Distributionally Robust Online Markov Game with Linear Function Approximation

Zewu Zheng, Yuanyuan Lin

Department of Statistics and Data Science, The Chinese University of Hongkong
zhengzw@link.cuhk.edu.hk, ylin@sta.cuhk.edu.hk

Abstract

The sim-to-real gap, where agents trained in a simulator face significant performance degradation during testing, is a fundamental challenge in reinforcement learning. Extensive works adopt the framework of distributionally robust RL, to learn a policy that acts robustly under worst case environment shift. Within this framework, our objective is to devise algorithms that are sample efficient with interactive data collection and large state spaces. By assuming d-rectangularity of environment dynamic shift, we identify a fundamental hardness result for learning in online Markov game, and address it by adopting minimum value assumption. Then, a novel least square value iteration type algorithm, DR-CCE-LSI, with exploration bonus devised specifically for multiple agents, is proposed to find an ϵ-approximate robust Coarse Correlated Equilibrium(CCE). To obtain sample efficient learning, we find that: when the feature mapping function satisfies certain properties, our algorithm, DR-CCE-LSI, is able to achieve ϵ-approximate CCE with a regret bound of $\mathcal{O}\{dH \min\{H, \frac{1}{\min\{\sigma_i\}}\}\sqrt{K}\}$, where K is the number of interacting episodes, H is the horizon length, d is the feature dimension, and σ_i represents the uncertainty level of player i. Our work introduces the first sample-efficient algorithm for this setting, matches the best result so far in single agent setting, and achieves minimax optimal sample complexity in terms of the feature dimension d. Meanwhile, we also conduct simulation study to validate the efficacy of our algorithm in learning a robust equilibrium.

Code — https://github.com/zewuzheng17/Markov_Game
Extended version — https://arxiv.org/abs/2511.07831

1 Introduction

Reinforcement learning (RL) has experienced significant advancements across various industrial applications, including video games, autonomous driving, and large language models, owing to its capacity to learn complex decision-making policies directly from data (Sutton and Barto 2018; Mnih et al. 2015; Silver et al. 2017; Ouyang et al. 2022). The complexity of RL tasks increases substantially when multiple agents operate in the same environment, a scenario commonly referred to as multi-agent reinforcement learning (MARL) (Vinyals et al. 2019; Berner et al. 2019). In MARL, the Markov Game framework (Littman 1994; Shapley 1953) is widely adopted, where each agent optimizes its individual reward function by learning a Markov policy that serves as the best response to the policies of other agents. The solution to these joint policies is typically framed as a stable equilibrium, such as Nash equilibrium (Nash 2024) or coarse correlated equilibrium (Aumann 1987). Extensive research has investigated learning equilibria in Markov Games across various settings, proposing algorithms with theoretical guarantees on sample efficiency.

Robustness is essential A critical challenge in reinforcement learning (RL) is the sim-to-real gap, characterized by discrepancies between simulated training environments and real-world deployment settings, which often result in degraded performance (Koos, Mouret, and Doncieux 2012; Jiang et al. 2021). This issue has motivated extensive research into distributionally robust reinforcement learning, where the objective is to develop policies that maintain robustness against variations in environmental dynamics (Iyengar 2005; Nilim and El Ghaoui 2005; Shi et al. 2023; Liu et al. 2022). The challenge is particularly pronounced in MARL, as highlighted by (Shi et al. 2024b), where the sensitivity of equilibrium solutions to minor environmental perturbations exacerbates the problem.

Large state space with linear function approximation When state and action space are large, function approximation are applied to mitigate the curse of dimensionality. The linear MDP framework, where the transition kernel and reward function are represented as linear functions of low-dimensional feature vectors, is analytically tractable and extensively studied by (Jin et al. 2020; Yang and Wang 2020; Cisneros-Velarde and Koyejo 2023; Wang, Yan, and Fan 2021; He et al. 2023; Li, Zhao, and Zhou 2024) in single agent RL. However, research on linear Markov Games remains limited (Chen, Zhou, and Gu 2022; Xie et al. 2020), particularly for multi-player general-sum Markov Games (Cisneros-Velarde and Koyejo 2023). To date, the study of online robust general-sum Markov Games with linear function approximation remains an open area, representing a critical gap in the existing literature. As a result, it is natural to ask:

Can we design a provably sample efficient algorithms for online robust general sum Markov Game with linear function approximation?

Our main contributions

- We investigate the inherent hardness of online learning in robust Markov games by constructing a specific two-player general-sum Markov game and establishing a lower bound that demonstrates the impossibility of learning without additional assumptions in this context. Subsequently, we adopt the minimum value assumption introduced by (Lu et al. 2024), providing analysis of its applicability.
- We address the additional challenges posed by distributionally robust Markov games compared to the single-agent setting (Liu and Xu 2024; Liu, Wang, and Xu 2024). From an algorithmic perspective, we introduce agent-specific bonus terms to ensure adequate exploration and maintain each agent's own risk preference. From a technical standpoint, when applying concentration arguments to establish uniform upper bounds for the CCE, we use the Find-CCE subroutine. This approach handles the issue that the CCEs of general-sum games are unstable (i.e., not Lipschitz continuous) with respect to changes in the payoff matrices, while remaining computationally feasible.
- We develop an algorithm for interactive data collection, utilizing ridge regression and deriving an instance-dependent upper bound via refined analysis. By exploiting the shrinkage properties of the robust value function and proposing a general assumption on feature mapping and transition kernel properties, our algorithm achieves a regret of order $\mathcal{O}\{dH\min\{H, \frac{1}{\min\{\sigma_i\}}\}\sqrt{K}\}$. This bound exhibits polynomial dependence on all key problem parameters and is minimax optimal with respect to the feature dimension d. The algorithm's efficacy is corroborated through further simulation studies.

2 Related Work

Robust Online Linear MDPs

The setting of online linear Markov Decision Processes (MDPs) (Yang and Wang 2019, 2020; Zanette et al. 2020; Jin et al. 2020; He, Zhou, and Gu 2021; He et al. 2023; Zhou, Gu, and Szepesvari 2021) has been extensively studied. The work of (He et al. 2023) achieved minimax optimality in this setting by incorporating variance-weighted ridge regression into their algorithm. In contrast, the study of online robust linear MDPs has only recently been explored in two works (Liu and Xu 2024; Liu, Wang, and Xu 2024). Specifically, (Liu, Wang, and Xu 2024) introduced a robust variant of variance-weighted ridge regression, achieving a regret bound of order $\mathcal{O}(dH\min\{1/\sigma, H\}\sqrt{K})$ under full data coverage assumption. However, this result still falls short of the constructed lower bound, which is of order $\Omega(dH^{\frac{1}{2}}\min\{1/\sigma, H\}\sqrt{K})$, highlighting that the single-agent counterpart of this setting remains insufficiently explored.

