ONLINE ROBUST MULTI-AGENT LEARNING WITH GENERAL FUNCTION APPROXIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributionally robust Markov games (DRMGs) provide a principled framework for handling model misspecification in multi-agent reinforcement learning, but existing methods often rely on restrictive assumptions or scale poorly to large-scale problems. In this work, we study online multi-agent reinforcement learning in general-sum DRMGs with general function approximation and ϕ-divergence uncertainty sets. We propose RoMEX-ϕ, a model-free framework that integrates equilibrium-based exploration with dual fitted learning. By reformulating the robust multi-agent Bellman operator via a functional dual representation, our approach enables tractable worst-case value estimation from nominal data through a unified objective based on a centered empirical robust discrepancy. To quantify the learning complexity, we introduce the *robust Multi-Agent Decoupling Coefficient* (robust MADC) to characterize intrinsic exploration complexity under adversarial dynamics. With this notion, we further establish sublinear regret guarantees of our RoMEX-ϕ, which rely explicitly on robust MADC instead of the large state and action spaces, and hence significantly improve the scalability and efficiency of robust multi-agent reinforcement learning in large-scaled multi-agent systems.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) (Albrecht et al., 2024) provides a fundamental framework for modeling sequential decision-making with multiple interacting agents. It has enabled significant advances in domains such as strategic games (Silver et al., 2016; Vinyals et al., 2019), autonomous systems (Shalev-Shwartz et al., 2016; Hua et al., 2024), and multi-robot coordination (Lowe et al., 2017; Matignon et al., 2012). Despite this progress, a key obstacle to real-world deployment is the *sim-to-real gap* (Zhao et al., 2020; Peng et al., 2018): policies trained under nominal or simulated dynamics often degrade significantly under even small model mismatches due to unmodeled effects, sensor noise, or environmental variations (Padakandla et al., 2020; Rajeswaran et al., 2016). This challenge is significantly exacerbated in multi-agent systems due to the *intrinsic coupling of agent behaviors*. A small perturbation affecting one agent can propagate through strategic interactions, inducing cascading adaptations across agents and resulting in highly non-stationary and unstable dynamics (Papoudakis et al., 2019; Canese et al., 2021; Wong et al., 2023). Consequently, even minor model inaccuracies can destabilize the entire system, making robustness a fundamental requirement for MARL.

To address such uncertainty, Distributionally Robust Markov Games (DRMGs) provide a principled framework (Zhang et al., 2020; Kardeş et al., 2011; Blanchet et al., 2023). DRMGs model transition uncertainty via some divergence-based ambiguity set of transition probabilities, and adopt the Principle of Pessimism to consider the worst-case expected returns over the sets. This pessimistic formulation ensures baseline performance guarantees under model misspecification and acts as an implicit regularizer, promoting policies that generalize better across uncertain environments (Vinitsky et al., 2020; Abdullah et al., 2019; Liu et al., 2025). Although promising, existing learning algorithms for DRMGs are still falling short in practical deployments due to two limitations. Firstly, most methods are limited to idealized data generation settings, assuming either access to a generative model (Shi et al., 2024b; Jiao & Li, 2024) or comprehensive offline datasets (Blanchet et al., 2023; Li et al., 2025). In many real-world applications—such as autonomous systems (Demontis et al., 2022) and healthcare (Alaa Eldin, 2023; Lu et al., 2021)—assumptions of high-fidelity simulator or high-quality dataset can be unrealistic, and agents must instead learn online via directly interactions.

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

However, the online DRMG setting introduces a fundamentally harder problem: agents face an off-dynamic (Holla, 2021; Eysenbach et al., 2020) setting: they interact and collect data from the nominal/training environment, whereas their goal is to find equilibria under the worst case. Therefore, they must balance exploration and exploitation under worst-case dynamics, which can be significantly more challenging that the previous settings.

The second limitation of existing studies is their poor scalability. In DRMGs, it is recently shown that online learning algorithms inherently suffer from the *curse of multi-agency* (Farhat et al., 2026; Shi et al., 2024b), i.e., the sample complexity depends on the joint action space size, which becomes exponentially large in large-scale systems. To break this curse, either additional uncertainty set structures (Shi et al., 2024a; Jiao & Li, 2024) or a worst performance estimation oracle (Ma et al., 2023) are required. A more recent work (Zheng & Lin, 2026) attempts to improve the scalability of robust MARL by utilizing the linear function approximation technique. However, due to the restrictive linear function class, the DRMG is assumed to be linear, which can be restrictive and hard to verify in practice. On the other hand, general function approximation aims to approximate the value function through some general function classes (like neural networks), and is well understood in standard RL through intrinsic complexity measures such as Bellman rank and Bellman–Eluder dimension (Jin et al., 2021; 2022). However, these measures rely on a single Bellman operator, which breaks down in general-sum robust multi-agent systems. Therefore, it is still unclear how to design provable complexity measurements and sample-efficient algorithms for robust MARL. Taken together, these limitations expose a fundamental research gap:

Can we design a robust MARL algorithm for DRMGs without restrictive assumptions and with a scalable sample complexity that breaks the multi-agency curse?

In this paper, we provide an affirmative answer to the above question by developing a unified framework for online robust MARL with general function approximation, together with scalable sample-efficiency guarantees. Our main contributions are summarized as follows.

Scalable dual robust fitted learning for online robust MARL with general function approximation. We propose RoMEX-ϕ, the first algorithm for *purely online* learning in general-sum DRMGs under general function approximation (instead of restrictive linear functions) and ϕ-divergence uncertainty sets. Our approach integrates robustness and exploration *via a dual reformulation of the robust multi-agent Bellman operator*, enabling policy evaluation directly from nominal data. Specifically, the algorithm computes a functional dual minimizer for each candidate joint policy and plugs it into a fitted robust Bellman recursion, yielding a *dual robust fitted-learning procedure* that converts intractable worst-case updates into tractable empirical objectives. This formulation both approximates the worst-case Bellman operator and enables a *global, function-class-level quantification of uncertainty* via the induced robust Bellman residual. Unlike tabular methods relying on per-state–action bonuses, RoMEX-ϕ employs a global regularization mechanism—through centered empirical robust discrepancy and regularized payoffs—that penalizes poor Bellman consistency and guides coordinated exploration across agents, while remaining computationally efficient in large-scale settings.

Robust MADC: intrinsic complexity for robust multi-agent learning. We introduce the *robust Multi-Agent Decoupling Coefficient (robust MADC)*, $d_{\mathrm{MADC}}^{\mathrm{rob,max}}$, as the intrinsic complexity measure governing exploration and regret in online DRMGs with general function approximation. It quantifies how empirical robust Bellman errors—computed from nominal data—control value prediction under worst-case dynamics across interacting agents and policy-induced distributions. Unlike BE-dimension-based measures (Ghosh et al., 2025b), which are defined for a single Bellman operator, robust MADC captures the strategic coupling and equilibrium structure in general-sum DRMGs while accounting for the statistical mismatch between nominal learning and adversarial evaluation. This yields a unified characterization of function complexity, extending intrinsic-complexity theory to the robust multi-agent setting.

