# SAMPLE-EFFICIENT DISTRIBUTIONALLY ROBUST MULTI-AGENT REINFORCEMENT LEARNING VIA ONLINE INTERACTION

**Anonymous authors**Paper under double-blind review

#### **ABSTRACT**

Well-trained multi-agent systems can fail when deployed in real-world environments due to model mismatches between the training and deployment environments, caused by environment uncertainties including noise or adversarial attacks. Distributionally Robust Markov Games (DRMGs) enhance system resilience by optimizing for worst-case performance over a defined set of environmental uncertainties. However, current methods are limited by their dependence on simulators or large offline datasets, which are often unavailable. This paper pioneers the study of online learning in DRMGs, where agents learn directly from environmental interactions without prior data. We introduce the *Multiplayer Optimistic Robust Nash Value Iteration (MORNAVI)* algorithm and provide the first provable guarantees for this setting. Our theoretical analysis demonstrates that the algorithm achieves low regret and efficiently finds the optimal robust policy for uncertainty sets measured by Total Variation divergence and Kullback-Leibler divergence. These results establish a new, practical path toward developing truly robust multi-agent systems.

#### 1 Introduction

Multi-agent reinforcement learning (MARL), along with its stochastic game-based mathematical formulation (Shapley, 1953; Littman, 1994), has emerged as a cornerstone paradigm for intelligent multi-agent systems capable of complex, coordinated behavior. It provides the theoretical and algorithmic foundation for enabling multiple agents to learn, adapt, and make sequential decisions in shared, dynamic environments. Its practical impacts span from strategic gaming, where MARL agents have achieved superhuman mastery (Silver et al., 2016; Vinyals et al., 2019); autonomous transportation, where it is used to coordinate fleets of vehicles to navigate complex traffic scenarios (Shalev-Shwartz et al., 2016; Hua et al., 2024); and distributed robotics, where teams of robots learn to execute tasks (Lowe et al., 2017; Matignon et al., 2012).

Despite the remarkable progress in MARL, a fundamental and pervasive challenge severely restricts its reliable deployment in the physical world: the *Sim-to-Real* gap (Zhao et al., 2020; Peng et al., 2018). A standard pipeline of RL involves training extensively within a high-fidelity simulator and then deploying in practice, as training directly in the real world can be prohibitively expensive, time-consuming, or dangerously unsafe. However, any simulator inevitably fails to capture the full richness and complexity of the real world, omitting subtle physical effects, unpredictable sensor noise, unmodeled system dynamics, or latent environmental factors (Padakandla et al., 2020; Rajeswaran et al., 2016). Consequently, a policy that appears optimal within the clean confines of a simulation can prove to be brittle and perform poorly—or even fail catastrophically—when deployed into the noisy, unpredictable environment it was designed for.

This vulnerability to model mismatch is magnified exponentially in the multi-agent context: this uncertainty is amplified through a cascading feedback loop of agent interactions. A minor, unmodeled perturbation that affects one agent can cause it to deviate from its expected behavior. This deviation alters the environment for its peers, who in turn must adapt their policies. Their adaptations further change the dynamics for all other agents, including the one first affected. This can trigger a chain of unpredictable responses, destabilizing the collective strategy and leading to a highly non-stationary

learning environment far more volatile than that caused by strategic adaptation alone (Papoudakis et al., 2019; Canese et al., 2021; Wong et al., 2023). The entire multi-agent system becomes fragile, as the intricate inter-agent dependencies act as amplifiers for even the smallest model inaccuracies.

To inoculate MARL agents against such environmental uncertainty, the framework of Distributionally Robust Markov Games (DRMGs) offers a principled and powerful solution (Zhang et al., 2020; Kardeş et al., 2011). Rather than trusting a single, nominal model of the environment (the simulator), the DRMG approach embraces a principle of pessimism. It defines an uncertainty set of plausible environment models centered around the nominal one. The agents' goal is to maximize the worst-case expected returns across the entire uncertainty set. This robust optimization strategy yields two profound benefits. First, it provides a formal performance guarantee: if the true environment lies within the uncertainty set, the policy's performance is guaranteed to be no worse than the optimized worst-case value. Second, it acts as a powerful regularizer, forcing agents to discover simpler and more generalizable policies that are inherently less sensitive to minor perturbations, thereby enhancing generalization even to environments outside the specified set (Vinitsky et al., 2020; Abdullah et al., 2019; Liu et al., 2025).

However, despite its theoretical appeal, the current body of research on DRMGs is built upon assumptions that create a critical disconnect from the realities of many high-stakes applications. The prevailing algorithmic frameworks fall into two main categories: those that assume access to a generative model (Shi et al., 2024b; Jiao & Li, 2024), which is tantamount to having a perfect, queryable oracle or simulator, and those designed for the offline setting (Li et al., 2025; Blanchet et al., 2023), which presuppose the existence of a large, static, and sufficiently comprehensive dataset collected beforehand. These assumptions are untenable in precisely the domains where robustness is most crucial. Consider applications in autonomous systems (Demontis et al., 2022) or personalized healthcare (Alaa Eldin, 2023; Lu et al., 2021). In these settings, creating a high-fidelity simulator is often impossible, and pre-collecting a dataset that covers all critical scenarios is infeasible. Agents have no choice but to learn online, through direct, sequential interaction with the complex and unknown real world. In this online paradigm, data is not a free commodity to be sampled at will; it is earned through experience, where every action has a real cost and naive exploration can lead to severe or irreversible outcomes. This necessitates a new class of algorithms that can navigate the exploration-exploitation tradeoff under the additional burden of worst-case environmental uncertainty.

We aim for robustness that survives contact with reality: agents must cope with misspecification while learning purely from experience. Without simulators or sizable offline datasets, existing approaches struggle to bridge theory and practice. This shortfall clarifies the gap we address and motivates our central question of our work: *How to design a provably effective online algorithms for distributionally robust Markov games?* 

In this paper, we answer the above question by designing a model-based online algorithm for DRMGs and providing corresponding theoretical guarantees. Our contributions are summarized as follows.

Hardness in Online DRMGs: We first revealed the inherent hardness of online learning in DRMGs. Specifically, we showed that the online learning can suffer from the support shifting issue, where the support of the worst-case kernel is not fully covered by the support of the nominal environment, by constructing a hard instance that achieve an  $\Omega(K\min\{H,\prod_i A_i\})$ -regret for any algorithm. Moreover, we use another example to show that even without the support shifting issue, the regret can still have a minimax lower bound of  $\Omega(\sqrt{K\prod_i A_i})$ . Here, K is the number of iteration episodes, H is the DRMG horizon, and  $\prod_i A_i$  is the size of the joint action space. These results directly imply the hardness of online learning, comparing to other well-posed learning schemes including generative model (Shi et al., 2024a; Jiao & Li, 2024) or offline learning (Li et al., 2025).

A Framework for Online Robust MARL: We introduce f-MORNAVI, a novel model-based meta-algorithm designed specifically for online learning in DRMGs. Our framework pioneers a dual approach that synergizes the *pessimism* required for robust optimization with the *optimism* essential for provably efficient online exploration. At its core, f-MORNAVI learns the nominal environment model from online interactions and then incorporates a carefully constructed, data-driven bonus term,  $\beta$ . This bonus term is uniquely tailored to the geometry of the chosen uncertainty set, guiding exploration while guaranteeing that the learned policy is robust to worst-case model perturbations. We further present two concrete instantiations of our framework for uncertainty sets defined by Total Variation (TV) distance and Kullback-Leibler (KL) divergence.