Robust Markov Game

While there has been extensive research on distributionally robust MDPs (Liu et al. 2022; Clavier, Pennec, and Geist 2023; Shi and Chi 2024; Shi et al. 2023; Wang et al. 2023a; Lu et al. 2024), the study of robust Markov games remains relatively underexplored. Existing works, such as (Kardeş, Ordóñez, and Hall 2011; Zhang et al. 2020), primarily focus on proving the existence of equilibria and analyzing convergence properties. In offline setting, a unified framework P^2M^2PO has been proposed by (Blanchet et al. 2023), with sample complexity of $\mathcal{O}(\frac{H^5|S|^2|A|^2}{\epsilon})$. (Shi et al. 2024a,b; Jiao and Li 2024) extended this framework to generative model setting, where samples can be obtained from any state-action pair. In particular, (Jiao and Li 2024) proposed a Q-FTRL type algorithm, demonstrating that it is minimax optimal and breaking the curse of multi-agency with a sample complexity of $\mathcal{O}(\frac{H^3|S|\sum_{i=1}^m|A_i|}{\varepsilon^2}\min\left\{H, \frac{1}{\sigma}\right\})$. In the more realistic online setting, the work most relevant to ours is (Ma et al. 2023). However, their approach requires the uncertainty level $\sigma_i \leq \max\left\{\frac{\varepsilon}{|S|H^2}, \frac{p_{\min}}{H}\right\}$ for all $i \in [n]$. This constraint limits the robustness of their framework, especially when high accuracy is required ($\varepsilon \to 0$) or the minimum positive transition probabilities ($p_{\min} \to 0$).

Online Linear Markov Games

The study of sample complexity in online linear Markov games encompasses both centralized learning (Xie et al. 2020; Chen, Zhou, and Gu 2022; Cisneros-Velarde and Koyejo 2023), which employs global linear function approximation, and decentralized learning, which relies on independent linear function approximation (Cui, Zhang, and Du 2023; Wang et al. 2023b; Dai, Cui, and Du 2024). While decentralized learning is often more favorable in tabular Markov games due to its ability to alleviate the curse of multi-agency, extending this approach to linear function approximation requires adopting independent linear function approximation. However, this deviates from the linear MDP setting commonly used in single-agent reinforcement learning. Furthermore, addressing robustness in such decentralized settings remains an open and challenging problem.

In the context of centralized learning, (Xie et al. 2020; Chen, Zhou, and Gu 2022) focus on two-player zero-sum games, which are less general compared to the multi-player general-sum games considered in (Cisneros-Velarde and Koyejo 2023). The work in (Cisneros-Velarde and Koyejo 2023) introduced the NQOVI algorithm, achieving a regret bound of $\mathcal{O}\left(\sqrt{d^3H^5K}\right)$. Notably, the incorporation of robustness into online linear Markov games has not yet been studied, underscoring the significance of our work in addressing this gap.

Notation Throughout this paper, we adopt the notation $[PV](s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}[V(s')]$ to represent the expected value under the transition dynamics. The set of integers $\{1, 2, \dots, n\}$ is denoted by $[n]$. For a vector where the i^{th} element is given by v_i, we use the notation $[v_i]_{i\in[d]}$. The eigenvalues of a square matrix A is denoted by $\lambda(A)$. To define norms, we write $||\phi||_A = \sqrt{\phi^\top A\phi}$, where $\phi \in \mathbb{R}^d$ is a vector and A is a positive semi-definite matrix. Moreover, with a set of parameters $\mathcal{X} := \{d, H, \{\sigma_i\}_{i=1}^n, 1/\delta\}$, the expression $f(\mathcal{X}) = \mathcal{O}(g(\mathcal{X}))$ signifies that there exists a constant C such that $f(\mathcal{X}) \leq Cg(\mathcal{X})$, and $f(\mathcal{X}) = \Omega(g(\mathcal{X}))$ indicates $f(\mathcal{X}) \geq Cg(\mathcal{X})$ for some constant C, and both of the expression omit logarithmic factors. The value $[V]_\alpha = \alpha$ if $V \geq \alpha$, otherwise $[V]_\alpha = V$. To further simplify notation, we let $\mathcal{Q} := \{(Q_1, Q_2, \cdots, Q_n) : \mathcal{S} \times \mathcal{A} \to R^n\}$ be the function class of estimated Q value in our algorithm, and $\mathcal{V}$ be its expectation w.r.t the CCE policy. Finally, we use π instead of π_h whenever there is no ambiguity.

3 Preliminaries

In this section, we begin by presenting the foundational concepts of distributionally robust general-sum Markov game, which serves as the basis for our analysis.

Distributionally Robust Markov Game

A distributionally robust Markov game can be represented as the tuples:

$$\mathcal{MG}_{rob} = (\mathcal{S}, \{\mathcal{A}_i\}_{1\leq i\leq n}, \{\mathcal{U}_\rho^{\sigma_i}(P^0)\}_{1\leq i\leq n}, r, H)$$

To clarify the notation here, $\mathcal{S}$ is the state space, either discrete or continuous, $\mathcal{A}_i = \{1, \cdots, A_i\}$ is the action space for player i, $P^0 = \{P_h^0\}_{1\leq h\leq H}$ is the nominal transition kernel in the simulator, $r = \{r_{i,h}\}_{1\leq i\leq n, 1\leq h\leq H} \in [0,1]$ is the reward functions, $H > 0$ is the horizon length. The joint action space for all players is defined as $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. By taking joint action $\boldsymbol{a} \in \mathcal{A}$ at state $s \in \mathcal{S}$ and time h, each player i receives his own deterministic scalar reward $r_{i,h}(s, \boldsymbol{a})$, then the environment transits to the next state s' with probability $P_h(s'|s, \boldsymbol{a})$. To incorporate robustness, the transition kernel $P = \{P_h\}_{1\leq h\leq H}$ in test environment is within a prescribed uncertainty set $\mathcal{U}_\rho^{\sigma_i}(P^0)$ for each player i, which is centered around a nominal transition kernel P^0 and will be introduced shortly.

Robust Linear Markov Game

This paper focuses on distributionally robust Markov games with linear function approximation. Accordingly, we introduce the fundamental assumptions of linear Markov game (Jin et al. 2020; Cisneros-Velarde and Koyejo 2023; He et al. 2023).