Sharp regret guarantees via robust intrinsic complexity. We then establish the first sublinear regret guarantee for online general-sum DRMGs with general function approximation, showing that the robust regret of RoMEX-ϕ scales as $\mathcal{O}\big(d_{\mathrm{MADC}}^{\mathrm{rob,max}}\sqrt{K} + Hd_{\mathrm{MADC}}^{\mathrm{rob,max}} + H(B_{\phi}^{\max})^2\sqrt{K\log(|\mathcal{F}||\mathcal{G}|KHm/\delta)} + \varepsilon_{\mathrm{dual}}^{\mathrm{approx}}\big)$. Our bound reveals two fundamental sources of difficulty: the *intrinsic exploration complexity*, characterized by the robust MADC $d_{\mathrm{MADC}}^{\mathrm{rob,max}}$, and the *robust Bellman estimation cost* incurred when learning the worst-case Bellman operator from

nominal interaction data, captured by $B_\phi^{\max}$ and the dual approximation error $\varepsilon_{\text{dual}}^{\text{approx}}$. Importantly, the regret has no explicit dependence on the state-space size $|\mathcal{S}|$ or the joint action-space size $|\mathcal{A}|$, thereby replacing the tabular dependence with intrinsic function-class complexity while explicitly accounting for the additional statistical cost of distributional robustness.

Numerical validation. We complement our theory with experiments on a scalable general-sum DRMG under total variation (TV) uncertainty (Appendix F). RoMEX-ϕ exhibits substantially greater resilience to transition shift than its non-robust counterpart and remains competitive with an exact tabular robust baseline, demonstrating the practical effectiveness of robust learning with function approximation.

2 Preliminaries and Problem Formulation

Robust MARL is generally formulated as a distributionally robust Markov game (DRMG). A finite-horizon n-player general-sum DRMG is specified as $\mathcal{MG}_{\text{rob}} = \big\{\mathcal{M}, \mathcal{S}, \mathcal{A}, H, P^\star, \{\mathcal{P}_i\}_{i\in\mathcal{M}}, \boldsymbol{r}, \rho_0\big\}$, where $\mathcal{M} = \{1, ..., m\}$ is the set of m agents, $\mathcal{S} = \{1, 2, \ldots, S\}$ 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 $\mathcal{A}_i = \{1, 2, \ldots, A_i\}$ is the action space of agent i, H denotes the horizon length, $P^\star$ is the nominal transition kernel, and $\boldsymbol{r} = \{r_i\}_{1\le i\le m}$ is the reward function where $r_i : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$. Agents in a DRMG maintain their own uncertainty sets of transition kernels $\mathcal{P}_i$, to capture the potential environment uncertainties from their perspective. At each step, the environment transitions follow an arbitrary kernel from the uncertainty set. We follow the standard *agent-wise* $(s, \boldsymbol{a})$*-rectangular uncertainty structure* (Shi et al., 2024b;a; Zhang et al., 2020; Farhat et al., 2026) due to its computational tractability (Iyengar, 2005; Wiesemann et al., 2013; Zhou et al., 2021; Shi et al., 2023)[1]. Specifically, for each agent i, $\mathcal{P}_i$ is defined independently across all horizons, states, and joint actions as $\mathcal{P}_i = \bigotimes_{(h,s,\mathbf{a})\in[H]\times\mathcal{S}\times\mathcal{A}} \mathcal{P}_{i,h,\phi}^{\sigma_i}(s, \mathbf{a})$, where $\otimes$ denotes the Cartesian product. We consider the ϕ-divergence uncertainty set standard in literature (Clavier et al., 2023; Shi et al., 2023; Panaganti et al., 2022; Yang et al., 2022a; Wang et al., 2024b; Zhang et al., 2025), that each local uncertainty set $\mathcal{P}_{i,h,\phi}^{\sigma_i}(s, \mathbf{a})$ is centered around a *nominal transition kernel* $P^\star$, through some ϕ-divergence, $\phi\big(P,\ P_h^\star(\cdot|s, \boldsymbol{a})\big) = \sum\limits_{s'\in\mathcal{S}} \phi\Big(P(s')\big/P_h^\star(s'|s, \boldsymbol{a})\Big) P_h^\star(s'|s, \boldsymbol{a})$:

Definition 1 (ϕ-Divergence Uncertainty Set)**.** The ϕ-divergence uncertainty set is defined as $\mathcal{P}_{i,h,\phi}^{\sigma_i}(s, \mathbf{a}) \triangleq \Big\{P \in \Delta(\mathcal{S}) : \phi\Big(P, P_h^\star(\cdot|s, \mathbf{a})\Big) \le \sigma_i, \text{Supp}(P) \subseteq \text{Supp}(P_h^\star(\cdot|s, \mathbf{a}))\Big\}$.

Support inclusion is essential for online robust RL because states outside the nominal support cannot be observed through online interaction, creating an information bottleneck that can lead to exponential regret (Lu et al., 2024; Ghosh et al., 2025b; Farhat et al., 2026). We consider TV, χ^2, and KL uncertainty sets, for which support inclusion holds automatically under χ^2 and KL, and under TV given Assumption 4.

Robust Value Functions. Considering a DRMG, we aim for each agent to maximize its performance against the worst-case transition kernels within its respective uncertainty set. Each agent follows its own policy defined as $\pi_i = \{\pi_{i,h} : \mathcal{S} \to \Delta(\mathcal{A}_i)\}_{h=1}^H$. The robust Q function for each agent $Q_{i,h}^{\pi,\sigma_i}$ for an initial state s and joint action $\boldsymbol{a}$ is defined as $Q_{i,h}^{\pi,\sigma_i}(s, \boldsymbol{a}) \triangleq \inf_{\tilde{P}\in\mathcal{P}_i} \mathbb{E}_{\pi,\tilde{\mathcal{P}}}\big[\sum_{t=h}^H r_{i,t}(s_t, \boldsymbol{a}_t)\big|s_h = s, \boldsymbol{a}_h = \boldsymbol{a}\big]$, where the expectation is taken over the randomness of the joint policy $\pi = \{\pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}_{h=1}^H$, and the kernel $\tilde{P}$ over its own uncertainty set, and one can similarly define the corresponding robust value function $V_{i,h}^{\pi,\sigma_i}(s)$ for an initial state s. For any joint policy π, the robust value functions to satisfy the robust Bellman equations (Iyengar, 2005)

$$Q_{i,h}^{\pi,\sigma_i}(s, \boldsymbol{a}) = r_{i,h}(s, \boldsymbol{a}) + \mathbb{E}_{\mathcal{P}_{i,h,\phi}^{\sigma_i}}[V_{i,h+1}^{\pi,\sigma_i}] \quad \text{and} \quad V_{i,h}^{\pi,\sigma_i}(s) = \mathbb{E}_{\boldsymbol{a}\sim\pi}[Q_{i,h}^{\pi,\sigma_i}(s, \boldsymbol{a})], \tag{1}$$

where $\mathbb{E}_{\mathcal{P}_{i,h,\phi}^{\sigma_i}(s,\boldsymbol{a})}[V_{i,h+1}^{\pi,\sigma_i}] \triangleq \inf_{P_h\in\mathcal{P}_i} \mathbb{E}_{s'\sim P_h}\Big[V_{i,h+1}^{\pi,\sigma_i}(s')\Big]$ is the worst-case expectation.

