Noname manuscript No.
(will be inserted by the editor)

Robustness and sample complexity of model-based MARL for general-sum Markov games

Jayakumar Subramanian · Amit Sinha · Aditya Mahajan

the date of receipt and acceptance should be inserted later

Abstract Multi-agent reinfocement learning (MARL) is often modeled using the framework of Markov games (also called stochastic games or dynamic games). Most of the existing literature on MARL concentrates on zero-sum Markov games but is not applicable to general-sum Markov games. It is known that the best-response dynamics in general-sum Markov games are not a contraction. Therefore, different equilibrium in general-sum Markov games can have different values. Moreover, the Q-function is not sufficient to completely characterize the equilibrium. Given these challenges, model based learning is an attractive approach for MARL in general-sum Markov games. In this paper, we investigate the fundamental question of *sample complexity* for model-based MARL algorithms in general-sum Markov games and show that $\tilde{\mathcal{O}}(|\mathcal{S}|\,|\mathcal{A}|(1-\gamma)^{-2}\alpha^{-2})$ samples are sufficient to obtain a α-approximate Markov perfect equilibrium with high probability, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the joint action space of all players, and γ is the discount factor, and the $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic terms. To obtain these results, we study the robustness of Markov perfect equilibrium to model approximations. We show that the Markov perfect equilibrium of an approximate (or perturbed) game is always an approximate Markov perfect equilibrium of the original game and

The work of Amit Sinha and Aditya Mahajan was supported in part by the Innovation for Defence Excellence and Security (IDEaS) Program of the Canadian Department of National Defence through grant CFPMN2-30.

A preliminary version of this work will appear in the 2021 Indian Control Conference.

J. Subramanian
Media and Data Science Research Lab, Digital Experience Cloud,
Adobe Inc., Noida, Uttar Pradesh, India.
E-mail: jasubram@adobe.com

A. Sinha and A. Mahajan
Department of Electrical and Computer Engineering,
McGill University, Montreal, Canada.
E-mail: amit.sinha@mail.mcgill.ca, aditya.mahajan@mcgill.ca

provide explicit bounds on the approximation error. We illustrate the results via a numerical example.

1 Introduction

Markov games (also called stochastic games or dynamic games) are a commonly used framework to model strategic interaction between multiple players interacting in a dynamic environment. Examples include applications in cyber-security (Sengupta et al. 2019), industrial organization (Ericson and Pakes 1995), political economics (Acemoglu and Robinson 2001), and many others (Başar and Zaccour 2018). Starting with the seminal work of Shapley (1953), several variations of Markov games have been considered in the literature. We refer the reader to Filar and Vrieze (1996) for an overview.

Overview of Markov games. In the basic setup of a dynamic game, the payoffs of players at any time not only depends on their current joint action profile but also on the current "state of the system". Furthermore, the state of the system evolves in a controlled Markov manner conditioned on the current action profile of the players. It is typically assumed that the state of the system and the action profile of all players are publicly monitored by all players. Although Markov games may be viewed as a special case of extensive form games with perfect information, rather than using the standard solution concept of sub-game perfect equilibrium, attention is often restricted to a refinement of sub-game perfect equilibrium called Markov perfect equilibrium (MPE) where all players play Markov strategies (i.e., choose their actions as a (possibly randomized) function of the current state) (Maskin and Tirole 1988a,b). MPE is an attractive refinement of sub-game perfect equilibrium, both from a computational as well as conceptual point of view, but has some limitations because it excludes some history dependent strategies (such as tit-for-tat and grim trigger) commonly used in the repeated games setup. See Maskin and Tirole (2001); Mailath and Samuelson (2006) for a discussion.

Games can also be classified based on the sum of per-step payoffs of players as zero-sum or general-sum games. The nature of results in these two cases are different as are the tools used to prove them. The differences stem from the fact that the best response mapping (called the Shapley operator) for two-player zero-sum games is a contraction (Shapley 1953). Therefore, zero-sum games have a unique value (i.e., all equilibria in zero-sum games have the same value). Moreover, the MPE (also called minimax equilibrium for the zero-sum case) can be computed via recursive operations of the Shapley operator (Shapley 1953; Hoffman and Karp 1966). In contrast, the best response mapping for general-sum games is not a contraction. Therefore, the existence of MPE needs to be proved using variations of Kakutani's fixed point theorem (Fink 1964; Takahashi 1964; Rogers 1969; Vrieze 1987). A consequence of this is that, in general, different MPEs do not have the same value, which makes it difficult to compute MPE. Various algorithms have been proposed to compute MPE, including non-linear programming (Filar et al. 1991) and homotopy methods (Herings et al. 2004; Herings and Peeters 2010).

Multi-agent reinforcement learning. In recent years, there has been significant interest in understanding interaction between strategic agents operating in unknown environments. Such multi-player problems are studied under the heading of multi-agent reinforcement learning (MARL) and often modeled as Markov games (Littman 1994; Busoniu et al. 2008; Zhang et al. 2021a). Although there have been significant recent successes in single agent RL, these do not directly translate into the multi-agent setting. Part of the difficulty is that when multiple agents are learning simultaneously, the "environment" as viewed by any single agent is non-stationary (Busoniu et al. 2008); so it is not possible to use the theoretical guarantees of single agent RL algorithms, which are derived for a stationary or time-homogeneous environment.

Nonetheless, MARL for two player zero-sum games is well understood due to two properties. First, if two strategies (π^1, π^2) and (μ^1, μ^2) are minimax equilibrium, then so are strategies (π^1, μ^2) and (μ^1, π^2). Therefore, to identify equilibrium strategies, it is sufficient to learn the action-value function (i.e., the Q-function). Second, the action-value function can be learnt using variants of Q-learning (called minimax Q-learning) because the Shapley operator is a contraction (Littman 1994, 2001). We refer the reader to Shoham et al. (2003) for an overview of MARL for zero-sum games.

However, the situation is different for general-sum MARL, where fewer convergence guarantees are available. Part of the difficulty is that the action-value function (or Q-function) is insufficient to characterize MPE (Zinkevich et al. 2006, Theorem 1).[1] For this reason, algorithms developed for two-player zero-sum games fail to converge to an MPE in general-sum games (Pérolat et al. 2017). There are some partial results, e.g., minimizing Bellman residual error to identify ε-MPE (Pérolat et al. 2017), using two-time scale stochastic approximation algorithms (Prasad et al. 2015), and using replicator dynamics based algorithms (Akchurina 2010). However, in general, developing MARL algorithms with convergence guarantees remains a challenging research direction.

Model based MARL, sample complexity, and robustness of equilibria. One potential approach to alleviate the difficulties in MARL for general-sum games is to use model based algorithms, which explicitly learn (or estimate) the system model and then use a "planning algorithm" to find the solution of the estimated model (Sutton 1990). There has been significant recent interest in model based RL for single agent system (see Wang et al. (2019) and references therein) and some interest in model-based approaches

[1] Zinkevich et al. (2006) construct two player general-sum games with the following properties. The game has two states: in state 1, player 1 has two actions and player 2 has one action; in state 2, player 1 has one action and player 2 has two actions. The transition probabilities are chosen such that there is a unique Markov perfect equilibrium in mixed strategies. This means that in state 1, both actions of player 1 maximize the Q-function; in state 2, both actions of player 2 minimize the Q-function. However, the Q-function in itself is insufficient to determine the randomizing probabilities for the mixed strategy MPE.

for MARL for zero-sum games (Krupnik et al. 2019; Sidford et al. 2020; Zhang et al. 2021b, 2020). However, as far as we are aware, there are no model based MARL algorithms for general-sum Markov games.

An important consideration in model-based RL is to determine how many samples are needed to identify an α-approximate solution (for a pre-specified accuracy level α). This is known as *sample complexity* of learning and is typically analyzed under the assumption that the learning agent has access to a generative model, i.e., a black box simulator that takes the current state and action profile as input and generates samples of the next state as output.

Starting with the work of Kearns and Singh (1999); Kakade (2003), there is an extensive literature on the sample complexity of Markov decision processes (MDPs) (Azar et al. 2013; Sidford et al. 2018; Agarwal et al. 2020; Li et al. 2020). The simplest approach in this setting is to use a plug-in estimator,[2] i.e., estimating the transition matrix using the generated samples and using the optimal policy corresponding to the estimated model in the true system. Recent results of Agarwal et al. (2020) show that the sample complexity of the plug-in estimator matches the lower bounds on sample complexity (Azar et al. 2013) modulo logarithmic factors. Recently, Zhang et al. (2020), build on this line of work to establish sample complexity bounds for zero-sum games. As far as we are aware, sample complexity of generative models for general-sum games hasn't been investigated before.

The analyses of model-based RL algorithms rely on the *robustness* of the "planning solution" to model approximations, i.e., *if the estimated model is close to the true model in some sense, does that imply that the strategy generated from the estimated model is approximately appropriate in some sense (optimality, equilibrium, etc.)?* This question is well understood for Markov decision processes (see Müller (1997) and follow-up work) and zero-sum Markov games (Tidball and Altman 1996; Tidball et al. 1997). In this paper, we address the question of robustness for general-sum Markov games. In particular, we show that if a dynamic game is approximated by another game such that the reward functions and transitions of the approximate game are close to those of the original game (in an appropriate sense), then a MPE of the approximate game is an approximate MPE of the original game. We quantify the exact relationship between the degree of approximation of the games and the approximation error in the MPE.

We then build up on these results to establish sample complexity bounds for learning with a generative model for general-sum Markov games. Our bound does not match the lower bound on sample complexity (obtained by translating the lower bound bound of Azar et al. (2013) for MDPs to games). However, we believe that this is a limitation of our proof technique. It might be possible to obtain tighter bounds by extending the variance reduction technique of Azar et al. (2013); Agarwal et al. (2020); Li et al. (2020) to general-sum games. See Sec. 6 for a more detailed discussion.

[2] The plug-in estimator is also known as a certainty equivalent controller in the stochastic control literature.

Other notions of robustness. Our notion of robustness is different from that of robust control (Başar and Bernhard 2008) and robust Markov perfect equilibrium (Jaśkiewicz and Nowak 2014), both of which are Markov decision processes with uncertain dynamics and are treated as zero-sum games where nature acts as an adversary and picks the worst-case realization of the transition dynamics. Our notion of robustness is also different from uniformly ε-equilibrium (Solan 2021), which captures robustness with respect to time-horizon and discount factor.

Our notion of robustness is similar in spirit to robust MPEs considered in Maskin and Tirole (2001), who defined a MPE to be robust if for any small perturbation of the payoffs, there exists a nearby MPE. Maskin and Tirole (2001) showed that almost all finite horizon games have a finite number of MPEs, all of which are robust. Our results are of a different nature and it is difficult to compare the two results because Maskin and Tirole (2001) considered an atypical model where the states are not specified exogenously but are rather determined as the payoff relevant component of the history. Consequently, perturbing the payoffs changes the state *space*, which is not the case for our model.

Related to Maskin and Tirole (2001) is the notion of strong stability considered in Doraszelski and Escobar (2010), using which it is shown that for almost all Markov games have finite number of MPEs and these equilibria can be approximated by equilibria of nearby games. The dynamics in Doraszelski and Escobar (2010) are exogenous and, therefore, their result does not have the same limitations as that of Maskin and Tirole (2001). The result of Doraszelski and Escobar (2010) is stronger than ours because we only show that equilibria of nearby games are approximate equilibria of the original game but we do not establish that they are also close to the equilibria of the original game. However, the results of Doraszelski and Escobar (2010) rely on continuity arguments and do not explicitly characterize bounds on the size of the neighborhood. In contrast, for any ε perturbation in payoffs and δ perturbation in dynamics, we explicitly characterize an α such that the MPE of the perturbed game is an α-MPE of the original game.

Perhaps the result most similar to ours is Whitt (1980), who consider a more general model and allow the approximate game to have a different state and action space than the original game. Their main result is to show that any α_1-MPE of the approximate game is an α_2-MPE of the original game and an explicit relationship between α_1 and α_2 is established. Our results are similar in spirit but the specific details are different.

