\ln A_i}{K}} + \frac{1}{N}\big) \\
&\le \textstyle\sum_{h=1}^{H} c\widetilde{\mathcal{O}}(\sqrt{Td})\cdot A_i H \cdot \Big(\sqrt{2d\ln(NH+1) + \ln(\frac{3TNHnK}{\delta})} + \sqrt{d\lambda}\Big) \\
&\quad + \textstyle\sum_{h=1}^{H}\sum_{t=1}^{T} c\big(\sqrt{\frac{\ln A_i}{K}} + \frac{1}{N}\big) \\
&\overset{\text{(i)}}{\le} \textstyle\sum_{h=1}^{H} c\widetilde{\mathcal{O}}(\sqrt{Td})\cdot A_i H \cdot \widetilde{\mathcal{O}}(\sqrt{d}) + \sum_{h=1}^{H}\sum_{t=1}^{T} c\big(\sqrt{\frac{\ln A_i}{K}} + \frac{1}{N}\big) \\
&\overset{\text{(ii)}}{\le} \widetilde{\mathcal{O}}\big(A_i H^2 d\sqrt{T}\big) + \widetilde{\mathcal{O}}(H\sqrt{T\ln A_i} + H) \overset{\text{(iii)}}{\le} \widetilde{\mathcal{O}}\big(A_i H^2 d\sqrt{T}\big)
\end{aligned}
$$

where (i) bounds $\sqrt{2d\ln(NH+1) + \ln(\frac{3TNHnK}{\delta})} + \sqrt{d\lambda}$ by $\widetilde{\mathcal{O}}(\sqrt{d})$ due to the requirement of $\lambda \ge c_0 \log(2dnKHT/\delta) = \widetilde{\mathcal{O}}(1)$ in Lemma 8, (ii) holds by setting $T = N = K$, and (iii) drops the residual because it is dominated by the leading $\widetilde{\mathcal{O}}(A_i H^2 d\sqrt{T})$.

Finally, we conclude that with probability at least $1-\delta$, Algorithm 3 guarantees that the regret is bounded by

$$
\mathrm{Regret}(T) = \max_{i\in[n]} \textstyle\sum_{t=1}^{T} \mathbb{E}_{\pi\sim\xi^t}\big[V_{i,1}^{\star,\pi_{-i}}(s_1)\big] - \mathbb{E}_{\pi\sim\xi^t}\big[V_{i,1}^{\pi}(s_1)\big] \le \widetilde{\mathcal{O}}\big(dH^2\sqrt{T}\cdot \max_{i\in[n]} A_i\big). \tag{68}
$$

Thus, we have finished the proof of the regret bound in Theorem 3.

Proof of the ε-CCE guarantee. It remains to show that, when $T > H^4 d^2 (\max_{i\in[n]} A_i)^2/\varepsilon^2$, Algorithm 3 outputs an ε-robust CCE specialized to any fixed initial state s_1 (cf. (8)). Define the uniform mixture of the output policy distributions of Algorithm 3 over the T iterates as

$$
\widehat{\xi} := \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \xi^t. \tag{69}
$$

For any fixed agent $i \in [n]$, by linearity of expectation, one has

$$
\begin{aligned}
&\mathbb{E}_{\pi\sim\widehat{\xi}}\big[V_{i,1}^{\star,\pi_{-i},\sigma_i}(s_1)\big] - \mathbb{E}_{\pi\sim\widehat{\xi}}\big[V_{i,1}^{\pi,\sigma_i}(s_1)\big] \\
&= \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\Big(\mathbb{E}_{\pi\sim\xi^t}\big[V_{i,1}^{\star,\pi_{-i},\sigma_i}(s_1)\big] - \mathbb{E}_{\pi\sim\xi^t}\big[V_{i,1}^{\pi,\sigma_i}(s_1)\big]\Big) \\
&\le \tfrac{1}{T}\max_{j\in[n]}\textstyle\sum_{t=1}^{T}\Big(\mathbb{E}_{\pi\sim\xi^t}\big[V_{j,1}^{\star,\pi_{-j},\sigma_j}(s_1)\big] - \mathbb{E}_{\pi\sim\xi^t}\big[V_{j,1}^{\pi,\sigma_j}(s_1)\big]\Big) \;=\; \tfrac{1}{T}\,\mathrm{Regret}(T),
\end{aligned}
$$

where the last equality follows from the definition of $\mathrm{Regret}(T)$ in (9). Plugging in the regret bound in (68) gives

$$
\max_{i\in[n]}\Big\{\mathbb{E}_{\pi\sim\widehat{\xi}}\big[V_{i,1}^{\star,\pi_{-i},\sigma_i}(s_1)\big] - \mathbb{E}_{\pi\sim\widehat{\xi}}\big[V_{i,1}^{\pi,\sigma_i}(s_1)\big]\Big\} \le \tfrac{\mathrm{Regret}(T)}{T} \le \widetilde{\mathcal{O}}\bigg(\frac{dH^2\,\max_{i\in[n]} A_i}{\sqrt{T}}\bigg) \le \varepsilon
$$

with probability at least $1-\delta$, as long as

$$
T \;>\; \tfrac{H^4 d^2\,(\max_{i\in[n]} A_i)^2}{\varepsilon^2},
$$

This is exactly the ε-robust CCE condition in (8), which completes the proof of Theorem 3.

