Journal of Machine Learning Research 24 (2023) 1-40 Submitted 3/23; Revised 11/23; Published 11/23

Decentralized Robust V-learning for Solving Markov Games with Model Uncertainty

Shaocong Ma S.MA@UTAH.EDU
Department of Electrical and Computer Engineering
University of Utah
Salt Lake City, UT 84112, USA

Ziyi Chen ZC286@CORNELL.EDU
Department of Electrical and Computer Engineering
University of Utah
Salt Lake City, UT 84112, USA

Shaofeng Zou SZOU3@BUFFALO.EDU
Department of Electrical Engineering
University at Buffalo, The State University of New York
Buffalo, NY 14260, USA

Yi Zhou YI.ZHOU@UTAH.EDU
Department of Electrical and Computer Engineering
University of Utah
Salt Lake City, UT 84112, USA

Editor: Martha White

Abstract

The Markov game is a popular reinforcement learning framework for modeling competitive players in a dynamic environment. However, most of the existing works on Markov games focus on computing a certain equilibrium following uncertain interactions among the players but ignore the uncertainty of the environment model, which is ubiquitous in practical scenarios. In this work, we develop a theoretical solution to Markov games with environment model uncertainty. Specifically, we propose a new and tractable notion of robust correlated equilibria for Markov games with environment model uncertainty. In particular, we prove that the robust correlated equilibrium has a simple modification structure, and its characterization of equilibria critically depends on the environment model uncertainty. Moreover, we propose the first fully-decentralized stochastic algorithm for computing such the robust correlated equilibrium. Our analysis proves that the algorithm achieves the polynomial episode complexity $\widetilde{\mathcal{O}}(SA^2H^5\epsilon^{-2})$ for computing an approximate robust correlated equilibrium with ϵ accuracy.

Keywords: robust Markov games, model uncertainty, robust correlated equilibrium, reinforcement learning

1 Introduction

The Markov game is a general and popular reinforcement learning framework for modeling multiple players competing with each other in a dynamic environment (Littman, 1994). In a

©2023 Shaocong Ma, Ziyi Chen, Shaofeng Zou, and Yi Zhou.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v24/23-0310.html.

Markov game, players interact with each other through a Markov decision process, and each player aims to improve its own decision-making to compete for more rewards. In particular, many important real-life applications fit into this framework, including multi-player games such as decentralized multi-agent robotic control (Brambilla et al., 2013) and distributed autonomous driving (Shalev-Shwartz et al., 2016).

One of the popular goals of Markov games is to achieve the Nash equilibrium (NE) among the players, that is, an optimal product policy so that no player can improve its gain by deviating from its own policy alone. The NE has been shown to exist for general Markov games (Filar and Vrieze, 2012). However, it turns out that finding NE of a Markov game is generally a PPAD-complete problem that is still unknown if it could be solved in polynomial time (Deng et al., 2021; Jin et al., 2022b), except for some special Markov games with either zero-sum rewards (Bai and Jin, 2020; Jin et al., 2018) or potential structures (Leonardos et al., 2021; Zhang et al., 2021). Besides Nash equilibria and its hardness, researchers have found multiple theoretically-tractable notions. For example, the correlated equilibrium (CE) (Aumann, 1974) (see Definition 4) and the coarse correlated equilibrium (CCE) (Moulin and Vial, 1978), which are similar to the NE but allows dependency among the players' policies, have been shown to have polynomial-complexity algorithms (Blum and Mansour, 2007). Also, the extensive-form correlated equilibrium (EFCE) is proposed in extensive-form games (Von Stengel and Forges, 2008), which is also the solution to the PPAD-complete problem of finding Nash equilibria in multi-player games with inperfect information. And the Stackelberg equilibrium is also proposed as a substituion for the NE due to its sample efficiency (Bai et al., 2021). Particularly, in recent years, the CE has attracted much attentions due to its polynomial computation complexity (Jin et al., 2022a; Liu et al., 2021; Li et al., 2021).

Although Markov games and these polynomially-efficient notions have been extensively investigated, this standard framework only considers the competition among the players but ignores the environment model uncertainty, which is a critical factor that often reduces players' gains and must be considered in theoretical analysis and practical applications. For example, many applications such as multi-UAV systems (Chen et al., 2023) naturally involve uncertain environments due to real-world noises, sensor errors, or dynamic changes. As another example, policies trained in a simulated environment often suffers from significant performance degradation when implemented in the real environment, due to model mismatch; therefore, it is much more preferred to train a robust agent in the simulated environment, especially in autonomous vehicle controls (Allamaa et al., 2022). In all these scenarios, it is much desired to learn an optimal robust policy against such model uncertainty. To address model uncertainty, numerous robust reinforcement learning approaches have been developed and extensively studied in the single-agent case (Wang and Zou, 2021; Li et al., 2022b,a; Neufeld and Sester, 2022). However, model uncertainty is still underexplored in the general case with multiple competing agents, where only two works (Kardeş et al., 2011; Zhang et al., 2020) exist to our knowledge. Specifically, Kardeş et al. (2011) applied the robust Markov game with model uncertainty to the application of queueing control. Zhang et al. (2020) proposed provably convergent Q-learning and actor-critic type algorithms to compute a certain robust variant of NE of robust Markov games. However, computing robust NE of general Markov games with model uncertainty is in general a PPAD-complete problem and therefore it remains open if any polynomial-time algorithms exist. Motivated by existing

studies proposing new notions that can be found in polynomial complexity, we notice that the robust version of correlated equilibria has not been studied. This motivates **the two major goals of this work**: (i) to propose a theoretically generalized robust equilibrium notion and study its fundamental properties; and (ii) to construct a fully decentralized, provably-convergent and polynomial-complexity algorithm for computing such the robust equilibrium.

1.1 Our Contributions

In this work, we study episodic Markov games in an uncertain environment, that is, the environment transition kernel in every time step is queried from an underlying uncertainty set. To find a robust equlibrium policy of such Markov games with model uncertainty,within polynomial time, we make the following technical contributions.

- We propose a new theoretically-tractable notion of robust correlated equilibrium (CE) for Markov games with model uncertainty (see Definition 6). Specifically, the robust CE generalizes the standard CE in that it is defined based on a robust value function (see eq. (2)), which corresponds to the worst-case value function achieved under model uncertainty.

- We study the fundamental properties of robust CE. Specifically, we show that robust CE can be equivalently defined using either stochastic modifications or deterministic modifications (see Proposition 7). This indicates that robust CE inherits the modification structure from the standard CE. Similar to the non-robust case, if a policy is a robust NE, then it must be a robust CE (see Proposition 8, item 1). Moreover, through an illustrative example (see Proposition 8, item 2), we prove that the characterization of equilibria of robust CE critically depends on the uncertainty set, that is, there exists some uncertainty sets for the same state and action spaces such that the set of robust CE strictly includes the set of robust NE.

- We develop a fully decentralized robust V-learning algorithm for finding robust CE of Markov games with model uncertainty. This algorithm is a generalization of the original V-learning algorithm (for solving standard Markov games) and adopts robust TD learning in its critic update. Under low-level of model uncertainty (that is, when the diameter of the uncertainty set D is not larger than a given threshold $\frac{\epsilon}{SH^2}$), we prove that this algorithm achieves a polynomial episode complexity $\widetilde{\mathcal{O}}(SA^2H^5\epsilon^{-2})$ for computing an approximate robust CE with ϵ accuracy. Under sufficient exploration with relatively high-level of model uncertainty (that is, when the diameter of the uncertainty set D is within the interval $[\frac{\epsilon}{SH^2}, \frac{p_{\min}}{H})$), the complexity becomes $\widetilde{\mathcal{O}}(SA^2H^5p_{\min}^{-2}\epsilon^{-2})$. This is the first non-asymptotic convergence result for solving Markov games with model uncertainty. Moreover, our analysis of robust V-learning is substantially different from that of the original V-learning. Please refer to the elaboration of technical novelty after Theorem 13 for more details. To briefly elaborate, this is because the robust TD update enables tracking the desired robust value function at the cost of introducing uncertainty to the state transitions when unrolling the iterative updates. Therefore, we need to bound the model uncertainty via a stronger convergence metric, which leads to solving a linear system that involves an upper triangular Toeplitz matrix.

1.2 Related Work

Markov games: Markov games, also known as stochastic games, are standard formalism in multi-agent RL (Littman, 1994). The existence of NE for multi-player general-sum Markov games has been established in Fink (1964). Various algorithms have been designed to find NE, such as Nash-Q learning (Hu and Wellman, 2003), FF-Q learning (Littman et al., 2001), and correlated-Q learning (Greenwald et al., 2003). The first polynomial-time algorithm for finding NE is developed in Hansen et al. (2013), but works only for zero-sum games. Recent studies showed that finding NE of general-sum multi-player games is PPAD-complete (Daskalakis, 2013), so there are currently no polynomial-time algorithms for solving them (Deng et al., 2021; Jin et al., 2022b). Another notable goal in Markov games is to find a weaker version of NE, such as the correlated equilibrium (CE) or coarse correlated equilibrium (CCE). Polynomial-time algorithms such as V-learning (Jin et al., 2022a; Mao and Başar, 2022; Song et al., 2021) and Nash value iteration (Liu et al., 2021) have been developed for computing these notions.

Robust reinforcement learning: Single-agent robust reinforcement learning has been widely explored (Nilim and Ghaoui, 2003; Nilim and El Ghaoui, 2005; Wiesemann et al., 2013; Satia and Lave Jr, 1973), which assume the environment transition kernel belongs to a given uncertainty set. Under a specific uncertainty model, Roy et al. (2017) and Wang and Zou (2021) developed model-free online robust Q-learning algorithms to solve the robust reinforcement learning problem. For robust multi-agent reinforcement learning, value iteration-based algorithm has been developed in Kardeş et al. (2011) but with no explicit analytical form. For cooperative multi-agent reinforcement learning with model uncertainty, Huang et al. (2021) proposed a robust policy iteration algorithm to maximize the gain of the whole group. For non-cooperative Markov games with model uncertainty, Zhang et al. (2020) introduced robust Q-learning and actor-critic algorithms with asymptotic convergence guarantees of finding robust NE. To the best of our knowledge, there is no existing polynomial-time algorithm for solving Markov games with model uncertainty.

Existence of Nash equilibria: The existence of Nash equilibria for the discounted stochastic games is provided by Fink (1964), which also implies the existence of correlated equilibria and coarse correlated equilibria. Though the existence of robust Nash equilibria doesn't hold in general, it is ensured under some mild regularity assumptions (Perchet, 2014, 2020; Kardeş et al., 2011).

2 Preliminaries of Markov Games

An episodic m-player Markov game is specified by the tuple $(H, \mathcal{S}, \mathcal{A}, \mathbb{P}, \{r^{(j)}\}_{j=1}^m)$, where H is the length of each episode, $\mathcal{S}$ and $\mathcal{A} := \bigtimes_{j=1}^m \mathcal{A}^{(j)}$ correspond to the state space and joint action space, respectively, and they are assumed to be finite. Moreover, $r^{(j)} : \mathcal{S} \times \mathcal{A} \to [0,1]$ denotes the reward function of the j-th player and $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^H$ corresponds to the collection of transition kernels at time steps $h = 1, ..., H$. At every time step h, the players observe a global state $s_h \in \mathcal{S}$ of the environment. Then, they take a joint action $a_h = [a_h^{(1)}, \ldots, a_h^{(m)}]$ following a joint stochastic policy $\pi_h(\cdot|s_{1:h}, a_{1:(h-1)})$, which corresponds to a distribution on the joint action space $\mathcal{A}$ that depends on the past states $s_{1:h} := \{s_t\}_{t=1}^h$ and past actions $a_{1:(h-1)} := \{a_t\}_{t=1}^{h-1}$. After that, the global state transfers to a new state s_{h+1} following the

state transition kernel $\mathbb{P}(\cdot|s_h, a_h)$, and each player j receives a local reward $r_h^{(j)}(s_h, a_h)$ from the environment.

In the above Markov game, each player j collects its own rewards over the episodes. In particular, we define $\pi := \{\pi_h\}_{h=1}^H$ as the collection of joint policies over the time steps (they may be correlated; that is, policies in different time steps could be dependent). Then we can define the following value function for the j-th player at state s and time step h under the policy π.

$$\text{(Value Function):} \quad \boldsymbol{v}_{\pi,h}^{(j)}(s) := \mathbb{E}\Big[\sum_{\ell=h}^{H} r_\ell^{(j)}(s_\ell, a_\ell)\Big| s_h = s, \pi, \mathbb{P}\Big], \tag{1}$$

which corresponds to the expected cumulative reward received by player j starting from state s at time step h under joint policy π. The goal of the player j is to optimize its own policy $\pi^{(j)} := \{\pi_h^{(j)}\}_{h=1}^H$ in order to maximize its associated value function. However, since every player's value function is also affected by the other players' policies and actions, the players must compete with each other to gain more rewards until they reach a certain equilibrium. Here, we introduce two popular equilibrium notions that will be discussed throughout the paper.

Definition 1 (Nash Equilibrium (NE)) *A joint policy π is called an NE if the following two conditions are met: (i) for any time step h, the joint policy π_h is a product of independent policies, that is, $\pi_h = \pi_h^{(1)} \times \ldots \times \pi_h^{(m)}$; (ii) For any player j with any associated policy $\widetilde{\pi}^{(j)}$, it holds that $\boldsymbol{v}_{\pi,1}^{(j)}(s) \geq \boldsymbol{v}_{\widetilde{\pi}^{(j)} \times \pi^{(\backslash j)},1}^{(j)}(s)$ for all states $s \in \mathcal{S}$. Here, $\pi^{(\backslash j)}$ denotes the joint policy of all the other players excluding the player j, and '$\times$' means that $\widetilde{\pi}^{(j)}$ is independent from $\pi^{(\backslash j)}$.*

In the existing literature, it has been shown that computing NE is in general a PPAD-complete problem (Deng et al., 2021; Jin et al., 2022b; Daskalakis, 2013), for which it is still unknown if it is possible to develop polynomial-time algorithms. This has motivated researchers to propose a surrogate correlated equilibrium (CE) notion (Aumann, 1974). Before introducing the formal definition of CE, we first define the following stochastic modification operator.

Definition 2 (Stochastic Modification) *At any time step h, denote $a_h^{(j)}$ as player j's action induced by joint policy π_h. Given the past states and actions $s_{1:h}, a_{1:(h-1)}$, a stochastic modification $\phi_h^{(j)}$ associated with player j randomly maps $a_h^{(j)}$ to another action $\widetilde{a}_h^{(j)}$, that is, $\widetilde{a}_h^{(j)} \sim \phi_h^{(j)}(\cdot|s_{1:h}, a_{1:h-1}, a_h^{(j)})$.*

Moreover, we denote $\phi_h^{(j)} \circ \pi_h$ as the joint policy modified by $\phi_h^{(j)}$, that is, π_h first generates a joint action $a_h := [a_h^{(j)}, a_h^{(\backslash j)}]$, and then $\phi_h^{(j)}$ maps $a_h^{(j)}$ to another $\widetilde{a}_h^{(j)}$.

Remark 3 *If $\phi_h^{(j)}$ places probability 1 on a single action, we say that it is a deterministic modification.*

Throughout, we denote $\phi^{(j)} := \{\phi_h^{(j)}\}_{h=1}^H$ and $\phi^{(j)} \circ \pi := \{\phi_h^{(j)} \circ \pi_h\}_{h=1}^H$ as the collections of stochastic modifications and modified policies over the episode, respectively. We are now ready to introduce the definition of correlated equilibrium (CE).

Definition 4 (Correlated Equilibrium (CE)) *A joint policy π is called a CE if for any player j and any stochastic modification $\phi^{(j)}$, it holds that $\boldsymbol{v}_{\pi,1}^{(j)}(s) \geq \boldsymbol{v}_{\phi^{(j)}\circ\pi,1}^{(j)}(s)$ for all states $s \in \mathcal{S}$.*

Intuitively, at CE, no player can improve its value function by modifying its own action induced by the joint CE policy. Compare to NE policies, CE policies do not require joint independence among all the players. In fact, it has been shown that any NE policy is guaranteed to be a CE policy (Jin et al., 2022a; Liu et al., 2021; Song et al., 2021), and hence CE is a weaker equilibrium notion than NE. Moreover, CE can be reformulated as linear programming and hence is tractable.

3 Markov Games with Model Uncertainty

In this section, we study episodic general-sum Markov games with uncertainty in the environment transition kernel. We aim to define a tractable notion of correlated equilibrium under such model uncertainty and study its fundamental properties.

3.1 Robust Correlated Equilibrium

We adopt the same episodic Markov game settings as described in Section 2, but consider an uncertain transition kernel. Specifically, at every time step h and for every state-action pair (s,a), the environment transition kernel $\widetilde{\mathbb{P}}_h(\cdot|s,a)$ is uncertain and belongs to a general uncertainty set $\mathcal{P}_h(s,a)$. Below we list some popular examples of uncertainty sets.

Example 1 (KL divergence) *The uncertainty set under KL divergence d_{KL} is defined as*

$$\mathcal{P}_h(s,a) := \big\{\widetilde{\mathbb{P}}_h(\cdot|s,a)\colon d_{KL}\big(\mathbb{P}_h(\cdot|s,a), \widetilde{\mathbb{P}}_h(\cdot|s,a)\big) \leq \rho\big\},$$

where $d_{KL}\big(\mathbb{P},\widetilde{\mathbb{P}}\big) := \sum_{s\in\mathcal{S}} \widetilde{\mathbb{P}}(s) \ln \frac{\widetilde{\mathbb{P}}(s)}{\mathbb{P}(s)}$ and $\mathbb{P}_h(\cdot|s,a)$ denotes a fixed transition kernel.

Example 2 (R-contamination model) *The uncertainty set of R-contamination model is*

$$\mathcal{P}_h(s,a) := \big\{(1-R)\mathbb{P}_h(\cdot|s,a) + Rq : q \in \Delta^{|\mathcal{S}|}\big\},$$

where $\mathbb{P}_h(\cdot|s,a)$ is a fixed transition kernel.

