RESEARCH ARTICLE

National Science Review
10: nwac256, 2023
https://doi.org/10.1093/nsr/nwac256
Advance access publication 22 November 2022

COMPUTER SCIENCE

On the complexity of computing Markov perfect equilibrium in general-sum stochastic games

Xiaotie Deng[1,2,*], Ningyuan Li[1,*], David Mguni[3], Jun Wang[4] and Yaodong Yang[2,*]

ABSTRACT

Similar to the role of Markov decision processes in reinforcement learning, Markov games (also called stochastic games) lay down the foundation for the study of multi-agent reinforcement learning and sequential agent interactions. We introduce approximate Markov perfect equilibrium as a solution to the computational problem of finite-state stochastic games repeated in the infinite horizon and prove its **PPAD**-completeness. This solution concept preserves the Markov perfect property and opens up the possibility for the success of multi-agent reinforcement learning algorithms on static two-player games to be extended to multi-agent dynamic games, expanding the reign of the **PPAD**-complete class.

Keywords: Markov game, multi-agent reinforcement learning, Markov perfect equilibrium, **PPAD**-completeness, stochastic game

[1]Center on Frontiers of Computing Studies, School of Computer Science, Peking University, Beijing 100091, China; [2]Center for Multi-Agent Research, Institute for AI, Peking University, Beijing 100091, China; [3]Huawei UK, London WC1E 6BT, UK and [4]Computer Science, University College London, London WC1E 6BT, UK

*__Corresponding authors__. E-mails: xiaotie@pku.edu.cn; liningyuan@pku.edu.cn; yaodong.yang@pku.edu.cn

Received 25 February 2022; **Revised** 11 August 2022; **Accepted** 6 October 2022

INTRODUCTION

Shapley [1] introduced stochastic games (SGs) to study the dynamic non-cooperative multi-player game, where each player simultaneously and independently chooses an action at each round for a reward. According to their current state and the chosen actions, the next state is determined by a probability distribution specified *a priori*. Shapley's work includes the first proof of the existence of a stationary strategy profile under which no agent has an incentive to deviate, in two-player zero-sum SGs. Next, the existence of equilibrium in stationary strategies was extended to multi-player general-sum SGs by Fink [2]. Such a solution concept (known as *Markov perfect equilibrium* (MPE) [3]) captures the dynamics of multi-player games.

Because of its generality, the framework of SGs has enlightened a sequence of studies [4] on a wide range of real-world applications ranging from advertising and pricing [5], species interaction game modeling in fisheries [6], traveling inspection [7] and gaming AIs [8]. As a result, developing algorithms to compute MPE in SGs has become one of the key subjects in this extremely rich research domain, using approaches from applied mathematics, economics, operations research, computer science and artificial intelligence (see, e.g., [9]).

The concept of the SG underpins many AI and machine learning studies. The optimal policy making of Markov decision processes (MDPs) captures the central problem of a single agent interacting with its environment, according to Sutton and Barto [10]. In multi-agent reinforcement learning (MARL) [11,12], SG extends to incorporate the dynamic nature in multi-agent strategic interactions, to study optimal decision making and subsequently equilibria in multi-player games [13,14].

For two-player zero-sum (discounted) SGs, the game-theoretical equilibrium is closely related to the optimization problem in MDPs as the opponent is purely adversarial [15,16]. On the other hand, solving general-sum SGs has been possible only under strong assumptions [2,17]. Zinkevich *et al.* [18] demonstrated that, for the entire class of value iteration methods, it is difficult to find stationary Nash equilibrium (NE) policies in general-sum SGs. This has led to few existing MARL algorithms to general-sum SGs. Known approaches have either studied special cases of SGs [19,20] or ignored the dynamic nature to limit the study to the weaker notion of Nash equilibrium [21].

Recently, Solan and Vieille [22] reconfirmed the importance of the existence of a stationary strategy profile as having several philosophical implications.

First, it is conceptually straightforward. Second, *past play affects the players' future behavior only through the current state*. Third, subsequently and most importantly, *equilibrium behavior does not involve non-credible threats, a property that is stronger than the equilibrium property, and viewed as highly desirable* [23].

Surprisingly, the complexity of finding an MPE in an SG remains an open problem, although an SG was proposed more than sixty years ago and despite its importance. While fruitful studies have been conducted on zero-sum SGs, we still know little about the complexity of solving general-sum SGs. It is clear that solving MPE in (infinite-horizon) SGs is at least **PPAD**-hard, since solving a two-player NE in one-shot SGs is already complete in this computational class [24,25] defined by Papadimitriou [26]. This suggests that it is unlikely to have polynomial-time algorithms in general-sum stochastic games for two players. Yet, with complications involved in the general-sum and dynamic settings, the unresolved challenge has been: Can solving MPE in general-sum SGs be anywhere complete in computational complexity classes?

We answer the above question in the positive, proving the computation of an approximate MPE in SGs equivalent to that of a Nash equilibrium in a single state setting, and subsequently showing its **PPAD**-completeness. It opens up the possibility to develop MARL algorithms to work for the general-sum SGs in the same way as for an ordinary Nash equilibrium computation.

Intuitions and a sketch of our main ideas

Computational studies on problem solving build understanding on various types of reduction. After all, computations carried out on computers are eventually reduced to AND/OR/NOT gates on electronic circuits.

To prove that a problem is **PPAD**-complete, we need to prove that it is in the class, and that it can be used as a base to solve any other problem in this class (for its hardness). More formally, the reduction needs to be carried out in polynomial (with respect to the input size) time. Nash equilibrium computation of two-player normal-form games [27] is arguably the most prominent **PPAD**-complete problem [24,25]. When one stochastic game has only one state and the discount factor $\gamma = 0$, then finding an MPE is equivalent to finding a Nash equilibrium in the corresponding normal-form game. The **PPAD**-hardness of finding an MPE follows immediately. Our main result is to prove the **PPAD** membership property of computing an approximate MPE (Lemma 2 below).

Firstly, we construct a function f on the strategy profile space, such that each strategy profile is a fixed point of f if and only if it is an MPE of the stochastic game (Theorem 2 below). Furthermore, we prove that the function f is continuous (λ-Lipschitz by Lemma 3 below), so that fixed points are guaranteed to exist by the Brouwer fixed point theorem.

Secondly, we prove that the function f has some 'good' approximation properties. Let $|\mathcal{SG}|$ be the input size of a stochastic game. If we can find a $\text{poly}(|\mathcal{SG}|)\epsilon^2$-approximate fixed point π of f, i.e. $\|f(\pi) - \pi\|_\infty \leq \text{poly}(|\mathcal{SG}|)\epsilon^2$, where π is a strategy profile, then π is an ϵ-approximate MPE for the stochastic game (combining Lemma 5 and Lemma 6 below). So our goal converts to finding an approximate fixed point of a Lipschitz function.

Finally, our **PPAD** membership follows from the theorem that computation of the approximate Brouwer fixed point of a Lipschitz function is **PPAD**-complete, as shown in the seminal paper by Papadimitriou [26].

Related work

In practice, MARL methods are most often applied to compute the MPE of an SG based on the interactions between agents and the environment. Their uses can be classified in two different settings: *online* and *offline*. In the offline setting (also known as the batch setting [21]), the learning algorithm controls all players in a centralized way, hoping that the learning dynamics can eventually lead to an MPE by using a limited number of interaction samples. In the online setting, the learner controls only one of the players to play with an arbitrary group of opponents in the game, assuming unlimited access to the game environment. The central focus is often on the *regret*: the difference between the learner's total reward during learning versus that of a benchmark measure in hindsight.

In the offline setting, two-player zero-sum (discounted) SGs have been extensively studied. Since the opponent is purely adversarial in zero-sum SGs, the process of seeking the worst-case optimality for each player can be thought of as solving MDPs. As a result, (approximate) dynamic programming methods [28,29] such as least-squares policy iteration [30] and fitted value iteration [31] or neural fitted Q iteration [32] can be adopted to solve SGs [33–36]. Under this setting, policy-based methods [37,38] can also be applied. However, directly applying existing MDP solvers on general-sum SGs is problematic. Since solving two-player NE in general-sum normal-form games (i.e. one-shot SGs) is well known to be **PPAD**-complete [24,25], the complexity of MPE in general-sum SGs is expected to be at least **PPAD**-hard. Although early attempts such as

Nash-Q learning [39], correlated-Q learning [40], friend-or-foe Q-learning [41] have been made to solve general-sum SGs under strong assumptions, Zinkevich *et al.* [18] demonstrated that none in the entire class of value iteration methods can find stationary NE policies in general-sum SGs. The difficulties on both the complexity side and the algorithmic side have led to few existing MARL algorithms for general-sum SGs. Successful approaches either assume knowing the complete information of the SG such that solving MPE can be turned into an optimization problem [42], or prove the convergence of batch RL methods to a weaker notion of NE [21].

In the online setting, agents aim to minimize their regret by trial and error. One of the most well-known online algorithms is R-MAX [43], which studies (average-reward) zero-sum SGs and provides a polynomial (in game size and error parameter) regret bound while competing against an arbitrary opponent. Following the same regret definition, UCSG [44] improved R-MAX and achieved a sublinear regret, but still in the two-player zero-sum SG setting. When it comes to MARL solutions, Littman [13] proposed a practical solution named Minimax-Q that replaces the max operator with the minimax operator. Asymptotic convergence results of Minimax-Q were developed in both tabular cases [45] and value function approximations [46]. To avoid the overly pessimism property by playing the minimax value for general-sum SGs, WoLF [47] was proposed to take variable steps to exploit an opponent's suboptimal policy for a higher reward on a variety of stochastic games. AWESOME [48] further generalized WoLF and achieved NE convergence in multi-player general-sum repeated games. However, outside the scope of zero-sum SGs, the question [43] of whether a polynomial time no-regret (near-optimal) MARL algorithm exists for general-sum SGs remains open.

Some recent works studied the sample complexity issue in RL and MARL algorithms, most of which considered a finite horizon. Jin *et al.* [49] proved that a variant of Q-learning with upper confidence bound exploration can achieve a near-optimal sample efficiency under episodic MDP setting. Zhang *et al.* [50] proposed a learning algorithm for episodic MDP with a regret bound close to the information theoretic lower bound. Li *et al.* [51] proposed a probably approximately correct learning algorithm for episodic RL with a sample complexity independent of the planning horizon. For general-sum MARL, Chen *et al.* [52] proved an exponential lower bound on the sample complexity of approximate Nash equilibrium even in n-player normal-form games. In the same direction, Song *et al.* [53] showed that correlated equilibrium (CE) and coarse correlated equilibrium (CCE) can be learned within a sample complexity polynomial in the maximum size of the action set of a player, rather than the size of the joint action space. Jin *et al.* [54] developed a decentralized MARL algorithm with polynomial sample complexity to learn CE and CCE.