Organization. The rest of the paper is organized as follows. In Sec. 2, we present our notion of approximation of a dynamic game and state our main results. In Sec. 3, we present background results on approximation of Markov decision processes. In Sec. 4, we provide the proof of our main results. We conclude in Sec. 6.

Notation. We use $\mathbb{R}$ to denote the set of real numbers, $\mathbb{P}(\cdot)$ to denote the probability of an event, $\mathbb{E}[\cdot]$ to denote the expectation of a random variable, and $\mathscr{P}(\cdot)$ denotes the set of probability measures on a set.

We use calligraphic letters (e.g., $\mathcal{S}$, $\mathcal{A}$, etc.) to denote sets, uppercase letters (e.g., S, A, etc.) to denote random variables and lowercase letters (e.g., s, a, etc.) to denote their realization. Superscripts index players and subscripts index time. For example, a^i_t denotes the action of player i at time t. For sequence of variables $\{s_t\}_{t \ge 1}$, we use the short hand notation $s_{1:t}$ to denote the sequence $(s_1, \dots, s_t)$.

Given a function $f \colon \mathcal{S} \to \mathbb{R}$, we use $\operatorname{span}(f)$ to denote the span seminorm of f, i.e., $\operatorname{span}(f) = \sup_{s \in S} f(s) - \inf_{s \in S} f(s)$. Given a metric space $(\mathcal{S}, d)$ and a function $f \colon \mathcal{S} \to \mathbb{R}$, we use $\operatorname{Lip}(f)$ to denote the Lipschitz constant of f, i.e.,

$$\operatorname{Lip}(f) = \sup_{s, s' \in \mathcal{S}} \frac{|f(s) - f(s')|}{d(s, s')}.$$

2 System model, robustness, and sample complexity

For ease of exposition, we restrict the discussion in this paper to models with finite state and action spaces. The results extend to models with continuous state and action spaces under standard technical assumptions on the existence of equilibria in that setting.

2.1 Dynamic games

An infinite horizon dynamic game (also called stochastic game or Markov game) is a tuple $\langle \mathcal{N}, \mathcal{S}, (\mathcal{A}^i)_{i \in \mathcal{N}}, \mathsf{P}, (r^i)_{i \in \mathcal{N}}, \gamma \rangle$ where:

- $\mathcal{N}$ is the (finite) set of players.
- $\mathcal{S}$ is the (finite) set of possible states of the game. We use $S_t \in \mathcal{S}$ to denote the state of the game at time t.
- $(\mathcal{A}^i)_{i \in \mathcal{N}}$ is the (finite) set of actions available to player i at each time. we also use $\mathcal{A} = \prod_{i \in \mathcal{N}} \mathcal{A}^i$ to denote the set of actions of all players. We use $A_t = (A^i_t)_{i \in \mathcal{N}}$ to denote the action profile of all players at time t. Given an action profile $A_t = (A^i_t)_{i \in \mathcal{N}}$ and a player $j \in \mathcal{N}$, we use the notation $A^{-j}_t = (A^i_t)_{i \in \mathcal{N} \setminus \{j\}}$ to denote the action profile of all players except j.
- $\mathsf{P} : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is the controlled transition probability of the state of the game. In particular, at any time t, given a realization $s_{1:t+1}$ of $S_{1:t+1}$ and choice of action profile $a_{1:t}$ of $A_{1:t}$, we have

$$\mathbb{P}(S_{t+1} = s_{t+1} \mid S_{1:t} = s_{1:t}, A_{1:t} = a_{1:t}) = \mathsf{P}(s_{t+1} \mid s_t, a_t).$$

- $r^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the per-step reward of player i.
- $\gamma \in (0, 1)$ is the discount factor.

We assume that all players have perfect monitoring. At time t, all players observe the current state S_t and simultaneously choose their respective actions. At the end of time period t, all players observe all the actions, and the state of the game evolves according to the transition kernel P.

Following Shapley (1953); Maskin and Tirole (1988a,b), we assume that each player chooses its action according to a time homogeneous Markov strategy. Let

$$\Pi^i = \{\pi^i : \mathcal{S} \to \mathscr{P}(\mathcal{A}^i)\}$$

denote the set of all Markov strategies for player i.

Given a strategy profile $\pi = (\pi^i)_{i\in\mathcal{N}}$, where $\pi^i \in \Pi^i$, and an initial state s_0, the expected discounted total reward of player i is given by:

$$V^i_{(\pi^i,\pi^{-i})}(s_0) = (1-\gamma)\mathbb{E}_{(\pi^i,\pi^{-i})}\left[\sum_{t=0}^{\infty}\gamma^t r^i(S_t, A_t) \;\middle|\; S_0 = s_0\right], \tag{1}$$

where the expectation is with respect to the joint measure on all the system variables induced by the choice of the strategy profile of all players.

Although the above model is formulated for infinite horizon, it can capture interactions for finite horizon by considering time as part of the state space and by assuming that, at the end of the horizon, the game moves to an absorbing state with zero rewards for all players. In the special case when the game has a single state, a dynamic game is equivalent to an infinitely repeated matrix game. In the special case when the game has only one player, a dynamic game is equivalent to a Markov decision process.

Remark 1 We follow the standard game theoretic convention of normalizing the expected total reward by pre-multiplying by $(1-\gamma)$. An immediate implication of this is that for any strategy π, $\|V^i_\pi\|_\infty \le \|r\|_\infty$. In some of the AI literature, the expected reward is not normalized. In such cases $\|V^i_\pi\|_\infty \le \|r\|_\infty/(1-\gamma)$.

There are two solution concepts commonly used for Markov games, which we state below.

Definition 1 (Markov perfect equilibrium) A Markov strategy profile $\pi = (\pi^i)_{i\in\mathcal{N}}$, where $\pi^i \in \Pi^i$, is called a Markov perfect equilibrium (MPE) if for every initial state $s \in \mathcal{S}$, and every player $i \in \mathcal{N}$,

$$V^i_{(\pi^i,\pi^{-i})}(s) \ge V^i_{(\tilde{\pi}^i,\pi^{-i})}(s), \quad \forall \tilde{\pi}^i \in \Pi^i. \tag{2}$$

A Markov perfect equilibrium can be viewed as a refinement of subgame perfect equilibrium where all players play Markov strategies (Maskin and Tirole 1988a,b). For games with finite state and action spaces, a Markov perfect equilibrium always exists (Fink 1964; Rogers 1969; Vrieze 1987; Filar and Vrieze 1996). For general state and action spaces, see Takahashi (1964).

Definition 2 (Approximate Markov perfect equilibrium) Given approximation level $\alpha = (\alpha^i)_{i\in\mathcal{N}}$, where α^i are positive constants, a strategy profile $\pi = (\pi^i)_{i\in\mathcal{N}}$, where $\pi^i \in \Pi^i$, is called an α-approximate Markov perfect equilibrium (α-MPE) if for every initial state $s \in \mathcal{S}$, and every player $i \in \mathcal{N}$,

$$V^i_{(\pi^i,\pi^{-i})}(s) \geq V^i_{(\tilde{\pi}^i,\pi^{-i})}(s) - \alpha^i, \quad \forall \tilde{\pi}^i \in \Pi^i. \tag{3}$$

When all α^i are identical and equal to say α', we simply call the approximate MPE an α'-MPE rather than $(\alpha', \dots, \alpha')$-MPE.

2.2 Integral probability metrics

Our results rely on a class of metrics on probability spaces known as integral probability metrics (IPMs) (Müller 1997).

Definition 3 Let $(\mathcal{X}, \mathscr{G})$ be a measurable space and $\mathfrak{F}$ denote a class of uniformly bounded measurable functions on $(\mathcal{X}, \mathscr{G})$. The integral probability metric (IPM) between two probability distributions $\mu, \nu \in \mathscr{P}(\mathcal{X})$ with respect to the function class $\mathfrak{F}$ is defined as

$$d_{\mathfrak{F}}(\mu, \nu) \coloneqq \sup_{f\in\mathfrak{F}} \left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right|.$$

Two specific forms of IPMs are used in this paper:

1. **Total variation distance**: If $\mathfrak{F}$ is chosen as $\mathfrak{F}^{\mathrm{TV}} \coloneqq \{f : \operatorname{span}(f) \leq 1\}$, then $d_{\mathfrak{F}}$ is the total variation distance[3].
2. **Wasserstein distance**: If $\mathcal{X}$ is a metric space and $\mathfrak{F}$ is chosen as $\mathfrak{F}^{\mathrm{W}} \coloneqq \{f : \operatorname{Lip}(f) \leq 1\}$ (where the Lipschitz constant is computed with respect to the metric on $\mathcal{X}$), then $d_{\mathfrak{F}}$ is the Wasserstein distance.

Our approximation results are stated in terms of the Minkowski functional of a function f (not necessarily in $\mathfrak{F}$) with respect to a function class $\mathfrak{F}$, which is defined as follows:

$$\rho_{\mathfrak{F}}(f) \coloneqq \inf\{\rho \in \mathbb{R}_{>0} : \rho^{-1} f \in \mathfrak{F}\}. \tag{4}$$

[3] If μ and ν are absolutely continuous with respect to some measure λ and let $p = d\mu/d\lambda$ and $q = d\nu/d\lambda$, then total variation is typically defined as $\frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)|\lambda(dx)$. This is consistent with our definition. Let $\bar{f} = (\sup f + \inf f)/2$. Then

$$\begin{aligned} \left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right| &= \left| \int_{\mathcal{X}} f(x) p(x) \lambda(dx) - \int_{\mathcal{X}} f(x) q(x) \lambda(dx) \right| \\ &= \left| \int_{\mathcal{X}} [f(x) - \bar{f}] \, [p(x) - q(x)] \, \lambda(dx) \right| \leq \|f - \bar{f}\|_\infty \int_{\mathcal{X}} \big|p(x) - q(x)\big| \lambda(dx) \\ &\leq \tfrac{1}{2} \operatorname{span}(f) \int_{\mathcal{X}} \big|p(x) - q(x)\big| \lambda(dx). \end{aligned}$$

A key implication of this definition is that for any function f,

$$\left| \int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu \right| \leq \rho_{\mathfrak{F}}(f) \cdot d_{\mathfrak{F}}(\mu, \nu), \tag{5}$$

The Minkowski functional of the two IPMs considered in this paper are as follows:

1. **Total variation distance**: If $\mathfrak{F}$ is chosen as $\mathfrak{F}^{\mathrm{TV}}$, $\left|\int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu\right| \leq \mathrm{span}(f) d_{\mathfrak{F}}(\mu, \nu)$. Thus, for total variation, $\rho_{\mathfrak{F}^{\mathrm{TV}}}(f) = \mathrm{span}(f)$.
2. **Wasserstein distance**: If $\mathfrak{F}$ is chosen as $\mathfrak{F}^{\mathrm{W}}$, $\left|\int_{\mathcal{X}} f d\mu - \int_{\mathcal{X}} f d\nu\right| \leq \mathrm{Lip}(f) \cdot d_{\mathfrak{F}}(\mu, \nu)$. Thus, for the Wasserstein distance, $\rho_{\mathfrak{F}^{\mathrm{W}}}(f) = \mathrm{Lip}(f)$.

2.3 Approximate game

Definition 4 Given a function class $\mathfrak{F}$ and positive constants (ε, δ), a game $\widehat{\mathscr{G}} = \langle \mathcal{N}, \mathcal{S}, (\mathcal{A}^i)_{i\in\mathcal{N}}, \widehat{\mathsf{P}}, (\hat{r}^i)_{i\in\mathcal{N}}, \gamma \rangle$ is an (ε, δ)-approximation of the game $\mathscr{G} = \langle \mathcal{N}, \mathcal{S}, (\mathcal{A}^i)_{i\in\mathcal{N}}, \mathsf{P}, (r^i)_{i\in\mathcal{N}}, \gamma \rangle$ if the following conditions are satisfied:

1. **Reward approximation**: For all $i \in \mathcal{N}$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$|r^i(s, a) - \hat{r}^i(s, a)| \leq \varepsilon. \tag{6}$$

2. **Transition approximation**: For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$d_{\mathfrak{F}}(\mathsf{P}(\cdot \mid s, a), \widehat{\mathsf{P}}(\cdot \mid s, a)) \leq \delta. \tag{7}$$

Our main result is the following.

