Gradient Play in Stochastic Games: Stationary Points, Convergence, and Sample Complexity

Runyu Zhang, Zhaolin Ren, and Na Li, Senior Member, IEEE

Abstract—In this article, we study the performance of the gradient play algorithm 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, and give a local convergence rate around strict NEs. Furthermore, for a subclass of SGs called Markov potential games (which includes the setting with identical rewards as an important special case), we design a sample-based reinforcement learning algorithm and give a nonasymptotic global convergence rate analysis for both exact gradient play and our sample-based learning algorithm. Our result shows that the number of iterations to reach an $\epsilon$-NE scales linearly, instead of exponentially, with the number of agents. Local geometry and local stability are also considered, where we prove that strict NEs are local maxima of the total potential function and fully mixed NEs are saddle points.

Index Terms—Multiagent systems, reinforcement learning.

I. INTRODUCTION

MULTIAGENT systems find applications in a wide range of societal systems, e.g., electric grids, traffic networks, smart buildings, and smart cities. Given the complexity of these systems, multiagent reinforcement learning (MARL) has gained increasing attention in recent years [1], [2]. 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 [3], [4], [5] have studied the performance of multiagent policy gradient algorithms, due to a lack of understanding of the optimization landscape in these multiagent 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 multiagent reinforcement learning (RL) setting.

In this article, we investigate this problem from a game-theoretic perspective. We model the multiagent system as a stochastic game (SG) where agents take independent stochastic policies and can have different reward functions. The study of SGs dates back to as early as the 1950s by Shapley [6] with a series of follow-up works on developing Nash equilibrium (NE)-seeking algorithms, especially in the RL setting (see, e.g., [7], [8], [9], and references therein). While well-known classical algorithms for solving SGs are mostly value based, such as Nash–Q learning [10], hyper-Q learning [11], and WoLF–PHC [12], gradient-based algorithms have also started to gain popularity in recent years due to their advantages as mentioned earlier [13], [14], [15].

In this work, we aim to gain a deeper understanding of the structure of first-order stationary points and the dynamical behavior 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) What is the stability of the individual NEs?
3) How do agents learn from samples in this environment?

These questions have already been widely discussed in other settings, e.g., one-shot (stateless) finite-action games [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], one-shot continuous games [26], zero-sum linear–quadratic (LQ) games [27], etc. There are both negative and positive results depending on the settings. For one-shot continuous games, Mazumdar et al. [26] proved a negative result suggesting that gradient flow has stationary points (even local maxima) that are not necessarily NEs. Conversely, Zhang et al. [27] 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: We consider the gradient play algorithm for the infinite-time-horizon discounted reward SGs with independent directly parameterized agents’ policies. Through generalizing the gradient domination property in [28] to the SG setting, we first establish the equivalence of first-order stationary policies and NEs (see Theorem 1). This result suggests that even if agents have an identical reward, the first-order stationary points are only equivalent to NEs, which are usually nonunique and have different reward values. This is fundamentally different from the centralized learning case [28], where it can be shown that the first-order stationary point is the unique global optimal solution.

Then, we study the convergence of gradient play for SGs. For general games, it is known that gradient play may fail to obtain global convergence [16], [17], [18], [19]. Thus, we first focus on characterizing some local properties for the general cases. In particular, we characterize the structure of strict NEs and show...
that gradient play locally converges to strict NEs within finite steps (see Theorem 2).

Next, we study a special class of SGs called Markov potential games (MPGs) [29], [30], [31], which includes identical reward multiagent RL [32], [33], [34], [35], [36], [37] as an important special case. Concurrently, this work and [31] have established the global convergence rate to an NE for gradient play under MPGs (see Theorem 3). However, the result does not specify which NE the policies converge to. Given the fact that there are many NEs that would have poor global value, global convergence results have a limited implication on the algorithm performance. This motivates us to study the local geometry around some specific types of NEs. 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 (see Theorem 4).

Then, we design a fully decentralized sample-based gradient play algorithm and prove that it can find an $\epsilon$-NE with high probability using $O\left(\frac{n\text{poly}(|S|, \max_i |A_i|)}{\epsilon^2}\right)$ samples (see Theorem 5; $|S|$ and $|A_i|$ denote the size of the state space and action space of agent $i$, respectively). The key enabler of our algorithm is the existence of an underlying averaged Markov decision process (MDP) for each agent when other agents’ policies are fixed. Our learning method can be viewed as a model-based policy evaluation method with respect to agents’ averaged MPDs. This averaged MDP concept could be applied to design many other MARL algorithms, especially policy-evaluation-based methods.

**Comparison to other works on NE learning for SGs:** There are some recent studies on general SGs with finite state–action spaces. However, either the structure of SGs or the methods they consider are different from our setting. For example, the authors of [39] and [40] consider learning correlated equilibria rather than NEs for finite-time-horizon general-sum games; the authors of [41] and [42] propose decentralized learning algorithms for the weakly acyclic games, which include the identical game as a special case, but only asymptotic convergence is considered; and the authors of [43] and [44] consider convergence to NE for two-player zero-sum games. In addition, the authors of [4], [5], and [45] consider slightly different MAML settings, where agents collaboratively maximize the summation of agents’ reward with either full or partial state observation. They also require communication between neighboring agents for a better global coordination.

For the MPG subclass, the authors of [31], [38], [39], [46], and [47] study the convergence to an NE. Note that [38] is the arXiv version of this article. Song et al. [39] design the Nash Coordinate Ascent algorithm, which requires agents to update sequentially and does not belong to the gradient-based algorithm class. While the authors of [46], [47], and [48] consider gradient-based algorithms, they study softmax policies or policies with function approximation, which are different from directly parameterized policies. The work in [31] is the most related work to this article, which studies the performance of gradient-based algorithms under direct parameterization. It studies the global convergence rate and develop sample complexity results for gradient play, but do not study the local geometry for general SGs. In addition, the sample-based algorithm considered in [31] is based on Monte Carlo gradient estimation, which might suffer from high variance in real implementation and is very different from our algorithm that estimates the gradient via estimating the "model" with respect to agents’ averaged MPDs. Moreover, our concept of "averaged" MPDs could also serve as a useful tool for the design and analysis of other MARL algorithms.

Finally, we remark here that some of the preliminary results have been published in the conference version [49] of this work. Compared to [49], this article includes the following additional materials: 1) the global convergence result for MPG; 2) the sample-based learning methods; 3) examples and detailed discussions on MPG; 4) all the proofs; and 5) more numerical studies and discussions. Due to space limit, some auxiliary proofs are deferred to the arXiv version [38] of this article.

## II. Problem Setting and Preliminaries

We consider an SG $\mathcal{G} = (N, \mathcal{S}, \mathcal{A} = A_1 \times \cdots \times A_n, P, r = (r_1, \ldots, r_n), \gamma, \rho)$ with $n$ agents that is specified by an agent set $N = \{1, 2, \ldots, n\}$, a finite state space $\mathcal{S}$, a finite action space $A_i$ 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 $a_i \in A_i$ is the action of agent $i$, agent $i$’s reward function $r_i : \mathcal{S} \times \mathcal{A} \rightarrow (0, 1]$, a discount factor $\gamma \in [0, 1]$, and an initial state distribution $\rho$ over $\mathcal{S}$.

A stochastic policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ (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., $\Pr(a_i|s_i) = \pi(a_i|s_i)$. A decentralized stochastic policy is a special subclass of stochastic policies, with $\pi = \pi_1 \times \cdots \times \pi_n$, where $\pi_i : \mathcal{S} \rightarrow \Delta(\mathcal{A}_i)$. For decentralized stochastic policies, each agent takes its action based on the current state $s$ independently of other agents’ choices of actions, i.e., $\Pr(a_i|s_i) = \pi(a_i|s_i) = \prod_{s_i \in \mathcal{S}} \pi_i(a_i|s_i), a_i = (a_{i,1}, \ldots, a_{i,n})$.

For notational simplicity, we define $\pi_i(a_i|s) := \prod_{s_i \in \mathcal{S}} \pi_i(a_i|s_i)$, where $I \subseteq N$ is an index set. Furthermore, we use the notation $-i$ to denote the indexed set $N\setminus\{i\}$.

We consider direct decentralized policy parameterization, where agent $i$’s policy is parameterized by $\theta_i$:

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

For notational simplicity, we abbreviate $\pi_{i,\theta_i}(a_i|s)$ as $\pi_{\theta_i}(a_i|s)$, and $\theta_i(a_i|s_i)$ as $\theta_{s_i,a_i}$. Here, $\theta_i \in \Delta(\mathcal{A}_i)$, i.e., $\theta_i$ is subject to the constraints $\theta_{s,a} \geq 0$ and $\sum_{s_i \in \mathcal{S}} \theta_{s_i,a_i} = 1$ for all $s \in \mathcal{S}$. The joint global policy is given by $\pi_{\theta_i}(a|s) = \prod_{i=1}^{n} \pi_{\theta_i}(a_i|s_i) = \prod_{i=1}^{n} \theta_{s_i,a_i}$. We use $X_i := \Delta(\mathcal{A}_i)$, $X := X_1 \times \cdots \times X_n$ to denote the feasible region of $\theta_i$ and $\theta$.

