Electronic Colloquium on Computational Complexity, Report No. 128 (2021)

On the Complexity of Computing Markov Perfect Equilibrium in General-Sum Stochastic Games

Xiaotie Deng* Yuhao Li* David Henry Mguni† Jun Wang‡ Yaodong Yang§

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 (MARL) and sequential agent interactions. In this paper, we introduce the solution concept, approximate Markov Perfect Equilibrium (MPE), to finite-state Stochastic Games repeated in the infinite horizon, and prove its **PPAD**-completeness in computational complexity. Technically, we adopt a function with a polynomially bounded description in the strategy space to convert the MPE computation to a fixed-point problem, even though the stochastic game may demand an exponential number of pure strategies, in the number of states, for each agent. The completeness result follows the reduction of the fixed-point problem to End of the Line.

Past works on the stochastic games are mostly zero-sum MARL algorithms. A P^{PPAD} algorithm for the general sum stochastic games in the finite horizon can be derived to establish an approximation algorithm for the general-sum stochastic games. That implies an approximate NE solution to the infinite-horizon setting. Such a possible extension suffers from three weaknesses: 1. the solution is time-dependent and hence not a perfect equilibrium; 2. the time-dependent solution suffers a weakness of noncredible threats; 3. the time complexity is not tight (lower bound **PPAD** and upper bound P^{PPAD}). Our result beats such a solution in all those three properties.

*Corresponding authors. Center on Frontiers of Computing Studies, Peking University. {xiaotie,yuhaoli.cs}@pku.edu.cn

†Huawei R&D UK. davidmguni@hotmail.com

‡University College London. jun.wang@cs.ucl.ac.uk

§King's College London. yaodong.yang@outlook.com

ISSN 1433-8092

1 Introduction

The seminal work of Shapley [42] defines Stochastic Games (SGs) to study the dynamic non-cooperative multi-player game, where each player simultaneously and independently chooses an action at each round, and the next state is determined by a probability distribution depending on the current state and the chosen joint actions. In two-player zero-sum SGs, Shapley [42] proved the existence of a stationary strategy profile in which no agent has an incentive to deviate; similarly, the existence of equilibrium in stationary strategies also holds in multi-player nonzero-sum SGs [16]. Such a solution concept (now also known as *Markov perfect equilibrium* (MPE) [29]) models the dynamic nature of multi-player games. As a refinement of Nash equilibrium [31] on SGs, MPE prevents non-payoff-relevant variables from affecting strategic behaviors, which allows researchers to identify the impact of state variables on outcomes.

Recently, Solan and Vieille [45] reconfirm the importance of the existence of a stationary strategy profile as having several 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 noncredible threats, a property that is stronger than equilibrium property, and viewed as highly desirable [see Selten [4]]."

Due to its generality, the framework of SGs has enlightened a sequence of studies [32] on a wide range of real-world applications ranging from advertising and pricing [2], fisheries modelling [44], football player selection [50], travelling inspection [14], and designing modern gaming AIs [34]. As a result, developing algorithms to compute MPE in SGs has become one of the key subjects in an extremely rich research domain, including but not limited to applied mathematics, economics, operations research, computer science and artificial intelligence [15, 40].

SGs underpin many AI/machine learning studies. For example, it is the key framework for studying adversarial training [17, 19] and modelling robustness [38, 1] in zero-sum setting. In reinforcement learning (RL), SG extends the Markov decision process (MDP) formulation to incorporate strategic interactions. Similar to the role of MDP in RL [46], SGs build the foundation for multi-agent reinforcement learning (MARL) techniques to study optimal decision makings in multi-player games [26]. In last decades, a wide variety of MARL algorithms have been developed to solved SGs [51].

Computing a MPE in (general-sum) SGs requires a perfect knowledge of the transition dynamics and the payoffs of the game [15], which is often infeasible in practice. To overcome this difficulty, MARL methods are often applied to learn the MPE of a SG based on the interactions between agents and the environment. MARL algorithms are generally considered under two settings: *online* and *offline*. In the offline setting (also known as the batch setting [37]), the learning algorithm controls all players in a centralised way, with the hope that the learning dynamics can eventually lead to a MPE by using limited number of interaction samples. In the online setting, the learner controls only one of the players to play with arbitrary opponents in the game, assuming having unlimited access to the game environment, and the central focus is often about the *regret*: the difference between a benchmark measure (often in hindsight) and the learner's total reward during learning.

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 for the worst-case optimality for each player can be thought of as solving MDPs. As a result, (approximate) dynamic programming methods [3, 47] such as LSPI [25] and FQI [30] / NFQI [41] can be adopted to solve SGs [35, 24, 36, 43, 23]. Under this setting, policy-based methods [11, 21] can also be applied. However, directly applying exiting MDP solvers on general-sum SGs are challenging. Since solving two-player NE in general-sum normal-form games (i.e., one-shot SGs) is well-known to be PPAD-complete

[12, 8], the complexity of MPE in general-sum SGs are expected to be at least PPAD. Although early attempts such as Nash-Q learning [22], Correlated-Q learning [20], Friend-or-Foe Q-Learning [27] have been made to solve general-sum SGs under strong assumptions, Zinkevich et. al. [52] demonstrated that the entire class of value iteration methods cannot find stationary NE policies in general-sum SGs. The difficulties on both the complexity side and the algorithmic side lead to very few existing MARL solutions to general-sum SGs; successful approaches either assumes knowing the complete information of the SG and thus solving MPE can be turned into an optimisation problem [39], or, proves the convergence of batch RL methods to a weaker notion of NE [37].

In the online setting, one of the most well-known algorithm is R-MAX [6], which studied (average-reward) zero-sum SGs and provided a polynomial (in terms of game size and error parameter) regret bound when competing with an arbitrary opponent. Under the same regret definition, recently, UCSG [49] improved R-MAX and achieved a sublinear regret, but still in two-player zero-sum SGs. When it comes to MARL solutions, Littman [26] proposed a practical solution named Minimax-Q that replaces the max operator with the minimax value. Asymptotic convergence results of Minimax-Q in both tabular cases [28] and value function approximations [13] have been shown. Yet, playing the minimax value could be overly pessimistic. If the adversary plays sub-optimally, the learner could achieve a higher reward. To account for this, WoLF [5] was proposed; and unlike Minimax-Q, WoLF is *rationale* in the sense that it can exploit opponent's policy. AWESOME [9] further generalised WoLF and achieve NE convergence in multi-player general-sum repeated games. However, outside the scope of zero-sum SGs, the question [6] of whether a polynomial time no-regret (near-optimal) RL/MARL algorithm exists for general-sum SGs is still unanswered.

Although SG has been proposed for more than 60 years and despite its importance, surprisingly, the complexity of finding a MPE in SG has never been answered. In fact, unlike the fruitful results on zero-sum SGs, we still know very little about the complexity of solving general-sum SGs. Two relevant results we know are that determining whether a pure-strategy NE exist in a SG is **PSPACE**-hard [10], and it is **NP**-hard to determine if there exists a memoryless ϵ-NE in *reachability* SGs [7]. It is long projected solving MPE in (infinite-horizon) SGs is at least **PPAD**-hard, since solving a two-player NE in one-shot SGs is already **PPAD**-hard [12, 8]. This suggests that under computational hardness assumption, it is unlike to have polynomial-time algorithms in even two-player stochastic games. Yet, the unresolved question is that

The key question that we try to address in this paper:

Can solving MPE in general-sum SGs be anywhere harder in the complexity class?

In this paper, we answer to the above question negatively by proving that computing a MPE in a finite-state discounted SG is **PPAD**-complete. Based on our result, we given an affirmative answer that finding an MPE in SGs is highly unlikely to be **NP**-hard under the circumstance that **NP**$\neq$ **co-NP**. We hope this result could encourage MARL researchers to work more on general-sum SGs, leading to fruitful MARL solutions as those currently on zero-sum SGs.