DEFINITIONS AND THE MAIN THEOREM

Definition 1 (Stochastic game). A stochastic game is defined by a tuple of six elements $\langle n, \mathbb{S}, \mathbb{A}, P, r, \gamma \rangle$.

- By n we denote the number of agents.
- By $\mathbb{S}$ we denote the set of finite environmental states. Let $S = |\mathbb{S}|$.
- By $\mathbb{A}^i$ we denote the action space of agent i. Note that each agent i can choose different actions under different states. Without loss of generality, we assume that, for each agent i, the action space $\mathbb{A}^i$ under each state is the same. Here $\mathbb{A} = \mathbb{A}^1 \times \cdots \times \mathbb{A}^n$ is the set of agents' joint action vector. Let $A^i = |\mathbb{A}^i|$ and $A_{\max} = \max_{i \in [n]} A^i$.
- By $P : \mathbb{S} \times \mathbb{A} \to \Delta(\mathbb{S})$ we denote the transition probability, that is, at each time step, given the agents' joint action vector $a \in \mathbb{A}$, then the transition probability from state s to state s' in the next time step is $P(s'|s, a)$.
- By $r = r^1 \times \cdots \times r^n : \mathbb{S} \times \mathbb{A} \to \mathcal{R}_+^n$ we denote the reward function, that is, when agents are at state s and play the joint action vector a, then agent i will get reward $r^i(s, a)$. We assume that the rewards are bounded by $r_{\max}$.
- By $\gamma \in [0, 1)$ we denote the discount factor that specifies the degree to which the agent's rewards are discounted over time.

Each agent aims to find a behavioral strategy with Markovian property, which is conditioned on the current state of the game.

The pure strategy space of agent i is $\prod_{s \in \mathbb{S}} \mathbb{A}^i$, which means that agent i needs to select an action at each state. Note that the size of the pure strategy space of each agent is $|\mathbb{A}^i|^S$, which is already exponential in the number of states. More generally, we define the mixed behavioral strategy as follows.

Definition 2 (Behavioral strategy). A behavioral strategy of agent i is $\pi^i : \mathbb{S} \to \Delta(\mathbb{A}^i)$. For all $s \in \mathbb{S}$, $\pi^i(s)$ is a probability distribution on $\mathbb{A}^i$.

In the rest of the paper, we focus on behavioral strategy and refer to it simply as a strategy for convenience. A strategy profile π is the Cartesian product of all agents' strategies, i.e. $\pi = \pi^1 \times \cdots \times \pi^n$. We denote the probability of agent i using action a^i at state s by $\pi^i(s, a^i)$. The strategy profile of all the

agents other than agent i is denoted by π^{-i}. We use π^i, π^{-i} to represent π, and a^i, a^{-i} to represent a.

Given π, the transition probability and the reward function depend only on the current state $s \in \mathbb{S}$. Let $r^{i,\pi}(s)$ denote $\mathbb{E}_{a\sim\pi(s)}[r^i(s,a)]$ and $P^{\pi}(s'|s)$ denote $\mathbb{E}_{a\sim\pi(s)}[P(s'|s,a)]$. Fix π^{-i}; the transition probability and the reward function depend only on the current state $s \in \mathbb{S}$ and player i's action a^i. Let $r^{i,\pi^{-i}}(s,a^i)$ denote $\mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}[r^i(s,(a^i,a^{-i}))]$ and $P^{\pi^{-i}}(s'|s,a^i)$ denote $\mathbb{E}_{a^{-i}\sim\pi^{-i}(s)}[P(s'|s,(a^i,a^{-i}))]$.

For any positive integer m, let $\Delta_m := \{x \in \mathcal{R}_+^m | \sum_{i=1}^m x_i = 1\}$. Define $\Delta_{A^i}^S := \Pi_{s\in\mathbb{S}}\Delta_{A^i}$. Then, for all $s \in \mathbb{S}$, $\pi^i(s) \in \Delta_{A^i}$, $\pi^i \in \Delta_{A^i}^S$ and $\pi \in \prod_{i=1}^n \Delta_{A^i}^S$.

Definition 3 (Value function). A value function for agent i under strategy profile π, $V_i^{\pi}(s) : \mathbb{S} \to R$, gives the expected sum of its discounted rewards when the starting state is s:

$$V_i^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t \cdot \mathbb{E}[r^{i,\pi}(s_t)|s_0 = s].$$

Here, $s_0, s_1, \ldots$ is the Markov chain such that the transition matrix is P^{π}, that is, $\Pr[s_{k+1} = s'|s_k = s] = P^{\pi}(s'|s)$ for all $k = 0, 1, \ldots$. Equivalently, the value function can be defined recursively via the Bellman policy equation

$$V_i^{\pi}(s) = \mathop{\mathbb{E}}_{a\sim\pi(s)}\left[r^i(s,a) + \gamma \sum_{s'\in\mathbb{S}} P(s'|s,a)V_i^{\pi}(s')\right].$$

Definition 4 (Markov perfect equilibrium). A behavioral strategy profile π is called a Markov perfect equilibrium if, for all $s \in \mathbb{S}$, all $i \in [n]$ and all $\tilde{\pi}^i \in \Delta_{A^i}^S$,

$$V_i^{\pi}(s) \geq V_i^{\tilde{\pi}^i,\pi^{-i}}(s),$$

where $V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$ is the value function of i when its strategy deviates to $\tilde{\pi}^i$ while the strategy profile of other agents is π^{-i}.

The Markov perfect equilibrium is a solution concept within SGs in which the players' strategies depend only on the current state but not on the game history.

Definition 5 (ϵ-approximate MPE). Given $\epsilon > 0$, a behavioral strategy profile π is called an ϵ-approximate MPE if, for all $s \in \mathbb{S}$, all $i \in [n]$ and all $\tilde{\pi}^i \in \Delta_{A^i}^S$,

$$V_i^{\pi}(s) \geq V_i^{\tilde{\pi}^i,\pi^{-i}}(s) - \epsilon.$$

We use Approximate MPE to denote the computational problem of finding an approximate Markov perfect equilibrium in stochastic games, where the inputs and outputs are as follows. The input instance of problem Approximate MPE is a pair $(\mathcal{SG}, L)$, where $\mathcal{SG}$ is a stochastic game and L is a positive integer. The output of problem Approximate MPE is a strategy profile $\pi \in \prod_{i=1}^n \Delta_{A^i}^S$, also dependent only at the current state but not on its history, such that π is a $1/L$-approximate MPE of $\mathcal{SG}$. We use the notation $|\mathcal{SG}|$ to denote the input size of the stochastic game $\mathcal{SG}$.

Theorem 1 (Main theorem). Approximate MPE *is PPAD-complete.*

We note that, when $|S| = 1$ and $\gamma = 0$, a stochastic game degenerates to an n-player normal-form game. At this time, any MPE of this stochastic game is a Nash equilibrium for the corresponding normal-form game. So we have the following hardness result immediately.

Lemma 1. Approximate MPE *is PPAD-hard.*

To derive Theorem 1, we focus on the proof of **PPAD** membership of Approximate MPE in the rest of the paper.

Lemma 2. Approximate MPE *is in PPAD.*

ON THE EXISTENCE OF MPE

The original proof of the existence of MPE is from [2] and based on Kakutani's fixed point theorem. Unfortunately, proofs that are based on Kakutani's fixed point theorem in general cannot be turned into **PPAD**-membership results. We develop a proof that uses Brouwer's fixed point theorem, based on which we also prove the **PPAD** membership of Approximate MPE.

Inspired by the continuous transformation defined by Nash to prove the existence of the equilibrium point [27], we define an updating function $f : \prod_{i=1}^n \Delta_{A^i}^S \to \prod_{i=1}^n \Delta_{A^i}^S$ to adjust the strategy profile of agents in a stochastic game to establish the existence of MPE.

Let $\pi \in \prod_{i=1}^n \Delta_{A^i}^S$ be the behavioral strategy profile under discussion.

Let $Q_i^{\pi^i,\pi^{-i}}(s,a^i)$ denote the expected sum of discounted rewards of agent i if agent i uses pure action a^i at state s at the first step, and then follows π^i after that, but every other agent j maintains its strategy π^j. Formally,

$$Q_i^{\pi^i,\pi^{-i}}(s,a^i) = r^{i,\pi^{-i}}(s,a^i) + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)V_i^{\pi^i,\pi^{-i}}(s').$$

For each player $i \in [n]$ with each action $a^i \in \mathbb{A}^i$ at each state $s \in \mathbb{S}$, we define a policy update function of $\pi^i(s, a^i)$ as

$$f(\pi)^i(s, a^i) = \frac{\pi^i(s, a^i) + \max\left(0,\ Q_i^{\pi^i, \pi^{-i}}(s, a^i) - V_i^{\pi^i, \pi^{-i}}(s)\right)}{1 + \sum_{b^i \in \mathbb{A}^i} \max\left(0,\ Q_i^{\pi^i, \pi^{-i}}(s, b^i) - V_i^{\pi^i, \pi^{-i}}(s)\right)}.$$

We consider the infinite norm distance of two strategy profiles π_1 and π_2, denoted by $\|\pi_1 - \pi_2\|_\infty$: $\|\pi_1 - \pi_2\|_\infty = \max_{i \in [n], s \in \mathbb{S}, a^i \in \mathbb{A}^i} |\pi_1^i(s, a^i) - \pi_2^i(s, a^i)|$.

We first prove that the function f satisfies a continuity property, namely, is λ-*Lipschitz* for λ equal to $11nS^2A_{\max}^2 r_{\max}/(1-\gamma)^2$.