F.1 Proof of auxiliary results.

F.1.1 Proof of Lemma 5

The proof follows the pipeline of that for Lemma 2 in Appendix E.3.1. Throughout the proof, $V : \mathcal{S} \to [0, H]$ is a fixed value function.

Step 1: Proof of the transition estimation gap. First, consider any fixed $(h, i, k, t) \in [H]\times[n]\times[K]\times[T]$. Recalling the auxiliary defintions in (31), and then applying strong-duality formulation [30] for the dual form of the robust Bellman operator [65, Lemma 1] yields

$$
\big|P_{i,h,s,a_i}^{\widehat{\pi}_{-i}^{k,t},V}V - \widehat{P}_{i,h,s,a_i}^{\widehat{\pi}_{-i}^{k,t},V}V\big| \le \max_{\alpha\in[\min_s V(s),\,\max_s V(s)]}\big|P_{h,s,a_i}^{\widehat{\pi}_{-i}^{k,t}}[V]_\alpha - P_{i,h,s,a_i}^{k,t}[V]_\alpha\big|,
$$

where $P^{k,t}_{i,h,s,a_i}(\cdot) \coloneqq \phi_i(s,a_i)^\top \widehat{\mu}^{k,t}_{i,h}(\cdot)$ is the estimated transition kernel defined in (16).

For any fixed α, the linear-MG assumption in (2) gives $P^{\widehat{\pi}^{k,t}_{-i}}_{h,s,a_i}[V]_\alpha = \phi_i(s,a_i)^\top \mu^{\widehat{\pi}^{k,t}_{-i}}_{i,h}[V]_\alpha$. Therefore, applying Theorem 5 to the online Gram matrix $\Lambda^{k,t}_{i,h}$ defined in (16) with $f = [V]_\alpha \in [0,H]$ yields that with probability at least $1-\delta$,

$$\big|P^{\widehat{\pi}^{k,t}_{-i}}_{h,s,a_i}[V]_\alpha - P^{k,t}_{i,h,s,a_i}[V]_\alpha\big| \le \big\|\phi_i(s,a_i)\big\|_{(\Lambda^{k,t}_{i,h})^{-1}} \big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3NH}{\delta}\big)} + H\sqrt{d\lambda}\big), \tag{70}$$

for all $(s,a_i) \in \mathcal{S}\times\mathcal{A}_i$.

To derive the union bound over $\alpha \in [0,H]$, we apply the same ε_1-net argument as in (52): for $\varepsilon_1 = 1/N$, we have $|N_{\varepsilon_1}| \le 3HN$. Observing the fact that the function $\alpha \mapsto |P^{\widehat{\pi}^{k,t}_{-i}}_{h,s,a_i}[V]_\alpha - P^{k,t}_{i,h,s,a_i}[V]_\alpha|$ is 1-Lipschitz, combined with the pointwise bound in (70) gives, with probability at least $1-\delta$, for all (s,a_i) and $\alpha \in [0,H]$,

$$\begin{aligned}&\big|P^{\widehat{\pi}^{k,t}_{-i}}_{h,s,a_i}[V]_\alpha - P^{k,t}_{i,h,s,a_i}[V]_\alpha\big| \\ &\quad\le \big\|\phi_i(s,a_i)\big\|_{(\Lambda^{k,t}_{i,h})^{-1}} \big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3NH\cdot 3HN}{\delta}\big)} + H\sqrt{d\lambda}\big) + \tfrac{1}{N}.\end{aligned}$$

With above pointwise bound for any fixed $(h,i,k,t) \in [H]\times[n]\times[K]\times[T]$, we replace δ by $\delta/(nKHT)$ to derive the union bound over $[H]\times[n]\times[K]\times[T]$: with probability at least $1-\delta$, for all (h,i,k,t) and all $(s,a_i)\in\mathcal{S}\times\mathcal{A}_i$,

$$\begin{aligned}&\big|P^{\widehat{\pi}^{k,t}_{-i},V}_{i,h,s,a_i}V - \widehat{P}^{\widehat{\pi}^{k,t}_{-i},V}_{i,h,s,a_i}V\big| \\ &\quad\le \big\|\phi_i(s,a_i)\big\|_{(\Lambda^{k,t}_{i,h})^{-1}} \big(2H\sqrt{d\ln(NH+1)+2\ln\big(\tfrac{3TKNHn}{\delta}\big)} + H\sqrt{d\lambda}\big) + \tfrac{1}{N}.\end{aligned}$$