**Near-Optimal Regret Bounds for Online DRMGs:** We establish the first known theoretical guarantees for online learning in general-sum DRMGs by providing rigorous, high-probability regret bounds for our algorithms. The regret measures the performance gap between our algorithm and an optimal robust policy, thus formally characterizing the sample complexity needed to solve the DRMG. We futher prove that our algorithms converge to an  $\epsilon$ -optimal robust policy with high sample efficiency (see Corollary 1). Our results are significant as they are the first to demonstrate that finding a robust equilibrium in a general-sum DRMG is achievable in a sample-efficient manner through online interaction, without requiring a simulator or a pre-collected dataset.

# 2 PROBLEM FORMULATION

#### 2.1 DISTRIBUTIONALLY ROBUST MARKOV GAMES

A Distributionally Robust Markov Game (DRMG) can be specified as  $\mathcal{MG}_{rob} = \{\mathcal{M}, \mathcal{S}, \mathcal{A}, H, \{\mathcal{P}_i\}_{i \in \mathcal{M}}, r\}$ , where  $\mathcal{M} = \{1, ..., m\}$  is the set of m agents,  $\mathcal{S} = \{1, 2, ..., S\}$  denotes the finite state space,  $\mathcal{A}$  denotes the joint action space for all agents as  $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$ , where  $\mathcal{A}_i = \{1, 2, ..., A_i\}$  being the action space of agent i, H denotes the horizon length. We consider non-stationary DRMGs, i.e., r is the reward function:  $r = \{r_{i,h}\}_{1 \leq i \leq m, 1 \leq h \leq H}$  with  $r_{i,h} : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$ . Specifically, for any  $(i,h,s,\mathbf{a}) \in \mathcal{M} \times [H] \times \mathcal{S} \times \mathcal{A}$ ,  $r_{i,h}(s,\mathbf{a})$  is the immediate (deterministic) reward received by the i-th agent in state s when the joint action profile is  $\mathbf{a}$ . The major difference between a DRMG and a standard Markov game is the transition kernel. Instead of having a fixed transition kernel, agents in a DRMG maintain their own uncertainty sets of transition kernels  $\mathcal{P}_i$ , to capture the potential environment uncertainties in their perspective. At each step, the environment does not transit following a fixed transition kernel, instead, it transits following an arbitrary kernel from the uncertainty set.

In this work, we mainly consider uncertainty sets specified by f-divergence (Sason & Verdú, 2016). Drawing inspiration from the rectangularity condition in robust single-agent RL (Iyengar, 2005; Wiesemann et al., 2013; Zhou et al., 2021; Shi et al., 2023), and following standard DRMG studies (Shi et al., 2024b;a; Zhang et al., 2020), we consider the agent-wise (s, a)-rectangular uncertainty set, due to its computational tractability. Namely, for each agent i, the DRMG specify an uncertainty set  $\mathcal{P}_i$ , which is independently defined over all horizons, states, and joint actions:

$$\mathcal{P}_{i} = \bigotimes_{(h,s,\mathbf{a})\in[H]\times\mathcal{S}\times\mathcal{A}} \mathcal{P}_{i,h,f}^{\rho_{i}}(s,\mathbf{a}), \tag{1}$$

where  $\otimes$  denotes the Cartesian product. At step h, if all agents take a joint action  $\mathbf{a_h}$  at the state  $s_h$ , each agent anticipates that the transition kernel is allowed to be chosen arbitrarily from the prescribed uncertainty set  $\mathcal{P}_{i,h,f}^{\rho_i}(s_h,\mathbf{a_h})$ . Here, the uncertainty set  $\mathcal{P}_{i,h,f}^{\rho_i}(s,\mathbf{a})$  is constructed centered on a nominal kernel  $P^*: \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ :

**Definition 1** (f-Divergence Uncertainty Set). The f-divergence uncertainty set is defined as:

$$\mathcal{P}_{i,h,f}^{\rho_i}(s,\mathbf{a}) = \left\{ P_h \in \Delta(\mathcal{S}) : f\left(P_h, P_h^{\star}(\cdot|s,\mathbf{a})\right) \le \rho_i \right\},\,$$

where the f-divergence is defined as  $f\left(P_h,\ P_h^\star(\cdot|s, \boldsymbol{a})\right) = \sum\limits_{s' \in \mathcal{S}} f\left(\frac{P_h(s')}{P_h^\star(s'|s, \boldsymbol{a})}\right) P_h^\star(s'|s, \boldsymbol{a}).$ 

The f-divergence uncertainty sets with different f have been extensively studied in distributionally robust RL (Clavier et al., 2023; Shi et al., 2023; Panaganti et al., 2022; Yang et al., 2022; Wang et al., 2024e; Zhang et al., 2025). In this work, we focus on TV and KL-divergence.

**Robust Value Functions.** For a DRMG, each agent aims to maximize its own worst-case performance over all possible transition kernels in its own (possibly different) prescribed uncertainty set. The strategy of agent i taking actions is captured by a policy  $\pi_i = \{\pi_{i,h} : \mathcal{S} \to \Delta(\mathcal{A}_i)\}_{h=1}^H$ . Since the immediate rewards and transition kernels are determined by the joint actions, the worst-case performance of the i-th agent over its own uncertainty set  $\mathcal{P}_i$  is determined by a joint policy  $\pi = \{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}_{h=1}^H$ , which we refer to as the robust value function  $V_{i,h}^{\pi,\rho_i}$  and the robust

Q-function  $Q_{i,h}^{\pi,\rho_i}$ , for an initial state s and initial action  ${\bf a}$ :

$$Q_{i,h}^{\pi,\rho_i}(s,\mathbf{a}) \triangleq \inf_{\tilde{P} \in \mathcal{P}_i} \mathbb{E}_{\pi,\tilde{P}} \left[ \sum_{t=h}^{H} r_{i,t}(s_t,\mathbf{a}_t) \mid s_h = s, \mathbf{a}_h = \mathbf{a} \right], V_{i,h}^{\pi,\rho_i}(s) \triangleq \sum_{\mathbf{a}} \pi(\mathbf{a}|s) Q_{i,h}^{\pi,\rho_i}(s,\mathbf{a}),$$

where the expectation is taken over the randomness of the joint policy  $\pi$  and the kernel  $\tilde{P}$ .

**Solutions to DRMGs.** Due to different objectives, the goal of a DRMG is to achieve some notions of equilibrium (Fudenberg & Tirole, 1991). We begin by formalizing the best-response policy.

For any given joint policy  $\pi$ , we use  $\pi_{-i}$  to represent the policies of all agents excluding the *i*-th agent. The agent *i*'s best response policy to  $\pi_{-i}$ ,  $\pi_i^{\dagger,\rho_i}(\pi_{-i})$ , is the policy that maximizes its own robust value function, at the give step h and state s:

$$\pi_i^{\dagger,\rho_i}(\pi_{-i}) \triangleq \arg \max_{\pi_i' \in \Delta(\mathcal{A}_i)} V_{i,h}^{(\pi_{-i} \times \pi_i'),\rho_i}(s). \tag{2}$$

The corresponding robust value function is denoted as

$$V_{i,h}^{\dagger,\pi_{-i},\rho_i}(s) \triangleq \max_{\pi_i' \in \Delta(\mathcal{A}_i)} V_{i,h}^{\pi_i' \times \pi_{-i},\rho_i}(s). \tag{3}$$

As noted, the objective in a DRMG is to compute an equilibrium policy (Fudenberg & Tirole, 1991): each agent's policy is a best response to the others, so no single agent can improve its robust value by deviating while the rest remain fixed. Standard notions of equilibrium include *robust Nash Equilibrium (NE)*, *robust Coarse Correlated Equilibrium (CCE)*, and *robust Correlated Equilibrium (CE)* (all of them exist (Blanchet et al., 2023)). A DRMG aims to find some approximated equilibrium:

**Robust**  $\varepsilon$ -NE. A product policy  $\pi \in \Delta(A_1) \times \cdots \times \Delta(A_m)$  is an *robust-\varepsilon NE* if for any  $s \in S$ :

$$\mathrm{gap}_{\mathrm{NE}}(\pi,s) \triangleq \max_{i \in \mathcal{M}} \left\{ V_{i,1}^{\dagger,\pi_{-i},\rho_i}(s) - V_{i,1}^{\pi,\rho_i}(s) \right\} \leq \varepsilon.$$

Robust NE ensures that, the agent i's policy induced by the NE is a best response policy to the remaining agents' joint policy (up to  $\epsilon$ ), thus no agent can improve its worst-case performance—evaluated over its own uncertainty set  $\mathcal{P}_i$ —by unilaterally deviating from the NE.

