Robustness of Markov perfect equilibrium to model approximations in general-sum dynamic games

Jayakumar Subramanian, Amit Sinha, Aditya Mahajan

***Abstract*— Dynamic games (also called stochastic games or Markov games) are an important class of games for modeling multi-agent interactions. In many situations, the dynamics and reward functions of the game are learnt from past data and are therefore approximate. In this paper, we study the robustness of Markov perfect equilibrium to approximations in reward and transition functions. Using approximation results from Markov decision processes, we show that the Markov perfect equilibrium of an approximate (or perturbed) game is always an approximate Markov perfect equilibrium of the original game. We provide explicit bounds on the approximation error in terms of three quantities: (i) the error in approximating the reward functions, (ii) the error in approximating the transition function, and (iii) a property of the value function of the MPE of the approximate game. The second and third quantities depend on the choice of metric on probability spaces. We also present coarser upper bounds which do not depend on the value function but only depend on the properties of the reward and transition functions of the approximate game. We illustrate the results via a numerical example.**

I. Introduction

Dynamic games (also called stochastic games or Markov games) are a commonly used framework to model strategic interaction between multiple players interacting in a dynamic environment. Examples include applications in industrial organization, finance, political economics, and many others [1]. Starting with the seminal work of [2], several variations of dynamic games have been considered in the literature [3].

In 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 is publicly monitored by all players. Attention is typically restricted to a refinement of subgame perfect equilibria known as Markov perfect equilibria (MPE), where all players play a Markov strategy (i.e., choose their actions as a (possibly randomized) function of the current state).

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 mappings (called the Shapley operator) for two-player zero-sum games is a contraction [2]. 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 in the zero-sum case) can be computed via recursive operations of the Shapley operator [2], [4]. 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 [5], [6]. Consequently, 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 [7] and homotopy methods [8], [9].

In many situations, the dynamics and reward functions of the game are learnt from past data and are therefore approximate. In this paper, we study the robustness of MPE to the approximations in reward and transition functions. Such robustness is well understood for Markov decision processes (see [10] and follow-up work) and zero-sum dynamic games [11], [12]. In this paper, we address the question of robustness for general-sum dynamic 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 the 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. These results are useful for developing numerical methods for dynamic games with continuous or large state and/or action spaces.

Our notion of robustness is different from that of robust control [13] and robust Markov perfect equilibrium [14], 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 similar in spirit to [15] (also see [16]), which shows that almost all dynamic games have a finite number of MPEs and these equilibria can be approximated by equilibria of nearby games. The result of [15] 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 [15] 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

Jayakumar Subramanian is with Media and Data Science Research Lab, Digital Experience Cloud, Adobe Inc., Noida, UP, India. Email: jasubram @adobe.com

Amit Sinha and Aditya Mahajan are with the Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada. Emails: amit.sinha@mail.mcgill.ca, aditya.mahajan@mcgill.ca

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 [17], 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 ε_1-MPE of the approximate game is an ε_2-MPE of the original game and an explicit relationship between ε_1 and ε_2 is established. Our results are similar in spirit but the specific details are different.

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

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_t^i denotes the action of player i at time t. For sequence of variables $\{s_t\}_{t\geq 1}$, 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.,

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

II. System model and main result

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.

A. 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 $A_t = (A_t^i)_{i\in\mathcal{N}}$ to denote the action profile of all players at time t. Given an action profile $A_t = (A_t^i)_{i\in\mathcal{N}}$ and a player $j \in \mathcal{N}$, we use the notation $A_t^{-j} = (A_t^i)_{i\in\mathcal{N}\setminus\{j\}}$ to denote the action profile of all players except j.
- $\mathsf{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ 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 [2], we assume that each player chooses its action according to a time homogeneous Markov strategy. Let $\Pi^i = \{\pi^i : \mathcal{S} \to \Delta(\mathcal{A}^i)\}$ 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.