Step 2: Proof of the reward estimation gap. The argument is identical to the generative case, except: (i) the Gram matrix is now $\Lambda^{k,t}_{i,h}$ from (16) with regularization $\lambda \ge 1$, so the ridge-regression bias term becomes $H\sqrt{d\lambda}$ (rather than $H\sqrt{d}$ in the generative case due to the $\lambda = 1$ specialization); (ii) we retain $\|\phi_i(s,a_i)\|_{(\Lambda^{k,t}_{i,h})^{-1}}$ explicitly; (iii) the union bound is over $(h,i,k,t)\in[H]\times[n]\times[K]\times[T]$, contributing an extra $\ln T$ factor inside the logarithmic term. Concretely, applying Theorem 5 to the scalar target $f = \mathbb{E}_{a_{-i}\sim\widehat{\pi}^{k,t}_{-i}(s)}[r_{i,h}(s,\mathbf{a})] \in [0,H]$ with the online Gram matrix and replacing δ by $\delta/(nKHT)$ in the resulting union bound, we have with probability at least $1-\delta$,

$$\begin{aligned}&\big|\mathbb{E}_{a_{-i}\sim\widehat{\pi}^{k,t}_{-i}(s)}[r_{i,h}(s,\mathbf{a})] - r^{k,t}_{i,h}(s,a_i)\big| \\ &\quad\le \big\|\phi_i(s,a_i)\big\|_{(\Lambda^{k,t}_{i,h})^{-1}} \big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3NTHnK}{\delta}\big)} + H\sqrt{d\lambda}\big),\end{aligned}$$

holds simultaneously for all (h,i,k,t) and (s,a_i).

F.1.2 Proof of Lemma 6

We will prove Lemma 6 through induction arguments. For the base case $h = H+1$, by definition,

$$\overline{V}^t_{i,H+1}(s) = V^{\star,\xi^t}_{i,H+1}(s) = 0 \tag{71}$$

for all $s\in\mathcal{S}$. Thus, assuming that the desired fact is satisfied at time step $h+1$, namely $\overline{V}^t_{i,h+1}(s) \ge V^{\star,\xi^t}_{i,h+1}(s)$ holds for all $s\in\mathcal{S}$, we will prove $\overline{V}^t_{i,h}(s) \ge V^{\star,\xi^t}_{i,h}(s)$ for all $s\in\mathcal{S}$.

Towards this, we first define several auxiliary notations as follows. For all agents $i\in[n]$ and $(t,k,s,h,a_i)\in[T]\times[K]\times\mathcal{S}\times[H]\times\mathcal{A}_i$, let $P^{k,t}_{i,h,s,a_i}(\cdot) \coloneqq \phi_i(s,a_i)^\top\widehat{\mu}^{k,t}_{i,h}(\cdot)$ denote the online estimated transition kernel from (16). Define the worst-case kernels under the true and the estimated nominal transitions, respectively, as

$$P^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i} \coloneqq \arg\min_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P^{\pi^{k,t}_{-i}}_{h,s,a_i}\right)} \mathcal{P}\overline{V}^t_{i,h+1},$$

$$\widehat{P}^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i} := \arg\min_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P^{k,t}_{i,h,s,a_i}\right)} \mathcal{P}\overline{V}^t_{i,h+1}.$$

To continue, let ξ^t be the uniform distribution over the K product policies $\{\pi^{k,t}\}_{k\in[K]}$ output by Algorithm 3 at iteration t, where each $\pi^{k,t} = \{\pi^{k,t}_h\}_{h\in[H]}$ with $\pi^{k,t}_h : \mathcal{S} \to \prod_{i\in[n]} \Delta(\mathcal{A}_i)$. Recalling the definition of the expected best-response value function $V^{\star,\xi^t}_{i,h}$ in (30) and (6), by the robust Bellman equation (23), we have

$$\begin{aligned}
V^{\star,\xi^t}_{i,h}(s) &= \max_{a_i\in\mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\Big[\mathbb{E}_{a_{-i}\sim\pi^{k,t}_{-i,h}(s)}\big[r_{i,h}(s,\mathbf{a})\big] + \inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P^{\pi^{k,t}_{-i}}_{h,s,a_i}\right)} \mathcal{P}V^{\star,\xi^t}_{i,h+1}\Big] \\
&\overset{\text{(i)}}{\le} \max_{a_i\in\mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\Big[\mathbb{E}_{a_{-i}\sim\pi^{k,t}_{-i,h}(s)}\big[r_{i,h}(s,\mathbf{a})\big] + \inf_{\mathcal{P}\in\mathcal{U}^{\sigma_i}\left(P^{\pi^{k,t}_{-i}}_{h,s,a_i}\right)} \mathcal{P}\overline{V}^t_{i,h+1}\Big] \\
&= \max_{a_i\in\mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\Big[\mathbb{E}_{a_{-i}\sim\pi^{k,t}_{-i,h}(s)}\big[r_{i,h}(s,\mathbf{a})\big] + P^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1}\Big] \\
&\le \max_{a_i\in\mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\Big[r^{k,t}_{i,h}(s,a_i) + \widehat{P}^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1} + \big|r^{k,t}_{i,h}(s,a_i) - \mathbb{E}_{a_{-i}\sim\pi^{k,t}_{-i,h}}[r_{i,h}(s,\mathbf{a})]\big| \\
&\qquad + \big|\widehat{P}^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1} - P^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1}\big|\Big],
\end{aligned}$$

where (i) holds by the induction assumption $V^{\star,\xi^t}_{i,h+1}(s) \le \overline{V}^t_{i,h+1}(s)$ for all $s\in\mathcal{S}$. In the next step, we apply a similar analysis as in Lemma 2, from which we obtain