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

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

where the expectation is with respect to the random trajectory $\tau = (s_0, a_1, r_1, \ldots)$, where $a_t \sim \pi_{\theta_i}(\cdot|s_t), s_{t+1} = P(\cdot|s_t, a_t)$.

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, NE is often used to characterize the performance of agents’ policies.

**Definition 1 (Nash equilibrium; cf. [51], [52]):** A policy $\pi^* = (\theta_1^*, \ldots, \theta_n^*)$ is called an NE if

$$J_i(\theta_i^*, \theta_i^*_{-i}) \geq J_i(\theta_i', \theta_i^*_{-i}) \quad \forall \theta_i' \in X_i, \quad i \in N.$$
The equilibrium is called a strict NE if the inequality holds strictly for all \( \theta^* \in \mathcal{X} \) and \( i \in N \). The equilibrium is called a pure NE if \( \theta^* \) corresponds to a deterministic policy. The equilibrium is called a mixed NE if it is not pure. Furthermore, the equilibrium is called a fully mixed NE if every entry of \( \theta^* \) is strictly positive, i.e., \( \theta^*_{-a_i} > 0 \forall a_i \in \mathcal{A}_i \forall s \in \mathcal{S}, i \in N \).

We define the discounted state visitation distribution \([50] \) \( d_\theta \) of a policy \( \pi_\theta \) given an initial state distribution \( \rho \) as

\[
d_\theta(s) := \mathbb{E}_{s_0 \sim \rho}(1 - \gamma)^i \sum_{i=0}^\infty \gamma^i \mathbb{P}_{\theta}^{i}(s_i = s|s_0) \tag{2}\]

where \( \mathbb{P}_{\theta}^{i}(s_i = s|s_0) \) is the state visitation probability that \( s_i = s \) when executing \( \pi_\theta \) starting at state \( s_0 \) at time step \( i \).

Throughout this article, we make the following assumption on the SGs we study.

**Assumption 1**: The SG \( \mathcal{M} \) 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 [28], [31], [44], [53]. Note that this assumption could be easily satisfied if the initial distribution \( \rho \) satisfies \( \rho(s) > 0 \forall s \in \mathcal{S} \).

Similar to centralized RL [50], define agent \( i \)'s \( Q \)-function \( Q_i^\theta \) and its advantage function \( A_i^\theta \) as

\[
Q_i^\theta(s, a) := \mathbb{E} \left[ \sum_{i=0}^\infty \gamma^i r_i(s_i, a_i) \mid \pi_\theta, s_0 = s, a_0 = a \right] \tag{6}
\]

\[
A_i^\theta(s, a) := Q_i^\theta(s, a) - V_i^\theta(s). \tag{7}
\]

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

\[
\overline{Q}_i^\theta(s, a_i) := \sum_{a_{-i}} \pi_{\theta_{-i}}(a_{-i}|s) Q_i^\theta(s, a_i, a_{-i}) \tag{8}
\]

\[
\overline{A}_i^\theta(s, a_i) := \sum_{a_{-i}} \pi_{\theta_{-i}}(a_{-i}|s) A_i^\theta(s, a_i, a_{-i}). \tag{9}
\]

Similarly, we define agent \( i \)'s “averaged” transition probability distribution \( \overline{P}_i : S \times S \times \mathcal{A} \to \mathbb{R} \) and “averaged” reward \( \overline{r}_i : S \times \mathcal{A} \to \mathbb{R} \) as

\[
\overline{P}_i(s'|s, a_i) := \sum_{a_{-i}} \pi_{\theta_{-i}}(a_{-i}|s) P_i(s'|s, a_i, a_{-i}) \tag{10}
\]

\[
\overline{r}_i(s'|s, a_i) := \sum_{a_{-i}} \pi_{\theta_{-i}}(a_{-i}|s) r_i(s'|s, a_i, a_{-i}). \tag{11}
\]

From its definition, the averaged \( Q \)-function satisfies the following Bellman equation.

**Lemma 1**: \( \overline{Q}_i^\theta \) satisfies

\[
\overline{Q}_i^\theta(s, a_i) = \overline{r}_i(s, a_i) + \gamma \sum_{s', a'_i} \pi_{\theta_{-i}}(a'_i|s') \overline{P}_i(s'|s, a_i) \overline{Q}_i^\theta(s', a'_i). \tag{12}
\]

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 \( \overline{P}_i, \overline{r}_i \) as its stage reward and transition probability, respectively. We define this MDP as the “averaged” MDP of agent \( i \), i.e., \( \mathcal{M}_i^\theta = (S, \mathcal{A}_i, \overline{P}_i, \overline{r}_i, \gamma, \rho) \). The notion of an “averaged” MDP will serve as an important intuition when designing the sample-based algorithm. Note that the “averaged” MDP is only well defined when the policies of the other agents \( \theta_{-i} \) 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. Thus, various classical policy evaluation RL algorithms can then be applied. In addition, we can apply the performance difference lemma [54] to the averaged MDP to derive a corresponding lemma for SGs, which is useful throughout this article.

**Lemma 2** (Performance difference lemma, for SGs; for proof, see Appendix A): Let \( \theta' = (\theta'_1, \ldots, \theta'_n) \)

\[
J_i(\theta'_i, \theta_{-i}) - J_i(\theta_i, \theta_{-i}) = \frac{1}{1 - \gamma} \sum_{s, a_i} d_\theta(s) \pi_{\theta_i}(a_i|s) \overline{A}_i^\theta(s, a_i). \tag{13}
\]

Note that in the single-agent case \( (n = 1) \), Lemma 2 is the same as the original performance difference lemma known in literature, e.g., [54, Lemma 6.1].

### III. GRADIENT PLAY FOR GENERAL SGs

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

\[
\theta_i(t+1) = \text{Proj}_{\mathcal{X}_i} \left( \theta_i(t) + \eta \nabla_{\theta_i} J_i \left( \theta_i(t) \right) \right), \quad \eta > 0. \tag{14}
\]

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 follows.

**Definition 2** (First-order stationary policy): A policy \( \theta^* = (\theta^*_1, \ldots, \theta^*_n) \) is called a first-order stationary policy if \( \theta^*_i - \theta^*_{-i} \) \( \nabla_{\theta_i} J_i(\theta^*) \leq 0 \forall \theta^*_i \in \mathcal{X}_i, i \in N \).

It is not hard to verify that \( \theta^* \) is a first-order stationary policy if and only if it is a fixed point under gradient play (5). 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 \( \theta^*_i \) is optimal among all possible \( \theta_i \) given \( \theta_{-i} \). 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 \( \nabla_{\theta_i} J_i \).

The policy gradient theorem [55] gives an efficient formula for the gradient, as follows:

\[
\nabla_{\theta_i} J_i(\theta) = \frac{1}{\gamma} \mathbb{E}_{s_0 \sim \rho, \theta^\theta} \left[ \log \pi_{\theta}(a|s) Q_i^\theta(s, a) \right] \tag{15}
\]

Applying (6), the gradient \( \nabla_{\theta_i} J_i \) can be written explicitly as follows:

**Lemma 3** (See [38, Appendix D] for proof): For direct distributed parameterization (1)

\[
\frac{\partial J_i(\theta)}{\partial \theta_{s, a_i}} = \frac{1}{1 - \gamma} d_\theta(s) \overline{Q}_i^\theta(s, a_i). \tag{16}
\]

### A. Gradient Dominance and the Equivalence Between NE and First-Order Stationary Policy

[28, Lemma 4.1] established gradient domination for centralized tabular MDP under direct parameterization. We can show that a similar property still holds for SGs.

**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 \( \theta^*_i \in \mathcal{X}_i, i \in N \):

\[
J_i(\theta_i^*, \theta_{-i}) - J_i(\theta_i, \theta_{-i}) \leq \left\| \frac{d \theta_i}{d \theta} \right\|_{\infty, \theta \in \mathcal{X}_i} \max_{\theta_i \in \mathcal{X}_i} (\theta_i - \theta_{-i}) \nabla_{\theta_i} J_i(\theta) \tag{17}
\]

where \( \left\| \frac{d \theta_i}{d \theta} \right\|_{\infty} := \max_{\theta \in \mathcal{X}_i} \frac{d \theta_i}{d \theta}(\theta) \) and \( \theta^* = (\theta^*_1, \ldots, \theta^*_n) \).