Assumption 3.1. (Linear Markov game) For any step h, and player i, there exist a known feature map function $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, such that the reward $r_{i,h}(s, \boldsymbol{a})$ and transition kernel $P_h(s'|s, \boldsymbol{a})$ can be represented in following form for any $(s, a, s', i, h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [n] \times [H]$:

$$\begin{aligned} r_{i,h}(s, \boldsymbol{a}) &= \langle \phi_{s\boldsymbol{a}}, \eta_{i,h} \rangle \\ P_h(s'|s, \boldsymbol{a}) &= \langle \phi_{s\boldsymbol{a}}, \mu_h^0(s') \rangle \end{aligned} \tag{1}$$

where $\eta_{i,h} \in \mathbb{R}^d$ represents a known d-dimensional vector that characterizes the reward of player i, and μ_h^0 is a d dimensional vector with each element being an unknown probability measure over $\mathcal{S}$. In addition, we assume that $\sum_{j=1}^d \phi_j(s, \boldsymbol{a}) = 1$, where $\phi_j(s, \boldsymbol{a}) \geq 0$ is the j^{th} element of ϕ and we represent $\phi(s, \boldsymbol{a})$ by $\phi_{s\boldsymbol{a}}$ for simplicity.

With the linearity structure of the transition kernel, we further adopt the d-rectangular uncertainty set structure in (Ma et al. 2022; Liu and Xu 2024; Goyal and Grand-Clement 2023).

Definition 3.2. (d-Rectangular robust linear Markov game) For a given robust Markov game instance $\mathcal{MG}_{rob} = (\mathcal{S}, \{\mathcal{A}_i\}_{1\leq i\leq n}, \{\mathcal{U}_{TV}^{\sigma_i}(P^0)\}_{1\leq i\leq n}, r, H)$, $\mathcal{U}_{TV}^{\sigma_i}(P^0)$ is the d-rectangular uncertainty set for player i under total variation distance:

$$\mathcal{U}_{TV}^{\sigma_i}(P^0) := \otimes_{[H],\mathcal{S},\mathcal{A}} \mathcal{U}_{TV}^{\sigma_i}(P_{h,s,\boldsymbol{a}}^0) \tag{2}$$

with

$$\begin{aligned} &\mathcal{U}_{TV}^{\sigma_i}(P_{h,s,\boldsymbol{a}}^0) : \{\phi(s, \boldsymbol{a})^\top \mu_h(\cdot) : \mu_h \in \mathcal{U}_{TV}^{\sigma_i}(\mu_h^0)\} \\ &\mathcal{U}_{TV}^{\sigma_i}(\mu_h^0) : \otimes_{j\in[d]} \{\mu_{h,j} : D_{\mathrm{TV}}(\mu_{h,j} \| \mu_{h,j}^0) \leq \sigma_i\} \end{aligned} \tag{3}$$

where $D_{\mathrm{TV}}(\mu_{h,j}, \mu_{h,j}^0) = \frac{1}{2}\int_{s\in\mathcal{S}} |\mu_{h,j}(s) - \mu_{h,j}^0(s)| ds$.

The d-rectangular robust linear Markov game is essential since it ensures that the robust action-value function remains linear, therefore avoiding completeness assumption imposed in the literature of RL with general function approximation, i.e, (Jin, Liu, and Yu 2022). In this paper, we focus on the total variation distance, which is widely used in the distributionally robust reinforcement learning literature (Shi et al. 2024b, 2023; Liu and Xu 2024; Liu, Wang, and Xu 2024; Wang, Shi, and Chi 2024; Lu et al. 2024).

Robust value function and Bellman Equation A joint Markovian policy $\pi = \{\pi_h(\cdot|s) : \mathcal{S} \to \Delta(\mathcal{A})\}_{h=1}^H$ takes any state $s \in \mathcal{S}$ as input, and output a probability simplex over the joint action space of all players. The robust value function $V_{i,h}^{\pi,\sigma}$ and action value function $Q_{i,h}^{\pi,\sigma}$ can be define as, for any $(i, h, s, a) \in [n] \times [H] \times \mathcal{S} \times \mathcal{A}$:

$$V_{i,h}^{\pi,\sigma} = \inf_{P\in\mathcal{U}_{TV}^{\sigma_i}(P^0)} V_{i,h}^{\pi,P}, \quad Q_{i,h}^{\pi,\sigma} = \inf_{P\in\mathcal{U}_{TV}^{\sigma_i}(P^0)} Q_{i,h}^{\pi,P} \tag{4}$$

where

$$V_{i,h}^{\pi,P}(s) = \mathbb{E}_{\pi,P}[\sum_{t=h}^H r_{i,t}(s_t, \boldsymbol{a}_t) | s_h = s]$$

$$Q_{i,h}^{\pi,P}(s, \boldsymbol{a}) = \mathbb{E}_{\pi,P}[\sum_{t=h}^H r_{i,t}(s_t, \boldsymbol{a}_t) | s_h = s, \boldsymbol{a}_h = \boldsymbol{a}]$$

Then, $V_{i,h}^{\pi,\sigma}$ and $Q_{i,h}^{\pi,\sigma}$ satisfy the robust Bellman Equation (Iyengar 2005; Liu and Xu 2024):

$$\begin{aligned} Q_{i,h}^{\pi,\sigma}(s, \boldsymbol{a}) &= r_{i,h}(s, \boldsymbol{a}) + \inf_{P\in\mathcal{U}_{TV}^{\sigma_i}(P_{h,s,\boldsymbol{a}}^0)} PV_{i,h+1}^{\pi,\sigma} \\ V_{i,h}^{\pi,\sigma}(s) &= \mathbb{E}_{\boldsymbol{a}\sim\pi_h(\cdot|s)}[Q_{i,h}^{\pi,\sigma}(s, \boldsymbol{a})] \end{aligned} \tag{5}$$

Solution concept and optimality In this work, we aim to find the robust CCE, which is defined as below.

Definition 3.3. A joint policy $\pi = \{\pi_h\}_{1\leq h\leq H}$ is said to be a robust CCE if:

$$V_{i,1}^{\pi,\sigma}(s) \geq V_{i,1}^{\star,\pi_{-i},\sigma}(s), \quad \forall (s, i) \in \mathcal{S} \times [n] \tag{6}$$

where π_{-i} be the joint policy of all players except for the i^{th} player, and

$$V_{i,h}^{\star,\pi_{-i},\sigma}(s) = V_{i,h}^{br(\pi_{-i})\times\pi_{-i},\sigma}(s) = \max_{\pi_i'\in\Pi} V_{i,h}^{\pi_i'\times\pi_{-i},\sigma} \tag{7}$$

is the robust best response policy of player i with respect to opponent policy π_{-i}. A robust CCE ensures that no player can gain by unilaterally deviating from their current strategy under worst-case transition, without the requirement that players should act independently.

The existence of a robust best-response policy and robust equilibrium, as introduced earlier, has been confirmed by (Blanchet et al. 2023). In MARL, achieving an ε-approximate robust CCE is computationally feasible. Consequently, to find such an equilibrium, our goal is to design sample-efficient algorithms that achieve sublinear regret, where the regret is defined as

$$Regret(K) = \max_{i\in[n]} \sum_{k=1}^{K} \left[V_{i,1}^{\star,\pi_{-i}^{k},\sigma}(s_1^k) - V_{i,1}^{\pi^k,\sigma}(s_1^k) \right] \quad (8)$$

4 Distributionally Robust Markov Game with Linear Function Approximation

Vanishing Minimal Value

The incorporation of robustness poses a fundamental challenge in online reinforcement learning and the Markov game framework. As discussed in detail by (Lu et al. 2024) in the single-agent case, with the existence of support shift problem, finding the optimal robust policy is infeasible without additional assumptions. We present the result in MARL in the theorem below. Relevant discussion can be found in the Appendix.

