# Roping in Uncertainty: Robustness and Regularization in Markov Games

**Jeremy McMahan** [1] **Giovanni Artiglio** [1] **Qiaomin Xie** [1]

## Abstract

We study robust Markov games (RMG) with $s$-rectangular uncertainty. We show a general equivalence between computing a robust Nash equilibrium (RNE) of a $s$-rectangular RMG and computing a Nash equilibrium (NE) of an appropriately constructed regularized MG. The equivalence result yields a planning algorithm for solving $s$-rectangular RMGs, as well as provable robustness guarantees for policies computed using regularized methods. However, we show that even for just reward-uncertain two-player zero-sum matrix games, computing an RNE is PPAD-hard. Consequently, we derive a special uncertainty structure called efficient player-decomposability and show that RNE for two-player zero-sum RMG in this class can be provably solved in polynomial time. This class includes commonly used uncertainty sets such as $L_1$ and $L_\infty$ ball uncertainty sets.

## 1. Introduction

Offline reinforcement learning (RL) and RL in simulated environments are effective ways to deal with situations where traditional online RL would be too risky or costly. However, these approaches suffer from the sim-to-real gap in which slight differences in the models can lead to policies with poor performance in the true environment. To combat the sim-to-real gap, robust policies were studied using the framework of robust Markov decision processes (RMDP) (Satia & Lave, 1973) and later robust Markov Games (RMG) (Zhang et al., 2020b). Robust approaches have been effective in the real world, especially for navigating UAVs in mission-critical multi-agent environments (Chen et al., 2023) and in queuing systems (Kardeş et al., 2011). In practice, regularization has been a popular approach to improving the robustness and convergence of multi-agent RL algorithms with empirical success.

[1]University of Wisconsin-Madison, USA. Correspondence to: Jeremy McMahan <jmcmahan@cs.wisc.edu>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

A robust Markov game $(\mathcal{S}, \{\mathcal{A}_i\}_{i\in[N]}, P^\star, \boldsymbol{r}^\star, H, \mathcal{U})$ is defined by a standard Markov game (MG) $G^\star := (P^\star, \boldsymbol{r}^\star)$, called the nominal game, and an uncertainty set $\mathcal{U} := \mathcal{P} \times \mathcal{U}^r$ that describes all possible models that could be realized. Here, $\mathcal{P}$ is the set of possible transition kernels, and $\mathcal{U}^r$ is the set of possible reward functions. A common solution concept for RMGs is the Markov-perfect robust Nash Equilibrium (MPRNE). A policy $\boldsymbol{\pi}$ is an MPRNE if, for each stage game $(h, s)$, $\boldsymbol{\pi}$ is a mutual best response assuming the worst-case model for each player.

Unlike classical RMDPs, solving RMGs is already difficult when only the reward function has uncertainty. Specifically, a single-stage game with reward uncertainty can capture arbitrary general-sum games and so is PPAD-hard to solve. Surprisingly, unlike traditional game theory, even the two-player zero-sum version of reward-uncertain RMGs with $|\mathcal{S}| = H = 1$ and the minimal $(s, a)$-rectangularity assumption is PPAD-hard to solve. Similarly, two-player zero-sum RMGs with only transition uncertainty and $H = 2$ periods are also PPAD-hard to solve. Thus, solving even simple RMGs is already a computational challenge.

Although many advances have been made for RMDPs, RMGs are much less understood. The seminal paper (Zhang et al., 2020b) devised algorithms to learn a robust NE (RNE) for RMGs but only proved asymptotic convergence of their methods. On the other hand, Blanchet et al. (2023) proposed provably sample-efficient algorithms to learn an RNE for the special case of $(s, a)$-rectangular RMGs, but their methods require an efficient planning oracle that does not currently exist in the literature. Creating such a planning oracle is one of the goals of this work. Lastly, adding a regularizer to the value function of an MG has shown promise to improve robustness empirically, but formal guarantees have not been shown in the multi-agent setting (Zhang et al., 2020a).

**Our Contributions.** We study the computational complexity of computing MPRNE for RMGs with $s$-rectangular uncertainty. We show that computing an MPRNE of an RMG with $s$-rectangular uncertainty can be done by computing a Markov-perfect NE (MPNE) of an appropriately designed regularized MG. In particular, the regularizer corresponds to the support function of the uncertainty set for the stage game. Furthermore, we show that for most well-known regularization functions, such as entropy and $\ell_p$ norm regularizers, the set of MPNE for a regularized game corresponds exactly to the set of MPRNE for an RMG with an interpretable uncertainty. This fact implies that for common classes of regularizers and uncertainty sets, the problems of solving RMGs and regularized MGs are polynomial-time equivalent. Thus, any efficient off-the-shelf algorithm for regularized MGs can be used to efficiently compute robust policies, confirming the empirical phenomenon mathematically.

We also show that many useful classes of RMGs with $s$-rectangular reward uncertainty can be solved in polynomial time. As in classical game theory, our first step is to extend the notion of zero-sum games to the robust setting. Although the zero-sum property does not guarantee efficiency, our proof of computational hardness reveals a key structural bottleneck to efficiency: general-sum behavior is simulated whenever the support function output of the reward uncertainty set involves a product of each player's policy. In contrast, if the support function decomposes into separate parts for both players, $\sigma(\boldsymbol{\pi}) = \Omega_1(\pi_1) + \Omega_2(\pi_2)$, we show the equivalent regularized MG is also zero-sum. Thus, our planning algorithm runs in polynomial time so long as the uncertainty satisfies what we call the *efficiently player-decomposable assumption*. This assumption permits many standard sets such as $L_1$ and $L_\infty$-ball uncertainty sets.

### 1.1. Related Work

**Robust MDPs.** Robust MDPs have been studied under many different uncertainty structures. The original structure, called $(s, a)$-rectangularity, was first introduced in (Satia & Lave, 1973; Nilim & El Ghaoui, 2003). Many attempts to generalize $(s, a)$-rectangularity have led to a rich family of rectangularity notions including $s$-rectangularity (Epstein & Schneider, 2003; Wiesemann et al., 2013), $r$-rectangularity (Goh et al., 2018; Goyal & Grand-Clément, 2023), $k$-rectangularity (Mannor et al., 2016), and $d$-rectangularity (Ma et al., 2023b). Many standard MDP techniques have also been extended to the robust setting including dynamic programming (Iyengar, 2005; Ho et al., 2018), policy iteration (Kaufman & Schaefer, 2013; Ho et al., 2021), policy gradient (Kumar et al., 2023; Wang et al., 2023), and function approximation (Lim & Autef, 2019; Tamar et al., 2014). Regularized MDP techniques also successfully solve robust MDPs due to a general equivalence for many uncertainty sets (Derman et al., 2021; 2023; Kumar et al., 2022) including both $(s, a)$ and $s$-rectangularity.

In the learning setting, standard RL approaches have been successfully "robustified" including model-based approaches (Wang & Zou, 2021), Q-learning (Liu et al., 2022), policy gradient (Wang & Zou, 2022; Badrinath & Kalathil, 2021), and kernel methods (Lim & Autef, 2019). Strong theoretical results have also pinned down the sample complexity of many methods (Panaganti & Kalathil, 2022; Shi & Chi, 2023; Yang et al., 2022). In fact, Pinto et al. (2017) showed that robust learning is equivalent to learning in adversarial games and this is further exploited using game-theoretical techniques (Hayashi et al., 2005).

**Robust MGs.** Robust normal form games were first introduced by Aghassi & Bertsimas (2006). Perchet (2020) showed that robust games can be reduced to general sum games, but computationally efficient methods have yet to be established. The notion of robustness has been extended to Markov games (Zhang et al., 2020b; Kardeş et al., 2011). A sample efficient approach for learning robust policies under $(s, a)$-rectangularity is derived by Blanchet et al. (2023), but their method relies on a planning oracle that has yet to be derived in the literature. Our work provides the efficient planning oracle needed to make those methods tractable and extends beyond just $(s, a)$-rectangularity. In contrast, Ma et al. (2023a) addresses the problem of learning a robust CCE with low sample complexity whereas we focus on computing the stronger solution concept of robust NE.

**Regularization in MDP and MGs.** Various regularization methods have been extensively used in MDPs (Kumar et al., 2023; Geist et al., 2019) and games (Grill et al., 2019; Cen et al., 2021; Zhang et al., 2023; Mertikopoulos & Sandholm, 2016), with diverse motivations, such as improved exploration (Lee et al., 2018), stability (Schulman et al., 2017) and convergence (Cen et al., 2021; Zhan et al., 2023). Popular regularizers include a variety of entropy functions and KL divergence. Recent works relate regularization to robustness in MDP/RL (Brekelmans et al., 2022; Eysenbach & Levine, 2021; Husain et al., 2021). In particular, Derman et al. (2023) provides an equivalence between regularization and robustness in MDPs. However, the robustness-regularization duality is much less understood in games. Our work fills this gap and opens the path to efficient planning and learning algorithms for achieving robustness in games via regularization.

**Notation.** For an integer $N$, we denote $[N] := \{1, 2, \ldots, N\}$. We define the extended reals by $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. For a given finite set $\mathcal{Z}$, we denote by $\Delta(\mathcal{Z})$ the probability simplex over $\mathcal{Z}$, and denote by $\mathbb{R}^{\mathcal{Z}}$ the class of real-valued functions defined over $\mathcal{Z}$. For a set $\mathcal{M} \subset \mathbb{R}^{\mathcal{Z}}$, the characteristic function $\delta_{\mathcal{M}} : \mathbb{R}^{\mathcal{Z}} \to \overline{\mathbb{R}}$ is defined as $\delta_{\mathcal{M}}(x) = 0$ if $x \in \mathcal{M}$ and $+\infty$ otherwise. The Legendre-Fenchel transform of $\delta_{\mathcal{M}}$ is the so-called support function $\sigma_{\mathcal{M}} : \mathbb{R}^{\mathcal{Z}} \to \overline{\mathbb{R}}$, with $\sigma_{\mathcal{M}}(y) := \sup_{x \in \mathcal{M}} \langle x, y \rangle$. For a compact set $\mathcal{W} \subset \mathbb{R}^n$, we denote the interior of $\mathcal{W}$ by $\text{int}(\mathcal{W}) \subset \mathcal{W}$, and the boundary of $\mathcal{W}$ by $\text{Bd}(\mathcal{W}) = \mathcal{W} \setminus \text{int}(\mathcal{W})$.

With slight overload of notation, $\langle \cdot, \cdot \rangle$ denotes the standard inner product when the inputs are vectors, and denotes the

Frobenius (also known as component-wise) inner product when the inputs are matrices of the same dimensions. We will overload the functions such as $\log(\cdot)$ and $\exp(\cdot)$ to take vector inputs, meaning the function is applied in an entry-wise manner. That is, for a vector $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, the notation $\exp(x) := (\exp(x_1), \ldots, \exp(x_n))^\top \in \mathbb{R}^n$. For a matrix $B \in \mathbb{R}^{m\times n}$, we denote by $\|B\|_{p\to q} = \sup\left\{ \|Bx\|_q \ : \ x \in \mathbb{R}^n \text{ with } \|x\|_p = 1\right\}$ the operator norm on the space $\mathbb{R}^{m\times n}$, where $\|\cdot\|_p$ denotes the vector $\ell_p$-norm. The dual to a norm $\|\cdot\|$ is defined as $\|v\|_* = \sup_{\|u\|\le 1}\langle u, v\rangle$. For any number $p \in [0,\infty]$, we use $p^* \in [0,\infty]$ to denote its conjugate satisfying $\frac{1}{p} + \frac{1}{p^*} = 1$. Therefore, the dual norm of $\|\cdot\|_p$ is $\|\cdot\|_{p^*}$.

## 2. Preliminaries

In this work, we consider $H$-horizon Markov games (MG) of $N$ players, with a finite state space $\mathcal{S}$, and a finite action space $\mathcal{A}_i$ for each player $i \in [N]$. We let $\mathcal{A} := \mathcal{A}_1 \times \mathcal{A}_2 \cdots \times \mathcal{A}_N$ denote the joint action space and let $\boldsymbol{a} = (a_1, \ldots, a_N) \in \mathcal{A}$ denote a joint action of $N$ players. Without loss of generality, we assume that the initial state $s_1$ is fixed [1]. For such a MG, we use $G = (P, \boldsymbol{r})$ to represent the game model, where $P = \{P_h : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})\}_{h\in[H]}$ is the transition kernel, and $\boldsymbol{r} = \{r_{i,h}\}_{i\in[N],h\in[H]}$ is the deterministic reward function, with $r_{i,h}(s,\boldsymbol{a})$ being the reward for player $i \in [N]$ given that the joint action $\boldsymbol{a}$ is applied at state $s$ at step $h$.

A Markovian policy for player $i$ is denoted by $\pi_i = \{\pi_{i,h} : \mathcal{S} \to \Delta(\mathcal{A}_i)\}_{h\in[H]}$, with $\pi_{i,h}(\cdot|s)$ being the the strategy of player $i$ at state $s$ at step $h$. We use $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_N)$ to represent a *product* joint policy of the $N$ players. For each player $i \in [N]$, $\boldsymbol{\pi}_{-i}$ denotes the joint policy of all players except player $i$. Let $\Pi_i$ be the set of all Markov policies for player $i$, and $\Pi := \Pi_1 \times \Pi_2 \times \cdots \times \Pi_N$ be the set of all product Markov joint policies of the $N$ players, and $\Pi_{-i} = \Pi_1 \times \cdots \times \Pi_{i-1} \times \Pi_{i+1} \cdots \times \Pi_N$. We overload the notation $\Delta(\mathcal{A}) := \Delta(\mathcal{A}_1) \times \cdots \Delta(\mathcal{A}_N)$ to represent the space of all product joint policy of the $N$ players at every single state.

### 2.1. Robust Markov Game

**Robust solution concept.** Recall that a robust Markov game (RMG) $(\mathcal{S}, \{\mathcal{A}_i\}_{i\in[N]}, P^\star, \boldsymbol{r}^\star, H, \mathcal{U})$ is defined by a standard MG $G^\star := (P^\star, \boldsymbol{r}^\star)$, called the nominal game, and an uncertainty set $\mathcal{U} := \mathcal{P} \times \mathcal{U}^r$ that describes all possible models. Here, $\mathcal{P}$ is the set of possible transition kernels, and $\mathcal{U}^r$ is the set of possible reward functions. Given a product joint policy $\boldsymbol{\pi}$, for each player $i \in [N]$ we define the *robust value functions* of $\boldsymbol{\pi}$ with respect to the uncertainty set $\mathcal{U}$ as follows: $\forall s \in \mathcal{S}, \mu \in \Delta(\mathcal{A}), h \in [H]$,

$$V^{\boldsymbol{\pi}}_{i,h}(s) := \inf_{G\in\mathcal{U}} \overline{V}^{\boldsymbol{\pi}}_{i,h}(s, G), \tag{1}$$

$$Q^{\boldsymbol{\pi}}_{i,h}(s,\mu) := \inf_{G\in\mathcal{U}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\overline{Q}^{\boldsymbol{\pi}}_{i,h}(s,\boldsymbol{a},G)\right], \tag{2}$$

where $\overline{V}^{\boldsymbol{\pi}}_{i,h}(\cdot, G)$ and $\overline{Q}^{\boldsymbol{\pi}}_{i,h}(\cdot,\cdot,G)$ are the standard value functions for the MG $G = (P, \boldsymbol{r})$ (cf. equations (10)-(11) in Appendix B). Similar to the Bellman equation for standard MG, we have a robust Bellman equation, as stated in Proposition B.1.

We let $V^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s)$ denote the optimal value for player $i$ starting from state $s$ at step $h$, given a product Markov joint policy $\boldsymbol{\pi}_{-i}$ of all players except player $i$, i.e.,

$$V^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s) = \sup_{\pi'_i\in\Pi_i} V^{\pi'_i\times\boldsymbol{\pi}_{-i}}_{i,h}(s), \quad \forall s \in \mathcal{S}, h \in [H]. \tag{3}$$

A policy $\pi_i$ that attains the optimal value $V^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s)$, for all $s \in \mathcal{S}, h \in [H]$ is a robust best response policy to a given $\boldsymbol{\pi}_{-i}$. It is well-known that when $\boldsymbol{\pi}_{-i}$ is Markovian, the best response amongst all history-dependent policies is Markovian, as it reduces to solving a single-agent robust MDP problem (Iyengar, 2005). Our work will focus on Markov policies $\pi_i$ as above. Compared with the best response policy in standard MGs, the robust best response policy maximizes the *robust* value function of each player $i$ given other players' policy $\boldsymbol{\pi}_{-i}$, leading to the notion of *robust Nash equilibrium* (Zhang et al., 2020b; Blanchet et al., 2023).

**Definition 2.1.** A joint product policy $\boldsymbol{\pi} = \{\boldsymbol{\pi}_h\}_{h\in[H]}$ is a *Markov perfect robust Nash equilibrium* if it holds that

$$V^{\boldsymbol{\pi}}_{i,h}(s) = V^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s), \qquad \forall i \in [N], h \in [H], s \in \mathcal{S}.$$

**Rectangularity.** It is common to choose the uncertainty set centered around the nominal model $G^\star := (P^\star, \boldsymbol{r}^\star)$. In this work, we consider reward uncertainty sets of the form $\mathcal{U}^r = \boldsymbol{r}^\star + \mathcal{R}$. We allow the reward uncertainty sets to potentially depend on the players' policy $\boldsymbol{\pi} \in \Pi$, denoted by $\mathcal{U}^r(\boldsymbol{\pi}) = \boldsymbol{r}^\star + \mathcal{R}(\boldsymbol{\pi})$. When the context is clear, we will drop the notation $(\boldsymbol{\pi})$ for the ease of exposition. As the robust MDP literature (Wiesemann et al., 2013; Derman et al., 2023; Nilim & El Ghaoui, 2003), we consider uncertainty sets that satisfy certain *rectangular condition*.

**Definition 2.2.** [$s$-rectangular Uncertainty Set] The uncertainty sets $\mathcal{U} := \mathcal{P} \times \mathcal{U}^r$ with $\mathcal{U}^r = \boldsymbol{r}^\star + \mathcal{R}$ satisfy

$$\mathcal{P} = \times_{(s,h)\in\mathcal{S}\times[H]}\mathcal{P}_{s,h}, \quad \mathcal{U}^r = \times_{(i,s,h)\in[N]\times\mathcal{S}\times[H]}\mathcal{U}^r_{i,s,h},$$

where $\mathcal{P}_{s,h} \subset \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and $\mathcal{U}^r_{i,s,h} = r^\star_{i,h}(s,\cdot) + \mathcal{R}_{i,s,h} \subset \mathbb{R}^{\mathcal{A}}$ are closed, convex sets.

[1] For the case that the initial state is stochastic, one can add $s_0$ as the initial state and set the initial state distribution as the transition kernel from $s_0$ to $s_1$.

**Definition 2.3.** [$(s,a)$-Rectangular Uncertainty Set] A special case of $s$-rectangularity is a $(s,a)$-rectangular uncertainty set:

$$\mathcal{P} = \times_{(s,\boldsymbol{a},h)\in\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{P}_{s,\boldsymbol{a},h},$$
$$\mathcal{U}^r = \times_{(i,s,\boldsymbol{a},h)\in[N]\times\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{U}^r_{i,s,\boldsymbol{a},h},$$

where $\mathcal{P}_{s,\boldsymbol{a},h} \subset \Delta(\mathcal{S})$ and $\mathcal{U}^r_{i,s,\boldsymbol{a},h} = r^\star_{i,h}(s,\boldsymbol{a}) + \mathcal{R}_{i,s,\boldsymbol{a},h} \subset \mathbb{R}$ are closed, convex sets.

We remark that the state-action value function defined in (2) is different from the classical setting. The new definition here allows us to provide a unified framework for both $s$-rectangular and $(s,a)$-rectangular uncertainty sets.

**Robust suboptimality gap.** Given a joint policy $\boldsymbol{\pi}$, for each player $i \in [N]$, we define the robust suboptimality gap at step $h$ as

$$\text{RNEGap}_{i,h}(\boldsymbol{\pi}, s) := V^{\dagger,\boldsymbol{\pi}_{-i}}_{i,h}(s) - V^{\boldsymbol{\pi}}_{i,h}(s). \tag{4}$$

That is, the RNE gap measures the suboptimality gap of player $i$'s policy $\pi_i$ against its robust best response policy given all other players' policy $\boldsymbol{\pi}_{-i}$. It is clear that $\text{RNEGap}_{i,h}(\boldsymbol{\pi}, s) \geq 0$ for all product joint $\boldsymbol{\pi}$ and all $s \in \mathcal{S}$ and $h \in [H]$. For any MPRNE policy $\boldsymbol{\pi}^*$, we have $\text{RNEGap}_{i,h}(\boldsymbol{\pi}^*, s) = 0$.

**Existence of RNE.** Recall that the reward uncertainty sets can depend on the players' policy $\boldsymbol{\pi} \in \Pi$, denoted by $\mathcal{U}^r(\boldsymbol{\pi})$. We consider the following assumption on the reward uncertainty sets, which ensures the existence of robust Nash equilibrium. We first define the continuity of a point-to-set function. Consider a function $f$ that maps each $z \in \mathbb{R}^{d_1}$ to a set in $\mathbb{R}^{d_2}$. The function $f$ is continuous at $z \in \mathbb{R}^{d_1}$ if it satisfies the following: for each $\epsilon > 0$, there exists $\delta(\epsilon) > 0$ such that for all $z' \in \mathbb{R}^{d_1}$ with $\|z - z'\|_\infty < \delta(\epsilon)$, it holds that $D\big(f(z), f(z')\big) := \max\big\{ \sup_{y\in f(z)} \inf_{y'\in f(z')} \|y - y'\|_\infty,$ $\sup_{y'\in f(z')} \inf_{y\in f(z)} \|y - y'\|_\infty \big\} < \epsilon$.