**Robust**  $\varepsilon$ **-CCE.** A (possibly correlated) joint policy  $\pi \in \Delta(\mathcal{A})$  is an *robust-\varepsilon CCE* if for any  $s \in \mathcal{S}$ :

$$\operatorname{gap}_{\operatorname{CCE}}(\pi,s) \triangleq \max_{i \in \mathcal{M}} \left\{ V_{i,1}^{\dagger,\pi_{-i},\rho_i}(s) - V_{i,1}^{\pi,\rho_i}(s) \right\} \leq \varepsilon.$$

Robust CCE relaxes the notion of NE by allowing for potentially correlated policies, while still ensuring that no agent has an incentive to unilaterally deviate from it.

**Robust**  $\varepsilon$ -CE. A joint policy  $\pi \in \Delta(A)$  is an *robust-\varepsilon CE* if for any  $s \in S$ :

$$\mathrm{gap}_{\mathrm{CE}}(\pi,s) \triangleq \max_{i \in \mathcal{M}} \left\{ \max_{\phi \in \Phi_i} V_{i,1}^{\phi \diamond \pi, \rho_i}(s) - V_{i,1}^{\pi,\rho_i}(s) \right\} \leq \varepsilon.$$

Here, a strategy modification  $\phi \triangleq \{\phi_{h,s}\}_{(h,s)\in [H]\times\mathcal{S}}$  for player i is a set of  $[H]\times\mathcal{S}$  functions from  $\mathcal{A}_i$  to itself. Let  $\Phi_i$  denote the set of all possible strategy modifications for player i. Given a joint policy  $\pi$ , applying a modification  $\phi$  yields a new joint policy  $\phi \diamond \pi$ , which matches  $\pi$  everywhere except that at each state s and timestep s, player s action s is replaced by s.

Online Learning in DRMGs. We consider online learning in DRMGs, aiming to compute equilibria  $\{\mathsf{NASH},\mathsf{CCE},\mathsf{CE}\}\$  via interaction with the nominal environment  $P^\star$  over  $K\in\mathbb{N}$  episodes. Each episode starts from  $s_1^k$ , proceeds with a policy  $\pi^k$  chosen from experience, and ends with an update for the next round. We use *robust regret* as our performance metric, which compares the learned outcome to the target equilibrium in the presence of model error.

**Definition 2** (Robust Regret). Let  $\pi^k$  be the execution policy in the  $k^{th}$  episode. After a total of K episodes, the corresponding robust regret is defined as

$$\operatorname{Regret}_{\{\mathsf{NASH},\mathsf{CCE},\mathsf{CE}\}}(K) = \sum_{k=1}^K \operatorname{gap}_{\{\mathsf{NASH},\mathsf{CCE},\mathsf{CE}\}}(\pi^k,s_1^k).$$

Notably, if an algorithm has a sub-linear regret, it achieves a robust equilibrium as  $K \to \infty$ .

217218

219

220

221

242243244245

246

247 248

249250

251

253254

255256

257258

259

260

261

262

263

264

265

266

267

268

269

# 3 OPTIMISTIC ROBUST NASH VALUE ITERATION

We then present Multiplayer Optimistic Robust Nash Value Iteration for f-Divergence Uncertainty Set (f-MORNAVI), a meta-algorithm for episodic, finite-horizon DRMGs with interactive data collection. f-MORNAVI handles general f-divergences, with emphasis on KL and TV.

```
222
             Algorithm 1: f-MORNAVI
223
               1: Input: Uncertainty level \rho_i > 0 for all i \in \mathcal{M}.
224
              2: Initialize: Dataset \mathbb{D} = \emptyset
225
              3: for episode k = 1, \dots, K do
226
                   * Nominal Transition Estimation *
                         Compute the transition kernel estimator \widehat{P}_h^k(s, \mathbf{a}, s') as given in eq. 4.
227
                        Optimistic Robust Planning *
228
                         Set \overline{V}_{H+1}^{k,\rho_i}(\cdot) = \underline{V}_{H+1}^{k,\rho_i}(\cdot) = 0 for all i \in \mathcal{M}. for step h = H, \dots, 1 do
229
              5:
230
              6:
                              For all (s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A} and i \in \mathcal{M}, update \overline{Q}_{i,h}^{k,\rho_i}(s, \mathbf{a}) [eq. 5] and \underline{Q}_{i,h}^{k,\rho_i}(s, \mathbf{a}) [eq. 6].
231
              7:
232
                              For all s \in \mathcal{S}, update \pi_h^k(\cdot|s) by eq. 7.
              8:
                              For all s \in \mathcal{S} and i \in \mathcal{M}, update \overline{V}_{i,h}^{k,\rho_i}(s) and \underline{V}_{i,h}^{k,\rho_i}(\cdot) by eq. 8.
              9:
             10:
235
                   * Execution of policy and data collection *
236
                         Receive initial State s_1^k \in \mathcal{S}
             11:
237
                         for step h = 1, \dots, H do
             12:
238
                              Take action \mathbf{a}_h^k \sim \pi_h^k(\cdot \mid s_h^k), observe reward r_h(s_h^k, \mathbf{a}_h^k) and next state s_{h+1}^k.
             13:
239
             14:
                         Set \mathbb{D} = \mathbb{D} \cup \{(s_h^k, \mathbf{a}_h^k, s_{h+1}^k)\}_{h=1}^H.
240
             15:
241
```

#### 3.1 ALGORITHM DESIGN

Our algorithm has the following three stages.

17: **Output:** Return policy  $\pi^{\text{out}} = \{\pi^k\}_{k=1}^K$ .

Stage 1: Nominal Transition Estimation (Line 4). At the start of each episode  $k \in [K]$ , we maintain an estimate of the nominal kernel  $P^*$  using the historical data  $\mathbb{D} = \{(s_h^\tau, \mathbf{a}_h^\tau, s_{h+1}^\tau)\}_{\tau=1,h=1}^{k-1,H}$  collected from past interactions with the training environment. Specifically, f-MORNAVI updates the empirical transition kernel for each tuple  $(h, s, \mathbf{a}, s') \in [H] \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$  as follows:

$$\widehat{P}_h^k(s'|s,\mathbf{a}) = \frac{N_h^k(s,\mathbf{a},s')}{N_h^k(s,\mathbf{a})} (\text{if } N_h^k(s,\mathbf{a}) > 0), \text{ and } \widehat{P}_h^k(s'|s,\mathbf{a}) = \frac{1}{|\mathcal{S}|} (\text{if } N_h^k(s,\mathbf{a}) = 0), \quad (4)$$

where  $N_h^k(s,\mathbf{a},s')$  and  $N_h^k(s,\mathbf{a})$ , are calculated on the current dataset  $\mathbb D$  by  $N_h^k(s,\mathbf{a},s')=\sum_{s=1}^{k-1}\mathbf 1\{(s_h^\tau,\mathbf a_h^\tau,s_{h+1}^\tau)=(s,\mathbf a,s')\}$ , and  $N_h^k(s,\mathbf a)=\sum_{s'\in\mathcal S}N_h^k(s,\mathbf a,s')$ . Note that we adopt a model-based approach that estimates transition kernels. Although this leads to higher memory consumption, model-free DRMGs are inherently challenging due to the non-linearity of worst-case expectation w.r.t. nominal kernels, which makes model-free estimators biased or sample-inefficient (Liu et al., 2022; Wang et al., 2023c; 2024d; Zhang et al., 2025).