1.1 Intuitions and a Sketch of Our Main Ideas

Like the classic complexity class **NP**, **PPAD** is a collection of computational problems. As the definition of **NP**-completeness, a problem is said to be **PPAD**-complete if it is in **PPAD**, and is at least as hard as every problem in **PPAD**. When one Stochastic Game has only one state and the discount factor $\gamma = 0$, then finding a Markov perfect equilibrium (MPE) is equivalent to finding a Nash equilibrium in the corresponding normal-form game, which is known to be **PPAD**-complete

[12, 8]. So the **PPAD**-hardness of finding MPE is relatively direct (Lemma 1).

To obtain the **PPAD**-complete result (Theorem 1), it is sufficient for us to prove the **PPAD** membership of MPE (Lemma 2).

i) The first key observation is that we can construct a function f of the strategy profile space, such that each strategy profile is a fixed point of f if and only of it is an MPE (Theorem 2). Further, we prove the function f is continuous (actually λ-Lipschitz by Lemma 3), so that fixed points are guaranteed to exist by the Brouwer fixed point theorem.

ii) We then prove the function f has some "good" approximation properties. Let $|\mathcal{SG}|$ be the input size of a stochastic game. If we can find a $\texttt{poly}(|\mathcal{SG}|)\epsilon^2$-approximate fixed point π of f, i.e., $\|f(\pi) - \pi\|_\infty \leq \texttt{poly}(|\mathcal{SG}|)\epsilon^2$, where π is a strategy profile, then π is an ϵ-approximate MPE for the Stochastic Game (combining Lemma 4 and Lemma 5). So our goal converts to finding an approximate fixed point.

iii) To prove the **PPAD** membership of finding an MPE, we will reduce it to the problem END OF THE LINE (whose formal definition is in Section 3), which is the first **PPAD**-complete problem introduced by Papadimitriou [33]. We will show, the reduction could be constructed in polynomial time, and every solution of the problem END OF THE LINE corresponds to a good approximate fixed point (Lemma 6), thus yields an ϵ-approximate MPE.

2 Stochastic Games

Definition 1 (Stochastic Game)**.** *A Stochastic Game is defined by a tuple of key elements* $\langle n, \mathbb{S}, \mathbb{A}, P, r, \gamma \rangle$*, where*

- n *is the number of agents.*
- $\mathbb{S}$ *is the set of finite environmental states. Suppose that* $|\mathbb{S}| = S$*.*
- $\mathbb{A} = \mathbb{A}^1 \times \cdots \times \mathbb{A}^n$ *is the set of agents' joint actions. Suppose that* $|\mathbb{A}^i| = A^i$ *and* $A_{\max} = \max_{i \in [n]} A^i$*.*
- $P : \mathbb{S} \times \mathbb{A} \to \Delta(\mathbb{S})$ *is the transition probability, that is, at each time step, given the agents' joint action* $a \in \mathbb{A}$*, then the transition probability from state* s *to state in the next time step* s' *is* $P(s'|s, a)$*.*
- $r = r^1 \times \cdots \times r^n : \mathbb{S} \times \mathbb{A} \to \mathcal{R}_+^n$ *is the reward function, that is, when an agents are at state* s *and play a joint action* a*, then the agent* i *will get reward* $r^i(s, a)$*. We assume that the rewards are uniformly bounded by* $R_{\max}$*.*
- $\gamma \in [0, 1)$ *is 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, meaning that each agent's strategy can be conditioned only on the current state of the game.

Note that behavioral strategy is different from mixed strategy. To be more clear, we give both definitions of mixed strategy and behavioral strategy.

The pure strategy space of an agent i is $\prod_{s \in \mathbb{S}} \mathbb{A}^i$, meaning that the agent i needs to select an action at each state. Note that the size of pure strategy space of each agent is $|\mathbb{A}^i|^S$, which is exponential in the number of states.

Definition 2 (Mixed Strategy). *The mixed strategy space is* $\Delta\left(\prod_{s\in\mathbb{S}}\mathbb{A}^i\right)$*, i.e., the probability distribution on pure strategy space* $\prod_{s\in\mathbb{S}}\mathbb{A}^i$*.*

Definition 3 (Behavioral Strategy). *A behavioral strategy of an agent* i *is* $\pi^i:\mathbb{S}\to\Delta(\mathbb{A}^i)$*, i.e.,* $\forall s\in\mathbb{S},\pi^i(s)$ *is a probability distribution on* $\mathbb{A}^i$*.*

In the rest of the paper, we will refer to a behavioral strategy simply as a strategy for convenience. A strategy profile π is the Cartesian product of all agents' strategy, i.e., $\pi=\pi^1\times\cdots\times\pi^n$.

We denote the probability of agents using the joint action a on state s by $\pi(s,a)$, the probability of agent i using the action a^i on state s by $\pi^i(s,a^i)$. The strategy profile other than agent i is denoted by π^{-i}. Given π, the transition probability and the reward function only depend on the current state $s\in\mathbb{S}$. So let $r^{i,\pi}(s)$ denote $\mathbb{E}_{a\sim\pi(s)}[r^i(s,a)]$ and let $P^\pi(s'|s)$ denote $\mathbb{E}_{a\sim\pi(s)}[P(s'|s,a)]$. Given π^{-i}, the transition probability and the reward function only depend on the current state $s\in\mathbb{S}$ and player i's action a^i. So 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 let $P^{\pi^{-i}}(s'|s,a)$ 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 $\Delta_{A^i}^k:=\times_{p=1}^k\Delta_{A^i}$. Then $\forall s\in\mathbb{S},\pi^i(s)\in\Delta_{A^i}$, $\pi^i\in\Delta_{A^i}^S$ and $\pi\in\prod_{i=1}^n\Delta_{A^i}^S$.

Definition 4 (Value Function). *A value function for a strategy profile* π *of an agent* i*, written* $V^{\pi^i,\pi^{-i}}:\mathbb{S}\to R$ *gives the expected sum of discounted rewards of the agent* i *when the starting state is* s*:*

$$V^{\pi^i,\pi^{-i}}(s)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r^i(s_t,a)\Big|s_0=s,a\sim\pi(s_t),s_{t+1}\sim P^\pi(s_t)\right].$$

Alternatively, the value function can also be defined recursively via the Bellman equation.

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

Definition 5 (Markov Perfect Equilibrium (MPE)). *A behavioral strategy profile* π *is called a Markov Perfect Equilibrium if*

$$\forall s\in\mathbb{S},i\in[n],\forall\tilde{\pi}^i\in\Delta_{A^i}^S,V^{\pi^i,\pi^{-i}}(s)\geq V^{\tilde{\pi}^i,\pi^{-i}}(s).$$

Definition 6 (ϵ-approximate MPE). *Given* $\epsilon>0$*, a behavioral strategy profile* π *is called an* ϵ*-approximate MPE if*

$$\forall s\in\mathbb{S},i\in[n],\forall\tilde{\pi}^i\in\Delta_{A^i}^S,V^{\pi^i,\pi^{-i}}(s)\geq V^{\tilde{\pi}^i,\pi^{-i}}(s)-\epsilon.$$

The Markov perfect equilibrium is a concept within SGs in which the players' strategies depend only on the current state and not the game history. So the state encodes all relevant information for the player's strategies.

3 The Class PPAD and Markov-Perfect Equilibrium Problem

The complexity class **PPAD** is introduced [33] to characterize the mathematical proof structure required in a class of mathematical problems based on a parity argument for a solution to exist as in the following problem of End of the Line. It has included Nash equilibrium computation [12, 8], as well as many other problems.

The problem is defined on a class of directed graphs consisting of an exponential number of vertices (numbered from 0^n to $2^n - 1$). Edges of this graph is defined by two polynomial-size circuits S and P, each with n input bits and n output bits. There is an edge from vertex u to vertex v if and only if $S(u) = v$ and $P(v) = u$. Note that each vertex has at most 1 indegree and at most 1 outdegree, which means that the graph only consists of paths, cycles, and isolated vertices.