**Assumption 2.1.** Consider $s$-rectangular uncertainty set. For each $\boldsymbol{\pi} \in \Pi$, the reward uncertainty set $\mathcal{U}^r(\boldsymbol{\pi}) = \times_{(i,s,h)\in[N]\times\mathcal{S}\times[H]}\mathcal{U}^r_{i,s,h}(\boldsymbol{\pi})$, where $\mathcal{U}^r_{i,s,h}(\boldsymbol{\pi}) = r^\star_{i,h}(s,\cdot) + \mathcal{R}_{i,s,h}(\boldsymbol{\pi}) \subset \mathbb{R}^{\mathcal{A}}$ satisfies the following for all $(i,s,h) \in [N]\times\mathcal{S}\times H$:

1. Bounded game value: There exists a constant $L_r$ such that for each $r \in \mathcal{U}^r_{i,s,h}(\boldsymbol{\pi})$, it holds that $L_r \leq \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}(\cdot|s)}\left[r(\boldsymbol{a})\right]$.
2. Convexity: The support function $\sigma_{\mathcal{R}_{i,s,h}(\boldsymbol{\pi})} : \mathbb{R}^{\mathcal{A}} \to (-\infty, +\infty]$ of $\mathcal{R}_{i,s,h}(\boldsymbol{\pi})$, defined as $\sigma_{\mathcal{R}_{i,s,h}(\boldsymbol{\pi})}(y) := \sup_{x\in\mathcal{R}_{i,s,h}(\boldsymbol{\pi})} \langle x, y\rangle$, satisfies that $\sigma_{\mathcal{R}_{i,s,h}(\boldsymbol{\pi})}\big(-\pi_{i,h}(s)\boldsymbol{\pi}^\top_{-i,h}(s)\big)$ is convex in $\pi_{i,h}(s)$ and continuous at all $\boldsymbol{\pi} \in \text{Bd}(\Pi)$.
3. Continuity: The set $\mathcal{R}_{i,s,h}(\cdot)$ is continuous at all $\boldsymbol{\pi} \in \text{int}(\Pi)$; and $\sup_{y\in\mathcal{R}_{i,s,h}(\boldsymbol{\pi})} \|y\|_\infty < \infty$ for all $\boldsymbol{\pi} \in \text{int}(\Pi)$.

We remark that a special reward uncertainty set that is policy-independent and bounded satisfies Assumption 2.1 (Aghassi & Bertsimas, 2006). Such uncertainty set has been considered in discounted Markov games (Kardeş et al., 2011; Zhang et al., 2020b). In this paper, we consider more general reward uncertainty sets. Importantly, the boundedness condition in Assumption 2.1 does not require the uncertainty set to be uniformly bounded. This relaxation allows us to consider uncertainty sets based on log-barrier functions.

The following theorem states that a robust Nash equilibrium (RNE) always exists. The proof is provided in Section C. A similar existence result has been proved for the case with only $(s,a)$-rectangular transition uncertainty (Blanchet et al., 2023).

**Theorem 2.1.** *(Existence of RNE). Given a RMG $(\mathcal{S}, \{\mathcal{A}_i\}_{i\in[N]}, P^\star, \boldsymbol{r}^\star, H, \mathcal{U})$ with $s$-rectangular uncertainty set $\mathcal{U}$ satisfying Assumption 2.1, the robust Nash equilibrium defined in Definition (2.1) always exists. Moreover, a joint policy $\boldsymbol{\pi}^\dagger = \{\boldsymbol{\pi}^\dagger_h\}_{h\in[H]}$ defined as follows is an MPRNE:*

$$\boldsymbol{\pi}^\dagger_h(\cdot \mid s) \in \text{NE}\left(\{Q^{\boldsymbol{\pi}^\dagger}_{i,h}(s,\cdot)\}_{i\in[N]}\right), \forall s \in \mathcal{S}, h \in [H],$$

*where* $\text{NE}(\cdot)$ *denotes the Nash equilibrium of a general-sum, normal-form game.*

### 2.2. Regularized Markov Games

A regularized Markov game with $N$ players can be described by a tuple $(\mathcal{S}, \{\mathcal{A}_i\}_{i\in[N]}, P, \boldsymbol{r}, H, \Omega)$, with $G = (P, \boldsymbol{r})$ being a standard MG model. Here $\Omega := (\Omega_{i,h})_{i\in[N],h\in[H]}$ is a set of policy regularization functions such that for all $(i,h) \in [N]\times[H]$, $\Omega_{i,h} : \mathcal{S}\times\Delta(\mathcal{A}) \to \mathbb{R}$ is convex in $\pi_{i,h}$ given a fixed $\boldsymbol{\pi}_{-i}$. Given a joint policy $\boldsymbol{\pi} \in \Pi$, the *regularized value functions* for each player $i$ are defined as follows: $\forall s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}, h \in [H]$,

$$\widetilde{V}^{\boldsymbol{\pi}}_{i,h}(s, G) = \mathbb{E}^{\boldsymbol{\pi}}_P\Big[\sum_{t=h}^{H} r_{i,t}(s_t, \boldsymbol{a}_t) - \Omega_{i,t}(s_t, \boldsymbol{\pi}_t) | s_h = s\Big],$$

$$\widetilde{Q}^{\boldsymbol{\pi}}_{i,h}(s, \boldsymbol{a}, G) = r_{i,h}(s, \boldsymbol{a}) + \mathbb{E}^{\boldsymbol{\pi}}_P\Big[\sum_{t=h+1}^{H} r_{i,t}(s_t, \boldsymbol{a}_t) - \Omega_{i,t}(s_t, \boldsymbol{\pi}_t) | s_h = s, \boldsymbol{a}_h = \boldsymbol{a}\Big].$$

A common solution concept for regularized MG is the *Markov perfect Nash equilibrium* (MPNE).

**Definition 2.4.** A joint product policy $\boldsymbol{\pi} = \{\boldsymbol{\pi}_h\}_{h\in[H]}$ is an MPNE for a regularized MG if it holds that $\widetilde{V}^{\boldsymbol{\pi}}_{i,h}(s) = \sup_{\pi'_i\in\Pi_i} \widetilde{V}^{\pi'_i\times\boldsymbol{\pi}_{-i}}_{i,h}(s), \forall i \in [N], h \in [H], s \in \mathcal{S}$.

## 3. Markov Games with Reward Uncertainty

In this section, we consider robust Markov games with only reward uncertainty. To solve such RMGs, we take inspiration from the single-player setting. We follow the ideas of Derman et al. (2023) to show that solving RMGs can be done through solving regularized MGs. In fact, for many common regularizers, the reverse is also true, which implies *regularization yields robust solutions*. Importantly, the equivalence results mean existing efficient, regularized MG solvers can be used off-the-shelf to efficiently solve robust MGs. Proofs for the results in this section are deferred to Appendix D.

### 3.1. Matrix Games

To build intuition for our results, we first study matrix games with $N$ players since they can be viewed as a simple Markov game with $S = H = 1$. Already through matrix games, we see that reward uncertainty is much more complex to handle than in the single-agent setting. Nevertheless, by understanding the support functions induced by uncertainty sets, we can derive equivalent regularized matrix games.

**Reward Structure.** We consider $s$-rectangular reward uncertainty set of the form $\mathcal{U} = \boldsymbol{r}^\star + \mathcal{R}$, where $\boldsymbol{r}^\star \in \mathcal{U}$ is the nominal model, and $\mathcal{R} = \times_{i\in[N]}\mathcal{R}_i$ with $\mathcal{R}_i \subset \mathbb{R}^{\mathcal{A}}$. Recall that the uncertainty set $\mathcal{R}$ can potentially depend on the players' joint policy $\boldsymbol{\pi} \in \Delta(\mathcal{A})$. We observe that the robust value for player $i$ can be simplified to

$$V_i^{\boldsymbol{\pi}} = \inf_{\boldsymbol{r}\in\mathcal{U}} \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}}\left[r_i(\boldsymbol{a})\right] = \inf_{\boldsymbol{r}\in\mathcal{U}} \pi_i^\top r_i \boldsymbol{\pi}_{-i}.$$

Here, we view $r_i$ as a matrix in $\mathbb{R}^{\mathcal{A}_i\times\mathcal{A}_{-i}}$, $\pi_i$ as a column vector in $\mathbb{R}^{\mathcal{A}_i}$ and $\boldsymbol{\pi}_{-i}$ as a column vector in $\mathbb{R}^{\mathcal{A}_{-i}}$. By Proposition C.2, we can equivalently rewrite the robust value function as follows:

$$V_i^{\boldsymbol{\pi}} = \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i}(-\pi_i\boldsymbol{\pi}_{-i}^\top), \tag{5}$$

where $\sigma_{\mathcal{R}_i}(y) = \sup_{x\in\mathcal{R}_i}\langle x, y\rangle$ denotes the support function of $\mathcal{R}_i$.

**Equivalence between Robustness and Regularization.** If we consider the support function $\sigma_{\mathcal{R}_i}$ as a regularizer, equation (5) shows an equivalence between player $i$'s robust game value $V_i^{\boldsymbol{\pi}}$ and the value $\widetilde{V}_i^{\boldsymbol{\pi}}(s, G^\star)$ in a $\sigma_{\mathcal{R}_i}$-regularized game. Note that the RNE of the matrix game is a product joint policy $\boldsymbol{\pi}^\dagger$ such that $\pi_i^\dagger$ is the robust best response policy to $\boldsymbol{\pi}_{-i}^\dagger$. That is,

$$\pi_i^\dagger \in \operatorname*{arg\,sup}_{\pi_i\in\Delta(\mathcal{A}_i)} V_i^{\pi_i\times\boldsymbol{\pi}_{-i}^\dagger} = \operatorname*{arg\,sup}_{\pi_i\in\Delta(\mathcal{A}_i)} \widetilde{V}_i^{\pi_i\times\boldsymbol{\pi}_{-i}^\dagger}(s, G^\star),$$

which implies an equivalence between solving a robust game and solving a regularized game.

**Theorem 3.1.** *Consider a robust matrix game $\mathfrak{G}$ with uncertainty set $\mathcal{U} = \boldsymbol{r}^\star + \mathcal{R}$ satisfying Assumption 2.1, where $\boldsymbol{r}^\star \in \mathcal{U}$ is the nominal model, and $\mathcal{R} = \times_{i\in[N]}\mathcal{R}_i$. Consider a regularized normal form game $\mathfrak{G}'$ with payoff matrix $\boldsymbol{r}^\star$ and the regularizer function $\Omega_i : \Delta(\mathcal{A}) \to \mathbb{R}$ for each player $i \in [N]$ defined as $\Omega_i(\boldsymbol{\pi}) := \sigma_{\mathcal{R}_i}(-\pi_i\boldsymbol{\pi}_{-i}^\top)$. Then, $\boldsymbol{\pi}$ is a RNE of robust game $\mathfrak{G}$ if and only if $\boldsymbol{\pi}$ is a NE of regularized game $\mathfrak{G}'$.*

We provide the proof of Theorem 3.1 in Appendix D.1.

**Corollary 3.2.** *Robust matrix games can be solved using any planning algorithm for regularized games.*

**Interpretable Equivalence.** Thus, we see that we can solve a given robust game by solving a particular regularized game. However, since our reduction maps robust games to very specialized regularized games, it is unclear whether commonly used regularized methods can be used to solve robust games. Fortunately, we can show for many common classes of regularizers, solutions to the regularized game correspond to the solutions of robust games with interpretable uncertainty sets.

**Theorem 3.3.** *Consider a regularized normal form game $\mathfrak{G}'$ with payoff matrix $\boldsymbol{r}^\star$ and the regularizer $\Omega_i : \Delta(\mathcal{A}) \to \mathbb{R}$ for each $i \in [N]$.*

1. *If $\Omega_i$ is $\ell_p/\ell_q$-norm regularization, i.e. $\Omega_i(\boldsymbol{\pi}) := \alpha_i \left\|-\pi_i\right\|_p \left\|\boldsymbol{\pi}_{-i}\right\|_q$ for each player $i \in [N]$, then solving for the NE of $\mathfrak{G}'$ is equivalent to solving for RNE of the robust game $\mathfrak{G}$ with $s$-rectangular ball uncertainty $\mathcal{U} = \boldsymbol{r}^\star + \times_{i\in[N]}\mathcal{R}_i$, where $\mathcal{R}_i = \left\{R_i \in \mathbb{R}^{\mathcal{A}_i\times\mathcal{A}_{-i}} : \left\|R_i\right\|_{q^*\to p} \le \alpha_i\right\}$.*
2. *If $\Omega_i$ is strongly convex and decomposable with kernel $\omega$, i.e., $\Omega_i(\boldsymbol{\pi}) := \tau_i \sum_{a_i} \pi_i(a_i)\omega_i(\pi_i(a_i))$ for each player $i \in [N]$ with $\tau_i \ge 0$, then solving for a NE of $\mathfrak{G}'$ is equivalent to solving for a RNE of robust game $\mathfrak{G}$ with $(s, a)$-rectangular, policy-dependent uncertainty set $\mathcal{U}(\boldsymbol{\pi}) = \boldsymbol{r}^\star + \times_{i\in[N],\boldsymbol{a}\in\mathcal{A}}\mathcal{R}_{i,\boldsymbol{a}}(\boldsymbol{\pi})$, where*

$$\begin{aligned}\mathcal{R}_{i,\boldsymbol{a}}(\boldsymbol{\pi}) = &\big[\tau_i\omega_i(\pi_i(a_i)) + g_i(\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i})),\\ &\overline{\omega}_i(\pi_i(a_i)) + \overline{g}_i(\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}))\big] \subset \mathbb{R},\end{aligned} \tag{6}$$

*with functions $\omega_i, \overline{\omega}_i : [0, 1] \to \mathbb{R}$ and $g_i, \overline{g}_i : [0, 1] \to \mathbb{R}$ are continuous.*

See Appendix D.2 for the proof of Theorem 3.3. We remark that the $s$-rectangular ball-constrained uncertainty set satisfies Assumption 2.1. The policy-dependent uncertainty set in (6) also satisfies Assumption 2.1 by properly choosing the functions $\omega_i, \overline{\omega}_i, g_i, \overline{g}_i$. Theorem 3.3 implies that the shape of the reward uncertainty set determines the equivalent regularizer function. For example, a ball-constrained uncertainty set corresponds to norm regularization. Also, the size of the uncertainty set, e.g., the radius parameter $\alpha_i$, determines the magnitude of the regularization factor.

**Corollary 3.4.** *For any regularizer considered in Theorem 3.3, solutions to the regularized game are provably robust.*

*Remark.* Observe that as long as the functions for regularization and uncertainty sets in Theorem 3.3 are efficiently computable, then given either a regularized game or a robust game, the construction of the equivalent game can be done in polynomial time. Therefore, Theorem 3.3 implies the problem of computing an RNE of a robust game for special classes of uncertainty is *polynomial-time equivalent* to the problem of computing an NE of a regularized game with special classes of regularizers. This means the computational complexity of both problems is the same. In particular, an efficient algorithm for one problem implies an efficient algorithm for the other.

**Examples of regularization.** A classical example of decomposable regularizers is the negative Shannon entropy $\Omega_i(\boldsymbol{\pi}) = \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i) \log \pi_i(a_i)$. Entropy regularization is applied extensively in both single-agent MDP and multi-agent games (Kumar et al., 2023; Grill et al., 2019; Geist et al., 2019; Zhan et al., 2023; Cen et al., 2021; Zhang et al., 2023), and has been shown to accelerate the convergence of many learning algorithms. Another example is KL divergence regularizer (Schulman et al., 2017) $\Omega_i(\boldsymbol{\pi}) = \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i) \log \frac{\pi_i(a_i)}{\mu_i(a_i)} = d_{\mathrm{KL}}(\pi_i, \mu_i)$, where $\mu_i \in \Delta(\mathcal{A}_i)$ is a given distribution. The literature has also considered the negative Tsallis entropy regularizer (Lee et al., 2018) $\Omega_i(\boldsymbol{\pi}) = \frac{1}{2}\left( \|\pi_i\|_2^2 - 1 \right) = \frac{1}{2} \sum_{a_i \in \mathcal{A}_i} \left( \pi_i(a_i)^2 - \pi_i(a_i) \right)$. Theorem 3.3 also applies to other regularizers, such as the Renyi entropy regularization (Mertikopoulos & Sandholm, 2016); See Section D.3 for additional discussion.

### 3.2. Markov Games

The intuition for matrix games extends directly to Markov games. The main additional ingredient needed for the analysis is backward induction. By applying Theorem 3.1 to each stage game, we can construct a regularized MG whose MPNE are all MPRNE for the RMG.

We consider Markov games with $s$-rectangular reward uncertainty of the form $\mathcal{U} = P^\star \times (\boldsymbol{r}^\star + \mathcal{R})$. Similar to matrix games, we see that each player $i$'s robust value function is equivalent to the value function of a regularized MG, as stated in the following proposition. The proof is provided in Appendix D.4. For notational simplicity, we define,

$$[P_h V](s, \boldsymbol{a}) := \mathbb{E}_{s' \sim P_h(\cdot|s,\boldsymbol{a})} \left[ V(s') \right].$$

**Proposition 3.5.** *Suppose the uncertainty is $s$-rectangular. Given any product joint policy $\boldsymbol{\pi} \in \Pi$, the robust value function $\left\{ V_{i,h}^{\boldsymbol{\pi}} \right\}_{h \in [H]}$ of each player $i \in [N]$ satisfies:*

$$\begin{aligned} V_{i,h}^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ r_{i,h}^\star(s, \boldsymbol{a}) + [P_h^\star V_{i,h+1}^{\boldsymbol{\pi}}](s, \boldsymbol{a}) \right], \\ - \sigma_{\mathcal{R}_{i,s,h}} \left( -\pi_{i,h}(s) \boldsymbol{\pi}_{-i,h}^\top(s) \right). \end{aligned} \quad (7)$$

Given the proposition, we can similarly construct a regularized Markov game as in the matrix game case.

**Theorem 3.6.** *Consider a robust MG $\mathfrak{G}$ with $s$-rectangular uncertainty set $\mathcal{U} = P^\star \times (\boldsymbol{r}^\star + \mathcal{R})$. Consider a regularized MG $\mathfrak{G}' = (\mathcal{S}, \{\mathcal{A}_i\}_{i \in [N]}, P^\star, \boldsymbol{r}^\star, H, \Omega)$, where the regularizer functions $\{\Omega_{i,h}\}_{h \in [H]}$ for each player $i \in [N]$ are given by $\Omega_{i,h}(s, \mu) := \sigma_{\mathcal{R}_{i,s,h}}\left( -\mu_i \mu_{-i}^\top \right), \forall s \in \mathcal{S}, h \in [H], \mu \in \Delta(\mathcal{A})$. Then, $\boldsymbol{\pi}$ is a MPRNE for $\mathfrak{G}$ if and only if $\boldsymbol{\pi}$ is a MPNE for $\mathfrak{G}'$.*

**Corollary 3.7.** *RMGs with $s$-rectangular reward uncertainty can be solved using any planning or learning algorithm for the equivalent regularized Markov games.*

*Remark.* The same results from Theorem 3.3 extend to the full RMG setting by simply using the same shape of uncertainty set for each stage $(s, h)$, as stated in Theorem D.1 in Appendix D.6. In particular, our results apply to the well-known regularized game, including $\ell_p/\ell_q$-norm regularization and decomposable regularizers such as entropy regularization and KL divergence. In the following, we give the example of widely employed entropy regularization.

**Example: Entropy Regularization.** Consider Shannon entropy regularized Markov game with

$$\Omega_{i,h}(s, \mu) = \tau_i \sum_{a_i \in \mathcal{A}_i} \mu_i(a_i) \log \mu_i(a_i), \quad \forall \mu \in \Delta(\mathcal{A}),$$

where $\tau_i > 0$ denotes regularization factor. The regularized NE is equivalent to the RNE of a robust MG with $(s, a)$-rectangular policy-dependent uncertainty set $\mathcal{U}(\boldsymbol{\pi}) = P^\star \times (\boldsymbol{r}^\star + \mathcal{R}(\boldsymbol{\pi}))$, where $\mathcal{R}(\boldsymbol{\pi}) = \times_{i,s,\boldsymbol{a},h} \mathcal{R}_{i,s,\boldsymbol{a},h}(\boldsymbol{\pi})$ with

$$\mathcal{R}_{i,s,\boldsymbol{a},h}(\boldsymbol{\pi}) = \left[ \tau_i \log \pi_{i,h}(a_i|s), \overline{\omega}_{i,s,h}(\pi_{i,h}(a_i|s)) \right],$$

where the function $\overline{\omega}_{i,s,h} : [0, 1] \to \mathbb{R}$ is continuous and non-negative to ensure that $\boldsymbol{r}^\star \in \mathcal{U}^{\boldsymbol{\pi}}$. The reader can find additional details of other regularizers in Section D.6.

## 4. Efficient Algorithms for Robust Zero-Sum MG

Although the results from Section 3 provide insights that allow us to solve RMGs, they do not yield efficient algorithms in general. Our reduction may result in general-sum regularized games which are hard to solve. As we will see shortly, even for a two-player zero-sum Markov game with reward uncertainty, computing the RNE is PPAD-hard in general. However, we show that when uncertainty sets satisfy a natural decomposition assumption, then an MPRNE of a two-player zero-sum Markov game can be computed in polynomial time. Proofs for the results in this section are deferred to Appendix E.

First, we extend the notion of zero-sum Markov games to the robust setting.

**Definition 4.1** (Two-Player Zero-Sum RMG)**.** A RMG $\mathfrak{G}$ is a *two-player zero-sum RMG* (TPZS RMG) with reward uncertainty if $N = 2$ and for each $h \in [H]$ and $s \in \mathcal{S}$, $\boldsymbol{r}^\star_{2,h}(s,\cdot) = -\boldsymbol{r}^\star_{1,h}(s,\cdot)$.

While this is a crucial first step, focusing on TPZS RMGs is insufficient to guarantee efficient algorithms. Surprisingly, even with the most restrictive notion of $(s,a)$-rectangularity, TPZS RMGs are PPAD-hard to solve. Observe this stands in sharp contrast to traditional game theory where two-player zero-sum games can be efficiently solved using LP methods.

**Theorem 4.1.** *Even restricted to the class of $(s,a)$-rectangular uncertainty sets, computing an RNE of a TPZS RMG is PPAD-hard even for $H = S = 1$.*

*Proof Sketch.* We present a poly-time reduction from the problem of computing an NE for a general-sum game to the problem of computing an RNE for a two-player zero-sum robust matrix game with $(s,a)$-rectangular reward uncertainty. Since computing an NE of a general-sum game is PPAD-hard, it then follows that computing an RNE for the aforementioned class of robust matrix games is also PPAD-hard. Let $(A, B)$ be an arbitrary general sum matrix game.