Theorem 4.1. *(Online regret lower bound for robust Markov game) There exist two players general sum robust Markov game $\mathcal{MG}_{rob}^{\theta}, \theta \in [2]$, such that the following lower bound holds:*

$$\inf_{\mathcal{ALG}} \sup_{\theta\in[2]} \mathbb{E}\left[\mathrm{Regret}_{\theta}^{\mathcal{ALG}}(K)\right] = \Omega(\sigma \cdot HK) \quad (9)$$

where $\mathrm{Regret}_{\theta}^{\mathcal{ALG}}(K)$ denote the online regret for algorithm $\mathcal{ALG}$ under robust Markov Game instance $\mathcal{MG}_{rob}^{\theta}$, σ is the shared uncertainty level of both players.

The support shift problem arises since the collected samples may not cover all possible trajectories related to the worst-case environment, leading to a lack of relevant information. Consequently, algorithms operating in such an environment may perform no better than random guessing in unobserved states, making it impossible to effectively learn a reliable equilibrium. This limitation may be less relevant in other settings, such as generative models or offline scenarios (Shi et al. 2023; Wang, Shi, and Chi 2024; Blanchet et al. 2023; Shi et al. 2024b). In generative models, algorithms have access to simulators and can query information starting from any state. While in offline setting, additional assumptions about dataset coverage are often imposed to mitigate support shift problems.

To address these challenges, we adopt the vanishing minimal value assumption introduced by (Lu et al. 2024) and adapt it to our framework, as stated below.

Assumption 4.2. (Vanishing minimal value for robust Markov game) We assume that the robust Markov game adheres to the following conditions:

$$\min_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma}(s) = 0$$

for $\forall(i,h,\pi) \in [n]\times[H]\times\Pi$. In addition, the initial state $s_0 \notin \arg\min_{s\in\mathcal{S}} V_{i,h}^{\pi,\sigma}(s)$.

We derive an intuitive implication of imposing the Vanishing minimal value assumption in our setting by the following proposition.

Proposition 4.3 (Equivalence of Optimization under Minimal Value Assumption)**.** *For any function with $V : \mathcal{S} \to [0, H]$, with $\min_{s\in\mathcal{S}} V(s) = 0$, we have $\forall(i,h,s,\boldsymbol{a}) \in [n]\times[H]\times\mathcal{S}\times\mathcal{A}$,*

$$\inf_{P\in\mathcal{U}_{TV}^{\sigma_i}(P_{h,s,\boldsymbol{a}}^0)} \mathbb{E}_P[V] = \sigma_i \mathbb{E}_{\tilde{P}_{h,s,\boldsymbol{a}}}[V] \quad (10)$$

The transition kernel can be represented as $\tilde{P}_{h,s,\boldsymbol{a}} = \langle\phi(s,\boldsymbol{a}), \tilde{\mu}_h\rangle$ where $\tilde{\mu}_h = \arg\inf_{\mu\in B_h^{\sigma_i}} \mathbb{E}_\mu[V]$ and

$$B_h^{\sigma_i} = \{\mu : \sup_{s'\in\mathcal{S}, j\in[d]} \{\frac{\mu_j(s')}{\mu_{h,j}^0(s')}\} \leq \frac{1}{\sigma_i}\} \quad (11)$$

Additionally, we have:

$$\sup_{s'\in\mathcal{S}} \{\frac{\tilde{P}_h(s'|s,\boldsymbol{a})}{P_h^0(s'|s,\boldsymbol{a})}\} \leq \frac{1}{\sigma_i} \quad (12)$$

Proposition 4.3 ensures that the transition kernels under the d-rectangular robust linear Markov game assumption are restricted to stay within the support of the nominal transition kernel, thereby avoiding the support shift problem. Additionally, the assumption can be fulfilled by augmenting the Markov game with an isolated absorbing fail state s_f in each time step, i.e, $P_h^0(s_f|s_f,\boldsymbol{a}) = 1, r_{i,h}(s_f,\boldsymbol{a}) = 0, \forall(i,h,\boldsymbol{a}) \in [n]\times[H]\times\mathcal{A}$. Adding a fail state does not affect the optimal value function and policy in the nominal environment, and therefore, it extends non-robust Markov game without loss of generality.

This assumption is reasonable, as it applies to many large-scale, real-world scenarios. For instance, in warfare, soldiers face daily risks to their lives; in hospitals, patients undergoing treatment may not survive each day; and in round-based computer games, players can fail in each round. When the "game" ends in every possible step, the associated minimal value is zero. Similar assumption has been made by (Panaganti et al. 2022; Liu and Xu 2024), to avoid solving for computationally inefficient optimization problem in both online and offline RL problems.

Distributionally Robust CCE Least Square Iteration

Training process In each interacting episode $k \in [K]$, players constructs their own robust action-value function estimate, $\{Q_{i,h}^k\}_{h=1}^H$, using the data from previous episodes, $\{(s_h^1, \boldsymbol{a}_h^1), \cdots, (s_h^{k-1}, \boldsymbol{a}_h^{k-1})\}_{h=1}^H$. Subsequently, the players update their joint policy π_h^k by computing a CCE of a n-player matrix game(Line 10 of Algorithm 1). The updated policy π_h^k is then used to interact with the environment, generating new data until the end of the current episode. This process is repeated until the maximum number of episodes, K, is reached. Under this algorithmic framework, we present and discuss on the key steps of our algorithm. In this framework, the robust Bellman Equation (5) can be expressed as:$\forall(i,h,s,a) \in [n]\times[H]\times\mathcal{S}\times\mathcal{A}$:

$$Q^{\pi,\sigma}_{i,h}(s,\boldsymbol{a}) = r_{i,h}(s,\boldsymbol{a}) + \inf_{P\in\mathcal{U}^{\sigma_i}_{TV}(P^0_{h,s,\boldsymbol{a}})} PV^{\pi,\sigma}_{i,h+1}$$
$$= r_{i,h}(s,\boldsymbol{a}) + \left\langle \phi_{s\boldsymbol{a}}, \inf_{\mu\in\mathcal{U}^{\sigma_i}_{TV}(\mu^0_h)} \mathbb{E}_\mu[V^{\pi,\sigma}_{i,h+1}] \right\rangle$$

then, by strong duality (Iyengar 2005; Wang, Shi, and Chi 2024) and Assumption 4.2, one has,

$$\inf_{\mu\in\mathcal{U}^{\sigma_i}_{TV}(\mu^0_{h,j})} \mathbb{E}_\mu[V^{\pi,\sigma}_{i,h+1}] =$$
$$\max_{\alpha\in\left[\min(V^{\pi,\sigma}_{i,h+1}),\max(V^{\pi,\sigma}_{i,h+1})\right]} \left\{\mathbb{E}_{\mu^0_{h,j}}[[V^{\pi,\sigma}_{i,h+1}]_\alpha] - \sigma_i\alpha\right\}$$