DRMG Equilibria. Due to potentially conflicting objectives among agents, the goal in a DRMG is to find an appropriate equilibrium policy under which no agent can improve its robust value through

[1]Robust MDPs without the rectangularity assumption can be NP-hard to solve (Wiesemann et al., 2013).

unilateral deviation (Fudenberg & Tirole, 1991). We consider three standard equilibrium notions adapted to the robust setting: *robust Nash Equilibrium (NE)*, *robust Coarse Correlated Equilibrium (CCE)*, and *robust Correlated Equilibrium (CE)*, whose existence has been established in prior work (Blanchet et al., 2023). We present the definition of robust NE below and defer those of robust CCE and CE to Appendix A. Since computing a Nash equilibrium is generally computationally intractable (Daskalakis et al., 2009), we also consider the more computationally tractable CCE and CE notions.

Definition 2 (ϵ-Robust NE)**.** Let $\pi \in \Delta(\mathcal{A}_1) \times \cdots \times \Delta(\mathcal{A}_m)$ be a product policy. It is an *ϵ-robust-NE* if $\mathrm{Gap}^{\mathrm{NE}}(\pi, s) \triangleq \max_{i \in \mathcal{M}} \left\{ V_{i,1}^{\dagger, \pi_{-i}, \sigma_i}(s_1) - V_{i,1}^{\pi, \sigma_i}(s_1) \right\} \leq \epsilon$.

In the online setting, agents need to learn from direct interaction with the nominal environment $P^\star$. The interaction will be deployed for K episodes, each starting from s_1^k, execute the policy π^k to collect new samples, and update the policy. To evaluate learning performance, we utilize the robust regret (Farhat et al., 2026), which measures the accumulated equilibrium gap.

Definition 3 (Robust Regret)**.** Let π^k be a policy executed at episode k. The robust regret over K episodes is defined as $\mathrm{Reg}^{\{\mathrm{NE,CE,CCE}\}}(K) \triangleq \sum_{k=1}^{K} \mathrm{Gap}^{\{\mathrm{NE,CE,CCE}\}}(\pi^k, s_1^k)$.

3 Robust MARL with Function Approximation

In this subsection, we discuss the structural and computational challenges of online DRMG with general function approximation, and present our approach to overcoming them.

Model-Free General Function Approximation. In DRMGs, large state–action spaces make tabular MARL methods computationally intractable (Farhat et al., 2026; Ma et al., 2023). To address this, we adopt a model-free function approximation framework, where each agent's robust action-value function is represented using a structured hypothesis class $\mathcal{F} = \bigotimes_{i \in \mathcal{M}} \mathcal{F}_i$, where $\mathcal{F}_i = \bigotimes_{h=1}^{H} \mathcal{F}_{i,h}$, where for each agent i and stage h, $\mathcal{F}_{i,h} = \{f_{i,h} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ consists of candidate functions (e.g., low-dimensional parametric models like neural networks) used to approximate the robust value functions. For any fixed agent i, we denote a candidate function as $f_i = \{f_{i,h}\}_{h=1}^{H}$, and denote the one-step robust Bellman backup under policy π as

$$\left[\mathcal{T}_{i,h}^{\pi, \phi, \sigma_i} f_{i,h+1}\right](s, \boldsymbol{a}) \triangleq r_{i,h}(s, \boldsymbol{a}) + \mathbb{E}_{\mathcal{P}_{i,h,\phi}^{\sigma_i}(s, \boldsymbol{a})}\left[V_{i,h+1}^{\pi, f}\right], \tag{2}$$

where the policy-averaged value function is given by $V_{i,h}^{\pi, f_i}(s') \triangleq \mathbb{E}_{\boldsymbol{a}' \sim \pi_h(\cdot|s')}[f_{i,h}(s', \boldsymbol{a}')]$. To enable statistically and computationally efficient learning with function approximation, prior work imposes *representation conditions* that characterize how the function class F interacts with the underlying (robust) MDP dynamics (Russo & Van Roy, 2013; Jiang et al., 2017; Sun et al., 2019; Wang et al., 2020; Jin et al., 2021; Panaganti et al., 2022). n standard RL, these include *realizability* and *completeness*. However, in the robust setting, realizability of the optimal robust value function is often overly restrictive due to the presence of worst-case expectations. For instance, in the linear function case (Zheng & Lin, 2026), the underlying DRMG is restrictively assumed to be linear to ensure realizability. Instead, following recent advances in robust RL (Ma et al., 2022; Liu & Xu, 2024a; Liu et al., 2024; Wang et al., 2024a), realizability is effectively relaxed and can be implicitly ensured through appropriate assumptions on the dual function class used to represent the robust Bellman operator. Therefore, in our setting, we adopt the standard *completeness* condition to guarantee that the function class remains closed under the robust Bellman backup:

Assumption 1 (Robust Completeness)**.** *Let π be any joint policy. For any $i \in \mathcal{M}$ and $f_{i,h+1} \in \mathcal{F}_{i,h+1}$, we have $\mathcal{T}_{i,h}^{\pi, \phi, \sigma_i} f_{i,h+1} \in \mathcal{F}_{i,h}$.*

This assumption ensures closure under the robust Bellman operator and directly implies robust realizability, eliminating the need for a separate realizability assumption (Ghosh et al., 2025b).

Dual Formulation of robust Bellman operator and functional optimization. Using standard duality (Yang et al., 2022b), the worst-case expectation in the robust Bellman operator is gven by

$$\mathbb{E}_{\mathcal{P}_{i,h,\phi}^{\sigma_i}(s, \boldsymbol{a})}[V_{i,h+1}^{\pi, f}] = \inf_{\lambda > 0, \nu \in \mathbb{R}} \left\{ \lambda \sigma_i - \nu + \lambda\, \mathbb{E}_{s' \sim P_h^\star}\left[\phi^\star\big((\nu - V_{i,h+1}^{\pi, f}(s'))/\lambda\big)\right] \right\}, \tag{3}$$

which yields the dual form of the robust Bellman operator

$$\left[\mathcal{T}_{i,h}^{\pi, \phi, \sigma_i} f_{i,h+1}\right](s, \boldsymbol{a}) = r_{i,h}(s, \boldsymbol{a}) - \inf_{\lambda, \nu} \mathbb{E}_{s' \sim P_h^\star}\left[\mathrm{Loss}_{i,h,\phi}(f_i; s, \boldsymbol{a}, s'; \lambda, \nu)\right], \tag{4}$$

where $\mathrm{Loss}_{i,h,\phi}$ denotes the pointwise dual integrand for a single next-state s' and joint action $\boldsymbol{a}$ for agent $i \in \mathcal{M}$: $\mathrm{Loss}_{i,h,\phi}(f_i; s, \boldsymbol{a}, s'; \lambda, \nu) \triangleq \lambda\sigma_i - \nu + \lambda\phi^\star\Big(-\big(V^{\pi,f}_{i,h+1}(s') - \nu\big)/\lambda\Big)$. While the dual reformulation removes the need to optimize over transition kernels, evaluating the resulting operator still involves solving an inner optimization over (λ, ν) for each state–action pair $(s, \boldsymbol{a})$. Consequently, the computational cost scales with the size of the state space, rendering direct evaluation of the robust Bellman operator impractical in large-scale settings. To address this, we adopt a *functional optimization* perspective (Panaganti et al., 2022; Ghosh et al., 2025b), replacing pointwise optimization with a single global problem over the state–action space. Let μ^π_h denote the state–joint action visitation distribution induced by policy π under the nominal kernel. We define the dual loss

$$\mathrm{LossDual}_{i,h,\phi}(g; f_i) = \mathbb{E}_{(s,\boldsymbol{a})\sim\mu^\pi_h}\Big[\mathbb{E}_{s'\sim P^\star_h}\mathrm{Loss}_{i,h,\phi}(f_i; s, \boldsymbol{a}, s'; g(s, \boldsymbol{a}))\Big], \tag{5}$$

where $g(s, \boldsymbol{a}) = (g^\lambda_i, g^\nu_i)$ denotes a dual function.[2] We can further show that minimizing the dual loss in equation 5 over $\mathcal{L}^1(\mu, \mathbb{R}^2)$ exactly recovers the robust Bellman operator in expectation under μ^π_h (refer to Lemma 2 in Appendix B for details). Thus, instead of solving an inner optimization for each $(s, \boldsymbol{a})$, we can optimize for a single dual function that captures worst-case transitions globally.