Definition 7 ((S, P)-Graph [18])**.** *An (S, P)-graph with parameter n is a graph on $\{0,1\}^n$ specified by circuits S and P, as described above, subject to the constraint that vertex 0^n has no incoming edge but does have an outgoing edge.*

Based on (S, P)-graphs, the problem End of the Line is to find a vertex other that 0^n such that the sum of its indegree and outdegree is one but Other End of this Line is to find the end of the particular path that starts at 0^n [18]. It turns out that the two problems are dramatically different in terms of their computational complexity. The former is **PPAD**-complete [33] but the latter is PSPACE-complete [18].

Here we give the definition of computational problem of finding a Markov Perfect Equilibrium in Stochastic Games.

Definition 8 (Markov-Perfect Equilibrium)**.** *The input instance of problem* Markov-Perfect Equilibrium *is a pair $(\mathcal{SG}, L)$ where $\mathcal{SG}$ is a Stochastic Game and L is a binary integer. The output of problem* Markov-Perfect Equilibrium *is a strategy profile $\pi \in \prod_{i=1}^{n} \Delta_{A^i}^{S}$ such that π is a $1/L$-approximate MPE.*

Theorem 1 (Main Theorem)**.** Markov-Perfect Equilibrium *is **PPAD**-complete.*

We note that when $|S| = 1$ and $\gamma = 0$, a Stochastic Game degenerates to an n-player matrix game. At this time, any Markov Perfect Equilibrium of this Stochastic Game is a Nash Equilibrium for the corresponding matrix game. So we have the following hardness result immediately:

Lemma 1. Markov-Perfect Equilibrium *is **PPAD**-hard.*

In the rest of the paper, we will mainly focus on the proof of **PPAD** membership of MPE.

Lemma 2. Markov-Perfect Equilibrium *is in **PPAD**.*

4 On the Existence of MPE

The original proof of the existence of MPE is from [16], mainly based on the Kakutani fixed point theorem. Here we give an alternative proof based on the Brouwer fixed point theorem, which also leads to our proof of **PPAD** membership of Markov-Perfect Equilibrium.

Inspired by the continuous transformation defined by Nash to prove the existence of equilibrium point [31], we define a new function $f : \prod_{i=1}^{n} \Delta_{A^i}^{S} \to \prod_{i=1}^{n} \Delta_{A^i}^{S}$ for a Stochastic Game to establish the existence of MPE. Let $V_{\pi^i(s,a^i)=1}^{\pi^i,\pi^{-i}}(s)$ denote the value function of agent i if agent i uses pure action a^i at state s, uses mixed actions $\pi^i(s')$ at state $s' \neq s$, and for any other agent $j \neq i$, agent j uses the strategy π^j.

Let $\pi \in \prod_{i=1}^{n} \Delta_{A^i}^{S}$ be a strategy profile. Then for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$, the modification of $\pi^i(s, a^i)$ is defined as follows:

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

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

We first prove the function f satisfies a continuity property namely λ-*Lipschitz*, where λ is defined as $\frac{9nS^2A_{\max}^2R_{\max}}{(1-\gamma)^2}$. The proof of Lemma 3 is challenging, because the value function $V^{\pi^i,\pi^{-i}}$ is defined recursively via Bellman equation. It could be written informally like $V^{\pi^i,\pi^{-i}} = (I - \gamma P^\pi)^{-1} r^{i,\pi}$, which is not linear even for each fixed π^{-i}. We refer the interested reader to Appendix A for a complete proof, whose techniques might be of independent interest.

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*

$$\left\| f(\pi_1) - f(\pi_2) \right\|_\infty \le \frac{9nS^2A_{\max}^2R_{\max}}{(1-\gamma)^2}\delta.$$

Now we could 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 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 the function f has at least one fixed point. Brouwer fixed point theorem states that for any continuous function mapping a compact convex set to itself, there is a fixed point. Notice that f is a function mapping a compact convex set to itself. Also, f is continuous by Lemma 3. So the function f has at least one fixed point.

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

The proof of the necessity part is immediate by the definition of MPE (Definition 5). If π is a MPE, then we have for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$, $V^{\pi^i,\pi^{-i}}(s) \ge V^{\pi^i,\pi^{-i}}_{\pi^i(s,a^i)=1}(s)$, which means $\max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,a^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) = 0$. Then for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$, $(f(\pi))^i(s,a^i) = \pi^i(s,a^i)$, which means π is a fixed point of f.

For the proof of the sufficiency part, suppose that π is a fixed point of f. Then we have for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$

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

Pick arbitrarily

$$a^{i,*} \in \operatorname*{arg\,min}_{b^i\in\mathbb{A}^i,\pi^i(s,b^i)>0} V^{\pi^i,\pi^{-i}}_{\pi^i(s,b^i)=1}(s).$$

It is not hard to prove $\max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,a^{i,*})=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) = 0$, which means

$$\begin{aligned}
&\pi^i(s,a^{i,*})\sum_{b^i\in\mathbb{A}^i\setminus\{a^{i,*}\}}\max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,b^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) = 0 \\
\implies &\sum_{b^i\in\mathbb{A}^i\setminus\{a^{i,*}\}}\max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,b^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) = 0 \\
\implies &\forall b^i \in \mathbb{A}^i\setminus\{a^{i,*}\}, \max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,b^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) = 0.
\end{aligned}$$

So we have $\forall b^i \in \mathbb{A}^i, \max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,b^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) = 0$, i.e., for any state $s \in \mathbb{S}$,

$$\pi^i \in \underset{\substack{\pi^{i,*}\in\Delta_A^S \\ \forall s'\neq s, \pi^{i,*}(s')=\pi^i(s)}}{\arg\max} V^{\pi^{i,*},\pi^{-i}}(s). \tag{1}$$

Note that if we fix the strategy profile other agent i, then for agent i, it is essentially a Markov decision process. By Equation (1), we know that π^i is an optimal policy of agent i, which means

$$\forall s \in \mathbb{S}, i \in [n], \forall \tilde{\pi}^i \in \Delta_A^S, V^{\pi^i,\pi^{-i}}(s) \geq V^{\tilde{\pi}^i,\pi^{-i}}(s),$$

i.e., π is a MPE of the Stochastic Game. □

5 PPAD Membership of Markov-Perfect Equilibrium

In this section, we will prove the **PPAD** membership of Markov-Perfect Equilibrium, by reducing it to End of the Line. We highlight our approximation guarantee proof (Section 5.1), which includes several innovative understanding of Markov Decision Processes and Stochastic Games. The construction of the graph of End of the Line is relatively standard and is from the simplicial approximation algorithm of Laan and Talman [48], which will be provided into Section 5.2.

5.1 The Approximation Guarantee

In Section 4, Theorem 2 states that f has a fixed point π if and only if π is an MPE for the Stochastic Game. Now we will prove f has some good approximation properties beyond that: if we find an ϵ-approximate fixed point π of f, then it is also a $\texttt{poly}(|\mathcal{SG}|)\sqrt{\epsilon}$-approximate MPE for the Stochastic Game (combining Lemma 4 and Lemma 5).

Moreover, we also get Corollary 1, which leads to better understanding for Markov Decision Process and might be of independent interest. The statement of Corollary 1 is as follows. Let $\epsilon > 0$ and π be a (not necessarily deterministic) policy. If for every starting state $s_0 \in \mathbb{S}$, the agent only changes the action of s_0 could gain at most ϵ more value, then the agent could gain at most $\epsilon/(1-\gamma)$ more value even if the agent changes its policy to the optimal policy, i.e., π is a good approximation of MDP.

The formal statements of lemmas and proofs are as follows. Proof of Lemma 4 is in Appendix B.1.

Lemma 4. *Let $\epsilon > 0$ and π be a strategy profile. If $\|f(\pi) - \pi\|_\infty \leq \epsilon$, then for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$, we have*

$$\max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,a^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) \leq A_{\max}\left(\frac{\sqrt{\epsilon'}}{1-\gamma} + R_{\max}\sqrt{\epsilon'} + \epsilon'\right),$$

where $\epsilon' = \epsilon\left(1 + \dfrac{A_{\max}R_{\max}}{1-\gamma}\right)$.