Stage 2: Optimistic Robust Planning (Lines 5–10). The f-MORNAVI constructs the episode policy  $\pi^k$  via optimistic robust planning based on the empirical model  $\widehat{P}^k$ . This involves estimating an upper bound on the robust value function, following the principle of Upper-Confidence-Bound (UCB) methods, which are well-established in online vanilla RL (Auer & Ortner, 2010; Azar et al., 2017; Zanette & Brunskill, 2019; Zhang et al., 2021b; Ménard et al., 2021; Zhang et al., 2024), and this optimism encourages exploration of less-visited state—action pairs.

To this end, f-MORNAVI maintains a bonus term at each episode k, capturing the gap between the robust value function under  $\widehat{P}^k$  and that under the true model. This bonus is added to the robust

Bellman estimate to ensure its optimism. Specifically, for each  $(h, s, \mathbf{a}) \in [H] \times \mathcal{S} \times \mathcal{A}$ , we set

$$\overline{Q}_{i,h}^{k,\rho_i}(s,\boldsymbol{a}) = \min \left\{ r_{i,h}(s,\boldsymbol{a}) + \sigma_{\widehat{\mathcal{P}}_{i,h,f}^{\rho_i}(s,\boldsymbol{a})} [\overline{V}_{i,h+1}^{k,\rho_i}] + \beta_{i,h,f}^{k}(s,\boldsymbol{a}), H \right\}.$$
 (5)

$$\underline{Q}_{i,h}^{k,\rho_i}(s,\boldsymbol{a}) = \max \left\{ r_{i,h}(s,\boldsymbol{a}) + \sigma_{\widehat{\mathcal{P}}_{i,h-f}^{\rho_i}(s,\boldsymbol{a})}[\underline{V}_{i,h+1}^{k,\rho_i}] - \beta_{i,h,f}^{k}(s,\boldsymbol{a}), 0 \right\},$$
(6)

here,  $\sigma_{\mathcal{P}}[V] = \inf_{P \in \mathcal{P}} \mathbb{E}_P[V]$  is the support function of V over the uncertainty set  $\mathcal{P}$ , and can be calculated through its dual representation (see Lemma 1);  $\widehat{\mathcal{P}}_{i,h,f}^{\rho_i}$  is the uncertainty set centered at  $\widehat{P}^k$  from eq. 4:  $\widehat{\mathcal{P}}_{i,h,f}^{\rho_i}(s,\mathbf{a}) = \left\{ P_h \in \Delta(\mathcal{S}) : f\left(P_h,\widehat{P}_h^k(\cdot|s,\mathbf{a})\right) \leq \rho_i \right\}$ .

Each of these estimates in eq. 5 and eq. 6 are based on estimated robust Bellman operators (see Appendix C for details) and a bonus term  $\beta_{i,h,f}^k(s,\mathbf{a}) \geq 0$ . The bonus term is constructed (we will discuss the construction later) to ensure the estimation becomes a confidence interval of the true robust value function, i.e.,  $Q_{i,h}^{\dagger,\pi-i,\rho_i}(s,\mathbf{a}) \in [\underline{Q}_{i,h}^{k,\rho_i}(s,\mathbf{a}),\overline{Q}_{i,h}^{k,\rho_i}(s,\mathbf{a})]$ , with high probability.

**EQUILIBRIUM subroutine (Line 8).** Given robust Q-function estimates  $\underline{Q}_{i,h}^{k,\rho_i}(s,\mathbf{a})$  and  $\overline{Q}_{i,h}^{k,\rho_i}(s,\mathbf{a})$  for  $i\in\mathcal{M}$  at step h, the sub-routine EQUILIBRIUM  $\in\{\mathsf{NASH},\mathsf{CCE},\mathsf{CE}\}$  finds a corresponding equilibrium  $\pi_h^k(\cdot|s)$  for the matrix-form game with pay-off matrices  $\{\overline{Q}_{i,h}^{k,\rho_i}(s,\cdot)\}_{i\in\mathcal{M}}$ :

$$\pi_h^k(\cdot|s) \leftarrow \text{EQUILIBRIUM}\Big(\Big\{\overline{Q}_{i,h}^{k,\rho_i}(s,\cdot)\Big\}_{i\in\mathcal{M}}\Big).$$
 (7)

Note that finding a NE can be PPAD-hard (Daskalakis et al., 2009), but computing CE or CCE remains tractable in polynomial time (Liu et al., 2021).

We then update the estimation of  $V_h^{\dagger,\pi_{-i},\rho}$  as

$$\overline{V}_{i,h}^{k,\rho_i}(s) = \mathbb{E}_{\boldsymbol{a} \sim \pi^k(\cdot|s)} \left[ \overline{Q}_{i,h}^{k,\rho_i}(s, \mathbf{a}) \right] \quad \text{ and } \quad \underline{V}_{i,h}^{k,\rho_i}(s) = \mathbb{E}_{\boldsymbol{a} \sim \pi^k(\cdot|s)} \left[ \underline{Q}_{i,h}^{k,\rho_i}(s, \mathbf{a}) \right]. \tag{8}$$

Note that while the lower estimate in eq. 6 does not influence policy execution directly, it plays a crucial role in constructing valid exploration bonuses and ensuring strong theoretical guarantees. By leveraging both upper and lower bounds, the algorithm performs optimistic robust planning, enabling structured, uncertainty-aware exploration that balances exploration, exploitation, and robustness.

Stage 3: Execution of Policy and Data Collection (Lines 11–17). After evaluating the policy  $\{\pi_h^k\}_{h=1}^H$  for episode k, the learner takes action based on  $\pi_h^k$  and observes the reward  $r_h(s_h^k, \mathbf{a}_h^k)$  and next state  $s_{h+1}^k$ , which get appended to the historical dataset collected till episode k-1.

# 4 HARDNESS OF ONLINE LEARNING

In this section, we aim to discuss the inherent hardness of online learning in DRMGs from two aspects: (1) When there is the support shift issue, no MARL algorithm can obtain a sub-linear regret on a certainty DRMG; (2) Even if there is no support shift issue, there exists a DRMG such that any online algorithm suffers from the curse of multi-agency. This is a separation between DRMGs with interactive data collection and generative model/offline data, and also between DRMGs with non-robust MGs, showing the inherent challenges of online DRMGs.

#### 4.1 HARDNESS WITH SUPPORT SHIFT

Support shift (Lu et al., 2024) refers to the case that the support of the worst-case transition kernel is not covered by the support of the nominal kernel. It can happen when, for instance, the uncertainty set is defined through TV. It will result in a challenge that, for those states that is not covered by the nominal kernel, there is no data available, so that the agent can never learn the optimal robust policy efficiently. Specifically, we derive the following result to illustrate the hardness.