**Proof**: According to Lemma 2, we have

\[
J_i(\theta_i^*, \theta_{-i}) - J_i(\theta_i, \theta_{-i}) = \frac{1}{1 - \gamma} \sum_{s, a_i} d_\theta(s) \pi_{\theta_i}(a_i|s) \overline{A}_i^\theta(s, a_i). \tag{18}
\]
From the definition of "averaged" advantage function, we have
\[ \sum_{a_i} \pi_{\theta_i}(a_i|s) \overline{A_i}(s, a_i) = 0 \quad \forall s \in S \]
which implies that \( \max_{a_i \in A_i} \overline{A_i}(s, a_i) \geq 0 \); thus, we have
\[ J_i(\theta_i', \theta_i) - J_i(\theta_i, \theta_i) = \frac{1}{1 - \gamma} \sum_s d_{\theta_i}(s) \pi_{\theta_i}(a_i|s) \overline{A_i}(s, a_i) \]
\[ \leq \frac{1}{1 - \gamma} \sum_s d_{\theta_i}(s) \pi_{\theta_i}(a_i|s) \overline{A_i}(s, a_i) \]
\[ \leq \frac{1}{1 - \gamma} \frac{1}{\max_{a_i \in A_i} \overline{A_i}(s, a_i)} \sum_s d_{\theta_i}(s) \max_{a_i \in A_i} \overline{A_i}(s, a_i). \]
(10)

We can rewrite \( \frac{1}{1 - \gamma} \sum_s d_{\theta_i}(s) \max_{a_i \in A_i} \overline{A_i}(s, a_i) \) as
\[ \frac{1}{1 - \gamma} \sum_s d_{\theta_i}(s) \max_{a_i \in A_i} \overline{A_i}(s, a_i) \]
\[ = \frac{1}{1 - \gamma} \max_{\theta_i \in \Theta} \sum_s d_{\theta_i}(s) \pi_{\theta_i}(a_i|s) \overline{A_i}(s, a_i) \]
\[ = \max_{\theta_i \in \Theta} \sum_s (\pi_{\theta_i}(a_i|s) - \pi_{\theta_i}(a_i|s)) \frac{1}{1 - \gamma} d_{\theta_i}(s) \overline{A_i}(s, a_i) \]
\[ = \max_{\theta_i \in \Theta} \sum_s (\pi_{\theta_i}(a_i|s) - \pi_{\theta_i}(a_i|s)) \frac{1}{1 - \gamma} d_{\theta_i}(s) \overline{A_i}(s, a_i) \]
\[ \left( \frac{1}{1 - \gamma} \sum_s d_{\theta_i}(s) V(s) \sum_{a_i} (\pi_{\theta_i}(a_i|s) - \pi_{\theta_i}(a_i|s)) \right) \]
\[ = \max_{\theta_i \in \Theta} \sum_s (\pi_{\theta_i}(a_i|s) - \pi_{\theta_i}(a_i|s)) \frac{1}{1 - \gamma} d_{\theta_i}(s) \overline{A_i}(s, a_i) \]
\[ = \max_{\theta_i \in \Theta} (\overline{A}_i - \overline{A}_i)^{\top} \nabla_{\theta_i} J_i(\theta_i). \]
(11)

Substituting this into (10), we may conclude that
\[ J_i(\theta_i', \theta_i) - J_i(\theta_i, \theta_i) \leq \left\| \frac{d_{\theta_i}(s)}{d_{\theta_i}(s)} \right\| \max_{\theta_i \in \Theta} (\overline{A}_i - \overline{A}_i)^{\top} \nabla_{\theta_i} J_i(\theta_i) \]
and this completes the proof.

For the single-agent case \((n = 1)\), (8) is consistent with the result in [28], i.e., \( J(\theta') - J(\theta) \leq \left\| \frac{d_{\theta}}{d_{\theta}} \right\| \max_{\theta_i \in X_i} (\overline{A}_i - \overline{A}_i)^{\top} \nabla_{\theta_i} J_i(\theta_i) \). However, when there are multiple agents, the condition is much weaker because the inequality requires \( \theta_i \) to be fixed. When \( n = 1 \), gradient domination rules out the existence of stationary points that are not global optima. For the multiagent 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 following theorem.

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

**Proof:** The definition of an NE naturally implies first-order stationarity, because for any \( \theta_i \in \mathcal{X}_i \):
\[ J_i((1 - \eta)\theta_i' + \eta\theta_i, \theta_i', \theta_{i-1}) - J_i(\theta_i', \theta_i') \leq \eta(\theta_i - \theta_i')^{\top} \nabla_{\theta_i} J_i(\theta_i') + o(\eta) \leq 0, \quad \forall \eta > 0 \]
Letting \( \eta \to 0 \) gives the first-order stationary condition:
\[ (\theta_i - \theta_i')^{\top} \nabla_{\theta_i} J_i(\theta_i') \leq 0, \quad \forall \theta_i \in \mathcal{X}_i \]
It remains to be shown that all first-order stationary policies are NEs. From Assumption 1, we know that for any pair of parameters \( \theta', \theta^* \), \( \|d_{\theta'}(s)\| \leq \infty \).

Take \( \theta' = (\theta_i', \theta_{i-1}) \), \( \theta^* = (\theta_i^*, \theta_{i-1}) \). According to Lemma 4, we have that for any first-order stationary policy \( \theta_i' \), we have
\[ J_i(\theta_i', \theta_{i-1}) - J_i(\theta_i^*, \theta_{i-1}) \leq \left\| \frac{d_{\theta_i}(s)}{d_{\theta_i}(s)} \right\| \max_{\theta_i \in \Theta} (\overline{A}_i - \overline{A}_i)^{\top} \nabla_{\theta_i} J_i(\theta_i') \leq 0 \]
which completes the proof.

We briefly note here that the equivalence between the first-order stationary points and NEs holds for all SGs that satisfy Assumption 1. One implication from the theorem is that for identical interest case where agents have the same rewards, we can only ensure the first-order stationary points to be NEs when the policies are decentralized policies. Note that NEs are often nonunique and often with different objective values. This is in contrast to the single-agent/centralized case where the first-order stationary point is equivalent to the global optimal point [28].

**B. 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 [16], [17], [18], [19]. One major difficulty is that the vector field \( \nabla_{\theta_i} J_i(\theta_i) = 0 \) is not a conservative vector field. 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 an SG \( M \), any strict NE \( \theta_i^* \) is pure, meaning that for each \( i \) and \( s \), there exist one \( a_i^*(s) \) such that \( \theta_i, a_i = \{a_i = a_i^*(s)\} \). In addition, we have
\[ \begin{align*}
\text{(i)} & \quad a_i^*(s) = \text{arg max}_{a_i} \overline{A}_i(s, a_i) \\
\text{(ii)} & \quad \overline{A}_i^0(s, a_i) = 0 \\
\text{(iii)} & \quad \overline{A}_i^0(s, a_i) < 0 \quad \forall a_i \neq a_i^*(s).
\end{align*} \]

Based on this lemma, we define the following for studying the local convergence of a strict NE \( \theta^* \):
\[ \Delta_i^0(s) := \min_{a_i \neq a_i^*(s)} \overline{A}_i^0(s, a_i) \]
\[ \Delta_i^0 := \min_{s} \frac{1}{\gamma} \sum_s d_i(s) \Delta_i^0(s) > 0. \]
(12)

**Theorem 2** (Local finite-time convergence around strict NE): Define the metric of policy parameters as \( D(\theta||\theta') := \max_{1 \leq s \leq n} \max_{s \in S} \|\theta_i - \theta_i'\|_1 \), where \( \|\cdot\|_1 \) denotes the \( \ell_1 \)-norm. Suppose that \( \theta^* \) is a strict NE; then, for any \( \theta(0) \) such that \( D(\theta(0)||\theta^*) \leq \frac{\Delta_i^0(1 - \gamma)^3}{8n(2n+1)\|\theta^*\|^6} \), running gradient play (5) will guarantee the convergence of \( D(\theta(t)||\theta^*) \leq \min\{D(\theta(t)||\theta^*) - \Delta_i^0, 0\} \), which means that gradient play is going to converge within \( \lceil 2D(\theta(0)||\theta^*) \rceil \) steps.

Proofs are deferred to Appendix B.

**Remark 1:** Note that the local convergence in Theorem 2 only requires a finite number of steps. The key insight of the proof is that the gradient always points toward \( \theta^* \), and that the algorithm projects the gradient update onto the probability simplex; thus, by picking the step size \( \eta \) arbitrarily large, exact convergence can be achieved by just one step. However, the caveat is that
we need to assume that the initial policy is sufficiently close to \( \theta^* \). For numerical stability considerations, one should pick reasonable step sizes to run the algorithm to accommodate random initializations. Theorem 2 also shows that the region of region of attraction for strict NEs is at least \( \frac{\Delta^o}{\sum_{i=1}^{n} |A_i|} \), and thus, \( \theta^* \) with a larger \( \Delta^o \), 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.

IV. GRADIENT PLAY FOR MPGS

We have discussed that the main problem for the global convergence of gradient play for general SGs is that the vector field \( \{\nabla_{\theta} J_i(\theta)\}_{i=1}^{n} \) is not conservative. Thus, in this section, we restrict our analysis to a special subclass where the vector field is conservative, which in turn enjoys global convergence. This subclass is generally referred to as an MPG in the literature.