Lemma 5. *Let $\epsilon > 0$ and π be a strategy profile. If for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$, $\max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,a^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) \leq \epsilon$, then π is an $\epsilon/(1-\gamma)$-approximate MPE.*

Proof. Pick any player $i \in [n]$, it is sufficient for us to prove $\forall s \in \mathbb{S}, \forall \tilde{\pi}^i \in \Delta_A^S, V^{\pi^i,\pi^{-i}}(s) \geq V^{\tilde{\pi}^i,\pi^{-i}}(s) - \epsilon$. Suppose that $\max_{a^i \in \mathbb{A}^i} \left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,a^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) = \epsilon(s)$. Consider the following linear program:

$$\begin{aligned} \min \quad & \sum_{s\in\mathbb{S}} V(s) \\ \text{s.t.,} \quad & V(s) \geq r^{i,\pi^{-i}}(s,a^i) + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)V(s') \quad \forall s \in \mathbb{S}, a^i \in \mathbb{A}^i. \end{aligned} \tag{2}$$

Let V^* be the solution of the linear program (2). It satisfies

$$V^*(s) = \max_{a^i\in\mathbb{A}^i} \left(r^{i,\pi^{-i}}(s,a^i) + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)V^*(s') \right),$$

which is also the value function of player i when she uses the optimal policy given others' strategy profile π^{-i}. (Note that when we are given π^{-i}, it is essentially a Markov Decision Process for player i. So we are using linear programming to solve this MDP.)

Now look at the other linear program:

$$\begin{aligned} \min \quad & \sum_{s\in\mathbb{S}} V(s) \\ \text{s.t.,} \quad & V(s) \geq r^{i,\pi^{-i}}(s,a^i) - \epsilon(s) + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)V(s') \quad \forall s \in \mathbb{S}, a^i \in \mathbb{A}^i. \end{aligned} \tag{3}$$

Let V' be the solution of the linear program (3). It satisfies

$$V'(s) = \max_{a^i\in\mathbb{A}^i} \left(r^{i,\pi^{-i}}(s,a^i) + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)V'(s') \right) - \epsilon(s),$$

which is also the value function for the strategy profile π for the player i.

Now it is sufficient for us to bound $V^*(s) - V'(s), \forall s \in \mathbb{S}$. Let $\epsilon_{\max} = \max_{s\in\mathbb{S}} \epsilon(s)$. Construct a new value vector for the player i: $\tilde{V}(s) = V'(s) + \epsilon_{\max}/(1-\gamma)$. Then we have

$$\begin{aligned} & V(s) \geq r^{i,\pi^{-i}}(s,a^i) - \epsilon(s) + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)V(s') \\ \iff \quad & V'(s) + \frac{\epsilon_{\max}}{1-\gamma} \geq r^{i,\pi^{-i}}(s,a^i) - \epsilon(s) + \frac{\epsilon_{\max}}{1-\gamma} + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)V(s') \\ \iff \quad & V'(s) + \frac{\epsilon_{\max}}{1-\gamma} \geq r^{i,\pi^{-i}}(s,a^i) - \epsilon(s) + \epsilon_{\max} + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i) \left(V(s') + \frac{\epsilon_{\max}}{1-\gamma} \right) \\ \iff \quad & \tilde{V}(s) \geq r^{i,\pi^{-i}}(s,a^i) - \epsilon(s) + \epsilon_{\max} + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)\tilde{V}(s) \\ \implies \quad & \tilde{V}(s) \geq r^{i,\pi^{-i}}(s,a^i) + \gamma \sum_{s'\in\mathbb{S}} P^{\pi^{-i}}(s'|s,a^i)\tilde{V}(s). \end{aligned}$$

So $\tilde{V}$ is a feasible solution of linear program (2), which means $V^*(s) \leq \tilde{V}(s)$ for any $s \in \mathbb{S}$. Then we have

$$V^*(s) - V'(s) \leq \tilde{V}(s) - V'(s) = \epsilon_{\max}/(1-\gamma),$$

i.e., the difference between the optimal value $V^*(s)$ and $V^{\pi^i,\pi^{-i}}$ is upper bounded by $\epsilon/(1-\gamma)$. The argument above applies to any player. So by the definition of ϵ-approximate MPE, we know that π is an $\epsilon/(1-\gamma)$-approximate MPE. □

Corollary 1. *Let $\epsilon > 0$ and π be a (not necessarily deterministic) policy of the agent. If for each state $s \in \mathbb{S}$ and each action $a \in \mathbb{A}$ (where $\mathbb{A}$ is the action space of the agent), $\max\left(0, V^{\pi}_{\pi(s,a)=1}(s) - V^{\pi}(s)\right) \leq \epsilon$, then π is an $\epsilon/(1-\gamma)$ approximation of MDP.*

5.2 Constructing the End of the Line Graph

In this section, we give an outline of our reduction from Markov-Perfect Equilibrium to End of the Line, with the help of the simplicial approximation algorithm of Laan and Talman [48]. We will focus on the correctness of reduction, leaving details about how to construct the vertices to the appendix.

Recall that the input instance of Markov-Perfect Equilibrium is a pair $(\mathcal{SG}, L)$. Let d be an integer, which will be defined later to make sure we can find an $1/L$-approximate MPE.

For each $i \in [n]$, define $\Delta_{A^i}(d)$ is the set of points of Δ_{A^i} induced by the regular grid of size d, i.e.,

$$\Delta_{A^i}(d) = \left\{ x \in \Delta_{A^i} | x_j = y_j/d, y_j \in \mathbb{Z}^+, \sum_{j=1}^{A^i} y_j = d \right\}.$$

Similarly, define $\Delta^k_{A^i}(d) := \times_{p=1}^{k} \Delta_{A^i}(d)$.

The Vertices of End of the Line Graph. The set of vertices Σ is a set of simplices defined on $\prod_{i=1}^{n} \Delta^S_{A^i}(d)$, which could be encoded with string $\{0,1\}^N$, where N is polynomial in $|\mathcal{SG}|$ and $\log d$. The formal definition of Σ is in Appendix B.2.

Labelling the Grid Points. We will give each point in $\prod_{i=1}^{n} \Delta^S_{A^i}(d)$ a label, which will be an element of the set $\mathcal{L} := \bigcup_{i\in[n], s\in\mathbb{S}, a^i\in\mathbb{A}^i}(i, s, a^i)$.

Without loss of generality, we assign a number to the state set $\mathbb{S}$ and action set $\mathbb{A}^i$ for each $i \in [n]$ arbitrarily for the purpose of labelling. Suppose that $\mathbb{S} = \{s_1, \cdots, s_S\}$ and $\mathbb{A}^i = \{a^i_1, \cdots, a^i_{A^i}\}$.

For each strategy profile $\pi \in \prod_{i=1}^{n} \Delta^S_{A^i}(d)$, π receives the label (i, s_j, a^i_k) if and only if (i, s_j, a^i_k) is the lexicographically least index such that $\pi^i(s_j, a^i_k) > 0$ and

$$(f(\pi))^i(s_j, a^i_k) - \pi^i(s_j, a^i_k) \leq (f(\pi))^{i'}(s_{j'}, a^{i'}_{k'}) - \pi^{i'}(s_{j'}, a^{i'}_{k'})$$

for all $i' \in [n], s_{j'} \in \mathbb{S}$ and $a^{i'}_{k'} \in \mathbb{A}^{i'}$.

Note that each strategy profile $\pi \in \prod_{i=1}^{n} \Delta^S_{A^i}(d)$ has exactly one label, which could be denoted by $l(\pi)$. Since the function f could be computed in time polynomial in N and $|\mathcal{SG}|$, the label could also computed in time polynomial in $|\mathcal{SG}|$ and $\log d$. Also the labelling rule is proper in the sense that $l(\pi) \neq (i, s_j, a^i_k)$ if $\pi^i(s_j, a^i_k) = 0$.

