Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-30 (2022) 73–78

Gradient Play in Stochastic Games: Stationary Points and Local Geometry

Runyu (Cathy) Zhang * **Zhaolin Ren** * **Na Li** *

* *Harvard University (e-mail: runyuzhang@fas.harvard.edu, zhaolinren@g.harvard.edu, nali@seas.harvard.edu)*

Abstract: We study the stationary points and local geometry of gradient play for stochastic games (SGs), where each agent tries to maximize its own total discounted reward by making decisions *independently* based on current state information which is shared between agents. Policies are directly parameterized by the probability of choosing a certain action at a given state. We show that Nash equilibria (NEs) and first-order stationary policies are equivalent in this setting by establishing a gradient domination condition for SGs. We characterize the structure of strict NEs and show that gradient play locally converges to strict NEs within finite steps. Further, for a subclass of SGs called Markov potential games, we prove that strict NEs are local maxima of the total potential function, thus locally stable under gradient play, and fully-mixed NEs are saddle points, thus unstable under gradient play.

Keywords: Stochastic games, Markov potential games, gradient play, Nash equilibrium

1. INTRODUCTION

Multi-agent systems find applications in a wide range of societal systems, e.g. electric grids, traffic networks, smart buildings and smart cities etc. Given the complexity of these systems, multi-agent reinforcement learning (MARL) has gained increasing attention in recent years (Shalev-Shwartz et al., 2016; Vidhate and Kulkarni, 2017; Xu et al., 2020). Among MARL algorithms, policy gradient-type methods are highly popular because of their flexibility and capability to incorporate structured state and action spaces. However, while many recent works (Zhang et al., 2018; Chen et al., 2018; Wai et al., 2018; Li et al., 2019; Qu et al., 2020) have studied the performance of multi-agent policy gradient algorithms, due to a lack of understanding of the optimization landscape in these multi-agent learning problems, most works can only show convergence to a first-order stationary point. Deeper understanding of the quality of these stationary points is missing even in the simple identical-reward multi-agent RL setting.

In this paper, we examine this problem from a game-theoretic perspective. We model the multi-agent system as a stochastic game (SG) where agents can have different reward functions, and study the dynamical behavior of first-order (gradient-based) learning methods. The study of SGs dates back to as early as the 1950s by Shapley (1953) with a series of followup works on developing NE-seeking algorithms, especially in the RL setting (e.g. (Littman, 1994; Buşoniu et al., 2010; Lanctot et al., 2017; Zhang et al., 2019a) and citations therein). While well-known classical algorithms for solving SGs are mostly value-based, such as Nash-Q learning (Hu and Wellman, 2003), Hyper-Q learning (Tesauro, 2003), and WoLF-PHC (Bowling and Veloso, 2001), gradient-based algorithms have also started to gain popularity in recent years due to their advantages as mentioned earlier (e.g. (Zhang and Lesser, 2010; Foerster et al., 2017)). In this work, we aim to gain a deeper understanding of the structure and quality of first-order stationary points for these gradient-based methods, with a particular focus on answering the following questions: 1) How do the first-order stationary points relate to the NEs of the underlying game?, 2) Do gradient-based algorithms guarantee convergence to a NE?, 3) What is the stability of the individual NEs?.

These questions have already been widely discussed in other settings, e.g., one-shot (stateless) finite-action games (Shapley, 1964; Jordan, 1993; Krishna and Sjöström, 1998; Van Damme, 1991), one-shot continuous games (Mazumdar et al., 2020), zero-sum linear quadratic (LQ) games Zhang et al. (2019b), etc. There are both negative and positive results depending on the settings. For one-shot continuous games, (Mazumdar et al., 2020) proved a negative result suggesting that gradient flow has stationary points (even local maxima) that are not necessarily NEs. Conversely, Zhang et al. (2019b) designed projected nested-gradient methods that provably converge to NEs in zero-sum LQ games. However, much less is known in the tabular setting of SGs with finite state-action spaces.

Contributions. In our paper, we consider the *gradient play* algorithm for the infinite time-discounted reward SG where an agent's local policy is directly parameterized by the probability of choosing an action from the agent's own action space at a given state. We focus on the tabular setting where state and action spaces are finite. Through generalizing the gradient domination property in (Agarwal et al., 2020) to the multi-agent setting studied in this paper, we first establish the equivalence of first-order stationary policies and Nash equilibria (Theorem 1).

Then we study the convergence of gradient play for SGs. For general games, it is known that gradient play may fail to have global convergence (Shapley, 1964; Crawford, 1985; Jordan, 1993; Krishna and Sjöström, 1998). Thus we firstly focus on characterizing some local properties for the

10.1016/j.ifacol.2022.11.031

general cases. In particular, we characterize the structure of strict NEs and show that gradient play locally converges to strict NEs within finite steps (Theorem 2).