Lemma 3. *The function f is λ-Lipschitz, i.e. for every $\pi_1, \pi_2 \in \prod_{i=1}^n \Delta_{A^i}^S$ such that $\|\pi_1 - \pi_2\|_\infty \le \delta$, we have*

$$\|f(\pi_1) - f(\pi_2)\|_\infty \le \frac{11nS^2A_{\max}^2 r_{\max}}{(1-\gamma)^2}\delta.$$

Proof. At any $s \in \mathbb{S}$, pick any player $i \in [n]$. For an action $a^i \in \mathbb{A}^i$, let $M_1(a^i)$ denote $\max(0,\ Q_i^{\pi_1^i, \pi_1^{-i}}(s, a^i) - V_i^{\pi_1^i, \pi_1^{-i}}(s))$ and $M_2(a^i) = \max(0,\ Q_i^{\pi_2^i, \pi_2^{-i}}(s, a^i) - V_i^{\pi_2^i, \pi_2^{-i}}(s))$. From the next claim (proof in the Appendix),

$$\begin{aligned}
&|f(\pi_1)^i(s, a^i) - f(\pi_2)^i(s, a^i)| \\
&= \left|\frac{\pi_1^i(s, a^i) + M_1(a^i)}{1 + \sum_{b^i \in \mathbb{A}^i} M_1(b^i)} - \frac{\pi_2^i(s, a^i) + M_2(a^i)}{1 + \sum_{b^i \in \mathbb{A}^i} M_2(b^i)}\right| \\
&\le |\pi_1^i(s, a^i) - \pi_2^i(s, a^i)| + |M_1(a^i) - M_2(a^i)| \\
&\quad + \left|\sum_{b^i \in \mathbb{A}^i} M_1(b^i) - \sum_{b^i \in \mathbb{A}^i} M_2(b^i)\right|.
\end{aligned}$$

Claim 1. *For any $x, x', y, y', z, z' \ge 0$ such that $(x+y)/(1+z) \le 1$ and $(x'+y')/(1+z') \le 1$, it holds that*

$$\left|\frac{x+y}{1+z} - \frac{x'+y'}{1+z'}\right| \le |x - x'| + |y - y'| + |z - z'|.$$

Take $\delta = \|\pi_1 - \pi_2\|_\infty$; then $|\pi_1^i(s, a^i) - \pi_2^i(s, a^i)| \le \delta$ for any $s \in \mathbb{S}$, $a^i \in \mathbb{A}^i$. Next, for any $a^i \in \mathbb{A}^i$, we estimate

$$\begin{aligned}
&|M_1(a^i) - M_2(a^i)| \\
&= \left|\max\left(0,\ Q_i^{\pi_1^i, \pi_1^{-i}}(s, a^i) - V_i^{\pi_1^i, \pi_1^{-i}}(s)\right)\right. \\
&\quad \left. - \max\left(0,\ Q_i^{\pi_2^i, \pi_2^{-i}}(s, a^i) - V_i^{\pi_2^i, \pi_2^{-i}}(s)\right)\right| \\
&\le \left|\left(Q_i^{\pi_1^i, \pi_1^{-i}}(s, a^i) - V_i^{\pi_1^i, \pi_1^{-i}}(s)\right)\right. \\
&\quad \left. - \left(Q_i^{\pi_2^i, \pi_2^{-i}}(s, a^i) - V_i^{\pi_2^i, \pi_2^{-i}}(s)\right)\right| \\
&\le \left|Q_i^{\pi_1^i, \pi_1^{-i}}(s, a^i) - Q_i^{\pi_2^i, \pi_2^{-i}}(s, a^i)\right| \\
&\quad + \left|V_i^{\pi_1^i, \pi_1^{-i}}(s) - V_i^{\pi_2^i, \pi_2^{-i}}(s)\right|.
\end{aligned}$$

We should first derive an upper bound on $|r^{i, \pi_1^{-i}}(s, b^i) - r^{i, \pi_2^{-i}}(s, b^i)|$.

Claim 2. *It holds that*

$$\left|r^{i, \pi_1^{-i}}(s, b^i) - r^{i, \pi_2^{-i}}(s, b^i)\right| \le (n-1)A_{\max} r_{\max}\delta.$$

This follows from the following claim (proof in the Appendix).

Claim 3. *It holds that*

$$\sum_{b^{-1} \in \mathbb{A}^{-1}} \left|\prod_{j=2}^n \pi_1^j(s, b^j) - \prod_{j=2}^n \pi_2^j(s, b^j)\right| \le (n-1)A_{\max}\delta.$$

Similarly, we have the following claim.

Claim 4. *It holds that*

$$\sum_{b^{-1} \in \mathbb{A}^{-1}} |P^{\pi_1^{-i}}(s'|s, b^i) - P^{\pi_2^{-i}}(s'|s, b^i)| \le (n-1)A_{\max}\delta.$$

To bound $|V_i^{\pi_1^i, \pi_1^{-i}}(s) - V_i^{\pi_2^i, \pi_2^{-i}}(s)|$ for every $s \in \mathbb{S}$, we denote by V_i^π the column vector $(V_i^\pi(s))_{s \in \mathbb{S}}$, and by $r^{i,\pi}$ the column vector $(r^{i,\pi}(s))_{s \in \mathbb{S}}$ and by P^π the matrix $P^\pi(s, s')_{s, s' \in \mathbb{S}}$. By the Bellman policy equation (Definition 3), we have

$$V_i^\pi = r^{i,\pi} + \gamma P^\pi V_i^\pi,$$

which means that

$$V_i^\pi = (I - \gamma P^\pi)^{-1} r^{i,\pi}.$$

We prove in Lemma 7 below that

$$|(I - \gamma P^{\pi_1})^{-1}(s'|s) - (I - \gamma P^{\pi_2})^{-1}(s'|s)| \le \frac{nSA_{\max}\delta}{(1-\gamma)^2}$$

for all $s, s' \in \mathbb{S}$.

Now we are ready to give an upper bound on $|V_i^{\pi_1^i,\pi_1^{-i}}(s) - V_i^{\pi_2^i,\pi_2^{-i}}(s)|$ for any $s \in \mathbb{S}$. We have

$$\begin{aligned}
&\left|V_i^{\pi_1^i,\pi_1^{-i}}(s) - V_i^{\pi_2^i,\pi_2^{-i}}(s)\right| \\
&= \left|\sum_{s'\in\mathbb{S}} r^{i,\pi_1}(s')(I-\gamma P^{\pi_1})^{-1}(s'|s)\right. \\
&\quad \left. - \sum_{s'\in\mathbb{S}} r^{i,\pi_2}(s')(I-\gamma P^{\pi_2})^{-1}(s'|s)\right| \\
&\le \sum_{s'\in\mathbb{S}} r^{i,\pi_1}(s')|(I-\gamma P^{\pi_1})^{-1}(s'|s) \\
&\qquad - (I-\gamma P^{\pi_2})^{-1}(s'|s)| \\
&\quad + \sum_{s'\in\mathbb{S}} (I-\gamma P^{\pi_2})^{-1}(s'|s) \\
&\qquad \times \left|r^{i,\pi_1}(s') - r^{i,\pi_2}(s')\right| \\
&\le \sum_{s'\in\mathbb{S}} \left(r_{\max}\frac{nSA_{\max}\delta}{(1-\gamma)^2} + \frac{1}{1-\gamma} nA_{\max}r_{\max}\delta\right) \\
&= \frac{SnA_{\max}r_{\max}}{1-\gamma}\left(\frac{S}{1-\gamma}+1\right)\delta \\
&\le \frac{2nS^2A_{\max}r_{\max}}{(1-\gamma)^2}\delta,
\end{aligned}$$

where the forth line follows from $|(I-\gamma P^{\pi_2})^{-1}(s'|s)| \le 1/(1-\gamma)$, in Lemma 7 below.

Similarly, we establish a bound for $|Q_i^{\pi_1^i,\pi_1^{-i}}(s,b^i) - Q_i^{\pi_2^i,\pi_2^{-i}}(s,b^i)|$:

$$\begin{aligned}
&\left|Q_i^{\pi_1^i,\pi_1^{-i}}(s,b^i) - Q_i^{\pi_2^i,\pi_2^{-i}}(s,b^i)\right| \\
&= \left|r^{i,\pi_1^{-i}}(s,b^i) + \gamma\sum_{s'\in\mathbb{S}} P^{\pi_1^{-i}}(s'|s,b^i)V_i^{\pi_1^i,\pi_1^{-i}}(s')\right. \\
&\left. - r^{i,\pi_2^{-i}}(s,b^i) - \gamma\sum_{s'\in\mathbb{S}} P^{\pi_2^{-i}}(s'|s,b^i)V_i^{\pi_2^i,\pi_2^{-i}}(s')\right| \\
&\le |r^{i,\pi_1^{-i}}(s,b^i) - r^{i,\pi_2^{-i}}(s,b^i)| \\
&\quad + \gamma\sum_{s'\in\mathbb{S}}\left|P^{\pi_1^{-i}}(s'|s,b^i)V_i^{\pi_1^i,\pi_1^{-i}}(s')\right. \\
&\quad \left. - P^{\pi_2^{-i}}(s'|s,b^i)V_i^{\pi_2^i,\pi_2^{-i}}(s')\right| \\
&\le nA_{\max}r_{\max}\delta \\
&\quad + \gamma\sum_{s'\in\mathbb{S}}\left(P^{\pi_1^{-i}}(s'|s,b^i)\left|V_i^{\pi_1^i,\pi_1^{-i}}(s')\right.\right. \\
&\quad \left.\left. - V_i^{\pi_2^i,\pi_2^{-i}}(s')\right| + V_i^{\pi_2^i,\pi_2^{-i}}(s')\left|P^{\pi_1^{-i}}(s'|s,b^i)\right.\right. \\
&\quad \left.\left. - P^{\pi_2^{-i}}(s'|s,b^i)\right|\right) \\
&\le nA_{\max}r_{\max}\delta \\
&\quad + \gamma\left(\frac{2nS^2A_{\max}r_{\max}}{(1-\gamma)^2}\delta + S\frac{r_{\max}}{1-\gamma}nA_{\max}\delta\right) \\
&= nA_{\max}r_{\max}\delta\left(1 + \frac{2\gamma S^2}{(1-\gamma)^2} + \frac{\gamma S}{1-\gamma}\right) \\
&\le \frac{3S^2nA_{\max}r_{\max}\delta}{(1-\gamma)^2}.
\end{aligned}$$

For any $b^i \in \mathbb{A}^i$, we have

$$\begin{aligned}
&|M_1(b^i) - M_2(b^i)| \\
&\quad \le \left|Q_i^{\pi_1^i,\pi_1^{-i}}(s,b^i) - Q_i^{\pi_2^i,\pi_2^{-i}}(s,b^i)\right| \\
&\qquad + \left|V_i^{\pi_1^i,\pi_1^{-i}}(s) - V_i^{\pi_2^i,\pi_2^{-i}}(s)\right| \\
&\quad \le \frac{5S^2nA_{\max}r_{\max}\delta}{(1-\gamma)^2}.
\end{aligned}$$

Thus, for any $s \in \mathbb{S}$ and any $a^i \in \mathbb{A}^i$, we obtain

$$\begin{aligned}
&|f(\pi_1)^i(s,a^i) - f(\pi_2)^i(s,a^i)| \\
&\quad \le |\pi_1^i(s,a^i) - \pi_2^i(s,a^i)| \\
&\qquad + |M_1(a^i) - M_2(a^i)| \\
&\quad + \sum_{b^i\in\mathbb{A}^i} |M_1(b^i) - M_2(b^i)| \\
&\quad \le \delta + \frac{5S^2nA_{\max}r_{\max}\delta}{(1-\gamma)^2} + A_{\max}\frac{5S^2nA_{\max}r_{\max}\delta}{(1-\gamma)^2} \\
&\quad \le \frac{11nS^2A_{\max}^2r_{\max}}{(1-\gamma)^2}\delta.
\end{aligned}$$

This completes the proof of Lemma 3. □

Now we can establish the existence of MPE by the Brouwer fixed point theorem.