A simplex $\sigma \in \Sigma$ will be called complete labelled if all its vertices[1] have a different label. A completely labelled simplex σ is called (i, s_j)-stopping if for each $a^i_k \in \mathbb{A}^i$, there exists $\pi \in \sigma$ such that $l(\pi) = (i, s_j, a^i_k)$. Further, a completely labelled simplex σ is called stopping if there exist $i \in [n]$ and $s_j \in \mathbb{S}$ such that σ is (i, s_j)-stopping.

The following lemma asserts that if we can find a stopping simplex, then we can find an $\texttt{poly}(|\mathcal{SG}|)/d$-approximate fixed point. The proof is in Appendix B.3.

[1] Please distinguish the vertices of a simplex and vertices of the End of the Line graph.

Lemma 6. *([48]) Suppose that a simplex $\sigma \in \Sigma$ is (i, s)-stopping for $i \in [n]$ and $s \in \mathbb{S}$. Then for any strategy profile $\pi \in \sigma$, we have*

$$\|f(\pi) - \pi\|_\infty \leq A_{\max}^2(\lambda + 1)\frac{1}{d}.$$

The Choice of d. Let

$$d = \frac{32A_{\max}^5 R_{\max}^3(\lambda + 1)}{(1-\gamma)^5}L^2.$$

It is easy to see d is $\texttt{poly}(|\mathcal{SG}|, L)$. The correctness of our choice is in Appendix B.4.

The Edges of End of the Line Graph. In the algorithm of Laan and Talman [48], they develop a partial one-to-one function $g : \Sigma' \to \Sigma'$ for $\Sigma' \subseteq \Sigma$ as well as a starting simplex $\sigma_0 \in \Sigma$, which have the following properties:

- $\sigma_0 \in \Sigma'$ and there is no $\sigma' \in \Sigma'$ such that $g(\sigma') = \sigma_0$;
- For any $\sigma \in \Sigma'$, if σ has no image, then σ is a stopping simplex. For any $\sigma \in \Sigma' \setminus \{\sigma_0\}$, if σ has no pre-image, then σ is a stopping simplex.
- the function g and g^{-1} could be computed in time polynomial in $|\mathcal{SG}|$ and $\log d$.

For the purpose of constructing the End of the Line graph, we complete the function g by letting $g(\sigma) = \sigma$ for any $\sigma \in \Sigma \setminus \Sigma'$. It is easy to verify our operation does not impact the properties of function g. So for any input instance $(|\mathcal{SG}|, L)$, we can reduce it to an instance of End of the Line, where the two circuits S and P correspond to g and g^{-1}. If we can find a solution of the End of the Line, by Lemma 6 we know that there is an $A_{\max}^2(\lambda + 1)\frac{1}{d}$-approximate fixed point in the solution simplex, thus an $1/L$-approximate MPE by Lemma 4, Lemma 5, and our choice of d.

6 Conclusion

Solving a Markov Perfect Equilibrium (MPE) in general-sum stochastic games (SG) 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 SGs is **PPAD**-complete. Our proof is novel in the sense that we adopt a function with a polynomial-bound description in the strategy space that effectively helps convert the MPE computation problem to a fixed-point problem, which, otherwise, would take a representation that requires an exponential number of pure strategies with respect to the number of states and the number of agents. Our completeness result indicates that computing MPE in SGs is highly unlikely to be **NP**-hard. We hope our results can encourage MARL researchers to study solving MPE in general-sum SGs, leading to more prosperous algorithmic developments as those currently on zero-sum SGs.

References

[1] Mohammed Amin Abdullah, Hang Ren, Haitham Bou Ammar, Vladimir Milenkovic, Rui Luo, Mingtian Zhang, and Jun Wang. Wasserstein robust reinforcement learning. *arXiv preprint arXiv:1907.13196*, 2019.

[2] S Christian Albright and Wayne Winston. A birth-death model of advertising and pricing. *Advances in Applied Probability*, pages 134–152, 1979.

[3] Dimitri P Bertsekas. Approximate dynamic programming. 2008.

[4] R Selten Bielefeld. Reexamination of the perfectness concept for equilibrium points in extensive games. In *Models of Strategic Rationality*, pages 1–31. Springer, 1988.

[5] Michael Bowling and Manuela Veloso. Rational and convergent learning in stochastic games. In *International joint conference on artificial intelligence*, volume 17, pages 1021–1026. Citeseer, 2001.

[6] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

[7] Krishnendu Chatterjee, Rupak Majumdar, and Marcin Jurdziński. On nash equilibria in stochastic games. In *International workshop on computer science logic*, pages 26–40. Springer, 2004.

[8] Xi Chen and Xiaotie Deng. Settling the complexity of two-player nash equilibrium. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 261–272. IEEE, 2006.

[9] Vincent Conitzer and Tuomas Sandholm. Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, 67(1-2):23–43, 2007.

[10] Vincent Conitzer and Tuomas Sandholm. New complexity results about nash equilibria. *Games and Economic Behavior*, 63(2):621–641, 2008.

[11] Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[12] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

[13] Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020.

[14] Jerzy Filar. Player aggregation in the traveling inspector model. *IEEE Transactions on Automatic Control*, 30(8):723–729, 1985.

[15] Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[16] Arlington M Fink. Equilibrium in a stochastic n-person game. *Journal of science of the hiroshima university, series ai (mathematics)*, 28(1):89–93, 1964.

[17] Barbara Franci and Sergio Grammatico. A game–theoretic approach for generative adversarial networks. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1646–1651. IEEE, 2020.

[18] Paul W Goldberg, Christos H Papadimitriou, and Rahul Savani. The complexity of the homotopy method, equilibrium selection, and lemke-howson solutions. *ACM Transactions on Economics and Computation (TEAC)*, 1(2):1–25, 2013.

[19] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[20] Amy Greenwald, Keith Hall, and Roberto Serrano. Correlated q-learning. In *ICML*, volume 3, pages 242–249, 2003.

[21] Thomas Dueholm Hansen, Peter Bro Miltersen, and Uri Zwick. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. *Journal of the ACM (JACM)*, 60(1):1–16, 2013.

[22] Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

[23] Zeyu Jia, Lin F Yang, and Mengdi Wang. Feature-based q-learning for two-player stochastic games. *arXiv preprint arXiv:1906.00423*, 2019.

[24] Michail G Lagoudakis and Ronald Parr. Value function approximation in zero-sum markov games. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 283–292, 2002.

[25] Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149, 2003.

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

[27] Michael L Littman. Friend-or-foe q-learning in general-sum games. In *ICML*, volume 1, pages 322–328, 2001.

[28] Michael L Littman and Csaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *ICML*, volume 96, pages 310–318. Citeseer, 1996.

[29] Eric Maskin and Jean Tirole. Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory*, 100(2):191–219, 2001.

[30] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

[31] John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

[32] Abraham Neyman and Sylvain Sorin. *Stochastic games and applications*, volume 570. ASIC, 2003.

[33] Christos H Papadimitriou. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and system Sciences*, 48(3):498–532, 1994.

[34] Peng Peng, Ying Wen, Yaodong Yang, Quan Yuan, Zhenkun Tang, Haitao Long, and Jun Wang. Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*, 2017.

[35] Julien Pérolat, Bilal Piot, Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. Softened approximate policy iteration for markov games. In *International Conference on Machine Learning*, pages 1860–1868. PMLR, 2016.

[36] Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning*, pages 1321–1329. PMLR, 2015.

[37] Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning nash equilibrium for general-sum markov games from batch data. In *Artificial Intelligence and Statistics*, pages 232–241. PMLR, 2017.

[38] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017.

[39] HL Prasad, Prashanth LA, and Shalabh Bhatnagar. Two-timescale algorithms for learning nash equilibria in general-sum stochastic games. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1371–1379, 2015.

[40] TES Raghavan and Jerzy A Filar. Algorithms for stochastic games—a survey. *Zeitschrift für Operations Research*, 35(6):437–472, 1991.

[41] Martin Riedmiller. Neural fitted q iteration–first experiences with a data efficient neural reinforcement learning method. In *European Conference on Machine Learning*, pages 317–328. Springer, 2005.

[42] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[43] Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 2992–3002. PMLR, 2020.