$$\begin{aligned}
&V^{\star,\xi^t}_{i,h}(s) \\
&\le \max_{a_i\in\mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\Big[r^{k,t}_{i,h}(s,a_i) + \widehat{P}^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1} + \big|r^{k,t}_{i,h}(s,a_i) - \mathbb{E}_{a_{-i}\sim\pi^{k,t}_{-i,h}}[r_{i,h}(s,\mathbf{a})]\big| \\
&\qquad + \big|\widehat{P}^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1} - P^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1}\big|\Big]. \\
&\overset{\text{(i)}}{\le} \max_{a_i\in\mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\big\|\phi(s,a_i)\big\|_{\left(\Lambda^{k,t}_{i,h}\right)^{-1}}\Big(2H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3TNHnK}{\delta}\big)} + H\sqrt{d}\Big) + \tfrac{1}{N} \\
&\quad + \max_{a_i\in\mathcal{A}_i} \sum_{k=1}^K \big[r^{k,t}_{i,h}(s,a_i) + \widehat{P}^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1}\big] \\
&\overset{\text{(ii)}}{\le} \max_{a_i\in\mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\big\|\phi(s,a_i)\big\|_{\left(\Lambda^{k,t}_{i,h}\right)^{-1}}\Big(4H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3TNHnK}{\delta}\big)} + 2H\sqrt{d}\Big) \\
&\quad + \tfrac{1}{N} + 2H\sqrt{\tfrac{\ln A_i}{K}} + \sum_{k=1}^K \tfrac{1}{K}\mathbb{E}_{a_i\sim\pi^{k,t}_{i,h}(s)}\big[r^{k,t}_{i,h}(s,a_i) + \widehat{P}^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1}\big] \\
&\quad = \beta^t_{i,h,1}(s) + \sum_{k=1}^K \tfrac{1}{K}\mathbb{E}_{a_i\sim\pi^{k,t}_{i,h}(s)}\big[r^{k,t}_{i,h}(s,a_i) + \widehat{P}^{\pi^{k,t}_{-i},\overline{V}}_{i,h,s,a_i}\overline{V}^t_{i,h+1}\big]. \qquad (72)
\end{aligned}$$

Here, (i) follows from Lemma 5, (ii) holds by the bound for standard FTRL in Theorem 5, and th final line follows from the definition of $\beta^t_{i,h,1}(s)$ in (19) as follows:

$$\begin{aligned}
\beta^t_{i,h,1}(s) &= \max_{a_i\in\mathcal{A}_i} \sum_{k=1}^K \tfrac{1}{K}\big\|\phi(s,a_i)\big\|_{\left(\Lambda^{k,t}_{i,h}\right)^{-1}}\Big(4H\sqrt{d\ln(NH+1)+\ln\big(\tfrac{3TNHnK}{\delta}\big)} + 2H\sqrt{d}\Big) \\
&\quad + \tfrac{1}{N} + 2H\sqrt{\tfrac{\ln A_i}{K}}.
\end{aligned}$$

Finally, recalling the definition of the optimistic estimation $\overline{V}^t_{i,h}(s)$ in (18)

$$\overline{V}^t_{i,h}(s) = \min\Big\{\sum_{k=1}^K \tfrac{1}{K}\big\langle\pi^{k,t}_{i,h}(\cdot\mid s), \overline{q}^{k,t}_{i,h}(s,\cdot)\big\rangle + \beta^t_{i,h,1}(s),\ H-h+1\Big\}, \qquad (73)$$

combining(72) with the elementary fact that $V^{\star,\xi^t}_{i,h}(s) \le H-h+1$ gives that

$$\forall(h,i,t,s)\in[H]\times[n]\times[T]\times\mathcal{S}: \quad V^{\star,\xi^t}_{i,h}(s) \le \overline{V}^t_{i,h}(s)$$

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

F.1.3 Proof of Lemma 7

We will prove Lemma 7 through induction arguments. For the base case $h = H + 1$, by definition,

$$\forall (i, s) \in [n] \times \mathcal{S}: \quad \underline{V}^t_{i,H+1}(s) = 0 = V^{\xi^t}_{i,H+1}(s).$$

Thus, assuming that the desired fact is satisfied at time step $h+1$, namely $\underline{V}^t_{i,h+1}(s) \le V^{\xi^t}_{i,h+1}(s)$ holds for all $s \in \mathcal{S}$, we will prove $\underline{V}^t_{i,h}(s) \le V^{\xi^t}_{i,h}(s)$ for all $s \in \mathcal{S}$.

Towards this, we first define several auxiliary notations as follows. For all agents $i \in [n]$ and $(t, k, s, h, a_i) \in [T] \times [K] \times \mathcal{S} \times [H] \times \mathcal{A}_i$, with $P^{k,t}_{i,h,s,a_i}$ the online estimated transition kernel from (16), we define the worst-case kernels (for the pessimistic value) under the true and the estimated nominal transitions, respectively, as

$$\begin{aligned} P^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} &:= \arg\min_{\mathcal{P} \in \mathcal{U}^{\sigma_i}\left(P^{\pi^{k,t}_{-i}}_{h,s,a_i}\right)} \mathcal{P}\underline{V}^t_{i,h+1}, \\ \widehat{P}^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} &:= \arg\min_{\mathcal{P} \in \mathcal{U}^{\sigma_i}\left(P^{k,t}_{i,h,s,a_i}\right)} \mathcal{P}\underline{V}^t_{i,h+1}. \end{aligned} \tag{74}$$