The idea is to construct a two-player zero-sum robust matrix game $\mathfrak{G}$ defined by $(\boldsymbol{r}^\star, \mathcal{R}_1, \mathcal{R}_2)$ so that the set of solutions to both games is the same. First, we choose matrices $\boldsymbol{r}^\star$, $R_1$ and $R_2$ satisfying $\boldsymbol{r}^\star + R_1 = A$ and $\boldsymbol{r}^\star + R_2 = -B$. Then, we choose $\mathcal{R}_1$ and $\mathcal{R}_2$ satisfying $-\sigma_{\mathcal{R}_1}(-\pi_1\pi_2^\top) = \pi_1^\top R_1 \pi_2$ and $-\sigma_{\mathcal{R}_1}(-\pi_1\pi_2^\top) = \pi_1^\top R_2 \pi_2$. Consequently, we see the robust suboptimality gap for $\mathfrak{G}$ simplifies to,

$$\begin{aligned}
&\mathrm{RNEGap}_1(\boldsymbol{\pi}) + \mathrm{RNEGap}_2(\boldsymbol{\pi}) \\
&= \max_{\pi_1' \in \Delta(\mathcal{A}_1)} \left\{ \pi_1'^\top (r^\star + R_1)\pi_2 \right\} - \pi_1^\top (r^\star + R_1)\pi_2 \\
&\quad + \max_{\pi_2' \in \Delta(\mathcal{A}_2)} \left\{ \pi_1^\top (-r^\star - R_2)\pi_2' \right\} - \pi_1^\top (-r^\star - R_2)\pi_2 \\
&= \max_{\pi_1' \in \Delta(\mathcal{A}_1)} \left\{ \pi_1'^\top A \pi_2 \right\} - \pi_1^\top A \pi_2 \\
&\quad + \max_{\pi_2' \in \Delta(\mathcal{A}_2)} \left\{ \pi_1^\top B \pi_2' \right\} - \pi_1^\top B \pi_2 .
\end{aligned}$$

We see the total suboptimality gap of $\boldsymbol{\pi}$ for $\mathfrak{G}$ exactly matches the suboptimality gap of $\boldsymbol{\pi}$ for $(A, B)$. This implies the set of solutions to both games is the same as desired. □

The proof of Theorem 4.1 is provided in Appendix E.1. Thanks to Theorem 4.1, we cannot hope to solve TPZS RMGs even under the simplest $(s,a)$-rectangularity assumption. However, the proof reveals key structural properties of the uncertainty set that lead to hardness. Specifically, we needed the uncertainty sets to satisfy $-\sigma_{\mathcal{R}_1}(-\pi_1\pi_2^\top) = \pi_1^\top R_1 \pi_2$. In particular, observe the support function involves a product of $\pi_1$ and $\pi_2$. This property allowed us to simulate a general sum game through the uncertainty sets.

One way to avoid such hardness is to consider uncertainty sets that break the support function into two separate pieces: one for $\pi_1$ and one for $\pi_2$. Formally, suppose that $\sigma_{\mathcal{R}_i}(-\pi_i \boldsymbol{\pi}_{-i}^\top) = \Omega_{i,i}(\pi_i) + \Omega_{i,-i}(\boldsymbol{\pi}_{-i})$. Immediately, we see that an uncertainty set satisfying $-\sigma_{\mathcal{R}_1}(-\pi_1\pi_2^\top) = \pi_1^\top R_1 \pi_2$ is no longer possible. Even better, by inspecting the RNEGap of $\mathfrak{G}$, we see that each of the $\Omega_{i,-i}(\boldsymbol{\pi}_{-i})$ terms cancel out, leading to the following RNEGap:

$$\begin{aligned}
&\max_{\pi_1' \in \Delta(\mathcal{A}_1)} \left\{ {\pi_1'}^\top r^\star \pi_2 - \Omega_{1,1}(\pi_1') + \Omega_{2,2}(\pi_2) \right\} \\
&- \min_{\pi_2' \in \Delta(\mathcal{A}_2)} \left\{ {\pi_1}^\top r^\star \pi_2' - \Omega_{1,1}(\pi_1) + \Omega_{2,2}(\pi_2') \right\} .
\end{aligned}$$

Observe this is exactly the suboptimality gap of $\boldsymbol{\pi}$ for a TPZS regularized game. Importantly, TPZS regularized games can be solved in polynomial time so long as the regularizer functions are strongly convex (Facchinei & Pang, 2003; Cherukuri et al., 2017). Thus, as long as the decomposition of $\sigma$ into $\Omega$ functions is known or can be computed efficiently, solving the robust matrix game can also be done efficiently by solving the zero-sum regularized game.

**Definition 4.2** (Efficiently Player-Decomposable)**.** Suppose $\mathfrak{G}$ is a TPZS RMG with $s$-rectangular reward uncertainty. We say that $\mathfrak{G}$ is *efficiently player-decomposable* if $\forall i \in \{1,2\}, s \in \mathcal{S}, h \in [H], \mu \in \Delta(\mathcal{A})$, $\sigma_{\mathcal{R}_{i,s,h}}(-\mu_i \mu_{-i}^\top) = \Omega^h_{i,i}(\mu_i) + \Omega^h_{i,-i}(\mu_{-i})$, where each $\Omega$ is a strongly convex function that is either known in advance or can be computed from $\mathfrak{G}$ in polynomial time.

**Lemma 4.2.** *Suppose $\mathfrak{G}$ is a robust matrix game that satisfies Definition 4.2 and $\mathfrak{G}'$ is the corresponding regularized game constructed in Theorem 3.1. Then, $\mathfrak{G}'$ is a TPZS regularized game with separate regularizers, $\Omega_i = \Omega_{i,i}$ for each player $i$. Therefore, computing an RNE for $\mathfrak{G}$ can be done in polynomial time by solving $\mathfrak{G}'$.*

The proof of Lemma 4.2 can be found in Appendix E.2. Lemma 4.2 implies we can solve each stage game so long as the decomposition holds for each stage. Thus, to solve a zero-sum RMG, we can solve each stage's regularized game using backward induction.

**Theorem 4.3.** *Suppose $\mathfrak{G}$ is an RMG that satisfies Definition 4.2 and $\mathfrak{G}'$ is the corresponding regularized MG constructed in Theorem 3.6. Then, $\mathfrak{G}'$ is a TPZS regularized MG with separate regularizers, $\Omega_{i,h}(s,\cdot) = \Omega^h_{i,i}(s,\cdot)$ for each player $i$ and stage $(s,h)$. Therefore, computing an*

*MPRNE for $\mathfrak{G}$ can be done in polynomial time by solving $\mathfrak{G}'$.*

The proof of Theorem 4.3 is provided in Appendix E.3. We immediately have the following result.

**Corollary 4.4.** *RMGs satisfying Definition 4.2 can be solved in polynomial time using any efficient planning or learning algorithm for TPZS regularized MGs.*

We point out that the classes of uncertainty set we discussed in Theorem 3.3 do not all satisfy Definition 4.2. Specifically, case 1. concerning general ball uncertainty does not satisfy our decomposition. However, we show very similar uncertainty sets do satisfy the conditions and thus can be solved efficiently, as stated in the following theorem.

**Theorem 4.5.** *Consider a regularized MG with payoff matrix $\boldsymbol{r}^\star$ and the regularizer $\Omega_{i,h} : \mathcal{S} \times \Pi \to \mathbb{R}$ for each player $i \in [N]$ and state $s \in \mathcal{S}$.*

1. *If $\Omega_{i,h}(s, \boldsymbol{\pi}_h) = \alpha_{i,s,h} \|\pi_{i,h}(s)\|_p$ is the p-norm regularizer for each $i \in [N], h \in [H], s \in \mathcal{S}$, then solving for MPNE of the regularized game is equivalent to solving for MPRNE of the robust game with ball constrained uncertainty set $\mathcal{R}_{1,s,h} = \{R_1 \in \mathbb{R}^{\mathcal{A}_1 \times \mathcal{A}_2} : \|R_1\|_{\infty \to p} \leq \alpha_{1,s,h}\}$ and $\mathcal{R}_{2,s,h} = \{R_2 \in \mathbb{R}^{\mathcal{A}_1 \times \mathcal{A}_2} : \|R_2^\top\|_{\infty \to p} \leq \alpha_{2,s,h}\}$.*
2. *The Markov game version of $(s, a)$-rectangular, policy-dependent reward uncertainty set from part 2. of Theorem 3.3 carries over without any additional restrictions.*

*In either case, both the RMG and regularized MG can be solved in polynomial time.*

*Remark.* As with Theorem 3.3, classical and common regularizers can be applied as special cases of Theorem 4.5. See Appendix D.3 for additional discussion.

## 5. Markov Games with Transition Uncertainty

Up to this point, we have focused on RMGs with reward uncertainty. In this section, we consider transition uncertainty. Similar to before, we show that RMGs with transition uncertainty can be solved using regularized MG methods. One can easily integrate the results of this section with those of Section 3 to characterize general RMGs with both reward and transition uncertainty. Proofs for the results in this section are deferred to Appendix F.

If $\mathfrak{G}$ is an RMG with $s$-rectangular transition uncertainty, then the uncertainty set takes the form $\mathcal{U} = \mathcal{P} \times \{\boldsymbol{r}^\star\}$. We can again derive a robust policy evaluation equation for $\mathfrak{G}$ similar to that in Proposition C.2 for reward uncertainty.

**Proposition 5.1.** *Consider a RMG $\mathfrak{G}$ with $s$-rectangular uncertainty set $\mathcal{U} = \mathcal{P} \times \{\boldsymbol{r}^\star\}$. Then for each product joint Markovian policy $\boldsymbol{\pi} \in \Pi$, the robust value function $\{V_{i,h}^{\boldsymbol{\pi}}\}_{h \in [H]}$ of each player $i \in [N]$ satisfies*

$$V_{i,h}^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ r_{i,h}^\star(s, \boldsymbol{a}) \right] - \sigma_{\mathcal{P}_{s,h}} \left( -V_{i,h+1}^{\boldsymbol{\pi}} \boldsymbol{\pi}_h(s)^\top \right) \tag{8}$$

*where $[V_{i,h+1}^{\boldsymbol{\pi}} \boldsymbol{\pi}_h(s)^\top](s', \boldsymbol{a}') = V_{i,h+1}^{\boldsymbol{\pi}}(s') \boldsymbol{\pi}_h(\boldsymbol{a}'|s)$ for all $s' \in \mathcal{S}, \boldsymbol{a}' \in \mathcal{A}$.*

Observe that (8) replaces the linear expected future value function in a standard MG with a support function $\sigma_{\mathcal{P}_{s,h}}(\cdot)$, which depends on both the policy and future robust value. Depending on the uncertainty set, the support function might involve a *non-linear* transformation of the policy and the value function.

To accommodate the non-linearity, we introduce a generalization of regularized Markov games where $\Omega := (\Omega_{s,h})_{s \in \mathcal{S}, h \in [H]}$ is a finite set of policy-value regularization functions such that for all $s \in \mathcal{S}, h \in [H]$, $\Omega_{s,h} : \Delta(\mathcal{S})^{\mathcal{A}} \times \mathbb{R}^{\mathcal{S}} \times \Delta(\mathcal{A}) \to \mathbb{R}$ satisfies that for each $P_h \in \Delta(\mathcal{S})^{\mathcal{A}}$ and for each $v \in \mathbb{R}^{\mathcal{S}}$, $\Omega_{s,h}(P_h, v, \cdot)$ is convex. Given a joint policy $\boldsymbol{\pi}$ for a MG with $G = (P, \boldsymbol{r})$, the general *policy-value regularized* value function is defined recursively as follows: $\forall s \in \mathcal{S}, h \in [H]$,

$$\begin{aligned} \widehat{V}_{i,h}^{\boldsymbol{\pi}}(s, G) := &\mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ r_{i,h}(s, \boldsymbol{a}) \right] \\ &- \Omega_{s,h} \left( P_h, -V_{i,h+1}^{\boldsymbol{\pi}}, \boldsymbol{\pi}_h(s) \right). \end{aligned} \tag{9}$$

Note that if $\Omega_{s,h}(P_h, -v, \mu) = \langle P_h, -v\mu^\top \rangle = -\mathbb{E}_{\boldsymbol{a} \sim \mu}\left[ [P_h v](s, \boldsymbol{a}) \right]$ for all $h \in [H]$, this regularized MG reduces to the standard MG.

Viewing the support function $\sigma_{\mathcal{P}_{s,h}}(\cdot)$ as the policy-value regularizer, Proposition 5.1 implies that the RNE of robust Markov games with transition uncertainty is equivalent to the NE of the regularized Markov game.

**Theorem 5.2.** *Consider a RMG $\mathfrak{G}$ with $s$-rectangular uncertainty set $\mathcal{U} = \mathcal{P} \times \{\boldsymbol{r}^\star\}$. Consider the policy-value regularized MG $\mathfrak{G}' = (\mathcal{S}, \{\mathcal{A}_i\}_{i \in [N]}, P^\star, \boldsymbol{r}^\star, H, \Omega)$, where the regularizer functions $\Omega := (\Omega_{s,h})_{s,h \in [H]}$ satisfies $\Omega_{s,h}(P_h^\star, -v, \mu) = \sigma_{\mathcal{P}_{s,h}}(-v\mu^\top), \forall v \in \mathbb{R}^{\mathcal{S}}, \forall \mu \in \Delta(\mathcal{A})$. Then, $\boldsymbol{\pi}$ is an MPRNE for $\mathfrak{G}$ if and only if $\boldsymbol{\pi}$ is an MPNE for $\mathfrak{G}'$.*

Depending on the uncertainty set $\mathcal{P}_{s,h}$, the support function can be further simplified, leading to efficient computation of regularized value functions. Here we give some examples of transition uncertainty sets that are commonly considered in the literature, including both $s$-rectangular and $(s, a)$-rectangular sets. We refer readers to Appendix F.3 for the proof and discussion of additional examples.

**Corollary 5.3.** *Consider a robust MG with uncertainty set $\mathcal{U} = \mathcal{P} \times \{\boldsymbol{r}^\star\}$, where $\mathcal{P} = \times_{(s,h) \in \mathcal{S} \times [H]} \mathcal{P}_{s,h}$ satisfies ball* *constraints:* $\mathcal{P}_{s,h} = \{P \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \|P - P^\star_{s,h}\|_{q^*\to p} \leq \beta_{s,h}\}$*. Then the equivalent policy-value regularized MG is associated with the convex regularizer function* $\Omega = \{\Omega_{s,h}\}_{s\in\mathcal{S},h\in[H]}$ *such that* $\forall s \in \mathcal{S}, v \in \mathbb{R}^{\mathcal{S}}, \mu \in \Delta(\mathcal{A})$,

$$\Omega_{s,h}(P^\star_h, -v, \mu) = \\ -\mathbb{E}_{\boldsymbol{a}\sim\mu}\left[[P^\star_h v](s,\boldsymbol{a})\right] + \beta_{s,h}\|-v\|_p \left\|\mu^\top\right\|_q .$$

We remark that for the policy-value regularized game in Corollary 5.3, the corresponding value function defined in (9) reduces to

$$\widehat{V}^{\boldsymbol{\pi}}_{i,h}(s, G^\star) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r^\star_{i,h}(s,\boldsymbol{a}) + [P^\star_h \widehat{V}^{\boldsymbol{\pi}}_{i,h+1}](s,\boldsymbol{a})\right] \\ - \beta_{s,h}\left\| - \widehat{V}^{\boldsymbol{\pi}}_{i,h+1}\right\|_p \|\boldsymbol{\pi}_h(s)\|_q .$$

Compared with standard MG, the value function involves an additional term that penalizes the $\ell_p$-norm of future rewards and $\ell_q$-norm of policy. Note that similar to the case with reward uncertainty, the size of the uncertainty set, i.e., the radius $\beta_{s,h}$, determines the penalty factor.

When the transition uncertainty set is $(s,a)$-rectangular, the regularizer functions for the equivalent policy-value regularized MG admits simpler forms.

**Corollary 5.4.** *Consider a robust MG with an uncertainty set* $\mathcal{U} = \mathcal{P} \times \{\boldsymbol{r}^\star\}$*, where* $\mathcal{P}$ *is* $(s,a)$*-rectangular of the form* $\mathcal{P} = \times_{(s,\boldsymbol{a},h)\in\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{P}_{s,\boldsymbol{a},h}$*, with* $\mathcal{P}_{s,\boldsymbol{a},h} \subset \Delta(\mathcal{S})$ *being compact and convex. Then the equivalent policy-value regularized MG is associated with convex regularizers* $\Omega = \{\Omega_{s,h}\}$ *such that:* $\forall s \in \mathcal{S}, v \in \mathbb{R}^{\mathcal{S}}, \mu \in \Delta(\mathcal{A})$,

$$\Omega_{s,h}(P^\star_h, -v, \mu) = \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\sigma_{\mathcal{P}_{s,\boldsymbol{a},h}}(-v)\right].$$

Various $(s,a)$-rectangular transition uncertainty sets have been considered for robust MDP (Shi et al., 2023; Iyengar, 2005). The uncertain transitions are typically of the form

$$\mathcal{P}_{s,\boldsymbol{a},h} = \left\{P \in \Delta(\mathcal{S}) : d(P, P^\star_{s,\boldsymbol{a},h}) \leq \beta_{s,\boldsymbol{a},h}\right\},$$

where $\beta_{s,\boldsymbol{a},h} > 0$ represents the level of uncertainty, and $d(\cdot,\cdot)$ is a distance metric between two probability distributions. Popular distance metrics include Total variation (TV) distance, KL distance, Chi-square distance, and Wasserstein distance. For each case, we can obtain equivalent policy-value regularizer functions. Here we consider the TV distance and defer the discussion of other distance metrics to Appendix F.3.3.

**Example.** Consider a TV uncertainty set given by $\mathcal{P} = \times_{(s,\boldsymbol{a},h)\in\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{P}^{\text{TV}}_{s,\boldsymbol{a},h}$ with $\mathcal{P}^{\text{TV}}_{s,\boldsymbol{a},h} = \{P \in \Delta(\mathcal{S}) : d_{\text{TV}}(P, P^\star_{s,\boldsymbol{a},h}) \leq \beta_{s,\boldsymbol{a},h}\}$. Here $d_{\text{TV}}(\eta,\eta') = \frac{1}{2}\|\eta-\eta'\|_1 = \frac{1}{2}\sum_s|\eta(s)-\eta'(s)|$ for any $\eta,\eta' \in \Delta(\mathcal{S})$. Then the equivalent policy-value regularized MG is associated with convex regularizers $\Omega = \{\Omega_{s,h}\}$ such that: $\forall s \in \mathcal{S}, v \in \mathbb{R}^{\mathcal{S}}, \mu \in \Delta(\mathcal{A})$,

$$\Omega_{s,h}(P^\star_h, -v, \mu) = \mathbb{E}_{\boldsymbol{a}\sim\mu}\Big[-[P^\star_h v](s,\boldsymbol{a}) + \frac{\beta_{s,\boldsymbol{a},h}}{2}\cdot \\ \min_{u\geq\mathbf{0}}\Big\{\max_{s'}\big(v(s')-u(s')\big) - \max_{s'}\big(v(s')-u(s')\big)\Big\}\Big].$$

In particular, the optimization in $\Omega_{s,h}$ is convex and can be computed in time $\mathcal{O}(|\mathcal{S}|\log|\mathcal{S}|)$ (Iyengar, 2005). Compared with standard MGs, the corresponding regularized value function (9) involves an additional penalty that depends on the policy and future rewards. Similar to the ball-constrained uncertainty set, the size of the uncertainty set, i.e., radius $\beta_{s,\boldsymbol{a}h}$ determines the penalty factor. We remark that in general, the regularizer function might not include the standard linear future value term $-[P^\star_h v](s,\boldsymbol{a})$, as shown in Appendix F.3.3 for various transition uncertainty sets.

**Computational Hardness.** Lastly, we note that even for TPSZ RMGs with $(s,a)$-rectangular transition uncertainty and $H = 2$, computing a MPRNE is PPAD-hard. Thus, dealing with transition uncertainty is generally difficult. The proof uses transition uncertainty to simulate the same hard reward uncertainty instances derived in Theorem 4.1.

**Theorem 5.5.** *Even restricted to the class of* $(s,a)$*-rectangular uncertainty sets, computing an MPRNE of a TPZS RMG with transition uncertainty is PPAD-hard even for* $S = H = 2$*.*

## 6. Conclusions

In this work, we study RMGs with $s$-rectangular uncertainty. We show that RMGs can be solved using regularized MG algorithms. This reduction yields a planning algorithm for computing an MPRNE of an RMG. We also show that for many commonly used regularizers, the set of MPNE of the regularized game is equal to the set of MPRNE of RMGs with well-behaved uncertainty sets. This gives proof that regularization methods do produce robust policies.

However, we show even for two-player robust matrix games with $(s,a)$-rectangularity, computing an MPRNE is PPAD-hard. This illustrates that the reward uncertainty case is already challenging compared to the single-agent setting. Despite this, we show whenever the support function of the uncertainty set decomposes into a sum of two parts, one corresponding to each player's policy, then our constructed regularized game is zero-sum. Consequently, we can compute an MPRNE for two-player zero-sum RMGs with efficient player-decomposable reward uncertainty in polynomial time.

## Impact Statement

This paper presents work whose goal is to advance the field of game theory. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## Acknowledgements

Q. Xie was supported in part by National Science Foundation Awards CNS-1955997 and EPCN-2339794. J. McMahan is supported in part by NSF grant 2023239.

## References

Aghassi, M. and Bertsimas, D. Robust game theory. *Math. Program.*, 107(1–2):231–273, jun 2006. ISSN 0025-5610. doi: 10.1007/s10107-005-0686-0. URL https://doi.org/10.1007/s10107-005-0686-0.

Badrinath, K. P. and Kalathil, D. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 511–520. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/badrinath21a.html.

Blanchet, J. and Murthy, K. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

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

Borwein, J. and Lewis, A. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. CMS Books in Mathematics. Springer New York, 2010. ISBN 9780387312569. URL https://books.google.com/books?id=eeL_Cdmowv8C.

Brekelmans, R., Genewein, T., Grau-Moya, J., Delétang, G., Kunesch, M., Legg, S., and Ortega, P. Your policy regularizer is secretly an adversary. *arXiv preprint arXiv:2203.12592*, 2022.

Cen, S., Wei, Y., and Chi, Y. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34: 27952–27964, 2021.

Chen, S., Liu, G., Zhou, Z., Zhang, K., and Wang, J. Robust multi-agent reinforcement learning method based on adversarial domain randomization for real-world dual-uav cooperation. *IEEE Transactions on Intelligent Vehicles*, pp. 1–13, 2023. doi: 10.1109/TIV.2023.3307134.

Cherukuri, A., Mallada, E., Low, S., and Cortés, J. The role of convexity in saddle-point dynamics: Lyapunov function and robustness. *IEEE Transactions on Automatic Control*, 63(8):2449–2464, 2017.

Derman, E., Geist, M., and Mannor, S. Twice regularized MDPs and the equivalence between robustness and regularization. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=x00mCNwbH8Q.

Derman, E., Men, Y., Geist, M., and Mannor, S. Twice regularized markov decision processes: The equivalence between robustness and regularization, 2023.

Epstein, L. and Schneider, M. Recursive multiple-priors. *Journal of Economic Theory*, 113(1):1–31, 2003. URL https://EconPapers.repec.org/RePEc:eee:jetheo:v:113:y:2003:i:1:p:1-31.

Eysenbach, B. and Levine, S. Maximum entropy rl (provably) solves some robust rl problems. In *International Conference on Learning Representations*, 2021.