In the above examples, $\mathbb{P}_h(\cdot|s,a)$ can be understood as the original stationary transition kernel, and the parameters $\rho > 0$ and $R > 0$ characterize the level of uncertainty. In a Markov game with model uncertainty, the state transitions are determined by uncertain transition kernels queried from the uncertainty sets. Therefore, it is possible that a certain transition kernel in the uncertainty set can lead to frequent low-reward state transitions, which are unacceptable to the players. Hence, under model uncertainty, each player aims to learn a robust optimal policy that maximizes its expected accumulated reward in the worst case. Motivated by this intuition, we define the following *robust value function* for the j-th player at state s and time step h under joint policy π. For simplicity of notation, we denote $\mathcal{P} := \bigotimes_{h,s,a} \mathcal{P}_h(s,a)$ as the product of uncertainty sets.

$$\text{(Robust Value Function):}\quad \boldsymbol{V}_{\pi,h}^{(j)}(s) := \inf_{\widetilde{\mathbb{P}}\in\mathcal{P}} \mathbb{E}\Big[\sum_{\ell=h}^{H} r_\ell^{(j)}(s_\ell,a_\ell)\Big| s_h = s, \pi, \widetilde{\mathbb{P}}\Big]. \tag{2}$$

Intuitively, the robust value function characterizes the minimum expected total reward one can obtain over all possible transition kernels in the uncertainty set. We note that the above robust value function is defined for every single player in the Markov game. In particular, the worst-case (adversarial) transition kernels associated with the players' robust value functions are generally different from each other. To deal with model uncertainty, the players aim to achieve a certain equilibrium in terms of the robust value function. Specifically, we define the following robust Nash equilibrium (NE).

Definition 5 (Robust NE) *A joint policy π is called robust NE if (i) for all h, π_h is a product policy; (ii) for any player j with any policy $\widetilde{\pi}^{(j)}$, we have $\boldsymbol{V}_{\pi,1}^{(j)}(s) \geq \boldsymbol{V}_{\widetilde{\pi}^{(j)}\times\pi^{(\setminus j)},1}^{(j)}(s)$ for all $s \in \mathcal{S}$.*

It can be seen that robust NE is similar to the NE defined in Definition 1, with the main difference being that robust NE is defined based on the robust value function. However, robust NE is generally more difficult to compute than NE. For example, NE is known to be tractable in zero-sum Markov games. As a comparison, in a zero-sum Markov game with model uncertainty, the environment model uncertainty can be viewed as a third adversarial player that competes with both players and breaks the zero-sum structure. Therefore, solving robust NE is PPAD-hard in general, and this further motivates us to define the following tractable surrogate notion of robust correlated equlibrium (CE).

Definition 6 (Robust CE) *A joint policy π is called a robust CE if for any player j and any stochastic modification $\phi^{(j)}$, it holds that $\boldsymbol{V}_{\pi,1}^{(j)}(s) \geq \boldsymbol{V}_{\phi^{(j)}\circ\pi,1}^{(j)}(s)$ for all states $s \in \mathcal{S}$.*

Although robust CE is a straightforward generalization of the standard CE defined in Definition 4, it incorporates model uncertainty into the nature of correlated equilibrium and turns out to have more complex structures than the standard CE as we elaborate in the next subsection.

3.2 Properties of Robust Correlated Equilibrium

Stochastic modification is the key element to define CE. In particular, it has been shown that the standard CE defined by stochastic modification is equivalent to that defined by deterministic modification. Our next result shows that robust CE inherits this property and we will leverage this property to build our convergence analysis later.

Proposition 7 *In a Markov game with model uncertainty, for any robust CE π and any player j, there exists a deterministic modification $\phi^{(j)}$ such that $\boldsymbol{V}_{\pi,1}^{(j)}(s) = \boldsymbol{V}_{\phi^{(j)}\circ\pi,1}^{(j)}(s)$ for all $s \in \mathcal{S}$.*

Our next result shows that for robust CE, its characterization of equilibrium can be different from that of robust NE and critically depends on the uncertainty set.

Proposition 8 *Robust CE and robust NE have the following relations.*

1. *In any robust Markov game, the set of robust CE includes the set of robust NE, or equivalently, any robust NE is a robust CE.*

2. There exists a robust CE (in some Markov games) which is not a robust NE.

Remark 9 *From Proposition 8 Item 1, the existence of robust CE is directly implied by the existence of robust NE, which is ensured under some mild regularity assumptions (Perchet, 2014, 2020; Kardeş et al., 2011).*

Based on the example given in Proposition 8 Item 2, we further illuminate the influence the uncertainty set on the set of robust CE. We consider the two-player coordination game described in Figure 1 in which there are five states $\mathcal{S} = \{s_i\}_{i=0}^4$ and each player has two actions $\mathcal{A} = \{a_i^{(1)}\}_{i=0}^1 \times \{a_i^{(2)}\}_{i=0}^1$. Its transition probability $\{\mathcal{P}_h\}_{h=1,2}$is parameterized by p. Let π_1 be the joint policy taken at the step $h = 1$ and π_2 be the joint policy taken at the step $h = 2$. The transition $\mathbb{P}_{h,p}$ is described in the proof of Proposition 8 Item 2. We consider the following three types of uncertainty sets:

1. $\mathcal{P}_h \subset \{\mathbb{P}_{h,p} : p \in [0, \frac{10}{29})\}$. In this case, the unique robust NE policy is given as $\pi_1(a = [0,0]|s = s_4) = 1$ and π_2 can be arbitrary (since the action taken at $h = 2$ does not change the state). Moreover, this robust NE is also the unique robust CE.
2. $\mathcal{P}_h \subset \{\mathbb{P}_{h,p} : p \in (\frac{10}{29}, \frac{1}{2})\}$. In this case, there are two robust NE, i.e., $\pi_1(a = [0,1]|s = s_4) = 1$ and $\pi_1(a = [1,0]|s = s_4) = 1$ (π_2 can be arbitrary). Moreover, any convex combination of these two policies is a robust CE. Therefore, the set of robust NE is a strict subset of the robust CE.
3. Let $\mathcal{P}_h$ be the union of the uncertainty sets in cases (1) and (2). Then, the robust NE and robust CE are the same as those of case (1).

When training the RL agent, the model mismatch problem is usually unavoidable. In this example, such problem could be described as the training environment has a large model shift (that is, the training environment has p larger than $\frac{10}{29}$ while the target environment has the parameter p less than $\frac{10}{29}$). In this case, we always need to learn the robust policy over a sufficiently large uncertainty set such that the it includes the target environment to ensure it achieves the robust NE. When the environment becomes more complicated, solving a robust NE could be costly, which motivates us to find a more efficient robust equilibrium (such as the robust CE) that could be solved in polynomial time.

In the next proposition, we consider the special case of a single player. When there is only a single player ($m = 1$), we can prove that every robust CE policy achieves the optimal robust value function, by noticing that for any given policies π, μ over the action space $\mathcal{A}$ there always exists a stochastic modification ϕ such that $\phi \circ \pi(a) = \mu(a)$ for all $a \in \mathcal{A}$. Therefore, the robust V-learning algorithm can be applied to single-agent reinforcement learning to address model uncertainty. We obtain the following proposition.

Proposition 10 *Let π be the policy such that $\boldsymbol{v}_{\pi,1}^{(1)}(s) \geq \boldsymbol{v}_{\phi\circ\pi,1}^{(1)}(s)$ for any stochastic modification ϕ and all states $s \in \mathcal{S}$, then $\boldsymbol{v}_{\pi,1}^{(1)}(s) \geq \boldsymbol{v}_{\mu,1}^{(1)}(s)$ for any policy μ and all states $s \in \mathcal{S}$.*

4 Decentralized Robust V-Learning

In this section, we develop a fully decentralized algorithm for finding robust CE of Markov games with model uncertainty. Our algorithm is inspired by the V-learning algorithm for solving standard Markov games (Jin et al., 2022a), and adopts some techniques from the

robust control in MDPs (Wang and Zou, 2021; Nilim and Ghaoui, 2003) to address model uncertainty.

4.1 Algorithm Design

We let every player j keep a value table $V_h^{(j)} \in \mathbb{R}^{|\mathcal{S}|}$ for each time step h, and denote $\{V_{k,h}^{(j)}\}_{h=1}^H$ as the value tables held by player j in the k-th episode. The main steps of our algorithm consist of a critic step and an actor step. In the critic step, we aim to learn the robust state value function associated with the current policy. To do so, we apply the following robust TD-learning type updates with every state-action transition sample (s, a, s') to update the players' value tables.

$$\widetilde{V}_{k,h}^{(j)}(s) = (1-\alpha_t)\widetilde{V}_{k-1,h}^{(j)}(s) + \alpha_t \left(r_h^{(j)} + \sigma_{\mathcal{P}_h(s,a)}(V_{k-1,h+1}^{(j)}) + \beta_t \right), \tag{3}$$

$$V_{k,h}^{(j)}(s) = \min\left\{H + 1 - h, \widetilde{V}_{k,h}^{(j)}(s)\right\}, \tag{4}$$

where the first update performs a robust TD type update and the second update performs a simple upper truncation. Here, $\alpha_t > 0$ is a learning rate parameter where $t := N_{k,h}(s)$ denotes that state s has been visited at step h for t times at the beginning of the k-th episode, and the value function mapping $\sigma_{\mathcal{P}_h(s,a)}(\cdot)$ is defined via the following linear program for any value table V.

$$\sigma_{\mathcal{P}_h(s,a)}(V) := \inf_{\widetilde{\mathbb{P}}_h(\cdot|s,a) \in \mathcal{P}_h(s,a)} \langle \widetilde{\mathbb{P}}_h(\cdot|s,a), V(\cdot) \rangle. \tag{5}$$

Intuitively, the above mapping corresponds to the worst-case expected state value of the next state. In particular, when there is no model uncertainty and the transition kernel is $\mathbb{P}_h(\cdot|s,a)$, it reduces to the expected state value at the next state, that is, $\mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)}[V(s')]$. Moreover, this linear program can be numerically solved for several important classes of uncertainty sets, as we elaborate below.

Example 3 (KL divergence) *Consider the uncertainty set $\mathcal{P}_h(s,a)$ defined under the KL divergence in Example 1. Then, the linear program* (5) *reduces to the following optimization problem, as proved in Theorem 1 of Hu and Hong (2013).*

$$\max_{\alpha \geq 0} \ -\alpha \ln \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)} \left[e^{-V(s')/\alpha} \right] - \alpha\eta. \tag{6}$$

In practice, we can query some samples to approximate the expectation involved in the above one-dimensional problem and solve it to obtain a sample-based estimator $\widehat{\sigma}_{\mathcal{P}_h(s,a)}(V)$.

Example 4 (R-contamination model) *Consider the uncertainty set $\mathcal{P}_h(s,a)$ defined by the R-contamination model in Example 2. Then, the linear program* (5) *can be approximated by the sample-based estimator $\widehat{\sigma}_{\mathcal{P}_h(s,a)}(V) = R \max_{s \in \mathcal{S}} V(s) + (1-R)V(s')$ (Wang and Zou, 2021).*

In the actor step, we leverage the adversarial bandit algorithm developed in Jin et al. (2022a) to update the current policy. To briefly explain, in step h of episode k, every player

j takes an action $a_h^{(j)} \sim \pi_{k,h}^{(j)}(\cdot|s_h)$ and observes an adversarial loss $1 - \frac{r_h^{(j)} + \widehat{\sigma}_{\mathcal{P}_h(s_h,a_h)}(V_{k,h+1}^{(j)})}{H}$, both of which are then fed into the adversarial bandit algorithm to produce the policy $\pi_{k+1,h}^{(j)}(\cdot|s_h)$. The specific updates of this algorithm are shown in Algorithm 3 in Appendix E. In particular, Jin et al. (2022a) proved that it achieves a regret bound in the order of $\widetilde{\mathcal{O}}(B\sqrt{H/t})$ (see Lemma 17 in the appendix).

The entire decentralized robust V-learning algorithm is summarized in Algorithm 1 below, where we use the estimator $\widehat{\sigma}_{\mathcal{P}_h(s,a)}$ instead of $\sigma_{\mathcal{P}_h(s,a)}$ in the critic update. For a specific uncertainty model, this estimator can be solved with arbitrary accuracy; for example, for KL divergence model, it suffices to solve eq. (6) with knowing its centroid $\mathbb{P}$ and the radius ρ. After obtaining all the policies $\{\pi_{k,h}\}_{k,h}$, the final non-Markov output policy $\hat{\pi}$ is defined by randomly selecting an episode k at each step h and taking an action $a_h \sim \pi_{k,h}$ (see Algorithm 2 for more details).

Algorithm 1: Decentralized Robust V-Learning (j-th player)

Initialize: Set $\widetilde{V}_{1,h}^{(j)}(s) = V_{1,h}^{(j)}(s) = H+1-h$, $\pi_{1,h}(a|s) = \frac{1}{|\mathcal{A}|}$, $N_{1,h}^{(j)}(s) = 0$ for all s, a, h

for episode $k = 1, \ldots, K$ **do**

  Observe initial state s_1, $V_{k,H+1}^{(j)}(s) = 0$ for all s

  for step $h = 1, \ldots, H$ **do**

    Take action $a_h^{(j)} \sim \pi_{k,h}^{(j)}(\cdot|s_h)$

    Transfer to next state $s_{h+1} \sim \widetilde{\mathbb{P}}_h(\cdot|s_h, a_h)$ with $\widetilde{\mathbb{P}}_h \in \mathcal{P}_h(s_h, a_h)$

    Let $\widetilde{V}_{k+1,h}^{(j)} \leftarrow \widetilde{V}_{k,h}^{(j)}$, $V_{k+1,h}^{(j)} \leftarrow V_{k,h}^{(j)}$, $\pi_{k+1,h}^{(j)} \leftarrow \pi_{k,h}^{(j)}$

    Receive reward $r_h^{(j)}$ and set $t := N_{k+1,h}^{(j)}(s_h) \leftarrow N_{k,h}^{(j)}(s_h) + 1$

$$\widetilde{V}_{k+1,h}^{(j)}(s_h) = (1-\alpha_t)\widetilde{V}_{k,h}^{(j)}(s_h) + \alpha_t \left(r_h^{(j)} + \widehat{\sigma}_{\mathcal{P}_h(s_h,a_h)}(V_{k,h+1}^{(j)}) + \beta_t^{(j)}\right) \quad (7)$$

$$V_{k+1,h}^{(j)}(s_h) = \min\{H+1-h, \widetilde{V}_{k+1,h}^{(j)}(s_h)\} \quad (8)$$

$$\pi_{k+1,h}^{(j)}(\cdot|s_h) = \texttt{ADV\_BANDIT}\Big(t, a_h, 1 - \frac{r_h^{(j)} + \widehat{\sigma}_{\mathcal{P}_h(s_h,a_h)}(V_{k,h+1}^{(j)})}{H}, \pi_{k,h}^{(j)}(\cdot|s_h)\Big) \quad (9)$$

  end

end

Output: Joint policy $\hat{\pi}$ defined by Algorithm 2

Here the hyperparameters α_t and $\beta_t^{(j)}$ are the learning rates given in (3). In practice, their values could be tuned using grid-search. We also provide a theoretical setting in Theorem 12 and Theorem 13 to ensure the polynomial-time sample complexity. After obtaining all $\pi_{k,h}$ by calling Algorithm 3 in Algorithm 1, we define the final output policy $\hat{\pi}_k := \hat{\pi}_{k,1}$ as Algorithm 2 below by following Algorithm 3 of Jin et al. (2022a). To facilitate the technical proof in Lemma 22 and Lemma 24, we use a slightly more general notation here as the policy $\hat{\pi}_{k,h}$ for all $k \in [K]$ and $h \in [H]$; when setting $h = 1$, Algorithm 2 exactly matches Algorithm 3 of Jin et al. (2022a).

Algorithm 2: Implement output policy $\hat{\pi}_{k,h}$. (Algorithm 3 from Jin et al. (2022a))

Input: Time step h, episode k, states $s_{h:H}$, policies $\{\pi_{k',h:H}\}_{k'=1}^{k}$ obtained from Algorithm 1.

for step $h' = h, h+1, \ldots, H$ **do**

Set $t \leftarrow N_{k,h'}(s_{h'})$ and let $\{k_{h'}^{i}(s_{h'})\}_{i=1}^{t}$ $(k_{h'}^{1}(s_{h'}) < k_{h'}^{2}(s_{h'}) < \ldots < k_{h'}^{t}(s_{h'}) < k)$ be the episodes where state $s_{h'}$ is visited at h'-th step, i.e., $s_{k_{h'}^{i}(s_{h'}),h} = s_{h'}$.

Randomly select $i \in [t]$ with prob. α_t^i (defined in (23)) and set $k \leftarrow k_{h'}^{i}(s_{h'})$.

$a_{h'} \sim \pi_{k,h'}(\cdot|s_{h'})$.

end

Output: Joint actions $a_{h:H}$.

4.2 Convergence and Complexity Analysis

For any joint policy π, we measure its optimality gap toward achieving exact robust CE as follows, where we define $\boldsymbol{V}^{(j)}_{\phi^*\circ\pi,1}(s) := \max_{\phi^{(j)}} \boldsymbol{V}^{(j)}_{\phi^{(j)}\circ\pi,1}(s)$ as player j's value function associated with the policy π modified by player j's best-response modification ϕ^*.

$$\text{(Optimality gap):} \quad \max_{j\in[J]} \max_{s\in\mathcal{S}} \left[\boldsymbol{V}^{(j)}_{\phi^*\circ\pi,1}(s) - \boldsymbol{V}^{(j)}_{\pi,1}(s)\right] \geq 0. \tag{10}$$

In particular, policy π is a robust CE if the gap vanishes. We also need the following definitions to characterize the impact of model uncertainty on Algorithm 1's convergence rate.