Theorem 1 *If game $\widehat{\mathscr{G}}$ is an (ε, δ)-approximation of game $\mathscr{G}$ and $\hat{\pi}$ is an MPE of $\widehat{\mathscr{G}}$, then $\hat{\pi}$ is also an α-MPE of $\mathscr{G}$, where $\alpha = (\alpha^i)_{i\in\mathcal{N}}$ can be bounded as*

$$\alpha^i \leq 2\left(\varepsilon + \frac{\gamma \Delta^i_{(\hat{\pi}^i, \hat{\pi}^{-i})}}{1 - \gamma}\right), \quad \forall i \in \mathcal{N}, \tag{8}$$

where

$$\begin{aligned} \Delta^i_{(\hat{\pi}^i, \hat{\pi}^{-i})} &= \max_{s\in S, a\in A} \left| \sum_{s'\in S} \left[\mathsf{P}(s'|s, a) \hat{V}^i_{(\hat{\pi}^i, \hat{\pi}^{-i})}(s') - \widehat{\mathsf{P}}(s'|s, a) \hat{V}^i_{(\hat{\pi}^i, \hat{\pi}^{-i})}(s') \right] \right| \\ &\leq \delta \rho_{\mathfrak{F}}(\hat{V}^i_{(\hat{\pi}^i, \hat{\pi}^{-i})}). \end{aligned}$$

Therefore, a looser upper bound is given by

$$\alpha^i \leq 2\left(\varepsilon + \frac{\gamma \delta \rho_{\mathfrak{F}}(V^i_{(\hat{\pi}^i, \hat{\pi}^{-i})})}{1 - \gamma}\right), \quad \forall i \in \mathcal{N}, \tag{9}$$

See Sec. 4 for the proof.

Remark 2 Both upper bounds (8) and (9) are instance dependent. The bound (8) depends on the transition probability P of the original game, while the only information about the original game $\mathscr{G}$ needed in (9) are the modeling errors (ε, δ). The approximation bound (9) also depends on the choice of IPM used to measure the approximation in the dynamics.

It is possible to obtain instance-independent bounds by using worst case upper bounds on $\rho_{\mathfrak{F}}(\hat{V}^i_{\hat{\pi}})$. In particular, the Minkowski functional $\rho_{\mathfrak{F}}(\hat{V}^i_{\hat{\pi}})$ is $\operatorname{span}(\hat{V}^i_{\hat{\pi}})$ when using the total variation distance and is $\operatorname{Lip}(\hat{V}^i_{\hat{\pi}})$ when using the Wasserstein distance. Using worst-case upper bounds on $\operatorname{span}(\hat{V}^i_{\hat{\pi}})$ and $\operatorname{Lip}(\hat{V}^i_{\hat{\pi}})$ gives us the following bounds.

Corollary 1 *When $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$, then*

$$\alpha^i \le 2\left(\varepsilon + \frac{\gamma\delta \operatorname{span}(\hat{r}^i)}{(1-\gamma)}\right), \quad \forall i \in \mathcal{N}. \tag{10}$$

The next bound holds for games where the transition matrix and reward function are Lipschitz.

Definition 5 Suppose the state space $\mathcal{S}$ is a metric space with metric d. Then, a game $\mathscr{G}$ is said to be (L_r, L_{P})-Lipschitz if for any $i \in \mathcal{N}$, $s_1, s_2 \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$\left| r^i(s_1, a) - r^i(s_2, a) \right| \le L_r d(s_1, s_2),$$

and

$$d_{\mathfrak{F}^{\mathrm{W}}}(\mathsf{P}(\cdot|s_1, a), \mathsf{P}(\cdot|s_2, a)) \le L_{\mathsf{P}} d(s_1, s_2),$$

where $d_{\mathfrak{F}^{\mathrm{W}}}$ denotes the Wasserstein distance.

Corollary 2 *When $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$ and $\widehat{\mathscr{G}}$ is (L_r, L_{P})-Lipschitz with $\gamma L_{\mathsf{P}} < 1$, then*

$$\alpha^i \le 2\left(\varepsilon + \frac{\gamma L_r \delta}{(1-\gamma L_{\mathsf{P}})}\right), \quad \forall i \in \mathcal{N}. \tag{11}$$

Remark 3 Although we have only elaborated on two specific choices of IPMs (total variation and Wasserstein distances), the result of Theorem 1 is applicable for *any* IPM. Many other IPMs have been considered in the literature including Kolmogorov distance, bounded Lipschitz metric, and maximum mean discrepancy. See, for example, Müller (1997); Subramanian et al. (2020). The choice of the metric often depends on the specific properties of the model.

2.4 Model based RL

Now, we consider the setting where the components $\langle \mathcal{S}, (\mathcal{A}^i)_{i\in\mathcal{N}}, (r^i)_{i\in\mathcal{N}}, \gamma\rangle$ of a game $\mathscr{G}$ are known but the transition probability matrix P is not known. Suppose we have access to a *generative model*, i.e., a black-box simulator which provides samples $S_+ \sim \mathsf{P}(\cdot|s,a)$ of the next state S_+ for a given state-action pair (s,a) as input. Suppose we call the simulator n times at each state-action pair[4] and estimate an empirical model $\widehat{\mathsf{P}}_n$ as $\widehat{\mathsf{P}}_n(s'|s,a) = \text{count}(s'|s,a)/n$, where $\text{count}(s'|s,a)$ is the number of times s' is sampled with the input is (s,a). The game $\widehat{\mathscr{G}_n} = \langle \mathcal{S}, (\mathcal{A}^i)_{i\in\mathcal{N}}, \widehat{\mathsf{P}}_n, (r^i)_{i\in\mathcal{N}}, \gamma\rangle$ may be viewed as an approximation of game $\mathscr{G}$. Let $\hat{\pi}_n$ be an MPE $\widehat{\mathscr{G}_n}$.

Note that we are assuming that there is a system planner which generates samples from the generative model and computes the MPE of $\widehat{\mathscr{G}_n}$. A more interesting setting is where each player generates independent samples from the generative model and computes a different MPE. This setting is challenging due to the multiplicity of MPE and is not discussed in this paper.

One fundamental question in this setting is the following. Given an $\alpha > 0$, how many samples n are needed from the generative model to ensure that $\hat{\pi}_n$ is an α-MPE for game $\mathscr{G}$. This is called the *sample complexity* of learning. Below, we obtain a bound on sample complexity using the approximation bounds of Theorem 1.

Theorem 2 *For any $\alpha > 0$ and $p > 0$, let*

$$n \ge \left\lceil \left(\frac{\gamma}{1-\gamma}\,\text{span}(r)\right)^2 \frac{2\log(2|\mathcal{S}|\left(\prod_{i\in\mathcal{N}}|\mathcal{A}^i|\right)|\mathcal{N}|/p)}{\alpha^2} \right\rceil.$$

Then, with probability $1-p$, an MPE $\hat{\pi}_n$ of game $\widehat{\mathscr{G}_n}$ is α-MPE for game $\mathscr{G}$.

See Sec. 4 for the proof.

Remark 4 In general, game $\widehat{\mathscr{G}_n}$ may have multiple MPE. The result of Theorem 2 is true for *every* MPE of game $\widehat{\mathscr{G}_n}$.

3 Preliminaries on MDPs

3.1 MDP, Bellman Operators, and Dynamic Programming

A Markov Decision Process (MDP) is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma\rangle$ where

- $\mathcal{S}$ is the (finite) set of states of the environment. The state at time t is denoted by S_t.
- $\mathcal{A}$ is the (finite) set of actions available to the agent. The action at time t is denoted by A_t.

[4] Note that since n samples are generated for every state-action pair, the total number of calls to the simulator equal $n|\mathcal{S}|\,|\mathcal{A}|$.

- $\mathsf{P} : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathcal{S})$ is the controlled transition probability. For any realization $s_{1:t+1}$ of $S_{1:t+1}$ and choice $a_{1:t}$ of $A_{1:t}$, we have

$$\mathbb{P}(S_{t+1} = s_{t+1} | S_{1:t} = s_{1:t}, A_{1:t} = a_{1:t}) = \mathsf{P}(s_{t+1} | s_t, a_t). \tag{12}$$

- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the per-step reward function.
- $\gamma \in (0, 1)$ is the discount factor.

It is assumed that the agent observes the state S_t and chooses the action A_t according to a Markov strategy $\pi : \mathcal{S} \to \mathscr{P}(\mathcal{A})$. The performance of a Markov strategy π starting from initial state $s_0 \in \mathcal{S}$ is given by:

$$V_\pi(s_0) = (1 - \gamma)\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \;\middle|\; S_0 = s_0 \right], \tag{13}$$

where the expectation is with respect to the joint measure on the system variables induced by the choice of strategy π. A strategy π is called optimal if for any other Markov strategy $\tilde{\pi}$, we have

$$V_\pi(s) \geq V_{\tilde{\pi}}(s), \quad \forall s \in \mathcal{S}. \tag{14}$$

In addition, given a positive constant α, a strategy π is called α-optimal if

$$V_\pi(s) \geq V_{\tilde{\pi}}(s) - \alpha, \quad \forall s \in \mathcal{S}. \tag{15}$$

Given an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma \rangle$ and a Markov strategy π, define the Bellman operators $\mathcal{B}_\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ and $\mathcal{B}_* : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ as follows: for any $v \in \mathbb{R}^{|\mathcal{S}|}$ and $s \in \mathcal{S}$

$$[\mathcal{B}_\pi v](s) = \sum_{a \in \mathcal{A}} \pi(a|s) \Big[(1 - \gamma) r(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathsf{P}(s'|s, a) v(s') \Big], \tag{16}$$

$$[\mathcal{B}_* v](s) = \max_{a \in \mathcal{A}} \Big[(1 - \gamma) r(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathsf{P}(s'|s, a) v(s') \Big]. \tag{17}$$

Then, optimal and approximately optimal strategies can be characterized using the Bellman operators as shown below. These are standard results. See Bertsekas (2017), for example.

Proposition 1 *A Markov strategy π is optimal if and only if there exists a value function $V \in \mathbb{R}^{|\mathcal{S}|}$ such that*

$$V = \mathcal{B}_\pi V \quad \textit{and} \quad V = \mathcal{B}_* V. \tag{18}$$

Remark 5 Note that an MDP can have more than one optimal strategy but all optimal strategies have the same performance and hence the same value function.

Proposition 2 *Given a Markov strategy π, let V_π be the unique fixed point of $V_\pi = \mathcal{B}_\pi V_\pi$ and let V_* be the unique fixed point of $V_* = \mathcal{B}_* V_*$. Then, the strategy π is α-optimal if and only if*

$$V_\pi \geq V_* - \alpha. \tag{19}$$

Remark 6 A sufficient condition to verify (19) in Proposition 2 is that

$$V_\pi \geq \mathcal{B}_* V_\pi - \alpha. \tag{20}$$

We now present some basic properties of the value function which are used later.

Lemma 1 *If V is the optimal value function of MDP $\mathcal{M}$, then*

$$\operatorname{span}(V) \leq \operatorname{span}(r).$$

Proof This result follows immediately by observing that the per-step reward $r(S_t, A_t) \in [\min(r), \max(r)]$. Therefore, $\max(V) \leq \max(r)$ and $\min(V) \geq \min(r)$. □

We now define the notion of a Lipschitz MDP.

Definition 6 Let d be a metric on the state space $\mathcal{S}$. The MDP $\mathcal{M}$ is said to be (L_r, L_{P})-Lipschitz if for any $s_1, s_2 \in \mathcal{S}$ and $a \in A$, the reward function r and transition kernel P of $\mathcal{M}$ satisfy the following

$$\left| r(s_1, a) - r(s_2, a) \right| \leq L_r d(s_1, s_2),$$

and

$$d_{\mathfrak{F}^{\mathrm{W}}}(\mathsf{P}(\cdot|s_1, a), \mathsf{P}(\cdot|s_2, a)) \leq L_{\mathsf{P}} d(s_1, s_2),$$

where $d_{\mathfrak{F}^{\mathrm{W}}}$ denotes the Wasserstein distance.