Next we study a special class of SGs called Markov potential games (MPGs) (González-Sánchez and Hernández-Lerma, 2013; Macua et al., 2018; Leonardos et al., 2021), which includes identical reward multi-agent RL (Tan, 1993; Panait and Luke, 2005) as an important special case. Though global convergence rate results were established recently by Zhang et al. (2021); Leonardos et al. (2021) for gradient play under MPGs, these results only bound the NE-gap (c.f. Definition 4) but do not say the convergence of the policies or which NE the policies converge to, even if they do converge. Given the fact that there are many NEs that would have poor global value, global convergence results has a limited implication on the algorithm performance. This motivate us study the local geometry around some specific types of NEs. In this paper, we show that strict NEs are local maxima of the total potential function, thus stable points under gradient play, and that fully mixed NEs are saddle points, thus unstable points under gradient play. Lastly, we note that results of this paper are posted online in our ArXiv report (Zhang et al., 2021), which contains all the proofs, numerical study and extra results.

2. PROBLEM SETTING AND PRELIMINARIES

We consider a stochastic game (SG) $\mathcal{M} = (N, \mathcal{S},\ \mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n,\ P,\ r = (r_1, \ldots, r_n),\ \gamma,\ \rho)$ with n agents (Shapley, 1953) which is specified by an agent set $N = \{1, 2, \ldots, n\}$, a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}_i$ for each agent $i \in N$, a transition model P where $P(s'|s, a) = P(s'|s, a_1, \ldots, a_n)$ is the probability of transitioning into state s' upon taking action $a := (a_1, \ldots, a_n)$ in state s where $a_i \in \mathcal{A}_i$ is action of agent i, agent i's reward function $r_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$, a discount factor $\gamma \in [0, 1)$, and an initial state distribution ρ over $\mathcal{S}$.

A stochastic policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ (where $\Delta(\mathcal{A})$ is the probability simplex over $\mathcal{A}$) specifies a strategy in which agents choose their actions *jointly* based on the current state in a stochastic fashion, i.e. $\Pr(a_t|s_t) = \pi(a_t|s_t)$. A distributed stochastic policy is a special subclass of stochastic policies, with $\pi = \pi_1 \times \ldots \times \pi_n$, where $\pi_i : \mathcal{S} \to \Delta(\mathcal{A}_i)$. For distributed stochastic policies, each agent takes its action based on the current state s *independently of* other agents' choices of actions, i.e.:

$$\Pr(a_t|s_t) = \pi(a_t|s_t) = \prod_{i=1}^{n} \pi_i(a_{i,t}|s_t), \quad a_t = (a_{1,t}, \ldots, a_{n,t}).$$

For notational simplicity, we define: $\pi_I(a_I|s) := \prod_{i \in I} \pi_i(a_i|s)$, where $I \subseteq N$ is an index set. Further, we use the notation $-i$ to denote the index set $N \backslash \{i\}$.

We consider *direct distributed policy parameterization*, where agent i's policy is parameterized by θ_i:

$$\pi_{i,\theta_i}(a_i|s) = \theta_{i,(s,a_i)}, \quad i = 1, 2, \ldots, n. \tag{1}$$

For notational simplicity, we abbreviate $\pi_{i,\theta_i}(a_i|s)$ as $\pi_{\theta_i}(a_i|s)$, and $\theta_{i,(s,a_i)}$ as θ_{s,a_i}. Here $\theta_i \in \Delta(\mathcal{A}_i)^{|\mathcal{S}|}$, i.e. θ_i is subject to the constraints $\theta_{s,a_i} \geq 0$ and $\sum_{a_i \in \mathcal{A}_i} \theta_{s,a_i} = 1$ for all $s \in \mathcal{S}$. The global joint policy is given by: $\pi_\theta(a|s) = \prod_{i=1}^n \pi_{\theta_i}(a_i|s) = \prod_{i=1}^n \theta_{s,a_i}$. We use $\mathcal{X}_i := \Delta(\mathcal{A}_i)^{|\mathcal{S}|}, \mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ to denote the feasible region of θ_i and θ.

Agent i's value function $V_i^\theta : \mathcal{S} \to \mathbb{R}, i \in N$ is defined as the discounted sum of future rewards starting at state s via executing π_θ, i.e.

$$V_i^\theta(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \Big|\ \pi_\theta, s_0 = s\right],$$

where the expectation is with respect to the random trajectory $\tau = (s_t, a_t, r_{i,t})_{t=0}^\infty$ where $a_t \sim \pi_\theta(\cdot|s_t), s_{t+1} = P(\cdot|s_t, a_t)$. We denote agent i's total reward starting from initial state $s_0 \sim \rho$ as:

$$J_i(\theta) = J_i(\theta_1, \ldots, \theta_n) := \mathbb{E}_{s_0 \sim \rho} V_i^\theta(s_0).$$

In the game setting, agent i's incentive is to maximize its own total reward J_i. A Nash equilibrium (NE) is often used to characterize the equilibrium where no agent has a unilateral incentive to deviate from it.