There are two solution concepts commonly used for dynamic 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 $i \in \mathcal{N}$,

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

A Markov perfect equilibrium can be viewed as a refinement of subgame perfect equilibrium where all players play Markov strategies. For games with finite state and action spaces, a Markov perfect equilibrium always exists [5].

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 $i \in \mathcal{N}$,

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

B. Integral probability metrics

Our results rely of a class of metrics on probability spaces known as integral probability metrics (IPMs) [18].

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 $\mu, \nu \in \Delta(\mathcal{X})$ with respect to the function class $\mathfrak{F}$ is defined as

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

Two specific forms of IPMs are used in this paper:

1) **Total variation distance**: If $\mathfrak{F}$ is chosen as $\mathfrak{F}^{\mathrm{TV}} := \{f : \operatorname{span}(f) \leq 1\}$, then $d_{\mathfrak{F}}$ is the total variation distance.

2) **Wasserstein distance**: If $\mathcal{X}$ is a metric space and $\mathfrak{F}$ is chosen as $\mathfrak{F}^{\mathrm{W}} := \{f : \mathrm{Lip}(f) \leq 1\}$ (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:

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

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, $\rho_{\mathfrak{F}^{\mathrm{W}}}(f) = \mathrm{Lip}(f)$.

C. 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 $a \in \mathcal{A}$, we have $|r^i(s,a) - \hat{r}^i(s,a)| \leq \varepsilon$.
2) **Transition approximation**: For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have $d_{\mathfrak{F}}(\mathsf{P}(\cdot \mid s,a), \widehat{\mathsf{P}}(\cdot \mid s,a)) \leq \delta$.

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*

$$\alpha^i \leq 2\left(\varepsilon + \frac{\gamma\delta\rho_{\mathfrak{F}}(\hat{V}^i_{(\hat{\pi}^i, \hat{\pi}^{-i})})}{1-\gamma}\right). \tag{6}$$

See Sec. IV for the proof.

Remark 1 The approximation bound (6) on α^i depends on the expected discounted payoff profile of strategy $\hat{\pi}$ in game $\widehat{\mathscr{G}}$. The only information needed about the original game $\mathscr{G}$ are the modeling errors (ε, δ). If the approximate game was estimated using sampling methods, then (ε, δ) can be computed in terms of the number of samples using concentration of measure arguments.

Next, we present easier to compute, but looser upper bounds on α^i.

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

$$\alpha^i \leq 2\left(\varepsilon + \frac{\gamma\delta\,\mathrm{span}(\hat{r}^i)}{(1-\gamma)}\right). \tag{7}$$

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 $a \in \mathcal{A}$,

$$\begin{aligned} \left|r^i(s_1,a) - r^i(s_2,a)\right| &\leq L_r d(s_1, s_2), \\ d_{\mathfrak{F}^{\mathrm{W}}}(\mathsf{P}(\cdot|s_1,a), \mathsf{P}(\cdot|s_2,a)) &\leq L_{\mathsf{P}} d(s_1, s_2). \end{aligned}$$

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*

$$\alpha^i \leq 2\left(\varepsilon + \frac{\gamma L_r \delta}{(1-\gamma L_{\mathsf{P}})}\right). \tag{8}$$

Remark 2 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, [18], [19]. The choice of the metric often depends on the specific properties of the model.

D. An illustrative example

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

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

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

$$\hat{r}(1) = \begin{bmatrix} (0.990.40) & (0.691.00) \\ (0.300.99) & (0.810.71) \end{bmatrix}, \hat{r}(2) = \begin{bmatrix} (0.590.70) & (0.690.61) \\ (0.300.80) & (0.190.21) \end{bmatrix},$$
$$\hat{r}(3) = \begin{bmatrix} (0.19,0.59) & (0.09,0.70) \\ (0.59,0.69) & (0.50,0.30) \end{bmatrix},$$