To continue, recall that ξ^t is the uniform distribution over product policies $\{\pi^{k,t}_h\}_{(h,k)\in[H]\times[K]}$ output by Algorithm 3. Recalling the definition of the expected version of the value function $V^{\xi^t}_{i,h}$ in (30), and referring to the robust Bellman equation in (23), we have

$$\begin{aligned} V^{\xi^t}_{i,h}(s) &= \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big[\mathbb{E}_{a_{-i} \sim \pi^{k,t}_{-i,h}(s)} \big[r_{i,h}(s, \mathbf{a}) \big] + \inf_{\mathcal{P} \in \mathcal{U}^{\sigma_i}\left(P^{\pi^{k,t}_{-i}}_{h,s,a_i}\right)} \mathcal{P} V^{\xi^t}_{i,h+1} \big] \\ &\overset{\text{(i)}}{\ge} \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big[\mathbb{E}_{a_{-i} \sim \pi^{k,t}_{-i,h}(s)} \big[r_{i,h}(s, \mathbf{a}) \big] + \inf_{\mathcal{P} \in \mathcal{U}^{\sigma_i}\left(P^{\pi^{k,t}_{-i}}_{h,s,a_i}\right)} \mathcal{P} \underline{V}^t_{i,h+1} \big] \\ &\overset{\text{(ii)}}{=} \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big[\mathbb{E}_{a_{-i} \sim \pi^{k,t}_{-i,h}(s)} \big[r_{i,h}(s, \mathbf{a}) \big] + P^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1} \big] \\ &\overset{\text{(iii)}}{\ge} \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big[r^{k,t}_{i,h}(s, a_i) + \widehat{P}^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1} \big] \\ &\quad - \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big| r^{k,t}_{i,h}(s, a_i) - \mathbb{E}_{a_{-i} \sim \pi^{k,t}_{-i,h}(s)} \big[r_{i,h}(s, \mathbf{a}) \big] \big| \\ &\quad - \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big| \widehat{P}^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1} - P^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1} \big|, \end{aligned} \tag{75}$$

where (i) holds by the induction hypothesis $\underline{V}^t_{i,h+1}(s) \le V^{\xi^t}_{i,h+1}(s)$ for all $s \in \mathcal{S}$, (ii) uses the notations defined in (74), and (iii) follows from the elementary inequality $x \ge y - |x - y|$ applied separately to the reward term $\mathbb{E}_{a_{-i} \sim \pi^{k,t}_{-i,h}(s)} \big[r_{i,h}(s, \mathbf{a}) \big]$ against its linear estimate $r^{k,t}_{i,h}(s, a_i)$ and to the transition kernel term $P^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1}$ against its empirical counterpart $\widehat{P}^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1}$.

In the next step, we apply Lemma 5, which simultaneously controls the reward and transition estimation errors, from which we obtain that, with probability at least $1 - \delta$:

$$\begin{aligned} V^{\xi^t}_{i,h}(s) &\ge \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big[r^{k,t}_{i,h}(s, a_i) + \widehat{P}^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1} \big] \\ &\quad - \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big| r^{k,t}_{i,h}(s, a_i) - \mathbb{E}_{a_{-i} \sim \pi^{k,t}_{-i,h}(s)} \big[r_{i,h}(s, \mathbf{a}) \big] \big| \\ &\quad - \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big| \widehat{P}^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1} - P^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1} \big| \\ &\overset{\text{(i)}}{\ge} \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big[r^{k,t}_{i,h}(s, a_i) + \widehat{P}^{\pi^{k,t}_{-i}, \underline{V}}_{i,h,s,a_i} \underline{V}^t_{i,h+1} \big] \\ &\quad - \textstyle\sum_{k=1}^K \frac{1}{K} \mathbb{E}_{a_i \sim \pi^{k,t}_{i,h}(s)} \big\| \phi(s, a_i) \big\|_{\left(\Lambda^{k,t}_{i,h}\right)^{-1}} \\ &\quad \cdot \Big(4H \sqrt{d \ln(NH + 1) + \ln\big(\tfrac{3TNHnK}{\delta}\big)} + 2H\sqrt{d} \Big) - \tfrac{1}{N} \end{aligned}$$

$$= \sum_{k=1}^{K} \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big[r_{i,h}^{k,t}(s, a_i) + \widehat{P}_{i,h,s,a_i}^{\pi_{-i}^{k,t}, \underline{V}} \underline{V}_{i,h+1}^{t} \big] - \beta_{i,h,2}^{t}(s), \tag{76}$$

where (i) follows from Lemma 5, and the final line follows from the definition of the pessimistic bonus $\beta_{i,h,2}^{t}(s)$ in (20), which we recall below:

$$\beta_{i,h,2}^{t}(s) = \sum_{k=1}^{K} \frac{1}{K} \mathbb{E}_{a_i \sim \pi_{i,h}^{k,t}(s)} \big\| \phi(s, a_i) \big\|_{\left(\Lambda_{i,h}^{k,t}\right)^{-1}} \cdot \Big(4H \sqrt{d \ln(NH+1) + \ln\big(\tfrac{3TNHnK}{\delta}\big)} + 2H\sqrt{d} \Big) + \tfrac{1}{N}.$$