[44] M Sobel. Stochastic fishery games with myopic equilibria. *Essays in the Economics of Renewable Resources*, 259:268, 1982.

[45] Eilon Solan and Nicolas Vieille. Stochastic games. *Proceedings of the National Academy of Sciences*, 112(45):13743–13746, 2015.

[46] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[47] Csaba Szepesvári and Michael L Littman. Generalized markov decision processes: Dynamic-programming and reinforcement-learning algorithms. In *Proceedings of International Conference of Machine Learning*, volume 96, 1996.

[48] Gerard van der Laan and A. J. J. Talman. On the computation of fixed points in the product space of unit simplices and an application to noncooperative N person games. *Math. Oper. Res.*, 7(1):1–13, 1982.

[49] Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4994–5004, 2017.

[50] W Winston and AV Cabot. A stochastic game model of football play selection. *Indiana University mimeograph*, 1984.

[51] Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.

[52] Martin Zinkevich, Amy Greenwald, and Michael Littman. Cyclic equilibria in markov games. *Advances in Neural Information Processing Systems*, 18:1641, 2006.

A Detailed Proofs from Section 4

A.1 Proof of Lemma 3

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*

$$\left\| f(\pi_1) - f(\pi_2) \right\|_\infty \le \frac{9nS^2 A_{\max}^2 R_{\max}}{(1-\gamma)^2}\delta.$$

Proof of Lemma 3. We first give an upper bound of $\left|r^{i,\pi_1}(s) - r^{i,\pi_2}(s)\right|$ for any $s \in \mathbb{S}$ and $i \in [n]$.

$$\begin{aligned}
& \left|r^{i,\pi_1}(s) - r^{i,\pi_2}(s)\right| \\
= & \left|\sum_{a\in\mathbb{A}} r^i(s,a)\pi_1(s,a) - \sum_{a\in\mathbb{A}} r^i(s,a)\pi_2(s,a)\right| \\
= & \left|\sum_{a\in\mathbb{A}} r^i(s,a)\prod_{i\in[n]}\pi_1^i(s,a^i) - \sum_{a\in\mathbb{A}} r^i(s,a)\prod_{i\in[n]}\pi_2^i(s,a^i)\right| \\
\le & \sum_{a\in\mathbb{A}} \left|r^i(s,a)\right| \left|\prod_{i\in[n]}\pi_1^i(s,a^i) - \prod_{i\in[n]}\pi_2^i(s,a^i)\right| \\
\le & R_{\max}\sum_{a\in\mathbb{A}} \left|\prod_{i\in[n]}\pi_1^i(s,a^i) - \prod_{i\in[n]}\pi_2^i(s,a^i)\right| \\
\le & nA_{\max}R_{\max}\delta,
\end{aligned}$$

where the last inequality follows

$$\begin{aligned}
& \sum_{a\in\mathbb{A}} \left|\prod_{i\in[n]}\pi_1^i(s,a^i) - \prod_{i\in[n]}\pi_2^i(s,a^i)\right| \\
= & \sum_{a\in\mathbb{A}} \left|\sum_{k=1}^{n}\prod_{i=1}^{k-1}\pi_1^i(s,a^i)\left(\pi_1^k(s,a^k) - \pi_2^k(s,a^k)\right)\prod_{i=k+1}^{n}\pi_2^i(s,a^i)\right| \\
\le & \delta\sum_{a\in\mathbb{A}} \left|\sum_{k=1}^{n}\prod_{i=1}^{k-1}\pi_1^i(s,a^i)\prod_{i=k+1}^{n}\pi_2^i(s,a^i)\right| \\
\le & nA_{\max}\delta.
\end{aligned}$$

Let $V^{\pi^i,\pi^{-i}}$ denote the column vector $(V^{\pi^i,\pi^{-i}}(s))_{s\in\mathbb{S}}$, $r^{i,\pi}$ denote the column vector $(r^{i,\pi}(s))_{s\in\mathbb{S}}$, and P^π denote the matrix $P^\pi(s,s')_{s,s'\in\mathbb{S}}$ respectively. By the Bellman equation (Definition 4), we have

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

which means

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

We will prove that $\left|(I-\gamma P^{\pi_1})^{-1}(s'|s) - (I-\gamma P^{\pi_2})^{-1}(s'|s)\right| \le \frac{nSA_{\max}\delta}{(1-\gamma)^2}$ for any $s, s' \in \mathbb{S}$ in the following lemma (Lemma 7).

Now we could give an upper bound of $\left|V^{\pi_1^i,\pi_1^{-i}}(s) - V^{\pi_2^i,\pi_2^{-i}}(s)\right|$ for any $s \in \mathbb{S}$.

$$
\begin{aligned}
& \left|V^{\pi_1^i,\pi_1^{-i}}(s) - V^{\pi_2^i,\pi_2^{-i}}(s)\right| \\
= & \left|\sum_{s'\in\mathbb{S}} r^{i,\pi_1}(s')(I-\gamma P^{\pi_1})^{-1}(s'|s) - \sum_{s'\in\mathbb{S}} r^{i,\pi_2}(s')(I-\gamma P^{\pi_2})^{-1}(s'|s)\right| \\
= & \left|\sum_{s'\in\mathbb{S}} r^{i,\pi_1}(s')\left((I-\gamma P^{\pi_1})^{-1}(s'|s) - (I-\gamma P^{\pi_2})^{-1}(s'|s)\right) + (I-\gamma P^{\pi_2})^{-1}(s'|s)\left(r^{i,\pi_1}(s') - r^{i,\pi_2}(s')\right)\right| \\
\leq & \sum_{s'\in\mathbb{S}}\left(R_{\max}\frac{nSA_{\max}\delta}{(1-\gamma)^2} + \frac{1}{1-\gamma}nA_{\max}R_{\max}\delta\right) \\
= & \frac{nSA_{\max}R_{\max}}{1-\gamma}\left(1+\frac{S}{1-\gamma}\right)\delta \\
\leq & \frac{2nS^2A_{\max}R_{\max}}{(1-\gamma)^2}\delta
\end{aligned}
$$

where the forth line follows from $|(I-\gamma P^{\pi_2})^{-1}(s'|s)| \leq \frac{1}{1-\gamma}$, which will also be proved in Lemma 7.

Let $D^{\pi^i,\pi^{-i}}_{\pi^i(s,a^i)=1}(s) = \max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,a^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right)$. Then we have the following upper bounds directly

$$
\left|D^{\pi_1^i,\pi_1^{-i}}_{\pi_1^i(s,a^i)=1}(s) - D^{\pi_2^i,\pi_2^{-i}}_{\pi_2^i(s,a^i)=1}(s)\right| \leq \frac{4nS^2A_{\max}R_{\max}}{(1-\gamma)^2}\delta,
$$

$$
\left|\sum_{b^i\in\mathbb{A}^i}\left(D^{\pi_1^i,\pi_1^{-i}}_{\pi_1^i(s,b^i)=1}(s) - D^{\pi_2^i,\pi_2^{-i}}_{\pi_2^i(s,b^i)=1}(s)\right)\right| \leq \frac{4nS^2A_{\max}^2R_{\max}}{(1-\gamma)^2}\delta.
$$

Finally, we could complete our proof of this lemma. For any player $i \in [n]$, any state $s \in \mathbb{S}$ and any action $a^i \in \mathbb{A}^i$, we have

$$
\begin{aligned}
& \left|(f(\pi_1))^i(s,a^i) - (f(\pi_2))^i(s,a^i)\right| \\
\leq & \left|\pi_1^i(s,a^i) - \pi_2^i(s,a^i)\right| + \left|D^{\pi_1^i,\pi_1^{-i}}_{\pi_1^i(s,a^i)=1}(s) - D^{\pi_2^i,\pi_2^{-i}}_{\pi_2^i(s,a^i)=1}(s)\right| + \left|\sum_{b^i\in\mathbb{A}^i} D^{\pi_1^i,\pi_1^{-i}}_{\pi_1^i(s,b^i)=1}(s) - D^{\pi_2^i,\pi_2^{-i}}_{\pi_2^i(s,b^i)=1}(s)\right| \\
\leq & \delta + \frac{4nS^2A_{\max}R_{\max}}{(1-\gamma)^2}\delta + \frac{4nS^2A_{\max}^2R_{\max}}{(1-\gamma)^2}\delta \\
\leq & \frac{9nS^2A_{\max}^2R_{\max}}{(1-\gamma)^2}\delta.
\end{aligned}
$$

□

A.2 Proof of Lemma 7

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

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

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

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

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

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 $\|P^{\pi_1} - P^{\pi_2}\|_{\max} \leq nA_{\max}\delta$.