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

[7]) are as follows:

$$\hat{\pi}^1 = \begin{bmatrix} 0.33 & 0.67 \\ 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix}, \qquad \hat{\pi}^2 = \begin{bmatrix} 0.13 & 0.87 \\ 1.00 & 0.00 \\ 1.00 & 0.00 \end{bmatrix}, \tag{9}$$

$$\hat{V}^1_{\hat{\pi}} = \begin{bmatrix} 0.6327 \\ 0.6170 \\ 0.6187 \end{bmatrix}, \qquad \hat{V}^2_{\hat{\pi}} = \begin{bmatrix} 0.7258 \\ 0.7148 \\ 0.7148 \end{bmatrix}. \tag{10}$$

In (9), the strategy is described as $\pi^i = [\pi^i(a^i|s)]_{s\in\mathcal{S}, a^i\in\mathcal{A}^i}$.

For strategy $\hat{\pi}$ in (9), 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. IV). These are given by

$$V^1_{\hat{\pi}} = \begin{bmatrix} 0.6341 \\ 0.6192 \\ 0.6209 \end{bmatrix}, \qquad V^2_{\hat{\pi}} = \begin{bmatrix} 0.7252 \\ 0.7142 \\ 0.7154 \end{bmatrix}, \tag{11}$$

$$V^1_{(*,\hat{\pi}^2)} = \begin{bmatrix} 0.6394 \\ 0.6222 \\ 0.6241 \end{bmatrix}, \qquad V^2_{(\hat{\pi}^1,*)} = \begin{bmatrix} 0.7280 \\ 0.7158 \\ 0.7171 \end{bmatrix}. \tag{12}$$

Note that

$$\alpha^1_* = \|V^1_{(*,\hat{\pi}^2)} - V^1_{\hat{\pi}}\|_\infty = 0.005300, \tag{13a}$$

$$\alpha^2_* = \|V^2_{(\hat{\pi}^1,*)} - V^2_{\hat{\pi}}\|_\infty = 0.002785. \tag{13b}$$

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.

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

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

and

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

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

$$\alpha \le 2\times 0.01 + 2\times 0.9\times 0.05 \begin{bmatrix} 0.015684 \\ 0.010990 \end{bmatrix} = \begin{bmatrix} 0.0341 \\ 0.0299 \end{bmatrix}.$$

Note that the above is an upper bound of α_* obtained in (13).

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

$$\alpha \le 2\times 0.01 + 2\times 0.9\times 0.10 \begin{bmatrix} 0.015684 \\ 0.010990 \end{bmatrix} = \begin{bmatrix} 0.0482 \\ 0.0398 \end{bmatrix}.$$

Note that the above is an upper bound of α_* obtained in (13).

The above example shows that even when (ε, δ) are significant, the bound of Theorem 1 is loose by only a small multiplicative factor of 6–15.

III. Preliminaries on MDPs

A. 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.
- $\mathsf{P} : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ 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

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

- $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ 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 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. The performance of a Markov strategy π starting from initial state $s_0 \in \mathcal{S}$ is given by:

$$V_\pi(s_0) = (1-\gamma)\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(S_t, A_t) \;\middle|\; S_0 = s_0\right], \tag{14}$$

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

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

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

$$V_\pi(s) \ge V_{\tilde{\pi}}(s) - \alpha, \quad \forall s \in \mathcal{S}. \tag{16}$$

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 $s \in \mathcal{S}$

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

$$[\mathcal{B}_* v](s) = \max_{a\in\mathcal{A}}\Big[(1-\gamma)r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathsf{P}(s'|s,a) v(s')\Big]. \tag{18}$$

Then, optimal and approximately optimal strategies can be characterized using the Bellman operators as shown below. These are standard results. See [20], 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 $V = \mathcal{B}_\pi V$ and $V = \mathcal{B}_* V$.*

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 $V_\pi \geq V_* - \alpha$.*

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 $r(S_t, A_t) \in [\min(r), \max(r)]$. ∎