Lemma 2 *If an MDP $\mathcal{M}$ is (L_r, L_{P})-Lipschitz and $\gamma L_{\mathsf{P}} < 1$, and V is the optimal value function of $\mathcal{M}$, then*

$$\operatorname{Lip}(V) \leq \frac{(1-\gamma) L_r}{1 - \gamma L_{\mathsf{P}}}.$$

Proof The result follows from (Hinderer 2005, Theorem 4.2). □

3.2 Robustness of MDPs to model approximation

Definition 7 Given a function class $\mathfrak{F}$ and positive constants (ε, δ), we say that an MDP $\widehat{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \widehat{\mathsf{P}}, \hat{r}, \gamma \rangle$ is an (ε, δ)-approximation of the MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma \rangle$ if it satisfies the following properties:

1. **Reward approximation**: For all $s \in \mathcal{S}$, and $a \in \mathcal{A}$,

$$|r(s, a) - \hat{r}(s, a)| \leq \varepsilon. \tag{21}$$

2. **Transition approximation**: For all $s \in \mathcal{S}$, and $a \in \mathcal{A}$,

$$d_{\mathfrak{F}}(\mathsf{P}(\cdot|s, a), \widehat{\mathsf{P}}(\cdot|s, a)) \leq \delta. \tag{22}$$

The main approximation result for MDPs relevant for our analysis is the following.

Theorem 3 *Given a function class $\mathfrak{F}$ and an MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathsf{P}, r, \gamma \rangle$, suppose $\widehat{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \widehat{\mathsf{P}}, \hat{r}, \gamma \rangle$ is an (ε, δ)-approximation of $\mathcal{M}$. Let $\hat{\pi}$ be an optimal strategy of $\widehat{\mathcal{M}}$ and $\hat{V}_{\hat{\pi}}$ be the corresponding value function. Then $\hat{\pi}$ is an α-optimal strategy of $\mathcal{M}$ with*

$$\alpha \leq 2\Big(\varepsilon + \frac{\gamma \Delta_{\hat{\pi}}}{(1-\gamma)}\Big), \tag{23}$$

where

$$\Delta_{\hat{\pi}} = \max_{s \in S, a \in A} \bigg| \sum_{s' \in S} \Big[\mathsf{P}(s'|s,a) \hat{V}_{\hat{\pi}}(s') - \widehat{\mathsf{P}}(s'|s,a) \hat{V}_{\hat{\pi}}(s') \Big] \bigg| \leq \delta \rho_{\mathfrak{F}}(\hat{V}_{\hat{\pi}}),$$

Therefore,

$$\alpha \leq 2\Big(\varepsilon + \frac{\gamma \delta \rho_{\mathfrak{F}}(\hat{V}_{\hat{\pi}})}{(1-\gamma)}\Big), \tag{24}$$

Proof The result follows along the same lines as (Subramanian et al. 2020, Theorem 27) applied to MDPs. For the sake of completeness, we provide a complete proof in Appendix A. □

The bounds in Theorem 3 are instance dependent. Instance-indepedent bounds can be obtained by using worst-case values of $\rho_{\mathfrak{F}}(\hat{V})$, as shown below.

Corollary 3 *If the function class $\mathfrak{F}$ in Theorem 3 is $\mathfrak{F}^{\mathrm{TV}}$, then*

$$\alpha \leq 2\Big(\varepsilon + \frac{\gamma \delta \operatorname{span}(\hat{r})}{(1-\gamma)}\Big).$$

Proof The result follows from the observation that $\rho_{d_{\mathfrak{F}^{\mathrm{TV}}}}(V) = \operatorname{span}(V)$ and then using Lemma 1 in Theorem 3. □

Corollary 4 *If the function class $\mathfrak{F}$ in Theorem 3 is $\mathfrak{F}^{\mathrm{W}}$, and the approximate MDP $\widehat{\mathcal{M}}$ is (L_r, L_{P})-Lipschitz with $\gamma L_{\mathsf{P}} < 1$, then*

$$\alpha \leq 2\left(\varepsilon + \frac{\gamma \delta L_r}{(1-\gamma L_{\mathsf{P}})}\right).$$

Proof The result follows from the observation that $\rho_{\mathfrak{F}^{\mathrm{W}}}(V) = \operatorname{Lip}(V)$ and then using Lemma 2 in Theorem 3. □

3.3 Generative models and PAC approximation bounds

In this section, we consider a setting similar to Sec. 2.4, but for MDPs. Suppose the components $\langle \mathcal{S}, \mathcal{A}, r, \gamma \rangle$ of an MDP $\mathcal{M}$ are known but the transition probability matrix P is not known. Suppose we have access to a *generative model*, i.e., a black-box simulator which provides samples $S_+ \sim \mathsf{P}(\cdot|s,a)$ of the next state S_+ for a given state-action pair (s,a) as input. Suppose we call the simulator n times at each state-action pair[5] and estimate an empirical model $\widehat{\mathsf{P}}_n$ as $\widehat{\mathsf{P}}_n(s'|s,a) = \text{count}(s'|s,a)/n$, where $\text{count}(s'|s,a)$ is the number of times s' is sampled with the input is (s,a). The MDP $\widehat{\mathcal{M}}_n = \langle \mathcal{S}, \mathcal{A}, \widehat{\mathsf{P}}_n, r, \gamma \rangle$ may be viewed as an approximation of MDP $\mathcal{M}$. Let $\hat{\pi}_n$ be the optimal strategy for $\widehat{\mathcal{M}}$.

As in Sec. 2.4, we are interested in the following question. Given an $\alpha > 0$, how many samples n are needed from the generative model to ensure that $\hat{\pi}_n$ is an α-optimal MDP. This is called the *sample complexity* of learning and a simple upper bound can be obtained using the approximation bounds of Theorem 3. For that matter, we state standard results on concentration inequalities.

Suppose $X \in \mathcal{X}$ is a random variable with distribution μ. Suppose $\{X_1, \dots, X_n\}$ is a sequence of random variables sampled according to μ. Let $\hat{\mu}_n$ denote the empirical measure constructed from $\{X_1, \dots, X_n\}$, i.e.,

$$\hat{\mu}_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{X_k = x\}}.$$

Then, we have the following.

Lemma 3 *For a given $H > 0$, let $\mathfrak{F}_H$ denote the set of functions $f\colon \mathcal{X} \to \mathbb{R}$ such that* $\text{span}(f) \le H$. *Then, for any $f \in \mathfrak{F}_H$ and $\Delta > 0$,*

$$\mathbb{P}\left(\left| \sum_{x \in \mathcal{X}} \big[\mu(x) f(x) - \mu_n(x) f(x) \big] \right| \ge \Delta \right) \le 2 \exp\left(-\frac{2n\Delta^2}{H^2} \right).$$

Proof Let $Z = f(X)$ and $Z_i = f(X_i)$. Then, $\{Z_1, \dots, Z_n\}$ is an i.i.d. sequence and $\text{Supp}(Z_i) \le H$. Then, by the Hoeffding inequality (Cesa-Bianchi and Lugosi 2006, Corollary A.1),

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mathbb{E}[Z] \right| \ge \Delta \right) \le 2 \exp\left(-\frac{2n\Delta^2}{H^2} \right).$$

The result then follows from observing that $\mathbb{E}[Z] = \sum_{x \in \mathcal{X}} \mu(x) f(x)$ and $(\sum_{i=1}^{n} Z_i)/n = \sum_{x \in \mathcal{X}} \mu_n(x) f(x)$. □

By combining Lemma 3 with Theorem 3, we get the following.

Theorem 4 *For any $\alpha > 0$ and $p > 0$, let*

$$n \ge \left\lceil \left(\frac{\gamma}{1-\gamma} \,\text{span}(r) \right)^2 \frac{2 \log(2|\mathcal{S}|\,|\mathcal{A}|/p)}{\alpha^2} \right\rceil.$$

Then, with probability $1-p$, the optimal strategy $\hat{\pi}_n$ of MDP $\widehat{\mathcal{M}}_n$ is α-optimal for $\mathcal{M}$.

[5] Note that since n samples are generated for every state-action pair, the total number of calls to the simulator equal $n|\mathcal{S}|\,|\mathcal{A}|$.

Proof Let $H = \operatorname{span}(r)$. Let $\hat{V}_n$ be the value function for $\widehat{\mathcal{M}}_n$. From Lemma 1, we know that $\hat{V}_n \in \mathfrak{F}_H$. Therefore, from Lemma 3, we have that for a given state-action pair (s, a) and $\Delta > 0$,

$$\mathbb{P}\left(\left|\sum_{s' \in \mathcal{S}} \mathsf{P}(s'|s,a)\hat{V}_n(s') - \sum_{s' \in \mathcal{S}} \widehat{\mathsf{P}}(s'|s,a)\hat{V}_n(s')\right| \geq \Delta\right) \leq 2\exp\left(-\frac{2n\Delta^2}{H^2}\right).$$

Therefore, by the union bound,

$$\mathbb{P}\left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\sum_{s' \in \mathcal{S}} \mathsf{P}(s'|s,a)\hat{V}_n(s') - \sum_{s' \in \mathcal{S}} \widehat{\mathsf{P}}(s'|s,a)\hat{V}_n(s')\right| \geq \Delta\right) \\ \leq 2|\mathcal{S}|\,|\mathcal{A}|\exp\left(-\frac{2n\Delta^2}{H^2}\right). \quad (25)$$

Now, choose Δ such that the right hand side of (25) equals p, i.e.,

$$\Delta = H\sqrt{\frac{\log(2|\mathcal{S}|\,|\mathcal{A}|/p)}{2n}}.$$

Then, Theorem 3 implies that with probability $1 - p$, $\hat{\pi}_n$ is a α_n-optimal strategy for $\mathcal{M}$, where

$$\alpha_n = 2\frac{\gamma\Delta}{1-\gamma} = \frac{\gamma}{1-\gamma}H\sqrt{\frac{2\log(2|\mathcal{S}|\,|\mathcal{A}|/p)}{n}}.$$

The result now follows by substituting the value of n. □

Remark 7 The sample complexity bound in Theorem 4 is not tight. It is shown in Azar et al. (2013) that finding an α-optimal policy with probability $1 - p$ requires at least

$$\Omega\left(|\mathcal{S}|\,|\mathcal{A}|\frac{\log(|\mathcal{S}|\,|\mathcal{A}|/\delta)}{(1-\gamma)\alpha^2}\right)$$

samples.[6] The upper bound in Theorem 4 is loose by a factor of $1/(1-\gamma)$. For the case of MDPs, tighter upper bounds which match the lower bound of Azar et al. (2013) (up to logarithmic factors) have been obtained in Sidford et al. (2018); Agarwal et al. (2020) by using Bernstein inequality rather than Hoeffding inequality and then providing a bound on the variance of the value function. We do not present these tighter bounds here because we have not been able to extend them to the setting of general-sum Markov games.

[6] Recall that we are working with normalized total expected reward (see Remark 1), while the results Azar et al. (2013) are derived for the unnormalized total reward. In the discussion above, we have normalized the results of Azar et al. (2013).

4 Proof of the main results

4.1 Bellman operators and characterization of Markov perfect equilibrium

Given a Markov strategy profile $\pi = (\pi^i)_{i\in\mathcal{N}}$, state $s \in \mathcal{S}$, and action profile $a = (a^i)_{i\in\mathcal{N}} \in \mathcal{A}$, we use the notation

$$\pi(a|s) = \prod_{i\in\mathcal{N}} \pi^i(a^i|s) \quad \text{and}$$
$$\pi^{-i}(a^{-i}|s) = \prod_{j\in\mathcal{N}\setminus\{i\}} \pi^j(a^j|s). \tag{26}$$

Given a player $i \in \mathcal{N}$ and a Markov strategy profile $\pi = (\pi^i, \pi^{-i})$, we define two Bellman operators as follows:

1. An operator $\mathcal{B}^i_{(\pi^i,\pi^{-i})} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ given as follows: for any $v \in \mathbb{R}^{|\mathcal{S}|}$ and $s \in \mathcal{S}$,
$$[\mathcal{B}^i_{(\pi^i,\pi^{-i})}v](s) = \sum_{a\in\mathcal{A}} \pi(a|s)\Big[(1-\gamma)r^i(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,a)v(s')\Big].$$
2. An operator $\mathcal{B}^i_{*,\pi^{-i}} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ given as follows: for any $v \in \mathbb{R}^{|\mathcal{S}|}$ and $s \in \mathcal{S}$,
$$[\mathcal{B}^i_{(*,\pi^{-i})}v](s) = \max_{a^i\in\mathcal{A}^i}\Bigg[\sum_{a^{-i}\in\mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s)\Big[(1-\gamma)r^i(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathbb{P}(s'|s,a)v(s')\Big]\Bigg].$$

Now, MPE and approximate MPE can be characterized using the Bellman operators. These are standard results. See, for example, Filar and Vrieze (1996).