Therefore, the robust Bellman Equation can be further expressed as,

$$Q^{\pi,\sigma}_{i,h}(s,\boldsymbol{a}) = r_{i,h}(s,\boldsymbol{a}) + \left\langle \phi_{s\boldsymbol{a}}, [\max_\alpha\{\nu_{i,h,j}(\alpha) - \sigma_i\alpha\}]_{j\in[d]} \right\rangle$$

where $\nu_{i,h,j}(\alpha) = \mathbb{E}_{\mu^0_{h,j}}[V^{\pi,\sigma}_{i,h+1}]_\alpha$. Given the collected data, we then apply ridge regression to estimate $\nu_{i,h}(\alpha) = (\nu_{i,h,1}(\alpha),\cdots,\nu_{i,h,d}(\alpha))$. The estimator is constructed by

$$\hat{\nu}^k_{i,h}(\alpha) = \arg\min_{\nu\in R^d} \sum_{\tau=1}^{k-1} (\nu^T\phi^\tau_h - [V^k_{i,h+1}]_\alpha)^2 + \lambda\|\nu\|^2_2$$
$$= (\Lambda^k_h)^{-1} \sum_{\tau=1}^{k-1} \phi^\tau_h [V^k_{i,h+1}]_\alpha(s^\tau_{h+1}) \tag{13}$$

where $\Lambda^k_h = \lambda I + \sum_{\tau=1}^{k-1}\phi^\tau_h(\phi^\tau_h)^T$ and $\phi^\tau_h = \phi(s^\tau_h, \boldsymbol{a}^\tau_h)$. To obtain an robust estimator, we further let $\hat{w}^k_{i,h,j} = \max_{\alpha\in[0,H]}\{\hat{\nu}^k_{i,h,j}(\alpha) - \sigma_i\alpha\}$, where $\hat{\nu}^k_{i,h,j}(\alpha)$ is the j^{th} element of vector $\hat{\nu}^k_{i,h}(\alpha)$. By incorporating the optimistic bonus term $\Gamma_{h,k}(s,\boldsymbol{a})$, we estimate $Q^{\pi^k,\sigma}_{i,h}$ by:

$$Q^k_{i,h}(s,\boldsymbol{a}) = \min\left\{r_{i,h}(s,\boldsymbol{a}) + \langle\phi_{s\boldsymbol{a}}, \hat{w}^k_{i,h}\rangle + \Gamma^i_{h,k}(s,a), \min\{H, \frac{1}{\sigma_i}\}\right\} \tag{14}$$

For more details about the training process, please refer to Algorithm 1.

About the bonus term In the estimation of $Q^{\pi^k,\sigma}_{i,h}$, we add a bonus term $\Gamma^i_{h,k}(s,\boldsymbol{a})$ to guide the exploration of the policy. It has the form $\Gamma^i_{h,k}(s,\boldsymbol{a}) = \beta_i\sum_{j=1}^d \sqrt{\phi_j(s,\boldsymbol{a})\mathbf{1}^T_j(\Lambda^k_h)^{-1}\mathbf{1}_j\phi_j(s,\boldsymbol{a})}$. The design of the bonus term arises from the need to address the optimization problem $\alpha_j = \arg\max_{\alpha\in[0,H]}\{\hat{\nu}^k_{i,h,j}(\alpha) - \sigma_i\alpha\}$, which depends on the j^{th} coordinate of $\hat{\nu}^k_{i,h}(\alpha)$. Solving this requires performing d ridge regression tasks for each agent i, with the objective being $[V^k_{i,h+1}(s^\tau_{h+1})]_{\alpha_j}, j\in[d]$. Thus, the bonus term consists of d upper confidence bonuses, which is different from the non-robust counterpart of linear MDP.

Algorithm 1: DR-CCE-LSI

Require: Parameter β,$\lambda > 0$
Initialize: $\Lambda^1_h = \lambda I$ for each step $h\in[H]$
for $k = 1$ **to** K **do**
  Receive initial state s^k_1
  Set $V^k_{i,H+1} = 0 \quad \forall i\in[n]$
  for step $h = H,\cdots,1$ **do**
    for $i = 1$ **to** n **do**
      Update $Q^k_{i,h}$ according to Equation (14)
    end for
    $\pi^k_h(s) \leftarrow$ Apply Find-CCE for the n-player game $\left(Q^k_{1,h}(s,\cdot),\cdots,Q^k_{n,h}(s,\cdot)\right)$
    for $i\in[n]$ **do**
      $V^k_{i,h}(s) = \mathbb{E}_{\pi^k_h}[Q^k_{i,h}(s,\boldsymbol{a})]$
    end for
  end for
  for Step $h = 1,\cdots,H$ **do**
    take action $\boldsymbol{a}^k_h \leftarrow \pi^k_h(s^k_h)$
    Receive next state s^k_{h+1}
    Update $\Lambda^{k+1}_h = \Lambda^k_h + \phi(s^k_h,\boldsymbol{a}^k_h)\phi(s^k_h,\boldsymbol{a}^k_h)^T$
  end for
end for

Algorithm 2: Find-CCE

1: **Input:** Matrix game $\left(Q^k_{1,h}(s,\cdot),\cdots,Q^k_{n,h}(s,\cdot)\right)$
2: Discretization parameter $\epsilon > 0$
3: Pick a tuple $(\tilde{Q}_1(s,\cdot),\tilde{Q}_2(s,\cdot),\cdots,\tilde{Q}_n(s,\cdot))$ from the $\varepsilon-$ *cover* of function class $\mathcal{Q}$ such that $\sup_{i\in[n],\boldsymbol{a}\in\mathcal{A}}\left|Q^k_{i,h}(s,\boldsymbol{a}) - \tilde{Q}_i(s,\boldsymbol{a})\right| \le \epsilon$
4: Let $\pi^k_h(s)$ be the CCE of the matrix game with payoff $(\tilde{Q}_1(s,\cdot),\tilde{Q}_2(s,\cdot),\cdots,\tilde{Q}_n(s,\cdot))$
5: **Return:** $\pi^k_h(s)$