Definition 11 *Regarding the uncertainty sets $\{\mathcal{P}_h(s,a)\}_{h,s,a}$, the value function mapping $\sigma_{\mathcal{P}_h(s,a)}$ and state exploration probability, we define the following quantities.*

- *Uncertainty diameter:* $D := \max_{h,s,a,a'} \max_{\mathbb{P}\in\mathcal{P}_h(s,a),\widetilde{\mathbb{P}}\in\mathcal{P}_h(s,a')} \|\mathbb{P}(\cdot) - \widetilde{\mathbb{P}}(\cdot)\|_\infty$.
- *Estimation error:* $\mathfrak{e} := \sup_{h,s,a,V} \left|\sigma_{\mathcal{P}_h(s,a)}(V) - \widehat{\sigma}_{\mathcal{P}_h(s,a)}(V)\right|$, *where the supremum is taken over all bounded value tables that satisfy $0 \leq V(s) \leq H+1$ for all s.*
- *State exploration:* $p_{\min} := \min_{s,h,k} \mathbb{P}(s_{k,h} = s)$, *which denotes the minimum probability of visiting an arbitrary state s at any step h of any episode k.*

The uncertainty diameter D defined above characterizes the diameter of the uncertainty set $\mathcal{P}_h$. That is, a larger D means that the transition kernel $\mathbb{P}_h$ can change over a wider range and therefore induces larger uncertainty. For example, for the uncertainty set defined by the R-contamination model in Example 3.2, the uncertainty diameter is analytically given by $D = R\max\left\{\max_{s'} \mathbb{P}_h(s'|s,a), 1 - \min_{s'} \mathbb{P}_h(s'|s,a)\right\}$, which monotonically increases with regard to the uncertainty set parameter R. When the uncertainty level is sufficiently small, we can simply bound the error caused by model uncertainty with a non-asymptotic error term $5DSH^2$. This bound also holds for any value of D but the robust V-learning algorithm may not achieve ϵ error when D is larger than $\frac{\epsilon}{SH^2}$. Our first theorem characterizes this situation.

Theorem 12 *Let $S := |\mathcal{S}|$ and $A := \max_{1\leq j\leq m} |\mathcal{A}^{(j)}|$ correspond to the size of the state space and action space, respectively. Choose $\beta_t^{(j)}$, α_t and α_t^i according to eqs.* (21)-(23).

The output policy $\hat{\pi}$ produced by Algorithm 1 satisfies the following convergence rate with probability at least $1 - c\delta$ for some constant $c > 0$. For any $D \geq 0$, we have the bound

$$\max_{j\in[J]} \max_{s\in\mathcal{S}} \left(\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi},1}(s) - \boldsymbol{V}^{(j)}_{\hat{\pi},1}(s)\right) \leq 5DSH^2 + \mathcal{O}\Big(H\Big(A\sqrt{\frac{H^3S}{K}\ln\frac{mKHSA^2}{\delta}} + \mathfrak{e}\Big)\Big).$$

Further, if the uncertainty diameter $D \leq \frac{\epsilon}{SH^2}$ and the approximation error $\mathfrak{e} = \mathcal{O}(\frac{\epsilon}{H})$, then the ϵ-accuracy is guaranteed with $K = \widetilde{\mathcal{O}}(SA^2H^5\epsilon^{-2})$ episodes.

In many situations, the uncertainty level cannot be guaranteed to be sufficiently small to the level $\mathcal{O}(\frac{\epsilon}{SH^2})$. In this case, we show that if the policy can maintain sufficient exploration during training, then the requirement on the diameter can be relatively relieved; more explicitly, if there is sufficient exploration $p_{\min} > \frac{\epsilon}{SH}$, the requirement on D of achieving polynomial complexity will be increased from $\frac{\epsilon}{SH^2}$ to $\frac{p_{\min}}{H}$.

Theorem 13 *Let $S := |\mathcal{S}|$ and $A := \max_{1\leq j\leq m} |\mathcal{A}^{(j)}|$ correspond to the size of the state space and action space, respectively. Choose $\beta_t^{(j)}$, α_t and α_t^i according to eqs.* (21)-(23)*. The output policy $\hat{\pi}$ produced by Algorithm 1 satisfies the following convergence rate with probability at least $1 - c\delta$ for some constant $c > 0$. For any D and $p_{\min}$ satisfying $\frac{\epsilon}{SH^2} \leq D < \frac{p_{\min}}{H}$, we have the bound*

$$\max_{j\in[J]} \max_{s\in\mathcal{S}} \left(\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi},1}(s) - \boldsymbol{V}^{(j)}_{\hat{\pi},1}(s)\right) \leq \mathcal{O}\Big(\frac{H}{p_{\min} - DH}\Big(A\sqrt{\frac{H^3S}{K}\ln\frac{mKHSA^2}{\delta}} + \mathfrak{e}\Big)\Big).$$

Further, if the state exploration $p_{\min} > \frac{\epsilon}{SH}$ and the approximation error $\mathfrak{e} = \mathcal{O}(\frac{\epsilon p_{\min}}{H})$, then the ϵ-accuracy is guaranteed with $K = \widetilde{\mathcal{O}}(SA^2H^5p_{\min}^{-2}\epsilon^{-2})$ episodes.

Theorem 12 and Theorem 13 characterize the convergence rate and episode complexity of decentralized robust V-learning. When the state exploration $p_{\min}$ is decreasing, the non-asymptotic bound will increase; also, we need to require a stronger restriction on the uncertainty diameter D (since the bound holds only for $D < \frac{p_{\min}}{H}$). We note that the optimality gap adopted in the above theorem takes the maximum over all the states and hence is stronger than that used in the original V-learning (Jin et al., 2022b). This is because we need to develop new techniques to address the state transition uncertainty caused by the environment model uncertainty. As a result, the environment model uncertainty diameter D should not be too large compared to the state exploration probability $p_{\min}$ and the target accuracy ϵ.

Technical novelty of Theorem 12 and Theorem 13. Our analysis leverages the following technical developments to address model uncertainty and establish the convergence rate.

- To address model uncertainty, our decentralized robust V-learning algorithm adopts the worst-case expected value function estimator $\widehat{\sigma}_{\mathcal{P}_h(s,a)}(V)$ in the critic update, as opposed to the exact value $V(s)$ used in the standard V-learning. Such a nonlinear operator allows us to track the robust value function in the analysis. In particular, we developed various important properties of this operator in Lemma 19, including boundedness, monotonicity, etc., which are crucial to establish the key Lemmas 20, 21, 22 and 24 that lead to the desired convergence rate result.

- The proof of the original V-learning algorithm (Jin et al., 2022a) tracks the upper bound of the optimality gap $\delta_{k,h} := V_{k,h}(s_{k,h}) - \underline{V}_{k,h}(s_{k,h})$ at a single state and builds a recursion on it. This approach cannot be applied to our case as this gap turns into an uncertainty form $\sigma_{\mathcal{P}_h(s,a)}(V_{k,h+1}) - \sigma_{\mathcal{P}_h(s,a)}(\underline{V}_{k,h+1})$, which inevitably involves all possible states. To address this issue, we decompose this term as $\sigma_{\mathcal{P}_h(s,a)}(V_{k,h+1}) - \sigma_{\mathcal{P}_h(s,a)}(\underline{V}_{k,h+1}) = \delta_{k,h+1} + \sigma_{\mathcal{P}_h(s,a)}(V_{k,h+1}) - \sigma_{\mathcal{P}_h(s,a)}(\underline{V}_{k,h+1}) - \delta_{k,h+1}$ to link it to the desired term $\delta_{k,h+1}$ (see eq. (14)). Consequently, we need to solve a more challenging recursion that we build in Lemma 26.

- The decomposition mentioned in the previous bullet point involves an error term $\sigma_{\mathcal{P}_h(s,a)}(V_{k,h+1}) - \sigma_{\mathcal{P}_h(s,a)}(\underline{V}_{k,h+1}) - \delta_{k,h+1}$ that critically depends on the level of uncertainty. For example, this error term vanishes when there is no model uncertainty. We develop an upper bound of this error term in Lemma 25 by leveraging the uncertainty diameter (Definition 11) and introducing a stronger convergence metric $\Delta_{k,h}^{(j)} := \sum_{s\in\mathcal{S}} \left(V_{k,h}^{(j)}(s) - \underline{V}_{k,h}^{(j)}(s)\right)$ compared to $\delta_{k,h}$.

- To derive the desired convergence rate, we build a recursion on the convergence metric $\{\sum_{k=1}^{K} \Delta_{k,h}^{(j)}\}_{h\in[H]}$ in Lemma 27. The key challenge in solving this recursion is to rewrite it as a vectorized linear system that involves an upper triangular Toeplitz matrix, whose spectrum can be characterized analytically and used to derive the final result.

5 Conclusion

In this work, we proposed a new and tractable notion of robust correlated equilibrium for Markov games with environment model uncertainty. We showed that the robust correlated equilibrium has a simple modification structure, and its characterization of equilibrium critically depends on the environment model uncertainty. Moreover, we proposed the first fully-decentralized robust V-learning algorithm for computing such robust correlated equilibrium and established a polynomial sample complexity for computing an approximate robust correlated equilibrium. We believe this work provides an initial solution to competitive multi-agent reinforcement learning in uncertain environment, and an interesting future direction is to explore if it is possible to establish convergence of the algorithm under relaxed requirements on the uncertainty diameter.

Acknowledgments and Disclosure of Funding

S. Ma and Y. Zhou's work are supported by the National Science Foundation under Grants CCF-2106216, DMS-2134223, ECCS-2237830 (CAREER).

S. Zou's work is supported by the National Science Foundation under Grants CCF-2106560, CCF-2007783.

This material is based upon work supported under the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award # 2229873 - National AI Institute for Exceptional Education. Any opinions, findings and conclusions or recommendations expressed in this material are those

of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

Appendix

Table of Contents

Appendix A. Proof Sketch of Theorem 12 and Theorem 13

Before giving the detailed proof for Theorem 12 and Theorem 13, we provide a sketch of proof with highlighting the technical novelty and the main differences from the standard convergence analysis of V-learning (Jin et al., 2022a) for solving the non-robust CE of Markov games.

- Instead of directly bounding $\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,h},h}(s) - \boldsymbol{V}^{(j)}_{\hat{\pi}_{k,h},h}(s)$, we follow the proof steps of V-learning (Jin et al., 2022a) to start with the gap between the optimistic and pessimistic estimation
$$\delta^{(j)}_{k,h} := V^{(j)}_{k,h}(s_{k,h}) - \underline{V}^{(j)}_{k,h}(s_{k,h}),$$
where $s_{k,h}$ is the state visited at k-th episode and h-th step. Since the update rule of robust V-learning algorithm depends on the worst-case expected value function estimator $\widehat{\sigma}_{\mathcal{P}_h(s,a)}(V)$ in the critic update, as opposed to the exact value $V(s)$ used in the standard V-learning. These optimistic and pessimistic estimations are not as straightforward as the original proof and require us to develop various important properties of this operator in Lemma 19, including boundedness, monotonicity, etc. These properties will be used to lead the desired pessimistic and optimistic estimations $\boldsymbol{V}^{(j)}_{\hat{\pi}_{k,h},h}(s) \geq \underline{V}^{(j)}_{k,h}(s)$ (Lemma 22) and $V^{(j)}_{k,h}(s) \geq \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,h},h}(s)$ (Lemma 24).

- With following the same steps as the proof of V-learing (Jin et al., 2022a), we obtain the following recursion:

$$\begin{aligned}\sum_{k=1}^{K}\delta_{k,h}^{(j)} \leq & \Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\Big(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V_{k',h+1}^{(j)})-\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}_{k',h+1}^{(j)})\Big)\\ & -\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\Big(V_{k',h+1}^{(j)}(s_{k',h+1})-\underline{V}_{k',h+1}^{(j)}(s_{k',h+1})\Big)\\ & +\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\delta_{k,h+1}^{(j)}+\Theta\Big(A_j\sqrt{H^3SK\ln\frac{mKHSA_j^2}{\delta}}\Big)+4K\mathfrak{e}.\end{aligned}$$

  For the non-robust V-learing (Jin et al., 2022a), the first two terms are directly canceled while we have to consider the influence of operator $\sigma_{\mathcal{P}_h(s,a)}$, which is one of the main differences in analyzing the robust V-learning.

- We developed Lemma 25 as one our of technical contributions to upper bound this difference term

$$\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V_{k',h+1}^{(j)})-\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}_{k',h+1}^{(j)})-V_{k',h+1}^{(j)}(s_{k',h+1})+\underline{V}_{k',h+1}^{(j)}(s_{k',h+1}).$$

  This lemma shows the operator $\sigma_{\mathcal{P}_h(s,a)}$ has Lipschitzness in L_1-norm with a high probability, which performs as the bridge between the robustness and non-robustness convergence analysis. After applying Lemma 25, we will be able to obtain the following bound (17):

$$\begin{aligned}\sum_{k=1}^{K}\delta_{k,h}^{(j)} \leq & D\Big(1+\frac{1}{H}\Big)\sum_{i=0}^{H-h}\Big(1+\frac{1}{H}\Big)^i\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V_{k,h+1+i}^{(j)}(s)-\underline{V}_{k,h+1+i}^{(j)}(s)\Big)\\ & +2H\Big[\Big(1+\frac{1}{H}\Big)\sqrt{32KH^2\ln\frac{2mKHSA}{\delta}}+\Theta\Big(A_j\sqrt{H^3SK\ln\frac{mKHSA_j^2}{\delta}}\Big)+4K\mathfrak{e}\Big].\end{aligned}$$

- We note that the first term in the above bound

$$D\Big(1+\frac{1}{H}\Big)\sum_{i=0}^{H-h}\Big(1+\frac{1}{H}\Big)^i\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V_{k,h+1+i}^{(j)}(s)-\underline{V}_{k,h+1+i}^{(j)}(s)\Big)$$

  will not appear in the non-robust V-learning since it characterizes the influence of diameter of uncertainty set on the convergence error of robust V-learning. For non-robust V-learning, the diameter of uncertainty set is 0, so this term will simply disappear. We propose two approaches to handle this error term and they will lead to different conditions on obtaining a polynomial-time complexity.

  - **Theorem 12.** In Theorem 12, we bound the tracked value functions by their upper bound to obtain the following characterization of the error term:

$$D\Big(1+\frac{1}{H}\Big)\sum_{i=1}^{H}\Big(1+\frac{1}{H}\Big)^{i-1}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V_{k,1+i}^{(j)}(s)-\underline{V}_{k,1+i}^{(j)}(s)\Big)$$

$$\leq 5DSKH^2.$$

It implies that if the uncertainty set is sufficiently small (that is, the diameter of uncertainty set D is less than the threshold $\frac{\epsilon}{5SKH^2}$), we could always control the error term caused by the model uncertainty by ϵ.

- **Theorem 13.** In many cases, we do not expect the uncertainty set to be too small. So we developed Theorem 13 by considering the expectation of this error term. With taking expectation, Equation (17) will be simplified to Equation (19)

$$p_{\min} \sum_{k=1}^{K} \Delta_{k,h}^{(j)} \leq D \sum_{i=0}^{H-h} (1 + \frac{1}{H})^{i+1} \sum_{k=1}^{K} \Delta_{k,h+1+i}^{(j)} + \mathscr{U}_j,$$

where $p_{\min}$ characterizes the exploration ability of the whole learning process. To solve this recursion, we developed Lemma 27 to characterize the solution of a upper triangular Toeplitz system, which is also a new technique in convergence analysis and will be potentially applied to other algorithms. To satisfy the condition of Lemma 27, we need to require the diameter of uncertainty set to be $\frac{\epsilon}{SH^2} \leq D < \frac{p_{\min}}{H}$. This condition indicates that for learning a robust CE of general-sum multi-agent Markov games, we cannot let the uncertainty set to be arbitrarily large.

Appendix B. Proof of Proposition 7

First, we re-state Proposition 7 here.

Proposition 14 *In a Markov game with model uncertainty, any robust CE policy π can be achieved by deterministic modifications, i.e., for any player j there exists a deterministic modification $\phi^{(j)}$ such that $\boldsymbol{V}_{\pi,1}^{(j)}(s) = \boldsymbol{V}_{\phi^{(j)}\circ\pi,1}^{(j)}(s)$.*

Proof

We will prove this proposition for a more general setting where each reward $r_h^{(j)} = r_h^{(j)}(s_{1:h}, a_{1:h})$ relies on all the past and current states $s_{1:h} := \{s_{h'}\}_{h'=1}^{h}$ and actions $a_{1:h} := \{a_{h'}\}_{h'=1}^{h}$. Then the conclusion directly applies to the special case of interest where $r_h^{(j)} = r_h^{(j)}(s_h, a_h)$.

We will find $\phi^{(j)}$ by applying mathematical induction to the horizon H.