Definition 1. (Nash equilibrium) A policy $\theta^* = (\theta_1^*, \ldots, \theta_n^*)$ is called a Nash equilibrium (NE) if

$$J_i(\theta_i^*, \theta_{-i}^*) \geq J_i(\theta_i', \theta_{-i}^*), \quad \forall \theta_i' \in \mathcal{X}_i, \quad i \in N$$

The equilibrium is called a strict NE if the inequality holds strictly for all $\theta_i' \in \mathcal{X}_i, \theta_i' \neq \theta_i$ and $i \in N$. The equilibrium is called a pure NE if θ^* corresponds to a deterministic policy. The equilibrium is called a mixed NE if it is not pure. Further, the equilibrium is called a fully mixed NE if every entry of θ^* is strictly positive, i.e.: $\theta_{s,a_i}^* > 0,\ \ \forall\ a_i \in \mathcal{A}_i,\ \ \forall\ s \in \mathcal{S},\ \ i \in N$

We define the *discounted state visitation distribution* d_θ of a policy π_θ given an initial state distribution ρ as:

$$d_\theta(s) := \mathbb{E}_{s_0 \sim \rho}(1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \Pr^\theta(s_t = s|s_0), \tag{2}$$

where $\Pr^\theta(s_t = s|s_0)$ is the state visitation probability that $s_t = s$ when executing π_θ starting at state s_0. Throughout the paper, we make the following assumption on the SGs we study.

Assumption 1. The stochastic game $\mathcal{M}$ satisfies: $d_\theta(s) > 0,\ \forall s \in \mathcal{S},\ \forall \theta \in \mathcal{X}$.

Assumption 1 requires that every state is visited with positive probability, which is a standard assumption for convergence proofs in the RL literature (e.g. (Agarwal et al., 2020; Mei et al., 2020)).

Similar to centralized RL, we define agent i's Q-function $Q_i^\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and its advantage function $A_i^\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as:

$$Q_i^\theta(s, a) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) \Big|\ \pi_\theta, s_0 = s, a_0 = a\right],$$
$$A_i^\theta(s, a) := Q_i^\theta(s, a) - V_i^\theta(s).$$

***'Averaged'* Markov decision process (MDP):** We further define agent i's *'averaged' Q-function* $\overline{Q_i^\theta} : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}$ and *'averaged' advantage-function* $\overline{A_i^\theta} : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}$ as:

$$\begin{aligned} \overline{Q_i^\theta}(s, a_i) &:= \sum_{a_{-i}} \pi_{\theta_{-i}}(a_{-i}|s) Q_i^\theta(s, a_i, a_{-i}), \\ \overline{A_i^\theta}(s, a_i) &:= \sum_{a_{-i}} \pi_{\theta_{-i}}(a_{-i}|s) A_i^\theta(s, a_i, a_{-i}). \end{aligned} \tag{3}$$

Similarly, we define agent i's *'averaged' transition probability distribution* $\overline{P_i^\theta} : \mathcal{S} \times \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}$, and *'averaged' reward* $\overline{r_i^\theta} : \mathcal{S} \times \mathcal{A}_i \to \mathbb{R}$ as:

$$\overline{P_i^\theta}(s'|s, a_i) := \sum_{a_{-i}} \pi_{\theta-i}(a_{-i}|s) P(s'|s, a_i, a_{-i}),$$
$$\overline{r_i^\theta}(s, a_i) := \sum_{a_{-i}} \pi_{\theta-i}(a_{-i}|s) r_i(s, a_i, a_{-i})$$

From its definition, the averaged Q-function satisfies the following Bellman equation:

Lemma 1. $\overline{Q_i^\theta}$ satisfies:

$$\overline{Q_i^\theta}(s, a_i) = \overline{r_i^\theta}(s, a_i) + \gamma \sum_{s', a_i'} \pi_{\theta_i}(a_i'|s') \overline{P_i^\theta}(s'|s, a_i) \overline{Q_i^\theta}(s', a_i')$$

Lemma 1 suggests that the averaged Q-function $\overline{Q_i^\theta}$ is indeed the Q-function for the MDP defined on action space $\mathcal{A}_i$, with $\overline{r_i^\theta}, \overline{P_i^\theta}$ as its stage reward and transition probability respectively. We define this MDP as the *'averaged' MDP* of agent i, i.e., $\mathcal{M}_i^\theta = (\mathcal{S}, \mathcal{A}_i, \overline{P_i^\theta}, \overline{r_i^\theta}, \gamma, \rho)$. Note that the 'averaged' MDP is only well-defined when the policies of the other agents θ_{-i} are kept fixed. When this is indeed the case, agent i can be treated as an independent learner with respect to its own 'averaged' MDP. This observation serves as an important intuition for our theoretical results, for example, we can apply performance difference lemma (Kakade and Langford, 2002) to the averaged MDP to derive a corresponding lemma for SGs which is useful throughout the paper (see Appendix C in our online supplementary material (Zhang et al., 2022) for more detail).