Technical consideration of Find-CCE In our algorithms, at each iteration k, the estimated value $V^k_{i,h}(s) = \mathbb{E}_{\pi^k_h}[Q^k_{i,h}(s,\boldsymbol{a})]$, is derived from the sample tuples $(s^\tau_h,\boldsymbol{a}^\tau_h)$ for all $\tau \le k-1$. Consequently, $V^k_{i,h}$ exhibits complex statistical dependencies on the previously collected samples. It is then unavoidable for us to employ a covering-number argument to decouple the dependency. Since $Q^k_{i,h}(s,\cdot)$ depends on problems parameters $\hat{w}^k_{i,h}$ and Λ^k_h and the norm of both parameters is bounded, we can find an $\epsilon-$covering set for $Q^k_{i,h}(s,\cdot)$. However, if the policy $\pi^k_h(s)$ is obtained by solving for CCE for the matrix game $\left(Q^k_{1,h}(s,\cdot),\cdots,Q^k_{n,h}(s,\cdot)\right)$, then constructing a covering set for the function class $\mathcal{V}$ is challenging due to the inherent instability of Coarse Correlated Equilibria. As highlighted in (Xie et al. 2020), a counterexample is provided, with details in Lemma 4.4 below. Their lemma demonstrates that there exist two matrix games, although the difference of their payoff is smaller than any constant ϵ, the expected return of their Coarse Correlated Equilibrium can still deviate.

Lemma 4.4. *((Xie et al. 2020), Lemma 19.(Modified)) For any $\epsilon > 0$, there exists a pair of games $Q(s,\cdot)$ and $Q'(s,\cdot)$, each with a unique CCE π and π', such that*

$$\|Q(s,\cdot) - Q'(s,\cdot)\|_\infty \le 2\epsilon \quad \textit{and} \quad \|V - V'\|_\infty \ge 1$$

where $V = \mathbb{E}_\pi[Q(s,\cdot)]$ and $V' = \mathbb{E}_{\pi'}[Q'(s,\cdot)]$

The intuition of the Find-CCE subroutine for solving this problem is straightforward. We observe that π in Lemma 4.4 is actually 2ϵ approximate CCE for the matrix game $Q'(s,\cdot)$. Therefore, for any $Q(s,\cdot)$ and $Q'(s,\cdot)$ with $\|Q(s,\cdot) - Q'(s,\cdot)\|_\infty \le 2\epsilon$, we can find a fix 2ϵ approximate CCE $\bar{\pi}$ for both games, while the infinity norm of the difference between $V = \mathbb{E}_{\bar{\pi}}[Q(s,\cdot)]$ and $V' = \mathbb{E}_{\bar{\pi}}[Q'(s,\cdot)]$ can be bounded by 2ϵ. In this way, we successfully construct 2ϵ covering set of V. As highlighted by (Xie et al. 2020), Find-CCE can be implemented efficiently with computational feasibility.

5 Theoretical Results

In this subsection, we are ready to provide our theoretical results for Algorithm 1. Proof can be found in the Appendix.

Theorem 5.1 (Instance Dependent Upper Bound)**.** *If we let $\lambda = 1, \beta_i = \min\{H, \frac{1}{\sigma_i}\}\sqrt{c_\beta nd\log(\frac{ndHK}{\delta})}, \epsilon = \frac{1}{KH}$ in Algorithm 1. Then, under Assumption 3.1, and 4.2, the suboptimality of DR-CCE-LSI satisfies below upper bound with probability at least $1-\delta$*

$$Regret(K) \le 8\min\{H, \frac{1}{\min\{\sigma_i\}}\}\sqrt{2HK\log(\frac{3n}{\delta})} + 4\max_i\{\beta_i\}\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{j=1}^{d}\sqrt{\phi_{h,j}^k \mathbf{1}_j^T(\Lambda_h^k)^{-1}\mathbf{1}_j\phi_{h,j}^k} \tag{15}$$

Where $\phi_{h,j}^k = \phi_j(s_h^k, a_h^k)$ be the j^{th} coordinate of ϕ_h^k, and δ is any fixed constant in $(0,1)$.

An interesting observation In our algorithm, each player may have different risk preferences by choosing various uncertainty levels σ_i. A player with a smaller uncertainty level is aggressive and willing to take risks, since he tries to find a policy that is less robust. The bound above inevitably depends on $\max_i\{\beta_i\}$, which may be large if at least one of the players is risk seeking. Our theoretical result implies that, given finite interacting episodes, players learning in the game should have a common sense of risk preference level to achieve the best sample efficiency.

Necessity for further assumption The term $\sum_{k=1}^K\sum_{h=1}^H\sum_{j=1}^d\sqrt{\phi_{h,j}^k\mathbf{1}_j^T(\Lambda_h^k)^{-1}\mathbf{1}_j\phi_{h,j}^k}$ is similar to (Liu, Wang, and Xu 2024; Liu and Xu 2024; Ma et al. 2022; Wang, Shi, and Chi 2024) in single agent RL. Since (Ma et al. 2022; Wang, Shi, and Chi 2024) consider offline setting, the term is further upper-bounded under certain data coverage assumption. While in online learning, (Liu, Wang, and Xu 2024; Liu and Xu 2024) proposed assumptions that $\mathbb{E}_\pi^{P^0}[\phi_{s\boldsymbol{a}}\phi_{s\boldsymbol{a}}^T] \ge c/d\cdot\mathbf{I}$ for some constant c. This assumption implicitly requires that the environment is exploratory enough, under certain structural properties of feature mapping ϕ and arbitrary policy π. In this paper, we are the first to show that such a bonus term has a learnability issue, followed by a fine-grained analysis on how the structure of feature mapping ϕ and properties of the transition kernel affect the sample complexity of our algorithm.

Theorem 5.2. *There exist an MDP instance, with three states, two actions, and horizon length 2, such that the term $\sum_{k=1}^K\sum_{j=1}^d\sqrt{\phi_j^k\mathbf{1}_j^T(\Lambda^k)^{-1}\mathbf{1}_j\phi_j^k}$ has an order of $\Omega(K)$.*

The Proof of Theorem 5.2 can be found at Appendix. The theorem implies that we can't upper bound $\sum_{k=1}^K\sum_{h=1}^H\sum_{j=1}^d\sqrt{\phi_{h,j}^k\mathbf{1}_j^T(\Lambda_h^k)^{-1}\mathbf{1}_j\phi_{h,j}^k}$ without imposing further assumptions, which serves as a unique challenges for online robust reinforcement learning with linear function approximation. We address this issue by providing a sufficient condition to ensure effective learning of our algorithm in the Corollary below, with the Proof postponed to the Appendix.