This observation allows us to replace the intractable pointwise optimization with a single functional optimization of the empirical dual loss of equation 5, $\widehat{\mathrm{LossDual}}_{i,h,\phi}(g; f_i)$. We further introduce a structured function class $\mathcal{G} = \bigotimes_{i,h}\mathcal{G}_{i,h}$, where each $\mathcal{G}_{i,h} = \{g_i = (g^\lambda_i, g^\nu_i) : \mathcal{S}\times\mathcal{A} \to \Theta_\phi \subseteq \mathbb{R}^2\}$, to replace $\mathcal{L}^1(\mu, \mathbb{R}^2)$ and to approximate the dual solution. We adopt the following assumption to quantify the approximation error (Panaganti et al., 2022; 2024; Ghosh et al., 2025b).

Assumption 2. *There exists a constant $\varepsilon^{\mathrm{approx}}_{\mathrm{dual}}$, such that for any agent $i \in \mathcal{M}$, horizon $h \in [H]$, function $f_i \in \mathcal{F}$, and policy π, $\inf_{g\in\mathcal{G}_i}\mathrm{LossDual}_{i,h,\phi}(g; f_i) - \inf_{g\in\mathcal{L}^1_i(\mu^\pi;\mathbb{R}^2)}\mathrm{LossDual}_{i,h,\phi}(g; f_i) \le \varepsilon^{\mathrm{approx}}_{\mathrm{dual}}$.*

Empirical dual minimization and operator approximation. Based on the discussion above, we can approximate the robust Bellman operator by first minimizing the empirical dual loss over $\mathcal{G}$:

$$\widehat{g_i}^f = \arg\min_{g\in\mathcal{G}}\widehat{\mathrm{LossDual}}_{i,h,\phi}(g; f_i), \tag{6}$$

and defining the approximate Bellman update by plugging $\widehat{g}^f_i$ into the dual formulation:

$$\big[\widehat{\mathcal{T}}^{\pi,\phi,\sigma_i}_{i,h} f_i\big](s, \boldsymbol{a}) = r_{i,h}(s, \boldsymbol{a}) - \mathbb{E}_{s'\sim P^\star_h}\big[\mathrm{Loss}_{i,h,\phi}(f_i; s, \boldsymbol{a}, s'; \widehat{g}^f_i(s, \boldsymbol{a}))\big]. \tag{7}$$

To characterize the approximation of the true robust Bellman operator, we adopt the following assumption, which holds for all considered uncertainty sets (Ghosh et al., 2025b).

Assumption 3. *We assume the existence of a compact set $\Theta_\phi \subseteq \mathbb{R}_+ \times \mathbb{R}$ and a finite constant $B_\phi(\sigma_i)$ such that the dual loss remains uniformly bounded. In particular, for all $h \in [H]$, $i \in \mathcal{M}$, any $f_{i,h+1} \in \mathcal{F}_{i,h+1}$, all $(\lambda, \nu) \in \Theta_\phi$, and every $(s, \boldsymbol{a}, s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, we have $\big|\mathrm{Loss}_{i,h,\phi}(f_i; s, \boldsymbol{a}, s'; \lambda, \nu)\big| \le B_\phi(\sigma_i)$. Additionally, we assume that the dual function class $\mathcal{G}$ is restricted to take values within the set Θ_ϕ.*

We then develop the following error quantification, whose proof is deferred to Appendix G.

Lemma 1 (Uniform dual-operator approximation under adaptive data)**.** *Fix an agent $i \in \mathcal{M}$ and a stage $h \in [H]$. Let $\{\pi^t\}_{t=1}^n$ be a predictable sequence of behavior policies, and let $\mathcal{D}_h$ be the stage-h dataset collected by executing $\{\pi^t\}_{t=1}^n$ over n consecutive episodes. Define the random historical occupancy mixture $\bar{\mu}^n_h := \sum_{t=1}^n \mu^{\pi^t}_h/n$. Assume Assumptions 1–3 hold and let the function classes $\mathcal{F}_i$ and $\mathcal{G}_i$ be finite. For each $\pi \in \Pi^{\mathrm{pur}}$ and $f_{i,h+1} \in \mathcal{F}_{i,h+1}$, let $\widehat{g}^{f,\pi}_i$ be an empirical minimizer of the dual loss on $\mathcal{D}_h$ as defined in equation 6. Then, for any $\delta \in (0, 1)$, with probability at least $1-\delta$:*

$$\sup_{f_{i,h+1}\in\mathcal{F}_{i,h+1},\pi\in\Pi^{\mathrm{pur}}}\Big\|\mathcal{T}^{\pi,\phi,\sigma_i}_{i,h} f_{i,h+1} - \mathcal{T}^{\pi,\phi,\sigma_i}_{\widehat{g}^{f,\pi}_i} f_{i,h+1}\Big\|_{1,\bar{\mu}^n_h}$$
$$\le CB_\phi(\sigma_i)\sqrt{\log\big(|\mathcal{F}_{i,h+1}||\mathcal{G}_{i,h}||\Pi^{\mathrm{pur}}|/\delta\big)/n} + \varepsilon^{\mathrm{approx}}_{\mathrm{dual}}.$$

[2]Here, g ranges over an appropriate function space (e.g., $\mathcal{L}^1(\mu^\pi)$), and $\mathrm{Loss}_{i,h,\phi}$ is the pointwise dual integrand as above. This formulation is equivalent to the original robust expectation under mild conditions; see (Panaganti et al., 2022; Ghosh et al., 2025b) for details.

Prior single-agent results (Panaganti et al., 2022; 2024) establish similar guarantees under a fixed offline distribution, whereas our result holds uniformly over all policies and their induced state–joint action distributions in the multi-agent setting. Lemma 1 shows that the empirical dual optimization uniformly approximates the robust Bellman operator under the $\mathcal{L}_i^1(\mu^\pi;\mathbb{R}^2)$ norm, with the approximation error controlled globally over the distribution μ^π rather than pointwise over individual $(s,\boldsymbol{a})$. This global control enables the construction of robust confidence sets and provides a unified error term underlying both our algorithm design and theoretical analysis.

4 Interactive Robust Multi-Agent Reinforcement Learning

We then present our *Robust Multi-Agent Exploration with ϕ-Divergence Uncertainty Set (RoMEX-ϕ)* (Algorithm 1). In each episode, we will learn value function approximations via fitted methods, where the fitting error is defined through our empirical robust operator. We describe our algorithm in this section, and defer our discussions on algorithmic novelties to Appendix E.