Theorem 2. *For any stochastic game* $\langle n, \mathbb{S}, \mathbb{A}, P, r, \gamma\rangle$*, a strategy profile* π *is an MPE if and only if it is a fixed point of the function f, i.e.* $f(\pi) = \pi$*. Furthermore, the function f has at least one fixed point.*

Proof. We first show that the function f has at least one fixed point. Brouwer's fixed point theorem states that, for any continuous function mapping a compact convex set to itself, there is a fixed point. Note that f is a function mapping a compact convex set to itself. Also, f is continuous by Lemma 3. Therefore, the function f has at least one fixed point.

We then prove that a strategy profile π is an MPE if and only if it is a fixed point of f.

$\Rightarrow$: For the necessity, suppose that π is an MPE; then, by Definition 4, we have, for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each policy

$\tilde{\pi}^i \in \Delta_{A^i}^S$, $V_i^{\pi^i,\pi^{-i}}(s) \geq V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$. By Lemma 4 to be proven next, we have, for any action $a^i \in \mathbb{A}^i$, $V_i^{\pi^i,\pi^{-i}}(s) \geq Q_i^{\pi^i,\pi^{-i}}(s,a^i)$, which implies that $\max(0, Q_i^{\pi^i,\pi^{-i}}(s,a^i) - V_i^{\pi^i,\pi^{-i}}(s)) = 0$. Then, for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$, $(f(\pi))^i(s,a^i) = \pi^i(s,a^i)$. It follows that π is a fixed point of f.

$\Leftarrow$: For the proof of the sufficiency part, let π be a fixed point of f. Then, for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$,

$$\begin{aligned}
&\pi^i(s,a^i) \\
&= (f(\pi))^i(s,a^i) \\
&= \frac{\pi^i(s,a^i) + \max\left(0,\, Q_i^{\pi^i,\pi^{-i}}(s,a^i) - V_i^{\pi^i,\pi^{-i}}(s)\right)}{1 + \sum_{b^i \in \mathbb{A}^i} \max\left(0,\, Q_i^{\pi^i,\pi^{-i}}(s,b^i) - V_i^{\pi^i,\pi^{-i}}(s)\right)}.
\end{aligned}$$

We first provide the following claim given the condition that π is a fixed point.

Claim 5. *For any* $a^i \in \mathbb{A}^i$, $Q_i^{\pi^i,\pi^{-i}}(s,a^i) - V_i^{\pi^i,\pi^{-i}}(s) \leq 0$.

Proof of Claim 5. Suppose for contradiction that there exists $i \in [n]$ and $d^i \in \mathbb{A}^i$ such that $Q_i^{\pi^i,\pi^{-i}}(s,d^i) > V_i^{\pi^i,\pi^{-i}}(s)$. The above fixed point equation implies that $\pi^i(s,d^i) > 0$.

Let $\mathbb{A}_+^i = \{a^i \in \mathbb{A}^i : \pi^i(s,a^i) > 0\}$; then $d^i \in \mathbb{A}_+^i$. Note that, by the recursive definition of $V_i^{\pi^i,\pi^{-i}}(s)$, we have

$$\begin{aligned}
V_i^{\pi^i,\pi^{-i}}(s) &= \mathop{\mathbb{E}}_{a^i \sim \pi^i(s)} \Bigg[r^{i,\pi^{-i}}(s,a^i) \\
&\qquad + \gamma \sum_{s' \in \mathbb{S}} P^{\pi^{-i}}(s'|s,a^i) V_i^{\pi^i,\pi^{-i}}(s') \Bigg] \\
&= \sum_{a^i \in \mathbb{A}^i} \pi^i(s,a^i) Q_i^{\pi^i,\pi^{-i}}(s,a^i) \\
&= \sum_{a^i \in \mathbb{A}_+^i} \pi^i(s,a^i) Q_i^{\pi^i,\pi^{-i}}(s,a^i).
\end{aligned}$$

Since $\sum_{a^i \in \mathbb{A}_+^i} \pi^i(s,a^i) = 1$, there must exist some $c^i \in \mathbb{A}_+^i$ such that $Q_i^{\pi^i,\pi^{-i}}(s,c^i) < V_i^{\pi^i,\pi^{-i}}(s)$, because otherwise we have $Q_i^{\pi^i,\pi^{-i}}(s,a^i) \geq V_i^{\pi^i,\pi^{-i}}(s)$ for all $a^i \in \mathbb{A}_+^i$, which, combined with $Q_i^{\pi^i,\pi^{-i}}(s,d^i) > V_i^{\pi^i,\pi^{-i}}(s)$, implies that $\sum_{a^i \in \mathbb{A}_+^i} \pi^i(s,a^i) Q_i^{\pi^i,\pi^{-i}}(s,a^i) > V_i^{\pi^i,\pi^{-i}}(s)$, a contradiction to the above equation. With some further calculation, we can have the equation

$$\begin{aligned}
&(f(\pi))^i(s,c^i) \\
&= \frac{\pi^i(s,c^i) + \max\left(0,\, Q_i^{\pi^i,\pi^{-i}}(s,c^i) - V_i^{\pi^i,\pi^{-i}}(s)\right)}{1 + \sum_{b^i \in \mathbb{A}^i} \max\left(0,\, Q_i^{\pi^i,\pi^{-i}}(s,b^i) - V_i^{\pi^i,\pi^{-i}}(s)\right)} \\
&= \frac{\pi^i(s,c^i)}{1 + \sum_{b^i \in \mathbb{A}^i} \max\left(0,\, Q_i^{\pi^i,\pi^{-i}}(s,b^i) - V_i^{\pi^i,\pi^{-i}}(s)\right)} \\
&\leq \frac{\pi^i(s,c^i)}{1 + Q_i^{\pi^i,\pi^{-i}}(s,d^i) - V_i^{\pi^i,\pi^{-i}}(s)} \\
&< \pi^i(s,c^i).
\end{aligned}$$

The above strict inequality follows because $1 + Q_i^{\pi^i,\pi^{-i}}(s,d^i) - V_i^{\pi^i,\pi^{-i}}(s) > 1$ as well as $\pi^i(s,c^i) > 0$.

This contradicts with the assumption that π is a fixed point of f. Therefore, it holds for any $a^i \in \mathbb{A}^i$ that $Q_i^{\pi^i,\pi^{-i}}(s,a^i) \leq V_i^{\pi^i,\pi^{-i}}(s)$. □

Combining Claim 5 and Lemma 4 (to be proven next), we find that, for any $\tilde{\pi}^i \in \Delta_{A^i}^S$ and any $s \in \mathbb{S}$, $V_i^{\pi^i,\pi^{-i}}(s) \geq V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$. Thus, π is an MPE by definition. This completes the proof of Theorem 2. □

Lemma 4. *For any player* $i \in [n]$, *given* π^{-i}, *for any* $\pi^i \in \Delta_{A^i}^S$, *the following two statements are equivalent*:

1. for all $s \in \mathbb{S}$ and all $a^i \in \mathbb{A}^i$, $V_i^{\pi^i,\pi^{-i}}(s) \geq Q_i^{\pi^i,\pi^{-i}}(s,a^i)$;
2. for all $s \in \mathbb{S}$ and all $\tilde{\pi}^i \in \Delta_{A^i}^S$, $V_i^{\pi^i,\pi^{-i}}(s) \geq V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$.

Proof. Let $\mathbb{V}$ denote the space of value functions $\mathbb{S} \to \mathcal{R}$, and define the l_∞ norm for any $v \in \mathbb{V}$ as $\|v\|_\infty = \max_{s \in \mathbb{S}} |v(s)|$.

Pick any player $i \in [n]$ and keep π^{-i} fixed. Define the Bellman operator $\Phi^i : \mathbb{V} \to \mathbb{V}$ such that, for any $v \in \mathbb{V}$ and any $s \in \mathbb{S}$,

$$\begin{aligned}
\Phi^i(v)(s) := \max_{a^i \in \mathbb{A}^i} \Bigg[& r^{i,\pi^{-i}}(s,a^i) \\
& + \gamma \sum_{s' \in \mathbb{S}} P^{\pi^{-i}}(s'|s,a^i) v(s') \Bigg].
\end{aligned}$$

Note that, for all $\tilde{\pi}^i \in \Delta_{A^i}^S$, $\Phi^i(V_i^{\tilde{\pi}^i,\pi^{-i}})(s) = \max_{a^i \in \mathbb{A}^i} Q_i^{\tilde{\pi}^i,\pi^{-i}}(s,a^i)$, since $Q_i^{\tilde{\pi}^i,\pi^{-i}}(s,a^i) = r^{i,\pi^{-i}}(s,a^i) + \gamma \sum_{s' \in \mathbb{S}} P^{\pi^{-i}}(s'|s,a^i) V_i^{\tilde{\pi}^i,\pi^{-i}}(s')$.

We first prove the equivalence between statements 1 and 2, based on Claim 6 below, which will be proved next for completeness.

2$\Rightarrow$1: From statement 2, for all $s \in \mathbb{S}$, $V_i^{\pi^i,\pi^{-i}}(s) = \max_{\tilde{\pi}^i \in \Delta_{A^i}^S} V_i^{\tilde{\pi}^i,\pi^{-i}}(s) = v^{i*}(s)$, which is the fixed point of Φ^i by Claim 6 below. That is, for all $s \in \mathbb{S}$, $V_i^{\pi^i,\pi^{-i}}(s) = \Phi^i(V_i^{\pi^i,\pi^{-i}})(s) = \max_{a^i \in \mathbb{A}^i} Q_i^{\pi^i,\pi^{-i}}(s,a^i)$, by definition of the Bellman operator Φ^i. Statement 1 holds.

1⇒2: If statement 1 holds, we have, for all $s \in \mathbb{S}$, $V_i^{\pi^i,\pi^{-i}}(s) \geq \max_{a^i\in\mathbb{A}^i} Q_i^{\pi^i,\pi^{-i}}(s,a^i)$. Since $V_i^{\pi^i,\pi^{-i}}(s) \leq \max_{a^i\in\mathbb{A}^i} Q_i^{\pi^i,\pi^{-i}}(s,a^i)$ by Claim 6 below, we get $V_i^{\pi^i,\pi^{-i}}(s) = \max_{a^i\in\mathbb{A}^i} Q_i^{\pi^i,\pi^{-i}}(s,a^i)$. $\Phi^i(V_i^{\pi^i,\pi^{-i}})(s) = \max_{a^i\in\mathbb{A}^i} Q_i^{\pi^i,\pi^{-i}}(s,a^i) = V_i^{\pi^i,\pi^{-i}}(s)$, implying that $V_i^{\pi^i,\pi^{-i}}$ is a fixed point of Φ^i. By Claim 6, the unique fixed point of Φ^i is $v^{i*} = V_i^{\pi^i,\pi^{-i}}$. Therefore, for all $s \in \mathbb{S}$, $V_i^{\pi^i,\pi^{-i}}(s) = \max_{\tilde{\pi}^i\in\Delta_{A^i}^S} V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$: statement 2 holds.