Proposition 3 *A Markov strategy profile $\pi = (\pi^i)_{i\in\mathcal{N}}$ is an MPE if and ony if there exist **value functions** $V^i \in \mathbb{R}^{|\mathcal{S}|}, i \in \mathcal{N}$, such that*

$$V^i = \mathcal{B}^i_{(\pi^i,\pi^{-i})}V^i, \qquad \forall i \in \mathcal{N}, \tag{27a}$$

and

$$V^i = \mathcal{B}^i_{(*,\pi^{-i})}V^i, \qquad \forall i \in \mathcal{N}. \tag{27b}$$

Proposition 4 *Given a Markov strategy profile $\pi = (\pi^i)_{i\in\mathcal{N}}$, for any $i \in \mathcal{N}$, let V^i_π be the unique fixed point of $V^i_\pi = \mathcal{B}^i_{(\pi^i,\pi^{-i})}V^i_\pi$ and let $V^i_{(*,\pi^{-i})}$ be the unique fixed point of $V^i_{(*,\pi^{-i})} = \mathcal{B}^i_{(*,\pi^{-i})}V^i_{(*,\pi^{-i})}$. Then, the strategy profile π is an α-MPE, $\alpha = (\alpha^i)_{i\in\mathcal{N}}$, if and only if*

$$V^i_\pi \geq V^i_{(*,\pi^{-i})} - \alpha^i, \quad \forall i \in \mathcal{N}. \tag{28}$$

4.2 Relationship between games and MDPs

Given a game $\mathscr{G} = \langle \mathcal{N}, \mathcal{S}, (\mathcal{A}^i)_{i\in\mathcal{N}}, \mathsf{P}, (r^i)_{i\in\mathcal{N}}, \gamma\rangle$ and a Markov strategy $\pi = (\pi^i)_{i\in\mathcal{N}}$, we can define MDPs $\{\mathcal{M}^i_{\pi^{-i}}\}_{i\in\mathcal{N}}$ as follows. For player $i \in \mathcal{N}$, MDP $\mathcal{M}^i_{\pi^{-i}} = \langle \mathcal{S}, \mathcal{A}^i, \mathsf{P}^i_{\pi^{-i}}, r^i_{\pi^{-i}}, \gamma\rangle$, where the transition matrix $\mathsf{P}^i_{\pi^{-i}} : \mathcal{S} \times \mathcal{A}^i \to \mathscr{P}(\mathcal{S})$ is given by

$$\mathsf{P}^i_{\pi^{-i}}(s'|s, a^i) = \sum_{a^{-i}\in\mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s)\mathsf{P}(s'|s, (a^i, a^{-i})), \tag{29}$$

and the reward function $r^i_{\pi^{-i}} : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}$ is given by

$$r^i_{\pi^{-i}}(s, a^i) = \sum_{a^{-i}\in\mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s)r^i(s, (a^i, a^{-i})). \tag{30}$$

Note the Bellman operators $\mathcal{B}^i_{(\pi^i,\pi^{-i})}$ and $\mathcal{B}^i_{(*,\pi^{-i})}$ corresponding to game $\mathscr{G}$ and strategy π are the same as Bellman operators of MDP $\mathcal{M}^i_{\pi^{-i}}$. Therefore, by combining Propositions 1 and 3, we have the following:

Corollary 5 *A Markov strategy profile $\pi = (\pi^i)_{i\in\mathcal{N}}$ is an MPE if and only if for every $i \in \mathcal{N}$, the strategy π^i is an optimal strategy for MDP $\mathcal{M}^i_{\pi^{-i}}$.*

Similarly, by combining Propositions 2 and 4, we have the following:

Corollary 6 *Given approximate levels $\alpha = (\alpha^i)_{i\in\mathcal{N}}$, $\alpha^i \in \mathbb{R}_{\geq 0}$, a Markov strategy profile $\pi = (\pi^i)_{i\in\mathcal{N}}$, is an α-MPE if and only if for every $i \in \mathcal{N}$, the strategy π^i is an α^i-optimal strategy for MDP $\mathcal{M}^i_{\pi^{-i}}$.*

4.3 Relationship between MDPs corresponding to a strategy profile

We first provide a preliminary result.

Lemma 4 *For any function $f : \mathcal{S} \to \mathbb{R}$, player $i \in \mathcal{N}$, and any strategy π^{-i} for players other than i, and any $(s, a^i) \in \mathcal{S} \times \mathcal{A}^i$,*

$$\begin{aligned}&\Big|\sum_{s'\in\mathcal{S}} f(s')\mathsf{P}^i_{\pi^{-i}}(s'|s, a^i) - \sum_{s'\in\mathcal{S}} f(s')\widehat{\mathsf{P}}^i_{\pi^{-i}}(s'|s, a^i)\Big| \\ &\qquad\le \max_{a^{-i}\in\mathcal{A}^{-i}} \Big|\sum_{s'\in\mathcal{S}} f(s')\mathsf{P}(s'|s, (a^i, a^{-i})) - \sum_{s'\in\mathcal{S}} f(s')\widehat{\mathsf{P}}(s'|s, (a^i, a^{-i}))\Big|.\end{aligned}$$

Therefore,

$$\begin{aligned}&\max_{s\in\mathcal{S},a^i\in\mathcal{A}^i} \Big|\sum_{s'\in\mathcal{S}} f(s')\mathsf{P}^i_{\pi^{-i}}(s'|s, a^i) - \sum_{s'\in\mathcal{S}} f(s')\widehat{\mathsf{P}}^i_{\pi^{-i}}(s'|s, a^i)\Big| \\ &\qquad\le \max_{s\in\mathcal{S},(a^i,a^{-i})\in\mathcal{A}} \Big|\sum_{s'\in\mathcal{S}} f(s')\mathsf{P}(s'|s, (a^i, a^{-i})) - \sum_{s'\in\mathcal{S}} f(s')\widehat{\mathsf{P}}(s'|s, (a^i, a^{-i}))\Big|.\end{aligned}$$

Proof For the first part, from definition of $\widehat{\mathsf{P}}^i_{\pi^{-i}}$, we have

$$
\begin{aligned}
&\left|\sum_{s'\in\mathcal{S}} f(s')\mathsf{P}^i_{\pi^{-i}}(s'|s,a^i) - \sum_{s'\in\mathcal{S}} f(s')\widehat{\mathsf{P}}^i_{\pi^{-i}}(s'|s,a^i)\right| \\
&\quad= \Bigg|\sum_{s'\in\mathcal{S}}\sum_{a^{-i}\in\mathcal{A}^{-i}} f(s')\pi^{-i}(a^{-i}|s)\mathsf{P}(s'|s,(a^i,a^{-i})) \\
&\qquad\qquad - \sum_{s'\in\mathcal{S}}\sum_{a^{-i}\in\mathcal{A}^{-i}} f(s')\pi^{-i}(a^{-i}|s)\widehat{\mathsf{P}}(s'|s,(a^i,a^{-i}))\Bigg| \\
&\quad\le \Bigg|\sum_{a^{-i}\in\mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \\
&\qquad\times\left[\sum_{s'\in\mathcal{S}} f(s')(\mathsf{P}(s'|s,(a^i,a^{-i})) - \widehat{\mathsf{P}}(s'|s,(a^i,a^{-i})))\right]\Bigg| \\
&\quad\le \sum_{a^{-i}\in\mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \\
&\qquad\times\left|\sum_{s'\in\mathcal{S}} f(s')(\mathsf{P}(s'|s,(a^i,a^{-i})) - \widehat{\mathsf{P}}(s'|s,(a^i,a^{-i})))\right| \\
&\quad\le \sum_{a^{-i}\in\mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \\
&\qquad\times\max_{\tilde{a}^{-i}\in\mathcal{A}^{-i}}\left|\sum_{s'\in\mathcal{S}} f(s')(\mathsf{P}(s'|s,(a^i,\tilde{a}^{-i})) - \widehat{\mathsf{P}}(s'|s,(a^i,\tilde{a}^{-i})))\right| \\
&\quad= \max_{\tilde{a}^{-i}\in\mathcal{A}^{-i}}\left|\sum_{s'\in\mathcal{S}} f(s')(\mathsf{P}(s'|s,(a^i,\tilde{a}^{-i})) - \widehat{\mathsf{P}}(s'|s,(a^i,\tilde{a}^{-i})))\right|.
\end{aligned}
$$

The second part following by taking a maximum over (s,a^i).

Suppose we are given a game $\mathscr{G}$ and its (ε,δ) approximation $\widehat{\mathscr{G}}$. Moreover, suppose $\hat{\pi} = (\hat{\pi}^i)_{i\in\mathcal{N}}$ is an MPE of $\widehat{\mathscr{G}}$.

Let $\{\widehat{\mathcal{M}}^i_{\hat{\pi}^{-i}}\}$ be the MDPs corresponding to game $\widehat{\mathscr{G}}$ and strategy $\hat{\pi}$. Similarly, let $\{\mathcal{M}^i_{\hat{\pi}^{-i}}\}$ be the MDPs corresponding to game $\mathscr{G}$ and strategy $\hat{\pi}$. Then, we have the following.

An immediate implication of Lemma 4 is the following.

Lemma 5 *For any player $i\in\mathcal{N}$, MDP $\widehat{\mathcal{M}}^i_{\hat{\pi}^{-i}}$ is an (ε,δ) approximation of MDP $\mathcal{M}^i_{\hat{\pi}^{-i}}$.*

Proof Consider

$$\begin{aligned}
&|r^i_{\hat{\pi}^{-i}}(s,a^i) - \hat{r}^i_{\hat{\pi}^{-i}}(s,a^i)| \\
&\overset{(a)}{\leq} \sum_{a^{-i}\in\mathcal{A}^{-i}} \hat{\pi}^{-i}(a^{-i}|s)|r^i(s,(a^i,a^{-i})) - \hat{r}^i(s,(a^i,a^{-i}))| \\
&\overset{(b)}{\leq} \sum_{a^{-i}\in\mathcal{A}^{-i}} \hat{\pi}^{-i}(a^{-i}|s)\varepsilon \\
&\overset{(c)}{=} \varepsilon,
\end{aligned} \tag{31}$$

where (a) follows from (30), (b) follows from (6) and (c) follows as ε is independent of a^{-i}. Furthermore,

$$\begin{aligned}
&\max_{s\in\mathcal{S},a^i\in\mathcal{A}^i} d_{\mathfrak{F}}(\mathsf{P}^i_{\hat{\pi}^{-i}}(\cdot|s,a^i), \widehat{\mathsf{P}}^i_{\hat{\pi}^{-i}}(\cdot|s,a^i)) \\
&\overset{(d)}{=} \sup_{f\in\mathfrak{F}} \max_{s\in\mathcal{S},a^i\in\mathcal{A}^i} \left| \sum_{s'\in\mathcal{S}} f(s')\mathsf{P}^i_{\hat{\pi}^{-i}}(s'|s,a^i) - \sum_{s'\in\mathcal{S}} f(s')\widehat{\mathsf{P}}^i_{\hat{\pi}^{-i}}(s'|s,a^i) \right| \\
&\overset{(e)}{\leq} \sup_{f\in\mathfrak{F}} \max_{\substack{s\in\mathcal{S}\\(a^i,a^{-i})\in\mathcal{A}}} \left| \sum_{s'\in\mathcal{S}} f(s')(\mathsf{P}(s'|s,(a^i,a^{-i})) - \widehat{\mathsf{P}}(s'|s,(a^i,a^{-i}))) \right| \\
&\overset{(f)}{=} \max_{\substack{s\in\mathcal{S}\\(a^i,a^{-i})\in\mathcal{A}}} d_{\mathfrak{F}}(\mathsf{P}(\cdot|s,(a^i,a^{-i})), \widehat{\mathsf{P}}(\cdot|s,(a^i,a^{-i}))) \\
&\overset{(g)}{=} \delta
\end{aligned} \tag{32}$$

where (d) and (f) follows from Definition 3, (e) follows from Lemma 4, and (g) follows from Definition 4.