Definition 3 (Markov potential game [31]): An SG \( M \) is called an MPG if there exists a potential function \( \phi: S \times A_1 \times \cdots \times A_n \rightarrow \mathbb{R} \) such that for any agent \( i \) and any pair of policy parameters \( (\theta_i^t, \theta_{-i}^t), (\theta_i^t, \theta_{-i}^t) \), we have

\[
E \sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) | \pi = (\theta_i^t, \theta_{-i}^t), s_0 = s \]

\[
- E \sum_{t=0}^{\infty} \gamma^t r_i(s_t, a_t) | \pi = (\theta_i^t, \theta_{-i}^t), s_0 = s \]

\[
= E \sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | \pi = (\theta_i^t, \theta_{-i}^t), s_0 = s \]

\[
- E \sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | \pi = (\theta_i^t, \theta_{-i}^t), s_0 = s \quad \forall s.
\]

As shown in the definition, the condition of an MPG is admittedly rather strong and difficult to verify for general SGs. The authors of [29] and [30] found that continuous MPGs can model applications such as the great fish war [56], the stochastic lake game [57], medium access control [30], etc. There are also efforts attempting to identify conditions such that an SG is an MPG [30, 31, 58]. In [38, Appendix A], we provide a more detailed discussion on MPGs, including a necessary condition of MPG, counterexamples of stagewise potential games that are not MPG, sufficient conditions for a SG to be an MPG, and application examples of MPG. Nevertheless, identifying sufficient and necessary conditions and broadening the applications of MPG are important future directions.

Given a policy \( \theta \), we define the "total potential function" \( ^2 \phi(\theta) := E_{\theta \sim \pi(\theta)} \left[ \sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | \pi_{\theta} \right] \) for an MPG. The following proposition guarantees that an MPG has at least one NE, and it is a pure NE (for proof, see [38, Appendix F]). We also define the quantity \( \Phi_{\max}, \Phi_{\min} \) as \( \Phi_{\max} := \frac{\Phi_{\max}}{1-\gamma} \), \( \Phi_{\min} := \frac{\Phi_{\min}}{1-\gamma} \), where \( \Phi_{\max} := \max_{a,s} \phi(s, a) \) and \( \Phi_{\min} := \min_{a,s} \phi(s, a) \). It can be easily verified that \( \Phi_{\min} \leq \Phi(\theta) \leq \Phi_{\max} \) for all \( \theta \).

Proposition 1: For an MPG, there is at least one global maximum \( \theta^* \) of the total potential function \( \Phi \), i.e., \( \theta^* \epsilon \arg \max_{\theta \in \mathcal{X}} \Phi(\theta) \) is a pure NE.

From the definition of the total potential function, we obtain the following relationship:

\[
J_i(\theta_i^t, \theta_{-i}^t) - J_i(\theta_i^t, \theta_{-i}^t) = \Phi(\theta_i^t, \theta_{-i}^t) - \Phi(\theta_i^t, \theta_{-i}^t).
\]

Thus

\[
\nabla_{\theta} J_i(\theta) = \nabla_{\theta} \Phi(\theta)
\]

which means that gradient play (5) is equivalent to running projected gradient ascent with respect to the total potential function \( \Phi \), i.e.,

\[
\theta^{t+1} = \text{Proj}_X(\theta^{t}) + \eta \nabla_{\theta} \Phi(\theta^{t}), \quad \eta > 0.
\]

A. Global Convergence

With the above property, we can establish the global convergence for gradient play to an \( \epsilon \)-NE for MPG, which was also established in [31] around the same time as the first arXiv version [38] of this article. For the sake of self-completeness, we include the theorem here. Before that, we define \( \epsilon \)-NE as follows.

Definition 4 (\( \epsilon \)-Nash equilibrium): Define the "NE-gap" of a policy \( \theta \) as

\[
\text{NE} - \text{NE}(\theta) := \max_{\theta_t^i, \theta_{-i}^t} \left[ J_i(\theta_i^t, \theta_{-i}^t) - J_i(\theta_i^t, \theta_{-i}^t) \right] - \Phi(\theta) \geq \epsilon.
\]

A policy \( \theta \) is an \( \epsilon \)-NE if \( \text{NE} - \text{NE}(\theta) \leq \epsilon \).

Theorem 3: Suppose that for all \( \theta \in \mathcal{X} \), with step size \( \eta \leq \frac{(1-\gamma)^2}{\Delta^o |\sum_{i=1}^{n} |A_i|} \), the NE asymptotically converges to 0 under gradient play (5), i.e., \( \lim_{t \rightarrow \infty} \text{NE} - \text{NE}(\theta(t)) = 0 \). Furthermore, we have

\[
\frac{1}{T} \sum_{1 \leq t \leq T} \text{NE} - \text{NE}(\theta(t))^2 \leq \epsilon^2
\]

whenever

\[
T \geq \frac{64 \Phi_{\max} - \Phi_{\min}}{\Delta^o} |\sum_{i=1}^{n} |A_i|} (1-\gamma)^2 \epsilon^2
\]

where \( \Phi := \max_{\theta, \theta \in \mathcal{X}} \| \frac{d \Phi}{d \theta} \|_{\infty} \) (by Assumption 1, we know that this quantity is well defined).

The factor \( M \) is also known as the distribution mismatch coefficient that characterizes how the state visitation varies with the policies. Given an initial state distribution \( \rho \) that has positive measure on every state, \( M \) can be at least bounded by

\[
M \leq \frac{1}{2} \max_{\theta, \theta \in \mathcal{X}} \| \frac{d \Phi}{d \theta} \|_{\infty} \leq \frac{1}{2} \left( \frac{1}{\min_{\theta} \rho(\theta)} \right). \]

The proof structure of Theorem 3 resembles the proof of convergence for single-agent MDPs in [28], where they leverage classical nonconvex optimization results [59, 60] and gradient domination to get the

\[
\text{NE} - \text{NE}(\theta(t))^2 \leq 3 \epsilon^2
\]

Note that our definition of MPG is slightly stronger than the definition in [31] as it requires the total potential function to take the particular form as a discounted sum of the potential function, i.e., \( \Phi(\theta) := E_{\theta \sim \pi(\theta)} \left[ \sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | \pi_{\theta} \right] \).

However, most of our results (see Theorems 3 and 5) still hold under the weaker definition in [31].

Theorem 4 is the only result that relies on the slightly stronger version of the definition.
convergence rate of $O\left(\frac{6\|\gamma\| \|\Delta\|^2 \|\gamma\|S\|A\|}{(1-\gamma)^3}\right)$ to the global optimum (see [38, Appendix G] for proof details). In fact, our result matches this bound when there is only one agent (the exponential factor on $(1-\gamma)$ looks slightly different because some factors are hidden implicitly in $M$ and $(\Phi_{\text{max}} - \Phi_{\text{min}})$ in our bound).

### B. Local Geometry of NEs

Theorem 3 suggests that gradient play is guaranteed to converge to an 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 I. 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 actions 1 and 2 randomly with probability $\gamma/2$. This NE is only of reward 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 and the local geometry around the NE, as a preliminary study, we characterize the local geometry and landscape for strict NEs and fully mixed NEs (stated in the following theorem). More future investigation is needed for nonstrict non-fully mixed NEs.

**Theorem 4:** For an MPG with $\Phi_{\text{min}} < \Phi_{\text{max}}$ (i.e., $\Phi$ is not a constant function), we have the following.

1. A strict NE $\theta^*$ is equivalent to a strict local maximum of the total potential function $\Phi$, i.e., $\exists \delta$ such that for all $\theta \in \mathcal{X}$, $\theta \neq \theta^*$ that satisfies $\|\theta - \theta^*\| \leq \delta$, we have $\Phi(\theta) < \Phi(\theta^*)$.

2. Any fully mixed NE $\theta^*$ is a saddle point of the total potential function $\Phi$, i.e., $\|\nabla \Phi(\theta^*)\| = 0$, and $\forall \delta > 0$, $\exists \theta \in \mathcal{X}$, such that $\theta - \theta^* \leq \delta$ and $\Phi(\theta) > \Phi(\theta^*)$.

The full proof of Theorem 4 is deferred to Appendix C.

**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 toward 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 [26].

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.

### C. Sample-Based Learning: Algorithm and Sample Complexity

In this section, we no longer assume access to the exact gradient, but instead need to estimate it via samples. Throughout the section, we make the following additional assumption on MPGs.

**Assumption 2** $(\tau, \sigma)\text{-sufficient exploration on states): There exist a positive integer $\tau$ and a $\sigma \in (0, 1)$ such that for any policy $\theta$ and any initial state–action pair $(s, a_i)$, we have $\Pr_t^{\theta}(s_i|s_0 = s, a_0 = a) \geq \sigma \forall s, a_i$ i.e., it poses a condition on the mixing time of the Markov chain induced by any policy $\theta$; there exists a sufficiently long time $\tau$, so the probability of being at any state at time $\tau$ is at least $\sigma$ for any initial state and action pair.