Finally, recalling the definition of the pessimistic estimation $\underline{V}_{i,h}^{t}(s)$ in (18)

$$\underline{V}_{i,h}^{t}(s) = \max \Big\{ \sum_{k=1}^{K} \frac{1}{K} \big\langle \pi_{i,h}^{k,t}(\cdot \mid s), \underline{q}_{i,h}^{k,t}(s, \cdot) \big\rangle - \beta_{i,h,2}^{t}(s),\ 0 \Big\}, \tag{77}$$

combining (76) with the elementary fact that $V_{i,h}^{\xi^t}(s) \geq 0$ gives that

$$\forall (h, i, t, s) \in [H] \times [n] \times [T] \times \mathcal{S} : \quad \underline{V}_{i,h}^{t}(s) \leq V_{i,h}^{\xi^t}(s)$$

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

F.1.4 Proof of Lemma 8

we define the centered symmetric matrices

$$W_m \coloneqq \phi_i^{(m)} (\phi_i^{(m)})^\top - X_{\tau(m),h}, \quad \text{so that} \quad \sum_{m=1}^{|\mathcal{D}_i|} W_m = \Lambda_{i,h}^{k,t} - \bar{\Lambda},$$

where each W_m are independent conditional on the past $t-1$ ronuds. Before continuing, note the basic fact

$$\big\| \phi_i^{(m)} \big\|_2 \leq 1 \qquad \text{and} \qquad \| \phi_i^{(m)} (\phi_i^{(m)})^\top \|_{\mathrm{op}} = \| \phi_i^{(m)} \|_2^2 \leq 1. \tag{78}$$

Therefore, we have

$$\| W_m \|_{\mathrm{op}} \leq \big\| \phi_i^{(m)} (\phi_i^{(m)})^\top \big\|_{\mathrm{op}} + \big\| X_{\tau(m),h} \big\|_{\mathrm{op}} \leq 2 \tag{79}$$

where the first inequality follows the triangle inequality, and the last inequality holds by (78) and

$$\| X_{\tau(m),h} \|_{\mathrm{op}} \leq \mathbb{E}[\| \phi_i^{(m)} (\phi_i^{(m)})^\top \|_{\mathrm{op}}] \leq 1, \tag{80}$$

which is achieved by applying Jensen's inequality.

In addition, for the variance term,

$$\mathbb{E}[W_m^2] = \mathbb{E}[(\phi_i^{(m)} (\phi_i^{(m)})^\top)^2] - X_{\tau(m),h}^2 \preceq \mathbb{E}\big[(\phi_i^{(m)} (\phi_i^{(m)})^\top)^2\big] \preceq \mathbb{E}\big[\phi_i^{(m)} (\phi_i^{(m)})^\top\big] = X_{\tau(m),h},$$

where the first inequality holds by the fact that $X_{\tau(m),h}^2 \succeq 0$, and the second inequality holds by $(\phi_i^{(m)} (\phi_i^{(m)})^\top)^2 = \| \phi_i^{(m)} \|_2^2 \, \phi_i^{(m)} (\phi_i^{(m)})^\top \preceq \phi_i^{(m)} (\phi_i^{(m)})^\top$, since $\| \phi_i^{(m)} \|_2 \leq 1$ by (78).

Consequently,

$$V \coloneqq \sum_{m=1}^{|\mathcal{D}_i|} \mathbb{E}[W_m^2] \preceq \sum_{m=1}^{|\mathcal{D}_i|} X_{\tau(m),h} = \bar{\Lambda} - \lambda I_d \preceq \bar{\Lambda}. \tag{81}$$

With above results for W_n and the variance of the sum in hand, we define the conjugated increments

$$Y_m \coloneqq \bar{\Lambda}^{-1/2} W_m \bar{\Lambda}^{-1/2}, \qquad \text{so that} \qquad \sum_{m=1}^{|\mathcal{D}_i|} Y_m = \bar{\Lambda}^{-1/2} \big(\Lambda_{i,h}^{k,t} - \bar{\Lambda} \big) \bar{\Lambda}^{-1/2}.$$

Notice that

$$\| Y_m \|_{\mathrm{op}} \leq \| \bar{\Lambda}^{-1/2} \|_{\mathrm{op}}^2 \, \| W_m \|_{\mathrm{op}} \leq \tfrac{2}{\lambda} \tag{82}$$

where the last inequality holds by (79) and the fact $\| \bar{\Lambda}^{-1/2} \|_{\mathrm{op}}^2 = \| \bar{\Lambda}^{-1} \|_{\mathrm{op}} \leq \lambda^{-1}$ since $\bar{\Lambda} \succeq \lambda I_d$. For the variance term,

$$\sum_{m=1}^{|\mathcal{D}_i|} \mathbb{E}[Y_m^2] = \bar{\Lambda}^{-1/2} \Bigg(\sum_{m=1}^{|\mathcal{D}_i|} \mathbb{E}\big[W_m \bar{\Lambda}^{-1} W_m \big] \Bigg) \bar{\Lambda}^{-1/2}$$

$$\overset{\text{(i)}}{\preceq} \tfrac{1}{\lambda}\,\bar{\Lambda}^{-1/2} V \bar{\Lambda}^{-1/2} \overset{\text{(ii)}}{\preceq} \tfrac{1}{\lambda}\,\bar{\Lambda}^{-1/2}\bar{\Lambda}\bar{\Lambda}^{-1/2} = \tfrac{1}{\lambda} I_d.$$

where (i) holds by $\bar{\Lambda}^{-1} \preceq \lambda^{-1} I_d$, and (ii) follows from (81). So we have

$$\sigma'^2 \coloneqq \Big\| \textstyle\sum_{m=1}^{|\mathcal{D}_i|} \mathbb{E}[Y_m^2] \Big\|_{\mathrm{op}} \le \tfrac{1}{\lambda}. \tag{83}$$