Corollary 5.3. *Without loss of generality, let $\mathcal{S}\times\mathcal{A}\subset R^m$ and $\phi: R^m \to R^d$ with $m \ge d$, if we assume:*

- *ϕ is non-degenerated. i.e. it does not map any m dimensional subset of R^m into d' dimensional subset in R^d, where $d' < d$.*
- *For any state-action pairs $\{s^j, \boldsymbol{a}^j\}_{j=1}^d$, if they satisfy the condition $\sum_{j=1}^d\phi(s^j,\boldsymbol{a}^j)\phi(s^j,\boldsymbol{a}^j)^T > \mathbf{0}$, then there exist a constant $c > 0$, such that $\sum_{j=1}^d\phi(s^j,\boldsymbol{a}^j)\phi(s^j,\boldsymbol{a}^j)^T \ge c\cdot\mathbf{I}$.*
- *$P_h^0(s'|s,\boldsymbol{a})$ is absolutely continuous w.r.t Lebesgue measure for all $(h,s,\boldsymbol{a},s') \in [H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}$*

Then, following the setting in Algorithm 1, the regret of DR-CCE-LSI is of order $\mathcal{O}\big(dH\min\{H, \frac{1}{\min\{\sigma_i\}}\}\sqrt{K}\big)$ with probability at least $1-\delta$.

To clarify, any d' dimensional subset $\mathcal{B}$ in R^d with $d' \le d$ means that, there should be exactly d' linearly independent vectors in R^d, to represent any vectors in $\mathcal{B}$ with linear combinations.

In Corollary 5.3, the structural assumption on the feature mapping ϕ carries significant intuitive implications. A non-degenerate feature mapping ensures that, when projecting from a high-dimensional space R^m to R^d, the representation fully utilizes the information in R^d without collapsing into a lower-dimensional subspace. Moreover, the condition $\sum_{j=1}^d\phi(s^j,\boldsymbol{a}^j)\phi(s^j,\boldsymbol{a}^j)^T \ge c\cdot\mathbf{I}$ guarantees that the feature mapping effectively explores the state-action space. This assumption is independent of the transition kernel P and policy π, which is not controllable during online learning process. In addition, we can reconstruct any tabular Markov game with linear MDP assumption, i.e. let $\phi(s,\boldsymbol{a}) = \mathbf{e}_{s\boldsymbol{a}} \in R^{|\mathcal{S}|\times|\mathcal{A}|}$ and the $(s,\boldsymbol{a})^{th}$ element of $\mu_h(s') \in R^{|\mathcal{S}|\times|\mathcal{A}|}$ be $P_h^0(s'|s,\boldsymbol{a})$. And verify that the feature mapping ϕ in the tabular case satisfies both assumptions in Corollary 5.3, implying applicability of our assumption to a broader range of Markov game instances.

Discussion on upper bound A recent study by (Liu, Wang, and Xu 2024) applies a variance-weighted ridge regression approach in the single-agent setting, obtaining an upper bound of $\mathcal{O}\big(dH\min\{H,\frac{1}{\sigma}\}\sqrt{K}\big)$. This is complemented by an information-theoretical lower bound of order $\Omega\big(dH^{1/2}\min\{H,\frac{1}{\sigma}\}\sqrt{K}\big)$. Our upper bound in the general-sum Markov game matches that of the single-agent case and is minimax optimal in terms of the feature dimension d. Notably, in online setting with linear function approximation, our algorithm is the first to achieve minimaxity in terms of feature dimension. For example, in non-robust counterpart, the minimax rate is $\sqrt{d^2H^3K}$, while the best result in online linear Markov game has an upper of $\sqrt{d^3H^4K}$ for two two-player zero sum game and $\sqrt{d^3H^5K}$ for multi-player general sum game. For more details, please refer to the discussion in the Appendix.

6 Simulation Study

We conduct numerical experiments to illustrate the effectiveness of our proposed algorithm, DR-CCE-LSI, to achieve robust equilibrium under model uncertainty. And compare it with the state-of-the-art algorithm NQOVI, which aims at solving non-robust online linear Markov game (Cisneros-Velarde and Koyejo 2023). The simulated linear Markov game consists of 5 states, two players, with a horizon length $H=3$. All the states are designed specifically to highlight the conflict of interest and environmental perturbation, such that a robust, stable equilibrium is crucial for good performance. For example, when considering personal interest only, player one will seek to reach state s_1 for reward maximization, while player two prefers reaching state s_2. The state s_f, is a self-absorbing fail state satisfying the minimum value assumption. The parameter $0\le\rho\le 1$ affects the transition probability of entering the fail state from s_1 and s_2. Therefore, when our nominal transition kernel has $\rho=0$, a larger value of uncertainty level ρ in the target Markov game implies a higher chance of transitioning to the fail state. As a result, the performance of any policy learned from an algorithm that does not consider robustness will degrade with the increase of uncertainty level, while any robust algorithms will present mild performance degradation. For more details of the experiment setup, please refer to the Appendix.

We can see from Figure 2 that, with increasing uncertainty level between the nominal Markov game and target Markov game, our algorithm's performance is significantly better than NQOVI, the state-of-the-art in online linear Markov game that does not account for robustness. This experiment's results validate the success of our algorithm in addressing the sim-to-real gap in our simulated linear Markov game example.

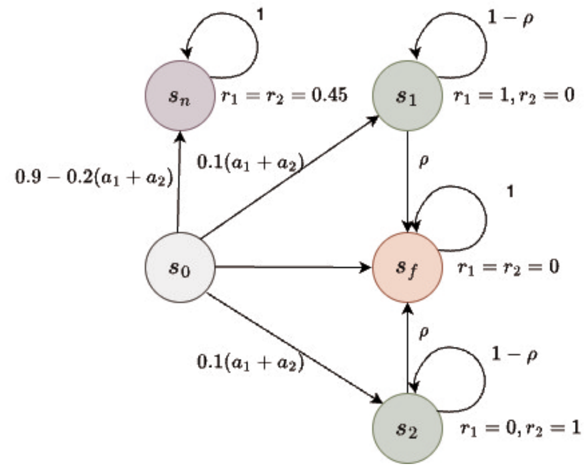

Figure 1: Constructed simulated Markov game

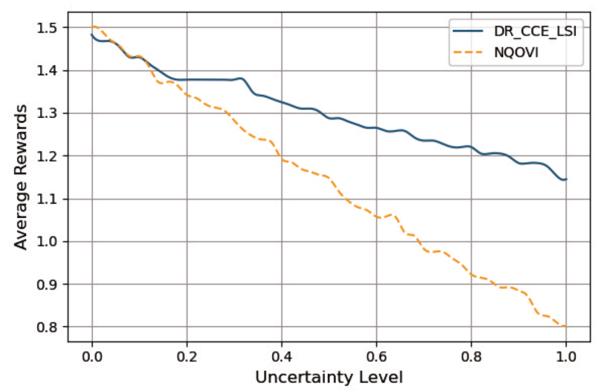

Figure 2: Comparison of the average reward under different uncertainty level.