Claim 6. *We have the following important properties.*

- *It holds that Φ^i is a γ-contraction mapping with respect to the l_∞ norm, and has a unique fixed point.*
- *For any $\tilde{\pi}^i \in \Delta_{A^i}^S$ and any $s \in \mathbb{S}$, $\Phi^i(V_i^{\tilde{\pi}^i,\pi^{-i}})(s) \geq V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$.*
- *Let $v^{i*} \in \mathbb{V}$ denote the fixed point of Φ^i; then v^{i*} is the optimal value function, i.e. for any $s \in \mathbb{S}$, $v^{i*}(s) = \max_{\tilde{\pi}^i\in\Delta_{A^i}^S} V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$.*

Proof of Claim 6. Define

$$Q_i^v(s,a^i) = r^{i,\pi^{-i}}(s,a^i) + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)v(s').$$

We have $\Phi^i(v)(s) = \max_{a^i\in\mathbb{A}^i} Q_i^v(s,a^i)$ for all $v \in \mathbb{V}$ and $s \in \mathbb{S}$.

We first prove that Φ^i is a γ-contraction mapping with respect to the l_∞ norm. For all $v_1, v_2 \in \mathbb{V}$, let $\delta = \|v_1 - v_2\|_\infty = \max_{s\in\mathbb{S}} |v_1(s) - v_2(s)|$. We show that $\|\Phi^i(v_1) - \Phi^i(v_2)\|_\infty \leq \gamma\delta$.

For all $s \in \mathbb{S}$ and all $a^i \in \mathbb{A}^i$, observe that

$$Q_i^{v_1}(s,a^i) - Q_i^{v_2}(s,a^i) = \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)(v_1(s') - v_2(s')),$$

so $|Q_i^{v_1}(s,a^i) - Q_i^{v_2}(s,a^i)| \leq \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)\delta = \gamma\delta$.

Without loss of generality, one can suppose that $\Phi^i(v_1)(s) \geq \Phi^i(v_2)(s)$. Taking arbitrary $a_1^i \in \arg\max_{a^i\in\mathbb{A}^i} Q_i^{v_1}(s,a^i)$, we have

$$\begin{aligned}\Phi^i(v_2)(s) &= \max_{a^i\in\mathbb{A}^i} Q_i^{v_2}(s,a^i)\\ &\geq Q_i^{v_2}(s,a_1^i)\\ &\geq Q_i^{v_1}(s,a_1^i) - \gamma\delta\\ &= \Phi^i(v_1)(s) - \gamma\delta;\end{aligned}$$

thus, $|\Phi^i(v_1)(s) - \Phi^i(v_2)(s)| \leq \gamma\delta$. By symmetry, the claim holds for the case in which $\Phi^i(v_1)(s) \leq \Phi^i(v_2)(s)$. Therefore, it holds that $\|\Phi^i(v_1) - \Phi^i(v_2)\|_\infty \leq \gamma\delta$. Thus, Φ^i is a γ-contraction mapping.

By the Banach fixed point theorem, we know that $\Phi^i : \mathbb{V} \to \mathbb{V}$ has a unique fixed point $v^{i*} \in \mathbb{V}$. Moreover, for any $v \in \mathbb{V}$, the point sequence $v, \Phi^i(v), \Phi^i(\Phi^i(v)), \ldots$ converges to v^{i*}, i.e. for all $s \in \mathbb{S}$, $\lim_{k\to\infty}(\Phi^i)^{(k)}(v)(s) = v^{i*}(s)$, where $(\Phi^i)^{(k)} = \Phi^i \circ (\Phi^i)^{(k-1)}$ is defined recursively with $(\Phi^i)^{(1)} = \Phi^i$.

Next, for all $\tilde{\pi}^i \in \Delta_{A^i}^S$ and all $s \in \mathbb{S}$, $\Phi^i(V_i^{\tilde{\pi}^i,\pi^{-i}})(s) \geq V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$, since

$$\begin{aligned}V_i^{\tilde{\pi}^i,\pi^{-i}}(s) &= \sum_{a^i\in\mathbb{A}^i} \tilde{\pi}^i(s,a^i) Q_i^{\tilde{\pi}^i,\pi^{-i}}(s,a^i)\\ &\leq \max_{a^i\in\mathbb{A}^i} Q_i^{\tilde{\pi}^i,\pi^{-i}}(s,a^i)\\ &= \Phi^i(V_i^{\tilde{\pi}^i,\pi^{-i}})(s)\end{aligned}$$

by definition.

Finally we prove that, for any $s \in \mathbb{S}$, $v^{i*}(s) = \max_{\tilde{\pi}^i\in\Delta_{A^i}^S} V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$. For any $\pi^i \in \Delta_{A^i}^S$, define the operator $\Psi_{\pi^i}^i : \mathbb{V} \to \mathbb{V}$, such that, for any $v \in \mathbb{V}$ and any $s \in \mathbb{S}$,

$$\Psi_{\pi^i}^i(v)(s) := \sum_{a^i\in\mathbb{A}^i} \pi^i(s,a^i) Q_i^v(s,a^i).$$

Note that, for any $\pi^i \in \Delta_{A^i}^S$, $\Psi_{\pi^i}^i$ is also a γ-contraction mapping. This is because, for any $v_1, v_2 \in \mathbb{V}$ such that $\|v_1 - v_2\|_\infty = \delta$, we have shown that, for any $s \in \mathbb{S}$ and any $a^i \in \mathbb{A}^i$, $|Q_i^{v_1}(s,a^i) - Q_i^{v_2}(s,a^i)| \leq \gamma\delta$, so

$$\begin{aligned}&\left|\Psi_{\pi^i}^i(v_1)(s) - \Psi_{\pi^i}^i(v_2)(s)\right|\\ &\leq \sum_{a^i\in\mathbb{A}^i} \pi^i(s,a^i)\left|Q_i^{v_1}(s,a^i) - Q_i^{v_2}(s,a^i)\right|\\ &\leq \gamma\delta,\end{aligned}$$

and then $\|\Psi_{\pi^i}^i(v_1) - \Psi_{\pi^i}^i(v_2)\|_\infty \leq \gamma\delta$.

For any $\pi^i \in \Delta_{A^i}^S$, we can observe that $V_i^{\pi^i,\pi^{-i}} = \Psi_{\pi^i}^i(V_i^{\pi^i,\pi^{-i}})$ by definition. By the Banach fixed point theorem, we know that $\Psi_{\pi^i}^i$ has a unique fixed point in $\mathbb{V}$, so $V_i^{\pi^i,\pi^{-i}}$ is exactly the unique fixed point of $\Psi_{\pi^i}^i$.

Now we arbitrarily take a policy $\pi_*^i \in \Delta_{A^i}^S$ such that, for all $s \in \mathbb{S}$, $\{a^i \in \mathbb{A}^i : \pi_*^i(s,a^i) > 0\} \subseteq \arg\max_{a^i\in\mathbb{A}^i} Q_i^{v^{i*}}(s,a^i)$. It can be seen that, for any $s \in \mathbb{S}$,

$$\begin{aligned}\Psi_{\pi_*^i}^i(v^{i*})(s) &= \max_{a^i\in\mathbb{A}^i} Q_i^{v^{i*}}(s,a^i)\\ &= \Phi^i(v^{i*})(s)\\ &= v^{i*}(s).\end{aligned}$$

It follows that $\Psi^i_{\pi^i_*}(v^{i*}) = v^{i*}$, so $v\mathrm{k}^{i*}$ is a fixed point of $\Psi^i_{\pi^i_*}$. Since the unique fixed point of $\Psi^i_{\pi^i_*}$ is $V_i^{\pi^i_*,\pi^{-i}}$, we have $V_i^{\pi^i_*,\pi^{-i}} = v^{i*}$. Thus, for any $s \in \mathbb{S}$, $\max_{\tilde{\pi}^i \in \Delta^S_{A^i}} V_i^{\tilde{\pi}^i,\pi^{-i}}(s) \geq V_i^{\pi^i_*,\pi^{-i}}(s) = v^{i*}(s)$.

To show that, for any $s \in \mathbb{S}$, $v^{i*}(s) \geq \max_{\tilde{\pi}^i \in \Delta^S_{A^i}} V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$, we observe that, given $v_1, v_2 \in \mathbb{V}$, if for all $s \in \mathbb{S}$, $v_1(s) \leq v_2(s)$, then, for any $s \in \mathbb{S}$ and any $a^i \in \mathbb{A}^i$, $Q_i^{v_1}(s, a^i) \leq Q_i^{v_2}(s, a^i)$. Therefore, $\Phi^i(v_1)(s) = \max_{a^i \in \mathbb{A}^i} Q_i^{v_1}(s, a^i) \leq \max_{a^i \in \mathbb{A}^i} Q_i^{v_2}(s, a^i) = \Phi^i(v_2)(s)$. As we have, for all $\tilde{\pi}^i \in \Delta^S_{A^i}$ and all $s \in \mathbb{S}$, $\Phi^i(V_i^{\tilde{\pi}^i,\pi^{-i}})(s) \geq V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$, we have, for any $k \in \mathcal{N}$ and any $s \in \mathbb{S}$, $(\Phi^i)^{(k+1)}(V_i^{\tilde{\pi}^i,\pi^{-i}})(s) \geq (\Phi^i)^{(k)}(V_i^{\tilde{\pi}^i,\pi^{-i}})(s)$ by induction. It follows that $V_i^{\tilde{\pi}^i,\pi^{-i}}(s) \leq (\Phi^i)^{(k+1)}(V_i^{\tilde{\pi}^i,\pi^{-i}})(s)$. Let $k \to \infty$; then we get $V_i^{\tilde{\pi}^i,\pi^{-i}}(s) \leq \lim_{k\to\infty}(\Phi^i)^{(k)}(V_i^{\tilde{\pi}^i,\pi^{-i}})(s) = v^{i*}(s)$. Thus, $v^{i*}(s) \geq \max_{\tilde{\pi}^i \in \Delta^S_{A^i}} V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$.

The claim that, for all $s \in \mathbb{S}$, $v^{i*}(s) = \max_{\tilde{\pi}^i \in \Delta^S_{A^i}} V_i^{\tilde{\pi}^i,\pi^{-i}}(s)$ follows.