When $H = 1$, the MDP does not involve transition kernel, so

$$\begin{aligned}
\boldsymbol{V}_{\phi^{(j)}\circ\pi,1}^{(j)}(s) &= \sum_{a_1} (\phi_1^{(j)} \circ \pi_1)(a_1|s) r_1^{(j)}(s, a_1) \\
&\overset{(i)}{=} \sum_{a_1, \widetilde{a}_1^{(j)}} \phi_1^{(j)}(a_1^{(j)}|s, \widetilde{a}_1^{(j)}) \pi_1([\widetilde{a}_1^{(j)}, a_1^{(\backslash j)}]|s) r_1^{(j)}(s, a_1) \\
&= \sum_{\widetilde{a}_1^{(j)}} \Big(\sum_{a_1^{(j)}} \phi_1^{(j)}(a_1^{(j)}|s, \widetilde{a}_1^{(j)}) \sum_{a_1^{(\backslash j)}} \pi_1([\widetilde{a}_1^{(j)}, a_1^{(\backslash j)}]|s) r_1^{(j)}(s, a_1) \Big)
\end{aligned}$$

$$\overset{(ii)}{\leq} \sum_{\widetilde{a}_1^{(j)}} \Big(\max_{a_1^{(j)}} \sum_{a_1^{(\backslash j)}} \pi_1([\widetilde{a}_1^{(j)}, a_1^{(\backslash j)}]|s) r_1^{(j)}(s, a_1) \Big),$$

where (i) uses the following formula that directly follows from the definition of stochastic modification $\phi^{(j)}$, and (ii) becomes "=" using the deterministic modification such that $\phi_1^{(j)}(a_1^{(j)}|s, \widetilde{a}_1^{(j)}) := 1$ for a certain $a_1^{(j)} \in \arg\max_{a_1^{(j)}} \sum_{a_1^{(\backslash j)}} \pi_1([\widetilde{a}_1^{(j)}, a_1^{(\backslash j)}]|s)) r_1^{(j)}(s, a_1)$.

$$(\phi_h^{(j)} \circ \pi_h)\big(a_h|s_{1:h}, a_{1:h-1}\big) = \sum_{\widetilde{a}_h^{(j)}} \phi_h^{(j)}(a_h^{(j)}|s_{1:h}, a_{1:h-1}, \widetilde{a}_h^{(j)}) \pi_h([\widetilde{a}_h^{(j)}, a_h^{(\backslash j)}]|s_{1:h}, a_{1:h-1}). \quad (11)$$

This proves the existence of the optimal deterministic solution $\phi^{(j)}$ for $H = 1$. Suppose it also exists for horizon $H - 1$. Then it suffices to prove the existence of $\phi^{(j)}$ for H as follows.

$$\begin{aligned}
&\boldsymbol{V}^{(j)}_{\phi^{(j)} \circ \pi, 1}(s) : \\
&= \inf_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\Big[\sum_{h=1}^{H} r_h^{(j)}(s_{1:h}, a_{1:h}) \Big| s_1 = s, \phi^{(j)} \circ \pi, \mathbb{P} \Big] \\
&= \inf_{\mathbb{P}_{1:H-2} \in \mathcal{P}_{1:H-2}} \Bigg(\mathbb{E}\Big[\sum_{h=1}^{H-1} r_h^{(j)}(s_{1:h}, a_{1:h}) \Big| s_1 = s, \{\phi_h^{(j)} \circ \pi_h\}_{h=1}^{H-1}, \mathbb{P}_{1:H-2} \Big] \\
&\quad + \inf_{\mathbb{P}_{H-1} \in \mathcal{P}_{H-1}} \mathbb{E}\Big[r_H^{(j)}(s_{1:H}, a_{1:H}) \Big| s_1 = s, \phi^{(j)} \circ \pi, \mathbb{P}_{1:H-1} \Big] \Bigg) \\
&= \inf_{\mathbb{P}_{1:H-2} \in \mathcal{P}_{1:H-2}} \Bigg(\mathbb{E}\Big[\sum_{h=1}^{H-1} r_h^{(j)}(s_{1:h}, a_{1:h}) \Big| s_1 = s, \{\phi_h^{(j)} \circ \pi_h\}_{h=1}^{H-1}, \mathbb{P}_{1:H-2} \Big] \\
&\quad + \sum_{s_{1:H-1}, a_{1:H-1}} \Pr(s_{1:H-1}, a_{1:H-1}|s_1 = s, \{\phi_h^{(j)} \circ \pi_h\}_{h=1}^{H-1}, \mathbb{P}_{1:H-2}) \\
&\quad \inf_{\substack{\mathbb{P}_{H-1}(\cdot|s_{H-1}, a_{H-1}) \\ \in \mathcal{P}_{H-1}(s_{H-1}, a_{H-1})}} \sum_{s_H} \mathbb{P}_{H-1}(s_H|s_{H-1}, a_{H-1}) \sum_{a_H} (\phi_H^{(j)} \circ \pi_H)(a_H|s_{1:H}, a_{1:H-1}) r_H^{(j)}(s_{1:H}, a_{1:H}) \Bigg) \\
&\overset{(i)}{=} \inf_{\mathbb{P}_{1:H-2} \in \mathcal{P}_{1:H-2}} \Bigg(\mathbb{E}\Big[\sum_{h=1}^{H-1} r_h^{(j)}(s_{1:h}, a_{1:h}) \Big| s_1 = s, \{\phi_h^{(j)} \circ \pi_h\}_{h=1}^{H-1}, \mathbb{P}_{1:H-2} \Big] \\
&\quad + \sum_{s_{1:H-1}, a_{1:H-1}} \Pr(s_{1:H-1}, a_{1:H-1}|s_1 = s, \{\phi_h^{(j)} \circ \pi_h\}_{h=1}^{H-1}, \mathbb{P}_{1:H-2}) \\
&\quad \inf_{\substack{\mathbb{P}_{H-1}(\cdot|s_{H-1}, a_{H-1}) \\ \in \mathcal{P}_{H-1}(s_{H-1}, a_{H-1})}} \sum_{s_H} \mathbb{P}_{H-1}(s_H|s_{H-1}, a_{H-1}) \\
&\quad \sum_{a_H, \widetilde{a}_H^{(j)}} \phi_H^{(j)}(a_H^{(j)}|s_{1:H}, a_{1:H-1}, \widetilde{a}_H^{(j)}) \pi_H([\widetilde{a}_H^{(j)}, a_H^{(\backslash j)}]|s_{1:H}, a_{1:H-1}) r_H^{(j)}(s_{1:H}, a_{1:H}) \Bigg) \\
&= \inf_{\mathbb{P}_{1:H-2} \in \mathcal{P}_{1:H-2}} \Bigg(\mathbb{E}\Big[\sum_{h=1}^{H-1} r_h^{(j)}(s_{1:h}, a_{1:h}) \Big| s_1 = s, \{\phi_h^{(j)} \circ \pi_h\}_{h=1}^{H-1}, \mathbb{P}_{1:H-2} \Big]
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{s_{1:H-1},a_{1:H-1}} \Pr(s_{1:H-1},a_{1:H-1}|s_1=s,\{\phi_h^{(j)}\circ\pi_h\}_{h=1}^{H-1},\mathbb{P}_{1:H-2}) \\
&\inf_{\substack{\mathbb{P}_{H-1}(\cdot|s_{H-1},a_{H-1})\\ \in\mathcal{P}_{H-1}(s_{H-1},a_{H-1})}} \sum_{s_H,\widetilde{a}_H^{(j)}} \mathbb{P}_{H-1}(s_H|s_{H-1},a_{H-1}) \sum_{a_H^{(j)}} \phi_H^{(j)}(a_H^{(j)}|s_{1:H},a_{1:H-1},\widetilde{a}_H^{(j)}) \\
&\sum_{a_H^{(\backslash j)}} \pi_H([\widetilde{a}_H^{(j)},a_H^{(\backslash j)}]|s_{1:H},a_{1:H-1}) r_H^{(j)}(s_{1:H},a_{1:H})\Bigg) \\
\overset{(ii)}{\leq} &\inf_{\mathbb{P}_{1:H-2}\in\mathcal{P}_{1:H-2}} \Bigg(\mathbb{E}\Big[\sum_{h=1}^{H-1} r_h^{(j)}(s_{1:h},a_{1:h})\Big|s_1=s,\{\phi_h^{(j)}\circ\pi_h\}_{h=1}^{H-1},\mathbb{P}_{1:H-2}\Big] \\
&+ \sum_{s_{1:H-1},a_{1:H-1}} \Pr(s_{1:H-1},a_{1:H-1}|s_1=s,\{\phi_h^{(j)}\circ\pi_h\}_{h=1}^{H-1},\mathbb{P}_{1:H-2}) \inf_{\substack{\mathbb{P}_{H-1}(\cdot|s_{H-1},a_{H-1})\\ \in\mathcal{P}_{H-1}(s_{H-1},a_{H-1})}} \\
&\sum_{s_H,\widetilde{a}_H^{(j)}} \mathbb{P}_{H-1}(s_H|s_{H-1},a_{H-1}) \max_{a_H^{(j)}} \sum_{a_H^{(\backslash j)}} \pi_H([\widetilde{a}_H^{(j)},a_H^{(\backslash j)}]|s_{1:H},a_{1:H-1}) r_H^{(j)}(s_{1:H},a_{1:H})\Bigg) \\
\overset{(iii)}{=} &\inf_{\mathbb{P}_{1:H-2}\in\mathcal{P}_{1:H-2}} \mathbb{E}\Big[\sum_{h=1}^{H-2} r_h^{(j)}(s_{1:h},a_{1:h}) + \widetilde{r}_{H-1}^{(j)}(s_{1:H-1},a_{1:H-1})\Big|s_1=s,\{\phi_h^{(j)}\circ\pi_h\}_{h=1}^{H-1},\mathbb{P}_{1:H-2}\Big],
\end{aligned} \tag{12}
$$

where we denote $\mathcal{P}_{1:h} := \{\mathcal{P}_{h'}\}_{h'=1}^{h}$ and $\mathbb{P}_{1:h} := \{\mathbb{P}_{h'}\}_{h'=1}^{h}$, (i) uses eq. (11), (ii) becomes "=" using the deterministic modification $\phi_H^{(j)}$ such that $\phi_H^{(j)}(a_H^{(j)}|s_{1:H},a_{1:H-1},\widetilde{a}_H^{(j)}) := 1$ for a certain $a_H^{(j)} \in \arg\max_{a_H^{(j)}} \sum_{a_H^{(\backslash j)}} \pi_H([\widetilde{a}_H^{(j)},a_H^{(\backslash j)}]|s_{1:H},a_{1:H-1}) r_H^{(j)}(s_{1:H},a_{1:H})$, and (iii) denotes the following surrogate reward at step $H-1$,

$$
\begin{aligned}
&\widetilde{r}_{H-1}^{(j)}(s_{1:H-1},a_{1:H-1}) \\
&= r_{H-1}^{(j)}(s_{1:H-1},a_{1:H-1}) + \inf_{\substack{\mathbb{P}_{H-1}(\cdot|s_{H-1},a_{H-1})\\ \in\mathcal{P}_{H-1}(s_{H-1},a_{H-1})}} \sum_{s_H,\widetilde{a}_H^{(j)}} \mathbb{P}_{H-1}(s_H|s_{H-1},a_{H-1}) \\
&\max_{a_H^{(j)}} \sum_{a_H^{(\backslash j)}} \pi_H([\widetilde{a}_H^{(j)},a_H^{(\backslash j)}]|s_{1:H},a_{1:H-1}) r_H^{(j)}(s_{1:H},a_{1:H}).
\end{aligned} \tag{13}
$$

Note that eq. (12) can be seen as the value function with horizon $H-1$, so there are deterministic modifications $\phi_h^{(j)*}$ for $h \in [H-1]$ that maximize eq. (12), which along with $\phi_H^{(j)}$ above forms the deterministic modification $\phi^{(j)} := \{\phi_h^{(j)}\}_{h\in[H]}$ that maximizes $\boldsymbol{V}_{\phi^{(j)}\circ\pi,1}^{(j)}(s)$. This completes the proof. ■

Appendix C. Proof of Proposition 8

Proposition 15 *Robust CE and robust NE have the following relations.*

1. *In any robust Markov game, the set of robust CE includes the set of robust NE, or equivalently, any robust NE is a robust CE.*

2. There exists a robust CE (in some Markov games) which is not a robust NE.

Proof First, we will prove item 1 that the set of robust CE always includes the set of robust NE. This part of proof is directly taken from Proposition 9 (Jin et al., 2022a) with changing the V-function to the robust V-function. Let $\pi = \pi_1 \times \pi_2 \times \cdots \times \pi_m$ be a robust Nash equilibrium; then

$$\begin{aligned}\max_{\phi_i} \boldsymbol{V}^{(i)}_{(\phi_i \circ \pi_i)\times \pi^{(\setminus i)}}(s) &\overset{(i)}{=} \max_{\pi_i'} \boldsymbol{V}^{(i)}_{\pi_i' \times \pi^{(\setminus i)}}(s) \\ &\overset{(ii)}{\leq} \boldsymbol{V}^{(i)}_{\pi}(s),\end{aligned}$$

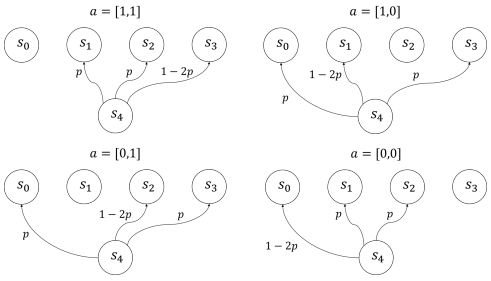

Figure 1: Transition kernel at $h = 1$

where (i) is because that π is a product policy and (ii) applies the definition of robust Nash equilibrium (see Definition 1). It implies that π is also a robust CE by Definition 4.

Next, we prove item 2. It suffices to give an example of a Markov game setting and a robust CE policy π that is not NE. Consider a two-player coordination game in which there are five states $\mathcal{S} = \{s_i\}_{i=0}^4$ and each player has two actions $\mathcal{A} = \{a_i^{(1)}\}_{i=0}^1 \times \{a_i^{(2)}\}_{i=0}^1$. At time step $h = 1$, Figure 1 depicts the transition kernel $\mathbb{P}_{1,p}$ parameterized by a parameter $p \in [0, \frac{1}{2})$. At time step $h = 2$, we set the transition kernel $\mathbb{P}_{2,p}(s|s,a) = 1$ for all s and a, i.e., players stay in their current state no matter what actions are taken. The rewards of both players are set as $r(s_0, a) = [0.5, 0.5]$, $r(s_1, a) = [0, 1]$, $r(s_2, a) = [1, 0]$, $r(s_3, a) = [0.95, 0.95]$, and $r(s_4, a) = [0, 0]$ for any action $a \in \mathcal{A}$. The initial state is fixed to be $s = s_4$. We consider the uncertainty set $\mathcal{P}_h = \{\mathbb{P}_{h,p} : p \in (\frac{10}{29}, \frac{1}{2})\}$ where there are two robust NE, i.e., $\pi_1(a = [0, 1]|s = s_4) = 1$ and $\pi_1(a = [1, 0]|s = s_4) = 1$ (π_2 can be arbitrary). Moreover, any convex combination of these two policies is a robust CE but not robust NE. ■

Appendix D. Proof of Proposition 10

Proposition 16 *Let π be the policy such that $\boldsymbol{v}^{(1)}_{\pi,1}(s) \geq \boldsymbol{v}^{(1)}_{\phi\circ\pi,1}(s)$ for any stochastic modification ϕ and all states $s \in \mathcal{S}$, then $\boldsymbol{v}^{(1)}_{\pi,1}(s) \geq \boldsymbol{v}^{(1)}_{\mu,1}(s)$ for any policy μ and all states $s \in \mathcal{S}$.*

Proof It suffices to prove that for the single-agent case, all stochastic modifications of a policy form the space of all policies. Let Π be all distributions over $\mathcal{A}$ and $\pi \in \Pi$ is given with $\pi(a) > 0$ for all $a \in \mathcal{A}$. We will prove that

$$\{\phi \circ \pi : \phi \text{ is a stochastic modification}\} = \Pi.$$

For any $\mu \in \Pi$, we can construct the desired ϕ as follows: define $\phi(\cdot|b) = \mu$ for all $b \in \mathcal{A}$. Then we will show that $\phi \circ \pi$ is same μ.

$$\begin{aligned}\phi \circ \pi(a) &= \sum_{b\in\mathcal{A}} \pi(b)\phi(a|b) \\ &= \sum_{b\in\mathcal{A}} \pi(b)\mu(a) = \mu(a).\end{aligned}$$

This implies that $\Pi = \{\phi \circ \pi : \phi \text{ is a stochastic modification}\}$. It concludes that enumerating all stochastic modifications of a given policy is equivalent to enumerating all policies. ■

Appendix E. Policies in Algorithm 1

In this section, we will elaborate how is $\pi_{k,h}$ in Algorithm 1 obtained from the adversarial bandit algorithm.

Obtaining $\pi_{k,h}$: In Algorithm 1, to output robust equilibrium (robust CE), we adopt V-learning algorithm (Jin et al., 2022a) for single-agent adversarial bandit to update the current policy. To elaborate, we denote the i-th iteration of this algorithm as $\pi_{i+1}(\cdot) \leftarrow \texttt{ADV\_BANDIT}(b_i, \ell_i(\cdot))$, where the player takes action $b_i \in \mathcal{B}$ following its own policy $\pi_i(\cdot)$ obtained from its previous iteration and observes the noisy bandit-feedback $\ell_i(b_i)$ with the loss function ℓ_i selected by the adversary. The procedure of implementing `ADV_BANDIT` algorithm for multiple iterations is shown in Algorithm 6 of Jin et al. (2022a) and we extracted the i-th iteration as shown in the following Algorithm 3.

Algorithm 3: Adversarial bandit algorithm (ADV_BANDIT)

Input: Iteration index i, action $\widetilde{b}$ and the corresponding bandit-feedback $\ell(\widetilde{b})$, the previous policy π_i.

for each action $b \in \mathcal{B}$ **do**

$\hat{\ell}_i(\widetilde{b}|b) \leftarrow \frac{\pi_i(b)\ell(\widetilde{b})}{\pi_i(\widetilde{b})+\gamma_i}$

$\hat{\ell}_i(b'|b) \leftarrow 0, \forall b' \in \mathcal{B}/\{\widetilde{b}\}$

$\widetilde{\pi}(\cdot|b) \propto \exp\big[-\frac{\eta_i}{w_i}\sum_{j=1}^{i} w_j \hat{\ell}_j(\cdot|b)\big]$, where $\hat{\ell}_j$ is obtained from the j-th iteration of ADV_BANDIT algorithm

end

Output: π_{i+1} obtained by solving the linear equation $\pi_{i+1}(\cdot) = \sum_{b\in\mathcal{B}} \pi_{i+1}(b)\widetilde{\pi}(\cdot|b)$

It has been proved by Corollary 25 of (Jin et al., 2022b) that Algorithm 3 has the following convergence rate.