Algorithm 1: Robust Multi-Agent Exploration with ϕ-Divergence Uncertainty Set (RoMEX-ϕ)

1: **Input:** Number of episodes K, function classes $\mathcal{F}=\otimes_{i=1}^M\mathcal{F}_i$ and $\mathcal{G}=\otimes_{i=1}^M\mathcal{G}_i$, regularization parameter $\beta>0$, equilibrium oracle $\mathrm{EqOracle}(\cdot;\mathrm{NE/CCE/CE})$
2: **Initialize:** $\mathcal{D}_h^0\leftarrow\emptyset$ for all $h\in[H]$
3: **for** $k=1,2,\ldots,K$ **do**
4: **for** each pure joint policy $\pi\in\Pi^{\mathrm{pur}}$ **do**
5: Set $f_{i,H+1}^{k,\pi}\equiv 0$ for each agent $i\in\mathcal{M}$.
6: **for** $h=H,H-1,\ldots,1$ **do**
7: Compute empirical dual minimizer $\widehat{g}_{i,h}^{k,\pi}$ as in equation 8 for each agent $i\in\mathcal{M}$.
8: Compute plug-in robust Bellman fit $f_{i,h}^{k,\pi}$ (by Step 2) for each agent $i\in\mathcal{M}$.
9: **end for**
10: Compute regularised robust payoff $\widehat{U}_i^k(\pi)$ as given in equation 10 for $i\in\mathcal{M}$.
11: **end for**
12: Construct the normal-form game $\widehat{\mathbf{\Gamma}}^k:=\left(\{\Pi_i^{\mathrm{pur}}\}_{i\in\mathcal{M}},\{\widehat{U}_i^k\}_{i\in\mathcal{M}}\right)$ and compute equilibrium policy $\pi^k\leftarrow\mathrm{EqOracle}(\widehat{\mathbf{\Gamma}}^k;\mathrm{NE/CCE/CE})$.
13: Execute π^k in the nominal environment $P^\star$ for one episode and collect $\tau^k=\{(s_h^k,\boldsymbol{a}_h^k,r_h^k,s_{h+1}^k)\}_{h=1}^H$ and update dataset $\mathcal{D}_h^k\leftarrow\mathcal{D}_h^{k-1}\cup\{(s_h^k,\boldsymbol{a}_h^k,s_{h+1}^k)\}$ for $h\in[H]$.
14: **end for**
15: **Return** $\pi^{\mathrm{out}}:=\{\pi^k\}_{k=1}^K$ or $\frac{1}{K}\sum_{k=1}^K\pi^k$.

Fitted estimation. Let Π_i^{pur} denote the class of pure policies for agent i, and let $\Pi^{\mathrm{pur}}:=\Pi_1^{\mathrm{pur}}\times\cdots\times\Pi_m^{\mathrm{pur}}$ be the set of pure joint policies. Fix an episode k, a candidate joint policy $\pi\in\Pi^{\mathrm{pur}}$, and an agent $i\in[M]$. The algorithm computes a fitted robust value sequence $\{f_{i,h}^{k,\pi}\}_{h=1}^{H+1}$ via backward dynamic programming on the collected dataset, with terminal condition $f_{i,H+1}^{k,\pi}\equiv 0$ (Line 5). For each stage $h=H,\ldots,1$, two coupled steps are performed.

Step 1: Empirical dual minimization (Line 7). Using the stage-h dataset $\mathcal{D}_h^{k-1}$ collected up to episode $k-1$, we compute the empirical dual minimizer

$$\widehat{g}_{i,h}^{k,\pi}\in\arg\min_{g\in\mathcal{G}_{i,h}}\widehat{\mathrm{LossDual}}_{i,h,\phi}^{k-1}(g;f_{i,h+1}^{k,\pi},\pi),\tag{8}$$

where $\widehat{\mathrm{LossDual}}_{i,h,\phi}^{k-1}$ is the empirical version of the dual loss defined in equation 5 using $\mathcal{D}_h^{k-1}$.

Step 2: Robust Bellman fitting with plug-in residual (Line 8). Once the dual minimizer $\widehat{g}_{i,h}^{k,\pi}$ has been obtained, the algorithm uses it to define a plug-in empirical approximation of the robust Bellman operator. It then computes the next-stage h fitted value function by minimizing the empirical plug-in robust Bellman residual: $f_{i,h}^{k,\pi}\in\arg\min_{f\in\mathcal{F}_{i,h}}\sum_{(s,\boldsymbol{a},s')\in\mathcal{D}_h^{k-1}}\widehat{\delta}_{i,h}(f_i;\widehat{g}_{i,h}^{k,\pi};\pi;s,\boldsymbol{a},s')^2$, where $\widehat{\delta}_{i,h}$ is the plug-in empirical discrepancy defined as

$$\widehat{\delta}_{i,h}(f_i;g_i;\pi;s,\boldsymbol{a},s'):=f_i(s,\boldsymbol{a})-r_{i,h}(s,\boldsymbol{a})+\mathrm{Loss}_{i,h,\phi}(f_i;s,\boldsymbol{a},s';g_i,\pi).\tag{9}$$

Note that $\widehat{\delta}_{i,h}$ penalizes hypotheses that do not match the robust Bellman equation well on the historical data. The dual minimizer enters precisely through this term: the Bellman discrepancy is no longer computed using the nominal transition expectation, but through the dual-induced robust integrand. Intuitively, the dual minimizer $\widehat{g}_{i,h}^{k,\pi}$ converts the robust Bellman backup into a tractable empirical target, and the fitted value function $f_{i,h}^{k,\pi}$ best fits this target on the observed data.

Regularized robust payoff (Line 10). For each episode $k \in [K]$, agent $i \in \mathcal{M}$, and candidate pure joint policy $\pi \in \Pi^{\mathrm{pur}}$, we define the revised empirical payoff by

$$\widehat{U}_i^k(\pi) := f_{i,1}^{k,\pi}(s_1^k) - \beta \widehat{\mathcal{K}}_i^{k-1}\Big(f_i^{k,\pi}; \widehat{g}_i^{k,\pi}; \pi\Big), \tag{10}$$

where $\beta > 0$ is the regularization parameter, $\widehat{g}_{i,h}^{k,\pi}$ is the empirical dual minimizer in equation 6 for $f_{i,h+1}^{k,\pi}$ and π, and $\widehat{\mathcal{K}}_i^{k-1}(f_i; g_i; \pi)$ denotes the cumulative centered empirical robust discrepancy:

$$\sum_{h=1}^{H}\Big[\sum_{j=1}^{k-1}\widehat{\delta}_{i,h}(f_i; g_{i,h}; \pi; s_h^j, \boldsymbol{a}_h^j, s_{h+1}^j)^2 - \inf_{f_i' \in \mathcal{F}_i}\sum_{j=1}^{k-1}\widehat{\delta}_{i,h}(f_i'; g_{i,h}; \pi; s_h^j, \boldsymbol{a}_h^j, s_{h+1}^j)^2\Big], \tag{11}$$

with $\widehat{\mathcal{K}}_i^0(f_i; g_i; \pi) = 0$. The first term in equation 10 favors policies with large estimated robust value, while the second penalizes large empirical robust Bellman discrepancy. The resulting objective acts as a conservative surrogate value, trading off value estimation and Bellman consistency rather than introducing an explicit optimistic bonus. Similar to MAMEX (Xiong et al., 2024), this discrepancy term serves as a global regularizer, and in our setting, it controls the gap between empirical robust Bellman error and robust value prediction via the robust MADC.