Note that similar assumptions are common in proving finite-time convergence of RL algorithms [5, 61, 62], where ergodicity of the Markov chain induced by certain policies is generally assumed, which results in every state being visited with positive probability in the stationary distribution.

We further introduce the *state transition probability under $\theta$* $P^\theta_S : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$ as $P^\theta_S(s'|s) := \sum_a \pi_{\theta}(a|s)P(s'|s,a)$.

We consider fully decentralized learning, where agent $i$’s observation only includes state $s_i$, its own action $a_i$, and its own reward $r_{i,t} := r_i(s_i, a_i)$ at time $t$. Such fully decentralized learning is plausible due to the fact that when $\theta_i$ is fixed, agent $i$ can be treated as an independent learner with the underlying MD being the “averaged” MD described in Section II. With this key observation, we design a “model-based” on-policy learning algorithm, where agents perform policy evaluation in the inner loop and gradient ascent at the outer loop. The algorithm is provided in Algorithm 1. Roughly, it consists of three main steps.

1. **(Inner loop)** Estimate the averaged transition probability and reward using on-policy samples $P^\theta_i$, $\overline{Q}^\theta_i$, and $P^\theta_S$.
2. **(Inner loop)** Calculate averaged $Q$-function $\overline{Q}^\theta_i$ and discounted state visitation distribution $d_\theta$ and compute the estimated gradient accordingly.
3. **(Outer loop)** Run projected gradient ascent with estimated gradients.

Before discussing our algorithm in more detail, we highlight that the idea of using the “averaged” MD can be used to design other learning methods including model-free methods, e.g., using the temporal difference methods to perform policy evaluation. One caveat is that the “averaged” MD is only well defined when all the other agents use fixed policies. This makes it difficult to extend the two-timescale framework (i.e., with an inner loop and outer loop) to single-timescale settings, which is an interesting future direction. Furthermore, note that the current algorithm requires full state observation; it remains an intriguing open question to extend it to the case with only partial observability. We would also like to point out that the algorithm initialization still requires extra consensus/coordination among the players to agree on the hyperparameters $T_j, T_{\nu}, C_{\nu}$, etc., which guarantees that agents go through the same equal-length phases to sample the trajectories, and compute gradient estimates.

**Step 1.** **Empirical estimation of $P^\theta_i, \overline{Q}^\theta_i$, and $P^\theta_S$:** Given a sequence $\{s_t, a_{i,t}, r_{i,t}, T_{i,t}\}_{t=0}^{T_{i,t}}$ generated by a policy $\theta := (\theta_i, \theta_{-i})$, the empirical estimation $\hat{P}^\theta_i$ of $P^\theta_i$ is given by

$$\hat{P}_i^\theta(s'|s, a_i) := \frac{\sum_{t=0}^{T_{i,t}} 1\{s_{t+1} = s', a_{i,t+1} = a_i\} \delta_i(s_{t+1}, a_{i,t+1})}{\sum_{t=1}^{T_{i,t}} 1\{s_{t+1} = s, a_{i,t+1} = a_i\}}$$

for $\sum_{t=1}^{T_{i,t}} 1\{s_{t+1} = s, a_{i,t+1} = a_i\} \geq 1$

for $\sum_{t=1}^{T_{i,t}} 1\{s_{t+1} = s, a_{i,t+1} = a_i\} = 0$.

(16)
Algorithm 1: Sample-Based Learning

Require: learning rate $\eta$, greedy parameter $\alpha$, sample trajectory length $T_j$, total iteration steps $T_G$
For each agent $i$
for $k = 0, 1, \ldots, T_G - 1$ do
for $t = 0, 1, \ldots, T_j$ do
Sample $s_0 \sim \rho$, implement policy $\theta(k)$ and collect trajectory $D^i_{t,k}$:
$$D^i_{t,k} = D^i_{t,k} \cup \{s_t, a_t, r_t, i_t\}, a_t \sim \pi_{\theta(s)}(\cdot | s_t)$$
end for
Estimate $\hat{Q}_i^\theta, \hat{d}_\theta$ by (16), (17), (18), respectively.
Calculate $\hat{Q}_i^\alpha, \hat{d}_\theta$ by (19), (20), respectively.
Estimate the gradient by (21):
Run projected gradient ascent as in (22)
end for

Here, we separately treat the special case where the state and action pair is not visited throughout the whole trajectory, i.e.,
$$\sum_{t=1}^{T_j} 1\{s_t = a_t, a_t = a_i\} = 0$$
make $P_i^\theta$ well defined.
Similarly, the estimates $\hat{r}_i^\theta, \hat{P}_i^\theta, \hat{\rho}_i^\theta, \hat{d}_\theta$ are given by
$$\hat{r}_i^\theta(s, a_i) := \left\{ \begin{array}{ll} \sum_{t=1}^{T_j} 1\{s_t = s, a_t = a_i\}, & \sum_{t=1}^{T_j} 1\{s_t = s, a_t = a_i\} \geq 1 \\ 0, & \text{otherwise} \end{array} \right. \tag{17}$$
$$\hat{P}_i^\theta(s'|s) := \left\{ \begin{array}{ll} \sum_{t=1}^{T_j} 1\{s_{t+1} = s', a_{t+1} = a\}, & \sum_{t=1}^{T_j} 1\{s_{t+1} = s', a_{t+1} = a\} \geq 1 \\ 1\{s' = s\}, & \sum_{t=1}^{T_j} 1\{s_t = s\} = 0 \end{array} \right. \tag{18}$$

Step 2. Estimation of $\hat{Q}_i^\theta$ and $\hat{d}_\theta$. We slightly abuse notation and use $\hat{Q}_i^\theta, \hat{r}_i^\theta \in \mathbb{R}^{|S||A_i|}$ to also denote the vectors corresponding to the averaged Q-function and reward function of agent $i$. Similarly, $\rho, \hat{d}_\theta \in \mathbb{R}^{|S_i|}$ are used to denote the vectors for $\rho(s)$ and $\hat{d}_\theta(s)$. Define $M_i^\theta \in \mathbb{R}^{|A_i||S|\times |S_i|}$,
$$M_i^\theta(s, a_i) := \pi_{\theta_i}(a_i | s) \hat{P}_i^\theta(s'|s, a_i).$$

Then, from Lemma 1, $\hat{Q}_i^\theta$ is given by
$$(I - \gamma M_i^\theta)\hat{Q}_i^\theta = \hat{r}_i^\theta \implies \hat{Q}_i^\theta = (I - \gamma M_i^\theta)^{-1}\hat{r}_i^\theta. \tag{19}$$
The estimated averaged Q-function $\hat{Q}_i^\theta$ is given by
$$\hat{Q}_i^\theta = (I - \gamma M_i^\theta)^{-1}\hat{r}_i^\theta$$
$$M_i^\theta(s, a_i) := \pi_{\theta_i}(a_i | s) \hat{P}_i^\theta(s'|s, a_i). \tag{19}$$

Similarly, from (2), we have that $d_\theta$ and $\hat{d}_\theta$ are given by (for derivation, see [38, Appendix C])
$$d_\theta = (1 - \gamma) (I - \gamma \hat{P}_i^\theta) \rho, \quad \hat{d}_\theta := (1 - \gamma) (I - \gamma \hat{P}_i^\theta) \rho. \tag{20}$$

4From the Perron–Frobenius theorem, we know that the absolute values of the eigenvalues of $M_i^\theta$ are upper bounded by 1, which guarantees that the matrix $I - \gamma M_i^\theta$ is invertible.

Table II: Game 1: Reward

| $a_2$ | Reward |
|-------|---------|
| $a_1 = 1$ | (-1,1) (-3,0) |
| $a_1 = 2$ | (0,-3) (-2,-2) |

Then, accordingly, the estimated gradient is computed as
$$\hat{d}_{\theta(s,a)} J_i(\theta(k)) = \frac{1}{1 - \gamma} \hat{d}_\theta(s) \hat{Q}_i^\theta(s, a_i). \tag{21}$$

Step 3. Projected gradient ascent onto the set of $\alpha$-greedy policies: Let $U_n = \{1, \ldots, n\} \in \Delta(n)$ denote the $n$-dimensional uniform distribution. Define $\Delta^\alpha(n) := \{\theta \in \Delta(n), \beta \neq \alpha \theta + \alpha U_n\}$. We use $X_i^\alpha := \Delta^\alpha(|A_i|)^{|S|}, X_i^{\alpha} := X_i^\alpha \times X_i^\alpha \times \cdots \times X_i^\alpha$ to denote the set of the $\alpha$-greedy policies for $\theta_i$ and $\theta$, respectively. Every step after doing gradient ascent, the parameter $\theta$ will further be projected onto $X_i^{\alpha}$, i.e.,
$$\theta_i^{(k+1)} = \text{Proj}_{X_i^{\alpha}}(\theta_i^{(k)} + \eta \hat{d}_{\theta(s,a)} J_i(\theta(k))). \tag{22}$$

The reason of projecting onto $X_i^{\alpha}$ instead of $X_i^\alpha$ is to make sure that every action has positive possibility of being selected in order to get a relatively accurate estimation of averaged Q-function. Intuitively, a larger $\alpha$ introduces a larger additional error in the NE-gap; however, a smaller $\alpha$ requires more samples to estimate the gradient. Thus, the choice of $\alpha$ is the tradeoff between the two effects.