Equations (31) and (32) imply that MDP $\widehat{\mathcal{M}}^i_{\hat{\pi}^{-i}}$ is an (ε,δ)-approximation of $\mathcal{M}^i_{\hat{\pi}^{-i}}$ (see Definition 7). □

4.4 Proof of Theorem 1

Arbitrarily fix a player $i \in \mathcal{N}$. Then, we have the following.

1. From Corollary 5, since $\hat{\pi}$ is an MPE of $\widehat{\mathscr{G}}$, we have that the strategy $\hat{\pi}^i$ is optimal for MDP $\widehat{\mathcal{M}}^i_{\hat{\pi}^{-i}}$.
2. From Lemma 5, we know that MDP $\widehat{\mathcal{M}}^i_{\hat{\pi}^{-i}}$ is an (ε,δ) approximation of MDP $\mathcal{M}^i_{\hat{\pi}^{-i}}$. Then, by Theorem 3, we get that strategy $\hat{\pi}^i$ is an α^i-optimal strategy for MDP $\mathcal{M}^i_{\hat{\pi}^{-i}}$, where α^i is given by Theorem 3. Lemma 4 shows that Δ for MDP $\mathcal{M}^i_{\hat{\pi}^{-i}}$ is upper-bounded by $\Delta^i_{(\hat{\pi}^i,\hat{\pi}^{-i})}$ given in Theorem 1.
3. Since the above results hold for all $i \in \mathcal{N}$, Corollary 6 implies that strategy profile $\hat{\pi}$ is an α-MPE of $\mathscr{G}$, where $\alpha = (\alpha^i)_{i\in\mathcal{N}}$ and α^i is given by Theorem 3.
4. The specific formulas for α in Corollaries 1 and 2 follow from Corollaries 3 and 4.

4.5 Proof of Theorem 2

The proof is similar to the proof of Theorem 4. Let $H = \operatorname{span}(r)$. Recall the definition of $\mathfrak{F}_H$ in Lemma 3. Arbitrarily pick a player $i \in \mathcal{N}$. Then, from Lemma 1, we know that $\hat{V}^i_{\hat{\pi}_n} \in \mathfrak{F}_H$. Therefore, from Lemma 3, we have that for a given state-action pair (s, a) and $\Delta > 0$,

$$\mathbb{P}\left(\left|\sum_{s' \in \mathcal{S}} \mathsf{P}(s'|s,a)\hat{V}^i_{\hat{\pi}_n}(s') - \sum_{s' \in \mathcal{S}} \widehat{\mathsf{P}}(s'|s,a)\hat{V}^i_{\hat{\pi}_n}(s')\right| \geq \Delta\right) \leq 2\exp\left(-\frac{2n\Delta^2}{H^2}\right).$$

Therefore, by the union bound,

$$\mathbb{P}\left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\sum_{s' \in \mathcal{S}} \mathsf{P}(s'|s,a)\hat{V}^i_{\hat{\pi}_n}(s') - \sum_{s' \in \mathcal{S}} \widehat{\mathsf{P}}(s'|s,a)\hat{V}^i_{\hat{\pi}_n}(s')\right| \geq \Delta\right) \leq 2|\mathcal{S}|\,|\mathcal{A}|\exp\left(-\frac{2n\Delta^2}{H^2}\right),$$

where $|\mathcal{A}| = \prod_{i\in\mathcal{N}} |\mathcal{A}^i|$.

Now, by choosing $\Delta = H\sqrt{(\log(2|\mathcal{S}|\,|\mathcal{A}|\,|\mathcal{N}|/p))/2n}$, we get

$$\mathbb{P}\left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|\sum_{s' \in \mathcal{S}} \mathsf{P}(s'|s,a)\hat{V}^i_{\hat{\pi}_n}(s') - \sum_{s' \in \mathcal{S}} \widehat{\mathsf{P}}(s'|s,a)\hat{V}^i_{\hat{\pi}_n}(s')\right| \geq H\sqrt{\frac{\log(2|\mathcal{S}|\,|\mathcal{A}|\,|\mathcal{N}|/p)}{2n}}\right) \leq \frac{p}{|\mathcal{N}|}. \quad (33)$$

Note that the first term inside the probability expression equals $\Delta^i_{\hat{\pi}_n}$. Then, by union bound

$$\mathbb{P}\left(\max_{i\in\mathcal{N}} \Delta^i_{\hat{\pi}_n} \geq H\sqrt{\frac{\log(2|\mathcal{S}|\,|\mathcal{A}|\,|\mathcal{N}|/p)}{2n}}\right) \leq p \quad (34)$$

Therefore, Theorem 1 implies that with probability $1-p$, $\hat{\pi}_n$ is a α_n-MPE of game $\mathscr{G}$, where

$$\alpha_n = 2\frac{\gamma\Delta^i_{\hat{\pi}_n}}{1-\gamma} \leq \frac{\gamma}{1-\gamma}H\sqrt{\frac{2\log(2|\mathcal{S}|\,|\mathcal{A}|\,|\mathcal{N}|/p)}{n}}.$$

The result then follows by substituting the value of n. $\square$

5 Numerical examples

In this section, we present two numerical examples to validate the main results of Theorem 1 and 2.

5.1 Robustness of Markov perfect equilibrium

Consider a setting where $\mathcal{N} = \{1, 2\}$, $\mathcal{S} = \{1, 2, 3\}$, $\mathcal{A}_1 = \mathcal{A}_2 = \{1, 2\}$, and $\gamma = 0.9$. We consider two games: original game $\mathscr{G}$ and approximate game $\widehat{\mathscr{G}}$ which differ in their reward functions and transition matrices. We describe the transition matrices as $\{\mathsf{P}(a)\}_{a\in\mathcal{A}}$, where $\mathsf{P}(a) = [\mathsf{P}(s' \mid s, a)]_{s,s'\in\mathcal{S}}$ and describe the reward functions as $\{r(s)\}_{s\in\mathcal{S}}$ where $r(s)$ is the bi-matrix $[(r_1(s, (a_1, a_2)), r_2(s, (a_1, a_2)))]_{(a_1,a_2)\in\mathcal{A}}$.

For the original game $\mathscr{G}$, we have

$$r(1) = \begin{bmatrix} (1.0, 0.4) & (0.7, 1.0) \\ (0.3, 1.0) & (0.8, 0.7) \end{bmatrix}, \quad r(2) = \begin{bmatrix} (0.6, 0.7) & (0.7, 0.6) \\ (0.3, 0.8) & (0.2, 0.2) \end{bmatrix},$$
$$r(3) = \begin{bmatrix} (0.2, 0.6) & (0.1, 0.7) \\ (0.6, 0.7) & (0.5, 0.3) \end{bmatrix},$$

and

$$\mathsf{P}((1,1)) = \begin{bmatrix} 0.40 & 0.40 & 0.20 \\ 0.10 & 0.50 & 0.40 \\ 0.40 & 0.10 & 0.50 \end{bmatrix}, \qquad \mathsf{P}((1,2)) = \begin{bmatrix} 0.30 & 0.40 & 0.30 \\ 0.20 & 0.20 & 0.60 \\ 0.30 & 0.35 & 0.35 \end{bmatrix},$$
$$\mathsf{P}((2,1)) = \begin{bmatrix} 0.25 & 0.25 & 0.50 \\ 0.30 & 0.30 & 0.40 \\ 0.20 & 0.20 & 0.60 \end{bmatrix}, \qquad \mathsf{P}((2,2)) = \begin{bmatrix} 0.10 & 0.20 & 0.70 \\ 0.20 & 0.10 & 0.70 \\ 0.40 & 0.20 & 0.40 \end{bmatrix}.$$

For the approximate game $\widehat{\mathscr{G}}$, we have

$$\hat{r}(1) = \begin{bmatrix} (0.99, 0.40) & (0.69, 1.00) \\ (0.30, 0.99) & (0.81, 0.71) \end{bmatrix}, \quad \hat{r}(2) = \begin{bmatrix} (0.59, 0.70) & (0.69, 0.61) \\ (0.30, 0.80) & (0.19, 0.21) \end{bmatrix},$$
$$\hat{r}(3) = \begin{bmatrix} (0.19, 0.59) & (0.09, 0.70) \\ (0.59, 0.69) & (0.50, 0.30) \end{bmatrix},$$

and

$$\widehat{\mathsf{P}}((1,1)) = \begin{bmatrix} 0.45 & 0.35 & 0.20 \\ 0.15 & 0.45 & 0.40 \\ 0.45 & 0.10 & 0.45 \end{bmatrix}, \quad \widehat{\mathsf{P}}((1,2)) = \begin{bmatrix} 0.25 & 0.45 & 0.30 \\ 0.25 & 0.15 & 0.60 \\ 0.35 & 0.30 & 0.35 \end{bmatrix},$$
$$\widehat{\mathsf{P}}((2,1)) = \begin{bmatrix} 0.25 & 0.30 & 0.45 \\ 0.35 & 0.30 & 0.35 \\ 0.25 & 0.20 & 0.55 \end{bmatrix}, \quad \widehat{\mathsf{P}}((2,2)) = \begin{bmatrix} 0.15 & 0.15 & 0.70 \\ 0.25 & 0.10 & 0.65 \\ 0.40 & 0.25 & 0.35 \end{bmatrix}.$$

A MPE of $\widehat{\mathscr{G}}$ and the corresponding value functions (computed by solving a non-linear program as described in Filar et al. (1991)) are as follows:

$$\hat{\pi}^1 = \begin{bmatrix} 0.33 & 0.67 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}, \qquad \hat{\pi}^2 = \begin{bmatrix} 0.13 & 0.87 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \end{bmatrix}, \tag{35}$$

$$\hat{V}^1_{\hat{\pi}} = \begin{bmatrix} 0.6327 \\ 0.6170 \\ 0.6187 \end{bmatrix}, \qquad \hat{V}^2_{\hat{\pi}} = \begin{bmatrix} 0.7258 \\ 0.7148 \\ 0.7148 \end{bmatrix}. \tag{36}$$

In (35), the strategy is described as $\hat{\pi}^i = [\hat{\pi}^i(a^i|s)]_{s\in\mathcal{S},a^i\in\mathcal{A}^i}$. For strategy $\hat{\pi}$ in (35), we compute the value functions $V^i_{\hat{\pi}}$ for game $\mathscr{G}$ as described in Proposition 3 and the value functions $V^i_{(*,\hat{\pi}^{-i})}$ as described in Proposition 4 (see Sec. 4). These are given by

$$V^1_{\hat{\pi}} = \begin{bmatrix} 0.6341 \\ 0.6192 \\ 0.6209 \end{bmatrix}, \qquad V^2_{\hat{\pi}} = \begin{bmatrix} 0.7252 \\ 0.7142 \\ 0.7154 \end{bmatrix}, \tag{37}$$

$$V^1_{(*,\hat{\pi}^2)} = \begin{bmatrix} 0.6394 \\ 0.6222 \\ 0.6241 \end{bmatrix}, \qquad V^2_{(\hat{\pi}^1,*)} = \begin{bmatrix} 0.7280 \\ 0.7158 \\ 0.7171 \end{bmatrix}. \tag{38}$$

Note that

$$\alpha^1_* = \|V^1_{(*,\hat{\pi}^2)} - V^1_{\hat{\pi}}\|_\infty = 0.005300, \tag{39a}$$

$$\alpha^2_* = \|V^2_{(\hat{\pi}^1,*)} - V^2_{\hat{\pi}}\|_\infty = 0.002785. \tag{39b}$$

Thus, $\hat{\pi}$ is a $(0.005300, 0.002785)$-MPE of $\mathscr{G}$.