Facchinei, F. and Pang, J.-S. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Goh, J., Bayati, M., Zenios, S. A., Singh, S., and Moore, D. Data uncertainty in markov chains: Application to cost-effectiveness analyses of medical innovations. *Operations Research*, 66(3):697–715, 2018. doi: 10.1287/opre.2017.1685. URL https://doi.org/10.1287/opre.2017.1685.

Goyal, V. and Grand-Clément, J. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023. doi: 10.1287/moor.2022.1259. URL https://doi.org/10.1287/moor.2022.1259.

Grill, J.-B., Darwiche Domingues, O., Menard, P., Munos, R., and Valko, M. Planning in entropy-regularized markov decision processes and games. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/

50982fb2f2cfa186d335310461dfa2be-Paper.pdf.

Hayashi, S., Yamashita, N., and Fukushima, M. Robust nash equilibria and second-order cone complementarity problems. *Journal of Nonlinear and Convex Analysis*, 6: 283–296, 2005.

Ho, C. P., Petrik, M., and Wiesemann, W. Fast Bellman updates for robust MDPs. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1979–1988. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ho18a.html.

Ho, C. P., Petrik, M., and Wiesemann, W. Partial policy iteration for l1-robust markov decision processes. *J. Mach. Learn. Res.*, 22(1), jan 2021. ISSN 1532-4435.

Husain, H., Ciosek, K., and Tomioka, R. Regularized policies are reward robust. In *International Conference on Artificial Intelligence and Statistics*, pp. 64–72. PMLR, 2021.

Iyengar, G. N. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/25151652.

Kakutani, S. A generalization of brouwer's fixed point theorem. *Duke Math. J.*, 8(1):457–459, 1941.

Kardeş, E., Ordóñez, F., and Hall, R. W. Discounted robust stochastic games and an application to queueing control. *Operations Research*, 59(2):365–382, 2011. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/23013175.

Kaufman, D. L. and Schaefer, A. J. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013. doi: 10.1287/ijoc.1120.0509. URL https://doi.org/10.1287/ijoc.1120.0509.

Kumar, N., Levy, K., Wang, K., and Mannor, S. Efficient policy iteration for robust markov decision processes via regularization, 2022.

Kumar, N., Derman, E., Geist, M., Levy, K., and Mannor, S. Policy gradient for rectangular robust markov decision processes, 2023.

Lee, K., Choi, S., and Oh, S. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018.

Lim, S. H. and Autef, A. Kernel-based reinforcement learning in robust Markov decision processes. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3973–3981. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lim19a.html.

Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z., and Zhou, Z. Distributionally robust $q$-learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13623–13643. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/liu22a.html.

Ma, S., Chen, Z., Zou, S., and Zhou, Y. Decentralized robust v-learning for solving markov games with model uncertainty. *Journal of Machine Learning Research*, 24(371): 1–40, 2023a. URL http://jmlr.org/papers/v24/23-0310.html.

Ma, X., Liang, Z., Blanchet, J., Liu, M., Xia, L., Zhang, J., Zhao, Q., and Zhou, Z. Distributionally robust offline reinforcement learning with linear function approximation, 2023b.

Mannor, S., Mebel, O., and Xu, H. Robust mdps with k-rectangular uncertainty. *Math. Oper. Res.*, 41(4): 1484–1509, nov 2016. ISSN 0364-765X.

Mertikopoulos, P. and Sandholm, W. H. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016.

Nilim, A. and El Ghaoui, L. Robustness in markov decision problems with uncertain transition matrices. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, pp. 839–846, Cambridge, MA, USA, 2003. MIT Press.

Panaganti, K. and Kalathil, D. Sample complexity of robust reinforcement learning with a generative model. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9582–9602. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/panaganti22a.html.

Perchet, V. Finding robust nash equilibria. In Kontorovich, A. and Neu, G. (eds.), *Proceedings of the*

*31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pp. 725–751. PMLR, 08 Feb–11 Feb 2020. URL https://proceedings.mlr.press/v117/perchet20a.html.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2817–2826. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/pinto17a.html.

Satia, J. K. and Lave, R. E. Markovian decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973. URL https://EconPapers.repec.org/RePEc:inm:oropre:v:21:y:1973:i:3:p:728-740.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shi, L. and Chi, Y. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity, 2023.

Shi, L., Li, G., Wei, Y., Chen, Y., Geist, M., and Chi, Y. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

Tamar, A., Mannor, S., and Xu, H. Scaling up robust mdps using function approximation. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 181–189, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/tamar14.html.

Wang, Q., Ho, C. P., and Petrik, M. Policy gradient in robust mdps with global convergence guarantee. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7193–7206. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/3a4496776767aaa99f9804d0905fe584-Paper.pdf.

Wang, Y. and Zou, S. Policy gradient method for robust reinforcement learning. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23484–23526. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wang22at.html.

Wiesemann, W., Kuhn, D., and Rustem, B. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013. doi: 10.1287/moor.1120.0566. URL https://doi.org/10.1287/moor.1120.0566.

Yang, W., Zhang, L., and Zhang, Z. Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics, 2022.

Zhan, W., Cen, S., Huang, B., Chen, Y., Lee, J. D., and Chi, Y. Policy mirror descent for regularized reinforcement learning: A generalized framework with linear convergence. *SIAM Journal on Optimization*, 33(2):1061–1091, 2023.

Zhang, F., Tan, V. Y. F., Wang, Z., and Yang, Z. Learning regularized monotone graphon mean-field games, 2023.

Zhang, K., Kakade, S., Basar, T., and Yang, L. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1166–1178. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/0cc6ee01c82fc49c28706e0918f57e2d-Paper.pdf.

Zhang, K., SUN, T., Tao, Y., Genc, S., Mallya, S., and Basar, T. Robust multi-agent reinforcement learning with model uncertainty. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10571–10583. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/774412967f19ea61d448977ad9749078-Paper.pdf.

## A. Useful Technical Results

In this section, we state some existing results that are useful in our analysis.

**Theorem A.1** (Fenchel-Rockafellar duality (Borwein & Lewis, 2010, Theorem 3.3.5)). *Consider the problems:*

$$(P) \qquad \min_x f(x) + g(Ax)$$
$$(D) \qquad \max_y -f^*(-A^*y) - g^*(y)$$

*where $f : \mathcal{X} \to \overline{\mathbb{R}}$ and $g : \mathcal{Y} \to \overline{\mathbb{R}}$ are proper, closed convex functions, and $A : \mathcal{X} \to \mathcal{Y}$ is a linear map. If the regularity condition $0 \in \operatorname{core}(\operatorname{dom}(g) - A(\operatorname{dom}(f)))$ holds*[2]*, then the primal $(P)$ and dual $(D)$ optimal values are equal.*

**Lemma A.2** (Induced norm of rank-1 matrices). *For any vectors $u, v$, we have $\left\|uv^\top\right\|_{p\to q} = \|u\|_q \|v\|_{p^*}$.*

*Proof.* By definition of matrix-induced norm, we have

$$\begin{aligned}
\left\|uv^\top\right\|_{p\to q} &= \sup_{x:x\neq 0} \frac{\left\|uv^\top x\right\|_q}{\|x\|_p} \\
&= \sup_{x:x\neq 0} \frac{\left|v^\top x\right| \cdot \|u\|_q}{\|x\|_p} && \text{positive homogeneity of norms} \\
&= \|u\|_q \cdot \sup_{x:x\neq 0} \frac{\left|v^\top x\right|}{\|x\|_p} \\
&= \|u\|_q \cdot \|v\|_{p^*} . && \text{definition of dual norm of } \|\cdot\|_p .
\end{aligned}$$

□

**Lemma A.3** (Dual norm of induced matrix norm). *For any matrix $R$, we have*

$$\sup_{R:\|R\|_{p\to q}\leq 1} \left\langle R, xy^\top \right\rangle = \|x\|_{q^*} \|y\|_p = \left\|xy^\top\right\|_{p^*\to q} .$$

*Proof.* By definition of dual norms, there exists a vector $u$ with $\|u\|_q = 1$ such that $\langle u, x\rangle = \|x\|_{q^*}$. Similarly, there exists $v$ with $\|v\|_{p^*} = 1$ such that $\langle v, y\rangle = \|y\|_p$. Then the matrix $R_0 := uv^\top$ satisfies

$$\left\langle R_0, xy^\top \right\rangle = \operatorname{Tr}\left(R_0^\top xy^\top\right) = \operatorname{Tr}\left(y^\top (vu^\top) x\right) = \langle u, x\rangle \langle v, y\rangle = \|x\|_{q^*} \|y\|_p$$

and, by Lemma A.2,

$$\|R_0\|_{p\to q} = \left\|uv^\top\right\|_{p\to q} = \|u\|_q \|v\|_{p^*} = 1.$$

So

$$\sup_{\|R\|_{p\to q}\leq 1} \left\langle R, xy^\top \right\rangle \geq \left\langle R_0, xy^\top \right\rangle = \|x\|_{q^*} \|y\|_p .$$

On the other hand, for any $R$ with $\|R\|_{p\to q} \leq 1$, we have

$$\left\langle R, xy^\top \right\rangle = x^\top R y \leq \|x\|_{q^*} \|Ry\|_q \leq \|x\|_{q^*} \|y\|_p ,$$

hence $\sup_{\|R\|_{p\to q}\leq 1} \left\langle R, xy^\top \right\rangle \leq \|x\|_{q^*} \|y\|_p$. This proves the first equality in the lemma.

The second equality follows from Lemma A.2. □

[2] Given $C \subset \mathbb{R}^d$, we say that $x \in \operatorname{core}(C)$ if for all $z \in \mathbb{R}^d$, there exists a small enough $t \in \mathbb{R}$ such that $x + tz \in C$ (Borwein & Lewis, 2010).

# B. Properties of Robust Markov Games

Recall that given a joint policy $\boldsymbol{\pi} \in \Pi$ for a MG $G = (P, \boldsymbol{r})$, the value function and the state-action function for each player $i$ is defined as: $\forall s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}, h \in [H]$,

$$\overline{V}^{\boldsymbol{\pi}}_{i,h}(s, G) := \mathbb{E}^{\boldsymbol{\pi}}_{P}\Big[\sum_{t=h}^{H} r_{i,t}(s_t, \boldsymbol{a}_t) \mid s_h = s\Big], \tag{10}$$

$$\overline{Q}^{\boldsymbol{\pi}}_{i,h}(s, \boldsymbol{a}, G) := \mathbb{E}^{\boldsymbol{\pi}}_{P}\Big[\sum_{t=h}^{H} r_{i,t}(s_t, \boldsymbol{a}_t) \mid s_h = s, \boldsymbol{a}_h = \boldsymbol{a}\Big], \tag{11}$$

where the expectation $\mathbb{E}^{\boldsymbol{\pi}}_{P}[\cdot]$ is taken with respect to the trajectory $\{s_h, a_h, \boldsymbol{r}_h\}_{h\in[H]}$ induced by the transition kernel $P$ and the joint policy $\boldsymbol{\pi}$, i.e., $\boldsymbol{a}_h \sim \boldsymbol{\pi}_h(s_h)$ and $s_{h+1} \sim P_h(\cdot|s_h, \boldsymbol{a}_h)$. It is convenient to set $\overline{V}^{\boldsymbol{\pi}}_{i,H+1}(s, G) \equiv \overline{Q}^{\boldsymbol{\pi}}_{i,H+1}(s, \boldsymbol{a}, G) \equiv 0$ for the terminal reward.

### B.1. Robust Markov Game Bellman Equation

Similar to the Bellman equation for standard MG, we have the following robust Bellman equation.

**Proposition B.1.** *(Robust Bellman Equation). Under Assumption 2.2, for any joint policy $\boldsymbol{\pi} = \{\boldsymbol{\pi}_h\}_{h\in[H]} \in \Pi$, the following Bellman equations hold:*

$$V^{\boldsymbol{\pi}}_{i,h}(s) = Q^{\boldsymbol{\pi}}_{i,h}\left(s, \boldsymbol{\pi}_h(\cdot \mid s)\right), \tag{12}$$

$$Q^{\boldsymbol{\pi}}_{i,h}(s, \mu) = \inf_{\boldsymbol{r}_h \in \mathcal{U}^r_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,h}(s, \boldsymbol{a})\right] + \inf_{P_h \in \mathcal{P}_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[[P_h V^{\boldsymbol{\pi}}_{i,h+1}](s, \boldsymbol{a})\right], \tag{13}$$

*where $\mathcal{U}^r_{s,h} := \times_{i\in[N]} \mathcal{U}^r_{i,s,h}$.*

*Proof.* Blanchet et al. (2023) established the robust Bellman Equation for robust Markov game with $\mathcal{S} \times \mathcal{A}$-rectangular transition uncertainty set. We extend the result to Markov games with general $\mathcal{S}$-uncertainty set, including both reward function uncertainty and transition uncertainty. Throught out the proof, for each game model $G = (P, \boldsymbol{r})$, we will use $G^h := \left(\{P_t\}_{t=h}^{H}, \{\boldsymbol{r}_t\}_{t=h}^{H}\right)$ to denote the model parameters from step $h$ to the terminal step $H$. With this notation, we note that for a standard Markov game with model $G$, the value functions satisfy

$$\overline{V}^{\boldsymbol{\pi}}_{i,h}(s, G) = \overline{V}^{\boldsymbol{\pi}}_{i,h}(s, G^h),$$
$$\overline{Q}^{\boldsymbol{\pi}}_{i,h}(s, \boldsymbol{a}, G) = \overline{Q}^{\boldsymbol{\pi}}_{i,h}(s, \boldsymbol{a}, G^h).$$

To facilitate the proof, we introduce the following shorthands to denote the uncertainty sets from step $h$ to the terminal step $H$:

$$\mathcal{P}^h := \times_{(s,t)\in\mathcal{S}\times\{t,t+1,\ldots,H\}} \mathcal{P}_{s,t},$$
$$\mathcal{U}^{r,h} := \times_{(i,s,t)\in[N]\times\mathcal{S}\times\{t,t+1,\ldots,H\}} \mathcal{U}^r_{i,s,t},$$
$$\mathcal{U}^h := \mathcal{P}^h \times \mathcal{U}^{r,h}.$$

Let $\mathcal{P}_h := \times_{s\in\mathcal{S}} \mathcal{P}_{s,h}$ and $\mathcal{U}^r_h := \times_{(i,s)\in[N]\times\mathcal{S}} \mathcal{U}^r_{i,s,h}$.

We will prove the following stronger results via induction from step $h = H$ to 1: given any $\boldsymbol{\pi} \in \Pi$, for each player $i \in [N]$, there exists a game model $\widehat{G} = (\widehat{P}, \widehat{\boldsymbol{r}})$ such that: (1) Robust Bellman equations (12)-(13) hold; (2) the robust value functions satisfy

$$V^{\boldsymbol{\pi}}_{i,h}(s) = \overline{V}^{\boldsymbol{\pi}}_{i,h}(s, \widehat{G}^h), \qquad , \forall s \in \mathcal{S}, \tag{14}$$

$$Q^{\boldsymbol{\pi}}_{i,h}(s, \mu) = \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\overline{Q}^{\boldsymbol{\pi}}_{i,h}(s, \boldsymbol{a}, \widehat{G}^h)\right], \qquad \forall s \in \mathcal{S}, \mu \in \Delta(\mathcal{A}). \tag{15}$$

1. (Base case): For $h = H$, by the definition of robust state-action value function in (2), it follows that (13) holds. We also have

$$Q^{\boldsymbol{\pi}}_{i,H}(s, \mu) = \inf_{G\in\mathcal{U}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\overline{Q}^{\boldsymbol{\pi}}_{i,H}(s, \boldsymbol{a}, G)\right] = \inf_{\boldsymbol{r}\in\mathcal{U}^r} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,H}(s, \boldsymbol{a})\right] = \inf_{\boldsymbol{r}_H(s,\cdot)\in\mathcal{U}^r_{s,H}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,H}(s, \boldsymbol{a})\right].$$

By Assumption 2.2, one can thus find a single reward function $\widehat{\boldsymbol{r}}_H \in \mathcal{U}^r_{s,H} = \times_i \mathcal{U}^r_{i,s,H}$ such that for each $s \in \mathcal{S}$,

$$\widehat{\boldsymbol{r}}_H(s,\cdot) \in \underset{\boldsymbol{r}_H(s,\cdot)\in\mathcal{U}^r_{s,H}}{\arg\inf} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,H}(s,\boldsymbol{a})\right].$$

Therefore,

$$Q^{\boldsymbol{\pi}}_{i,H}(s,\mu) = \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\widehat{r}_{i,H}(s,\boldsymbol{a})\right] = \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\overline{Q}^{\boldsymbol{\pi}}_{i,H}(s,\boldsymbol{a},\widehat{G}^H)\right]. \tag{16}$$

By definition of robust value function in (1), we have

$$V^{\boldsymbol{\pi}}_{i,H}(s) = \inf_{G\in\mathcal{U}} \overline{V}^{\boldsymbol{\pi}}_{i,H}(s,G) \overset{(I)}{=} \inf_{G\in\mathcal{U}} \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_H(s)}\left[\overline{Q}^{\boldsymbol{\pi}}_{i,H}(s,\boldsymbol{a},G)\right] \overset{(II)}{=} Q^{\boldsymbol{\pi}}_{i,H}(s,\boldsymbol{\pi}_H(s)) \overset{(III)}{=} \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_H(s)}\left[\overline{Q}^{\boldsymbol{\pi}}_{i,H}(s,\boldsymbol{a},\widehat{G}^H)\right]$$
$$\overset{(VI)}{=} \overline{V}^{\boldsymbol{\pi}}_{i,H}(s,\widehat{G}^H),$$

where equalities $(I)$ and $(VI)$ follow from the Bellman equations (10)-(11) of standard Markov game, equality $(II)$ follows from the definition of robust Q-function in (2), and $(III)$ holds due to (16). We complete the proof of the base case.

2. (Induction step): Now suppose that (12)-(13) and (14)-(15) hold for all $t > h$. Thus there exists $\widehat{G}^{h+1} := \left(\{\widehat{P}_t\}_{t=h+1}^H, \{\widehat{\boldsymbol{r}}_t\}_{t=h+1}^H\right)$ such that

$$V^{\boldsymbol{\pi}}_{i,h+1}(s) = \overline{V}^{\boldsymbol{\pi}}_{i,h+1}(s,\widehat{G}^{h+1}), \qquad ,\forall s\in\mathcal{S}. \tag{17}$$

By the definition of robust state-action value function in (2), we have: $\forall s\in\mathcal{S}, \mu\in\Delta(\mathcal{A})$,

$$\begin{aligned}
Q^{\boldsymbol{\pi}}_{i,h}(s,\mu) &:= \inf_{G^h\in\mathcal{U}^h} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\overline{Q}^{\boldsymbol{\pi}}_{i,h}(s,\boldsymbol{a},G^h)\right] \\
&= \inf_{\{P_t\}_{t=h}^H\in\mathcal{P}^h, \{\boldsymbol{r}_t\}_{t=h}^H\in\mathcal{U}^{r,h}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,h}(s,\boldsymbol{a}) + \mathbb{E}_{s'\sim P_h(\cdot|s,\boldsymbol{a})}\left[\overline{V}^{\boldsymbol{\pi}}_{i,h+1}(s',G^{h+1})\right]\right] \\
&\overset{(I)}{=} \inf_{\boldsymbol{r}_h(s,\cdot)\in\mathcal{U}^r_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,h}(s,\boldsymbol{a})\right] + \inf_{\{P_t\}_{t=h}^H\in\mathcal{P}^h, \{\boldsymbol{r}_t\}_{t=h+1}^H\in\mathcal{U}^{r,h+1}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\mathbb{E}_{s'\sim P_h(\cdot|s,\boldsymbol{a})}\left[\overline{V}^{\boldsymbol{\pi}}_{i,h+1}(s',G^{h+1})\right]\right] \\
&\le \inf_{\boldsymbol{r}_h(s,\cdot)\in\mathcal{U}^r_{s,h}} \underbrace{\mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,h}(s,\boldsymbol{a})\right]}_{T_1} + \inf_{P_h\in\mathcal{P}_h} \underbrace{\mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\mathbb{E}_{s'\sim P_h(\cdot|s,\boldsymbol{a})}\left[\overline{V}^{\boldsymbol{\pi}}_{i,h+1}(s',\widehat{G}^{h+1})\right]\right]}_{T_2},
\end{aligned} \tag{18}$$

where the equality $(I)$ follows from the rectangular uncertainty set assumption and the fact that the first only depends on the reward function at state $s$ and step $h$. We thus can find a single reward function $\widehat{\boldsymbol{r}}_h \in \mathcal{U}^r_h$ that attains the minimum value of the term $T_1$ for each state $s\in\mathcal{S}$; we can also find a single transition kernel $\widehat{P}_h \in \mathcal{P}_h$ that it attains the minimum value of $T_2$.

On the other hand, by (17) and the definition of robust value function in (1), we have

$$\begin{aligned}
Q^{\boldsymbol{\pi}}_{i,h}(s,\mu) &\le \inf_{\boldsymbol{r}_h(s,\cdot)\in\mathcal{U}^r_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,h}(s,\boldsymbol{a})\right] + \inf_{P_h\in\mathcal{P}_h} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\mathbb{E}_{s'\sim P_h(\cdot|s,\boldsymbol{a})}\left[V^{\boldsymbol{\pi}}_{i,h+1}(s)\right]\right] \qquad (19)\\
&= \inf_{\boldsymbol{r}_h(s,\cdot)\in\mathcal{U}^r_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,h}(s,\boldsymbol{a})\right] \\
&\quad + \inf_{P_h\in\mathcal{P}_h} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\mathbb{E}_{s'\sim P_h(\cdot|s,\boldsymbol{a})}\left[\inf_{\{P_t\}_{t=h+1}^H\in\mathcal{P}^{h+1}, \{\boldsymbol{r}_t\}_{t=h+1}^H\in\mathcal{U}^{r,h+1}} \overline{V}^{\boldsymbol{\pi}}_{i,h+1}(s',G^{h+1})\right]\right] \\
&= \inf_{\boldsymbol{r}_h(s,\cdot)\in\mathcal{U}^r_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,h}(s,\boldsymbol{a})\right] + \inf_{\{P_t\}_{t=h}^H\in\mathcal{P}^h, \{\boldsymbol{r}_t\}_{t=h+1}^H\in\mathcal{U}^{r,h+1}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\mathbb{E}_{s'\sim P_h(\cdot|s,\boldsymbol{a})}\left[\overline{V}^{\boldsymbol{\pi}}_{i,h+1}(s',G^{h+1})\right]\right] \\
&= Q^{\boldsymbol{\pi}}_{i,h}(s,\mu),
\end{aligned}$$

where the last equality follows from the definition of robust state-action value function in (2). Therefore, all the inequality above are equalities. In particular, equation (19) proves the robust Bellman equation (13) for step $h$. In addition, from (18), we have

$$Q^{\boldsymbol{\pi}}_{i,h}(s,\mu) = \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\widehat{r}_{i,h}(s,\boldsymbol{a})\right] + \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\mathbb{E}_{s'\sim\widehat{P}_h(\cdot|s,\boldsymbol{a})}\left[\overline{V}^{\boldsymbol{\pi}}_{i,h+1}(s',\widehat{G}^{h+1})\right]\right] = \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\overline{Q}^{\boldsymbol{\pi}}_{i,h}(s,\boldsymbol{a},\widehat{G}^h)\right], \tag{20}$$

which proves (15) for step $h$.