Theorem 5 (Sample complexity): Assume that the MPG satisfies Assumption 2. Let $M := \max_{\theta, \alpha} \|D_{\theta} \|_\infty$. In Algorithm 1, for $\eta \leq \frac{(1 - \gamma)\delta}{M^2 |S| \alpha \delta^2}$ and $T_j \geq \frac{20\theta^2 |S| |A_i| \log \left( \frac{16 |S| \log |S| \sum |A_i|}{\delta} \right)}{(1 - \gamma)\delta^2 |S| \alpha^2}$

$$T_G \geq \frac{64 |S|^2 M^2 (\Phi_{\max} - \Phi_{\min})}{\eta^2 \delta^{3}}$$

with probability at least $1 - \delta$, we have that
$$\frac{1}{T_G} \sum_{k=1}^{T_G} \text{NE} - \text{gap}(\theta(k))^2 \leq \epsilon^2.$$

That is, with a proper choice of step size, e.g., $\eta = \frac{(1 - \gamma)^3}{8 |S| \sum |A_i|}$, the algorithm can find an $\epsilon$-NE with probability at least $1 - \delta$ with
$$T_j T_G \sim O \left( \frac{n}{\epsilon^2} \text{poly} \left( \frac{1}{1 - \gamma}, |S|, \max |A_i| \right) \right) \tag{23}$$
samples, where $O$ hides log factors.

We would like to first compare our result with one related work with sample complexity of learning MPGs [31]. Interestingly, both sample complexities are $O(1/\epsilon^4)$. It is an interesting question to study whether such dependence is fundamental or not for learning with simultaneous-updating agents. Yet, Leonardos et al. [31] consider Monte Carlo model-free gradient estimation, while our algorithm takes a model-based approach that suffers less from high variance and the notion of “averaged” MDP can potentially be extrapolated to other settings.

Proof sketch: The proof of Theorem 5 consists of three major steps. The first step is to bound the estimation error of parameters $P_i^\theta, \hat{r}_i^\theta$, and $\hat{d}_\theta$ of the “averaged” MDP. This step leverages Assumption 1 and Azuma–Hoeffding inequality to get high

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
probability bounds for the parameters. The second step translates the estimation error of the “averaged” MDP into the gradient estimation error. Then, the third step treats gradient estimation step as an oracle that gives biased gradient information, where the bias is the estimation error. The final result is obtained by analyzing the performance of biased projected gradient ascent algorithm. The detailed proofs are provided in Appendix D.

Comparison with centralized learning: The best known sample complexity bound for single-agent/centralized MDP is $O\left(\frac{|S|\cdot|A|}{(1-\gamma)^2\epsilon^4}\right)$ [63]. Compared with (23), the centralized bound scales better with respect to $\epsilon$, $|S|$, $|A|$, $\frac{1}{1-\gamma}$. However, as argued in the previous subsection, the total action space $|A| = \prod_{i=1}^n |A_i|$ in the centralized bound scales exponentially with the number of agent $n$, while our complexity bound only scales linearly. Here, we briefly state the fundamental difficulties of learning in the SG setting compared with centralized learning, which also explains why our bound scales worse with respect to the factors $\epsilon$, $|S|$, $|A|$, $\frac{1}{1-\gamma}$.

1) First, the optimization landscape in the SG setting is more complicated. For centralized learning, the gradient domination property is stronger and accelerated gradient methods (e.g., via natural policy gradient or entropy regularization) can speed up the convergence of exact gradient from $O(\frac{1}{\epsilon})$ to $O(\frac{1}{\epsilon^4})$ [28], or even $O(\log(\frac{1}{\epsilon}))$ [53]. In contrast, for multiagent settings, due to the more complicated optimization landscape, these methods can no longer improve the dependency on $\epsilon$ and, thus, makes the outer loop complexity $T_C$ larger.

2) Second, the behavior of other agents makes the environment nonstationary, i.e., the averaged $Q$-function $Q_i^o$ as well as the averaged transition probability distribution $P_i^o$ depends on the policy of other agent $\theta_{-i}$, $\epsilon$. Thus, unlike centralized learning, where the state transition probability matrix can be estimated in an off-policy or even online manner, i.e., using data samples from different policies, $P_i^o$ can only be estimated in an online manner, using samples generated by exactly the same policy $\theta$, which increases the inner loop complexity $T_J$.

3) Third, the complicated interactions among agents necessitate more care during the learning process. Algorithms designed for centralized learning that achieve near-optimal sample complexity are generally $Q$-learning-type algorithms. However, in SGs, it can be shown that having each agent maximize its own averaged $Q$-function may actually lead to nonconvergent behavior. Thus, we need to consider algorithms that update in a less aggressive manner, e.g., soft $Q$-learning, or policy gradient (which is considered in this article).

V. NUMERICAL SIMULATIONS

This section studies three numerical examples to corroborate our theoretical results. The multistage prisoner’s dilemma (Game 1) confirms the local stability results for general-sum SGs; the coordination game (Game 2) considers local stability as well as convergence rate of exact gradient play for MPG; the state-based coordination game (Game 3) tests the performance of the sample-based algorithm proposed in Section IV-C.

A. Game 1: Multistage Prisoner’s Dilemma

The first example—multistage prisoner’s dilemma model [41]—studies exact gradient play for general SGs. It is a two-agent SG, with $S = A_1 = A_2 = \{1, 2\}$. The reward for each agent $r_i(s, a_1, a_2), i \in \{1, 2\}$, is independent of state $s$ and is given by Table II. The state transition probability is determined by agents’ previous actions $P(s_{t+1} = 1|a_1, a_2) = (1, 1) = 1 - \epsilon$ $P(s_{t+1} = 1|a_1, a_2, s_i) = (1, 1) = \epsilon$.

Here, action $a_i = 1$ means that agent $i$ chooses to cooperate and $a_i = 2$ means betray. The state $s$ serves as a noisy indicator, with error rate $\epsilon$, of whether both agents cooperated ($s_i = 1$) or not ($s_i = 2$) in the previous stage $t - 1$.

The single-stage game corresponds to the famous Prisoner’s dilemma, and it is well known that there is a unique NE $(a_1, a_2) = (2, 2)$, where both agents decide to betray. The dilemma arises from the fact that there exists a joint non-NE strategy $(1, 1)$ such that both players obtain a higher reward than what they get under the NE. However, in the multistage case, the introduction of an additional state $s$ allows agents to make decisions based on whether they have cooperated before. It turns out that cooperation can be achieved given that the discounting $\gamma$ is close to 1 and that the indicator for $s$ is accurate enough, i.e., $\epsilon$ is close to 0. Apart from the fully betray strategy, where both agents will betray regardless of $s$, there is another strict NE $\theta^*$ that is $\theta^* = 1, a_1 = 1, \theta^* = 2, a_1 = 2$.

We simulate gradient play for this model and mainly focus on the convergence to the cooperative equilibrium $\theta^*$. We fix $\gamma = 0.95$. The initial policy is set as $\theta_{s=1, a_1=1} = 1 - 0.4\delta_1$ and $\theta_{s=2, a_1=1} = 0$, where the $\delta$’s are uniformly sampled from [0,1]. The initialization implies that at the beginning, both agents are willing to cooperate to some extent given that they cooperated at the previous stage. Fig. 1 shows a trial converging to the NE starting from a randomly initialized policy. Then, we study the size of attraction region for $\theta^*$ and how it varies with the indicator’s error rate $\epsilon$, which is shown in Table III. The size of the region of attraction for $\theta^*$ can be reflected by the ratio of convergence $\frac{\# \text{Trials that converge to } \theta^*}{\# \text{Total number of trials}}$ for multiple trials with different initial points. Here, we calculate one ratio.

![Fig. 1](image-url)
Table IV

| \( s_2 = 1 \) | \( s_2 = 2 \) |
|---|---|
| \( s_1 = 1 \) | 2 | 0 |
| \( s_1 = 2 \) | 0 | 1 |

Fig. 2. (Game 2) Starting from a close neighborhood of a fully mixed NE.

Fig. 3. (Game 2) Total reward for multiple runs.

using 100 trials, and the mean and standard deviation (std) are calculated by computing the ratio ten times using different trials. An empirical estimate of the volume of the region is the convergence ratio times the volume of the uniform sampling area; hence, the larger the ratio, the larger the region of attraction. Intuitively speaking, the more accurately the state \( s \) represents the cooperation situation of the agents, the less incentive agents will have for betraying when observing \( s = 1 \), that is, the larger \( \Delta^0(s=1) \) will become, and thus, the larger the convergence ratio will be. This intuition matches the simulation result as well as the theoretical guarantees on the local convergence around a strict NE in Theorem 2.