Lemma 17 *Implement Algorithm 3 for iterations $i = 1, \ldots, t$ with hyperparameter choices $w_t = \frac{\alpha_t}{\prod_{i=2}^{t}(1-\alpha_i)}$ (α_i is defined in eq. (23)), $\gamma_t = \eta_t = \sqrt{(H \ln B)/t}$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\min_{\phi} \sum_{i=1}^{t} \alpha_t^i \big[\langle \pi_i(\cdot), l_i(\cdot)\rangle - \langle (\phi \circ \pi_i)(\cdot), l_i(\cdot)\rangle\big] \le 10B\sqrt{\frac{H\ln(B^2/\delta)}{t}},$$

where $\phi \circ \pi_i$ can be defined by reducing the definition of stochastic modification in Definition 2 to single-agent policy π_i.

Appendix F. Proof of Theorem 12

Theorem 12 *Let $S := |\mathcal{S}|$ and $A := \max_{1\le j\le m} |\mathcal{A}^{(j)}|$ correspond to the size of the state space and action space, respectively. Choose $\beta_t^{(j)}$, α_t and α_t^i according to eqs. (21)-(23).*

The output policy $\hat{\pi}$ produced by Algorithm 1 satisfies the following convergence rate with probability at least $1 - c\delta$ for some constant $c > 0$. For any $D \geq 0$, we have the bound

$$\max_{j\in[J]} \max_{s\in\mathcal{S}} \left(\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi},1}(s) - \boldsymbol{V}^{(j)}_{\hat{\pi},1}(s)\right) \leq 5DSH^2 + \mathcal{O}\Big(H\Big(A\sqrt{\frac{H^3S}{K}\ln\frac{mKHSA^2}{\delta}} + \mathfrak{e}\Big)\Big).$$

Further, if the uncertainty diameter $D \leq \frac{\epsilon}{SH^2}$ and the approximation error $\mathfrak{e} = \mathcal{O}(\frac{\epsilon}{H})$, then the ϵ-accuracy is guaranteed with $K = \widetilde{\mathcal{O}}(SA^2H^5\epsilon^{-2})$ episodes.

Proof Note that $V^{(j)}_{k,h}(s)$ (defined by eqs. (7) & (8)) and $\underline{V}^{(j)}_{k,h}(s)$ (defined by eqs. (27) & (28)) are respectively the upper bound and the lower bound of $\boldsymbol{V}^{(j)}_{\hat{\pi}_{k,h},h}(s)$ (Since $V^{(j)}_{k,h}(s) \geq \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,h},h}(s) \geq \boldsymbol{V}^{(j)}_{\hat{\pi}_{k,h},h}(s) \geq \underline{V}^{(j)}_{k,h}(s)$ based on Lemmas 22 & 24). Denote the gap between the upper bound and the lower bound as follows

$$\delta^{(j)}_{k,h} := V^{(j)}_{k,h}(s_{k,h}) - \underline{V}^{(j)}_{k,h}(s_{k,h}) \geq 0.$$

Let $s_{k,h}$, $a_{k,h}$ respectively be the state and action at the h-th step in the k-th episode. Let $\{k^i_{k,h}\}_{1\leq i\leq n_{k,h}}$ ($k^1_{k,h} < k^2_{k,h} < \ldots < k^{n_{k,h}}_{k,h} < k$) be the set of episodes in which the state $s_{k,h}$ is visited at the h-th step. $n_{k,h} := N_{k,h}(s_{k,h})$ is the number of such visits.

Then we unroll the update rule for both $V^{(j)}_{k,h}(s_{k,h})$ and $\underline{V}^{(j)}_{k,h}(s_{k,h})$ along k as follows.

$$\begin{aligned}
\delta^{(j)}_{k,h} &\overset{(i)}{\leq} \widetilde{V}^{(j)}_{k,h}(s_{k,h}) - \underline{V}^{(j)}_{k,h}(s_{k,h}) \\
&\overset{(ii)}{=} \alpha^0_{n_{k,h}}(H-h+1) \\
&\quad + \sum_{i=1}^{n_{k,h}} \alpha^i_{n_{k,h}} \left(\widehat{\sigma}_{\mathcal{P}_h(s_{k,h},a_{k^i_{k,h},h})}(V^{(j)}_{k^i_{k,h},h+1}) - \widehat{\sigma}_{\mathcal{P}_h(s_{k,h},a_{k^i_{k,h},h})}(\underline{V}^{(j)}_{k^i_{k,h},h+1}) + 2\beta^{(j)}_i\right) \\
&\overset{(iii)}{\leq} \sum_{i=1}^{n_{k,h}} \alpha^i_{n_{k,h}} \left(\sigma_{\mathcal{P}_h(s_{k,h},a_{k^i_{k,h},h})}(V^{(j)}_{k^i_{k,h},h+1}) - \sigma_{\mathcal{P}_h(s_{k,h},a_{k^i_{k,h},h})}(\underline{V}^{(j)}_{k^i_{k,h},h+1})\right) \\
&\quad + 2\sum_{i=1}^{n_{k,h}} \alpha^i_{n_{k,h}}\beta^{(j)}_i + 2\mathfrak{e}\sum_{i=1}^{n_{k,h}} \alpha^i_{n_{k,h}} \\
&\overset{(iv)}{\leq} \sum_{i=1}^{n_{k,h}} \alpha^i_{n_{k,h}} \left(\sigma_{\mathcal{P}_h(s_{k,h},a_{k^i_{k,h},h})}(V^{(j)}_{k^i_{k,h},h+1}) - \sigma_{\mathcal{P}_h(s_{k,h},a_{k^i_{k,h},h})}(\underline{V}^{(j)}_{k^i_{k,h},h+1})\right) \\
&\quad + \Theta\Big(A_j\sqrt{\frac{H^3}{n_{k,h}}\ln\frac{mKHSA^2_j}{\delta}}\Big) + 4\mathfrak{e},
\end{aligned}$$

where (i) uses eqs. (8) & (28), (ii) unrolls the update rules (7) & (27), (iii) uses eq. (29) and $\alpha^0_t = 0$ (eq. (23)), and (iv) uses eqs. (24) & (26). By summing over k, we obtain the following recursion:

$$\sum_{k=1}^{K}\delta^{(j)}_{k,h} \leq \sum_{k=1}^{K}\sum_{i=1}^{n_{k,h}} \alpha^i_{n_{k,h}} \left(\sigma_{\mathcal{P}_h(s_{k,h},a_{k^i_{k,h},h})}(V^{(j)}_{k^i_{k,h},h+1}) - \sigma_{\mathcal{P}_h(s_{k,h},a_{k^i_{k,h},h})}(\underline{V}^{(j)}_{k^i_{k,h},h+1})\right)$$

$$
\begin{aligned}
&+\sum_{k=1}^{K}\Theta\Big(A_j\sqrt{\frac{H^3}{n_{k,h}}\ln\frac{mKHSA_j^2}{\delta}}\ \Big)+4K\mathfrak{e}\\
&\overset{(i)}{\leq}\sum_{k'=1}^{K}\Big(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V^{(j)}_{k',h+1})-\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}^{(j)}_{k',h+1})\Big)\sum_{i=n_h^{k'}+1}^{\infty}\alpha_i^{n_h^{k'}}\\
&\quad+\sum_{k=1}^{K}\Theta\Big(A_j\sqrt{\frac{H^3}{n_{k,h}}\ln\frac{mKHSA_j^2}{\delta}}\ \Big)+4K\mathfrak{e}\\
&\overset{(ii)}{\leq}\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\Big(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V^{(j)}_{k',h+1})-\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}^{(j)}_{k',h+1})\Big)\\
&\quad+\sum_{s}\sum_{n=1}^{N_{K+1,h}(s)}\Theta\Big(A_j\sqrt{\frac{H^3}{n}\ln\frac{mKHSA_j^2}{\delta}}\ \Big)+4K\mathfrak{e}\\
&\overset{(iii)}{\leq}\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\Big(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V^{(j)}_{k',h+1})-\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}^{(j)}_{k',h+1})\Big)\\
&\quad+S\cdot\frac{1}{S}\sum_{s}\Theta\Big(A_j\sqrt{H^3N_{K+1,h}(s)\ln\frac{mKHSA_j^2}{\delta}}\ \Big)+4K\mathfrak{e}\\
&\overset{(iv)}{\leq}\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\Big(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V^{(j)}_{k',h+1})-\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}^{(j)}_{k',h+1})\Big)\\
&\quad-\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\Big(V^{(j)}_{k',h+1}(s_{k',h+1})-\underline{V}^{(j)}_{k',h+1}(s_{k',h+1})\Big)\\
&\quad+\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\Big(V^{(j)}_{k',h+1}(s_{k',h+1})-\underline{V}^{(j)}_{k',h+1}(s_{k',h+1})\Big)\\
&\quad+S\Theta\Big(A_j\sqrt{H^3\frac{1}{S}\sum_{s}N_{K+1,h}(s)\ln\frac{mKHSA_j^2}{\delta}}\ \Big)+4K\mathfrak{e}\\
&\overset{(v)}{=}\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\Big(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V^{(j)}_{k',h+1})-\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}^{(j)}_{k',h+1})\Big)\\
&\quad-\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\Big(V^{(j)}_{k',h+1}(s_{k',h+1})-\underline{V}^{(j)}_{k',h+1}(s_{k',h+1})\Big)\\
&\quad+\Big(1+\frac{1}{H}\Big)\sum_{k'=1}^{K}\delta^{(j)}_{k,h+1}+\Theta\Big(A_j\sqrt{H^3SK\ln\frac{mKHSA_j^2}{\delta}}\ \Big)+4K\mathfrak{e}
\end{aligned}
\tag{14}
$$

where (i) changes the order of summation by following Jin et al. (2018) and Jin et al. (2022a), (ii) uses eq. (25) and pigeonhole argument, (iii) uses $\sum_{n=1}^{N_{K+1,h}(s)}\sqrt{1/n}=\Theta\big(N_{K+1,h}(s)\big)$, (iv) applies Jensen's inequality to the convex function $\sqrt{\cdot}$, and (v) is by the definition of

$\delta_{k,h+1}$. Now we apply Lemma 25 to the first two terms above on the right-hand side,

$$\sum_{k'=1}^{K}\Big(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V^{(j)}_{k',h+1})-\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}^{(j)}_{k',h+1})\Big)$$
$$-\sum_{k'=1}^{K}\Big(V^{(j)}_{k',h+1}(s_{k',h+1})-\underline{V}^{(j)}_{k',h+1}(s_{k',h+1})\Big). \tag{15}$$

Then we obtain the recursion

$$\sum_{k=1}^{K}\delta^{(j)}_{k,h}\leq D\Big(1+\frac{1}{H}\Big)\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V^{(j)}_{k,h+1}(s)-\underline{V}^{(j)}_{k,h+1}(s)\Big)+\Big(1+\frac{1}{H}\Big)\sqrt{32KH^2\ln\frac{2mKHSA}{\delta}}$$
$$+\Big(1+\frac{1}{H}\Big)\sum_{k=1}^{K}\delta^{(j)}_{k,h+1}+\Theta\Big(A_j\sqrt{H^3SK\ln\frac{mKHSA_j^2}{\delta}}\Big)+4K\mathfrak{e}. \tag{16}$$

We apply Lemma 26 to solve this recursive relation by setting

$$\mathsf{a}_h=\sum_{k=1}^{K}\delta^{(j)}_{k,h},\mathsf{b}_h=\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V^{(j)}_{k,h}(s)-\underline{V}^{(j)}_{k,h}(s)\Big),\mathsf{C}_1=1+\frac{1}{H},\mathsf{C}_2=D\Big(1+\frac{1}{H}\Big),$$
$$\mathsf{C}_3=\Big(1+\frac{1}{H}\Big)\sqrt{32KH^2\ln\frac{2mKHSA}{\delta}}+\Theta\Big(A_j\sqrt{H^3SK\ln\frac{mKHSA_j^2}{\delta}}\Big)+4K\mathfrak{e}.$$

Then we obtain

$$\sum_{k=1}^{K}\delta^{(j)}_{k,h}\leq D\Big(1+\frac{1}{H}\Big)\sum_{i=0}^{H-h}\Big(1+\frac{1}{H}\Big)^i\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V^{(j)}_{k,h+1+i}(s)-\underline{V}^{(j)}_{k,h+1+i}(s)\Big) \tag{17}$$
$$+2H\Big[\Big(1+\frac{1}{H}\Big)\sqrt{32KH^2\ln\frac{2mKHSA}{\delta}}+\Theta\Big(A_j\sqrt{H^3SK\ln\frac{mKHSA_j^2}{\delta}}\Big)+4K\mathfrak{e}\Big],$$

where we also apply $(1+\frac{1}{H})^{H-h+1}<3$. The first term of the recursion eq. (17),

$$D\Big(1+\frac{1}{H}\Big)\sum_{i=0}^{H-h}\Big(1+\frac{1}{H}\Big)^i\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V^{(j)}_{k,h+1+i}(s)-\underline{V}^{(j)}_{k,h+1+i}(s)\Big)$$

measures the influence of diameter of uncertainty set on the convergence error of robust V-learning. In Theorem 12, we aim to build the upper bound with dependencies on the uncertainty diameter D without introducing other conditions. To do so, we can estimate this term with a universal upper bound; that is, for $h=1$,

$$D\Big(1+\frac{1}{H}\Big)\sum_{i=1}^{H}\Big(1+\frac{1}{H}\Big)^{i-1}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V^{(j)}_{k,1+i}(s)-\underline{V}^{(j)}_{k,1+i}(s)\Big)$$
$$\leq D\Big(1+\frac{1}{H}\Big)\sum_{i=0}^{H-1}\Big(1+\frac{1}{H}\Big)^{i}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(H-i\Big)$$

$$
\begin{aligned}
=&DSK\sum_{i=0}^{H-1}\Big(1+\frac{1}{H}\Big)^i(H-i)\\
\leq&5DSKH^2.
\end{aligned}
$$

Then eq. (17) can be bounded by

$$
\begin{aligned}
\sum_{k=1}^{K}\delta_{k,1}^{(j)} \leq& 2H\Big[\Big(1+\frac{1}{H}\Big)\sqrt{32KH^2\ln\frac{2mKHSA}{\delta}}+\Theta\Big(A_j\sqrt{H^3SK\ln\frac{mKHSA_j^2}{\delta}}\ \Big)+4K\mathfrak{e}\Big]\\
&+5DSKH^2.
\end{aligned}
\tag{18}
$$

The derived upper bound of optimality gap based on this inequality becomes

$$
\begin{aligned}
&\max_{j\in[m]}[\boldsymbol{V}_{\phi^*\circ\hat{\pi},1}^{(j)}(s_1)-\boldsymbol{V}_{\hat{\pi},1}^{(j)}(s_1)]\\
\leq&5DSH^2+2H\Big[2\sqrt{\frac{32H^2}{K}\ln\frac{2mKHSA}{\delta}}+\Theta\Big(A\sqrt{\frac{H^3S}{K}\ln\frac{mKHSA^2}{\delta}}\ \Big)+4\mathfrak{e}\Big],
\end{aligned}
$$

where s_1 is the initial state. Since the initial state can be any state over $\mathcal{S}$ due to the initialization, we obtain the desired bound. If $D\leq\frac{\epsilon}{SH^2}$ and the estimation error $\mathfrak{e}$ is sufficiently small, then this bound implies the number of episodes for achieving ϵ-approximation of robust correlated equilibrium is $K=\widetilde{\mathcal{O}}(SA^2H^5\epsilon^{-2})$.

■

Appendix G. Proof of Theorem 13

Definition 18 *Let $s_{k,h}$ be the state visited at h-th step k-th episode. The density of $s_{k,h}$ is universally bounded below by $p_{\min}$; that is*

$$
p_{\min}=\inf_{s\in\mathcal{S},k\in\mathbb{N},h\in[H]}\mathbb{P}(s_{k,h}=s).
$$

Theorem 13 *Let $S:=|\mathcal{S}|$ and $A:=\max_{1\leq j\leq m}|\mathcal{A}^{(j)}|$ correspond to the size of the state space and action space, respectively. Choose $\beta_t^{(j)}$, α_t and α_t^i according to eqs. (21)-(23). The output policy $\hat{\pi}$ produced by Algorithm 1 satisfies the following convergence rate with probability at least $1-c\delta$ for some constant $c>0$. For any D and $p_{\min}$ satisfying $\frac{\epsilon}{SH^2}\leq D<\frac{p_{\min}}{H}$, we have the bound*

$$
\max_{j\in[J]}\max_{s\in\mathcal{S}}\big(\boldsymbol{V}_{\phi^*\circ\hat{\pi},1}^{(j)}(s)-\boldsymbol{V}_{\hat{\pi},1}^{(j)}(s)\big)\leq\mathcal{O}\Big(\frac{H}{p_{\min}-DH}\Big(A\sqrt{\frac{H^3S}{K}\ln\frac{mKHSA^2}{\delta}}\ +\mathfrak{e}\Big)\Big).
$$

Further, if the state exploration $p_{\min}>\frac{\epsilon}{SH}$ and the approximation error $\mathfrak{e}=\mathcal{O}(\frac{\epsilon p_{\min}}{H})$, then the ϵ-accuracy is guaranteed with $K=\widetilde{\mathcal{O}}(SA^2H^5p_{\min}^{-2}\epsilon^{-2})$ episodes.