Lemma 2. (Performance difference lemma, for SGs) Let $\theta' = (\theta_i', \theta_{-i})$

$$J_i(\theta_i', \theta_{-i}) - J_i(\theta_i, \theta_{-i}) = \frac{1}{1-\gamma} \sum_{s, a_i} d_{\theta'}(s) \pi_{\theta_i'}(a_i|s) \overline{A_i^\theta}(s, a_i).$$

Note that in the single agent case ($n = 1$), Lemma 2 is the same as the original performance difference lemma known in literature.

3. GRADIENT PLAY FOR GENERAL STOCHASTIC GAMES

Under direct distributed parameterization, the gradient play algorithm is given by:

$$\theta_i^{(t+1)} = Proj_{\mathcal{X}_i}(\theta_i^{(t)} + \eta \nabla_{\theta_i} J_i(\theta_i^{(t)})), \eta > 0. \quad (4)$$

Gradient play can be viewed as a 'better response' strategy, where agents update their own parameters by gradient ascent with respect to their own rewards. A first-order stationary point is defined as such:

Definition 2. (First-order stationary policy) A policy $\theta^* = (\theta_1^*, \ldots, \theta_n^*)$ is called a first-order stationary policy if

$$(\theta_i' - \theta_i^*)^\top \nabla_{\theta_i} J_i(\theta^*) \le 0, \ \forall \theta_i' \in \mathcal{X}_i, \ i \in N.$$

It is not hard to verify that θ^* is a first-order stationary policy if and only if it is a fixed point under gradient play (4). Comparing Definition 1 (of NE) and Definition 2, we know that NEs are first-order stationary policies, but not necessarily vice versa. For each agent i, first-order stationarity does not imply that θ_i^* is optimal among all possible θ_i given θ_{-i}^*. However, interestingly, we will show that NEs are equivalent to first-order stationary policies due to a gradient domination property that we will show later. Before that, we first calculate the explicit form of the gradient $\nabla_{\theta_i} J_i$.

Policy gradient theorem (Sutton et al., 1999) gives an efficient formula for the gradient:

$$\nabla_\theta \mathbb{E}_{s_0 \sim \rho} V_i^\theta(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta(\cdot|s)} [\nabla_\theta \log \pi_\theta(a|s) Q_i^\theta(s, a)]. \quad (5)$$

Applying (5), the gradient $\nabla_{\theta_i} J_i$ can be written explicitly as follows:

Lemma 3. (Proof in Appendix B of our online supplementary material (Zhang et al., 2022)) For direct distributed parameterization (1),

$$\frac{\partial J_i(\theta)}{\partial \theta_{s, a_i}} = \frac{1}{1-\gamma} d_\theta(s) \overline{Q_i^\theta}(s, a_i) \quad (6)$$

Gradient domination and the equivalence between NE and first-order stationary policy. Lemma 4.1 in Agarwal et al. (2020) established gradient domination for centralized tabular MDP under direct parameterization. We can show that a similar property still holds for stochastic games.

Lemma 4. (Gradient domination) For direct distributed parameterization (1), we have that for any $\theta = (\theta_1, \ldots, \theta_n) \in \mathcal{X}$ and any $\theta_i' \in \mathcal{X}_i, \ i \in N$:

$$J_i(\theta_i', \theta_{-i}) - J_i(\theta_i, \theta_{-i}) \le \left\| \frac{d_{\theta'}}{d_\theta} \right\|_\infty \max_{\overline{\theta}_i \in \mathcal{X}_i} (\overline{\theta}_i - \theta_i)^\top \nabla_{\theta_i} J_i(\theta), \quad (7)$$

where $\left\| \frac{d_{\theta'}}{d_\theta} \right\|_\infty := \max_s \frac{d_{\theta'}(s)}{d_\theta(s)}$, and $\theta' = (\theta_i', \theta_{-i})$.

The proof of Lemma 4 resembles the proof technique in Agarwal et al. (2020). Agarwal et al. (2020) leverage performance difference lemma for centralized MDP to derive their result, while we can replace the performance difference lemma by Lemma 2 to prove Lemma 4. The detailed proof can be found in Appendix C of our online supplementary material (Zhang et al., 2022).