B. Game 2: Coordination Game

It is an identical-reward game, which is one special class of MPG. Consider a two-agent identical reward coordination game problem with state space \( S = S_1 \times S_2 \) and action space \( A = A_1 \times A_2 \), where \( S_1 = S_2 = A_1 = A_2 = \{1, 2\} \). The state transition probability is given by

\[
P(s_{i,t+1} = 1|a_{i,t} = 1) = 1 - \epsilon, \quad P(s_{i,t+1} = 1|a_{i,t} = 2) = \epsilon
\]

where \( i = 1, 2 \). The reward table is given by Table IV. Here, we can view the actions \( \{1, 2\} \) as two different social networks that agents can choose. They are rewarded only if they are in the same network. Network 1 has a higher reward than network 2. The state \( s_i \) stands for the network that agent \( i \) is really at after taking an action, \( \epsilon \) stands for the randomness in reaching a network after taking the action.

There is at least one fully mixed NE where both agents join network 1 with probability \( \frac{1-3\epsilon}{3(1-2\epsilon)} \) regardless of the current occupancy of networks, and there are 13 different strict NEs that can be verified numerically. Fig. 2 shows a gradient play trajectory whose initial point lies in a close neighborhood of the mixed NE. As the algorithm progresses, we see that the trajectory in Fig. 2 diverges from the mixed NE, indicating that the fully mixed NE is indeed a saddle point. This corroborates our finding in Theorem 4. Fig. 3 shows the evolution of total reward \( J(\theta(t)) \) for gradient play for different random initial points \( \theta(0) \). Different initial points converge to one of 13 different strict NEs each with a different total reward (some strict NEs with relatively small region of attraction are omitted in the figure). While the total rewards are different, as shown in Fig. 4, we see that the NE-gap of each trajectory (corresponding to same initial points in Fig. 3) converges to 0. This suggests that the algorithm is indeed able to converge to an NE. Notice that the NE-gaps do not decrease monotonically.

C. Game 3: State-Based Coordination Game

Our third numerical example studies the empirical performance of the sample-based learning algorithm, i.e., Algorithm 1. Here, we consider a generalization of coordination game (Game 2) where the two players now try to coordinate on a 2-D grid. The two-player state-based coordination game on a \( 3 \times 3 \) grid is defined as follows: the state space is given by \( S = S_1 = S_2 = \{x_a, x_b, x_c\} \times \{y_a, y_b, y_c\} \), action space is given by \( A = A_1 \times A_2 \), \( A_1 = A_2 = \{\text{Stay, Left, Right, Up, Down}\} \), i.e., agent can choose to stay at current grid or move left/right/up/down to its neighboring grids (as illustrated in Fig. 5). We assume that there is random noise during the transition, where the agent might end up in a neighboring grid of the target location with error probability \( \epsilon \). The reward is given by

\[
r(s_1, s_2) = 1\{s_1 = s_2 = \{x_a, y_a\} \text{ or } \{x_c, y_c\}\}
\]

i.e., the two agents are only rewarded if they stay at the upper-left or lower-right corner at the same time.

For numerical simulation, we take \( T_c = 300, T_I = 10000, \alpha = 0.1, \eta = 10, \) and \( \epsilon = 0.1 \); the numerical results are as displayed in Figs. 6–8. Fig. 6 shows that total reward increases as the number of iterations increases, and Fig. 8 shows that the NE-gap converges to a value close to zero. However, because we project the policy to the \( \alpha \)-greedy set \( \mathcal{X}^\alpha \), the NE-gap cannot converge to exactly zero. Fig. 7 visualizes the discounted state visitation distribution. To make the visualization more intuitive, we look at the marginalized discounted state visitation distribution \( d^\pi_0 \) defined as follows:

\[
d^\pi_0(s_1, x_2, s_2, x_2) = \sum_{s_1, y_1, s_2} d_\theta(s_1, x_1, y_1, s_2, x_2, s_2, x_2).
\]
Fig. 7: Marginal distribution $d^n_0(s_{1,x}, s_{2,x})$ and $d^n_0(s_{1,y}, s_{2,y})$.

Fig. 8: NE − gap converges to a value close to zero. Here, the NE − gap is measured by $\max\{\max_{(s,a_1)} A_i^n(s, a_i)\}.$

d^n_0$ is defined similarly. From Fig. 7, we can see that most of the probability measure concentrates on $\{(x_a, x_a), (x_c, x_c)\}, \{(y_a, y_b), (y_c, y_c)\}$, indicating that the two agents are able to coordinate most of the time.

VI. CONCLUSION AND DISCUSSION

This article studies the optimization landscape and convergence of gradient play for SGs. For general SGs, we establish local convergence for strict NEs. For MPG, we establish the global convergence with respect to NE-gap and the local stability results for strict NEs as well as fully mixed ones. A sample-based NE-learning algorithm with sample complexity guarantee is also proposed under this setting. There are many interesting future directions. First, the current assumption of MPGs is relatively strong compared with the notion of potential games in the one-shot setup, which might restrict its application to broader settings. More effort would be needed to identify other special types of SGs that facilitate efficient learning. It would also be meaningful to investigate real-life applications, such as dynamic congestion and routing. Second, other sample-based learning methods, such as actor–critic, natural policy gradient, Gauss–Newton methods, could also be considered, which might improve the sample complexity.

APPENDIX

A. Proof of Lemma 2

Proof: According to the performance difference lemma (cf. [54]), let $\theta' := \langle \theta'_1, \theta'_2, \ldots, \theta'_n \rangle$, and

\[
J_i(\theta'_i, \theta'_{-i}) - J_i(\theta_i, \theta_{-i}) = \frac{1}{1 - \gamma} \sum_{s,a} d_\theta(s) \pi_{\theta'}(a|s) A_i^n(s, a)
\]

\[
= \frac{1}{1 - \gamma} \sum_{s,a_i} d_\theta(s) \pi_{\theta'}(a_i|s) \sum_{a_{-i}} \pi_{\theta_{-i}}(a_{-i}|s) A_i^n(s, a_i, a_{-i})
\]

\[
= \frac{1}{1 - \gamma} \sum_{s,a_i} d_\theta(s) \pi_{\theta'}(a_i|s) A_i^n(s, a_i).
\]

B. Proof of Theorem 2 and Lemma 5

Proof: (Lemma 5) For a given strict NE $\theta^*$, randomly set $a_i^*(s) \in \arg \max_{a_i} A_i^n(s, a_i)$, and let $\theta_i$ be $\theta_{s,a_i} = 1 \{ a_i = a_i^*(s) \}$. Let $\theta := \langle \theta_1, \theta_2, \ldots, \theta_n \rangle$. From Lemma 2, we have

\[
J_i(\theta_i, \theta_{-i}) - J_i(\theta^*_i, \theta^*_{-i}) = \sum_{s,a_i} d_\theta(s) \pi_{\theta}(a_i|s) A_i^n(s, a_i)
\]

\[
= \sum_{s} d_\theta(s) \max_{a_i} A_i^n(s, a_i) \geq 0.
\]

Because $\theta^*$ is a strict NE, this inequality above forces $\theta^*_i = \theta$, and that $\max_{a_i} A_i^n(s, a_i) = 0$. The uniqueness of $\theta^*$ also implies the uniqueness of $a_i^*(s)$, and thus

\[
A_i^n(s, a_i) < 0 \forall a_i \neq a_i^*(s)
\]

which completes the proof of the lemma.

The proof of Theorem 2 relies on the following auxiliary lemma, whose proof can be found in [38, Appendix L].

Lemma 6: Let $\mathcal{X}$ denote the probability simplex of dimension $n$. Suppose $\theta \in \mathcal{X}, g \in \mathbb{R}^n$ and that there exists $i^* \in \{ 1, 2, \ldots, n \}$ and $\Delta > 0$ such that $\theta_{i^*} = 1, g_i, g_i \geq g_i + \Delta \forall i \neq i^*$. Let $\theta' = \text{Proj}_{\mathcal{X}}(\theta + g)$; then

\[
\theta_{i^*}' \geq \min \left\{ 1, \theta_{i^*} + \frac{\Delta}{2} \right\}.
\]

Proof: (Theorem 2) For a fixed agent $i$ and state $s$, the gradient play (5) update rule of policy $\theta_{i,s}$ is given by

\[
\theta_{i,s}^{(t+1)} = \text{Proj}_{\Delta(|A_i|)} \left( \theta_{i,s}^{(t)} + \frac{\eta}{1 - \gamma} d_{\theta}(s) Q_i^{(t)}(s, \cdot) \right) \tag{24}
\]

where $\Delta(|A_i|)$ denotes the probability simplex in the $|A_i|$th dimension and $Q_i^{(t)}(s, \cdot)$ is an $|A_i|$th dimensional vector with the $a_i$th element equal to $Q_i^{(t)}(s, a_i)$. We will show that this update rule satisfies the conditions in Lemma 6, which will then allow us to prove that

\[
D(\theta^{(t+1)}||\theta^*) \leq \max \left\{ 0, D(\theta^{(t)}||\theta^*) - \frac{\eta \Delta}{2} \right\}.
\]