Now, we compare α_* with the bounds that we obtain using Theorem 1. We first consider the bound in terms of $\Delta^i_{\hat{\pi}}$. Note that

$$\max_{a\in\mathcal{A}} \max_{s\in\mathcal{S}} |r(s,a) - \hat{r}(s,a)| = 0.01.$$

Thus, $\varepsilon = 0.01$. Moreover,

$$\Delta^1_{\hat{\pi}} = \max_{s\in S, a\in A} \left| \sum_{s'\in S} \left[\mathsf{P}(s'|s,a)\hat{V}^1_{\hat{\pi}}(s') - \widehat{\mathsf{P}}(s'|s,a)\hat{V}^1_{\hat{\pi}}(s') \right] \right| = 0.000784$$

$$\Delta^2_{\hat{\pi}} = \max_{s\in S, a\in A} \left| \sum_{s'\in S} \left[\mathsf{P}(s'|s,a)\hat{V}^2_{\hat{\pi}}(s') - \widehat{\mathsf{P}}(s'|s,a)\hat{V}^2_{\hat{\pi}}(s') \right] \right| = 0.000550$$

Then, by Theorem 1, we have that

$$\alpha \le 2\left(\varepsilon + \frac{\gamma \Delta_{\hat{\pi}}}{1-\gamma}\right) = 2\times 0.01 + \frac{2\times 0.9}{0.1} \begin{bmatrix} 0.000784 \\ 0.000550 \end{bmatrix} = \begin{bmatrix} 0.034112 \\ 0.029900 \end{bmatrix}.$$

Now, we consider the upper bound on $\Delta^i_{\hat{\pi}}$ in terms of $\rho_F(\hat{V}^i_{\hat{\pi}})$.

1. We first consider the case when $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$. Note that

$$\max_{a\in\mathcal{A}} \max_{s\in\mathcal{S}} d_{\mathfrak{F}^{\mathrm{TV}}}(\mathsf{P}(\cdot|s,a), \widehat{\mathsf{P}}(\cdot|s,a)) = 0.05,$$

   Thus when $\mathfrak{F} = \mathfrak{F}^{\mathrm{TV}}$, $\widehat{\mathscr{G}}$ is a $(0.01, 0.05)$-approximation of game $\mathscr{G}$. Also note that $\mathrm{span}(\hat{V}^1_{\hat{\pi}}) = 0.015684$ and $\mathrm{span}(\hat{V}^2_{\hat{\pi}}) = 0.010990$. Then, from Theorem 1, we have that

$$\alpha \le 2\left(\varepsilon + \frac{\gamma\delta\,\mathrm{span}(\hat{V}^i_{\hat{\pi}})}{1-\gamma}\right) = 2\times 0.01 + \frac{2\times 0.9\times 0.05}{0.1} \begin{bmatrix} 0.015684 \\ 0.010990 \end{bmatrix} = \begin{bmatrix} 0.034116 \\ 0.029903 \end{bmatrix}.$$

2. Now we equip the state space $\mathcal{S}$ with a metric d where $d(s, s') = |s - s'|$ and consider the case $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$. Note that

$$\max_{a \in \mathcal{A}} \max_{s \in \mathcal{S}} d_{\mathfrak{F}^{\mathrm{W}}}(\mathsf{P}(\cdot|s,a), \widehat{\mathsf{P}}(\cdot|s,a)) = 0.10.$$

Thus when $\mathfrak{F} = \mathfrak{F}^{\mathrm{W}}$, $\widehat{\mathscr{G}}$ is a $(0.01, 0.10)$-approximation of game $\mathscr{G}$. Also note that $\mathrm{Lip}(\hat{V}^1_{\hat{\pi}}) = 0.015684$ and $\mathrm{Lip}(\hat{V}^2_{\hat{\pi}}) = 0.010990$. Then, from Theorem 1, we have that

$$\alpha \le 2\biggl(\varepsilon + \frac{\gamma\delta \,\mathrm{Lip}(\hat{V}^i_{\hat{\pi}})}{1-\gamma}\biggr) = 2 \times 0.01 + \frac{2 \times 0.9 \times 0.10}{0.1} \begin{bmatrix} 0.015684 \\ 0.010990 \end{bmatrix} = \begin{bmatrix} 0.048231 \\ 0.039782 \end{bmatrix}.$$

The above example shows that with the given (ε, δ), the bound of Theorem 1 is loose by only a small multiplicative factor of approximately 6 to 15.

5.2 Sample complexity of generative models

We now consider the setting for model based MARL. Consider the game $\mathscr{G}$ described in Sec. 5.1 but suppose that the transition matrix P is not known but we have access to a generative model which can generate samples from P. We assume that we are interested in identifying an α-MPE, where $\alpha = 0.1$ with probability $1 - p = 0.99$. From Theorem 2, we have an upper bound on the number n of samples for each state-action pair as

$$\begin{aligned} n &\ge \left\lceil \left(\frac{\gamma}{1-\gamma}\,\mathrm{span}(r)\right)^2 \frac{2\log(2|\mathcal{S}|\bigl(\prod_{i\in\mathcal{N}} |\mathcal{A}^i|\bigr)|\mathcal{N}||/p)}{\alpha^2} \right\rceil. \\ &= \left\lceil \left(\frac{0.9}{1-0.9} \times 0.9\right)^2 \frac{2\log(2 \times 3 \times 4 \times 2/0.01)}{0.1^2} \right\rceil = 111,227. \end{aligned}$$

We now verify this result via simulation. We run $M = 1,000$ experiments. For each experiment, we generate generate $n = 111227$ samples for each state-action pair, and estimate an empirical model $\widehat{\mathsf{P}}_n(s'|s,a) = \mathrm{count}(s'|s,a)/n$. We compute the MPE $\hat{\pi}_n$ for the approximate game $\widehat{\mathscr{G}} = \langle \mathcal{S}, \{\mathcal{A}\}_{i\in\mathcal{N}}, \widehat{\mathsf{P}}, r, \gamma \rangle$. Then, using Proposition 4, we compute $\alpha_n = (\alpha^1_n, \alpha^2_n)$ such that $\hat{\pi}$ is a α-MPE of game $\mathscr{G}$. The scatter plot of $\alpha_n = (\alpha^1_n, \alpha^2_n)$ along with the empirical distribution of α^1_n and α^2_n are shown in Fig. 1. Note that for most cases, both α^1_n and α^2_n are smaller than 10^{-4}, which is much less than our target α of 0.1. This highlights the looseness of the upper bound in Theorem 2.

6 Discussion and Conclusion

In this paper, we show that MPE are robust to model approximation and provide sample complexity bounds for computing an approximate MPE. In particular, we show that any MPE for an approximate or perturbed game is an approximate MPE of the original game. We provide bounds on the degree of

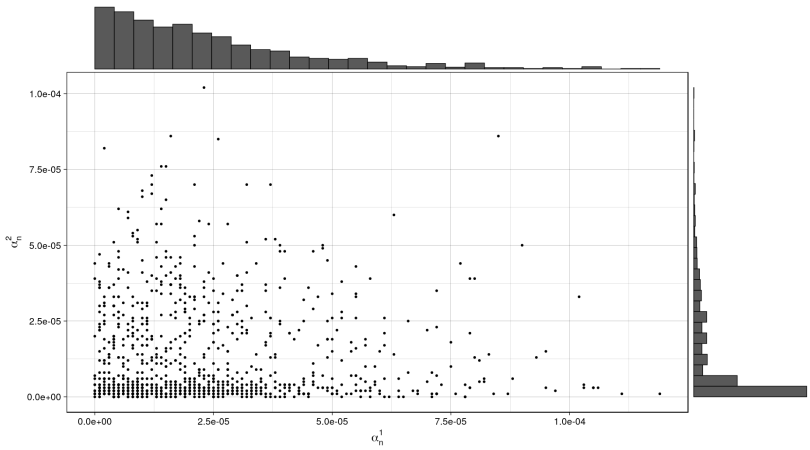

Fig. 1 Scatter plot of (α_n^1, α_n^2) such that the policy $\hat{\pi}_n$ of game $\widehat{\mathscr{G}}_n$ is a (α_n^1, α_n^2)-MPE of game $\mathscr{G}$ for $M = 1,000$ independently generated experiments. The histogram of the marginal probability distribution for α_n^1 and α_n^2 are shown on the top and the right.

approximation based on the approximation error in the reward and transition functions and properties of the value function of the MPE. We also present coarser, instance independent upper bounds, which do not depend of the value function but only depend on the properties of the reward and transition function of the approximate game.

An interesting feature of the results is that the approximation bounds depend on the choice of the metric on probability spaces. We work with a class of metrics known as IPMs and specialize our results for two specific choices of IPMs: total variation distance and Wasserstein distance. However, the results are applicable to any IPM. For games with high-dimensional state spaces, metrics such as maximum mean discrepancy (Sriperumbudur et al. 2008) might be more appropriate.

Using these robustness bounds, we derive an upper bound on sample complexity of model based MARL using a generative model. In particular, we show that for any $\alpha > 0$ and $p > 0$, any MPE of an approximate game constructed using an approximate model based on

$$|\mathcal{S}|\Big(\prod_{i=1}^{n} |\mathcal{A}^i|\Big)n \geq |\mathcal{S}|\Big(\prod_{i=1}^{n} |\mathcal{A}^i|\Big)\Big(\frac{\gamma}{1-\gamma}\,\mathrm{span}(r)\Big)^2 \frac{2\log(2|\mathcal{S}|\big(\prod_{i\in\mathcal{N}} |\mathcal{A}^i|\big)|\mathcal{N}||/p)}{\alpha^2}$$

samples is an α-MPE of the original game with probability $1-p$.

The upper bound is not tight. As argued in Zhang et al. (2020), the lower bound on the sample complexity for MDPs obtained in Azar et al. (2013) (see Remark 7) can be directly translated to games to provide a lower bound of

$$\Omega\bigg(|\mathcal{S}|\Big(\prod_{i=1}^{n} |\mathcal{A}^i|\Big)\frac{\log(|\mathcal{S}|\big(\sum_{i\in\mathcal{N}} |\mathcal{A}^i|\big)/\delta)}{(1-\gamma)\alpha^2}\bigg).$$

Thus, our upper bound is tight in $|\mathcal{S}|$, but loose in the logarithmic factor in $\{|\mathcal{A}^i|\}$ ($\sum_{i\in\mathcal{N}} |\mathcal{A}^i|$ vs $\prod_{i\in\mathcal{N}} |\mathcal{A}^i|$), and loose by a factor of $(1-\gamma)$ in the discount factor. The sample complexity bounds for zero-sum games obtained in Zhang et al. (2020) were also loose in the action space. Zhang et al. (2020) also highlight the difficulty in tightening the lower bound to $\Omega(\log(\prod_{i\in\mathcal{N}} |\mathcal{A}^i|))$. Since, zero-sum games are a special case of general-sum games, the same difficulties hold for the latter as well.

For MDPs, as mentioned in Remark 7, recent papers Sidford et al. (2018); Agarwal et al. (2020); Li et al. (2020) have obtained upper bounds which are tight in $(1-\gamma)$. Here we highlight the challenges in extending the analysis to general-sum games. The analysis in Sidford et al. (2018) and Li et al. (2020) assumes that specific algorithms are used to obtain a near-optimal solution of the sampled model $\widehat{\mathcal{M}}_n$. These algorithms are not directly applicable for general-sum games. The analysis in Agarwal et al. (2020) does not make any assumptions on the algorithm used to compute the near-optimal solution of the sampled model. This analysis is generalized to zero-sum games in Zhang et al. (2020), which obtains sample complexity bounds which are tight in $(1-\gamma)$. The key step in the analysis in Agarwal et al. (2020); Zhang et al. (2020) is the construction of an absorbing MDP, which relies on the uniqueness of the optimal or minimax value function. In general-sum games, the value functions corresponding to different MPE are different, so the aborbing Markov chain construction does not work. Therefore, a more nuanced argument is needed to obtain a tight sample complexity bound for general-sum games.