Our result (7) is consistent with the result in Agarwal et al. (2020) for the single-agent case ($n = 1$), i.e.: $J(\theta') - J(\theta) \le \left\| \frac{d_{\theta'}}{d_\theta} \right\|_\infty \max_{\overline{\theta} \in \mathcal{X}} (\overline{\theta} - \theta)^\top \nabla J(\theta)$. However, when there are multiple agents, the condition is much weaker because the inequality requires θ_{-i} to be fixed. When $n = 1$, gradient domination rules out the existence of stationary points that are not global optima. For the multi-agent case, the property can no longer guarantee the equivalence between first-order stationarity and global optimality; instead, it links the stationary points with NEs as shown in the next theorem.

Theorem 1. Under Assumption 1, first-order stationary policies and NEs are equivalent.

The proof of Theorem 1 is given in Appendix C in the online supplementary material (Zhang et al., 2022). Before moving on to the local convergence results, we would like to point that the equivalence established in Theorem 1 cannot be generalized to settings other than tabular SGs. For example, Mazumdar et al. (2020) construct counterexamples for continuous games using quadratic functions. Our result does not contradict their work because

their counterexamples are not tabular SG as studied in this paper and the utility functions there may not be gradient dominant.

Local convergence for strict NEs Although the equivalence of NEs and stationary points under gradient play has been established, it is in fact difficult to show that gradient play converges to these stationary points. Even in the simpler static (stateless) game setup, gradient play might fail to converge (Shapley, 1964; Crawford, 1985; Jordan, 1993; Krishna and Sjöström, 1998). One major difficulty is that the vector field $\{\nabla_{\theta_i} J_i(\theta)\}_{i=1}^n$ is not a conservative vector field (definition see e.g. Marsden and Tromba (2003)). Accordingly, its dynamics may display complicated behavior. Thus, as a preliminary study, instead of looking at global convergence, we focus on the local convergence and restrict our study to a special subset of NEs - the strict NEs. We begin by giving the following characterization of strict NEs:

Lemma 5. Given a stochastic game $\mathcal{M}$, any strict NE θ^* is pure, meaning that for each i and s, there exist one $a_i^*(s)$ such that $\theta^*_{s,a_i} = \mathbf{1}\{a_i = a_i^*(s)\}$. Additionally,

$$\begin{aligned} &i)\ \ a_i^*(s) = \arg\max_{a_i} \overline{A_i^{\theta^*}}(s, a_i), \\ &ii)\ \overline{A_i^{\theta^*}}(s, a_i^*(s)) = 0; \\ &iii)\ \ \overline{A_i^{\theta^*}}(s, a_i) < 0,\ \forall\ a_i \neq a_i^*(s) \end{aligned} \tag{8}$$

Based on this lemma, we define the following for studying the local convergence of a strict NE θ^*:

$$\begin{aligned} &\Delta_i^{\theta^*}(s) := \min_{a_i \neq a_i^*(s)} \left| \overline{A_i^{\theta^*}}(s, a_i) \right|, \\ &\Delta^{\theta^*} := \min_i \min_s \frac{1}{1-\gamma} d_{\theta^*}(s) \Delta_i^{\theta^*}(s) > 0. \end{aligned} \tag{9}$$

Theorem 2. (Local finite time convergence around strict NE) Define the metric of policy parameters as: $D(\theta||\theta') := \max_{1\leq i\leq n} \max_{s\in\mathcal{S}} \|\theta_{i,s} - \theta'_{i,s}\|_1$, where $\|\cdot\|_1$ denote the ℓ_1-norm. Suppose θ^* is a *strict* Nash equilibrium, then for any $\theta^{(0)}$ such that $D(\theta^{(0)}||\theta^*) \leq \frac{\Delta^{\theta^*}(1-\gamma)^3}{8n|\mathcal{S}|(\sum_{i=1}^n |\mathcal{A}_i|)}$, running gradient play (4) will guarantee $D(\theta^{(t+1)}||\theta^*) \leq \max\left\{D(\theta^{(t)}||\theta^*) - \frac{\eta\Delta^{\theta^*}}{2}, 0\right\}$, which means that gradient play is going to converge within $\lceil \frac{2D(\theta^{(0)}||\theta^*)}{\eta\Delta^{\theta^*}} \rceil$ steps.

Proofs of Lemma 5 and Theorem 2 are given in Appendix D of our online supplementary material (Zhang et al., 2022).

Remark 1. Note that the local convergence in Theorem 2 only requires a finite number of steps and the stepsize η can be chosen arbitrarily large so that exact convergence can happen in even just one step after projection to the feasible set $\mathcal{X}$. However, the caveat is that we need to assume that the initial policy is sufficiently close to θ^*. For numerical stability considerations, one should pick reasonable stepsizes to run the algorithm to accommodate random initializations. Theorem 2 also shows that the radius of region of attraction for strict NEs is at least $\frac{\Delta^{\theta^*}(1-\gamma)^3}{8n|\mathcal{S}|(\sum_{i=1}^n |\mathcal{A}_i|)}$, and thus θ^* with a larger Δ^{θ^*}, i.e., a larger value gap between the optimal action and other actions, will have a larger region of attraction. We would like to further remark that Theorem 2 only focuses on the local convergence property; hence, we can interpret the theorem in the following way: if there exists a strict NE, then it is locally asymptotically stable under gradient play. However, it does not claim to solve the global existence or convergence of the strict NEs.