Next we will show that $V_{i,h}^{\boldsymbol{\pi}}(s)$ satisfies (12) and (14). By definition of robust value function in (1), we have

$$\begin{aligned}
V_{i,h}^{\boldsymbol{\pi}}(s) &= \inf_{\{P_t\}_{t=h}^H \in \mathcal{P}^h, \{\boldsymbol{r}_t\}_{t=h}^H \in \mathcal{U}^{r,h}} \mathbb{E}_P^{\boldsymbol{\pi}}\left[\sum_{t=h}^H r_{i,h}(s_t, \boldsymbol{a}_t) \mid s_h = s\right] \\
&= \inf_{\{P_t\}_{t=h}^H \in \mathcal{P}^h, \{\boldsymbol{r}_t\}_{t=h}^H \in \mathcal{U}^{r,h}} \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[\mathbb{E}_P^{\boldsymbol{\pi}}\left[\sum_{t=h}^H r_{i,h}(s_t, \boldsymbol{a}_t) \mid s_h = s, \boldsymbol{a}_h = \boldsymbol{a}\right]\right] \\
&\overset{(I)}{=} \inf_{\{P_t\}_{t=h}^H \in \mathcal{P}^h, \{\boldsymbol{r}_t\}_{t=h}^H \in \mathcal{U}^{r,h}} \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[\overline{Q}_{i,h}^{\boldsymbol{\pi}}(s, \boldsymbol{a}, G^h)\right] \\
&\leq \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[\overline{Q}_{i,h}^{\boldsymbol{\pi}}(s, \boldsymbol{a}, \widehat{G}^h)\right] \\
&\overset{(II)}{=} Q_{i,h}^{\boldsymbol{\pi}}(s, \boldsymbol{\pi}_h(s)) \\
&\overset{(III)}{=} \inf_{G^h \in \mathcal{U}^h} \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[\overline{Q}_{i,h}^{\boldsymbol{\pi}}(s, \boldsymbol{a}, G^h)\right] \\
&\overset{(VI)}{=} \inf_{G^h \in \mathcal{U}^h} \overline{V}_{i,h}^{\boldsymbol{\pi}}(s, G^h) \\
&\overset{(V)}{=} V_{i,h}^{\boldsymbol{\pi}}(s),
\end{aligned}$$

where $(I)$ follows from the definition of state-action value function definition (11), $(II)$ follows from (20), $(III)$ holds due to the definition of robust state-action value function definition (2), $(VI)$ is true from Bellman equation of Markov game, and $(V)$ holds by definition in (1). Therefore, the inequality above is equality. Note that $(II)$ proves the Bellman equation (12). In addition, we have

$$V_{i,h}^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[\overline{Q}_{i,h}^{\boldsymbol{\pi}}(s, \boldsymbol{a}, \widehat{G}^h)\right] = \overline{V}_{i,h}^{\boldsymbol{\pi}}(s, \widehat{G}^h).$$

We complete the proof of step $h$.

This finishes the proof of Proposition B.1. □

## C. Proof of Existence of Robust NE

In this section, we provide the proof of Theorem 2.1 on the existence of robust Nash equilibrium. We first consider matrix games in Section C.1 and then proceed to Markov games in Section C.2.

### C.1. Matrix Games with Reward Uncertainty

We first consider matrix games with reward uncertainty. In particular, the reward uncertainty set is of the form $\mathcal{U} = \boldsymbol{r}^\star + \mathcal{R}$ with $\mathcal{R} = \times_{i\in[N]}\mathcal{R}_i$, where $\mathcal{R}_i \subset \mathbb{R}^{\mathcal{A}}$ contains the possible reward functions for player $i$. Here we allow the reward uncertainty sets to potentially depend on the players' policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ with $\pi_i \in \Delta(\mathcal{A}_i)$, denoted by $\mathcal{U}(\boldsymbol{\pi}) = \boldsymbol{r}^\star + \mathcal{R}(\boldsymbol{\pi})$ with $\mathcal{R}(\boldsymbol{\pi}) = \times_{i\in[N]}\mathcal{R}_i(\boldsymbol{\pi})$. The existence of RNE has been established for the setting where the uncertainty set is fixed and policy-independent (Zhang et al., 2020b; Kardeş et al., 2011; Aghassi & Bertsimas, 2006). Here we extend the result to more general uncertainty sets.

Our proof uses Kakutani's Fixed Point Theorem (Kakutani, 1941). We first state a relevant definition, followed by Kakutani's theorem below. For a mapping from a closed, bounded, convex set $E$ in a Euclidean space into the family of all closed, covex subsets of $E$, upper semicontinuity is defined as follows.

**Definition C.1.** A point-to-set mapping $\phi : E \to 2^E$ is called upper semicontinuous (u.s.c.) if

$$\begin{aligned}
y_n \in \phi(x_n), &\qquad n = 1, 2, 3, \ldots \\
\lim_{n\to\infty} x_n = x&, \\
\lim_{n\to\infty} y_n = y&,
\end{aligned}$$

imply that $y \in \phi(x)$.

**Theorem C.1** (Kakutani's Fixed Point Theorem (Kakutani, 1941))**.** *If $E$ is a closed, bounded, and convex set in a Euclidean space, and $\phi$ is an upper semicontinuous point-to-set mapping of $E$ into the family of closed, convex subsets of $E$, then $\exists x \in E$ s.t. $x \in \phi(x)$.*

### C.1.1. Properties of Robust Value Functions

To apply Kakutani's Fixed Point Theorem, we first establish some properties of the worst-case expected payoff functions, i.e., robust value function, defined as

$$F_i(\boldsymbol{\pi}) \triangleq \inf_{\boldsymbol{r}\in\mathcal{U}(\boldsymbol{\pi})} \pi_i^\top r_i \boldsymbol{\pi}_{-i}.$$

Here we view $r_i \in \mathcal{U}(\boldsymbol{\pi})$ as a matrix in $\mathbb{R}^{\mathcal{A}_i \times \mathcal{A}_{-i}}$, $\pi_i \in \mathbb{R}^{\mathcal{A}_i}$ as a column vector in and $\boldsymbol{\pi}_{-i} \in \mathbb{R}^{\mathcal{A}_{-i}}$. We begin by deriving an equivalent equation for the robust value function.

**Proposition C.2.** *For each player $i \in [N]$, given any product joint policy $\boldsymbol{\pi} \in \Pi$ of all players, the robust value for player $i$ satisfies*

$$F_i(\boldsymbol{\pi}) = \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i(\boldsymbol{\pi})}(-\pi_i \boldsymbol{\pi}_{-i}^\top), \tag{21}$$

*where $\sigma_{\mathcal{R}_i(\boldsymbol{\pi})}(\cdot)$ is the support function of the reward uncertainty set $\mathcal{R}_i(\boldsymbol{\pi})$.*

*Proof.* Recall that the characteristic function $\delta_{\mathcal{R}_i(\boldsymbol{\pi})} : \mathbb{R}^{\mathcal{A}} \to \{0, \infty\}$ over a set $\mathcal{R}_i(\boldsymbol{\pi}) \subseteq \mathbb{R}^{\mathcal{A}}$ is defined as $\delta_{\mathcal{R}_i(\boldsymbol{\pi})}(x) = 0$ if $x \in \mathcal{R}_i(\boldsymbol{\pi})$ and $+\infty$ otherwise. The Legendre-Fenchel transform of $\delta_{\mathcal{R}_i(\boldsymbol{\pi})}$, i.e., the support function $\sigma_{\mathcal{R}_i(\boldsymbol{\pi})} : \mathbb{R}^{\mathcal{A}} \to (-\infty, +\infty]$, is defined as

$$\sigma_{\mathcal{R}_i(\boldsymbol{\pi})}(y) := \sup_{x\in\mathcal{R}_i(\boldsymbol{\pi})} \langle x, y\rangle = \sup_{x\in\mathbb{R}^{\mathcal{A}}} \left\{\langle y, x\rangle - \delta_{\mathcal{R}_i(\boldsymbol{\pi})}(x)\right\}. \tag{22}$$

By the form of the uncertainty set $\mathcal{U}(\boldsymbol{\pi}) = \boldsymbol{r}^\star + \times\mathcal{R}_i(\boldsymbol{\pi})$, given any $\boldsymbol{\pi} \in \Pi$, we have

$$\begin{aligned} F_i(\boldsymbol{\pi}) = \inf_{\boldsymbol{r}\in\mathcal{U}(\boldsymbol{\pi})} \pi_i^\top r_i \boldsymbol{\pi}_{-i} &= \inf_{x\in\mathcal{R}_i(\boldsymbol{\pi})} \pi_i^\top (r_i^\star + x)\boldsymbol{\pi}_{-i} \\ &= \inf_{x\in\mathcal{R}_i(\boldsymbol{\pi})} \pi_i^\top x \boldsymbol{\pi}_{-i} + \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} \\ &= \inf_{x\in\mathbb{R}^{\mathcal{A}}} \{\pi_i^\top x\boldsymbol{\pi}_{-i} + \delta_{\mathcal{R}_i(\boldsymbol{\pi})}(x)\} + \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i}. \end{aligned}$$

We now proceed to apply Fenchel-Rockafellar duality theorem (Theorem A.1) to the minimization term. Fix player $i$'s policy $\pi_i$ and all other players' policy $\boldsymbol{\pi}_{-i}$. We define the function $f : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$ as $f(x) = \pi_i^\top x \boldsymbol{\pi}_{-i}$ for each $x \in \mathbb{R}^{\mathcal{A}}$. Consider the identity mapping $\mathrm{Id}_{\mathcal{A}} : \mathbb{R}^{\mathcal{A}} \to \mathbb{R}^{\mathcal{A}}$. Then $\mathrm{dom}(f) = \mathbb{R}^{\mathcal{A}}$, $\mathrm{dom}(\delta_{\mathcal{R}_i(\boldsymbol{\pi})}) = \mathcal{R}_i(\boldsymbol{\pi})$, and thus $\mathrm{core}(\mathrm{dom}(\delta_{\mathcal{R}_i(\boldsymbol{\pi})}) - \mathrm{Id}_{\mathcal{A}}(\mathrm{dom}(f))) = \mathrm{core}(\mathcal{R}_i(\boldsymbol{\pi}) - \mathbb{R}^{\mathcal{A}}) = \mathrm{core}(\mathbb{R}^{\mathcal{A}}) = \mathbb{R}^{\mathcal{A}}$. Note that $0 \in \mathbb{R}^{\mathcal{A}}$. We now can apply Fenchel-Rockafellar duality, noting that $(\mathrm{Id}_{\mathcal{A}})^* = \mathrm{Id}_{\mathcal{A}}$ and $(\delta_{\mathcal{R}_i(\boldsymbol{\pi})})^*(y) = \sigma_{\mathcal{R}_i(\boldsymbol{\pi})}(y)$:

$$\inf_{x\in\mathbb{R}^{\mathcal{A}}} \{f(x) + \delta_{\mathcal{R}_i(\boldsymbol{\pi})}(x)\} = -\inf_{y\in\mathbb{R}^{\mathcal{A}}} \{f^*(-y) + \delta^*_{\mathcal{R}_i(\boldsymbol{\pi})}(y)\} = -\inf_{y\in\mathbb{R}^{\mathcal{A}}} \{f^*(-y) + \sigma_{\mathcal{R}_i(\boldsymbol{\pi})}(y)\}.$$

We have

$$\begin{aligned} f^*(-y) &= \sup_{x\in\mathbb{R}^{\mathcal{A}}} \{\langle x, -y\rangle - \pi_i^\top x \boldsymbol{\pi}_{-i}\} \\ &= \sup_{x\in\mathbb{R}^{\mathcal{A}}} \sum_{a_i\in\mathcal{A}_i} \sum_{\boldsymbol{a}_{-i}\in\mathcal{A}_{-i}} x(a_i, \boldsymbol{a}_{-i}) \left[-y(a_i, \boldsymbol{a}_{-i}) - \pi_i(a_i)\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i})\right] \\ &= \begin{cases} 0 & \text{if } -y(a_i, \boldsymbol{a}_{-i}) - \pi_i(a_i)\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}) = 0 \ \forall a_i \in \mathcal{A}_i, \boldsymbol{a}_{-i} \in \mathcal{A}_{-i} \\ +\infty & \text{otherwise} \end{cases} \end{aligned}$$

Note that $y \in \mathbb{R}^{\mathcal{A}}$ such that $y(a_i, \boldsymbol{a}_{-i}) = -\pi_i(a_i)\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i})$ is exactly the negative outer product of $\pi_i$ and $\boldsymbol{\pi}_{-i}$, i.e., $y = -\pi_i \boldsymbol{\pi}_{-i}^\top$. Therefore, we have

$$F_i(\boldsymbol{\pi}) = -\inf_{y\in\mathbb{R}^{\mathcal{A}}} \{f^*(-y) + \sigma_{\mathcal{R}_i(\boldsymbol{\pi})}(y)\} + \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} = \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i(\boldsymbol{\pi})}(-\pi_i \boldsymbol{\pi}_{-i}^\top).$$

□

**Lemma C.3.** *Under Assumption 2.1, for each agent $i \in [N]$, $F_i(\boldsymbol{\pi})$ is continuous on $\Pi$.*

*Proof.* Here we focus on the case where the uncertainty set $\mathcal{U}(\boldsymbol{\pi})$ is well-defined for each $\boldsymbol{\pi} \in \text{int}(\Pi)$. The proof for the setting where $\mathcal{U}(\boldsymbol{\pi})$ is well-defined for all $\boldsymbol{\pi} \in \Pi$ follows similarly. Recall that for an arbitrary $\boldsymbol{\pi} \in \text{int}(\Pi)$, we have $\sigma_{\mathcal{R}_i(\boldsymbol{\pi})}(-\pi_i \boldsymbol{\pi}_{-i}^\top) := \sup_{x \in \mathcal{R}_i(\boldsymbol{\pi})} \langle x, -\pi_i \boldsymbol{\pi}_{-i}^\top \rangle = \sup_{x \in \mathbb{R}^{\mathcal{A}}} \left\{ \langle -\pi_i \boldsymbol{\pi}_{-i}^\top, x \rangle - \delta_{\mathcal{R}_i(\boldsymbol{\pi})}(x) \right\}$. For each $x \in \mathcal{R}_i(\boldsymbol{\pi})$, by Assumption 2.1, we have

$$\langle x, -\pi_i \boldsymbol{\pi}_{-i}^\top \rangle = -\mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}} [x(\boldsymbol{a})] \le -L_r < +\infty.$$

Therefore, the sup attains its maximum on the closed set $\mathcal{R}_i(\boldsymbol{\pi})$. We denote the maximizer by $x^*(\boldsymbol{\pi}) \in \mathcal{R}_i(\boldsymbol{\pi})$.

Let us consider $\epsilon > 0$. Consider $\delta(\epsilon, \boldsymbol{\pi})$ given by

$$\bar{\delta}(\epsilon, \boldsymbol{\pi}) = \min \left\{ \frac{1}{2} \inf_{z \in \text{Bd}(\Pi)} \|\boldsymbol{\pi} - z\|_\infty , \delta\left(\frac{\epsilon}{3}\right), \frac{\min\{\epsilon, 1\}}{6|\mathcal{A}| \cdot \max\{r_{\max}, M\}} \right\},$$

where $M := \max_{x \in \mathcal{R}_i(\boldsymbol{\pi})} \|x\|_\infty$. For each $\boldsymbol{\pi}' \in \Pi$ such that $\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_\infty < \bar{\delta}(\epsilon, \boldsymbol{\pi})$, we have $\boldsymbol{\pi}' \in \text{int}(\Pi)$. By Assumption 2.1, $D\big(\mathcal{R}_i(\boldsymbol{\pi}), \mathcal{R}_i(\boldsymbol{\pi}')\big) < \frac{\epsilon}{3}$. Thus there exist $\hat{x}' \in \mathcal{R}_i(\boldsymbol{\pi}')$ and $\hat{x} \in \mathcal{R}_i(\boldsymbol{\pi})$ such that $\|\hat{x} - x^*(\boldsymbol{\pi}')\|_\infty < \frac{\epsilon}{3}$ and $\|\hat{x}' - x^*(\boldsymbol{\pi})\|_\infty < \frac{\epsilon}{3}$.

By Proposition C.2, we have

$$\begin{aligned}
F_i(\boldsymbol{\pi}') - F_i(\boldsymbol{\pi}) &= (\pi_i')^\top r_i^\star \boldsymbol{\pi}'_{-i} - \sup_{x \in \mathcal{R}_i(\boldsymbol{\pi}')} \langle x, -\pi_i' (\boldsymbol{\pi}'_{-i})^\top \rangle - \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} + \sup_{x \in \mathcal{R}_i(\boldsymbol{\pi})} \langle x, -\pi_i \boldsymbol{\pi}_{-i}^\top \rangle \\
&\le \langle r^\star, \pi_i' (\boldsymbol{\pi}'_{-i})^\top - \pi_i \boldsymbol{\pi}_{-i}^\top \rangle - \langle \hat{x}', -\pi_i' (\boldsymbol{\pi}'_{-i})^\top \rangle + \langle x^*(\boldsymbol{\pi}), -\pi_i \boldsymbol{\pi}_{-i}^\top \rangle \\
&= \langle r^\star, \pi_i' (\boldsymbol{\pi}'_{-i})^\top - \pi_i \boldsymbol{\pi}_{-i}^\top \rangle - \langle \hat{x}' - x^*(\boldsymbol{\pi}), -\pi_i' (\boldsymbol{\pi}'_{-i})^\top \rangle + \langle x^*(\boldsymbol{\pi}), \pi_i' (\boldsymbol{\pi}'_{-i})^\top - \pi_i \boldsymbol{\pi}_{-i}^\top \rangle \\
&\le |\mathcal{A}| \, \|r_i^\star\|_\infty \left\| \pi_i' (\boldsymbol{\pi}'_{-i})^\top - \pi_i \boldsymbol{\pi}_{-i}^\top \right\|_\infty + \|\hat{x}' - x^*(\boldsymbol{\pi})\|_\infty + |\mathcal{A}| \, \|x^*(\boldsymbol{\pi})\|_\infty \left\| \pi_i \boldsymbol{\pi}_{-i}^\top - \pi_i' (\boldsymbol{\pi}'_{-i})^\top \right\|_\infty \\
&\le |\mathcal{A}| r_{\max} \cdot 2 \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_\infty + \|\hat{x}' - x^*(\boldsymbol{\pi}')\|_\infty + |\mathcal{A}| \cdot M \cdot 2 \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_\infty \\
&< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\
&= \epsilon.
\end{aligned}$$

For the other direction, we have

$$\begin{aligned}
F_i(\boldsymbol{\pi}) - F_i(\boldsymbol{\pi}') &= \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i(\boldsymbol{\pi})}(-\pi_i \boldsymbol{\pi}_{-i}^\top) - (\pi_i')^\top r_i^\star \boldsymbol{\pi}'_{-i} + \sigma_{\mathcal{R}_i(\boldsymbol{\pi}')}(-\pi_i' (\boldsymbol{\pi}'_{-i})^\top) \\
&= \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sup_{x \in \mathcal{R}_i(\boldsymbol{\pi})} \langle x, -\pi_i \boldsymbol{\pi}_{-i}^\top \rangle - (\pi_i')^\top r_i^\star \boldsymbol{\pi}'_{-i} + \sup_{x \in \mathcal{R}_i(\boldsymbol{\pi}')} \langle x, -\pi_i' (\boldsymbol{\pi}'_{-i})^\top \rangle \\
&\le \langle r^\star, \pi_i \boldsymbol{\pi}_{-i}^\top - \pi_i' (\boldsymbol{\pi}'_{-i})^\top \rangle - \langle \hat{x}, -\pi_i \boldsymbol{\pi}_{-i}^\top \rangle + \langle x^*(\boldsymbol{\pi}'), -\pi_i' (\boldsymbol{\pi}'_{-i})^\top \rangle \\
&= \langle r^\star, \pi_i \boldsymbol{\pi}_{-i}^\top - \pi_i' (\boldsymbol{\pi}'_{-i})^\top \rangle - \langle \hat{x}, \pi_i' (\boldsymbol{\pi}'_{-i})^\top - \pi_i \boldsymbol{\pi}_{-i}^\top \rangle + \langle x^*(\boldsymbol{\pi}') - \hat{x}, -\pi_i' (\boldsymbol{\pi}'_{-i})^\top \rangle \\
&\le |\mathcal{A}| \, \|r_i^\star\|_\infty \left\| \pi_i \boldsymbol{\pi}_{-i}^\top - \pi_i' (\boldsymbol{\pi}'_{-i})^\top \right\|_\infty + |\mathcal{A}| \, \|\hat{x}\|_\infty \left\| \pi_i \boldsymbol{\pi}_{-i}^\top - \pi_i' (\boldsymbol{\pi}'_{-i})^\top \right\|_\infty + \|x^*(\boldsymbol{\pi}') - \hat{x}\|_\infty \\
&\le |\mathcal{A}| r_{\max} \cdot 2 \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_\infty + |\mathcal{A}| \cdot M \cdot 2 \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_\infty + \|\hat{x} - x^*(\boldsymbol{\pi}')\| \\
&< \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} \\
&= \epsilon.
\end{aligned}$$

Therefore, $F_i(\cdot)$ is continuous in $\text{int}(\Pi)$. By Assumption 2.1, the support function $\sigma_{\mathcal{R}_i(\boldsymbol{\pi})}\big(-\pi_i \boldsymbol{\pi}_{-i}^\top\big)$ is continuous in the boundary of the compact set $\Pi$. From Proposition C.2, we have that $F_i(\cdot)$ is also continuous in $\text{Bd}(\Pi)$. Therefore, $F_i(\cdot)$ is continuous in $\Pi$. $\square$

**Lemma C.4.** *Under Assumptions 2.2 and 2.1, for each agent $i \in [N]$, $F_i(\boldsymbol{\pi})$ is concave in $\pi_i$ given a fixed $\boldsymbol{\pi}_{-i}$.*

*Proof.* By Proposition C.2, the concavity of $F_i(\boldsymbol{\pi})$ in $\pi_i$ follows by the convexity of the support function $\sigma_{\mathcal{R}_i(\boldsymbol{\pi})}\big(-\pi_i \boldsymbol{\pi}_{-i}^\top\big)$. $\square$

### C.1.2. Existence of RNE

**Theorem C.5.** *Under Assumptions 2.2 and 2.1, the robust matrix game has an equilibrium.*

*Proof.* We first construct a point-to-set mapping. Note that $\Pi$ is closed, bounded and convex. We define $\phi : \Pi \to 2^{\Pi}$ as

$$\phi(\boldsymbol{\pi}) := \left\{ z \in \Pi \mid z_i \in \arg\max_{u_i \in \Delta(\mathcal{A}_i)} F_i(u_i, \boldsymbol{\pi}_{-i}), i = 1, \ldots, N \right\}.$$

We next show that $\phi$ satisfies all the conditions in the Kakutani's Fixed Point Theorem.