**Theorem 1.** There exists a TV-DRMG, such that any online learning algorithm suffers the following regret lower bound:

$$\inf_{\mathcal{ALG}} \mathbb{E}[\mathsf{Regret}_{\mathsf{NASH}}(K)] \geq \Omega \Big( \rho K \cdot \min\{H, \prod_{i \in \mathcal{M}} A_i\} \Big).$$

Our construction is deferred to Example 1 in Appendix. This regret bound is linear in the number of episodes K, creating a combinatorial explosion that makes the problem information-theoretically intractable. Moreover, our result shows that when the game horizon H is large enough, the minimax lower bound depends on the joint action space, showing the hardness of online learning compared to generative models and offline settings.

#### 4.2 HARDNESS WITHOUT SUPPORT SHIFT

We then illustrate the hardness of online DRMGs when there is no support shift. Note that when the uncertainty set is defined through, e.g., KL divergence, the worst-case support will be covered by the nominal one, so there will not be any support shift. However, we construct another example to show that, even without the support shift, the online learning can still be challenging and inefficient.

**Theorem 2** (Lower Bound for Robust Learning without Support Shift). There exists a DRMG, such that any learning algorithm suffers the following cumulative regret lower bound over K episodes:

$$\inf_{\mathcal{ALG}} \mathbb{E}[\mathsf{Regret}_{\mathsf{NASH}}(K)] \geq \Omega\Big(\sqrt{K\prod_{i \in \mathcal{M}} A_i}\Big).$$

Our construction is in Example 2 in Appendix. This result illustrates that, even without any support shift, some hard instance can require at least  $\Omega\left(\sqrt{K\prod_i A_i}\right)$  regret. Our result hence suggests that the dependence on the joint action space may be inevitable in online robust learning, which suffer from the curse of multi-agency.

# 5 THEORETICAL GUARANTEES

We then develop the theoretical results of our algorithm under both TV and KL sets.

# 5.1 REGRET BOUND FOR TOTAL VARIATION

As discussed in Section 4, no efficient algorithm can be expected due to the support shifting issue. We hence adopt a standard fail-state assumption (Lu et al., 2024; Liu et al., 2024) to ensure the worst-case kernel support will be covered by the nominal one, bypassing the issue.

**Assumption 1** (Failure States). For any agent i, there exists an (agent-specified) set of failure states  $\mathcal{S}_f^i \subseteq \mathcal{S}$ , such that  $r_i(s, \boldsymbol{a}) = 0$ , and  $P_h^{\star}(s'|s, \boldsymbol{a}) = 1$ ,  $\forall \boldsymbol{a} \in \mathcal{A}, \forall s \in \mathcal{S}_f^i, \forall s' \in \mathcal{S}_f^i$ .

This assumption is only for TV case. Assumption 1 is a standard assumption in single-agent robust RL studies (Panaganti et al., 2022; Lu et al., 2024), and we adapt it to multi-agent cases.

We then present our threotical guarantees.

**Theorem 1** (Upper bound of TV-MORNAVI). Denote  $\rho_{\min} := \min_{i \in \mathcal{M}} \rho_i$ . For any  $\delta \in (0, 1)$ ,

we set 
$$\beta_{i,h,f}^{k}(s,\mathbf{a})$$
 as  $\sqrt{\frac{c_{1}\iota \operatorname{Var}_{\hat{P}_{h}^{k}(\cdot|s,\mathbf{a})}\left[\overline{V}_{i,h+1}^{k,\rho_{i}}+\underline{V}_{i,h+1}^{k,\rho_{i}}\right]}{N_{h}^{k}(s,\mathbf{a})\vee 1}} + \frac{c_{2}H^{2}S\iota}{\sqrt{N_{h}^{k}(s,\mathbf{a})\vee 1}} + \frac{2\mathbb{E}_{\hat{P}_{h}^{k}(\cdot|s,\mathbf{a})}\left[\overline{V}_{i,h+1}^{k,\rho_{i}}-\underline{V}_{i,h+1}^{k,\rho_{i}}\right]}{H}} + \frac{1}{2}\mathcal{E}_{\hat{P}_{h}^{k}(\cdot|s,\mathbf{a})}\left[\overline{V}_{i,h+1}^{k,\rho_{i}}-\underline{V}_{i,h+1}^{k,\rho_{i}}\right]}{H} + \frac{1}{2}\mathcal{E}_$ 

 $\frac{1}{\sqrt{K}}$ , where  $\iota = \log \left( S^2(\prod_{i=1}^m A_i) H^2 K^{3/2}/\delta \right)$  and  $c_1, c_2$  are absolute constants. Then under Assumption 1, for EQUILIBRIUM being one of {NASH, CE, CCE}, with probability at least  $1 - \delta$ , the regret of our TV-MORNAVI algorithm can be bounded as:

$$\operatorname{Regret}_{\{\mathsf{NASH},\mathsf{CCE},\mathsf{CE}\}}(K) = \tilde{\mathcal{O}}\left(\sqrt{\min\left\{\rho_{\min}^{-1},H\right\}H^2SK\Big(\prod_{i\in\mathcal{M}}A_i\Big)}\right),$$

where  $f(K) = \tilde{\mathcal{O}}(g(K))$  means  $f(K) \leq \textbf{Poly}(\log(K)) \cdot g(K)$  for sufficiently large K and some polynomial of  $\log(K)$ .

#### 5.2 REGRET BOUND FOR KL-DIVERGENCE

We then study the regret bound of KL-divergence set. As discussed, KL set is free from supporting issue hence no additional assumption is required. Our regret bound result is as follows.

**Theorem 2.** For any  $\delta$ , set  $\beta_{i,h,f}^k(s,\boldsymbol{a})$  in KL-DRMG as  $\frac{2c_fH}{\rho_i}\sqrt{\frac{\iota}{\left(N_h^k(s,\boldsymbol{a})\vee 1\right)}\widehat{P}_{\min,h}^k(s,\boldsymbol{a})}+\sqrt{\frac{1}{K}}$ , where  $\widehat{P}_{\min,h}^k(s,\boldsymbol{a})=\min_{s'\in\mathcal{S}}\{\widehat{P}_h^k(s'|s,\boldsymbol{a}):\widehat{P}_h^k(s'|s,\boldsymbol{a})>0\}$ ,  $\iota=\log\left(S^2(\prod_{i=1}^mA_i)H^2K^{3/2}/\delta\right)$ , and  $c_f$  is an absolute constant. Then for Equilibrium being one of {NASH, CE, CCE}, with probability at least  $1-\delta$ , it holds that

$$\operatorname{Regret}_{\{\mathsf{NASH},\mathsf{CCE},\mathsf{CE}\}}(K) = \tilde{\mathcal{O}}\left(\sqrt{H^4 \exp(2H^2)KS\Big(\prod_{i \in \mathcal{M}} A_i\Big)\Big(\rho_{\min}^2 P_{\min}^{\star}\Big)^{-1}}\right), \tag{9}$$

here,  $P_{\min}^{\star} \triangleq \min_{(s, \boldsymbol{a}, s', h): P_h(s'|s, \boldsymbol{a}) > 0} P(s'|s, \boldsymbol{a})$  is the smallest positive entry of the nominal kernel.

We note that  $\exp(H)$  term in KL results can be replaced by  $P_{\min}^{-1}$  (Panaganti & Kalathil, 2022; Blanchet et al., 2023), and both of these terms are inevitable.

#### 5.3 SAMPLE COMPLEXITY

As a direct corollary, we derive the sample complexity to learn an  $\varepsilon$ -equilibrium. Using a standard online-to-batch conversion (Cesa-Bianchi et al., 2001), we have the following results.