THE APPROXIMATION GUARANTEE

Theorem 2 states that π is a fixed point of f if and only if π is an MPE for the stochastic game. Now we prove that f has some good approximation properties: if we find an ϵ-approximate fixed point π of f then it is also a $\mathrm{poly}(|\mathcal{SG}|)\sqrt{\epsilon}$-approximate MPE for the stochastic game (combining the following Lemma 5 and Lemma 6). This implies the **PPAD**-membership of Approximate MPE.

Lemma 5. *Let $\epsilon > 0$ and π be a strategy profile. If $\|f(\pi) - \pi\|_\infty \leq \epsilon$ then, for each player $i \in [n]$ and each state $s \in \mathbb{S}$, we have*

$$\max_{a^i \in \mathbb{A}^i} \left(Q_i^{\pi^i,\pi^{-i}}(s, a^i) - V_i^{\pi^i,\pi^{-i}}(s) \right) \leq A_{\max} \left(r_{\max}\sqrt{\epsilon'} + \frac{\sqrt{\epsilon'}}{1-\gamma} + \epsilon' \right),$$

where

$$\epsilon' = \epsilon\left(1 + \frac{A_{\max} r_{\max}}{1-\gamma}\right).$$

Proof. Pick any player $i \in [n]$ and any state $s \in \mathbb{S}$. For simplicity, for any $a^i \in \mathbb{A}^i$, define $Q(a^i) = Q_i^{\pi^i,\pi^{-i}}(s, a^i)$ and $M(a^i) = \max(0, Q(a^i) - V_i^{\pi^i,\pi^{-i}}(s))$.

First we give an upper bound on $M(a^i)$. For any $a^i \in \mathbb{A}^i$, it can be easily seen that

$$M(a^i) \leq Q(a^i) = Q_i^{\pi^i,\pi^{-i}}(s, a^i) \leq \frac{r_{\max}}{1-\gamma}.$$

By the condition $\|f(\pi) - \pi\|_\infty \leq \epsilon$, for any $a^i \in \mathbb{A}^i$, we have

$$\begin{aligned} & \pi^i(s, a^i) - \frac{\pi^i(s, a^i) + M(a^i)}{1 + \sum_{b^i \in \mathbb{A}^i} M(b^i)} \leq \epsilon \\ \Rightarrow \quad & \pi^i(s, a^i) \sum_{b^i \in \mathbb{A}^i} M(b^i) - M(a^i) \\ & \leq \left(1 + \sum_{b^i \in \mathbb{A}^i} M(b^i)\right)\epsilon \\ \Rightarrow \quad & \pi^i(s, a^i) \sum_{b^i \in \mathbb{A}^i} M(b^i) \\ & \leq M(a^i) + \left(1 + \frac{A_{\max} r_{\max}}{1-\gamma}\right)\epsilon. \end{aligned}$$

Set $\epsilon' = (1 + A_{\max} r_{\max}/(1-\gamma))\epsilon$; then we have the crucial inequality

$$\pi^i(s, a^i) \sum_{b^i \in \mathbb{A}^i} M(b^i) \leq M(a^i) + \epsilon'. \tag{1}$$

Let $\mathbb{A}^i_-$ denote $\{a^i \in \mathbb{A}^i : M(a^i) = 0\}$ or, equivalently, $\{a^i \in \mathbb{A}^i : Q(a^i) - V_i^{\pi^i,\pi^{-i}}(a^i) \leq 0\}$. Let $t = \sum_{a^i \in \mathbb{A}^i_-} \pi^i(s, a^i)$.

Case 1: $t \geq \sqrt{\epsilon'}/r_{\max}$. By inequality (1) we have

$$\begin{aligned} & \sum_{a^i \in \mathbb{A}^i_-} \pi^i(s, a^i) \sum_{b^i \in \mathbb{A}^i} M(b^i) \\ & \leq \sum_{a^i \in \mathbb{A}^i_-} (M(a^i) + \epsilon') \\ \Rightarrow \quad & t \sum_{b^i \in \mathbb{A}^i} M(b^i) \leq \sum_{a^i \in \mathbb{A}^i_-} \epsilon' \\ \Rightarrow \quad & \sum_{b^i \in \mathbb{A}^i} M(b^i) \\ & \leq A_{\max}\epsilon'/t = A_{\max} r_{\max}\sqrt{\epsilon'}. \end{aligned}$$

Case 2: $t < \sqrt{\epsilon'}/r_{\max}$. By inequality (1) we have

$$\begin{aligned} & \pi^i(s, a^i) \sum_{b^i \in \mathbb{A}^i} M(b^i) \leq M(a^i) + \epsilon' \\ & \text{for all } a^i \in \mathbb{A}^i \\ \Rightarrow \quad & \sum_{a^i \in \mathbb{A}^i} (\pi^i(s, a^i))^2 \sum_{b^i \in \mathbb{A}^i} M(b^i) \\ & \leq \sum_{a^i \in \mathbb{A}^i} \pi^i(s, a^i) M(a^i) + \epsilon' \end{aligned}$$

$$\Rightarrow \sum_{a^i\in\mathbb{A}^i}(\pi^i(s,a^i))^2\sum_{b^i\in\mathbb{A}^i}M(b^i) \le \sum_{a^i\in\mathbb{A}^i\setminus\mathbb{A}^i_-}\pi^i(s,a^i)M(a^i)+\epsilon'. \quad (2)$$

As $V_i^{\pi^i,\pi^{-i}}(s)=\sum_{a^i\in\mathbb{A}^i}\pi^i(s,a^i)Q(a^i)$ and, for all $a^i\in\mathbb{A}^i\setminus\mathbb{A}^i_-$, $M(a^i)=Q(a^i)-V_i^{\pi^i,\pi^{-i}}(s)$,

$$\begin{aligned}0&=\sum_{a^i\in\mathbb{A}^i}\pi^i(s,a^i)(Q(a^i)-V_i^{\pi^i,\pi^{-i}}(s))\\&=\sum_{a^i\in\mathbb{A}^i\setminus\mathbb{A}^i_-}\pi^i(s,a^i)M(a^i)\\&\quad+\sum_{a^i\in\mathbb{A}^i_-}\pi^i(s,a^i)(Q(a^i)-V_i^{\pi^i,\pi^{-i}}(s)).\end{aligned}$$

Therefore,

$$\begin{aligned}&\sum_{a^i\in\mathbb{A}^i\setminus\mathbb{A}^i_-}\pi^i(s,a^i)M(a^i)\\&=\sum_{a^i\in\mathbb{A}^i_-}\pi^i(s,a^i)(V_i^{\pi^i,\pi^{-i}}(s)-Q(a^i))\\&\le\frac{r_{\max}}{1-\gamma}t\\&<\frac{\sqrt{\epsilon'}}{1-\gamma}.\end{aligned}$$

Moreover, observe that $\sum_{a^i\in\mathbb{A}^i}(\pi^i(s,a^i))^2\ge 1/A_{\max}$. Substituting these into inequality (2), we get

$$\frac{1}{A_{\max}}\sum_{b^i\in\mathbb{A}^i}M(b^i)\le\frac{\sqrt{\epsilon'}}{1-\gamma}+\epsilon'.$$

It follows that $\sum_{b^i\in\mathbb{A}^i}M(b^i)\le A_{\max}(\sqrt{\epsilon'}/(1-\gamma)+\epsilon')$.

In conclusion, combining the two cases, we get

$$\sum_{b^i\in\mathbb{A}^i}M(b^i)\le A_{\max}\left(r_{\max}\sqrt{\epsilon'}+\frac{\sqrt{\epsilon'}}{1-\gamma}+\epsilon'\right).$$

Thus, for each $a^i\in\mathbb{A}^i$, we have

$$\begin{aligned}Q_i^{\pi^i,\pi^{-i}}(s,a^i)-V_i^{\pi^i,\pi^{-i}}(s)&\le M(a^i)\\&\le\sum_{b^i\in\mathbb{A}^i}M(b^i)\\&\le A_{\max}\left(r_{\max}\sqrt{\epsilon'}+\frac{\sqrt{\epsilon'}}{1-\gamma}+\epsilon'\right),\end{aligned}$$

which completes the proof. □

Lemma 6. *Let $\epsilon>0$ and π be a strategy profile. If, for each player $i\in[n]$ and each state $s\in\mathbb{S}$, $\max_{a^i\in\mathbb{A}^i}Q_i^{\pi^i,\pi^{-i}}(s,a^i)-V_i^{\pi^i,\pi^{-i}}(s)\le\epsilon$ then π is an $\epsilon/(1-\gamma)$-approximate MPE.*

Proof. Recall the mapping $\Phi^i:\mathbb{V}\to\mathbb{V}$, defined as the Bellman operator, from the proof of Lemma 4. Let $v^{i*}\in\mathbb{V}$ be the unique fixed point of Φ^i and recall that, for all $s\in\mathbb{S}$, $v^{i*}(s)=\max_{\tilde\pi^i\in\Delta^S_{A^i}}V_i^{\tilde\pi^i,\pi^{-i}}(s)$.

Pick any player $i\in[n]$; by assumption, for each state $s\in\mathbb{S}$, we have $\max_{a^i\in\mathbb{A}^i}Q_i^{\pi^i,\pi^{-i}}(s,a^i)-V_i^{\pi^i,\pi^{-i}}(s)\le\epsilon$. On the other hand, $V_i^{\pi^i,\pi^{-i}}(s)\le\max_{a^i\in\mathbb{A}^i}Q_i^{\pi^i,\pi^{-i}}(s,a^i)$, so we have $|V_i^{\pi^i,\pi^{-i}}(s)-\max_{a^i\in\mathbb{A}^i}Q_i^{\pi^i,\pi^{-i}}(s,a^i)|\le\epsilon$, i.e. $\|V_i^{\pi^i,\pi^{-i}}-\Phi^i(V_i^{\pi^i,\pi^{-i}})\|_\infty\le\epsilon$.