We conclude by noting that the results presented in this paper were restricted to Markov games with perfect information. An interesting future direction is to develop similar approximation bounds for Markov games with imperfect information as well as specific classes of dynamic games such as mean-field games and their variants.

7 Acknowledgements

The work of Amit Sinha and Aditya Mahajan was supported in part by the Innovation for Defence Excellence and Security (IDEaS) Program of the Canadian Department of National Defence through grant CFPMN2-30.

A Proof of Theorem 3

Let V denote the value function for MDP $\mathcal{M}$ and $V_{\hat{\pi}}$ denote the value function for policy $\hat{\pi}$ in MDP $\mathcal{M}$. From triangle inequality, we have

$$\|V - V_{\hat{\pi}}\|_\infty \le \|V - \hat{V}_{\hat{\pi}}\|_\infty + \|V_{\hat{\pi}} - \hat{V}_{\hat{\pi}}\|_\infty. \tag{40}$$

Now we bound the two terms separately. For the first term, we have

$$\begin{aligned}
\|V - \hat{V}_{\hat{\pi}}\|_\infty &\overset{(a)}{\leq} \max_{s\in\mathcal{S}} \Big| \max_{a\in\mathcal{A}} \Big[(1-\gamma) r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathsf{P}(s'|s,a) V(s') \\
&\qquad - (1-\gamma)\hat{r}(s,a) - \gamma \sum_{s'\in\mathcal{S}} \widehat{\mathsf{P}}(s'|s,a)\hat{V}_{\hat{\pi}}(s') \Big] \Big| \\
&\overset{(b)}{\leq} (1-\gamma) \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big| r(s,a) - \hat{r}(s,a) \big| \\
&\quad + \gamma \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big| \sum_{s'\in\mathcal{S}} \mathsf{P}(s'|s,a)V(s') - \mathsf{P}(s'|s,a)\hat{V}_{\hat{\pi}}(s') \Big| \\
&\quad + \gamma \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big| \sum_{s'\in\mathcal{S}} \mathsf{P}(s'|s,a)\hat{V}_{\hat{\pi}}(s') - \widehat{\mathsf{P}}(s'|s,a)\hat{V}_{\hat{\pi}}(s') \Big| \\
&\leq (1-\gamma)\varepsilon + \gamma \|V - \hat{V}_{\hat{\pi}}\|_\infty + \gamma \Delta_{\hat{\pi}},
\end{aligned}$$

where (a) relies on the fact that $\max f(x) \leq \max|f(x) - g(x)| + \max g(x)$, (b) follows from triangle inequality, and (c) follows from the definition of ε and $\Delta_{\hat{\pi}}$. Therefore,

$$\|V - \hat{V}_{\hat{\pi}}\|_\infty \leq \varepsilon + \frac{\gamma \Delta_{\hat{\pi}}}{1-\gamma}. \tag{41}$$

For the second term of (40), we have

$$\begin{aligned}
\|V_{\hat{\pi}} - \hat{V}_{\hat{\pi}}\|_\infty &= \max_{s\in\mathcal{S}} \Big| \sum_{a\in\mathcal{A}} \hat{\pi}(a|s) \Big[(1-\gamma) r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathsf{P}(s'|s,a) V_{\hat{\pi}}(s') \\
&\qquad - (1-\gamma)\hat{r}(s,a) - \gamma \sum_{s'\in\mathcal{S}} \widehat{\mathsf{P}}(s'|s,a)\hat{V}_{\hat{\pi}}(s') \Big] \Big| \\
&\overset{(d)}{\leq} (1-\gamma) \max_{s\in\mathcal{S}} \Big| \sum_{a\in\mathcal{A}} \hat{\pi}(a|s) \big[r(s,a) - \hat{r}(s,a) \big] \Big| \\
&\quad + \gamma \max_{s\in\mathcal{S}} \Big| \sum_{a\in\mathcal{A}} \hat{\pi}(a|s) \Big[\sum_{s'\in\mathcal{S}} \big[\mathsf{P}(s'|s,a)V_{\hat{\pi}}(s') - \mathsf{P}(s'|s,a)\hat{V}_{\hat{\pi}}(s') \big] \Big] \Big| \\
&\quad + \gamma \max_{s\in\mathcal{S}} \Big| \sum_{a\in\mathcal{A}} \hat{\pi}(a|s) \Big[\sum_{s'\in\mathcal{S}} \big[\mathsf{P}(s'|s,a)\hat{V}_{\hat{\pi}}(s') - \widehat{\mathsf{P}}(s'|s,a)\hat{V}_{\hat{\pi}}(s') \big] \Big] \Big| \\
&\overset{(e)}{\leq} (1-\gamma)\varepsilon + \gamma \|V_{\hat{\pi}} - \hat{V}_{\hat{\pi}}\|_\infty + \Delta_{\hat{\pi}}
\end{aligned}$$

where (d) follows from triangle inequality and (e) follows from the definition of ε and $\Delta_{\hat{\pi}}$. Therefore,

$$\|V_{\hat{\pi}} - \hat{V}_{\hat{\pi}}\|_\infty \leq \varepsilon + \frac{\gamma \Delta_{\hat{\pi}}}{1-\gamma}. \tag{42}$$

The result then follows by substituting (41) and (42) in (40). □

References

D. Acemoglu and J. A. Robinson. A theory of political transitions. *American Economic Review*, 91(4):938–963, 2001.

A. Agarwal, S. Kakade, and L. F. Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

N. Akchurina. *Multi-agent reinforcement learning algorithms.* PhD thesis, University of Paderborn, 2010.

M. G. Azar, R. Munos, and H. J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

T. Başar and P. Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach.* Springer Science & Business Media, 2008.

T. Başar and G. Zaccour, editors. *Handbook of Dynamic Game Theory.* Springer International Publishing, 2018.

D. P. Bertsekas. Dynamic programming and optimal control. *Belmont, MA: Athena Scientific*, 2017.

L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.* Cambridge university press, 2006.

U. Doraszelski and J. F. Escobar. A theory of regular Markov perfect equilibria in dynamic stochastic games: Genericity, stability, and purification. *Theoretical Economics*, 5(3): 369–402, 2010. ISSN 1555-7561.

R. Ericson and A. Pakes. Markov-perfect industry dynamics: A framework for empirical work. *The Review of economic studies*, 62(1):53–82, 1995.

J. Filar and K. Vrieze. *Competitive Markov Decision Processes.* Springer, New York, NY, 1996. ISBN 978-1-4612-8481-9 978-1-4612-4054-9.

J. A. Filar, T. A. Schultz, F. Thuijsman, and O. Vrieze. Nonlinear programming and stationary equilibria in stochastic games. *Mathematical Programming*, 50(1):227–237, 1991.

A. M. Fink. Equilibrium in a stochastic n-person game. *Hiroshima Mathematical Journal*, 28(1), 1964.

P. J.-J. Herings and R. Peeters. Homotopy methods to compute equilibria in game theory. *Economic Theory*, 42(1):119–156, 2010.

P. J.-J. Herings, R. J. Peeters, et al. Stationary equilibria in stochastic games: Structure, selection, and computation. *Journal of Economic Theory*, 118(1):32–60, 2004.

K. Hinderer. Lipschitz continuity of value functions in Markovian decision processes. *Mathematical Methods of Operations Research*, 62(1):3–22, Sep 2005. ISSN 1432-5217.

A. J. Hoffman and R. M. Karp. On nonterminating stochastic games. *Management Science*, 12(5):359–370, 1966.

A. Jaśkiewicz and A. S. Nowak. Robust Markov perfect equilibria. *Journal of Mathematical Analysis and Applications*, 419(2):1322–1332, 2014.

S. M. Kakade. *On the sample complexity of reinforcement learning.* PhD thesis, University College, London, 2003.

M. Kearns and S. Singh. Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in neural information processing systems*, pages 996–1002, 1999.

O. Krupnik, I. Mordatch, and A. Tamar. Multi-agent reinforcement learning with multi-step generative models. *arXiv preprint arXiv:1901.10251*, Nov. 2019.

G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33, 2020.

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

M. L. Littman. Value-function reinforcement learning in Markov games. *Cognitive systems research*, 2(1):55–66, 2001.

G. J. Mailath and L. Samuelson. *Repeated Games and Reputations: Long-Run Relationships.* Oxford University Press, 2006.

E. Maskin and J. Tirole. A theory of dynamic oligopoly, I: Overview and quantity competition with large fixed costs. *Econometrica: Journal of the Econometric Society*, pages 549–569, 1988a.

E. Maskin and J. Tirole. A theory of dynamic oligopoly, II: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica: Journal of the Econometric Society*,

pages 571–599, 1988b.
E. Maskin and J. Tirole. Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory*, 100(2):191–219, 2001.
A. Müller. How does the value function of a Markov decision process depend on the transition probabilities? *Mathematics of Operations Research*, 22(4):872–885, 1997.
A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.
J. Pérolat, F. Strub, B. Piot, and O. Pietquin. Learning Nash equilibrium for general-sum Markov games from batch data. In *Artificial Intelligence and Statistics*, pages 232–241. PMLR, 2017.
H. Prasad, P. LA, and S. Bhatnagar. Two-timescale algorithms for learning Nash equilibria in general-sum stochastic games. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1371–1379, 2015.
P. D. Rogers. *Nonzero-sum stochastic games.* PhD thesis, University of California, Berkeley, 1969.
S. Sengupta, A. Chowdhary, D. Huang, and S. Kambhampati. General sum markov games for strategic detection of advanced persistent threats using moving target defense in cloud networks. In *International Conference on Decision and Game Theory for Security*, pages 492–512. Springer, 2019.
L. S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100, 1953.
Y. Shoham, R. Powers, and T. Grenager. Multi-agent reinforcement learning: a critical survey. Technical report, Stanford University, 2003.
A. Sidford, M. Wang, X. Wu, L. F. Yang, and Y. Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5192–5202, 2018.
A. Sidford, M. Wang, L. Yang, and Y. Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002. PMLR, 2020.
E. Solan. *A Course in Stochastic Game Theory.* Cambridge University Press, 2021.
B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. R. G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *Conference on Learning Theory*, 2008.
J. Subramanian, A. Sinha, R. Seraj, and A. Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *arXiv preprint arXiv:2010.08843*, 2020.
R. S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *International Conference on Machine Learning*, pages 216–224. San Francisco (CA), 1990.
M. Takahashi. Equilibrium points of stochastic non-cooperative n-person games. *Hiroshima Mathematical Journal*, 28(1), Jan. 1964.
M. M. Tidball and E. Altman. Approximations in dynamic zero-sum games I. *SIAM Journal on Control and Optimization*, 34(1):311–328, Jan. 1996.
M. M. Tidball, O. Pourtallier, and E. Altman. Approximations in dynamic zero-sum games II. *SIAM Journal on Control and Optimization*, 35(6):2101–2117, Nov. 1997.
O. J. Vrieze. *Stochastic games with finite state and action spaces.* CWI, Jan. 1987. ISBN 978-90-6196-313-4.
T. Wang, X. Bao, I. Clavera, J. Hoang, Y. Wen, E. Langlois, S. Zhang, G. Zhang, P. Abbeel, and J. Ba. Benchmarking Model-Based Reinforcement Learning. *arXiv preprint arXiv:1907.02057*, July 2019.
W. Whitt. Representation and approximation of noncooperative sequential games. *SIAM Journal on Control and Optimization*, 18(1):33–48, 1980.
K. Zhang, S. Kakade, T. Basar, and L. Yang. Model-based multi-agent rl in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33, 2020.
K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages

321–384, 2021a.

W. Zhang, X. Wang, J. Shen, and M. Zhou. Model-based Multi-agent Policy Optimization with Adaptive Opponent-wise Rollouts. In *International Joint Conference on Artificial Intelligence*. Montreal, Canada, 2021b.

M. Zinkevich, A. Greenwald, and M. Littman. Cyclic equilibria in Markov games. In *Neural Information Processing Systems*, pages 1641–1648, 2006.