Proof We follow the same steps of Theorem 12 to obtain eq. (17):

$$\sum_{k=1}^{K}\delta_{k,h}^{(j)} \leq D\Big(1+\frac{1}{H}\Big)\sum_{i=0}^{H-h}\Big(1+\frac{1}{H}\Big)^{i}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V_{k,h+1+i}^{(j)}(s)-\underline{V}_{k,h+1+i}^{(j)}(s)\Big)$$
$$+2H\Big[\Big(1+\frac{1}{H}\Big)\sqrt{32KH^2\ln\frac{2mKHSA}{\delta}}+\Theta\Big(A_j\sqrt{H^3SK\ln\frac{mKHSA_j^2}{\delta}}\Big)+4K\mathfrak{e}\Big].$$

In Theorem 13, we need to consider the case where D cannot be sufficiently small but we still expect to obtain the ϵ-accuracy. To do so, we need to introduce the state exploration parameter $p_{\min}$ (Definition 18). For convenience, we define

$$\Delta_{k,h}^{(j)} := \sum_{s\in\mathcal{S}}\Big(V_{k,h}^{(j)}(s)-\underline{V}_{k,h}^{(j)}(s)\Big)$$

and

$$\mathscr{U}_j := 4H\Big[\Big(1+\frac{1}{H}\Big)\sqrt{32KH^2\ln\frac{2mKHSA}{\delta}}+\Theta\Big(A_j\sqrt{H^3SK\ln\frac{mKHSA_j^2}{\delta}}\Big)+4K\mathfrak{e}\Big].$$

Then we can have the following compact form of eq. (17):

$$\sum_{k=1}^{K}\delta_{k,h}^{(j)} \leq D\sum_{i=0}^{H-h}(1+\frac{1}{H})^{i+1}\sum_{k=1}^{K}\Delta_{k,h+1+i}^{(j)}+\mathscr{U}_j.$$

We take expectation on both sides and obtain

$$\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\mathbb{P}(s_{k,h}=s)\Big(V_{k,h}^{(j)}(s)-\underline{V}_{k,h}^{(j)}(s)\Big) \leq D\sum_{i=0}^{H-h}(1+\frac{1}{H})^{i+1}\sum_{k=1}^{K}\Delta_{k,h+1+i}^{(j)}+\mathscr{U}_j.$$

The left-hand side can be further lower bounded by the definition of $p_{\min}$. We have

$$p_{\min}\sum_{k=1}^{K}\Delta_{k,h}^{(j)} \leq D\sum_{i=0}^{H-h}(1+\frac{1}{H})^{i+1}\sum_{k=1}^{K}\Delta_{k,h+1+i}^{(j)}+\mathscr{U}_j. \tag{19}$$

Now we assume $p_{\min}>0$. Then this recursion can be solved by applying Lemma 27 with setting

$$\mathsf{a}_h=\sum_{k=1}^{K}\Delta_{k,h}^{(j)},\ \mathsf{C}_1=\frac{\mathscr{U}_j}{p_{\min}},\ \text{and}\ \mathsf{C}_2=\frac{D}{p_{\min}}.$$

Due to the requirement given in Lemma 27, we require $\mathsf{C}_2 := \frac{D}{p_{\min}} < \frac{1}{H}$. Under this condition, we obtain the following upper bound:

$$\max_{h\in[H]}\sum_{k=1}^{K}\Delta_{k,h}^{(j)} \leq \frac{\mathscr{U}_j}{p_{\min}-HD}. \tag{20}$$

Lastly, we bound the optimality gap at the step h. In expectation with respect choosing k, we have

$$\begin{aligned}
&\max_{j\in[m]}\max_{s\in\mathcal{S}}[\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi},1}(s)-\boldsymbol{V}^{(j)}_{\hat{\pi},1}(s)]\\
\le&\max_{j\in[m]}\sum_{s\in\mathcal{S}}[\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi},1}(s)-\boldsymbol{V}^{(j)}_{\hat{\pi},1}(s)]\\
\overset{(i)}{\le}&\max_{j\in[m]}\frac{1}{K}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}[\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,1},1}(s)-\boldsymbol{V}^{(j)}_{\hat{\pi}_{k,1},1}(s)]\\
\overset{(ii)}{\le}&\max_{j\in[m]}\frac{1}{K}\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\Big(V^{(j)}_{k,1}(s)-\underline{V}^{(j)}_{k,1}(s)\Big)\\
\overset{(iii)}{=}&\max_{j\in[m]}\frac{1}{K}\sum_{k=1}^{K}\Delta^{(j)}_{k,1}\\
\overset{(iv)}{\le}&\frac{3H}{p_{\min}-HD}\Big[2\sqrt{\frac{32H^2}{K}\ln\frac{2mKHSA}{\delta}}+\Theta\Big(A\sqrt{\frac{H^3S}{K}\ln\frac{mKHSA^2}{\delta}}\ \Big)+4\mathfrak{e}\Big]
\end{aligned}$$

where (i) uses the definitions of $\hat{\pi}$ and $\hat{\pi}_{k,h}$ given by Algorithms 2, (ii) uses Lemmas 22 & 24 and sampling rule of k given in Algorithm 2, (iii) is by the definition of $\Delta^{(j)}_{k,h}$, and (iv) uses eq. (20) and $A:=\max_{1\le j\le m}A_j$. It completes the proof.

When evaluating the sample complexity, we require the estimation error to satisfy $\mathfrak{e}\le\frac{p_{\min}-HD}{24H}\epsilon$. Then the last term is bounded by $\epsilon/2$. Then we let

$$\frac{1}{p_{\min}-HD}\sqrt{H^5SA^2\ln\frac{mKHSA^2}{\delta}/K}\le\frac{1}{2}\epsilon.$$

It solves the number of episodes for achieving ϵ-approximation of robust correlated equilibrium is $K=\widetilde{\mathcal{O}}(SA^2H^5p_{\min}^{-2}\epsilon^{-2})$.

■

Appendix H. Supporting Lemmas

Hyperparameter choices Throughout this subsection, we use the following hyperparameter choices

$$\beta^{(j)}_t:=cA_j\sqrt{\frac{H^3}{t}\ln\frac{mKHSA_j^2}{\delta}}+\mathfrak{e},\tag{21}$$

$$\alpha_t:=\frac{H+1}{H+t},\tag{22}$$

$$\alpha^0_t=0,\quad \alpha^i_t=\alpha_i\prod_{j=i+1}^{t}(1-\alpha_j)(i\ge 1)\tag{23}$$

where $c > 0$ is an absolute constant. It can be seen that the above hyperparameters satisfy the following conditions. (Eq. (24) is obvious and eqs. (25) & (26) are proved in Lemma 10 of Jin et al. (2022a).)

$$\sum_{i=1}^{t} \alpha_t^i = 1 \tag{24}$$

$$\sum_{t=i}^{\infty} \alpha_t^i = 1 + \frac{1}{H} \tag{25}$$

$$\sum_{i=1}^{t} \alpha_t^i \beta_i^{(j)} = \Theta\left(A_j \sqrt{\frac{H^3}{t} \ln \frac{mKHSA_j^2}{\delta}}\right) + \mathfrak{e} \tag{26}$$

Pessimistic estimation of V function: To facilitate the proof, we also provide a pessimistic estimator of V functions denoted as $\underline{V}_{k,h}^{(j)}$, which is constructed by the following update rules with initial values $\underline{V}_{1,h}^{(j)}(s) = \underline{V}_{k,H+1}^{(j)}(s) = 0$ for all s, h, k, j

$$\underline{V}_{k+1,h}^{(j)}(s_h) \leftarrow (1-\alpha_t)\underline{V}_{k,h}^{(j)}(s_h) + \alpha_t \left(r_h^{(j)} + \widehat{\sigma}_{\mathcal{P}_h(s_h,a_h)}(\underline{V}_{k,h+1}^{(j)}) - \beta_t^{(j)}\right); \tag{27}$$

$$\underline{V}_{k+1,h}^{(j)}(s_h) \leftarrow \max\{0, \underline{V}_{k+1,h}^{(j)}(s_h)\}. \tag{28}$$

The above update rules are similar to those for optimistic estimation in eqs. (7) & (8), with the major difference that $+\beta_t^{(j)} > 0$ in eq. (7) yields optimism (i.e. $V_{k,h}^{(j)}(s) \geq \boldsymbol{V}_{\phi^* \circ \hat{\pi}_{k,h},h}^{(j)}(s) \geq \boldsymbol{V}_{\hat{\pi}_{k,h},h}^{(j)}(s)$ as shown by Lemma 24) while $-\beta_t^{(j)} < 0$ in eq. (27) yields pessimism (i.e. $\underline{V}_{k,h}^{(j)}(s) \leq \boldsymbol{V}_{\hat{\pi}_{k,h},h}^{(j)}(s)$ as shown by Lemma 22)

Lemma 19 *The operator $\sigma_{\mathcal{P}_h(s,a)}$ defined in eq. (5) has the following properties:*

1. *Boundedness:* $\inf_{s'} V(s') \leq \sigma_{\mathcal{P}_h(s,a)}(V) \leq \sup_{s'} V(s')$.
2. *Monotonicity:* $\sigma_{\mathcal{P}_h(s,a)}(V') \leq \sigma_{\mathcal{P}_h(s,a)}(V)$ *for any V-tables* V, V' *such that* $V'(s) \leq V(s), \forall s$.
3. *Estimation bound: The estimator $\widehat{\sigma}_{\mathcal{P}_h(s,a)}$ has the following bounds for any V function V.*

$$\sigma_{\mathcal{P}_h(s,a)}(V) - \mathfrak{e} \leq \widehat{\sigma}_{\mathcal{P}_h(s,a)}(V) \leq \sigma_{\mathcal{P}_h(s,a)}(V) + \mathfrak{e}. \tag{29}$$

Proof Proof of boundedness: The upper bound $\sigma_{\mathcal{P}_h(s,a)}(V) \leq \sup_{s'} V(s')$ can be directly proved based on eq. (5) as follows.

$$\begin{aligned}
\sigma_{\mathcal{P}_h(s,a)}(V) &= \inf_{\widetilde{\mathbb{P}}_h(\cdot|s,a) \in \mathcal{P}_h(s,a)} \sum_{s' \in \mathcal{S}} \widetilde{\mathbb{P}}_h(s'|s,a) V(s') \\
&\leq \inf_{\widetilde{\mathbb{P}}_h(\cdot|s,a) \in \mathcal{P}_h(s,a)} \sum_{s' \in \mathcal{S}} \widetilde{\mathbb{P}}_h(s'|s,a) \sup_{s'' \in \mathcal{S}} V(s'')
\end{aligned}$$

$$\overset{(i)}{=} \sup_{s''\in\mathcal{S}} V(s''),$$

where (i) uses $\sum_{s'\in\mathcal{S}} \widetilde{\mathbb{P}}_h(s'|s,a) = 1$. The proof logic for the lower bound $\inf_{s'} V(s') \le \sigma_{\mathcal{P}_h(s,a)}(V)$ is similar.

Proof of monotonicity: Suppose $p \in \mathcal{P}_h(s,a)$ achieves the infimum in $\sigma_{\mathcal{P}_h(s,a)}(V)$ defined by eq. (5), i.e.

$$\sigma_{\mathcal{P}_h(s,a)}(V) = \sum_{s'\in\mathcal{S}} p(s')V(s'). \tag{30}$$

Then the monotonicity can be proved as follows.

$$\begin{aligned}
&\sigma_{\mathcal{P}_h(s,a)}(V) - \sigma_{\mathcal{P}_h(s,a)}(V') \\
=&\sum_{s'\in\mathcal{S}} p(s')V(s') - \inf_{\widetilde{\mathbb{P}}_h(\cdot|s,a)\in\mathcal{P}_h(s,a)} \sum_{s'\in\mathcal{S}} \widetilde{\mathbb{P}}_h(s'|s,a)V'(s') \\
\overset{(i)}{\ge}&\sum_{s'\in\mathcal{S}} p(s')V(s') - \sum_{s'\in\mathcal{S}} p(s')V'(s') \overset{(ii)}{\ge} 0,
\end{aligned} \tag{31}$$

where (i) will be used later and (ii) uses $p(s') \ge 0$ and $V(s') \ge V'(s')$ for all $s' \in \mathcal{S}$.

Proof of estimation bound: $\mathfrak{e} := \sup_{h,s,a,V} \left|\sigma_{\mathcal{P}_h(s,a)}(V) - \widehat{\sigma}_{\mathcal{P}_h(s,a)}(V)\right|$ defined by Definition 11 directly implies eq. (29). ■

Lemma 20 *For any player j and all $s \in \mathcal{S}$, the V-table $\widetilde{V}_{k,h}^{(j)}$ and $\underset{\sim}{V}_{k,h}^{(j)}$ tracked by Algorithm 1 at the h-th step in the k-th episode satisfying $\widetilde{V}_{k,h}^{(j)}(s) \ge 0$ and $\underset{\sim}{V}_{k,h}^{(j)}(s) \le H+1-h$.*

Proof We will only prove $\underset{\sim}{V}_{k,h}^{(j)}(s) \le H+1-h$ since the proof logic for $\widetilde{V}_{k,h}^{(j)}(s) \ge 0$ is similar. For $k = 1$, the initial value $\underset{\sim}{V}_{1,h}^{(j)}(s) := 0 \le H+1-h$. Then we assume $\underset{\sim}{V}_{k,h}^{(j)}(s) \le H+1-h$ for a certain fixed $k \ge 1$ and we prove $\underset{\sim}{V}_{k+1,h}^{(j)}(s) \le H+1-h$ as follows.

$$\begin{aligned}
\underset{\sim}{V}_{k+1,h}^{(j)}(s_h) &\overset{(i)}{=} (1-\alpha_t)\underset{\sim}{V}_{k,h}^{(j)}(s_h) + \alpha_t\big(r_h^{(j)}(s_h,a_h) + \widehat{\sigma}_{\mathcal{P}_h(s_h,a_h)}(\underline{V}_{k,h+1}^{(j)}) - \beta_t^{(j)}\big) \\
&\overset{(ii)}{\le} (1-\alpha_t)(H+1-h) + \alpha_t\big(1 + \sigma_{\mathcal{P}_h(s_h,a_h)}(\underline{V}_{k,h+1}^{(j)}) - \beta_t^{(j)} + \mathfrak{e}\big) \\
&\overset{(iii)}{\le} H+1-h,
\end{aligned}$$

where (i) uses the update rules (27) & (28), (ii) uses $\underset{\sim}{V}_{k,h}^{(j)}(s) \le H+1-h$ and eq. (29), and (iii) uses $\beta_t^{(j)} \ge \mathfrak{e}$ based on eq. (21) and the following inequality based on item 1 of Lemma 19. This concludes the proof.

$$\sigma_{\mathcal{P}_h(s_h,a_h)}(\underline{V}_{k,h+1}^{(j)}) \le \max_s \underline{V}_{k,h+1}^{(j)}(s) \le \max_s \left[\max\left(0, \underline{V}_{k,h+1}^{(j)}(s)\right)\right] \le H-h.$$

■

The following lemma says that the tracked upper confidence bound $V_{k,h}^{(j)}(s)$ is always larger than the lower confidence bound $\underline{V}_{k,h}^{(j)}(s)$.

Lemma 21 *For any player j and all $s \in \mathcal{S}$, the V-tables $V_{k,h}^{(j)}(s)$ and $\underline{V}_{k,h}^{(j)}(s)$ tracked by Algorithm 1 at the h-th step in the k-th episode satisfy the following inequality*

$$V_{k,h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s). \tag{32}$$

Proof It suffices to show $\widetilde{V}_{k,h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s)$ since it implies eq. (32) as follows

$$\begin{aligned}
&V_{k,h}^{(j)}(s) - \underline{V}_{k,h}^{(j)}(s) \\
&\overset{(i)}{=} \min\{H+1-h, \widetilde{V}_{k,h}^{(j)}(s_h)\} - \max\{0, \underline{V}_{k,h}^{(j)}(s_h)\} \\
&\overset{(ii)}{=} \min\{H+1-h, \max[0, \widetilde{V}_{k,h}^{(j)}(s_h)]\} - \min\{H+1-h, \max[0, \underline{V}_{k,h}^{(j)}(s_h)]\} \\
&\overset{(iii)}{\geq} 0,
\end{aligned} \tag{33}$$

where (i) uses eqs. (8) and (28), (ii) uses Lemma 20, and (iii) uses $\widetilde{V}_{k,h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s)$.

Then we prove $\widetilde{V}_{k,h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s)$ via induction with regards to k. For $k=1$, $\widetilde{V}_{1,h}^{(j)}(s) = H+1-h \geq \underline{V}_{1,h}^{(j)}(s) = 0$ due to initialization. Suppose $\widetilde{V}_{k,h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s)$ and thus eq. (32) holds for a certain fixed k. Then we aim to prove $\widetilde{V}_{k+1,h}^{(j)}(s) \geq \underline{V}_{k+1,h}^{(j)}(s)$ (i.e., eq. (32) also holds for $k+1$). It suffices to consider the case where s is the state visited at the h-th step in the k-th episode, that is, $s = s_{k,h}$. Otherwise, $\widetilde{V}_{k+1,h}^{(j)}(s) = \widetilde{V}_{k,h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s) = \underline{V}_{k+1,h}^{(j)}(s)$. When $s = s_{k,h}$, the update rules (7) & (27) imply that

$$\begin{aligned}
&\widetilde{V}_{k+1,h}^{(j)}(s) - \underline{V}_{k+1,h}^{(j)}(s) \\
=&(1-\alpha_t)\left(\widetilde{V}_{k,h}^{(j)}(s) - \underline{V}_{k,h}^{(j)}(s)\right) + \alpha_t\left(\widehat{\sigma}_{\mathcal{P}_h(s,a)}(V_{k,h+1}^{(j)}) - \widehat{\sigma}_{\mathcal{P}_h(s,a)}(\underline{V}_{k,h+1}^{(j)})\right) + 2\alpha_t\beta_t^{(j)} \\
\overset{(i)}{\geq}&\alpha_t\left(\sigma_{\mathcal{P}_h(s,a)}(V_{k,h+1}^{(j)}) - \sigma_{\mathcal{P}_h(s,a)}(\underline{V}_{k,h+1}^{(j)}) - 2\mathfrak{e} + 2\beta_t^{(j)}\right) \overset{(ii)}{\geq} 0
\end{aligned} \tag{34}$$

where (i) uses $\widetilde{V}_{k,h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s)$ and eq. (29), and (ii) uses $V_{k,h+1}^{(j)} \geq \underline{V}_{k,h+1}^{(j)}$, the monotonicity of $\sigma_{\mathcal{P}_h(s,a)}$ (see item 2 of Lemma 19) and $\beta_t^{(j)} \geq \mathfrak{e}$ (see eq. (21)). This concludes the proof.