**Corollary 1** (Sample Complexity). With probability at least  $1-\delta$ , and under the settings of Theorem 1 and Theorem 2, the number of samples required to find an  $\epsilon$ -approximate equilibrium is bounded as:

$$KH = \begin{cases} \tilde{\mathcal{O}}\left(\epsilon^{-2}\min\left\{\rho_{\min}^{-1}, H\right\} H^3 S\left(\prod_{i \in \mathcal{M}} A_i\right)\right), & \textit{for TV-DRMG} \\ \tilde{\mathcal{O}}\left(\epsilon^{-2} H^5 \exp(2H^2) S\left(\prod_{i \in \mathcal{M}} A_i\right) \left(\rho_{\min}^2 P_{\min}^{\star}\right)^{-1}\right), & \textit{for KL-DRMG} \end{cases}.$$

Our results hence implies that, despite the inherent hardness of online learning in DRMGs, our algorithm is able to learn an equilibrium with efficient sample complexity. As we shall discussed in the next section, our complexity bounds are near-optimal (expect the term  $\prod_{i \in \mathcal{M}} A_i$ ), which hence implies the efficiency of our method.

### 6 COMPARISON WITH PRIOR WORKS AND DISCUSSION

We then compare our results with prior works (the detailed Comparisons are shown in Table 1).

A substantial body of research on DRMGs has focused on two primary settings: (i) generative model setting, where the agents can freely sample from all state-action pairs (Shi et al., 2024a;b; Jiao & Li, 2024); (ii) offline setting, which relies on a comprehensive, pre-collected dataset (Blanchet et al., 2023; Li et al., 2025). As we discuss in Section 4, both of these avoid exploration and are therefore easier than the online regime we consider. Despite this added difficulty, our algorithm attains complexities comparable to those reported for the generative and offline settings.

For both uncertainty sets, our results match or improve upon previous results and the minimax lower bound in all parameters except for the action-product term,  $\prod_i A_i$ , under the generative model setting. In the offline setting, if the dataset is generated uniformly, the convergence coefficients  $C_{u/p}^{\star}$  from (Li et al., 2025; Blanchet et al., 2023) introduce an additional  $\prod_i A_i$  term into the sample complexity. Consequently, our results also match or surpass the offline complexity in all parameter dependence. This raises an important open question:

# Can any online DRMG learning algorithm (or even under generative model settings) overcome the curse of multi-agency and eliminate the dependence on $\prod_i A_i$ ?

While some works (Shi et al., 2024a; Jiao & Li, 2024; Li et al., 2025; Ma et al., 2023) have achieved independence from  $\prod_i A_i$ , it remains unclear whether these improvements are applicable to general DRMGs. Specifically, the results in (Shi et al., 2024a) and (Jiao & Li, 2024) are developed for special uncertainty sets with desirable properties. For instance, the fictitious TV uncertainty set in (Shi et al., 2024a) allows the global transition kernel to be estimated from a single agent's local information; And robust RL under contamination models is known to be equivalent to a non-robust problem with a

Table 1: Comparison with prior results.  $C_{u/p}^{\star}$  are coverage coefficients for offline learning. In (Li et al., 2025),  $f(H,\rho) = (H\rho - 1 + (1-\rho)^H)/\rho^2$ . The  $\exp(H)$  term in KL results can be replaced by  $P_{\min}^{-1}$  directly (Panaganti & Kalathil, 2022; Blanchet et al., 2023).

G 44	Γ	1
Setting &	Uncertainty Set	Sample Complexity
Algorithm		r r r r
Generative	TV	$\tilde{\mathcal{O}}\left(\epsilon^{-2}H^3S(\prod_{i\in\mathcal{M}}A_i)\min\left\{\rho_{\min}^{-1},H\right\}\right)$
(Shi et al., 2024b)		$\mathcal{O}\left(\mathcal{O}\left(\Pi_{i} \otimes \Pi_{i}\right) \prod_{i \in \mathcal{M}} \Pi_{i}\right) \prod_{i \in \mathcal{M}} \left(\mathcal{O}_{\min}, \Pi_{i}\right)$
Generative	Contamination	$\tilde{\mathcal{O}}(\epsilon^{-2}H^3S(\sum_{i\in\mathcal{M}}A_i)\min\left\{\rho_{\min}^{-1},H\right\})$
(Jiao & Li, 2024)		
Generative	TV (fictitious)	$\tilde{\mathcal{O}}\left(s^{-4} H^6 S(\sum A) \min \left(s^{-1} H\right)\right)$
(Shi et al., 2024a)		$\tilde{\mathcal{O}}\left(\epsilon^{-4}H^6S(\sum_{i\in\mathcal{M}}A_i)\min\left\{\rho_{\min}^{-1},H\right\}\right)$
Offline	KL	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\rho_{\min}^{-2}C_u^{\star}H^4\exp(H)S^2(\prod_{i\in\mathcal{M}}A_i)\right)$
(Blanchet et al., 2023)	TV	$\tilde{\mathcal{O}}\Big(\epsilon^{-2}C_u^{\star}H^4S^2(\prod_{i\in\mathcal{M}}A_i)\Big)$
Offline	TV	$\tilde{O}(-2C) \times II^4C(\sum^m A) \cdot (f(II))$
(Li et al., 2025)		$\tilde{\mathcal{O}}\left(\epsilon^{-2}C_p^{\star}H^4S(\sum_{i=1}^m A_i)\min\left\{f(H,\rho),H\right\}\right)$
Online	KL	$ ilde{\mathcal{O}}(\epsilon^{-2}H^5S(\max_i\{A_i\})^2)$ (with an oracle)
(Ma et al., 2023)		$O(\epsilon - \Pi^* S(\max_i \{A_i\}))$ (with all oracle)
Online	TV	$\tilde{\mathcal{O}}\left(\epsilon^{-2}H^3S(\prod_{i\in\mathcal{M}}A_i)\min\left\{\rho_{\min}^{-1},H\right\}\right)$
(Our work)	KL	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\rho_{\min}^{-2}(P_{\min}^{\star})^{-1}H^{5}\exp(2H^{2})S\left(\prod_{i\in\mathcal{M}}A_{i}\right)\right)$
Generative		
Lower bound	TV	$\Omega\left(\epsilon^{-2}H^3S(\max_{i\in\mathcal{M}}A_i)\min\left\{\rho_{\min}^{-1},H\right\}\right)$
(Shi et al., 2024b)		( " ( min ) )
		I .

specific discount factor (Wang et al., 2023a). And the improvement in the offline setting is attributed to the benefits of the coverage coefficient.

The only online method (which also breaks the curse of multi-agency) is presented in (Ma et al., 2023). However, their algorithm relies on additional assumptions about uncertainty sets and a powerful oracle. This oracle is required to provide an  $\epsilon$ -accurate estimation of the worst-case performance,  $\sigma_{\mathcal{P}_i}[V]$  (see Theorem 12 of (Ma et al., 2023)), without any need for exploration. A central challenge in the analysis of robust learning algorithms is precisely quantifying this estimation error, as demonstrated in works like (Shi et al., 2023; Xu et al., 2023; Panaganti & Kalathil, 2022; Liu & Xu, 2024). By assuming the existence of such an oracle, they bypass this core challenge, which significantly reduces their sample complexity. Moreover, their results need additional assumptions on the radius  $\rho$ . For instance, it is assumed that  $\rho \leq \frac{P_{\min}^*}{\mu}$ , whereas ours do not require any of them.

Therefore, it is still uncertain whether the complexity reduction in these papers is a blessing of their specific uncertainty set structures, the properties of offline coverage coefficients, or the use of an estimation oracle. Furthermore, based on our discussion in Section 4, it is not clear whether the minimax lower bound for online DRMGs is independent of the size of the joint action space. We, therefore, leave the exploration of this direction, including whether practical relaxations and techniques can avoid it, for future work.