4. GRADIENT PLAY FOR MARKOV POTENTIAL GAMES

We have discussed that the main problem for the global convergence of gradient play for general SGs is that the vector field $\{\nabla_{\theta_i} J_i(\theta)\}_{i=1}^n$ is not conservative. Thus, in this section, we restrict our analysis to a special subclass, Markov Potential Games (MPG), where the vector field is conservative, which in turn enjoys global convergence (Zhang et al., 2021; Leonardos et al., 2021).

Definition 3. (Markov potential game (Macua et al., 2018; Zhang et al., 2021; Leonardos et al., 2021)) A stochastic game $\mathcal{M}$ is called a Markov potential game if there exists a potential function $\phi : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \to \mathbb{R}$ such that for any agent i and any pair of policy parameters $(\theta'_i, \theta_{-i}), (\theta_i, \theta_{-i})$:

$$\begin{aligned} &\mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t r_i(s_t, a_t) \big| \pi = (\theta'_i, \theta_{-i}), s_0 = s\right] \\ -&\mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_i(s_t, a_t) \big| \pi = (\theta_i, \theta_{-i}), s_0 = s\right] \\ =&\mathbb{E}\left[\textstyle\sum_{t=0}^\infty \gamma^t \phi(s_t, a_t) \big| \pi = (\theta'_i, \theta_{-i}), s_0 = s\right] \\ -&\mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \phi(s_t, a_t) \big| \pi = (\theta_i, \theta_{-i}), s_0 = s\right], \quad \forall\ s. \end{aligned}$$

More discussions on MPG including conditions to verify a SG is a MPG could be found in Macua et al. (2018); Zhang et al. (2021); Leonardos et al. (2021). We defer readers to these references. Note that identical interest game where agents share a same reward function naturally satisfies the above condition and serves as one important special case of MPG.

Given a policy θ, we define the *'total potential function'* $\Phi(\theta) := \mathbb{E}_{s_0\sim\rho(\cdot)}\left[\sum_{t=0}^\infty \gamma^t \phi(s_t, a_t) \big|\ \pi_\theta\right]$ for a MPG. From the definition of the total potential function we obtain the following relationship

$$J_i(\theta'_i, \theta_{-i}) - J_i(\theta_i, \theta_{-i}) = \Phi(\theta'_i, \theta_{-i}) - \Phi(\theta_i, \theta_{-i}). \tag{10}$$

Thus,

$$\nabla_{\theta_i} J_i(\theta) = \nabla_{\theta_i} \Phi(\theta),$$

which means that gradient play (4) is equivalent to running projected gradient ascent with respect to the total potential function Φ, i.e.:

$$\theta^{(t+1)} = Proj_{\mathcal{X}}(\theta^{(t)} + \eta \nabla_\theta \Phi(\theta_i^{(t)})),\ \eta > 0.$$

With this property people have established the global convergence for gradient play to a ϵ-NE for MPG Zhang et al. (2021); Leonardos et al. (2021). We cite the definition of ϵ-NE and the theorem here for the sake of self-completeness.

Definition 4. (ϵ-Nash equilibrium) Define the 'NE-gap' of a policy θ as:

$$\begin{aligned} \texttt{NE-gap}_i(\theta) &:= \max_{\theta'_i \in \mathcal{X}_i} J_i(\theta'_i, \theta_{-i}) - J_i(\theta_i, \theta_{-i}); \\ \texttt{NE-gap}(\theta) &:= \max_i \texttt{NE-gap}_i(\theta). \end{aligned}$$

A policy θ is an ϵ-Nash equilibrium if: $\texttt{NE-gap}(\theta) \leq \epsilon$.

Theorem 3. ((Zhang et al., 2021; Leonardos et al., 2021)) Suppose that total potential function Φ is bounded, i.e.,

for all $\theta \in \mathcal{X}$, $\Phi_{\min} \le \Phi(\theta) \le \Phi_{\max}$, then with stepsize $\eta = \frac{(1-\gamma)^3}{2\sum_{i=1}^n |\mathcal{A}_i|}$, the NE-gap of $\theta^{(t)}$ asymptotically converge to 0 under gradient play (4), i.e., $\lim_{t\to\infty} \texttt{NE-gap}(\theta^{(t)}) = 0$. Further, we have:

$$\frac{1}{T}\sum_{1\le t\le T} \texttt{NE-gap}(\theta^{(t)})^2 \le \epsilon^2, \quad \text{whenever } T \ge \frac{64M^2(\Phi_{\max}-\Phi_{\min})|\mathcal{S}|\sum_{i=1}^n |\mathcal{A}_i|}{(1-\gamma)^3\epsilon^2}, \tag{11}$$

where $M := \max_{\theta,\theta'\in\mathcal{X}} \left\| \frac{d_\theta}{d_{\theta'}} \right\|_\infty$ (by Assumption 1, we know that this quantity is well-defined, and if the initial state distribution satisfies $\rho(s) > 0$ for all s, M can be bounded by $M \le 1/(1-\gamma)\min_s \rho(s)$).