Policy selection and data collection (Line 12-15). We then define a normal-form game $\widehat{\mathbf{\Gamma}}^k := \Big(\{\Pi_i^{\mathrm{pur}}\}_{i\in[M]}, \{\widehat{U}_i^k\}_{i\in[M]}\Big)$, whose pure-strategy profile is $\pi \in \Pi^{\mathrm{pur}}$ and whose payoff to agent i at profile π is $\widehat{U}_i^k(\pi)$. Once the normal-form game $\widehat{\mathbf{\Gamma}}^k$ is formed, the algorithm computes the desired equilibrium through some oracle: $\pi^k \leftarrow \mathrm{EqOracle}(\widehat{\mathbf{\Gamma}}^k; \mathrm{NE/CCE/CE})$. As we mainly focus on statistical complexity, we assume such oracles. Specifically, the NE oracle is generally NP-hard (Daskalakis, 2013), while CE and CCE can be computed with polynominal complexity (Zheng & Lin, 2026). The policy π^k is then executed for one episode in the nominal environment $P^\star$, producing the trajectory $\tau^k = \{(s_h^k, \boldsymbol{a}_h^k, r_h^k, s_{h+1}^k)\}_{h=1}^H$. For each stage h, the transition tuple $(s_h^k, \boldsymbol{a}_h^k, s_{h+1}^k)$ is appended to the dataset $\mathcal{D}_h^{k-1}$ to form $\mathcal{D}_h^k$, for the following algorithm executions. The final output is either the sequence $\{\pi^k\}_{k=1}^K$ or the uniform mixture $\bar{\pi} = \frac{1}{K}\sum_{k=1}^K \pi^k$, depending on the equilibrium notion under consideration.

5 Theoretical Guarantee

We now develop our regret analysis. Inspired by the non-robust studies (Xiong et al., 2024), we first introduce a fundamental complexity measure, named *robust multi-agent decoupling coefficient* (Robust MADC), to quantify the inherent difficulty of exploration in DRMGs. This notion will enable us to develop our analysis from an optimization perspective.

Definition 4 (Robust Multi-Agent Decoupling Coefficient)**.** For each agent $i \in \mathcal{M}$, policy $\pi \in \Pi^{\mathrm{pur}}$, and stage $h \in [H]$, let μ_h^π denote the stage-h state-joint action visitation distribution induced by policy π under the nominal kernel $P^\star$. Then, for any $i \in \mathcal{M}$, $\mu, c > 0$, $f_i^k \in \mathcal{F}_i$, and joint policy $\{\pi^k\}_{k=1}^K$, the robust MADC for agent $i \in \mathcal{M}$ is the smallest constant $1 \le d_{i,\mathrm{MADC}}^{\mathrm{rob}} < \infty$ such that:

$$\sum_{k=1}^{K}(V_{i,1}^{\pi^k, f_i^k}(s_1^k) - V_{i,1}^{\pi^k,\sigma_i}(s_1^k)) \le \frac{1}{\mu}\sum_{k=1}^{K}\sum_{s=1}^{k-1} e_{i,1}^s(f_i^k, \pi^k) + \mu\, d_{i,\mathrm{MADC}}^{\mathrm{rob}} + cHd_{i,\mathrm{MADC}}^{\mathrm{rob}}, \tag{12}$$

where $e_{i,1}^s(f, \pi)$ is the discrepancy function defined as the mean-squared robust Bellman error:

$$e_{i,1}^s(f,\pi) \triangleq \sum_{h=1}^{H}\mathbb{E}_{(s_h,\boldsymbol{a}_h)\sim\mu_h^{\pi^s}}\left[\big(f_{i,h}(s_h,\boldsymbol{a}_h) - [\mathcal{T}_{i,h}^{\pi,\phi,\sigma_i} f_{i,h+1}](s_h,\boldsymbol{a}_h)\big)^2\right], \forall s \in [K], \forall f \in \mathcal{F}_i.$$

Interpretation of robust MADC. Our robust MADC extends the decoupling coefficient in (Dann et al., 2021; Xiong et al., 2024) to robust settings, capturing the joint difficulty of strategic interactions and transition uncertainty. At a high level, $d^{\mathrm{rob}}_{i,\mathrm{MADC}}$ quantifies the decoupling between (i) the *training error*, given by the cumulative empirical robust Bellman residual, and (ii) the *prediction error*, defined as the gap between the learned value and the true robust value under worst-case dynamics. Definition 4 ensures that the cumulative prediction error is controlled by the cumulative training error up to a term governed by $d^{\mathrm{rob}}_{i,\mathrm{MADC}}$. Unlike the non-robust setting (Xiong et al., 2024), the prediction error is evaluated under adversarial dynamics while the training error is computed from nominal interaction data. Thus, $d^{\mathrm{rob}}_{i,\mathrm{MADC}}$ characterizes how effectively nominal interaction data controls value prediction under worst-case dynamics, capturing the intrinsic exploration complexity of DRMGs.

Structural perspective. Beyond the decoupling interpretation, robust MADC also provides a structural view: a smaller $d^{\mathrm{rob}}_{i,\mathrm{MADC}}$ indicates that small empirical robust Bellman error translates into small value prediction error, reflecting better alignment of the function class with the robust Bellman operator. This parallels Bellman-error-based complexity measures such as the BE dimension, while additionally capturing the mismatch between nominal data and adversarial evaluation. This perspective extends naturally to structured function classes, such as bilinear models (Du et al., 2021b), where low-dimensional representations enable accurate estimation of the robust Bellman operator from nominal data, thereby controlling the discrepancy terms, yielding a bounded robust MADC, and improving generalization under worst-case dynamics.

In previous RL studies, the (multi-agent) Bellman-Eluder (MABE) dimension is another widely used complexity measure in RL (Jin et al., 2021; Ghosh et al., 2025b). In the non-robust setting, a low BE dimension implies a low MADC (Xiong et al., 2024), making MADC a more general complexity measure. In the robust setting, the robust BE dimension was introduced for single-agent RL in (Ghosh et al., 2025b) and can be extended to multi-agent settings. Our robust MADC instead directly quantifies how robust Bellman errors estimated from nominal interactions control value prediction under uncertain transition kernels. Moreover, Theorem 3 shows that a low robust MABE, extending the robust BE dimension of (Ghosh et al., 2025b), implies a low robust MADC, establishing robust MADC as a more general complexity measure.

Extending results from the non-robust setting (Jin et al., 2021; 2022), the inherent robust MABE can be substantially smaller than the tabular complexity $S\prod_{i=1}^{m} A_i$ for structured function classes. By Theorem 3, the robust MADC can therefore also be much smaller than $S\prod_{i=1}^{m} A_i$ under function approximation. Since our sample-complexity guarantees are characterized by the robust MADC, this enables improved scalability over tabular approaches.

We now present the theoretical guarantees of RoMEX-ϕ; full details are deferred to Appendix H.