#### 7 CONCLUSION

In this paper, we introduced the Multiplayer Optimistic Robust Nash Value Iteration (MORNAVI) algorithm, pioneering the study of online learning in DRMGs. Our work provides the first provable guarantees for this challenging setting, demonstrating that MORNAVI achieves low regret and efficiently identifies optimal robust policies for TV-divergence and KL-divergence uncertainty sets. This research establishes a practical path toward developing truly robust multi-agent systems that learn directly from environmental interactions. Despite the inherent hardness of online DRMGs, our algorithm achieves complexity results comparable to generative model and offline settings. This work also highlights a critical open question: whether online DRMG learning algorithms can overcome the curse of multi-agency and eliminate the dependence on the joint action space size. Future work will explore this fundamental challenge to advance the scalability of robust MARL.

# REFERENCES

- Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian Zhang, and Jun Wang. Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*, 2019.
  - Baraa Alaa Eldin. Why applying deep reinforcement learning in healthcare is hard. https://medium.com/@baraa.alaa.eldin/why-applying-deep-reinforcement-learning-in-healthcare-is-hard-ffc6e05ab7ca, 2023. Accessed: 2025-07-28.
  - Peter Auer and Ronald Ortner. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.
  - Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.
  - Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *International conference on machine learning*, pp. 263–272. PMLR, 2017.
  - Kishan Panaganti Badrinath and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *Proc. International Conference on Machine Learning (ICML)*, pp. 511–520. PMLR, 2021.
  - Yu Bai and Chi Jin. Provable Self-Play Algorithms for Competitive Reinforcement Learning. In *International conference on machine learning*, pp. 551–560. PMLR, 2020.
  - 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.
  - Alexander Bukharin, Yan Li, Yue Yu, Qingru Zhang, Zhehui Chen, Simiao Zuo, Chao Zhang, Songan Zhang, and Tuo Zhao. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 68121–68133, 2023.
  - Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.
  - Lorenzo Canese, Gian Carlo Cardarilli, Luca Di Nunzio, Rocco Fazzolari, Daniele Giardino, Marco Re, and Sergio Spanò. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences*, 11(11):4948, 2021.
  - Nicoló Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the Generalization Ability of On-Line Learning Algorithms. *Advances in neural information processing systems*, 14, 2001.
  - Zixiang Chen, Dongruo Zhou, and Quanquan Gu. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In Sanjoy Dasgupta and Nika Haghtalab (eds.), *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 227–261. PMLR, 29 Mar–01 Apr 2022. URL https://proceedings.mlr.press/v167/chen22d.html.
  - Pierre Clavier, Erwan Le Pennec, and Matthieu Geist. Towards Minimax Optimality of Model-based Robust Reinforcement Learning. *arXiv preprint arXiv:2302.05372*, 2023.
  - Qiwen Cui, Kaiqing Zhang, and Simon Du. 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*, pp. 2651–2652. PMLR, 2023.
  - Constantinos Daskalakis. On the complexity of approximating a nash equilibrium. *ACM Transactions on Algorithms (TALG)*, 9(3):1–35, 2013.

- Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.
- Ambra Demontis, Maura Pintor, Luca Demetrio, Kathrin Grosse, Hsiao-Ying Lin, Chengfang Fang, Battista Biggio, and Fabio Roli. A survey on reinforcement learning security with application to autonomous driving, 2022. URL https://arxiv.org/abs/2212.06123.
  - Xiaotie Deng, Ningyuan Li, David Mguni, Jun Wang, and Yaodong Yang. On the complexity of computing markov perfect equilibrium in general-sum stochastic games. *National Science Review*, 10(1):nwac256, 2023.
- Jing Dong, Jingwei Li, Baoxiang Wang, and Jingzhao Zhang. Online policy optimization for robust mdp. *arXiv preprint arXiv:2209.13841*, 2022.
  - Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.
  - Songtao Feng, Ming Yin, Yu-Xiang Wang, Jing Yang, and Yingbin Liang. Improving sample efficiency of model-free algorithms for zero-sum markov games. *arXiv preprint arXiv:2308.08858*, 2023.
  - Arlington M Fink. Equilibrium in a stochastic *n*-person game. *Journal of science of the hiroshima university, series ai (mathematics)*, 28(1):89–93, 1964.
    - Drew Fudenberg and Jean Tirole. Game theory. MIT press, 1991.
    - Debamita Ghosh, George K. Atia, and Yue Wang. Provably near-optimal distributionally robust reinforcement learning in online settings, 2025. URL https://arxiv.org/abs/2508.03768.
  - Amy Greenwald, Keith Hall, Roberto Serrano, et al. Correlated q-learning. In *ICML*, volume 3, pp. 242–249, 2003.
  - Songyang Han, Sanbao Su, Sihong He, Shuo Han, Haizhao Yang, Shaofeng Zou, and Fei Miao. What is the solution for state-adversarial multi-agent reinforcement learning? *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=HyqSwNhM3x.
  - Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.
  - Sihong He, Songyang Han, Sanbao Su, Shuo Han, Shaofeng Zou, and Fei Miao. Robust multi-agent reinforcement learning with state uncertainty, 2023. URL https://arxiv.org/abs/2307.16212.
  - Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.
  - Min Hua, Dong Chen, Xinda Qi, Kun Jiang, Zemin Eitan Liu, Quan Zhou, and Hongming Xu. Multi-agent reinforcement learning for connected and automated vehicles control: Recent advancements and future prospects, 2024. URL https://arxiv.org/abs/2312.11084.
  - Garud N Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.
  - Yuchen Jiao and Gen Li. Minimax-optimal multi-agent robust reinforcement learning. *arXiv preprint arXiv:2412.19873*, 2024.
    - Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4868–4878, 2018.
    - 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.

- Yujia Jin, Vidya Muthukumar, and Aaron Sidford. The complexity of infinite-horizon general-sum stochastic games. arXiv preprint arXiv:2204.04186, 2022.
  - 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.
  - Na Li, Yuchen Jiao, Hangguan Shan, and Shefeng Yan. Provable memory efficient self-play algorithm for model-free reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.
  - Na Li, Zewu Zheng, Wei Ni, Hangguan Shan, Wenjie Zhang, and Xinyu Li. Sample efficient robust offline self-play for model-based reinforcement learning. Manuscript, OpenReview preprint, 2025. URL https://openreview.net/forum?id=31XZjsir0e.
  - Shihui Li, Yi Wu, Xinyue Cui, Honghua Dong, Fei Fang, and Stuart Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proc. Conference on Artificial Intelligence (AAAI)*, volume 33, pp. 4213–4220, 2019.
  - Yan Li, Guanghui Lan, and Tuo Zhao. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022.
  - Zhipeng Liang, Xiaoteng Ma, Jose Blanchet, Jiheng Zhang, and Zhengyuan Zhou. Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*, 2023.
  - Jieyu Lin, Kristina Dzeparoska, Sai Qian Zhang, Alberto Leon-Garcia, and Nicolas Papernot. On the robustness of cooperative multi-agent reinforcement learning, 2020. URL https://arxiv.org/abs/2003.03722.
  - Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.
  - Michael L Littman and Csaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *ICML*, volume 96, pp. 310–318, 1996.
  - Michael L Littman et al. Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pp. 322–328, 2001.
  - Guangyi Liu, Suzan Iloglu, Michael Caldara, Joseph W Durham, and Michael M. Zavlanos. Distributionally robust multi-agent reinforcement learning for dynamic chute mapping. In *Proc. International Conference on Machine Learning (ICML)*, 2025.
  - Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *Proc. International Conference on Machine Learning (ICML)*, pp. 7001–7010. PMLR, 2021.
  - Zhishuai Liu and Pan Xu. Minimax Optimal and Computationally Efficient Algorithms for Distributionally Robust Offline Reinforcement Learning. *Advances in Neural Information Processing Systems*, 37:86602–86654, 2024.
  - Zhishuai Liu, Weixin Wang, and Pan Xu. Upper and lower bounds for distributionally robust off-dynamics reinforcement learning. *arXiv preprint arXiv:2409.20521*, 2024.
  - Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust Q-learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 13623–13643. PMLR, 2022.
  - Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 6379–6390, 2017.
  - Miao Lu, Han Zhong, Tong Zhang, and Jose Blanchet. Distributionally robust reinforcement learning with interactive data collection: Fundamental hardness and near-optimal algorithm. *The Thirty-eighth Annual Conference on Neural Information Processing Systemss*, 2024.