■

Lemma 22 *The V-tables $\boldsymbol{V}_{\hat{\pi}_{k,h},h}^{(j)}$ and $\underline{V}_{k,h}^{(j)}$ satisfy the following inequality with probability at least $1-\delta$ for all players $j \in [m]$, episodes $k \in [K]$, time steps $h \in [H]$ and states $s \in \mathcal{S}$ and any $\delta \in (0, \frac{1}{2})$,*

$$\boldsymbol{V}_{\hat{\pi}_{k,h},h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s).$$

Proof It suffices to prove $\boldsymbol{V}_{\hat{\pi}_{k,h},h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s)$; it implies $\boldsymbol{V}_{\hat{\pi}_{k,h},h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s)$ because

$$\boldsymbol{V}_{\hat{\pi}_{k,h},h}^{(j)}(s) \geq 0$$

by its definition (2) (note that $r_\ell^{(j)}(s_\ell, a_\ell) \geq 0$) and

$$\underline{V}_{k,h}^{(j)}(s) := \max\{0, \underline{V}_{k,h}^{(j)}(s_h)\}.$$

Now we start to prove $\boldsymbol{V}_{\hat{\pi}_{k,h},h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s)$ by induction with respect to h backward. When $h = H+1$, the proof is trivial as $\boldsymbol{V}_{\pi,H+1}^{(j)}(s) = \underline{V}_{0,H+1}^{(j)}(s) = 0$ for any policy π. Then suppose $\boldsymbol{V}_{\hat{\pi}_{k,h+1},h+1}^{(j)}(s) \geq \underline{V}_{k,h+1}^{(j)}(s)$ holds so $\boldsymbol{V}_{\hat{\pi}_{k,h+1},h+1}^{(j)}(s) \geq \underline{V}_{k,h+1}^{(j)}(s)$ for a certain fixed h, and we will prove that $\boldsymbol{V}_{\hat{\pi}_{k,h},h}^{(j)}(s) \geq \underline{V}_{k,h}^{(j)}(s)$. Let $\{k_i\}_{1\leq i\leq t}$ ($k_1 < k_2 < \ldots < k_t < k$) be the set of episodes where the state s is visited at the h-th step. Then we unroll the update rule (27) of $\underline{V}_{k,h}^{(j)}(s)$ as follows with respect to the episode k.

$$\begin{aligned}
\underline{V}_{k,h}^{(j)}(s) &= \sum_{i=1}^{t} \alpha_t^i \left[r_h^{(j)}(s, a_{k^i,h}) + \widehat{\sigma}_{\mathcal{P}_h(s,a_{k^i,h})}(\underline{V}_{k^i,h+1}^{(j)}) - \beta_i^{(j)} \right] \\
&\overset{(i)}{\leq} \sum_{i=1}^{t} \alpha_t^i \left[r_h^{(j)}(s, a_{k^i,h}) + \sigma_{\mathcal{P}_h(s,a_{k^i,h})}(\underline{V}_{k^i,h+1}^{(j)}) + \mathfrak{e} - \beta_i^{(j)} \right], \qquad (35)
\end{aligned}$$

where (i) uses eq. (29). Let

$$X_i := \alpha_t^i \left[r_h^{(j)}(s, a_{k^i,h}) + \sigma_{\mathcal{P}_h(s_h,a_{k^i,h})}(\underline{V}_{k^i,h+1}^{(j)}) \right].$$

Then X_i always has the following bound since $\underline{V}_{k^i,h+1}^{(j)} = \max\{0, \underline{V}_{k^i,h+1}^{(j)}(s_h)\} \leq H - h$ based on Lemma 20.

$$0 \leq X_i \leq \alpha_t^i (H + 1 - h).$$

Then by using Azuma's inequality and applying union bound to all $j \in [m], i \in [t] \subset [K], h \in [H], s \in \mathcal{S}$, we have the following bound with probability at least $1 - \delta$.

$$\sum_{i=1}^{t} X_i - \mathbb{E}\Big[\sum_{i=1}^{t} X_i\Big] \leq \sqrt{\frac{1}{2}\Big(\ln\frac{2mKHS}{\delta}\Big)\sum_{i=1}^{t}(\alpha_t^i)^2(H+1-h)^2} \overset{(i)}{\leq} \sqrt{\frac{H^3}{t}\ln\frac{2mKHS}{\delta}}, \quad (36)$$

where (i) uses $\sum_{i=1}^{t}(\alpha_t^i)^2 \leq \frac{2H}{t}$ in Lemma 10 of Jin et al. (2022a). Therefore, with probability at least $1 - \delta$, the following inequality holds for all j, k, h, s.

$$\begin{aligned}
\underline{V}_{k,h}^{(j)}(s) &\overset{(i)}{\leq} \sum_{i=1}^{t} X_i \\
&\overset{(ii)}{\leq} \sum_{i=1}^{t} \alpha_t^i \mathbb{E}_{\pi_{k^i,h}} \left[r_h^{(j)}(s, a_{k^i,h}) + \sigma_{\mathcal{P}_h(s_h,a_{k^i,h})}(\underline{V}_{k^i,h+1}^{(j)}) \right] \\
&\quad + \sqrt{\frac{H^3}{t}\ln\frac{2mKHS}{\delta}} - \sum_{i=1}^{t} \alpha_t^i (\beta_i^{(j)} - \mathfrak{e}) \\
&\overset{(iii)}{\leq} \sum_{i=1}^{t} \alpha_t^i \mathbb{E}_{\pi_{k^i,h}} \left[r_h^{(j)}(s, a_{k^i,h}) + \sigma_{\mathcal{P}_h(s_h,a_{k^i,h})}(\boldsymbol{V}_{\hat{\pi}_{k^i,h+1},h+1}^{(j)}) \right]
\end{aligned}$$

$$\overset{(iv)}{=} \boldsymbol{V}^{(j)}_{\hat{\pi}_{k,h},h}(s),$$

where (i) uses eq. (35), (ii) uses eq. (36), (iii) uses eq. (26) and the assumption that $\boldsymbol{V}^{(j)}_{\hat{\pi}_{k^i},h+1}(s) \geq \underline{V}^{(j)}_{k^i,h+1}(s)$ holds for $\ell = k^i \leq k-1$, and (iv) uses the robust Bellman equation and the definition of $\hat{\pi}_{k,h}$ given by Algorithm 2. ■

H.1 Lemmas on Robust Correlated equilibrium

The following lemma follows Lemma 15 of (Jin et al., 2022a), with the bandit input changed from $(a_h, \frac{H-r_h-V_{h+1}(s_{h+1})}{H})$ to $(a_h, \frac{H-r_h-\widehat{\sigma}_{\mathcal{P}_h(s_h,a_h)}(V_{h+1})}{H})$.

Lemma 23 *Let $\pi_{k+1,h}$ be the policy given by the ADV_BANDIT_UPDATE algorithm at the h-th step of the k-th episode. Then the following bound holds for all $j \in [m]$, $k \in [K]$, $h \in [H]$ and $s \in \mathcal{S}$ with probability at least $1-\delta$ under Lemma 17.*

$$\begin{aligned}
&\max_{\phi^{(j)}} \sum_{i=1}^{t} \alpha_t^i \Big[\mathbb{E}_{a\sim\phi^{(j)}\circ(\pi_{k^i,h})}[r_h^{(j)}(s,a) + \sigma_{\mathcal{P}_h(s,a)}(V^{(j)}_{k^i,h+1})]\Big] \\
\leq &\sum_{i=1}^{t} \alpha_t^i \Big[\mathbb{E}_{a\sim\pi_{k^i,h}}[r_h^{(j)}(s,a) + \sigma_{\mathcal{P}_h(s,a)}(V^{(j)}_{k^i,h+1})]\Big] + 10A_j\sqrt{\frac{H^3}{t}\ln\frac{mKHSA_j^2}{\delta}}, \qquad (37)
\end{aligned}$$

Proof By applying Lemma 17 to the loss function $l_i(a) = \frac{H-r_h(s,a)-\sigma_{\mathcal{P}_h(s,a)}(V_{h+1})}{H}$ for any s, we obtain that the following bound holds for all $k \in [K]$, $h \in [H]$ and $s \in \mathcal{S}$ with probability $1-\delta$ (we replace δ in the bound in Lemma 17 with $\frac{\delta}{KHS}$ by applying union bound to all $k \in [K]$, $h \in [H]$ and $s \in \mathcal{S}$), which is equivalent to the above bound and thus concludes the proof.

$$\begin{aligned}
&\max_{\phi^{(j)}} \sum_{i=1}^{t} \alpha_t^i \mathbb{E}_{a\sim\pi_{k^i,h}}\Big[\frac{H - r_h^{(j)}(s,a) - \sigma_{\mathcal{P}_h(s,a)}(V^{(j)}_{k^i,h+1})}{H}\Big] \\
\leq &\max_{\phi^{(j)}} \sum_{i=1}^{t} \alpha_t^i \mathbb{E}_{a\sim\phi^{(j)}\circ\pi_{k^i,h}}\Big[\frac{H - r_h^{(j)}(s,a) - \sigma_{\mathcal{P}_h(s,a)}(V^{(j)}_{k^i,h+1})}{H}\Big] \\
&+ 10A_j\sqrt{\frac{H}{t}\ln\frac{mKHSA_j^2}{\delta}}.
\end{aligned}$$

■

Lemma 24 *For the j-th player, the V-tables $\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,h},h}(s) := \max_{\phi^{(j)}} \boldsymbol{V}^{(j)}_{\phi^{(j)}\circ\hat{\pi}_{k,h},h}(s)$ and $V^{(j)}_{k,h}(s)$ at h-th step in the k-th episode, satisfy the following inequality with probability at least $1-2\delta$ for any $\delta \in (0, \frac{1}{2})$*

$$V^{(j)}_{k,h}(s) \geq \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,h},h}(s)$$

for all $s \in \mathcal{S}$.

Proof It suffices to prove $\widetilde{V}^{(j)}_{k,h}(s) \geq \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,h},h}(s)$; it implies $V^{(j)}_{k,h}(s) \geq \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,h},h}(s)$, since

$$V^{(j)}_{k,h}(s_h) = \min\{H+1-h, \widetilde{V}^{(j)}_{k,h}(s_h)\}$$

due to eq. (8), and

$$\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,h},h}(s) := \max_{\phi^{(j)}} \boldsymbol{V}^{(j)}_{\phi^{(j)}\circ\hat{\pi}_{k,h},h}(s) \leq H+1-h$$

since $r^{(j)}_h \leq 1$ for all j,h. Let $\{k_i\}_{1\leq i\leq t}$ ($k_1 < k_2 < \ldots < k_t < k$) be the set of episodes where the state s is visited at the h-th step. Then we unroll the update rule (7) of $\widetilde{V}^{(j)}_{k,h}(s)$ with respect to the episode k as follows.

$$\begin{aligned}\widetilde{V}^{(j)}_{k,h}(s) &= \alpha_t^0(H-h+1) + \sum_{i=1}^t \alpha_t^i \left[r^{(j)}_h(s,a_{k^i,h}) + \widehat{\sigma}_{\mathcal{P}_h(s,a_{k^i,h})}(V^{(j)}_{k^i,h+1}) + \beta^{(j)}_i\right] \\ &\geq \sum_{i=1}^t \alpha_t^i \left[r^{(j)}_h(s,a_{k^i,h}) + \sigma_{\mathcal{P}_h(s,a_{k^i,h})}(V^{(j)}_{k^i,h+1}) + \beta^{(j)}_i - \mathfrak{e}\right].\end{aligned}$$

where the above $\geq$ uses $\alpha_t^0 = 0$ (see eq. (23)) and eq. (29). Substituting (36) which holds with probability at least $1-\delta$ into the above inequality, we obtain that the following bound holds for all j,k,h,s with probability at least $1-\delta$,

$$\begin{aligned}\widetilde{V}^{(j)}_{k,h}(s) \geq &\sum_{i=1}^t \alpha_t^i \mathbb{E}_{\pi_h^{k^i}}\left[r^{(j)}_h(s,a_{k^i,h}) + \sigma_{\mathcal{P}_h(s,a_{k^i,h})}(V^{(j)}_{k^i,h+1})\right] \\ &+ \sum_{i=1}^t \alpha_t^i(\beta^{(j)}_i - \mathfrak{e}) - \sqrt{\frac{H^3}{t}\ln\frac{2mKHS}{\delta}}. \end{aligned} \tag{38}$$

By substituting eq. (37) into eq. (38), we obtain the following bound which holds for all j,k,h,s with probability at least $1-2\delta$ (since eq. (37) holds with probability at least $1-\delta$ and so does eq. (38)).

$$\begin{aligned}\widetilde{V}^{(j)}_{k,h}(s) \geq &\max_{\phi^{(j)}} \sum_{i=1}^t \alpha_t^i \mathbb{E}_{\phi^{(j)}\circ\pi_{k^i,h}}\left[r^{(j)}_h(s,a_{k^i,h}) + \sigma_{\mathcal{P}_h(s,a_{k^i,h})}(V^{(j)}_{k^i,h+1})\right] \\ &+ \sum_{i=1}^t \alpha_t^i(\beta^{(j)}_i - \mathfrak{e}) - \sqrt{\frac{H^3}{t}\ln\frac{2mKHS}{\delta}} - 10A_j\sqrt{\frac{H^3}{t}\ln\frac{mKHSA_j^2}{\delta}} \\ \overset{(i)}{\geq} &\max_{\phi_h^{(j)}} \sum_{i=1}^t \alpha_t^i \mathbb{E}_{a\sim\phi_h^{(j)}\circ\pi_{k^i,h}}\left[r^{(j)}_h(s,a) + \sigma_{\mathcal{P}_h(s,a)}(V^{(j)}_{k^i,h+1})\right] \end{aligned} \tag{39}$$

where (i) holds using eq. (26).

Then we can apply induction to h backward to prove $\widetilde{V}^{(j)}_{k,h}(s) \geq \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k^i},h}$. For the base case $h = H+1$, it can be easily seen that $\widetilde{V}^{(j)}_{k,H+1}(s) = \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_k,H+1}(s) = 0$ based on Algorithm 1 and the definition of $\boldsymbol{V}^{(j)}_{\pi,h}$ given by eq. (2). Suppose $\widetilde{V}^{(j)}_{k,h+1}(s) \geq \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_k,h+1}(s)$

for a certain fixed h and all j, k, s, so $V^{(j)}_{k,h+1}(s) \geq \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_k,h+1}(s)$. Then eq. (39) further implies the following inequality, which concludes the proof. (Note that the whole induction builds on eq. (39) which holds for all k, h, j, s with probability at least $1-2\delta$.)

$$\widetilde{V}^{(j)}_{k,h}(s) \geq \max_{\phi^{(j)}_h} \sum_{i=1}^{t} \alpha_t^i \mathbb{E}_{a\sim\phi^{(j)}_h\circ\pi_{k^i,h}}\Big[r^{(j)}_h(s,a) + \sigma_{\mathcal{P}_h(s,a)}(\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k^i,h+1},h+1})\Big] \geq \boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k^i,h},h},$$

where the second $\leq$ uses the following inequality obtained via the same proof logic as Lemma 13 of Jin et al. (2022a) (see the beginning of page 19 of Jin et al. (2022a)).

$$\begin{aligned}
&\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k,h},h}(s) := \max_{\phi^{(j)}} \boldsymbol{V}^{(j)}_{\phi^{(j)}\circ\hat{\pi}_{k,h},h}(s)\\
&\overset{(i)}{=} \max_{\phi^{(j)}_h} \max_{\phi^{(j)}_{(h+1):H}} \mathbb{E}_{a\sim\phi^{(j)}_h\circ[\hat{\pi}_{k,h}]_h}\Big(r^{(j)}_h(s,a) + \sigma_{\mathcal{P}_h(s,a)}\big(\boldsymbol{V}^{(j)}_{\phi^{(j)}_{(h+1):H}\circ\hat{\pi}_{k,h+1},h+1}\big)\Big)\\
&\overset{(ii)}{=} \max_{\phi^{(j)}_h} \max_{\phi^{(j)}_{(h+1):H}} \sum_{i=1}^{t} \alpha_t^i \mathbb{E}_{a\sim\phi^{(j)}_h\circ\pi_{k^i,h}}\Big(r^{(j)}_h(s,a) + \sigma_{\mathcal{P}_h(s,a)}\big(\boldsymbol{V}^{(j)}_{\phi^{(j)}_{(h+1):H}\circ\hat{\pi}_{k^i,h+1},h+1}\big)\Big)\\
&\overset{(iii)}{\leq} \max_{\phi^{(j)}_h} \sum_{i=1}^{t} \alpha_t^i \mathbb{E}_{a\sim\phi^{(j)}_h\circ\pi_{k^i,h}}\Big(r^{(j)}_h(s,a)\\
&\qquad + \max_{\phi^{(j)}_{(h+1):H}} \inf_{\widetilde{\mathbb{P}}_h(\cdot|s,a)\in\mathcal{P}_h(s,a)} \sum_{s'} \widetilde{\mathbb{P}}_h(s'|s,a)\boldsymbol{V}^{(j)}_{\phi^{(j)}_{(h+1):H}\circ\hat{\pi}_{k^i,h+1},h+1}(s')\Big)\\
&\leq \max_{\phi^{(j)}_h} \sum_{i=1}^{t} \alpha_t^i \mathbb{E}_{a\sim\phi^{(j)}_h\circ\pi_{k^i,h}}\Big(r^{(j)}_h(s,a)\\
&\qquad + \inf_{\widetilde{\mathbb{P}}_h(\cdot|s,a)\in\mathcal{P}_h(s,a)} \sum_{s'} \widetilde{\mathbb{P}}_h(s'|s,a) \max_{\phi^{(j)}_{(h+1):H}} \boldsymbol{V}^{(j)}_{\phi^{(j)}_{(h+1):H}\circ\hat{\pi}_{k^i,h+1},h+1}(s')\Big)\\
&\overset{(iv)}{=} \max_{\phi^{(j)}_h} \sum_{i=1}^{t} \alpha_t^i \mathbb{E}_{a\sim\phi^{(j)}_h\circ\pi_{k^i,h}}\Big(r^{(j)}_h(s,a) + \sigma_{\mathcal{P}_h(s,a)}\big(\boldsymbol{V}^{(j)}_{\phi^*\circ\hat{\pi}_{k^i,h+1},h+1}\big)\Big),
\end{aligned}$$

where (i) uses robust Bellman equation and denotes $[\hat{\pi}_{k,h}]_h$ as the marginal distribution of a_h based on policy $\hat{\pi}_{k,h}$ defined by Algorithm 2, (ii) uses the definition of $\hat{\pi}_{k,h}$ given by Algorithm 2, (iii) and (iv) use the definition of $\sigma_{\mathcal{P}_h(s,a)}$ given by eq. (5). ■