Quality of NEs. Theorem 3 suggests that gradient play is guaranteed to converge to a NE, however, which exact NE it converges to is not specified in the theorem.The qualities of NEs can vary significantly. For example, consider a simple two-agent identical-interest normal form game with reward table given in Table 1. There are three NEs. Two of them are strict NEs, where both agents choose the same action, i.e. $a_1 = a_2 = 1$ or 2. Both NEs are of reward 1. Another NE is a fully mixed NE, where both agents choose action 1 and 2 randomly with probability $\frac{1}{2}$. This NE is only of reward $\frac{1}{2}$. This significant quality difference between different types of NEs motivates us to further understand whether gradient play can find NEs with relatively good qualities. Since the NE that gradient play converges to depends on the initialization as well as the local geometry around the NE, as a preliminary study, we characterize the local geometry and landscape for strict NEs as well as fully mixed NEs (stated in the following theorem). More future investigation is needed for non-strict, non-fully-mixed NEs.

	$a_2 = 1$	$a_2 = 2$
$a_1 = 1$	1	0
$a_1 = 2$	0	1

Table 1.

Theorem 4. For a Markov potential game with $\Phi_{\min} < \Phi_{\max}$ (i.e., Φ is not a constant function):

- A strict NE θ^* is equivalent to a strict local maximum of the total potential function Φ, i.e.: $\exists\ \delta$, such that for all $\theta \in \mathcal{X}, \theta \ne \theta^*$ that satisfies $\|\theta - \theta^*\| \le \delta$, we have $\Phi(\theta) < \Phi(\theta^*)$.
- Any fully mixed NE θ^* is a saddle point with regard to the total potential function Φ that satisfies: $\forall\ \delta > 0$, $\exists\ \theta \in \mathcal{X}$, such that $\|\theta - \theta^*\| \le \delta$ and $\Phi(\theta) > \Phi(\theta^*)$.

The proof of Theorem 4 is given in Appendix E in our online supplementary material (Zhang et al., 2022).

Remark 2. Theorem 4 implies that strict NEs are asymptotically locally stable under first-order methods such as gradient play; while the fully mixed NEs are unstable under gradient play. Note that the theorem does not claim stability or instability for other types of NEs, e.g., pure NEs or non-fully mixed NEs. Nonetheless, we believe that these preliminary results can serve as a valuable platform towards a better understanding of the geometry of the problem. We remark that the conclusion about strict NEs in Theorem 4 does not hold for settings other than tabular MPG; for instance, for continuous games, one can use quadratic functions to construct simple counterexamples (Mazumdar et al., 2020). Also, similar to Remark 1, this theorem focuses on the local geometry of the NEs but does not claim the global existence or convergence of either strict NEs or fully mixed NEs.

5. CONCLUSIONS AND DISCUSSIONS

This paper studies the optimization landscape of multi-agent reinforcement learning through a game theoretic point of view. Specifically, we look into the tabular stochastic game problem and prove that all first order stationary policies are NEs under this setting. We also give a local convergence rate around strict NEs. For a special subclass of stochastic games called the Markov potential game, we have shown that strict NEs are the local maxima of the total potential function and fully mixed NEs are saddle points.

We believe that this is a fruitful research direction with many interesting open questions. For instance, one could explore generalizing our work (which assumes access to exact gradients) to a setting where gradients are estimated using data samples. Extending our results beyond direct policy parameterization to e.g. softmax parameterization (cf. (Agarwal et al., 2020)), is another interesting topic. Other interesting questions include local stability analysis in more general games (beyond Markov potential games), faster algorithm design (via e.g. natural policy gradient, Gauss-Newton methods), and online algorithm design for stochastic learning.

ACKNOWLEDGEMENTS

Runyu Zhang and Zhaolin Ren are supported by NSF CAREER: ECCS-1553407, NSF AI institute: 2112085 and ONR YIP: N00014-19-1-2217.

REFERENCES

Agarwal, A., Kakade, S.M., Lee, J.D., and Mahajan, G. (2020). On the theory of policy gradient methods: Optimality, approximation, and distribution shift.

Arslan, G. and Yüksel, S. (2016). Decentralized q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, 62(4), 1545–1558.

Bertrand, N., Markey, N., Sadhukhan, S., and Sankur, O. (2020). Dynamic network congestion games. *arXiv preprint arXiv:2009.13632*.