- MingYu Lu, Zachary Shahn, Daby Sow, Finale Doshi-Velez, and Li-wei H Lehman. Is deep reinforcement learning ready for practical applications in healthcare? A sensitivity analysis of duel-DDQN for hemodynamic management in sepsis patients. In *AMIA annual symposium proceedings*, volume 2020, pp. 773, 2021.
  - Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou. Decentralized robust v-learning for solving markov games with model uncertainty. *Journal of Machine Learning Research*, 24(371):1–40, 2023.
  - Xiaoteng Ma, Zhipeng Liang, Jose Blanchet, Mingwen Liu, Li Xia, Jiheng Zhang, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally Robust Offline Reinforcement Learning with Linear Function Approximation. *arXiv preprint arXiv:2209.06620*, 2022.
  - Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, 13(1):165–186, 2023.
  - Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.
  - Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization. *arXiv preprint arXiv:0907.3740*, 2009.
  - Pierre Ménard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. UCB Momentum Q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pp. 7609–7618. PMLR, 2021.
  - Arnab Nilim and Laurent El Ghaoui. Robust Control of Markov Decision Processes with Uncertain Transition Matrices. *Operations Research*, 53(5):780–798, 2005.
  - Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.
  - Sindhu Padakandla, Prabuchandran KJ, and Shalabh Bhatnagar. Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, 50(11):3590–3606, 2020.
  - Kishan Panaganti and Dileep Kalathil. Sample Complexity of Robust Reinforcement Learning with a Generative Model. In *International Conference on Artificial Intelligence and Statistics*, pp. 9582–9602. PMLR, 2022.
  - Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *arXiv preprint arXiv:2208.05129*, 2022.
  - Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.
  - Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In 2018 IEEE international conference on robotics and automation (ICRA), pp. 3803–3810. IEEE, 2018.
  - Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles. *arXiv preprint arXiv:1610.01283*, 2016.
  - Shyam Sundhar Ramesh, Pier Giuseppe Sessa, Yifan Hu, Andreas Krause, and Ilija Bogunovic. Distributionally robust model-based reinforcement learning with large state spaces, 2023. URL https://arxiv.org/abs/2309.02236.
  - Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.
  - Shai Shalev-Shwartz et al. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv* preprint arXiv:1610.03295, 2016.

- Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.
  - 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.
  - 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.
  - Laixi Shi, Jingchu Gai, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Breaking the curse of multiagency in robust multi-agent reinforcement learning. *arXiv preprint arXiv:2409.20067*, 2024a.
  - Laixi Shi, 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*, 2024b.
  - David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016. doi: 10.1038/nature16961.
  - 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.
  - Eugene Vinitsky, Yuqing Du, Kanaad Parvate, Kathy Jang, Pieter Abbeel, and Alexandre Bayen. Robust reinforcement learning using adversarial populations. *arXiv preprint arXiv:2008.01825*, 2020.
  - Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575:350 354, 2019. URL https://api.semanticscholar.org/CorpusID:204972004.
  - He Wang, Laixi Shi, and Yuejie Chi. Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*, 2024a.
  - Qiuhao Wang, Chin Pang Ho, and Marek Petrik. Policy gradient in robust mdps with global convergence guarantee. In *Proc. International Conference on Machine Learning (ICML)*, pp. 35763–35797. PMLR, 2023a.
  - Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A Finite Sample Complexity Bound for Distributionally Robust Q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3370–3398. PMLR, 2023b.
  - Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample Complexity of Variance-Reduced Distributionally Robust Q-Learning. *Journal of Machine Learning Research*, 25(341):1–77, 2024b.
  - Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. On the foundation of distributionally robust reinforcement learning, 2024c. URL https://arxiv.org/abs/2311.09018.
  - Yudan Wang, Yue Wang, Yi Zhou, Alvaro Velasquez, and Shaofeng Zou. Data-driven robust multi-agent reinforcement learning. In 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE, 2022.

- Yudan Wang, Shaofeng Zou, and Yue Wang. Model-free robust reinforcement learning with sample complexity analysis. In *Proc. International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024d.
  - Yue Wang and Shaofeng Zou. Online Robust Reinforcement Learning with Model Uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.
  - Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. Model-free robust average-reward reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, pp. 36431–36469. PMLR, 2023c.
  - Yue Wang, Zhongchang Sun, and Shaofeng Zou. A Unified Principle of Pessimism for Offline Reinforcement Learning under Model Mismatch. *Advances in Neural Information Processing Systems*, 37:9281–9328, 2024e.
  - Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov Decision Processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.
  - Annie Wong, Thomas Bäck, Anna V Kononova, and Aske Plaat. Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review*, 56(6):5023–5056, 2023.
  - Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 3674–3682. PMLR, 2020.
  - Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved Sample Complexity Bounds for Distributionally Robust Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 9728–9754. PMLR, 2023.
  - Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward Theoretical Understandings of Robust Markov Decision Processes: Sample Complexity and Asymptotics. *The Annals of Statistics*, 50(6): 3223–3248, 2022.
  - Wenhao Yang, Han Wang, Tadashi Kozuno, Scott M Jordan, and Zhihua Zhang. Robust markov decision processes without model estimation. *arXiv preprint arXiv:2302.01248*, 2023.
  - Andrea Zanette and Emma Brunskill. Tighter Problem-Dependent Regret Bounds in Reinforcement Learning without Domain Knowledge using Value Function Bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.
  - Chi Zhang, Zain Ulabedeen Farhat, George K. Atia, and Yue Wang. Model-free offline reinforcement learning with enhanced robustness. In *Proc. International Conference on Learning Representations (ICLR)*, 2025.
  - Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.
  - Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pp. 321–384, 2021a.
  - Runyu Zhang, Yang Hu, and Na Li. Soft Robust MDPs and Risk-Sensitive MDPs: Equivalence, Policy Gradient, and Sample Complexity. *arXiv preprint arXiv:2306.11626*, 2023.
  - Zihan Zhang, Xiangyang Ji, and Simon Du. Is Reinforcement Learning More Difficult Than Bandits? A Near-optimal Algorithm Escaping the Curse of Horizon. In *Conference on Learning Theory*, pp. 4528–4531. PMLR, 2021b.
  - Zihan Zhang, Yuxin Chen, Jason D Lee, and Simon S Du. Settling the sample complexity of online reinforcement learning. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 5213–5219. PMLR, 2024.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In 2020 IEEE symposium series on computational intelligence (SSCI), pp. 737–744. IEEE, 2020.

Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-Sample Regret Bound for Distributionally Robust Offline Tabular Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339. PMLR, 2021.

Ziyuan Zhou, Guanjun Liu, and Mengchu Zhou. A robust mean-field actor-critic reinforcement learning against adversarial perturbations on agent states. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14370–14381, October 2024. ISSN 2162-2388. doi: 10.1109/tnnls. 2023.3278715. URL http://dx.doi.org/10.1109/TNNLS.2023.3278715.