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

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

where the sixth line follows the following facts:

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

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

$$\|Q^2\|_1 \leq 1 + \gamma \|P^{\pi_2} Q^2\|_1 \leq 1 + \gamma \|P^{\pi_2}\|_1 \|Q^2\|_1 \leq 1 + \gamma \|Q^2\|_1,$$

which leads to the fourth fact.

So we have

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

□

B Detailed Proofs from Section 5

B.1 Proof of Lemma 4

Lemma 4. *Let $\epsilon > 0$ and π be a strategy profile. If $\|f(\pi) - \pi\|_\infty \leq \epsilon$, then for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$, we have*

$$\max\left(0, V^{\pi^i,\pi^{-i}}_{\pi^i(s,a^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) \leq A_{\max}\left(\frac{\sqrt{\epsilon'}}{1-\gamma} + R_{\max}\sqrt{\epsilon'} + \epsilon'\right),$$

where $\epsilon' = \epsilon\left(1 + \dfrac{A_{\max}R_{\max}}{1-\gamma}\right)$.

Proof of Lemma 4. Pick any player $i \in [n]$ and state $s \in \mathbb{S}$ in this proof. Suppose that the action space of player i is $\mathbb{A}^i = \{a^i_1, \cdots, a^i_{A^i}\}$. For the simplicity of notations, for any $a^i_j \in \mathbb{A}^i$, let

$$V_{a^i_j}(s) := V^{\pi^i,\pi^{-i}}_{\pi^i(s,a^i_j)=1}(s),$$

and

$$D_{a^i_j}(s) := \max\left(0, V_{a^i_j}(s) - V^{\pi^i,\pi^{-i}}(s)\right).$$

Without loss of generality, assume that

$$V_{a^i_1}(s) \geq V_{a^i_2}(s) \geq \cdots \geq V_{a^i_k}(s) \geq V^{\pi^i,\pi^{-i}}(s) \geq V_{a^i_{k+1}}(s) \geq \cdots \geq V_{a^i_{A^i}}(s).$$

We first give an upper bound of $D_{a^i_j}(s)$.

$$D_{a^i_j}(s) = \max\left(0, V_{a^i_j}(s) - V^{\pi^i,\pi^{-i}}(s)\right) \leq V_{a^i_j}(s) \leq R_{\max}/(1-\gamma).$$

For any action $a^i_j \in \mathbb{A}^i$, by the condition $\|f(\pi) - \pi\|_\infty \leq \epsilon$, we know that

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

Setting $\epsilon' = \epsilon\left(1 + \frac{A_{\max}R_{\max}}{1-\gamma}\right)$, we have the following crucial inequality:

$$\pi^i(s,a^i_j) \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \leq D_{a^i_j}(s) + \epsilon'. \tag{4}$$

Let $t := \sum_{j=k+1}^{A^i} \pi^i(s,a^i_j)$.
Case 1: $t \geq \sqrt{\epsilon'}/R_{\max}$.

Note that for $k+1 \le j \le A^i, D_{a_j^i}(s) = 0$. By the inequality (4), we have

$$\begin{aligned}
&\sum_{j=k+1}^{A^i} \left(\pi^i(s, a_j^i) \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \right) \le \sum_{j=k+1}^{A^i} \left(D_{a_j^i}(s) + \epsilon' \right) \\
\implies \quad & t \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \le A_{\max} \epsilon' \\
\implies \quad & D_{a_1^i}(s) \le \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \le A_{\max} R_{\max} \sqrt{\epsilon'}.
\end{aligned}$$

Case 2: $t \le \sqrt{\epsilon'}/R_{\max}$.
By the inequality (4), we have

$$\begin{aligned}
&\pi^i(s, a_j^i) \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \le D_{a_j^i}(s) + \epsilon' \\
\implies \quad & \pi^i(s, a_j^i)^2 \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \le \pi^i(s, a_j^i) \left(D_{a_j^i}(s) + \epsilon' \right) \\
\implies \quad & \sum_{j=1}^{A^i} \left(\pi^i(s, a_j^i)^2 \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \right) \le \sum_{j=1}^{A^i} \left(\pi^i(s, a_j^i) \left(D_{a_j^i}(s) + \epsilon' \right) \right) \\
\implies \quad & \sum_{j=1}^{A^i} \pi^i(s, a_j^i)^2 \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \le \sum_{j=1}^{k} \left(\pi^i(s, a_j^i) D_{a_j^i}(s) \right) + \epsilon' \\
\implies \quad & \sum_{j=1}^{A^i} \pi^i(s, a_j^i)^2 \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \le \frac{R_{\max}}{1-\gamma} \sum_{j=1}^{k} \pi^i(s, a_j^i) + \epsilon' \\
\implies \quad & \sum_{j=1}^{A^i} \pi^i(s, a_j^i)^2 \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \le \frac{R_{\max}}{1-\gamma} \frac{\sqrt{\epsilon'}}{R_{\max}} + \epsilon' \\
\implies \quad & \frac{1}{A^i} \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \le \frac{\sqrt{\epsilon'}}{1-\gamma} + \epsilon' \\
\implies \quad & D_{a_1^i}(s) \le \sum_{b^i \in \mathbb{A}^i} D_{b^i}(s) \le A_{\max} \left(\frac{\sqrt{\epsilon'}}{1-\gamma} + \epsilon' \right).
\end{aligned}$$

Note that the argument above could be applied to any player and any state, so for each player $i \in [n]$, each state $s \in \mathbb{S}$ and each action $a^i \in \mathbb{A}^i$, we have

$$\max \left(0, V^{\pi^i, \pi^{-i}}_{\pi^i(s,a^i)=1}(s) - V^{\pi^i, \pi^{-i}}(s) \right) \le A_{\max} \left(\frac{\sqrt{\epsilon'}}{1-\gamma} + R_{\max} \sqrt{\epsilon'} + \epsilon' \right).$$

□

B.2 Definition of the Set of Vertices Σ

Note that the strategy profile space is $\prod_{i=1}^n \Delta_{A^i}^S$, which is a production of unit simplices. We will adopt the techniques in [48] to triangulate the strategy profile space, where the set of vertices of END OF THE LINE graph will correspond to a set of simplices after triangulation.

We use $Q_{\mathbb{A}^i}$ to denote the $A^i \times A^i$ matrix

$$Q_{\mathbb{A}^i} = \begin{bmatrix} -1 & 0 & . & . & . & 1 \\ 1 & -1 & 0 & . & . & 0 \\ . & . & . & & & . \\ . & & . & . & & . \\ . & & & . & . & . \\ 0 & . & . & 0 & 1 & -1 \end{bmatrix}.$$

For each player $i \in [n]$, we use Q_i to denote the $A^i S \times A^i S$ matrix, which is a block diagonal matrix

$$Q_i = \left.\begin{bmatrix} Q_{\mathbb{A}^i} & 0 & \dots & 0 \\ 0 & Q_{\mathbb{A}^i} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Q_{\mathbb{A}^i} \end{bmatrix}\right\} S.$$

Finally, we use Q to denote the block diagonal matrix

$$Q = \begin{bmatrix} Q_1 & 0 & \dots & 0 \\ 0 & Q_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Q_n \end{bmatrix}.$$

For any agent $i \in [n]$, state $s \in \mathbb{S}$ and action $a^i \in \mathbb{A}^i$, we use $Q(i, s, a^i)$ to denote the corresponding column. Let v^0 be an arbitrary (starting) point in $\prod_{i=1}^n \Delta_{A^i}^S(d)$.