H.2 Key Lemmas to Handle Uncertainty

Lemma 25 *With the probability at least* $1-\delta$,

$$\begin{aligned}
&\sum_{k'=1}^{K}\Big(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V^{(j)}_{k',h+1}) - \sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}^{(j)}_{k',h+1})\Big)\\
&- \sum_{k'=1}^{K}\Big(V^{(j)}_{k',h+1}(s_{k',h+1}) - \underline{V}^{(j)}_{k',h+1}(s_{k',h+1})\Big)
\end{aligned}$$

$$\leq D\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\left(V_{k,h+1}^{(j)}(s)-\underline{V}_{k,h+1}^{(j)}(s)\right)+\sqrt{32KH^2\ln\frac{2mKHSA}{\delta}}. \tag{40}$$

Proof For the k'-th episode, let

$$\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}_{k',h+1}^{(j)})=\sum_s p_{k',h}(s)\underline{V}_{k',h+1}^{(j)}(s)=p_{k',h}^T\underline{V}_{k',h+1}^{(j)},$$

i.e., the minimum is achieved at $p_{k',h}^T$. Then

$$\begin{aligned}
&\sum_{k'=1}^{K}\left(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V_{k',h+1}^{(j)})-\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}_{k',h+1}^{(j)})\right)\\
&-\sum_{k'=1}^{K}\left(V_{k',h+1}^{(j)}(s_{k',h+1})-\underline{V}_{k',h+1}^{(j)}(s_{k',h+1})\right)\\
\leq&\sum_{k'=1}^{K}p_{k',h}^T\left(V_{k',h+1}^{(j)}-\underline{V}_{k',h+1}^{(j)}\right)-\sum_{k'=1}^{K}\left(V_{k',h+1}^{(j)}(s_{k',h+1})-\underline{V}_{k',h+1}^{(j)}(s_{k',h+1})\right)\\
\leq&\underbrace{\sum_{k'=1}^{K}p_{k',h}^T\left(V_{k',h+1}^{(j)}-\underline{V}_{k',h+1}^{(j)}\right)-\sum_{k'=1}^{K}\mathbb{E}\left(V_{k',h+1}^{(j)}(s_{k',h+1})-\underline{V}_{k',h+1}^{(j)}(s_{k',h+1})\Big|s_{k',h},a_{k',h}\right)}_{(A)}\\
&\underbrace{+\sum_{k'=1}^{K}\mathbb{E}\Big(V_{k',h+1}^{(j)}(s_{k',h+1})-\underline{V}_{k',h+1}^{(j)}(s_{k',h+1})\Big|s_{k',h},a_{k',h}\Big)-\sum_{k'=1}^{K}\left(V_{k',h+1}^{(j)}(s_{k',h+1})-\underline{V}_{k',h+1}^{(j)}(s_{k',h+1})\right)}_{(B)}.
\end{aligned}$$

Then we will bound terms (A) and (B), respectively. For term (A), we define $p'_{k,h}(s):=\mathbb{P}_h(s_{k,h+1}=s|s_{k,h},a_{k,h})$ for some distribution sampled from the uncertainty set $\mathcal{P}_h(s_{k,h},a_{k,h})$. Then we obtain

$$\begin{aligned}
&\sum_{k'=1}^{K}p_{k',h}^T\left(V_{k',h+1}^{(j)}-\underline{V}_{k',h+1}^{(j)}\right)-\sum_{k'=1}^{K}\mathbb{E}\left(V_{k,h+1}^{(j)}(s_{k',h+1})-\underline{V}_{k',h+1}^{(j)}(s_{k',h+1})\Big|s_{k',h},a_{k',h}\right)\\
\overset{(i)}{=}&\sum_{k'=1}^{K}p_{k',h}^T\left(V_{k',h+1}^{(j)}-\underline{V}_{k',h+1}^{(j)}\right)-\sum_{k'=1}^{K}p_{k',h}'^T\left(V_{k',h+1}^{(j)}-\underline{V}_{k',h+1}^{(j)}\right)\\
\overset{(ii)}{=}&\sum_{k'=1}^{K}(p_{k',h}-p'_{k',h})^T\left(V_{k',h+1}^{(j)}-\underline{V}_{k',h+1}^{(j)}\right)\\
\overset{(iii)}{\leq}&\max_{s\in\mathcal{S}}|p_{k',h}(s)-p'_{k,h}(s)|\sum_{k'=1}^{K}\sum_{s\in\mathcal{S}}\left(V_{k',h+1}^{(j)}(s)-\underline{V}_{k',h+1}^{(j)}(s)\right)\\
\overset{(iv)}{\leq}&D\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\left(V_{k,h+1}^{(j)}(s)-\underline{V}_{k,h+1}^{(j)}(s)\right),
\end{aligned}$$

where (i) expands the conditional expectation, (ii) combines the same term together, (iii) applies the Hölder's inequality $\langle u, v\rangle \le \|u\|_\infty \|v\|_1$, and (iv) uses $\mathfrak{e} := \sup_{h,s,a,V} \left|\sigma_{\mathcal{P}_h(s,a)}(V) - \widehat{\sigma}_{\mathcal{P}_h(s,a)}(V)\right|$ defined by Definition 11.

Now we turn to bound term (B). Let

$$Y_{k'} = \mathbb{E}\left(V^{(j)}_{k',h+1}(s_{k',h+1}) - \underline{V}^{(j)}_{k',h+1}(s_{k',h+1}) \middle| s_{k',h}, a_{k',h}\right) - \left(V^{(j)}_{k',h+1}(s_{k',h+1}) - \underline{V}^{(j)}_{k',h+1}(s_{k',h+1})\right).$$

Then $\sum_{k'=1}^{k} Y_{k'}$ forms a martingale with $|Y_{k'}| \le 4H$. By Azuma-Hoeffding inequality, with a probability at least $1-\delta$, for any episode k, any $(s_{k',h+1}, a_{k',h}) \in \mathcal{S}\times\mathcal{A}$, any agent j, and any step $h \in [H]$,

$$\sum_{k'=1}^{K} Y_{k'} \le \sqrt{32KH^2 \ln \frac{2mKHSA}{\delta}}.$$

Combining the bounds of (A) and (B), we obtain the upper bound of eq. (15):

$$\begin{aligned}
&\sum_{k'=1}^{K}\left(\sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(V^{(j)}_{k',h+1}) - \sigma_{\mathcal{P}_h(s_{k,h},a_{k',h})}(\underline{V}^{(j)}_{k',h+1})\right)\\
\le& D\sum_{k=1}^{K}\sum_{s\in\mathcal{S}}\left(V^{(j)}_{k,h+1}(s) - \underline{V}^{(j)}_{k,h+1}(s)\right) + \sqrt{32KH^2\ln\frac{2mKHSA}{\delta}}.
\end{aligned}$$

■

This lemma gives a more general version of recursion used in Jin et al. (2022a). When setting $b_h \equiv 0$ and iterating to $h = 1$, this result is reduced to Jin et al. (2022a).

Lemma 26 *Suppose the sequence $\{\mathsf{a}_h, \mathsf{b}_h\}_{h\in[H+1]}$ satisfies the following recursion:*

$$\begin{aligned}
&\mathsf{a}_{H+1} = \mathsf{b}_{H+1} = 0,\\
&\mathsf{a}_h \le \mathsf{C}_1\mathsf{a}_{h+1} + \mathsf{C}_2\mathsf{b}_{h+1} + \mathsf{C}_3.
\end{aligned}$$

Then for any $h \in [H]$,

$$\mathsf{a}_h \le \mathsf{C}_2\sum_{i=0}^{H-h}\mathsf{C}_1^i\mathsf{b}_{h+1+i} + \Big(\frac{\mathsf{C}_1^{H-h+1}-1}{\mathsf{C}_1-1}\Big)\mathsf{C}_3.$$

Proof We prove it by induction with respect to h backward. For $h = H$, the statement obviously holds. Then assuming the statement holds for $h+1$ (for some $h < H$), we consider the upper bound of a_h:

$$\begin{aligned}
\mathsf{a}_h &\overset{(i)}{\le} \mathsf{C}_1\mathsf{a}_{h+1} + \mathsf{C}_2\mathsf{b}_{h+1} + \mathsf{C}_3\\
&\overset{(ii)}{\le} \mathsf{C}_1\Big[\mathsf{C}_2\sum_{i=0}^{H-h-1}\mathsf{C}_1^i\mathsf{b}_{h+2+i} + \Big(\frac{\mathsf{C}_1^{H-h}-1}{\mathsf{C}_1-1}\Big)\mathsf{C}_3\Big] + \mathsf{C}_2\mathsf{b}_{h+1} + \mathsf{C}_3\\
&\overset{(iii)}{=} \mathsf{C}_2\Big[\sum_{i=0}^{H-h-1}\mathsf{C}_1^{i+1}\mathsf{b}_{h+2+i} + \mathsf{b}_{h+1}\Big] + \Big[\mathsf{C}_1\Big(\frac{\mathsf{C}_1^{H-h}-1}{\mathsf{C}_1-1}\Big) + 1\Big]\mathsf{C}_3
\end{aligned}$$

$$= \mathsf{C}_2 \sum_{i=0}^{H-h} \mathsf{C}_1^i \mathsf{b}_{h+1+i} + \Big(\frac{\mathsf{C}_1^{H-h+1}-1}{\mathsf{C}_1-1}\Big)\mathsf{C}_3,$$

where (i) uses the recursion, (ii) applies the induction hypothesis, and (iii) rearranges the order of each term. It completes the proof.

■

Lemma 27 *Suppose the sequence* $\{\mathsf{a}_h, \mathsf{b}_h\}_{h\in[H+1]}$ *satisfies the following recursion:*

$$\mathsf{a}_{H+1} = 0,$$
$$\mathsf{a}_h \le \mathsf{C}_1 + \mathsf{C}_2 \sum_{i=0}^{H-h} \Big(1+\frac{1}{H}\Big)^{i+1} \mathsf{a}_{h+1+i}.$$

If $\mathsf{C}_2 < 1/H$*, then*

$$\max_h \mathsf{a}_h \le \frac{\mathsf{C}_1}{1-H\mathsf{C}_2}.$$

Proof We re-write the recursion in matrix form. Here inequality holds for entry-wise.

$$\begin{bmatrix} \mathsf{a}_1 \\ \mathsf{a}_2 \\ \vdots \\ \mathsf{a}_H \end{bmatrix} \le \mathsf{C}_1 \mathbf{1}_H + \mathsf{C}_2 \begin{bmatrix} 0 & \big(1+\frac{1}{H}\big) & \big(1+\frac{1}{H}\big)^2 & \dots & \big(1+\frac{1}{H}\big)^{H-1} \\ 0 & 0 & \big(1+\frac{1}{H}\big) & \dots & \big(1+\frac{1}{H}\big)^{H-2} \\ 0 & 0 & 0 & \dots & \big(1+\frac{1}{H}\big)^{H-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \mathsf{a}_1 \\ \mathsf{a}_2 \\ \vdots \\ \mathsf{a}_H \end{bmatrix}.$$

Denote the upper triangular Toeplitz matrix by T. Then we take $\|\cdot\|_\infty$ on both sides and obtain

$$\begin{aligned} \max_h \mathsf{a}_h &\le \mathsf{C}_1 + \mathsf{C}_2 \|\mathsf{T}\|_\infty \max_h \mathsf{a}_h \\ &\le \mathsf{C}_1 + \mathsf{C}_2 \sum_{h'=1}^{H-1} \Big(1+\frac{1}{H}\Big)^{h'} \max_h \mathsf{a}_h \\ &\le \mathsf{C}_1 + \mathsf{C}_2 H\Big[\Big(1+\frac{1}{H}\Big)^H - 1\Big] \max_h \mathsf{a}_h \\ &\le \mathsf{C}_1 + H\mathsf{C}_2 \max_h \mathsf{a}_h. \end{aligned}$$

When $\mathsf{C}_2 < 1/H$, it solves the upper bound of $\max_h \mathsf{a}_h$ as

$$\max_h \mathsf{a}_h \le \frac{\mathsf{C}_1}{1-H\mathsf{C}_2}.$$

■

References

Jean Pierre Allamaa, Panagiotis Patrinos, Herman Van der Auweraer, and Tong Duy Son. Sim2real for autonomous vehicle control using executable digital twin. IFAC-PapersOnLine, 55(24):385–391, 2022.

Robert J Aumann. Subjectivity and correlation in randomized strategies. Journal of Mathematical Economics, 1(1):67–96, 1974.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In ICML, pages 551–560. PMLR, 2020.

Yu Bai, Chi Jin, Huan Wang, and Caiming Xiong. Sample-efficient learning of Stackelberg equilibria in general-sum games. NeurIPS, 34:25799–25811, 2021.

Avrim Blum and Yishay Mansour. From external to internal regret. Journal of Machine Learning Research, 8(6), 2007.

Manuele Brambilla, Eliseo Ferrante, Mauro Birattari, and Marco Dorigo. Swarm robotics: a review from the swarm engineering perspective. Swarm Intelligence, 7(1):1–41, 2013.

Shutong Chen, Guanjun Liu, Ziyuan Zhou, Kaiwen Zhang, and Jiacun Wang. Robust multi-agent reinforcement learning method based on adversarial domain randomization for real-world dual-uav cooperation. IEEE Transactions on Intelligent Vehicles, 2023.

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

Xiaotie Deng, Yuhao Li, David Henry Mguni, Jun Wang, and Yaodong Yang. On the complexity of computing Markov perfect equilibrium in general-sum stochastic games. arXiv preprint arXiv:2109.01795, 2021.

Jerzy Filar and Koos Vrieze. Competitive Markov Decision Processes. Springer Science & Business Media, 2012.

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.

Amy Greenwald, Keith Hall, Roberto Serrano, et al. Correlated Q-learning. In ICML, volume 3, pages 242–249, 2003.

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.

Junling Hu and Michael P Wellman. Nash Q-learning for general-sum stochastic games. Journal of Machine Learning Research, 4(Nov):1039–1069, 2003.

Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. Optimization Online, pages 1695–1724, 2013.

Feng Huang, Ming Cao, and Long Wang. Optimal control of robust team stochastic games. arXiv preprint arXiv:2105.07405, 2021.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? NeurIPS, 31, 2018.

Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. In ICLR 2022 Workshop on Gamification and Multiagent Solutions, 2022a.

Yujia Jin, Vidya Muthukumar, and Aaron Sidford. The complexity of infinite-horizon general-sum stochastic games. arXiv preprint arXiv:2204.04186, 2022b.

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.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in Markov potential games. arXiv preprint arXiv:2106.01969, 2021.

Jialian Li, Tongzheng Ren, Dong Yan, Hang Su, and Jun Zhu. Policy learning for robust Markov decision process with a mismatched generative model. arXiv preprint arXiv:2203.06587, 2022a.

Tianjiao Li, Ziwei Guan, Shaofeng Zou, Tengyu Xu, Yingbin Liang, and Guanghui Lan. Faster algorithm and sharper analysis for constrained Markov decision process. arXiv preprint arXiv:2110.10351, 2021.

Yan Li, Tuo Zhao, and Guanghui Lan. First-order policy optimization for robust Markov decision process. arXiv preprint arXiv:2209.10579, 2022b.

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

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

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In ICML, pages 7001–7010. PMLR, 2021.

Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum Markov games. Dynamic Games and Applications, pages 1–22, 2022.

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

Ariel Neufeld and Julian Sester. Robust Q-learning algorithm for Markov decision processes under wasserstein uncertainty. arXiv preprint arXiv:2210.00898, 2022.

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

Arnab Nilim and Laurent Ghaoui. Robustness in Markov decision problems with uncertain transition matrices. NeurIPS, 16, 2003.

Vianney Perchet. A note on robust nash equilibria with uncertainties. RAIRO-Operations Research-Recherche Opérationnelle, 48(3):365–371, 2014.

Vianney Perchet. Finding robust nash equilibria. In Algorithmic Learning Theory, pages 725–751. PMLR, 2020.

Aurko Roy, Huan Xu, and Sebastian Pokutta. Reinforcement learning under model mismatch. NeurIPS, 30, 2017.

Jay K Satia and Roy E Lave Jr. Markovian decision processes with uncertain transition probabilities. Operations Research, 21(3):728–740, 1973.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. arXiv preprint arXiv:1610.03295, 2016.

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.

Bernhard Von Stengel and Françoise Forges. Extensive-form correlated equilibrium: Definition and computational complexity. Mathematics of Operations Research, 33(4):1002–1022, 2008.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. NeurIPS, 34:7193–7206, 2021.

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

Kaiqing Zhang, Tao Sun, Yunzhe Tao, Sahika Genc, Sunil Mallya, and Tamer Basar. Robust multi-agent reinforcement learning with model uncertainty. NeurIPS, 33:10571–10583, 2020.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in multi-agent Markov stochastic games: Stationary points and convergence. arXiv preprint arXiv:2106.00198, 2021.