Applying matrix Bernstein [72, Theorem 6.1.1] to the independent, mean-zero, symmetric matrices $\{Y_m\}_{m=1}^{|\mathcal{D}_i|}$ with $R' = \frac{2}{\lambda}$ and $\sigma'^2 = \frac{1}{\lambda}$ as in (82) and (83), we obtain, for any $\theta > 0$,

$$\mathbb{P}\Big[\Big\| \textstyle\sum_{m=1}^{|\mathcal{D}_i|} Y_m \Big\|_{\mathrm{op}} \ge \theta\Big] \le 2d \exp\Big(- \tfrac{\theta^2/2}{\sigma'^2 + R'\theta/3} \Big).$$

Setting the RHS to $\delta' \coloneqq \delta/(nKHT)$ and inverting yields that with probability at least $1 - \delta'$,

$$\begin{aligned} \Big\| \textstyle\sum_{m=1}^{|\mathcal{D}_i|} Y_m \Big\|_{\mathrm{op}} &= \Big\| \bar{\Lambda}^{-1/2} \big(\Lambda_{i,h}^{k,t} - \bar{\Lambda} \big) \bar{\Lambda}^{-1/2} \Big\|_{\mathrm{op}} \le \sqrt{2\sigma'^2 \log(2d/\delta')} + \tfrac{2R' \log(2d/\delta')}{3} \\ &\le \sqrt{\tfrac{2\log(2d/\delta')}{\lambda}} + \tfrac{4\log(2d/\delta')}{3\lambda} \le \tfrac{1}{2} \end{aligned} \tag{84}$$

where the last inequality holds by setting $\lambda \ge c_0 \log(2dnKHT/\delta)$ for a sufficiently large universal constant c_0, so that both terms on the right-hand side of (84) are at most $1/4$.

Regarding (84), we have $-\frac{1}{2} I_d \preceq \bar{\Lambda}^{-1/2}(\Lambda_{i,h}^{k,t} - \bar{\Lambda})\bar{\Lambda}^{-1/2} \preceq \frac{1}{2} I_d$, which directly gives $-\frac{1}{2}\bar{\Lambda} \preceq \Lambda_{i,h}^{k,t} - \bar{\Lambda} \preceq \frac{1}{2}\bar{\Lambda}$, yielding

$$\tfrac{1}{2}\,\bar{\Lambda} \;\preceq\; \Lambda_{i,h}^{k,t} \;\preceq\; \tfrac{3}{2}\,\bar{\Lambda}. \tag{85}$$

F.1.5 Proof of Lemma 9

Denote $\mu_1, \ldots, \mu_d \ge 0$ denote the eigenvalues of the positive definite matrix $S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}$, since $X_{t,h}$ is positive semi-definite. We have

$$\|S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}\|_{\mathrm{op}} \le \|S_{t,h}^{-1}\|_{\mathrm{op}} \|X_{t,h}\|_{\mathrm{op}} \le 1 \tag{86}$$

since the fact $S_{t,h} \succeq \lambda I_d$ in (61) with $\lambda \ge 1$ gives $\|S_{t,h}^{-1}\|_{\mathrm{op}} \le 1$, and $\|X_{t,h}\|_{\mathrm{op}} \le \mathsf{tr}(X_{t,h}) = \mathbb{E}_{s_h,a_i}[\|\phi_i(s_h,a_i)\|_2^2] \le 1$. Therefore, we have $0 \le \mu_i \le 1$ and thus

$$\log(1+\mu_i) \ge \tfrac{\mu_i}{2} \tag{87}$$

due to $\log(1+x) \ge x/2$ for $x \in [0,1]$. Applying the fact to each eigenvalue, we arrive at

$$\mathsf{tr}\big(S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}\big) = \textstyle\sum_{i=1}^d \mu_i \le 2 \sum_{i=1}^d \log\det(1+\mu_i) = 2\log\det(I_d + S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}). \tag{88}$$

Since $S_{t,h} \succ 0$, both $S_{t,h}^{1/2}$ and $S_{t,h}^{-1/2}$ are well-defined. Therefore, we have

$$\begin{aligned} \det(S_{t+1,h}) &= \det(S_{t,h} + X_{t,h}) \\ &= \det\Big(S_{t,h}^{1/2} \Big(I_d + S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2} \Big) S_{t,h}^{1/2} \Big) \\ &= \det(S_{t,h}^{1/2}) \det(I_d + S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}) \det(S_{t,h}^{1/2}) \\ &= \det(S_{t,h}) \det(I_d + S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2}), \end{aligned} \tag{89}$$

where the first equality uses the update rule $S_{t+1,h} = S_{t,h} + X_{t,h}$, and the last two equalities use the multiplicativity of the determinant and $\det(S_{t,h}^{1/2})^2 = \det(S_{t,h})$.