Since Φ^i is a γ-contraction mapping,

$$\begin{aligned}&\|v^{i*}-\Phi^i(V_i^{\pi^i,\pi^{-i}})\|_\infty\\&=\|\Phi^i(v^{i*})-\Phi^i(V_i^{\pi^i,\pi^{-i}})\|_\infty\\&\le\gamma\|v^{i*}-V_i^{\pi^i,\pi^{-i}}\|_\infty.\end{aligned}$$

In addition, by the triangle inequality we have

$$\begin{aligned}&\|v^{i*}-\Phi^i(V_i^{\pi^i,\pi^{-i}})\|_\infty\\&\quad+\|V_i^{\pi^i,\pi^{-i}}-\Phi^i(V_i^{\pi^i,\pi^{-i}})\|_\infty\\&\ge\|v^{i*}-V_i^{\pi^i,\pi^{-i}}\|_\infty,\end{aligned}$$

so it follows that

$$\begin{aligned}&\|V_i^{\pi^i,\pi^{-i}}-\Phi^i(V_i^{\pi^i,\pi^{-i}})\|_\infty\\&\ge(1-\gamma)\|v^{i*}-V_i^{\pi^i,\pi^{-i}}\|_\infty.\end{aligned}$$

Thus, we have

$$\begin{aligned}\|v^{i*}-V_i^{\pi^i,\pi^{-i}}\|_\infty&\le\frac{1}{1-\gamma}\|V_i^{\pi^i,\pi^{-i}}\\&\quad-\Phi^i(V_i^{\pi^i,\pi^{-i}})\|_\infty\\&\le\frac{\epsilon}{1-\gamma}.\end{aligned}$$

It follows that, for any $s\in\mathbb{S}$ and any $\tilde\pi^i\in\Delta^S_{A^i}$,

$$\begin{aligned}&V_i^{\tilde\pi^i,\pi^{-i}}(s)-V_i^{\pi^i,\pi^{-i}}(s)\\&\le v^{i*}(s)-V_i^{\pi^i,\pi^{-i}}(s)\\&\le\|v^{i*}-V_i^{\pi^i,\pi^{-i}}\|_\infty\\&\le\frac{\epsilon}{1-\gamma}.\end{aligned}$$

By definition, it follows that π is an $\epsilon/(1-\gamma)$-approximate MPE. □

To conclude, Lemma 2 is proven by combining Lemma 5 and Lemma 6, which completes the proof of Lemma 1.

CONCLUSION

Solving an MPE in general-sum SGs has long expected to be at least **PPAD**-hard. In this paper, we prove that computing an MPE in a finite-state infinite horizon discounted SG is **PPAD**-complete. Our completeness result also immediately implies the **PPAD**-completeness of computing an MPE in action-free SGs and single-controller SGs. We hope that our results can encourage MARL researchers to study solving an MPE in general-sum SGs, proposing a sample-efficient MARL solution, which leads to more prosperous algorithmic developments than those currently on zero-sum SGs.

APPENDIX

Proof of Claim 1

We have

$$\begin{aligned}
&|(x+y)(1+z')-(x'+y')(1+z)| \\
&\quad\le |(x+y)(1+z')-(x'+y')(1+z')| \\
&\qquad +|(x'+y')(1+z')-(x'+y')(1+z)| \\
&\quad= (1+z')|(x+y)-(x'+y')| \\
&\quad +(x'+y')|z'-z| \\
&\quad\le (1+z')(|x-x'+y-y'|+|z-z'|) \\
&\quad\le (1+z')(|x-x'|+|y-y'|+|z-z'|) \\
&\le (1+z)(1+z')(|x-x'|+|y-y'|+|z-z'|).
\end{aligned}$$

The first and third inequalities follow by the triangle inequality, the second inequality holds because $x'+y'\le 1+z'$ and the last inequality follows because $1+z\ge 1$. It immediately follows that

$$\begin{aligned}
&\left|\frac{x+y}{1+z}-\frac{x'+y'}{1+z'}\right| \\
&\quad= \frac{|(x+y)(1+z')-(x'+y')(1+z)|}{(1+z)(1+z')} \\
&\quad\le |x-x'|+|y-y'|+|z-z'|.
\end{aligned}$$

Proof of Claim 2

We have

$$\begin{aligned}
&|r^{i,\pi_1^{-i}}(s,b^i)-r^{i,\pi_2^{-i}}(s,b^i)| \\
&= \Big|\sum_{b^{-i}\in\mathbb{A}^{-i}} r^i(s,b^i,b^{-i})\pi_1^{-i}(s,b^{-i}) \\
&\quad -\sum_{b^{-i}\in\mathbb{A}^{-i}} r^i(s,b^i,b^{-i})\pi_2^{-i}(s,b^{-i})\Big| \\
&= \Big|\sum_{b^{-i}\in\mathbb{A}^{-i}} r^i(s,b^i,b^{-i})(\pi_1^{-i}(s,b^{-i}) \\
&\qquad -\pi_2^{-i}(s,b^{-i}))\Big| \\
&= \Big|\sum_{b^{-i}\in\mathbb{A}^{-i}} r^i(s,b^i,b^{-i}) \\
&\qquad \times\Big(\prod_{j\neq i}\pi_1^j(s,b^j)-\prod_{j\neq i}\pi_2^{-i}(s,b^j)\Big)\Big| \\
&\le \sum_{b^{-i}\in\mathbb{A}^{-i}} r^i(s,b^i,b^{-i})\Big|\prod_{j\neq i}\pi_1^j(s,b^j) \\
&\qquad -\prod_{j\neq i}\pi_2^j(s,b^j)\Big| \\
&\le r_{\max}\sum_{b^{-i}\in\mathbb{A}^{-i}}\Big|\prod_{j\neq i}\pi_1^j(s,b^j)-\prod_{j\neq i}\pi_2^j(s,b^j)\Big| \\
&\le (n-1)A_{\max}r_{\max}\delta,
\end{aligned}$$

where the last inequality follows from the next claim.

Proof of Claim 3

We have

$$\begin{aligned}
&\sum_{b^{-1}\in\mathbb{A}^{-1}}\Big|\prod_{j=2}^n\pi_1^j(s,b^j)-\prod_{j=2}^n\pi_2^j(s,b^j)\Big| \\
&\quad= \sum_{b^{-1}\in\mathbb{A}^{-1}}\Big|\sum_{k=2}^n\Big(\prod_{l=2}^{k-1}\pi_1^l(s,b^l)\Big) \\
&\qquad\times(\pi_1^k(s,b^k)-\pi_2^k(s,b^k))\prod_{l=k+1}^n\pi_2^l(s,b^l)\Big| \\
&\quad\le \sum_{k=2}^n\sum_{b^{-1}\in\mathbb{A}^{-1}}\Big(\prod_{l=2}^{k-1}\pi_1^l(s,b^l)\Big) \\
&\qquad\times|\pi_1^k(s,b^k)-\pi_2^k(s,b^k)|\prod_{l=k+1}^n\pi_2^l(s,b^l) \\
&\quad= \sum_{k=2}^n\sum_{b^k\in\mathbb{A}^k}|\pi_1^k(s,b^k)-\pi_2^k(s,b^k)| \\
&\quad\le (n-1)A_{\max}\delta.
\end{aligned}$$

Proof of Claim 4

We have

$$\begin{aligned}
&\sum_{b^{-1}\in\mathbb{A}^{-1}} |P^{\pi_1^{-i}}(s'|s,b^i) - P^{\pi_2^{-i}}(s'|s,b^i)| \\
&\le (n-1)A_{\max}\delta |P^{\pi_1^{-i}}(s'|s,b^i) - P^{\pi_2^{-i}}(s'|s,b^i)| \\
&= \Big| \sum_{b^{-i}\in\mathbb{A}^{-i}} P(s'|s,b^i,b^{-i})\pi_1^{-i}(s,b^{-i}) \\
&\quad - \sum_{b^{-i}\in\mathbb{A}^{-i}} P(s'|s,b^i,b^{-i})\pi_2^{-i}(s,b^{-i}) \Big| \\
&= \Big| \sum_{b^{-i}\in\mathbb{A}^{-i}} P(s'|s,b^i,b^{-i})(\pi_1^{-i}(s,b^{-i}) \\
&\quad - \pi_2^{-i}(s,b^{-i})) \Big| \\
&\le \sum_{b^{-i}\in\mathbb{A}^{-i}} P(s'|s,b^i,b^{-i}) \Big| \prod_{j\neq i}\pi_1^j(s,b^j) \\
&\quad - \prod_{j\neq i}\pi_2^j(s,b^j) \Big| \\
&\le \sum_{b^{-i}\in\mathbb{A}^{-i}} \Big| \prod_{j\neq i}\pi_1^j(s,b^j) - \prod_{j\neq i}\pi_2^j(s,b^j) \Big| \\
&\le (n-1)A_{\max}\delta.
\end{aligned}$$

Lemma 7 and its proof

Lemma 7. *For every* $\pi_1,\pi_2 \in \prod_{i=1}^n \Delta_{A^i}^S$ *such that* $\|\pi_1-\pi_2\|_\infty \le \delta$, *we have*

$$\begin{aligned}
&|(I-\gamma P^{\pi_1})^{-1}(s'|s) - (I-\gamma P^{\pi_2})^{-1}(s'|s)| \\
&\le \frac{nSA_{\max}\delta}{(1-\gamma)^2}
\end{aligned}$$

for any $s, s' \in \mathbb{S}$.

Proof. We first give an upper bound on $|P^{\pi_1}(s'|s) - P^{\pi_2}(s'|s)|$ for any $s, s' \in \mathbb{S}$:

$$\begin{aligned}
&|P^{\pi_1}(s'|s) - P^{\pi_2}(s'|s)| \\
&= \Big| \sum_{a\in\mathbb{A}} P(s'|s,a) \prod_{i\in[n]} \pi_1^i(s,a^i) \\
&\quad - \sum_{a\in\mathbb{A}} P(s'|s,a) \prod_{i\in[n]} \pi_2^i(s,a^i) \Big| \\
&\le \sum_{a\in\mathbb{A}} P(s'|s,a) \Big| \prod_{i\in[n]} \pi_1^i(s,a^i) - \prod_{i\in[n]} \pi_2^i(s,a^i) \Big| \\
&\le nA_{\max}\delta.
\end{aligned}$$

Now we view P^π as an $S\times S$ matrix. For any two $S\times S$ matrices M^1, M^2, we use $\|M^1-M^2\|_{\max}$ to denote $\max_{i,j}|M^1(i,j)-M^2(i,j)|$, i.e. the max norm. Then we have $\|P^{\pi_1}-P^{\pi_2}\|_{\max} \le nA_{\max}\delta$.

Let $Q^1 = (I-\gamma P^{\pi_1})^{-1}$ and $Q^2 = (I-\gamma P^{\pi_2})^{-1}$. (Note that the inverse of $(I-\gamma P^\pi)$ must exist because $\gamma<1$.)