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

$$\begin{aligned} \big|r(s_1,a) - r(s_2,a)\big| &\leq L_r d(s_1,s_2), \\ d_{\mathfrak{F}^{\mathrm{W}}}(\mathsf{P}(\cdot|s_1,a), \mathsf{P}(\cdot|s_2,a)) &\leq L_{\mathsf{P}} d(s_1,s_2). \end{aligned}$$

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 [21, Theorem 4.2]. ∎

B. 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 $a \in \mathcal{A}$, we have $|r(s,a) - \hat{r}(s,a)| \leq \varepsilon$.
2) **Transition approximation**: For all $s \in \mathcal{S}$, and $a \in \mathcal{A}$, we have $d_{\mathfrak{F}}(\mathsf{P}(\cdot|s,a), \widehat{\mathsf{P}}(\cdot|s,a)) \leq \delta$.

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

Theorem 2 *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 the optimal strategy of $\widehat{\mathcal{M}}$ and $\hat{V}$ be the corresponding value function. Then $\hat{\pi}$ is an α-optimal strategy of $\mathcal{M}$ with*

$$\alpha \leq 2\Big(\varepsilon + \frac{\gamma \delta \rho_{\mathfrak{F}}(\hat{V})}{(1-\gamma)}\Big). \tag{19}$$

PROOF The result follows from [19, Theorem 24] applied to MDPs. Also see, [10, Theorem 4.2]. ∎

We now present two instances of the result of Theorem 2, which follow from Lemmas 1 and 2.

Corollary 3 *If the function class $\mathfrak{F}$ in Theorem 2 is $\mathfrak{F}^{\mathrm{TV}}$, then*

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

Corollary 4 *If the function class $\mathfrak{F}$ in Theorem 2 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*

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

IV. PROOF OF THE MAIN RESULTS

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

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 $s \in \mathcal{S}$,

$$\begin{aligned} [\mathcal{B}^i_{(\pi^i,\pi^{-i})} v](s) = \sum_{a \in \mathcal{A}} \pi(a|s) \Big[(1-\gamma) r^i(s,a) \\ + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s,a) v(s') \Big]. \end{aligned}$$

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 $s \in \mathcal{S}$,

$$\begin{aligned} [\mathcal{B}^i_{(*,\pi^{-i})} v](s) = \max_{a^i \in \mathcal{A}^i} \Bigg[\sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \\ \times \Big[(1-\gamma) r^i(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s'|s,a) v(s') \Big] \Bigg]. \end{aligned}$$

Now, MPE and approximate MPE can be characterized using the Bellman operators. These are standard results [3].

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 for all $i \in \mathcal{N}$, $V^i = \mathcal{B}^i_{(\pi^i, \pi^{-i})} V^i$, and $V^i = \mathcal{B}^i_{(*, \pi^{-i})} V^i$.*

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 for all $i \in \mathcal{N}$, $V^i_\pi \geq V^i_{(*,\pi^{-i})} - \alpha^i$.*

B. 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 $\mathsf{P}^i_{\pi^{-i}} : \mathcal{S} \times \mathcal{A}^i \to \Delta(\mathcal{S})$ is given by

$$\mathsf{P}^i_{\pi^{-i}}(s'|s,a^i) = \sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) \mathsf{P}(s'|s,(a^i,a^{-i})), \tag{20}$$

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

$$r^i_{\pi^{-i}}(s, a^i) = \sum_{a^{-i} \in \mathcal{A}^{-i}} \pi^{-i}(a^{-i}|s) r^i(s, (a^i, a^{-i})). \quad (21)$$

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 $\mathcal{M}^i_{\pi^{-i}}$.*

C. Relationship between MDPs for a strategy profile

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.

Lemma 3 *For any player $i \in \mathcal{N}$, MDP $\widehat{\mathcal{M}}^i_{\hat{\pi}^{-i}}$ is an (ε, δ) approximation of MDP $\mathcal{M}^i_{\hat{\pi}^{-i}}$.*

The proof is omitted due to space constraints.