Theorem 1. *Assume Assumptions 1–3 and the function classes $\mathcal{F}_i$ and dual function class $\mathcal{G}_i$ are finite*[3]*. Let $\{\pi^k\}_{k=1}^{K}$ be policies produced by Algorithm 1 with different oracles, and denote $d^{\mathrm{rob,max}}_{\mathrm{MADC}} = \max_{i\in\mathcal{M}} d^{\mathrm{rob}}_{i,\mathrm{MADC}}$, $B^{\max}_{\phi} = \max_{i\in\mathcal{M}} B_\phi(\sigma_i)$ and $\iota^{\delta}_{\mathcal{F},\mathcal{G}} := \log\big(\frac{|\mathcal{F}||\mathcal{G}||\Pi^{\mathrm{pur}}|KHm}{\delta}\big)$. Then, with probability at least $1-\delta$, the corresponding robust regret satisfies*

$$\mathrm{Reg}(K) \le \mathcal{O}\Big(d^{\mathrm{rob,max}}_{\mathrm{MADC}}\sqrt{K} + H d^{\mathrm{rob,max}}_{\mathrm{MADC}} + H(B^{\max}_{\phi})^2\sqrt{K\iota^{\delta}_{\mathcal{F},\mathcal{G}}} + \varepsilon^{\mathrm{approx}}_{\mathrm{dual}}\Big).$$

Theorem 1 shows that RoMEX-ϕ efficiently learns robust equilibria in general-sum DRMGs with general function approximation. The bound decomposes into four terms: (i) the *exploration term* $d^{\mathrm{rob,max}}_{\mathrm{MADC}}\sqrt{K}$, governed by the robust MADC, which—analogous to (Xiong et al., 2024)—links cumulative in-sample robust Bellman discrepancy to out-of-sample robust value prediction error and thus captures the intrinsic difficulty of exploration. Optimizing the decoupling parameter yields the $\sqrt{K}$ rate. While small cumulative discrepancy ensures that the exploration-induced component of regret is sublinear, the overall regret additionally depends on the robust estimation term arising from approximating the worst-case Bellman operator from nominal data; (ii) a *burn-in horizon-dependent term* $H\, d^{\mathrm{rob,max}}_{\mathrm{MADC}}$, which reflects horizon-dependent initialization effects and does not affect the asymptotic rate; (iii) a *robust estimation term* $H\,(B^{\max}_{\phi})^2\sqrt{K\,\iota^{\delta}_{\mathcal{F},\mathcal{G}}}$, which is unique to the robust

[3]This assumption is made for technical simplicity. While extensions to infinite function classes may be possible using covering arguments (e.g., ϵ-nets (Panaganti et al., 2022)), establishing such results rigorously in the robust multi-agent setting remains non-trivial and is left for future work.

setting. This term captures the statistical cost of estimating the robust Bellman operator from nominal data via the dual formulation. Here, $B_\phi^{\max}$ controls the dual loss envelope, $\iota^{\delta}_{\mathcal{F},\mathcal{G}}$ reflects the complexity of the primal and dual function classes; and (iv) the *dual approximation error* $\varepsilon_{\text{dual}}^{\text{approx}}$, which captures the representational error of the dual function class and vanishes under realizability of $\mathcal{G}$. Overall, the bound reveals two core challenges: exploration under strategic interaction, captured by $d_{\text{MADC}}^{\text{rob,max}}$ and robust Bellman estimation under distributional mismatch, reflected in by $(B_\phi^{\max})^2\iota^{\delta}_{\mathcal{F},\mathcal{G}}$. Refer to Appendix E for a detailed discussion on the technical novelties of RoMEX-ϕ.

As a consequence of our regret guarantees, applying a standard online-to-batch conversion (Cesa-Bianchi et al., 2001) gives the first sample-complexity guarantees for learning an ε-optimal policy with RoMEX-ϕ. As corollary, we obtain regret bound under total variation, χ^2- and KL-divergence uncertainty sets, presented in Appendix H.1 and Appendix I. To our knowledge, these results constitute the first sample-complexity characterization for distributionally robust MARL with general function approximation. Crucially, the bounds are governed by the robust MADC and contain no explicit dependence on the state-space size $|\mathcal{S}|$ or the joint action-space size $|\mathcal{A}|$. Thus, compared with tabular robust MARL, our analysis replaces the unfavorable joint-action-space dependence with an intrinsic function-class complexity, improving scalability and mitigating the curse of multi-agency (Shi et al., 2024b; Farhat et al., 2026).

To further interpret the sharpness and generality of our results, we specialize the general bound to representative settings. In particular, we consider the tabular case, robust MABE/BE-type characterizations, and linear function approximation. A more detailed and comprehensive comparison with prior work is provided in Appendix D.

Example 1 (Tabular setting) We can choose $\mathcal{F}$ and $\mathcal{G}$ as full classes over $\mathcal{S} \times \prod_{i=1}^m \mathcal{A}_i$ to reduce to tabular settings, which yields $d_{\text{MADC}}^{\text{rob,max}}$ and $\log(|\mathcal{F}||\mathcal{G}|)$ of order $\widetilde{\mathcal{O}}\big(S\prod_{i=1}^m A_i\big)$. Substituting them into our general bound implies the tabular results: e.g., for KL set, it achieves $\widetilde{\mathcal{O}}\Big(\frac{H^7\sigma_{\max}^4 S^2(\prod_{i=1}^m A_i)^2}{\varepsilon^2}\Big)$, which is comparable to the tabular result in (Farhat et al., 2026). We highlight that although this complexity from direct reduction also depends on the joint action space, in more structured function approximation settings, the approximation function class will have a much smaller $d_{\text{MADC}}^{\text{rob,max}}$ and smaller complexity. Moreover, the minimax lower bound for robust MARL developed in (Farhat et al., 2026) depends on the joint action space, thus this dependence is inevitable in tabular settings.

Example 2 (Linear function approximation) Our results can also be reduced to linear function approximation settings (details are deferred to Section K). Specifically, under the robust bilinear structure, Theorem 4 yields $d_{\text{MADC}}^{\text{rob,max}} = \widetilde{\mathcal{O}}(H^3 d_{\text{lin}})$. Substituting into the general bound gives $\widetilde{\mathcal{O}}\big(\max\{H^6 d_{\text{lin}}^2/\varepsilon^2,\ H^7\Psi_\phi(\sigma)/\varepsilon^2\}\big)$. For χ^2 and KL uncertainty sets, the robustness term scales polynomially as $(1+\sqrt{\sigma_{\max}})^4$ and $\max\{1,\sigma_{\max}^4\}$. Although compared to linear-specific analyses (Zheng & Lin, 2026), our results have a higher-ordered dependence on the horizon H, our results enjoy a much broader applicability and the additional complexity arises from the general function-approximation framework, reflecting a unified Bellman-error-based theory that extends beyond linear settings and accommodates a broad class of ϕ-divergence uncertainty sets.

6 Conclusion

We study online distributionally robust MARL in general-sum Markov games under general function approximation. We propose RoMEX-ϕ, a model-free framework that uses a dual reformulation of the robust Bellman operator to enable tractable worst-case evaluation from nominal data via an equilibrium oracle. We introduce the *robust Multi-Agent Decoupling Coefficient* (robust MADC) as an intrinsic complexity measure and establish the first sublinear regret guarantees for learning robust equilibria. Our results decompose sample complexity into intrinsic exploration, governed by robust MADC, and robust Bellman estimation from nominal data, while avoiding explicit dependence on the state and joint action spaces. Robust MADC further provides a unified view connecting intrinsic complexity with structured function classes. An important direction for future work is extending the analysis to richer function classes, such as infinite or neural parameterizations, as well as to decentralized or partially observed multi-agent settings.