By definition, we have $Q^1 = I+\gamma P^{\pi_1}Q^1$ and $Q^2 = I+\gamma P^{\pi_2}Q^2$. Then

$$\begin{aligned}
&\|Q^1-Q^2\|_{\max} \\
&= \gamma\|P^{\pi_1}Q^1 - P^{\pi_2}Q^2\|_{\max} \\
&= \gamma \max_{i,j} \Big| \sum_k P^{\pi_1}(i,k)Q^1(k,j) \\
&\qquad - \sum_k P^{\pi_2}(i,k)Q^2(k,j) \Big| \\
&\le \gamma \max_{i,j} \sum_k \Big| P^{\pi_1}(i,k)Q^1(k,j) \\
&\qquad - P^{\pi_2}(i,k)Q^2(k,j) \Big| \\
&\le \gamma \max_{i,j} \Big(\sum_k P^{\pi_1}(i,k)|Q^1(k,j)-Q^2(k,j)| \\
&+ \sum_k |Q^2(k,j)||P^{\pi_1}(i,k)-P^{\pi_2}(i,k)| \Big) \\
&\le \gamma \max_{i,j} \Big(\max_k |Q^1(k,j)-Q^2(k,j)| \\
&\qquad + \sum_k \frac{nA_{\max}\delta}{1-\gamma} \Big) \\
&= \gamma\Big(\|Q^1-Q^2\|_{\max} + \frac{nSA_{\max}\delta}{1-\gamma} \Big),
\end{aligned}$$

where the sixth line follows the following facts:

1. $\sum_k P^{\pi_1}(i,k) = 1$;
2. $|Q^1(k,j)-Q^2(k,j)| \le \max_k |Q^1(k,j)-Q^2(k,j)|$;
3. $|P^{\pi_1}(i,k)-P^{\pi_2}(i,k)| \le nA_{\max}\delta$;
4. $|Q^2(k,j)| \le \|Q^2\|_1 \le 1/(1-\gamma\|P^{\pi_2}\|_1) \le 1/(1-\gamma)$.

Note that $Q^2 = I+\gamma P^{\pi_2}Q^2$. Since the 1-norm is submultiplicative, we have

$$\begin{aligned}
\|Q^2\|_1 &\le 1+\gamma\|P^{\pi_2}Q^2\|_1 \\
&\le 1+\gamma\|P^{\pi_2}\|_1\|Q^2\|_1 \\
&\le 1+\gamma\|Q^2\|_1,
\end{aligned}$$

which leads to the fourth fact. So we have

$$|Q^1 - Q^2|_{\max} \leq \frac{nSA_{\max}\delta}{(1-\gamma)^2}.$$

This completes the proof.

ACKNOWLEDGEMENT

We would like to thank Yuhao Li for his early work, when he was an undergraduate student at Peking University.

FUNDING

This work was partially supported by the Science and Technology Innovation 2030—'New Generation of Artificial Intelligence' Major Project (2018AAA0100901).

AUTHOR CONTRIBUTIONS

X.D., D.M. and J.W. designed the research; X.D. and D.M. identified the research problem; X.D. and N.L. performed the research; D.M. and Y.Y. coordinated the team; X.D., N.L. and Y.Y. wrote the paper.

Conflict of interest statement. None declared.

REFERENCES

1. Shapley LS. Stochastic games. *Proc Natl Acad Sci USA* 1953; **39**: 1095–100.
2. Fink AM. Equilibrium in a stochastic *n*-person game. *J Sci Hiroshima Univ Ser A-I Math* 1964; **28**: 89–93.
3. Maskin E and Tirole J. Markov perfect equilibrium: I. observable actions. *J Econ Theory* 2001; **100**: 191–219.
4. Neyman A and Sorin S. *Stochastic Games and Applications*, vol. 570. New York: Springer Science and Business Media, 2003.
5. Albright SC and Winston W. A birth-death model of advertising and pricing. *Adv Appl Probab* 1979; **11**: 134–52.
6. Sobel MJ. Myopic solutions of Markov decision processes and stochastic games. *Oper Res* 1981; **29**: 995–1009.
7. Filar J. Player aggregation in the traveling inspector model. *IEEE Trans Automat Contr* 1985; **30**: 723–9.
8. Perez-Nieves N, Yang Y and Slumbers O *et al*. Modelling behavioural diversity for learning in open-ended games. In: Meila M and Zhang T (eds). *Proceedings of the 38th International Conference on Machine Learning*, vol. 139. Cambridge, MA: PMLR, 2021, 8514–24.
9. Filar J and Vrieze K.*Competitive Markov Decision Processes*. New York: Springer Science and Business Media, 2012.
10. Sutton RS and Barto AG. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2018.
11. Busoniu L, Babuska R and De Schutter B. A comprehensive survey of multiagent reinforcement learning. *IEEE Trans Syst Man Cybern C Appl Rev* 2008; **38**: 156–72.
12. Yang Y and Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective. arXiv: 2011.00583.
13. Littman ML. Markov games as a framework for multi-agent reinforcement learning. In: Cohen WW and Hirsh H (eds). *Proceedings of the 11th International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers, 1994, 157–63,
14. Hu J and Wellman MP. Multiagent reinforcement learning: theoretical framework and an algorithm. In: *Proceedings of the 15th International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers, 1998, 242–50,
15. Cesa-Bianchi N and Lugosi G. *Prediction, Learning, and Games*. Cambridge: Cambridge University Press, 2006.
16. Bubeck S and Cesa-Bianchi N. *Regret Analysis Of Stochastic and Nonstochastic Multi-Armed Bandit Problems*. Delft: Now Publishers Inc, 2012.
17. Takahashi M. Stochastic games with infinitely many strategies. *J Sci Hiroshima Univ Ser A-I Math* 1962; **26**: 123–34.
18. Zinkevich M, Greenwald A and Littman M. Cyclic equilibria in Markov games. In: *Proceedings of the 18th International Conference on Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2005, 1641–8.
19. Yang Y, Luo R and Li M *et al*. Mean field multi-agent reinforcement learning. In: *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. Cambridge, MA: PMLR, 2018, 5571–80.
20. Guo X, Hu A and Xu R *et al*. Learning mean-field games. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, vol. 80. Red Hook, NY: Curran Associates, 2019, 4966–76.
21. Pérolat J, Strub F and Piot B *et al*. *Learning nash equilibrium for general-sum Markov games from batch data*. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54. Cambridge, MA: PMLR, 2017, 232–41.
22. Solan E and Vieille N. Stochastic games. *Proc Natl Acad Sci USA* 2015; **112**: 13743–6.
23. Selten R. Reexamination of the perfectness concept for equilibrium points in extensive games. *Internat J Game Theory* 1975; **4**: 25–55.
24. Daskalakis C, Goldberg PW and Papadimitriou CH. The complexity of computing a Nash equilibrium. *SIAM J Comput* 2009; **39**: 195–259.
25. Chen X, Deng X and Teng SH. Settling the complexity of computing two-player nash equilibria. *J ACM* 2009; **56**: 1–57.
26. Papadimitriou CH. On the complexity of the parity argument and other inefficient proofs of existence. *J Comput Syst Sci* 1994; **48**: 498–532.
27. Nash J. Non-cooperative games. *Ann Math* 1951; **54**: 286–95.
28. Bertsekas DP. Approximate dynamic programming. In: Sammut C and Webb GI (eds). *Encyclopedia of Machine Learning*. Boston, MA: Springer, 2010, 39.
29. Szepesvári C and Littman ML. Generalized Markov decision processes: dynamic-programming and reinforcement-learning algorithms. Technical Report. Brown University, 1997.
30. Lagoudakis MG and Parr R. Least-squares policy iteration. *J Mach Learn Res* 2003; **4**: 1107–49.
31. Munos R and Szepesvári C. Finite-time bounds for fitted value iteration. *J Mach Learn Res* 2008; **9**: 815–57.

32. Riedmiller M. Neural fitted Q iteration - first experiences with a data efficient neural reinforcement learning method. In: Gama J, Camacho R and Brazdil PB *et al* (eds). *Machine Learning: ECML 2005*, vol. 3720. Berlin: Springer, 2005, 317–28.
33. Pérolat J, Piot B and Geist M *et al. Softened approximate policy iteration for Markov games*. In: *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48. Cambridge, MA: PMLR, 2016, 1860–8.
34. Lagoudakis MG and Parr R. Value function approximation in zero-sum Markov games. In: *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann Publishers, 2002, 283–92.
35. Pérolat J, Scherrer B and Piot B *et al. Approximate dynamic programming for two-player zero-sum Markov games*. In: *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37. Cambridge, MA: PMLR, 2015, 1321–9.
36. Sidford A, Wang M and Yang L *et al*. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In: Chiappa S and Calandra R (eds). *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, vol. 108. Cambridge, MA: PMLR, 2020, 2992–3002.
37. Daskalakis C, Foster DJ and Golowich N. Independent policy gradient methods for competitive reinforcement learning. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates, 2020, 5527–40.
38. Hansen TD, Miltersen PB and Zwick U. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *J ACM* 2013; **60**: 1–16.
39. Hu J and Wellman MP. Nash Q-learning for general-sum stochastic games. *J Mach Learn Res* 2003; **4**: 1039–69.
40. Greenwald A and Hall K. Correlated-Q learning. In: *Proceedings of the 20th International Conference on International Conference on Machine Learning*. Washington, DC: AAAI Press, 2003, 242–49.
41. Littman ML. Friend-or-foe Q-learning in general-sum games. In: *Proceedings of the 18th International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers, 2001, 322–8.
42. Prasad H, LA P and Bhatnagar S. 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*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2015, 1371–9.
43. Brafman RI and Tennenholtz M. R-MAX – A general polynomial time algorithm for near-optimal reinforcement learning. *J Mach Learn Res* 2002; **3**: 213–31.
44. Wei CY, Hong YT and Lu CJ. Online reinforcement learning in stochastic games. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates, 2017, 4994–5004.
45. Littman ML and Szepesvári C. A generalized reinforcement-learning model: convergence and applications. In: *Proceedings of the 13th International Conference on International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers, 1996, 310–8.
46. Fan J, Wang Z and Xie Y *et al*. A theoretical analysis of deep Q-learning. In: *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, vol. 120. Cambridge, MA: PMLR, 2020, 486–9.
47. Bowling M and Veloso M. Rational and convergent learning in stochastic games. In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann Publishers, 2001, 1021–6.
48. Conitzer V and Sandholm T. Awesome: a general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Mach Learn* 2007; **67**: 23–43.
49. Jin C, Allen-Zhu Z and Bubeck S *et al*. Is Q-learning provably efficient? In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates, 2018, 4868–78.
50. Zhang Z, Zhou Y and Ji X. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates, 2020, 15198–207.
51. Li Y, Wang R and Yang LF. Settling the horizon-dependence of sample complexity in reinforcement learning. In: *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. Piscataway, NJ: IEEE Press, 2022, 965–76.
52. Chen X, Cheng Y and Tang B. Well-supported versus approximate Nash equilibria: query complexity of large games. In: *Proceedings of the 2017 ACM Conference on Innovations in Theoretical Computer Science*. New York, NY: Association for Computing Machinery, 2017, 57.
53. Song Z, Mei S and Bai Y. When can we learn general-sum Markov games with a large number of players sample-efficiently? In: *International Conference on Learning Representations*. La Jolla, CA: OpenReview, 2022.
54. Jin C, Liu Q and Wang Y *et al*. V-learning – a simple, efficient, decentralized algorithm for multiagent RL. arXiv: 2110.14555.