7 Conclusion

In this paper, we identify the hardness result for online learning in Markov game, with theoretical validation. To address this issue, we adopt the minimal value assumption, and provide an intuitive explanation of this assumption by Proposition 4.3. Then, we propose a novel algorithm, DR-CCE-LSI, with an instance-dependent upper bound. By further illustration of impossibility of learning and studying an explicit relation between the structure properties of feature mapping ϕ, transition kernel and the instance dependent upper bound, our algorithm is provably sample-efficient in finding an $\epsilon-$approximate robust CCE (by online to batch conversion) with the regret of order $\mathcal{O}\big(dH\min\{H,\frac{1}{\min\{\sigma_i\}}\}\sqrt{K}\big)$. Simulation studies are conducted to validate the effectiveness of our algorithm in dealing with the sim-to-real gap.

Building on our findings, future work will focus on refining our upper bound, particularly w.r.t the horizon length H, to align with the information-theoretical lower bound. One possible approach is to incorporate variance-weighted ridge regression into our algorithm. However, applying a variance weighted analysis framework to Markov game setting is highly non-trivial, both theoretically and algorithmically, since the fundamental requirement of learning a monotonic value function in this framework is not accessible in multi agent case. More fine-grained analysis and algorithm design are required for further improvement in the upper bound.

Acknowledgments

We are grateful to the anonymous reviewers for their insightful feedback and valuable suggestions that helped improve this paper. Yuanyuan Lin's research was partially supported by the Hong Kong Research Grants Council (No.14304523) and Direct Grants for Research, The Chinese University of Hong Kong.

References

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

Berner, C.; Brockman, G.; Chan, B.; Cheung, V.; Dębiak, P.; Dennison, C.; Farhi, D.; Fischer, Q.; Hashme, S.; Hesse, C.; et al. 2019. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*.

Blanchet, J.; Lu, M.; Zhang, T.; and Zhong, H. 2023. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. volume 36, 66845–66859.

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

Cisneros-Velarde, P.; and Koyejo, S. 2023. Finite-sample guarantees for nash Q-learning with linear function approximation. In *Uncertainty in Artificial Intelligence*, 424–432. PMLR.

Clavier, P.; Pennec, E. L.; and Geist, M. 2023. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*.

Cui, Q.; Zhang, K.; and Du, S. 2023. Breaking the curse of multiagents in a large state space: Rl in markov games with independent linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, 2651–2652. PMLR.

Dai, Y.; Cui, Q.; and Du, S. S. 2024. Refined sample complexity for markov games with independent linear function approximation. *arXiv preprint arXiv:2402.07082*.

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

He, J.; Zhao, H.; Zhou, D.; and Gu, Q. 2023. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, 12790–12822. PMLR.

He, J.; Zhou, D.; and Gu, Q. 2021. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, 4171–4180. PMLR.

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

Jiang, Y.; Zhang, T.; Ho, D.; Bai, Y.; Liu, C. K.; Levine, S.; and Tan, J. 2021. Simgan: Hybrid simulator identification for domain adaptation via adversarial reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2884–2890. IEEE.

Jiao, Y.; and Li, G. 2024. Minimax-Optimal Multi-Agent Robust Reinforcement Learning. *arXiv preprint arXiv:2412.19873*.

Jin, C.; Liu, Q.; and Yu, T. 2022. The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*, 10251–10279. PMLR.

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

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

Koos, S.; Mouret, J.-B.; and Doncieux, S. 2012. The transferability approach: Crossing the reality gap in evolutionary robotics. *IEEE Transactions on Evolutionary Computation*, 17(1): 122–145.

Li, L.-F.; Zhao, P.; and Zhou, Z.-H. 2024. Near-Optimal Dynamic Regret for Adversarial Linear Mixture MDPs. *arXiv preprint arXiv:2411.03107*.

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

Liu, Z.; Bai, Q.; Blanchet, J.; Dong, P.; Xu, W.; Zhou, Z.; and Zhou, Z. 2022. Distributionally Robust Q-Learning. In *International Conference on Machine Learning*, 13623–13643. PMLR.

Liu, Z.; Wang, W.; and Xu, P. 2024. Upper and Lower Bounds for Distributionally Robust Off-Dynamics Reinforcement Learning. *arXiv preprint arXiv:2409.20521*.

Liu, Z.; and Xu, P. 2024. Distributionally robust off-dynamics reinforcement learning: Provable efficiency with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, 2719–2727. PMLR.

Lu, M.; Zhong, H.; Zhang, T.; and Blanchet, J. 2024. Distributionally Robust Reinforcement Learning with Interactive Data Collection: Fundamental Hardness and Near-Optimal Algorithm. arXiv:2404.03578.

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

Ma, X.; Liang, Z.; Blanchet, J.; Liu, M.; Xia, L.; Zhang, J.; Zhao, Q.; and Zhou, Z. 2022. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*.

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

Nash, J. F. 2024. Non-cooperative games. In *The Foundations of Price Theory Vol 4*, 329–340. Routledge.

Nilim, A.; and El Ghaoui, L. 2005. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5): 780–798.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Panaganti, K.; Xu, Z.; Kalathil, D.; and Ghavamzadeh, M. 2022. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35: 32211–32224.

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

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

Shi, L.; Gai, J.; Mazumdar, E.; Chi, Y.; and Wierman, A. 2024a. Breaking the Curse of Multiagency in Robust Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2409.20067*.

Shi, L.; Li, G.; Wei, Y.; Chen, Y.; Geist, M.; and Chi, Y. 2023. The Curious Price of Distributional Robustness in Reinforcement Learning with a Generative Model. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 79903–79917. Curran Associates, Inc.

Shi, L.; Mazumdar, E.; Chi, Y.; and Wierman, A. 2024b. Sample-Efficient Robust Multi-Agent Reinforcement Learning in the Face of Environmental Uncertainty. *arXiv preprint arXiv:2404.18909*.

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

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

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

Wang, B.; Yan, Y.; and Fan, J. 2021. Sample-efficient reinforcement learning for linearly-parameterized mdps with a generative model. *Advances in neural information processing systems*, 34: 23009–23022.

Wang, H.; Shi, L.; and Chi, Y. 2024. Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*.

Wang, S.; Si, N.; Blanchet, J.; and Zhou, Z. 2023a. A finite sample complexity bound for distributionally robust q-learning. In *International Conference on Artificial Intelligence and Statistics*, 3370–3398. PMLR.

Wang, Y.; Liu, Q.; Bai, Y.; and Jin, C. 2023b. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, 2793–2848. PMLR.

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

Yang, L.; and Wang, M. 2019. Sample-optimal parametric q-learning using linearly additive features. In *International conference on machine learning*, 6995–7004. PMLR.

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

Zanette, A.; Brandfonbrener, D.; Brunskill, E.; Pirotta, M.; and Lazaric, A. 2020. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, 1954–1964. PMLR.

Zhang, K.; Sun, T.; Tao, Y.; Genc, S.; Mallya, S.; and Basar, T. 2020. Robust multi-agent reinforcement learning with model uncertainty. *Advances in neural information processing systems*, 33: 10571–10583.

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