Summing over $t = 1, \ldots, T$, we obtain

$$\begin{aligned} \textstyle\sum_{t=1}^T \mathsf{tr}\Big(S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2} \Big) &\le 2 \textstyle\sum_{t=1}^T \log\det\Big(I_d + S_{t,h}^{-1/2} X_{t,h} S_{t,h}^{-1/2} \Big) \\ &= 2 \textstyle\sum_{t=1}^T \big[\log\det(S_{t+1,h}) - \log\det(S_{t,h}) \big] \end{aligned}$$

$$= 2\left[\log\det(S_{T+1,h}) - \log\det(S_{1,h})\right]$$
$$= 2\left[\log\det(S_{T+1,h}) - d\log\lambda\right], \tag{90}$$

where the last equality uses $S_{1,h} = \lambda I_d$. It remains to upper bound $\log\det(S_{T+1,h})$. Let $\nu_1, \ldots, \nu_d > 0$ be the eigenvalues of $S_{T+1,h}$. By he arithmetic mean–geometric mean inequality (AM–GM) and the update $S_{T+1,h} = \lambda I_d + \sum_{t=1}^{T} X_{t,h}$,

$$\begin{aligned}\log\det(S_{T+1,h}) &= \log\textstyle\prod_{i=1}^{d}\nu_i \le d\log\Big(\frac{1}{d}\sum_{i=1}^{d}\nu_i\Big)\\ &= d\log\Big(\tfrac{\mathsf{tr}(S_{T+1,h})}{d}\Big)\\ &= d\log\left(\frac{\mathsf{tr}(\lambda I_d) + \sum_{t=1}^{T}\mathsf{tr}(X_{t,h})}{d}\right)\\ &\le d\log\big(\tfrac{d\lambda+T}{d}\big)\,, \end{aligned}\tag{91}$$

where the last inequality uses $\mathsf{tr}(X_{t,h}) \le 1$ for all $t \in T$. Therefore,

$$\begin{aligned}\textstyle\sum_{t=1}^{T}\mathsf{tr}\Big(S_{t,h}^{-1/2}X_{t,h}S_{t,h}^{-1/2}\Big) &\le 2\left[d\log\big(\tfrac{d\lambda+T}{d}\big) - d\log\lambda\right]\\ &= 2d\log\big(1+\tfrac{T}{d\lambda}\big) = \widetilde{O}(d).\end{aligned}\tag{92}$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state our two main contributions (sample-complexity bound $\tilde{\mathcal{O}}(H^9 d^2/\varepsilon^4)$ in the generative-model setting and regret bound $\tilde{\mathcal{O}}\big(d \max_{i\in[n]} A_i\, H^2\sqrt{T}\big)$ in the online interactive setting), which are formalized in Theorems 2 and 3. Table 1 positions these claims relative to prior work, and the scope (R-LMGs with total-variation uncertainty under fictitious (s, a_i)-rectangularity) is explicitly stated.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The scope of our results is delimited by three modeling choices that are stated and discussed where introduced: (i) we assume independent linear function approximation for the nominal kernel and reward (Section 2, Eq. (2)); (ii) we work with the fictitious uncertainty set under (s, a_i)-rectangularity defined via total variation (Section 2); (iii) the online algorithm relies on access to an adversarial sampling oracle inside the uncertainty set (Definition 1), motivated by adversarial-training and sim-to-real practices but stronger than nominal-only access. The current results do not extend to more general divergences (e.g., KL, Wasserstein), to centralized linear structures, or to non-rectangular uncertainty sets, and they are theoretical only (no empirical evaluation).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theorem and lemma is numbered and cross-referenced; assumptions are stated either in the theorem statement or in the corresponding section. The full proofs of Theorems 2 and 3 are deferred to the appendix (Sections B and onward), with intermediate lemmas (e.g., the ridge-regression error bound, the optimistic/pessimistic estimate inequalities) stated and proved separately.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper is purely theoretical and does not contain experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper is purely theoretical; no experiments, code, or data are produced. The proposed algorithms are fully specified in pseudocode (Algorithms 2 and 3).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: The paper does not contain experiments. Theoretical guarantees are stated as high-probability bounds with the failure probability δ explicitly tracked in the theorem statements.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This is a purely theoretical study of sample complexity and regret for distributionally robust Markov games; the work involves no human subjects, no data collection, no released models, and no deployment, and we have reviewed and conformed to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is foundational theoretical research on sample-efficient algorithms for distributionally robust multi-agent reinforcement learning; it does not propose any specific application, deployable system, dataset, or model. Any downstream societal impact would be mediated by future applications built on top of robust MARL (e.g., safer autonomous-driving controllers or more reliable energy-system policies), and we are not aware of a direct path from our theoretical guarantees to specific positive or negative societal harms.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any data, pretrained models, generators, or scraped assets; it contributes only theoretical algorithms and proofs.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use any external code, datasets, or pretrained models. All prior work that we build upon (e.g., [35, 66, 75, 89]) is cited in the references.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release any new code, dataset, or model.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so no IRB or equivalent review is required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core method, theoretical analysis, or any experiments in this paper. Their usage was limited to writing/editing assistance, which the policy explicitly states does not require declaration.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.