For each agent $i \in [n]$ and $s \in \mathbb{S}$, let $I_{i,s} = \{(i, s, a^i) | a^i \in \mathbb{A}^i\}$. Let $\mathcal{I}$ be a collection of all subsets T of $\bigcup_{i\in[n],s\in\mathbb{S}} I_{i,s}$ such that for each i and s there is at least one element (i, s, a^i) not in T.

For all $T \in \mathcal{I}$, we define $A(T)$, which is a subset of $\prod_{i=1}^n \Delta_{A^i}^S$, as follows.

$$A(T) = \left\{ x \in \prod_{i=1}^n \Delta_{A^i}^S | x = v^0 + \sum_{(i,s,a^i)\in T} \lambda(i, s, a^i) Q(i, s, a^i) \ \texttt{for} \ \lambda(i, s, a^i) \geq 0 \right\}.$$

Let us fix some $T \in \mathcal{T}$, $\phi : [|T|] \to T$ be a permutation of T, and $w^0 \in A(T) \cap \prod_{i=1}^n \Delta_{A^i}^S(d)$. We use $\Delta(w^0, \phi)$ to denote the convex hull of $|T| + 1$ vertices $\{w^0, w^1, \cdots, w^{|T|}\}$ (which is a simplex), where

$$w^i = w^{i-1} + Q(\phi(i)), \quad i \in [|T|].$$

Define

$$\Sigma_T = \{\Delta(w^0, \phi) | \Delta(w^0, \phi) \in A(T) \cap \prod_{i=1}^n \Delta_{A^i}^S(d), \phi \ \texttt{is a permutation of} \ T\}.$$

Then we have the following lemma.

Lemma 8 ([48]). *For each $T \in \mathcal{I}$, Σ_T triangulates $A(T)$.*

The Vertices of END OF THE LINE **Graph** are $\Sigma := \bigcup_{T\in\mathcal{I}} \Sigma_T$.

B.3 Proof of Lemma 6

Lemma 6. *([48]) Suppose that a simplex $\sigma \in \Sigma$ is (i,s)-stopping for $i \in [n]$ and $s \in \mathbb{S}$. Then for any strategy profile $\pi \in \sigma$, we have*

$$\|f(\pi) - \pi\|_\infty \leq A_{\max}^2(\lambda+1)\frac{1}{d}.$$

Proof of Lemma 6. Because of the triangulation, we know that for any simplex $\delta \in \Sigma$ and two strategy profiles $\pi, \pi' \in \delta$, $\|\pi - \pi'\|_\infty \leq 1/d$.

Now let the simplex $\delta \in \Sigma$ be (i,s)-stopping. By the definition, we know for any $a^i \in \mathbb{A}^i$, there is a strategy profile, denoted by $\pi_{a^i} \in \prod_{i=1}^n \Delta_{A^i}^S$, whose label is (i,s,a^i). Then

$$\left(f(\pi_{a^i})\right)^i(s,a^i) - \pi_{a^i}^i(s,a^i) \leq 0 \quad \forall a^i \in \mathbb{A}^i.$$

Then for any $\pi \in \delta$, $\forall a^i \in \mathbb{A}^i$, we have $\pi_{a^i}^i(s,a^i) - \pi^i(s,a^i) \leq \frac{1}{d}$ and f is λ-Lipschitz, which means

$$\left(f(\pi)\right)^i(s,a^i) - \pi^i(s,a^i) \leq \left(f(\pi_{a^i})\right)^i(s,a^i) - \pi_{a^i}^i(s,a^i) + (\lambda+1)\frac{1}{d} \leq (\lambda+1)\frac{1}{d}.$$

Using $\sum_{b^i \in \mathbb{A}^i} \pi^i(s,b^i) = \sum_{b^i \in \mathbb{A}^i} \left(f(\pi^i)\right)(s,b^i) = 1$, we have

$$\begin{aligned}
& \left(f(\pi)\right)^i(s,a^i) - \pi^i(s,a^i) \\
= & \sum_{b^i \in \mathbb{A}^i, b^i \neq a^i} \pi^i(s,b^i) - \sum_{b^i \in \mathbb{A}^i, b^i \neq a^i} \left(f(\pi^i)\right)^i(s,b^i) \\
\geq & -(A_{\max}-1)(\lambda+1)\frac{1}{d}.
\end{aligned}$$

Pick $a^i \in \mathbb{A}^i$ arbitrarily. Combine with the definition of labelling rule, for any $\pi \in \delta$, $j \in [n]$, $v \in \mathbb{S}$ and $b^j \in \mathbb{A}^j$, we have

$$\begin{aligned}
& \left(f(\pi)\right)^j(v,b^j) - \pi^j(v,b^j) \\
\geq & \left(f(\pi_{a^i})\right)^j(v,b^j) - \pi_{a^i}^j(v,b^j) - (\lambda+1)\frac{1}{d} \\
\geq & \left(f(\pi_{a^i})\right)^i(s,a^i) - \pi_{a^i}^i(s,a^i) - (\lambda+1)\frac{1}{d} \\
\geq & -A_{\max}(\lambda+1)\frac{1}{d},
\end{aligned}$$

which finishes the lower bound of this lemma.

Also, by the similar argument $\sum_{b^j \in \mathbb{A}^j} \pi^j(v,b^j) = \sum_{b^j \in \mathbb{A}^j} \left(f(\pi^j)\right)^j(v,b^j) = 1$, we know that

$$\begin{aligned}
& \left(f(\pi)\right)^j(v,b^j) - \pi^j(v,b^j) \\
\leq & A_{\max}(A_{\max}-1)(\lambda+1)\frac{1}{d} \\
\leq & A_{\max}^2(\lambda+1)\frac{1}{d},
\end{aligned}$$

which finished the upper bound of this lemma. □

B.4 Correctness of Our Choice of d

By Lemma 6, we will find a π such that

$$\|f(\pi) - \pi\|_\infty \leq \frac{(1-\gamma)^5}{32A_{\max}^3 R_{\max}^3} \frac{1}{L^2}.$$

In Lemma 4, we will have

$$\epsilon' = \frac{(1-\gamma)^5}{32A_{\max}^3 R_{\max}^3} \frac{1}{L^2} \left(1 + \frac{A_{\max} R_{\max}}{1-\gamma}\right) \leq \frac{(1-\gamma)^5}{32A_{\max}^3 R_{\max}^3} \frac{1}{L^2} \frac{2A_{\max} R_{\max}}{1-\gamma} \leq \frac{(1-\gamma)^4}{16A_{\max}^2 R_{\max}^2} \frac{1}{L^2},$$

which means

$$\begin{aligned}
& \max\left(0, V^{\pi^i, \pi^{-i}}_{\pi^i(s,a^i)=1}(s) - V^{\pi^i,\pi^{-i}}(s)\right) \\
\leq\ & A_{\max}\left(\frac{\sqrt{\epsilon'}}{1-\gamma} + R_{\max}\sqrt{\epsilon'} + \epsilon'\right) \\
\leq\ & A_{\max}\left(\frac{2R_{\max}}{1-\gamma}\sqrt{\epsilon'} + \epsilon'\right) \\
\leq\ & A_{\max}\left(\frac{2R_{\max}}{1-\gamma}\frac{(1-\gamma)^2}{4A_{\max}R_{\max}}\frac{1}{L} + \frac{(1-\gamma)^4}{16A_{\max}^2 R_{\max}^2}\frac{1}{L^2}\right) \\
\leq\ & \frac{(1-\gamma)}{2}\frac{1}{L} + \frac{(1-\gamma)^4}{16A_{\max}R_{\max}^2}\frac{1}{L^2}.
\end{aligned}$$

By Lemma 5, we know π is an $\frac{1}{2L} + \frac{(1-\gamma)^3}{16A_{\max}R_{\max}^2}\frac{1}{L^2}$-approximate MPE, so π is a $1/L$-approximate MPE.

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il