D. Proof of the Theorem 1

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

1) From Cor. 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 3, 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 Thm 2, 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 Thm. 2. (Also see Cor. 3 and 4).
3) Since the above results hold for all $i \in \mathcal{N}$, Cor. 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 Thm. 2.
4) The specific formulas for α follow from Cor. 3 and 4.

V. Conclusion

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

Finding bounds on approximation error as a function of model approximation is a critical step in developing sample complexity bounds of model-based RL algorithms. Therefore, we believe that the approximation results presented in this paper can serve as a building block for model-based MARL. The results are also useful to develop numerical methods for computing approximate MPEs in games with continuous or large state and/or action spaces.

VI. 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.

References

[1] T. Başar and G. Zaccour, Eds., *Handbook of Dynamic Game Theory*. Springer International Publishing, 2018.
[2] L. S. Shapley, "Stochastic games," *Proceedings of the National Academy of Sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.
[3] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*. New York, NY: Springer, 1996.
[4] A. J. Hoffman and R. M. Karp, "On nonterminating stochastic games," *Management Science*, vol. 12, no. 5, pp. 359–370, 1966.
[5] A. M. Fink, "Equilibrium in a stochastic n-person game," *Hiroshima Mathematical Journal*, vol. 28, no. 1, 1964.
[6] M. Takahashi, "Equilibrium points of stochastic non-cooperative n-person games," *Hiroshima Mathematical Journal*, vol. 28, no. 1, Jan. 1964.
[7] J. A. Filar, T. A. Schultz, F. Thuijsman, and O. Vrieze, "Nonlinear programming and stationary equilibria in stochastic games," *Mathematical Programming*, vol. 50, no. 1, pp. 227–237, 1991.
[8] P. J.-J. Herings, R. J. Peeters *et al.*, "Stationary equilibria in stochastic games: Structure, selection, and computation," *Journal of Economic Theory*, vol. 118, no. 1, pp. 32–60, 2004.
[9] P. J.-J. Herings and R. Peeters, "Homotopy methods to compute equilibria in game theory," *Economic Theory*, vol. 42, no. 1, pp. 119–156, 2010.
[10] A. Müller, "How does the value function of a Markov decision process depend on the transition probabilities?" *Mathematics of Operations Research*, vol. 22, no. 4, pp. 872–885, 1997.
[11] M. M. Tidball and E. Altman, "Approximations in dynamic zero-sum games I," *SIAM Journal on Control and Optimization*, vol. 34, no. 1, pp. 311–328, Jan. 1996.
[12] M. M. Tidball, O. Pourtallier, and E. Altman, "Approximations in dynamic zero-sum games II," *SIAM Journal on Control and Optimization*, vol. 35, no. 6, pp. 2101–2117, Nov. 1997.
[13] T. Başar and P. Bernhard, *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.
[14] A. Jaśkiewicz and A. S. Nowak, "Robust Markov perfect equilibria," *Journal of Mathematical Analysis and Applications*, vol. 419, no. 2, pp. 1322–1332, 2014.
[15] U. Doraszelski and J. F. Escobar, "A theory of regular Markov perfect equilibria in dynamic stochastic games: Genericity, stability, and purification," *Theoretical Economics*, vol. 5, no. 3, pp. 369–402, 2010.
[16] E. Maskin and J. Tirole, "Markov perfect equilibrium: I. observable actions," *Journal of Economic Theory*, vol. 100, no. 2, pp. 191–219, 2001.
[17] W. Whitt, "Representation and approximation of noncooperative sequential games," *SIAM Journal on Control and Optimization*, vol. 18, no. 1, pp. 33–48, 1980.
[18] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, vol. 29, no. 2, pp. 429–443, 1997.
[19] 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.
[20] D. P. Bertsekas, "Dynamic programming and optimal control," *Belmont, MA: Athena Scientific*, 2017.
[21] K. Hinderer, "Lipschitz continuity of value functions in Markovian decision processes," *Mathematical Methods of Operations Research*, vol. 62, no. 1, pp. 3–22, Sep 2005.