By Lemma C.3, $F_i(u_i, \boldsymbol{\pi}_{-i})$ is continuous. By Weierstrass' Theorem, the maximum of this continuous function on a compact set $\Delta(\mathcal{A}_i)$ exists, i.e., $\arg\max_{u_i \in \Delta(\mathcal{A}_i)} F_i(u_i, \boldsymbol{\pi}_{-i}) \neq \emptyset$. We thus have $\phi(\boldsymbol{\pi}) \neq \emptyset$ for each $\boldsymbol{\pi} \in \Pi$.

Next, we show that $\phi(\boldsymbol{\pi})$ is a convex set for each $\boldsymbol{\pi} \in \Pi$. Suppose that $z, w \in \phi(\boldsymbol{\pi})$. By the definition of $\phi$, for each $i \in [N]$ and $\forall y_i \in \Delta(\mathcal{A}_i)$, we have

$$F_i(z_i, \boldsymbol{\pi}_{-i}) = F_i(w_i, \boldsymbol{\pi}_{-i}) \geq F_i(y_i, \boldsymbol{\pi}_{-i}).$$

Hence, for each $\lambda \in [0, 1]$, we have

$$\lambda F_i(z_i, \boldsymbol{\pi}_{-i}) + (1 - \lambda) F_i(w_i, \boldsymbol{\pi}_{-i}) \geq F_i(y_i, \boldsymbol{\pi}_{-i}).$$

By the concavity of $F_i$ from Lemma C.4,

$$F_i(\lambda z_i + (1 - \lambda) w_i, \boldsymbol{\pi}_{-i}) \geq \lambda F_i(z_i, \boldsymbol{\pi}_{-i}) + (1 - \lambda) F_i(w_i, \boldsymbol{\pi}_{-i}) \geq F_i(y_i, \boldsymbol{\pi}_{-i}).$$

Therefore, $\lambda z + (1 - \lambda) w \in \phi(\boldsymbol{\pi})$.

We now show that $\phi$ is upper semi-continuous. Suppose that $\boldsymbol{\pi}^n \in \Pi$ with $\lim_{n\to\infty} \boldsymbol{\pi}^n = \boldsymbol{\pi}$, and $z^n \in \phi(\boldsymbol{\pi}^n)$ with $\lim_{n\to\infty} z^n = z$. By the definition of $\phi$,for each $n$ and $i \in [N]$ and $\forall y_i \in \Delta(\mathcal{A}_i)$, we have

$$F_i(z_i^n, \boldsymbol{\pi}_{-i}^n) \geq F_i(y_i, \boldsymbol{\pi}_{-i}^n).$$

By the continuity of $F_i$, if we take the limit on both sides, we have

$$F_i(z_i, \boldsymbol{\pi}_{-i}) \geq F_i(y_i, \boldsymbol{\pi}_{-i}).$$

Hence $z \in \phi(\boldsymbol{\pi})$. Therefore, $\phi$ is upper semi-continuous.

Together, we show that $\phi$ satisfies all the conditions of Kakutani's Fixed Point Theorem (Theorem C.1). Therefore, there exists $\boldsymbol{\pi} \in \Pi$, such that $\boldsymbol{\pi} \in \phi(\boldsymbol{\pi})$. That is, there exists an equilibrium in the robust matrix game. ☐

### C.2. Markov Games: Proof of Theorem 2.1

*Proof.* We define,

$$\boldsymbol{\pi}_h^\dagger(s) \in NE(Q_h^{\boldsymbol{\pi}^\dagger}(s, \cdot)). \tag{23}$$

We will show that the above policy $\boldsymbol{\pi}^\dagger$ is well defined. By definition of NE and (12), this means that,

$$V_{i,h}^{\boldsymbol{\pi}^\dagger}(s) = Q_{i,h}^{\boldsymbol{\pi}^\dagger}(s, \boldsymbol{\pi}_h^\dagger(s)) = \sup_{u \in \Delta(\mathcal{A}_i)} Q_{i,h}^{\boldsymbol{\pi}^\dagger}(s, (u, \boldsymbol{\pi}_{-i,h}^\dagger(s))).$$

We show the stronger claim that for the policy $\boldsymbol{\pi}^\dagger$ defined above,

$$V_{i,h}^{\boldsymbol{\pi}^\dagger}(s) = \sup_{\tilde{\pi}_i \in \Pi_i} V_{i,h}^{(\tilde{\pi}_i, \boldsymbol{\pi}_{-i}^\dagger)}(s), \qquad \forall h \in [H], s \in \mathcal{S}, i \in [N], \tag{24}$$

where $\Pi_i$ denotes the set of policies for agent $i$. We proceed by induction on $h$.

1. (*Base Case*) Suppose that $h = H$. For each $s \in \mathcal{S}$, by Theorem C.5, there exists an equilibrium for the robust matrix game with reward uncertainty set $\times_{i\in[N]}\mathcal{U}_{i,s,H}$. Therefore, $\boldsymbol{\pi}^\dagger_H(s)$ in (23) is well defined. We also have,

$$\begin{aligned}
V^{\boldsymbol{\pi}^\dagger}_{i,H}(s) &= \sup_{u\in\Delta(\mathcal{A}_i)} Q^{\boldsymbol{\pi}^\dagger}_{i,H}\big(s,(u,\boldsymbol{\pi}^\dagger_{-i,H}(s))\big) \\
&= \sup_{u\in\Delta(\mathcal{A}_i)} \inf_{\boldsymbol{r}_H\in\mathcal{U}^r_{s,H}} \mathbb{E}_{\boldsymbol{a}\sim(u,\boldsymbol{\pi}^\dagger_{-i,H}(s))}\left[r_{i,H}(s,\boldsymbol{a})\right] \\
&= \sup_{\tilde\pi_i\in\Pi_i} \inf_{\boldsymbol{r}_H\in\mathcal{U}^r_{s,H}} \mathbb{E}_{\boldsymbol{a}\sim(\tilde\pi_{i,H}(s),\boldsymbol{\pi}^\dagger_{-i,H}(s))}\left[r_{i,H}(s,\boldsymbol{a})\right] \\
&= \sup_{\tilde\pi_i\in\Pi_i} V^{(\tilde\pi_i,\boldsymbol{\pi}^\dagger_{-i})}_{i,H}(s).
\end{aligned}$$

The second equality follows from (13) and the fourth equality follows from (12). The third equality follows since $\boldsymbol{\pi}^\dagger$ is Markovian.

Overall, we see that (24) holds for $h = H$.

2. (*Inductive Step*) Suppose that $h < H$. For any $i \in [N]$ and $s \in \mathcal{S}$, for each $\mu \in \Delta(\mathcal{A})$, we define

$$\begin{aligned}
F_{i,s,h}(\mu) &= Q^{\boldsymbol{\pi}^\dagger}_{i,h}(s,\mu) \\
&= \inf_{\boldsymbol{r}_h\in\mathcal{U}^r_{s,h}(\mu)} \inf_{P_h\in\mathcal{P}_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[r_{i,h}(s,\boldsymbol{a}) + [P_h V^{\boldsymbol{\pi}^\dagger}_{i,h+1}](s,\boldsymbol{a})\right].
\end{aligned}$$

Following the same line of argument for the proof of Theorem C.5 for a robust matrix game, we can show that $F_{i,s,h}(\mu)$ is continuous and concave, and consequently, there exists an equilibrium for the stage game with the expected payoff $\{F_{i,s,h}\}_{i\in[N]}$. Therefore, the NE $\boldsymbol{\pi}^\dagger_h(s)$ in (23) is well defined. We see that,

$$\begin{aligned}
V^{\boldsymbol{\pi}^\dagger}_{i,h}(s) &= \sup_{u\in\Delta(\mathcal{A}_i)} Q^{\boldsymbol{\pi}^\dagger}_{i,h}\big(s,(u,\boldsymbol{\pi}^\dagger_{-i,h}(s))\big) \\
&= \sup_{u\in\Delta(\mathcal{A}_i)} \inf_{\boldsymbol{r}_h\in\mathcal{U}^r_{s,h}} \inf_{P_h\in\mathcal{P}_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim(\tilde\pi_{i,h}(s),\boldsymbol{\pi}^\dagger_{-i,h}(s))}\left[r_{i,h}(s,\boldsymbol{a}) + [P_h V^{\boldsymbol{\pi}^\dagger}_{i,h+1}](s,\boldsymbol{a})\right] \\
&\le \sup_{\tilde\pi_i\in\Pi_i} \inf_{\boldsymbol{r}_h\in\mathcal{U}^r_{s,h}} \inf_{P_h\in\mathcal{P}_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim(\tilde\pi_{i,h}(s),\boldsymbol{\pi}^\dagger_{-i,h}(s))}\left[r_{i,h}(s,\boldsymbol{a}) + [P_h V^{(\tilde\pi_i,\boldsymbol{\pi}^\dagger_{-i})}_{i,h+1}](s,\boldsymbol{a})\right] \\
&= \sup_{\tilde\pi_{i,h}(s)\in\Delta(\mathcal{A}_i)} \inf_{\boldsymbol{r}_h\in\mathcal{U}^r_{s,h}} \inf_{P_h\in\mathcal{P}_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim(\tilde\pi_{i,h}(s),\boldsymbol{\pi}^\dagger_{-i,h}(s))}\left[r_{i,h}(s,\boldsymbol{a}) + \sup_{\tilde\pi_i\in\Pi_i}[P_h V^{(\tilde\pi_i,\boldsymbol{\pi}^\dagger_{-i})}_{i,h+1}](s,\boldsymbol{a})\right] \\
&\le \sup_{\tilde\pi_{i,h}(s)\in\Delta(\mathcal{A}_i)} \inf_{\boldsymbol{r}_h\in\mathcal{U}^r_{s,h}} \inf_{P_h\in\mathcal{P}_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim(\tilde\pi_{i,h}(s),\boldsymbol{\pi}^\dagger_{-i,h}(s))}\left[r_{i,h}(s,\boldsymbol{a}) + \mathbb{E}_{s'\sim P_h(s,a)}\left[\sup_{\tilde\pi_i\in\Pi_i} V^{(\tilde\pi_i,\boldsymbol{\pi}^\dagger_{-i})}_{i,h+1}(s')\right]\right] \\
&= \sup_{\tilde\pi_{i,h}(s)\in\Delta(\mathcal{A}_i)} \inf_{\boldsymbol{r}_h\in\mathcal{U}^r_{s,h}} \inf_{P_h\in\mathcal{P}_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim(\tilde\pi_{i,h}(s),\boldsymbol{\pi}^\dagger_{-i,h}(s))}\left[r_{i,h}(s,\boldsymbol{a}) + \mathbb{E}_{s'\sim P_h(s,a)}\left[V^{\boldsymbol{\pi}^\dagger}_{i,h+1}(s')\right]\right] \\
&= \sup_{\tilde\pi_{i,h}(s)\in\Delta(\mathcal{A}_i)} Q^{\boldsymbol{\pi}^\dagger}_{i,h}\big(s,(\tilde\pi_{i,h}(s),\boldsymbol{\pi}^\dagger_{-i,h}(s))\big) \\
&= V^{\boldsymbol{\pi}^\dagger}_{i,h}(s).
\end{aligned}$$

The first two equalities use (23) and (13), respectively. The first inequality follows by allowing player $i$ to also deviate at future steps. The second inequality uses Jensen's inequality. The following equality follows from the induction hypothesis. The last two equalities use (13) and (23) respectively.

Next, since the starting and ending terms are the same, all inequalities must be equalities. In particular, the first inequality is equality, which is the relationship we wanted to show after applying (23) and (13). Since this relationship holds for arbitrary $s \in \mathcal{S}$ and $i \in [N]$, we see that (24) holds at time $h$.

This completes the proof of (24). The fact that $\boldsymbol{\pi}^\dagger$ is an NE then immediately follows from the $h = 1$ case.

□

# D. Analysis of Markov Games with Reward Uncertainty

## D.1. Proof of Theorem 3.1

Let us first recall the definition of *regularized value functions* for a regularized Markov game $(\mathcal{S}, \{\mathcal{A}_i\}_{i\in[N]}, P, \boldsymbol{r}, H, \Omega)$, with $G = (P, \boldsymbol{r})$. Given a joint policy $\boldsymbol{\pi} \in \Pi$, for each player $i$, $\forall s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}, h \in [H]$,

$$\widetilde{V}_{i,h}^{\boldsymbol{\pi}}(s, G) = \mathbb{E}_P^{\boldsymbol{\pi}}\Big[\sum_{t=h}^{H} r_{i,t}(s_t, \boldsymbol{a}_t) - \Omega_{i,t}(\boldsymbol{\pi}_t(s_t)) | s_h = s\Big], \tag{25}$$

$$\widetilde{Q}_{i,h}^{\boldsymbol{\pi}}(s, \boldsymbol{a}, G) = r_{i,h}(s, \boldsymbol{a}) + \mathbb{E}_P^{\boldsymbol{\pi}}\Big[\sum_{t=h+1}^{H} (r_{i,t}(s_t, \boldsymbol{a}_t) - \Omega_{i,t}(\boldsymbol{\pi}_t(s_t)))\, | s_h = s, \boldsymbol{a}_h = \boldsymbol{a}\Big]. \tag{26}$$

*Proof.* By the definition of the expected payoff $\widetilde{V}_i^{\boldsymbol{\pi}}$ in (25) for each player $i$ in a regularized game, for each product joint policy $\boldsymbol{\pi} \in \Pi$, we have

$$\widetilde{V}_i^{\boldsymbol{\pi}}(\boldsymbol{r}^\star) = \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \Omega_i(\boldsymbol{\pi}) = \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i}(-\pi_i \boldsymbol{\pi}_{-i}^\top).$$

For the robust game, by Proposition C.2, for each $\boldsymbol{\pi} \in \Pi$, the robust value of player $i$ satisfies:

$$V_i^{\boldsymbol{\pi}} = \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i}(-\pi_i \boldsymbol{\pi}_{-i}^\top) = \widetilde{V}_i^{\boldsymbol{\pi}}(\boldsymbol{r}^\star).$$

Consider any RNE $\boldsymbol{\pi}^\dagger$ of the robust game. By definition, we have

$$\pi_i^\dagger \in \arg\max_{\pi_i \in \Delta(\mathcal{A}_i)} V_i^{\pi_i \times \boldsymbol{\pi}^\dagger_{-i}} = \arg\max_{\pi_i \in \Delta(\mathcal{A}_i)} \widetilde{V}_i^{\pi_i \times \boldsymbol{\pi}^\dagger_{-i}}(\boldsymbol{r}^\star),$$

which implies that $\boldsymbol{\pi}^\dagger$ is an NE of the regularized game. Following a similar argument, we can conclude that any NE of the regularized game is an RNE of the robust game. □

## D.2. Proof of Theorem 3.3

*Proof.* For ball constrained uncertainty set $\mathcal{R}_i := \Big\{R_i \in \mathbb{R}^{\mathcal{A}_i \times \mathcal{A}_{-i}} : \|R_i\|_{q^* \to p} \le \alpha_i\Big\}$, we have

$$\begin{aligned}
\sigma_{\mathcal{R}_i}(-\pi_i \boldsymbol{\pi}_{-i}^\top) &= \sup_{R_i \in \mathcal{R}_i} \langle R_i, -\pi_i \boldsymbol{\pi}_{-i}^\top\rangle \\
&= \sup_{R_i : \|R_i\|_{q^*\to p} \le \alpha_i} \langle R_i, -\pi_i \boldsymbol{\pi}_{-i}^\top\rangle \\
&= \alpha_i \sup_{R_i : \|R_i\|_{q^*\to p} \le 1} \langle R_i, -\pi_i \boldsymbol{\pi}_{-i}^\top\rangle \\
&= \alpha_i \|-\pi_i\|_p \|\boldsymbol{\pi}_{-i}\|_q \\
&= \Omega_i(\boldsymbol{\pi}),
\end{aligned}$$

where the second to last equality follows from Lemma A.3 on the dual norm of matrix operator norm. The equivalence between the robust game and the regularized games immediately follow from Theorem 3.1.

For the $(s, a)$-rectangular policy-dependent uncertainty set, let $\mathcal{R}_i(\boldsymbol{\pi}) := \times_{\boldsymbol{a} \in \mathcal{A}} \mathcal{R}_{i,\boldsymbol{a}}(\boldsymbol{\pi})$. We have

$$\begin{aligned}
\sigma_{\mathcal{R}_i^{\boldsymbol{\pi}}}(-\pi_i \boldsymbol{\pi}_{-i}^\top) &= \sup_{R_i \in \mathcal{R}_i(\boldsymbol{\pi})} \langle R_i, -\pi_i \boldsymbol{\pi}_{-i}^\top\rangle \\
&= -\sum_{a_i} \sum_{\boldsymbol{a}_{-i}} \left[-\tau_i \omega_i(\pi_i(a_i)) - g_i(\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}))\right] \pi_i(a_i) \boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}) \\
&= \tau_i \sum_{a_i} \pi_i(a_i) \omega_i(\pi_i(a_i)) + \sum_{\boldsymbol{a}_{-i}} g_i(\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i})) \boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}).
\end{aligned}$$

By Proposition C.2, the robust best response policy $\mathsf{br}(\boldsymbol{\pi}_{-i})$ for player $i$ is given by the following optimization problem:

$$\arg \sup_{\pi_i \in \Delta(\mathcal{A}_i)} V_i^{\pi_i \times \boldsymbol{\pi}_{-i}}$$

$$= \arg \sup_{\pi_i \in \Delta(\mathcal{A}_i)} \left\{ \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i^\pi}(-\pi_i \boldsymbol{\pi}_{-i}^\top) \right\}$$

$$= \arg \sup_{\pi_i \in \Delta(\mathcal{A}_i)} \left\{ \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \tau_i \sum_{a_i} \pi_i(a_i)\omega_i(\pi_i(a_i)) - \sum_{\boldsymbol{a}_{-i}} g_i(\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}))\boldsymbol{\pi}_{-i}(\boldsymbol{a}_{-i}) \right\}$$

$$\equiv \arg \sup_{\pi_i \in \Delta(\mathcal{A}_i)} \left\{ \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \tau_i \sum_{a_i} \pi_i(a_i)\omega_i(\pi_i(a_i)) \right\}$$

$$= \arg \sup_{\pi_i \in \Delta(\mathcal{A}_i)} \left\{ \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \Omega_i(\boldsymbol{\pi}) \right\}$$

which gives the best response policy w.r.t. $\boldsymbol{\pi}_{-i}$ for the regularized game with regularizer $\Omega = \{\Omega_i\}$. Therefore, solving the RNE of the robust game is equivalent to solving the NE of the regularized normal-form game. □

### D.3. Examples of Game Regularization

As pointed out in Section 3.1, we can apply Theorem 3.3 to popular regularization schemes in games, including negative Shannon entropy regularization, KL divergence regularization, and Tsallis entropy regularization. Here we provide details of the $\{\omega_i\}$ functions for the reward function uncertainty set and two more examples of regularizers studied in games.

- The negative Shannon entropy: $\Omega_i(\boldsymbol{\pi}) = \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i) \log \pi_i(a_i)$. Thus we can define $\omega_i(\pi_i(a_i)) := \log \pi_i(a_i)$.
- The KL divergence regularizer: $\Omega_i(\boldsymbol{\pi}) = \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i) \log \frac{\pi_i(a_i)}{\mu_i(a_i)} = d_{\text{KL}}(\pi_i, \mu_i)$, where $\mu_i \in \Delta(\mathcal{A}_i)$ is a given distribution. We can let $\omega_i(\pi_i(a_i)) := \log \frac{\pi_i(a_i)}{\mu_i(a_i)}$.
- The Tsallis entropy regularizer $\Omega_i(\boldsymbol{\pi}) = \frac{1}{2} \sum_{a_i \in \mathcal{A}_i} (\pi_i(a_i)^2 - \pi_i(a_i))$. Thus we can define $\omega_i(\pi_i(a_i)) := \frac{1}{2}(\pi_i(a_i) - 1)$.
- The Renyi (negative) entropy regularizer: $\Omega_i(\boldsymbol{\pi}) = -(1-q)^{-1} \log \left( \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i)^q \right)$ for a given $q \in (0, 1)$, and thus we can let (with a slight abuse of notation) $\omega_i(\pi_i, a_i) := -(1-q)^{-1} \log \left( \sum_{a_i' \in \mathcal{A}_i} \pi_i(a_i')^q \right)$.

### D.4. Proof of Proposition 3.5

*Proof.* From Proposition B.1 we have that

$$V_{i,h}^{\boldsymbol{\pi}}(s) = \inf_{\boldsymbol{r}_h \in \mathcal{U}_{s,h}^r} \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ r_{i,h}(s, \boldsymbol{a}) \right] + \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ [P_h^\star V_{i,h+1}^{\boldsymbol{\pi}}](s, \boldsymbol{a}) \right].$$

We proceed with backward induction over $h$. In the base case at step $h = H$, there is no further future transitions, thus $\forall s \in \mathcal{S}$ and each $i \in [N]$

$$\begin{aligned} V_{i,H}^{\boldsymbol{\pi}}(s) &= \inf_{\boldsymbol{r}_H \in \mathcal{U}_{s,H}^r} \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_H(s)} \left[ r_{i,H}(s, \boldsymbol{a}) \right] \\ &= \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_H(s)} \left[ \inf_{\boldsymbol{r}_H \in \mathcal{U}_{s,H}^r} r_{i,H}(s, \boldsymbol{a}) \right] \\ &= \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_H(s)} \left[ r_{i,H}^\star(s, \boldsymbol{a}) \right] - \sigma_{\mathcal{R}_{i,s,H}} \left( -\pi_{i,H}(s) \boldsymbol{\pi}_{-i,H}^\top(s) \right) \end{aligned}$$

where the last equality holds from Proposition C.2.