Bowling, M. and Veloso, M. (2001). Rational and convergent learning in stochastic games. In *International joint conference on artificial intelligence*, volume 17, 1021–1026. Citeseer.

Buşoniu, L., Babuška, R., and De Schutter, B. (2010). Multi-agent reinforcement learning: An overview. *Innovations in multi-agent systems and applications-1*, 183–221.

Chen, T., Zhang, K., Giannakis, G.B., and Başar, T. (2018). Communication-efficient policy gradient methods for distributed reinforcement learning. *arXiv preprint arXiv:1812.03239*.

Crawford, V.P. (1985). Learning behavior and mixed-strategy nash equilibria. *Journal of Economic Behavior & Organization*, 6(1), 69–78.

Foerster, J.N., Chen, R.Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2017). Learning with opponent-learning awareness.

González-Sánchez, D. and Hernández-Lerma, O. (2013). *Discrete–time stochastic control and dynamic potential games: the Euler–Equation approach*. Springer Science & Business Media.

Hu, J. and Wellman, M.P. (2003). Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov), 1039–1069.

Jordan, J.S. (1993). Three problems in learning mixed-strategy nash equilibria. *Games and Economic Behavior*, 5(3), 368–386.

Kakade, S.M. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In C. Sammut and A.G. Hoffmann (eds.), *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, 267–274. Morgan Kaufmann.

Krishna, V. and Sjöström, T. (1998). On the convergence of fictitious play. *Mathematics of Operations Research*, 23(2), 479–511.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D., and Graepel, T. (2017). A unified game-theoretic approach to multiagent reinforcement learning. *arXiv preprint arXiv:1711.00832*.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. (2021). Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*.

Li, Y., Tang, Y., Zhang, R., and Li, N. (2019). Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach.

Littman, M.L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, 157–163. Elsevier.

Macua, S.V., Zazo, J., and Zazo, S. (2018). Learning parametric closed-loop policies for markov potential games. *CoRR*, abs/1802.00899.

Marsden, J.E. and Tromba, A. (2003). *Vector calculus*. Macmillan.

Mazumdar, E., Ratliff, L.J., and Sastry, S.S. (2020). On gradient-based learning in continuous games. *SIAM Journal on Mathematics of Data Science*, 2(1), 103–131.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In H.D. III and A. Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6820–6829. PMLR.

Mguni, D. (2020). Stochastic potential games. *arXiv preprint arXiv:2005.13527*.

Monderer, D. and Shapley, L.S. (1996). Potential games. *Games and economic behavior*, 14(1), 124–143.

Panait, L. and Luke, S. (2005). Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems*, 11(3), 387–434.

Qu, G., Wierman, A., and Li, N. (2019). Scalable reinforcement learning of localized policies for multi-agent networked systems.

Qu, G., Wierman, A., and Li, N. (2020). Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Learning for Dynamics and Control*, 256–266. PMLR.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *ArXiv*, abs/1610.03295.

Shapley, L. (1964). Some topics in two-person games. *Advances in game theory*, 52, 1–29.

Shapley, L.S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, 39(10), 1095–1100.

Sutton, R.S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Tan, M. (1993). Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, 330–337.

Tesauro, G. (2003). Extending q-learning to general adaptive multi-agent systems. *Advances in neural information processing systems*, 16, 871–878.

Van Damme, E. (1991). *Stability and perfection of Nash equilibria*, volume 339. Springer.

Vidhate, D.A. and Kulkarni, P. (2017). Cooperative multi-agent reinforcement learning models (cmrlm) for intelligent traffic control. In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, 325–331. IEEE.

Wai, H.T., Yang, Z., Wang, Z., and Hong, M. (2018). Multi-agent reinforcement learning via double averaging primal-dual optimization. NIPS'18, 9672–9683. Curran Associates Inc., Red Hook, NY, USA.

Wang, W. and Carreira-Perpiñán, M.Á. (2013). Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *CoRR*, abs/1309.1541.

Xu, X., Jia, Y., Xu, Y., Xu, Z., Chai, S., and Lai, C.S. (2020). A multi-agent reinforcement learning-based data-driven method for home energy management. *IEEE Transactions on Smart Grid*, 11(4), 3201–3211.

Zhang, C. and Lesser, V. (2010). Multi-agent learning with policy prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24.

Zhang, K., Yang, Z., and Başar, T. (2019a). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*.

Zhang, K., Yang, Z., and Basar, T. (2019b). Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Basar, T. (2018). Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, 5872–5881. PMLR.

Zhang, R., Ren, Z., and Li, N. (2021). Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*.

Zhang, R., Ren, Z., and Li, N. (2022). Gradient play in stochastic games: Stationary points and local geometry (supplementary material). `https://drive.google.com/file/d/1quEJpEUGGvc5zqCdkbUBaqwTzG3saiRb/view`.