AI USE STATEMENT

We used ChatGPT solely as a general-purpose assistant for language polishing and typesetting. Specifically, it assisted with grammar, style, and readability improvements, as well as LaTeX formatting tasks, including algorithm placement, bibliography and citation formatting, and resolving compilation issues. All research ideas, theoretical developments, derivations, experimental results, and scientific claims were developed, verified, and validated by the authors. All AI-assisted content was reviewed by the authors, who take full responsibility for the final content of this paper.

ETHICS STATEMENT

This work develops theoretical and algorithmic methods for distributionally robust multi-agent reinforcement learning and does not involve human subjects, personal data, or sensitive datasets. While the proposed methods may have broader applications to multi-agent decision-making systems, application-specific ethical, safety, fairness, and societal considerations are beyond the scope of this work.

REPRODUCIBILITY STATEMENT

We provide complete proofs of all theoretical results and detailed discussions of the assumptions in the appendices. The algorithm and its theoretical components are specified in the main text, while Appendix F provides the experimental setup, implementation details, baselines, evaluation protocol, and ablation studies. These materials are intended to provide the necessary details for reproducing the theoretical and empirical results reported in this work.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved Algorithms for Linear Stochastic Bandits. In *NIPS*, volume 11, pp. 2312–2320, 2011.

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.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1638–1646. PMLR, 2014.

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.

Stefano V Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024.

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.

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.

Fan Chen, Song Mei, and Yu Bai. Unified Algorithms for RL with Decision-Estimation Coefficients: No-Regret, PAC, and Reward-Free Learning . 2022a.

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 2022b. URL https://proceedings.mlr.press/v167/chen22d.html.

Pedro Cisneros-Velarde and Sanmi Koyejo. Finite-sample Guarantees for Nash Q-learning with Linear Function Approximation . In *Uncertainty in Artificial Intelligence*, pp. 424–432. PMLR, 2023.

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.

Yan Dai, Qiwen Cui, and Simon S Du. Refined Sample Complexity for Markov Games with Independent Linear Function Approximation. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1260–1261. PMLR, 2024.

Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A Provably Efficient Model-Free Posterior Sampling Method for Episodic Reinforcement Learning. *Advances in neural information processing systems*, 34:12040–12051, 2021.

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. *Commun. ACM*, 52(2):89–97, February 2009. ISSN 0001-0782. doi: 10.1145/1461928.1461951.

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.

Simon Du, Sham Kakade, Jason Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear Classes: A Structural Framework for Provable Generalization in RL. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2826–2836. PMLR, July 2021a.

Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021b.

Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-Dynamics Reinforcement Learning: Training for Transfer with Domain Classifiers. *arXiv preprint arXiv:2006.13916*, 2020.

Zain Ulabedeen Farhat, Debamita Ghosh, George K. Atia, and Yue Wang. Sample-efficient distributionally robust multi-agent reinforcement learning via online interaction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hIBMCDOwMI.

Dean Foster, Dylan J Foster, Noah Golowich, and Alexander Rakhlin. On the complexity of multi-agent decision making: From learning in games to partial monitoring. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2678–2792. PMLR, 2023.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Drew Fudenberg and Jean Tirole. *Game Theory*. MIT press, 1991.

Debamita Ghosh, George K. Atia, and Yue Wang. ORVIT: Near-Optimal Online Distributionally Robust Reinforcement Learning, 2025a. URL https://arxiv.org/abs/2508.03768.

Debamita Ghosh, George K Atia, and Yue Wang. Scaling Online Distributionally Robust Reinforcement Learning: Sample-Efficient Guarantees with General Function Approximation. *arXiv preprint arXiv:2512.18957*, 2025b.

Jiafan He, Heyang Zhao, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal reinforcement learning for linear markov decision processes. In *International Conference on Machine Learning*, pp. 12790–12822. PMLR, 2023.

Yiting He, Zhishuai Liu, Weixin Wang, and Pan Xu. Sample Complexity of Distributionally Robust Off-Dynamics Reinforcement Learning with Online Interaction. In *Forty-second International Conference on Machine Learning*, 2025.

Joshua Arvind Holla. *On the off-dynamics approach to reinforcement learning*. McGill University (Canada), 2021.

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.

Baihe Huang, Jason D Lee, Zhaoran Wang, and Zhuoran Yang. Towards general function approximation in zero-sum markov games. *arXiv preprint arXiv:2107.14702*, 2021.

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

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1704–1713. PMLR, 2017.

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

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Chi Jin, Qinghua Liu, and Tiancheng Yu. The power of exploiter: Provable multi-agent rl in large state spaces. In *Proc. International Conference on Machine Learning (ICML)*, pp. 10251–10279. PMLR, 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, 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=3lXZjsir0e.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research (PMLR)*, pp. 7001–7010. PMLR, 18–24 Jul 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, 2024a.

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

Zhishuai Liu, Weixin Wang, and Pan Xu. Upper and lower bounds for distributionally robust off-dynamics reinforcement learning. *arXiv preprint arXiv:2409.20521*, 2024.

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.

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.

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

Sindhu Padakandla, Prabuchandran KJ, and Shalabh Bhatnagar. Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence*, 50(11):3590–3606, 2020.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust Reinforcement Learning using Offline Data. *Advances in neural information processing systems*, 35:32211–32224, 2022.

Kishan Panaganti, Adam Wierman, and Eric Mazumdar. Model-Free Robust ϕ-Divergence Reinforcement Learning Using Both Offline and Online Data. *arXiv preprint arXiv:2405.05468*, 2024.

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.

Zachary Roch and Yue Wang. Distributionally robust markov games with average reward. *arXiv preprint arXiv:2508.03136*, 2025.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz et al. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

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.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

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.

Ruosong Wang, Ruslan Salakhutdinov, and Lin F. Yang. Reinforcement Learning with General Value Function Approximation: Provably Efficient Approach via Bounded Eluder Dimension. *arXiv:2005.10804 [cs, math, stat]*, June 2020. arXiv: 2005.10804.

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

Yue Wang and Shaofeng Zou. Online Robust Reinforcement Learning with Model Uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

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, 2024b.

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.

Nuoya Xiong, Zhihan Liu, Zhaoran Wang, and Zhuoran Yang. Sample-Efficient Multi-Agent RL: An Optimization Perspective. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

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, 2022a.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Toward Theoretical Understandings of Robust Markov Decision Processes: Sample Complexity and Asymptotics. *The Annals of Statistics*, 50, 12 2022b. doi: 10.1214/22-AOS2225.

Wenhao Zhan, Jason D Lee, and Zhuoran Yang. Decentralized optimistic hyperpolicy mirror descent: Provably no-regret learning in markov games. *arXiv preprint arXiv:2206.01588*, 2022.

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.

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.

Zewu Zheng and Yuanyuan Lin. Distributionally robust online markov game with linear function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28892–28900, 2026.

Han Zhong, Wei Xiong, Sirui Zheng, Liwei Wang, Zhaoran Wang, Zhuoran Yang, and Tong Zhang. A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962*, 2(3), 2022.

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.