Now suppose that (7) holds for all steps $t > h$. Then $\forall s \in \mathcal{S}$ and each $i \in [N]$, we have

$$\begin{aligned} V_{i,h}^{\boldsymbol{\pi}}(s) &= \inf_{\boldsymbol{r}_h \in \mathcal{U}_{s,h}^r} \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ r_{i,h}(s, \boldsymbol{a}) \right] + \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ [P_h^\star V_{i,h+1}^{\boldsymbol{\pi}}](s, \boldsymbol{a}) \right] \\ &= \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ \inf_{\boldsymbol{r}_h \in \mathcal{U}_{s,h}^r} r_{i,h}(s, \boldsymbol{a}) \right] + \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ [P_h^\star V_{i,h+1}^{\boldsymbol{\pi}}](s, \boldsymbol{a}) \right] \\ &= \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_h(s)} \left[ r_{i,h}^\star(s, \boldsymbol{a}) + [P_h^\star V_{i,h+1}^{\boldsymbol{\pi}}](s, \boldsymbol{a}) \right] - \sigma_{\mathcal{R}_{i,s,h}} \left( -\pi_{i,h}(s) \boldsymbol{\pi}_{-i,h}^\top(s) \right), \end{aligned}$$

which completes the proof for the induction step. □

### D.5. Proof of Theorem 3.6

*Proof.* For the robust MG, by Proposition 3.5, for each $\boldsymbol{\pi} \in \Pi$, the robust value functions of player $i$ satisfy:

$$V_{i,h}^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r_{i,h}^\star(s,\boldsymbol{a}) + [P_h^\star V_{i,h+1}^{\boldsymbol{\pi}}](s,\boldsymbol{a})\right] - \sigma_{\mathcal{R}_{i,s,h}}\left(-\pi_{i,h}(s)\boldsymbol{\pi}_{-i,h}^\top(s)\right).$$

We can use backward induction from $h = H$ to 1 to show that

$$V_{i,h}^{\boldsymbol{\pi}}(s) = \widetilde{V}_{i,h}^{\boldsymbol{\pi}}(s, G^\star), \quad \forall s \in \mathcal{S}, \forall h \in [H].$$

At the base step $h = H$, from Proposition 3.5, we have for $\forall s \in \mathcal{S}$ and each $i \in [N]$

$$\begin{aligned} V_{i,H}^{\boldsymbol{\pi}}(s) &= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_H(s)}\left[r_{i,H}^\star(s,\boldsymbol{a})\right] - \sigma_{\mathcal{R}_{i,s,H}}\left(-\pi_{i,H}(s)\boldsymbol{\pi}_{-i,H}^\top(s)\right) \\ &= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_H(s)}\left[r_{i,H}^\star(s,\boldsymbol{a})\right] - \Omega_{i,H}(\boldsymbol{\pi}, s) \\ &= \widetilde{V}_{i,H}^{\boldsymbol{\pi}}(s, G^\star) \end{aligned}$$

Now assume that for $t > h$ that

$$V_{i,t}^{\boldsymbol{\pi}}(s) = \widetilde{V}_{i,t}^{\boldsymbol{\pi}}(s, G^\star), \quad \forall s \in \mathcal{S}, \forall i \in [N]$$

and that $\boldsymbol{\pi}^\dagger = (\pi_1^\dagger, \ldots, \pi_N^\dagger)$ is an NE for the regularized game at steps $t > h$. Then $\forall s \in \mathcal{S}$ and each $i \in [N]$

$$\begin{aligned} V_{i,h}^{\boldsymbol{\pi}}(s) &= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r_{i,h}^\star(s,\boldsymbol{a}) + [P_h^\star V_{i,h+1}^{\boldsymbol{\pi}}](s,\boldsymbol{a})\right] - \sigma_{\mathcal{R}_{i,s,h}}\left(-\pi_{i,h}(s)\boldsymbol{\pi}_{-i,h}^\top(s)\right) \\ &= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r_{i,h}^\star(s,\boldsymbol{a}) + [P_h^\star \widetilde{V}_{i,h+1}^{\boldsymbol{\pi}}](s,\boldsymbol{a})\right] - \Omega_{i,h}(\boldsymbol{\pi}, s) \\ &= \widetilde{V}_{i,h}^{\boldsymbol{\pi}}(s, G^\star) \end{aligned}$$

Thus $V_{i,h}^{\boldsymbol{\pi}}(s) = \widetilde{V}_{i,h}^{\boldsymbol{\pi}}(s, G^\star), \forall s \in \mathcal{S}, \forall h \in [H], \forall i \in [N]$.

Consider any RNE $\boldsymbol{\pi}^\dagger$ of the robust MG. By definition, we have

$$\pi_i^\dagger \in \operatorname*{arg\,max}_{\pi_i\in\Delta(\mathcal{A}_i)} V_{i,1}^{\pi_i\times\boldsymbol{\pi}^\dagger_{-i}}(s_1) = \operatorname*{arg\,max}_{\pi_i\in\Delta(\mathcal{A}_i)} \widetilde{V}_{i,1}^{\pi_i\times\boldsymbol{\pi}^\dagger_{-i}}(s_1, G^\star),$$

which implies that $\boldsymbol{\pi}^\dagger$ is an NE of the regularized MG. Following a similar argument, we can conclude that any NE of the regularized MG is an RNE of the robust MG. ☐

### D.6. Examples of Regularized Markov Game and Equivalent Robust Markov Game

**Theorem D.1.** *Consider a regularized MG $\mathfrak{G}' = (\mathcal{S}, \{\mathcal{A}_i\}_{i\in[N]}, P^\star, \boldsymbol{r}^\star, H, \Omega)$ with regularizer functions $\Omega := (\Omega_{i,h})_{i\in[N],h\in[H]}$.*

1. *If $\Omega_{i,h}(s,\mu) = \alpha_{i,s,h}\left\|\mu_i\right\|_p\left\|\mu_{-i}\right\|_q$ is the $\ell_p/\ell_q$-norm regularizer for each $i \in [N], h \in [H], s \in \mathcal{S}$ and $\forall \mu \in \Delta(\mathcal{A})$, then solving for MPNE of the regularized game $\mathfrak{G}'$ is equivalent to solving for MPRNE of the robust game $\mathfrak{G}$ with s-rectangular ball constrained reward uncertainty set $\mathcal{U}^r = \boldsymbol{r}^\star + \times_{i,s,h}\mathcal{R}_{i,s,h}$,*

   $$\mathcal{R}_{i,s,h} = \left\{R_i \in \mathbb{R}^{\mathcal{A}_i\times\mathcal{A}_{-i}} : \left\|R_i\right\|_{q^*\to p} \le \alpha_{i,s,h}\right\},$$

   *where $q^*$ satisfies $\frac{1}{q^*} + \frac{1}{q} = 1$.*

2. *If $\Omega_{i,h}$ is decomposable with kernel $\omega$, i.e., $\Omega_{i,h}(s,\mu) := \tau_{i,s,h}\sum_{a_i\in\mathcal{A}_i}\mu_i(a_i)\omega_{i,s,h}\big(\mu_i(a_i)\big)$, $i \in [N], h \in [H], s \in \mathcal{S}$, with $\tau_{i,s,h} \ge 0$, and $\Omega_{i,h}(s,\mu)$ is convex in $\mu_i$ for each given $\mu_{-i}$. Then solving for NE of the regularized game $\mathfrak{G}'$ is equivalent to solving for RNE of robust game with $(s,a)$-rectangular policy-dependent uncertainty set $\mathcal{U}(\boldsymbol{\pi}) = \boldsymbol{r}^\star + \times_{i,s,\boldsymbol{a},h}\mathcal{R}_{i,s,\boldsymbol{a},h}(\boldsymbol{\pi})$, where*

   $$\begin{aligned} \mathcal{R}_{i,s,\boldsymbol{a},h}(\boldsymbol{\pi}) = \Big[&\tau_{i,s,h}\omega_{i,s,h}\left(\pi_{i,h}(a_i|s)\right) + g_{i,s,h}\left(\boldsymbol{\pi}_{-i,h}(\boldsymbol{a}_{-i}|s)\right), \\ &\overline{\omega}_{i,s,h}\left(\pi_{i,h}(a_i|s)\right) + \overline{g}_{i,s,h}(\boldsymbol{\pi}_{-i,h}(\boldsymbol{a}_{-i}|s))\Big] \subset \mathbb{R}, \end{aligned}$$

   *with functions $\omega_{i,s,h}, \overline{\omega}_{i,s,h} : [0,1] \to \mathbb{R}$ and $g_{i,s,h}, \overline{g}_{i,s,h} : [0,1] \to \mathbb{R}$ are continuous.*

*The proof follows easily from Theorem 3.3 using backward induction.*

#### D.6.1. EXAMPLES OF MARKOV GAME REGULARIZATION

We can apply Theorem D.1 above to popular regularization schemes in Markov games. Below we provide four examples with the corresponding $\{\omega_i\}$ functions for the reward function uncertainty set.

- The negative Shannon entropy: $\Omega_{i,h}(s,\mu) = \sum_{a_i \in \mathcal{A}_i} \mu_i(a_i) \log \mu_i(a_i)$. Thus we can define $\omega_{i,s,h}(\pi_{i,h}(a_i|s)) := \log \pi_{i,h}(a_i|s)$.
- The KL divergence regularizer: $\Omega_{i,h}(s,\mu) = \sum_{a_i \in \mathcal{A}_i} \mu_i(a_i) \log \frac{\mu_i(a_i)}{\nu_i(a_i)} = d_{\mathrm{KL}}(\mu_i, \nu_i)$, where $\nu_i \in \Delta(\mathcal{A}_i)$ is a given distribution. We can let $\omega_{i,s,h}(\pi_{i,h}(a_i|s)) := \log \frac{\pi_{i,h}(a_i|s)}{\nu_i(a_i)}$.
- The Tsallis entropy regularizer $\Omega_{i,h}(s,\mu) = \frac{1}{2}\sum_{a_i \in \mathcal{A}_i} (\mu_i(a_i)^2 - \mu_i(a_i))$. Thus we can define $\omega_{i,s,h}(\pi_{i,h}(a_i|s)) := \frac{1}{2}(\pi_{i,h}(a_i|s) - 1)$.
- The Renyi (negative) entropy regularizer: $\Omega_{i,h}(s,\mu) = -(1-q)^{-1} \log\left(\sum_{a_i \in \mathcal{A}_i} \mu_i(a_i)^q\right)$ for a given $q \in (0,1)$. Thus we can let (with a slight abuse of notation) $\omega_{i,s,h}(\pi_{i,h}(\cdot|s), a_i) := -(1-q)^{-1} \log\left(\sum_{a_i' \in \mathcal{A}_i} \pi_{i,h}(a_i'|s)^q\right)$.

## E. Analysis of Robust Zero-Sum Markov Games

### E.1. Proof of Theorem 4.1

*Proof.* We present a poly-time reduction from the problem of computing an NE for a general-sum game to the problem of computing an RNE for a two-player zero-sum robust matrix game with $(s,a)$-rectangular reward uncertainty. Since computing an NE of a general-sum game is PPAD-hard, it then follows that computing an RNE for the aforementioned class of robust matrix games is also PPAD-hard. Let $(A,B)$ be an arbitrary general sum matrix game. WLOG we can further assume that $A, B \leq 0$. To construct the robust matrix game instance, we first define $\overline{r} = -\frac{A+B}{2}$ and $\underline{r} = \frac{A+B}{2}$. Then, we map $(A,B)$ to the robust matrix game $\mathfrak{G}$ defined by $r^\star = A - \underline{r}$, $\mathcal{R}_1 = \{r \in \mathbb{R}^{n_1 \times n_2} \mid \underline{r} \leq r \leq \overline{r}\}$, and $\mathcal{R}_2 = -\mathcal{R}_1$.

To prove the reduction is correct, we show that $\pi$ is an NE for $(A,B)$ if and only if $\pi$ is an RNE for $\mathfrak{G}$. First, we observe that for player 1,

$$\begin{aligned}
-\sigma_{\mathcal{R}_1}(-\pi_1\pi_2^\top) &= -\sup_{\underline{r} \leq R \leq \overline{r}} \langle R, -\pi_1\pi_2^\top \rangle \\
&= -\sup_{\underline{r} \leq R \leq \overline{r}} -\pi_1^\top R \pi_2 \\
&= \inf_{\underline{r} \leq R \leq \overline{r}} \pi_1^\top R \pi_2 \\
&= \pi_1^\top \underline{r} \pi_2.
\end{aligned}$$

Using Proposition C.2, we see the robust suboptimality gap for player 1 under $\pi$ is exactly,

$$D_1(\pi) = \max_{\pi_1' \in \Delta(\mathcal{A}_1)} \left\{\pi_1'^\top (r^\star + \underline{r})\pi_2\right\} - (\pi_1^\top (r^\star + \underline{r})\pi_2).$$

Similarly, we observe that $-\sigma_{-\mathcal{R}_1}(-\pi_1\pi_2^\top) = \pi_1^\top(-\overline{r})\pi_2$. Thus, the robust suboptimality gap for player 2 under $\pi$ is exactly,

$$D_2(\pi) = \max_{\pi_2' \in \Delta(\mathcal{A}_2)} \left\{\pi_1^\top (-r^\star - \overline{r})\pi_2'\right\} - (\pi_1^\top (-r^\star - \overline{r})\pi_2).$$

Putting these together, the RNE gap for $\pi$ is,

$$\begin{aligned}
\text{RNEGap}(\pi) &= \max_{\pi_1' \in \Delta(\mathcal{A}_1)} \left\{\pi_1'^\top (r^\star + \underline{r})\pi_2\right\} + \max_{\pi_2' \in \Delta(\mathcal{A}_2)} \left\{\pi_1^\top (-r^\star - \overline{r})\pi_2'\right\} - \pi_1^\top (r^\star + \underline{r})\pi_2 - \pi_1^\top (-r^\star - \overline{r})\pi_2 \\
&= \max_{\pi_1' \in \Delta(\mathcal{A}_1)} \left\{\pi_1'^\top A \pi_2\right\} + \max_{\pi_2' \in \Delta(\mathcal{A}_2)} \left\{\pi_1^\top B \pi_2'\right\} - \pi_1^\top A \pi_2 - \pi_1^\top B \pi_2.
\end{aligned}$$

Observe this last term is exactly the NE gap for $(A,B)$. Thus, minimizing the optimality gaps for both games is equivalent which implies the set of RNEs for $\mathfrak{G}$ is exactly the set of NEs for $(A,B)$. In particular, this means that $\pi$ is an NE for $(A,B)$ if and only if $\pi$ is an RNE for $\mathfrak{G}$. Thus, the reduction is correct.

Lastly, we note that $\mathfrak{G}$ can easily be computed in linear time in the size of $(A, B)$ just by computing the average and difference of matrices. Thus, the reduction can be done in polynomial time.

By simply defining $\mathcal{R}_1(\boldsymbol{a}) = [\underline{r}(\boldsymbol{a}), \overline{r}(\boldsymbol{a})]$ and $\mathcal{R}_2(\boldsymbol{a}) = [-\overline{r}(\boldsymbol{a}), -\underline{r}(\boldsymbol{a})]$ the same proof applies to $(s, a)$-rectangularity. □

### E.2. Proof of Lemma 4.2

Proposition C.2 implies that, the robust value for player $i$ satisfies,

$$V_i^{\boldsymbol{\pi}} = \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i}(-\pi_i \boldsymbol{\pi}_{-i}^\top)$$

Now, suppose that the characteristic function could be decomposed into $\sigma_{\mathcal{R}_i}(-\pi_i \boldsymbol{\pi}_{-i}^\top) = \Omega_{i,i}(\pi_i) + \Omega_{i,-i}(\boldsymbol{\pi}_{-i})$. Then the robust suboptimality gap for player $i$ takes the form,

$$\begin{aligned} V_i^{\dagger, \boldsymbol{\pi}_{-i}}(s) - V_i^{\boldsymbol{\pi}}(s) &= \max_{\pi_i' \in \Delta(\mathcal{A}_1)} \left\{ {\pi_i'}^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i}(-\pi_i' \boldsymbol{\pi}_{-i}^\top) \right\} - \left( \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \sigma_{\mathcal{R}_i}(-\pi_i \boldsymbol{\pi}_{-i}^\top) \right) \\ &= \max_{\pi_i' \in \Delta(\mathcal{A}_1)} \left\{ {\pi_i'}^\top r_i^\star \boldsymbol{\pi}_{-i} - \Omega_{i,i}(\pi_i') - \Omega_{i,-i}(\boldsymbol{\pi}_{-i}) \right\} - \left( \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \Omega_{i,i}(\pi_i) - \Omega_{i,-i}(\boldsymbol{\pi}_{-i}) \right) \\ &= \max_{\pi_i' \in \Delta(\mathcal{A}_1)} \left\{ {\pi_i'}^\top r_i^\star \boldsymbol{\pi}_{-i} - \Omega_{i,i}(\pi_i') \right\} - \left( \pi_i^\top r_i^\star \boldsymbol{\pi}_{-i} - \Omega_{i,i}(\pi_i) \right). \end{aligned}$$

Next, we define $\Omega_i(\pi_i) := \Omega_{i,i}(\pi_i)$. We see the RNE gap takes the form,

$$\begin{aligned} \text{RNEGap} = &\max_{\pi_1' \in \Delta(\mathcal{A}_1)} \left\{ {\pi_1'}^\top r^\star \pi_2 - \Omega_1(\pi_1') \right\} - \left( \pi_1^\top r_1^\star \pi_2 - \Omega_1(\pi_1) \right) \\ &+ \max_{\pi_2' \in \Delta(\mathcal{A}_2)} \left\{ {\pi_1}^\top (-r^\star) \pi_2' - \Omega_2(\pi_2') \right\} - \left( \pi_1^\top (-r^\star) \pi_2 - \Omega_2(\pi_2) \right) \\ = &\max_{\pi_1' \in \Delta(\mathcal{A}_1)} \left\{ {\pi_1'}^\top r^\star \pi_2 - \Omega_1(\pi_1') + \Omega_2(\pi_2) \right\} \\ &- \min_{\pi_2' \in \Delta(\mathcal{A}_2)} \left\{ {\pi_1}^\top r^\star \pi_2' - \Omega_1(\pi_1) + \Omega_2(\pi_2') \right\} \end{aligned}$$

This exactly matches the RNEGap for the zero-zum regularized game with regularization functions $\Omega_1$ and $\Omega_2$. Thus, solving the original two-player zero-sum robust game is equivalent to solving the two-player zero-sum regularized game with regularization functions $\Omega_1$ and $\Omega_2$. Furthermore, since $\Omega_{1,1}$ and $\Omega_{2,2}$ are strongly convex by assumption, we know this regularized game can be solved in polynomial time (Zhang et al., 2020a).

### E.3. Proof of Theorem 4.3

*Proof.* The proof is nearly immediate given the proof of Theorem 3.6, Lemma 4.2, and the definition of regularized MGs. For completeness, we give a formal proof below. We follow the proof of Theorem 3.6 and show the constructed regularized game is TPZS. We proceed by backward induction on $h$. For the base case, we consider $h = H$. Fix any policy $\boldsymbol{\pi}$. For any $s \in \mathcal{S}$ and $i \in [N]$, we know by (7),

$$V_{i,H}^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_H(s)} \left[ r_{i,H}^\star(s, \boldsymbol{a}) \right] - \sigma_{\mathcal{R}_{i,s,H}} \left( -\pi_{i,H}(s) \boldsymbol{\pi}_{-i,H}^\top(s) \right).$$

Since $\sigma_{\mathcal{R}_{i,s,H}} \left( -\pi_{i,H}(s) \boldsymbol{\pi}_{-i,H}^\top(s) \right) = \Omega_{i,i}^H(\pi_{i,H}(s)) + \Omega_{i,-i}^H(\boldsymbol{\pi}_{-i,H}^\top(s))$ by Definition 4.2, Lemma 4.2 implies an NE for the robust stage game is equivalent to an NE for the corresponding TPZS regularized game.

For the inductive step, suppose that $h < H$. Fix any policy $\boldsymbol{\pi}$. For any $s \in \mathcal{S}$ and $i \in [N]$, again we know by (7),

$$V_{i,H}^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{a} \sim \boldsymbol{\pi}_H(s)} \left[ r_{i,h}^\star(s, \boldsymbol{a}) + [P_h^\star V_{i,h+1}^{\boldsymbol{\pi}}](s, \boldsymbol{a}) \right] - \sigma_{\mathcal{R}_{i,s,h}} \left( -\pi_{i,h}(s) \boldsymbol{\pi}_{-i,h}^\top(s) \right).$$

From the proof of Theorem 3.6, we know that $[P_h^\star V_{i,h+1}^{\boldsymbol{\pi}}](s, \boldsymbol{a}) = [P_h^\star \tilde{V}_{i,h+1}^{\boldsymbol{\pi}}](s, \boldsymbol{a})$. Furthermore, the induction hypothesis implies the future game is TPZS and so admits a unique NE value $\tilde{V}_{i,h+1}^{\boldsymbol{\pi}}(s') = \tilde{V}_{i,h+1}^*(s')$ for each $s' \in \mathcal{S}$. Define

$r'_{i,h}(s,\boldsymbol{a}) := r^\star_{i,h}(s,\boldsymbol{a}) + [P^\star_h \tilde{V}^*_{i,h+1}](s,\boldsymbol{a})$. Since $\sigma_{\mathcal{R}_{i,s,h}}\left(-\pi_{i,h}(s)\boldsymbol{\pi}^\top_{-i,h}(s)\right) = \Omega^h_{i,i}(\pi_{i,h}(s)) + \Omega^h_{i,-i}(\boldsymbol{\pi}^\top_{-i,h}(s))$ by Definition 4.2, we see that solutions to the robust stage game correspond to a TPZS regularized game by Lemma 4.2. Furthermore, this regularized game is exactly the regularized stage game of $\mathfrak{G}'$. This completes the proof. □

### E.4. Proof of Theorem 4.5

*Proof.* The two cases of efficiently decomposable structures are

1. For the ball constrained uncertainty case $\mathcal{R}_{1,s,h} = \left\{R_1 \in \mathbb{R}^{\mathcal{A}_1\times\mathcal{A}_2} : \|R_1\|_{\infty\to p} \le \alpha_{1,s,h}\right\}$ and $\mathcal{R}_{2,s,h} = \left\{R_2 \in \mathbb{R}^{\mathcal{A}_1\times\mathcal{A}_2} : \left\|R_2^\top\right\|_{\infty\to p} \le \alpha_{2,s,h}\right\}$.

   We can show the decomposability for each player. For player 1, we have

$$\begin{aligned}\sigma_{\mathcal{R}_{1,s,h}}(-\pi_{1,h}(s)\pi^\top_{2,h}(s)) &= \sup_{R_1\in\mathcal{R}_{1,s,h}} \langle R_1, -\pi_{1,h}(s)\pi^\top_{2,h}(s)\rangle \\ &= \sup_{R_1:\|R_1\|_{\infty\to p}\le\alpha_{1,s,h}} \langle R_1, -\pi_{1,h}(s)\pi^\top_{2,h}(s)\rangle \\ &= \alpha_{1,s,h} \sup_{R_1:\|R_1\|_{\infty\to p}\le 1} \langle R_1, -\pi_{1,h}(s)\pi^\top_{2,h}(s)\rangle \\ &= \alpha_{1,s,h}\left\|-\pi_{1,h}(s)\right\|_p\end{aligned}$$

   Similarly, for player 2, we have

$$\begin{aligned}\sigma_{\mathcal{R}_{2,s,h}}(-\pi_{1,h}(s)\pi^\top_{2,h}(s)) &= \sup_{R_2\in\mathcal{R}_{2,s,h}} \langle R_2, -\pi_{1,h}(s)\pi^\top_{2,h}(s)\rangle \\ &= \sup_{R_2:\left\|R_2^\top\right\|_{\infty\to p}\le\alpha_{2,s,h}} \langle R_2, -\pi_{1,h}(s)\pi^\top_{2,h}(s)\rangle \\ &= \alpha_{2,s,h} \sup_{R_2:\left\|R_2^\top\right\|_{\infty\to p}\le 1} \langle R_2, -\pi_{1,h}(s)\pi^\top_{2,h}(s)\rangle \\ &= \alpha_{2,s,h}\left\|-\pi_{2,h}(s)\right\|_p\end{aligned}$$

   where the last equality for both player views follows from Lemma A.3 on the dual norm of matrix operator norm. Thus we obtain decombosable $\Omega_{i,h}(s,\pi_{i,h}(s)) = \alpha_{i,s,h}\left\|\pi_{i,h}(s)\right\|_p$.

2. For the decomposable kernel case,

$$\mathcal{R}_{i,s,\boldsymbol{a},h}(\boldsymbol{\pi}) = \Big[\tau_{i,s,h}\omega_{i,s,h}\left(\pi_{i,h}(a_i|s)\right) + g_{i,s,h}\left(\boldsymbol{\pi}_{-i,h}(\boldsymbol{a}_{-i}|s)\right), \overline{\omega}_{i,s,h}\left(\pi_{i,h}(a_i|s)\right) + \overline{g}_{i,s,h}(\boldsymbol{\pi}_{-i,h}(\boldsymbol{a}_{-i}|s))\Big] \subset \mathbb{R},$$

   with parameters $\tau_{i,s,h} \ge 0$, functions $\omega_{i,s,h}, \overline{\omega}_{i,s,h} : [0,1] \to \mathbb{R}$ and $g_{i,s,h}, \overline{g}_{i,s,h} : [0,1] \to \mathbb{R}$.

   This is the same structure as in Theorem D.1, and thus we know the decomposable regularizer is $\Omega_{i,h}(s,\boldsymbol{\pi}) = \tau_{i,s,h}\sum_{a_i}\pi_{i,h}(a_i|s)\omega_{i,s,h}(\pi_{i,h}(a_i|s))$.

With each case of $\Omega_{i,h}$ shown to be decomposable, we can proceed as in Theorem D.1. The MPRNE of the RMG, and MPNE of the corresponding regularized MG, are found by solving for the respective NE for each $s \in \mathcal{S}$ at each $h \in [H]$. For either case, for each $s \in \mathcal{S}$, we have from Theorem 3.3 that at each step $h = H$ we can efficiently solve the RNE through the corresponding regularized NE. Then we can proceed via backward induction to solve the NE for all $h$, as done in Theorem 3.6. □

## F. Analysis of Markov Games with Transition Uncertainty

### F.1. Proof of Proposition 5.1

*Proof.* We will prove the proposition via an induction from step $h = H$ to 1.

For the base case $h = H$, as there is no transition, (8) holds by the definition of the robust value function.

Now suppose that (8) holds for all $t \geq h+1$. By Proposition B.1, we have

$$\begin{aligned} V_{i,h}^{\boldsymbol{\pi}}(s) &= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r_{i,h}^{\star}(s,\boldsymbol{a})\right] + \inf_{P_h\in\mathcal{P}_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[[P_h V_{i,h+1}^{\boldsymbol{\pi}}](s,\boldsymbol{a})\right] \\ &= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r_{i,h}^{\star}(s,\boldsymbol{a})\right] + \inf_{P_h\in\mathcal{P}_{s,h}} \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s),s'\sim P_h(\cdot|s,\boldsymbol{a})}\left[V_{i,h+1}^{\boldsymbol{\pi}}(s')\right] \\ &= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r_{i,h}^{\star}(s,\boldsymbol{a})\right] + \inf_{P_h\in\mathcal{P}_{s,h}} \langle P_h, V_{i,h+1}^{\boldsymbol{\pi}}\boldsymbol{\pi}_h(s)^{\top}\rangle \\ &= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r_{i,h}^{\star}(s,\boldsymbol{a})\right] - \sup_{P_h\in\mathcal{P}_{s,h}} \langle P_h, -V_{i,h+1}^{\boldsymbol{\pi}}\boldsymbol{\pi}_h(s)^{\top}\rangle \\ &= \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r_{i,h}^{\star}(s,\boldsymbol{a})\right] - \sigma_{\mathcal{P}_{s,h}}\left(-V_{i,h+1}^{\boldsymbol{\pi}}\boldsymbol{\pi}_h(s)^{\top}\right). \end{aligned}$$

□

## F.2. Proof of Theorem 5.2

*Proof.* For the robust MG, by Proposition 5.1, for each $\boldsymbol{\pi}\in\Pi$, the robust value functions of player $i$ satisfy:

$$V_{i,h}^{\boldsymbol{\pi}}(s) = \mathbb{E}_{\boldsymbol{a}\sim\boldsymbol{\pi}_h(s)}\left[r_{i,h}^{\star}(s,\boldsymbol{a})\right] - \sigma_{\mathcal{P}_{s,h}}\left(-V_{i,h+1}^{\boldsymbol{\pi}}\boldsymbol{\pi}_h(s)^{\top}\right), \quad \forall s\in\mathcal{S}, \forall h\in[H].$$

We can use backward induction from $h = H$ to 1 to show that

$$V_{i,h}^{\boldsymbol{\pi}}(s) = \widehat{V}_{i,h}^{\boldsymbol{\pi}}(s, G^{\star}), \quad \forall s\in\mathcal{S}, \forall h\in[H].$$

Consider any RNE $\boldsymbol{\pi}^{\dagger}$ of the robust MG. By definition, we have

$$\pi_i^{\dagger} \in \operatorname*{arg\,max}_{\pi_i\in\Delta(\mathcal{A}_i)} V_{i,1}^{\pi_i\times\boldsymbol{\pi}^{\dagger}_{-i}}(s_1) = \operatorname*{arg\,max}_{\pi_i\in\Delta(\mathcal{A}_i)} \widehat{V}_{i,1}^{\pi_i\times\boldsymbol{\pi}^{\dagger}_{-i}}(s_1, G^{\star}),$$

which implies that $\boldsymbol{\pi}^{\dagger}$ is an NE of the regularized MG. Following similar argument, we can conclude that any NE of the regularized MG is an RNE of the robust MG. □

## F.3. Examples of Transition Uncertainty Sets

### F.3.1. Proof of Corollary 5.3

*Proof.* For ball constrained uncertainty set $\mathcal{P}_{s,h} = \left\{P\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}} : \left\|P - P_{s,h}^{\star}\right\|_{q^*\to p} \leq \beta_{s,h}\right\}$, we have

$$\begin{aligned} \sigma_{\mathcal{P}_{s,h}}\left(-v\mu^{\top}\right) &= \sup_{P_h\in\mathcal{P}_{s,h}} \langle P_h, -v\mu^{\top}\rangle \\ &= \langle P_h^{\star}, -v\mu^{\top}\rangle + \sup_{\left\|P_h'\right\|_{p\to q^*}\leq\beta_{s,h}} \langle P_h', -v\mu^{\top}\rangle \\ &= -\mathbb{E}_{\boldsymbol{a}\sim\mu}\left[[P_h^{\star}v](s,\boldsymbol{a})\right] + \beta_{s,h} \sup_{\left\|P_h'\right\|_{p\to q^*}\leq 1} \langle P_h', -v\mu^{\top}\rangle \\ &\overset{(I)}{=} -\mathbb{E}_{\boldsymbol{a}\sim\mu}\left[[P_h^{\star}v](s,\boldsymbol{a})\right] + \beta_{s,h}\left\|-v\right\|_p \left\|\mu^{\top}\right\|_q \\ &= \Omega_{s,h}\left(P_h^{\star}, -v, \mu\right) \end{aligned}$$

where equality $(I)$ follows from Lemma A.3 on the dual norm of matrix operator norm. The equivalence between the robust game and the regularized game immediately follows from Theorem 5.2. □

### F.3.2. Proof of Corollary 5.4

*Proof.* By Theorem 5.2, it is sufficient to verify that the support function $\sigma_{\mathcal{P}_{s,h}}(\cdot)$ is equivalent to the specified regularizer function $\Omega_{s,h}$ for each uncertainty set. When the transition uncertainty set is $(s,a)$-rectangular with the form $\mathcal{P} =$

$\times_{(s,\boldsymbol{a},h)\in\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{P}_{s,\boldsymbol{a},h}$, for each $s\in\mathcal{S}, h\in[H]$, the support function $\sigma_{\mathcal{P}_{s,h}}$ can be simplified. In particular, for each $v\in\mathbb{R}^{\mathcal{S}}, \mu\in\Delta(\mathcal{A})$, we have

$$\begin{aligned}\sigma_{\mathcal{P}_{s,h}}(-v\mu^\top) &= \sup_{P\in\mathcal{P}_{s,h}}\langle P, -v\mu^\top\rangle \\ &= \sup_{P=\{P_{\boldsymbol{a}'}\}_{\boldsymbol{a}'\in\mathcal{A}}:P_{\boldsymbol{a}'}\in\mathcal{P}_{s,\boldsymbol{a}',h}} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\langle P_{\boldsymbol{a}}, -v\rangle\right] \\ &\overset{(I)}{=} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\sup_{P_{\boldsymbol{a}}\in\mathcal{P}_{s,\boldsymbol{a},h}}\langle P_{\boldsymbol{a}}, -v\rangle\right] \\ &\overset{(II)}{=} \mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\sigma_{\mathcal{P}_{s,\boldsymbol{a},h}}(-v)\right], \end{aligned} \tag{27}$$

where $(I)$ holds due to the independence of uncertainty sets $\{\mathcal{P}_{s,\boldsymbol{a},h}\}_{\boldsymbol{a}\in\mathcal{A}}$ and $(II)$ follows from the definition of support function. □

#### F.3.3. Examples of $\mathcal{S}\times\mathcal{A}$-rectangular Transition Uncertainty Sets

We first formally introduce various distance metric for distributions.

1. Total variation distance: for any $\eta,\nu\in\Delta(\mathcal{S})$, $d_{\text{TV}}(\eta,\nu)=\frac{1}{2}\|\eta-\nu\|_1=\frac{1}{2}\sum_s|\eta(s)-\nu(s)|$.
2. Kullback-Leibler (KL) distance: for any $\eta,\nu\in\Delta(\mathcal{S})$, $d_{\text{KL}}(\eta,\nu)=\sum_{s\in\mathcal{S}}\eta(s)\log\frac{\eta(s)}{\nu(s)}$.
3. Chi-square distance: for any $\eta,\nu\in\Delta(\mathcal{S})$, $d_{\chi^2}(\eta,\nu)=\sum_s\frac{(\eta(s)-\nu(s))^2}{\nu(s)}$.
4. Wasserstein distance: consider $p$-Wasserstein metric w.r.t. to a metric $\rho(\cdot)$, i.e., for any $\eta,\nu\in\Delta(\mathcal{S})$, $d_{W_p}(\eta,\nu):=\left(\inf_{\gamma\sim\Gamma(\eta,\nu)}\mathbb{E}_{(x,y)\sim\gamma}\left[\rho(x,y)^p\right]\right)^{1/p}$, where $\Gamma(\eta,\nu)$ denotes the set of all couplings of $\eta$ and $\nu$.

**Corollary F.1.** *Consider a robust MG $(\mathcal{S},\{\mathcal{A}_i\}_{i\in[N]},P^\star,\boldsymbol{r}^\star,H,\mathcal{U})$ with uncertainty set $\mathcal{U}=\mathcal{P}\times\{\boldsymbol{r}^\star\}$, where $\mathcal{P}$ is $(s,a)$-rectangular.*

1. *If $\mathcal{P}$ is a KL uncertainty set given by $\mathcal{P}=\times_{(s,\boldsymbol{a},h)\in\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{P}^{KL}_{s,\boldsymbol{a},h}$ with $\mathcal{P}^{KL}_{s,\boldsymbol{a},h}=\left\{P\in\Delta(\mathcal{S}):d_{KL}(P,P^\star_{s,\boldsymbol{a},h})\le\beta_{s,\boldsymbol{a},h}\right\}$, then the equivalent policy-value regularized MG $(\mathcal{S},\{\mathcal{A}_i\}_{i\in[N]},P^\star,\boldsymbol{r}^\star,H,\Omega)$ is associated with the following convex regularizer: $\forall s\in\mathcal{S}, v\in\mathbb{R}^{\mathcal{S}},\mu\in\Delta(\mathcal{A})$,*
$$\Omega_{s,h}\left(P^\star_h,-v,\mu\right)=\mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\min_{\lambda\ge0}\left\{\beta_{s,\boldsymbol{a},h}\lambda+\lambda\log\left((P^\star_{s,\boldsymbol{a},h})^\top\exp\left(-\frac{v}{\lambda}\right)\right)\right\}\right]. \tag{28}$$
2. *If $\mathcal{P}$ is a Chi-square uncertainty set given by $\mathcal{P}=\times_{(s,\boldsymbol{a},h)\in\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{P}^{\chi^2}_{s,\boldsymbol{a},h}$ with $\mathcal{P}^{\chi^2}_{s,\boldsymbol{a},h}=\left\{P\in\Delta(\mathcal{S}):d_{\chi^2}(P,P^\star_{s,\boldsymbol{a},h})\le\beta_{s,\boldsymbol{a},h}\right\}$, then the equivalent policy-value regularized MG $(\mathcal{S},\{\mathcal{A}_i\}_{i\in[N]},P^\star,\boldsymbol{r}^\star,H,\Omega)$ is associated with the following convex regularizer: $\forall s\in\mathcal{S}, v\in\mathbb{R}^{\mathcal{S}},\mu\in\Delta(\mathcal{A})$,*
$$\Omega_{s,h}\left(P^\star_h,-v,\mu\right)=\mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\max_{u\ge\mathbf{0}}\left\{(P^\star_{s,\boldsymbol{a},h})^\top(v-u)-\sqrt{\beta_{s,\boldsymbol{a},h}\operatorname{Var}_{P^\star_{s,\boldsymbol{a},h}}(v-u)}\right\}\right]; \tag{29}$$
3. *If $\mathcal{P}$ is a total variation uncertainty set given by $\mathcal{P}=\times_{(s,\boldsymbol{a},h)\in\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{P}^{TV}_{s,\boldsymbol{a},h}$ with $\mathcal{P}^{TV}_{s,\boldsymbol{a},h}=\left\{P\in\Delta(\mathcal{S}):d_{TV}(P,P^\star_{s,\boldsymbol{a},h})\le\beta_{s,\boldsymbol{a},h}\right\}$, then the equivalent policy-value regularized MG $(\mathcal{S},\{\mathcal{A}_i\}_{i\in[N]},P^\star,\boldsymbol{r}^\star,H,\Omega)$ is associated with the following convex regularizer: $\forall s\in\mathcal{S}, v\in\mathbb{R}^{\mathcal{S}},\mu\in\Delta(\mathcal{A})$,*
$$\Omega_{s,h}\left(P^\star_h,-v,\mu\right)=\mathbb{E}_{\boldsymbol{a}\sim\mu}\left[-[P^\star_h v](s,\boldsymbol{a})+\frac{\beta_{s,\boldsymbol{a},h}}{2}\min_{u\ge\mathbf{0}}\left\{\max_{s'}\left(v(s')-u(s')\right)-\max_{s}\left(v(s)-u(s)\right)\right\}\right]; \tag{30}$$
4. *Wasserstein uncertainty set: If the uncertainty set $\mathcal{P}$ is $\mathcal{S}\times\mathcal{A}$-rectangular given by $\mathcal{P}=\times_{(s,\boldsymbol{a},h)\in\mathcal{S}\times\mathcal{A}\times[H]}\mathcal{P}^{W_p}_{s,\boldsymbol{a},h}$ with $\mathcal{P}^{W_p}_{s,\boldsymbol{a},h}=\left\{P\in\Delta(\mathcal{S}):d_{W_p}(P,P^\star_{s,\boldsymbol{a},h})\le\beta_{s,\boldsymbol{a},h}\right\}$,*
$$\Omega_{s,h}\left(P^\star_h,-v,\mu\right)=\mathbb{E}_{\boldsymbol{a}\sim\mu}\left[\inf_{\lambda\ge0}\left\{\lambda\beta_{s,\boldsymbol{a},h}+\mathbb{E}_{\tilde{s}\sim P^\star_{s,\boldsymbol{a},h}}\left[\sup_{s'\in\mathcal{S}}\left\{-v(s')-\lambda\rho(\tilde{s},s')\right\}\right]\right\}\right]. \tag{31}$$

*Proof.* By Corollary 5.4, it is sufficient to analyze the support function $\sigma_{\mathcal{P}_{s,\boldsymbol{a},h}}$ for each case.

1. KL uncertainty set: From the strong duality result on KL constrained set (Iyengar, 2005, Lemma 4.1) the optimization problem in the support function $\sigma_{\mathcal{P}_{s,\boldsymbol{a},h}}(-v)$ is equivalent to

$$\min_{\lambda \geq 0} \left\{ \beta_{s,\boldsymbol{a},h}\lambda + \lambda \log \left( (P^\star_{s,\boldsymbol{a},h})^\top \exp\left(-\frac{v}{\lambda}\right) \right) \right\}, \tag{32}$$

   which is convex in $\lambda$ and can be solved efficiently. By (27), we have $\sigma_{\mathcal{P}_{s,h}}(-v\mu^\top) = \Omega_{s,h}\left(P^\star_h, -v, \mu\right)$ with $\Omega_{s,h}$ defined in (28).

2. Chi-square uncertainty set: From the strong duality result on $\chi^2$-distance constrained set (Iyengar, 2005, Lemma 4.2), the optimization problem in the support function $\sigma_{\mathcal{P}_{s,\boldsymbol{a},h}}(-v)$ is equivalent to

$$\min_{u \geq \mathbf{0}} \left\{ -(P^\star_{s,\boldsymbol{a},h})^\top (v-u) + \sqrt{\beta_{s,\boldsymbol{a},h} \operatorname{Var}_{P^\star_{s,\boldsymbol{a},h}}(v-u)} \right\}, \tag{33}$$

   where $\operatorname{Var}_{P^\star_{s,\boldsymbol{a},h}}(v-u) = (P^\star_{s,\boldsymbol{a},h})^\top (v-u)^2 - \left( (P^\star_{s,\boldsymbol{a},h})^\top (v-\mu) \right)^2$ and the convex optimization problem (33) can be solved with complexity $\mathcal{O}(|\mathcal{S}| \log |\mathcal{S}|)$. By (27), we have $\sigma_{\mathcal{P}_{s,h}}(-v\mu^\top) = \Omega_{s,h}\left(P^\star_h, -v, \mu\right)$ with $\Omega_{s,h}$ defined in (29).

3. TV uncertainty set: From the strong duality result on TV-constrained set (Iyengar, 2005, Lemma 4.3), the optimization problem in the support function $\sigma_{\mathcal{P}_{s,\boldsymbol{a},h}}(-v)$ is equivalent to

$$-(P^\star_{s,\boldsymbol{a},h})^\top v + \frac{\beta_{s,\boldsymbol{a},h}}{2} \min_{u \geq \mathbf{0}} \left\{ \max_{s'} \left( v(s') - u(s') \right) - \max_{s'} \left( v(s') - u(s') \right) \right\}, \tag{34}$$

   which can be solved with complexity $\mathcal{O}(|\mathcal{S}| \log |\mathcal{S}|)$. By (27), we have $\sigma_{\mathcal{P}_{s,h}}(-v\mu^\top) = \Omega_{s,h}\left(P^\star_h, -v, \mu\right)$ with $\Omega_{s,h}$ defined in (30).

4. Wasserstein uncertainty set: From the strong duality result (Blanchet & Murthy, 2019, Theorem 1), it holds that

$$\sigma_{\mathcal{P}_{s,\boldsymbol{a},h}}(-v) = \sup_{P \in \mathcal{P}^{W_p}_{s,\boldsymbol{a},h}} \mathbb{E}_P[-v] = \inf_{\lambda \geq 0} \left\{ \lambda \beta_{s,\boldsymbol{a},h} + \mathbb{E}_{\tilde{s} \sim P^\star_{s,\boldsymbol{a},h}} \left[ \sup_{s' \in \mathcal{S}} \left\{ -v(s') - \lambda \rho(\tilde{s}, s') \right\} \right] \right\},$$

   which yields the equivalent regualrizer function $\Omega_{s,h}$ defined in (30).

□

### F.4. Proof of Theorem 5.5

*Proof.* Given a general-sum game $(A, B)$ we construct a transition-uncertain RMG that recovers the same properties as the reward-uncertain RMG from the proof of Theorem 4.1. Then, the proof of hardness follows exactly as before.

We define $\underline{r}$ and $\overline{r}$ exactly as before. Then, we define the nominal reward model by $r^\star_1(s_1, \boldsymbol{a}) = A(\boldsymbol{a}) - \underline{r}(\boldsymbol{a})$, $r^\star_2(s_1, \boldsymbol{a}) = \underline{r}(\boldsymbol{a})$, and $r^\star_2(s_2, \boldsymbol{a}) = \overline{r}(\boldsymbol{a})$. Here, we use the subscript to denote the time step, not the player number since the game is zero-sum. Lastly, we define $\mathcal{P}_{s,a} = \Delta(\mathcal{S})$ to allow all possible transitions. Note, that we only need to define the transition uncertainty for the first step since $H = 2$. We assume the start state is $s_1$ so let $r^\star_2(s_1, \boldsymbol{a})$ be arbitrary.

It is then clear that the worst model for the first player deterministically sends it to state $s_1$, which yields a reward of $\pi_1^\top (r^\star + \underline{r})\pi_2$ as before. Similarly, the worst model for the second player deterministically sends it to state $s_2$ which yields a reward of $\pi_1^\top (-r^\star - \underline{r})\pi_2$ as before. Thus, the earlier proof then applies and shows the problem is PPAD-hard to solve.

□