Let $a_i^*(s)$ be the same definition as Lemma 5; then, we have

\[
\frac{1}{1 - \gamma} d_{\theta}(s) Q_i^{(t)}(s, a_i^*(s)) - \frac{1}{1 - \gamma} d_{\theta}(s) Q_i^{(t)}(s, a_i)
\]

\[
\geq \frac{1}{1 - \gamma} d_{\theta}(s) Q_i^{(t)}(s, a_i^*(s)) - \frac{1}{1 - \gamma} d_{\theta}(s) Q_i^{(t)}(s, a_i^*(s))
\]

\[
- \frac{1}{1 - \gamma} d_{\theta}(s) Q_i^{(t)}(s, a_i^*(s)) - \frac{1}{1 - \gamma} d_{\theta}(s) Q_i^{(t)}(s, a_i^*(s))
\]

\[
\geq \frac{d_{\theta}(s)}{1 - \gamma} \left( A_i^n(s, a_i^*(s)) - A_i^n(s, a_i) \right) - 2 |\nabla_{a_i} J_i(\theta(t)) - \nabla_{a_i} J_i(\theta^*)|
\]

\[
\geq \Delta_{\theta} - \frac{4}{(1 - \gamma)^3} \sum_{i=1}^{n} |A_i| \|\theta(t) - \theta^*\| \tag{25}
\]

\[
\geq \Delta_{\theta} - \frac{4}{(1 - \gamma)^3} \sum_{i=1}^{n} |A_i| \sum_{s} \|\theta(t,s) - \theta^*_s\| \tag{26}
\]

where (25) and (26) use the smoothness property in Lemma 9.
We use the proof of induction as supposed for \( t \leq t - 1 \); then, we have
\[
D(\theta^{(t+1)}||\theta^*) \leq \max\{D(\theta^{(t)}||\theta^*) - \frac{\eta \Delta^{\theta^*}}{2}, 0\}.
\]
Thus
\[
D(\theta^{(t)}||\theta^*) \leq D(\theta(0)||\theta^*) \leq \frac{\Delta^{\theta^*}(1 - \gamma)^2}{8m|S|\sum_{i=1}^{|A_i|}}.
\]
Then, we can further conclude that
\[
(1 - \gamma)d_{\theta(1)}(s)Q^{(o)}_\theta(s, a_1^*) - (1 - \gamma)d_{\theta(1)}(s)Q^{(o)}_\theta(s, a_i) \geq \Delta^{\theta^*} - \frac{4}{(1 - \gamma)^2}T|S| \sum_{i=1}^{|A_i|} D(\theta^{(t)}||\theta^*) \geq \Delta^{\theta^*}/2 \ \forall a_i \neq a_1^* (s).
\]
In addition, for \( D(\theta^{(t)}||\theta^*) \leq \frac{\Delta^{\theta^*}(1 - \gamma)^2}{8m|S|\sum_{i=1}^{|A_i|}} \), we may conclude
\[
\theta^{(t)}_{s,a_1^*} \geq 1/2 \geq \theta^{(t)}_{s,a_i}, \ \forall a_i \neq a_1^* (s).
\]
Then, by applying Lemma 26 to (34), we have
\[
\theta^{(t)}_{s,a_1^*} \geq \min \left\{ 1, \theta^{(t)}_{s,a_1^*} + \frac{\eta \Delta^{\theta^*}}{4} \right\}
\]
\[
\Rightarrow \left| \theta^{(t)}_{s,a_1^*} - \theta^{(t)}_{s,a_i} \right| = 2 \left( 1 - \theta^{(t)}_{s,a_1^*} \right)
\]
\[
\leq \max \left\{ 0, \left| \theta^{(t)}_{s,a_1^*} - \theta^{(t)}_{s,a_i} \right| - \frac{\eta \Delta^{\theta^*}}{2} \right\} \ \forall s \in S
\]
\[
\Rightarrow D(\theta^{(t+1)}||\theta^*) \leq \max \left\{ 0, D(\theta^{(t)}||\theta^*) - \frac{\eta \Delta^{\theta^*}}{2} \right\}
\]
which completes the proof. \( \square \)

C. Proof of Theorem 4

For the sake of space, please refer to Appendix H of [38].

D. Proof of Theorem 5

This section gives a more detailed proof for Theorem 5, which follows the proof sketch in Section IV-C.

1) Bound the Estimation Error of the “Averaged” MDP:

We first look into how to bound the estimation error of parameters \( \overline{P}\theta_i, \overline{r}^\theta_i \), and \( \overline{P}\theta_S \) of the “averaged” MDP. Assumption 2 has introduced \((\tau, \sigma_S)\)-sufficient exploration on states; however, this assumption alone cannot guarantee \( \overline{P}\theta_i \) and \( \overline{r}^\theta \) to be estimated accurately. Especially, if there is one action \( a_i \) such that \( \pi_{\theta_0}(a_i|s) = 0 \) for all \( s \), then this action will never be visited, and thus, its corresponding \( \overline{P}\theta_i \) and \( \overline{r}^\theta \) cannot be estimated effectively. Thus, in this section, we introduce a similar definition of \((\tau, \sigma_S)\)-sufficient exploration.

Definition 5 ((\(\tau, \sigma_S\))-sufficient exploration): An SG and a policy \( \theta \) are said to satisfy \((\tau, \sigma_S)\)-sufficient exploration condition if there exists a positive integer \( \tau \) and \( \sigma \in (0, 1) \) such that for policy \( \theta \) and any initial state-action pair \( (s, a_i) \) \( \forall i \), we have
\[
\text{Pr}(s_t, a_t, s_0 = s, a_0 = a) \geq \sigma \ \forall s_t, a_t, \tau.
\]

All the result in this step builds on the \((\tau, \sigma_S)\)-sufficient exploration assumption. Note that \((\tau, \sigma_S)\)-sufficient exploration” is a stronger condition compared with \((\tau, \sigma_S)\)-sufficient exploration on states.” However, it is not hard to verify that for any SG that satisfies \((\tau, \sigma_S)\)-sufficient exploration on states, and any

\[ \theta \in \mathcal{X} \alpha \], it will also satisfy \((\tau, \frac{\alpha \sigma_S}{\max_i |A_i|})\)-sufficient exploration condition.

Lemma 7: Assume that the SG with policy \( \theta \) satisfies \((\sigma, \gamma)\)-sufficient exploration condition (see Definition 5); then, fix \( s', s \), and \( a_i, \epsilon \leq 1 \)
\[
\text{Pr} \left( \left| \overline{P}\theta(s'|s, a_i) - \overline{P}\theta(s'|s, a_i) \right| \geq \epsilon \right) \leq 4 \exp \left( -\frac{\epsilon^2 \sigma^2 |T|}{32} \right)
\]
\[
\text{Pr} \left( \left| \overline{r}^\theta(s, a_i) - \overline{r}^\theta(s, a_i) \right| \geq \epsilon \right) \leq 4 \exp \left( -\frac{\epsilon^2 \sigma^2 |T|}{32} \right)
\]
\[
\text{Pr} \left( \left| \overline{P}\theta(s'|s) - \overline{P}\theta(s'|s) \right| \geq \epsilon \right) \leq 4 \exp \left( -\frac{\epsilon^2 \sigma^2 |T|}{32} \right).
\]

Detailed proofs are deferred to Appendix E1. Based on Lemma 7, we have the following corollary on bounding the estimation error of \( \overline{P}\theta_i \) and \( \overline{d}_\theta \), which will be helpful in the next step, i.e., bounding the gradient estimation error.

Corollary 1 (Estimation error of \( \overline{P}\theta_i \) and \( \overline{d}_\theta \)): Assume that the SG with policy \( \theta \) satisfies \((\sigma, \gamma)\)-sufficient exploration condition (see Definition 5); then, for a fixed \( i \), running Algorithm 1 will guarantee that
\[
\text{Pr} \left( \exists i, \left| \overline{P}\theta_i - \overline{P}\theta_i \right| \geq \epsilon \right) \leq 8 \sigma |S|^2 \sum_{i=1}^{|A_i|} \exp \left( -\frac{(1 - \gamma)^4 \epsilon^2 |S|^2 |T|}{32 |S|^2} \right)
\]
\[
\text{Pr} \left( \left| \overline{d}_\theta - d_\theta \right| \geq \epsilon \right) \leq 4 \sigma |S|^2 \exp \left( -\frac{(1 - \gamma)^4 \epsilon^2 |S|^2 |T|}{32 |S|^2} \right).
\]

2) Bound the Gradient Estimation Error: Since gradient can be written as a function on \( \overline{P}\theta_i \) and \( \overline{d}_\theta \) [see (7)], from Corollary 1, we have the following lemma (see Appendix E2 for proof).

Lemma 8 (Error bound for gradient estimation): Assume that the SG that satisfies Assumption 2. In Algorithm 1, for
\[
T_j \geq 32(1 + \alpha)^2 |S|^2 \sum_{i=1}^{|A_i|} \max_i |A_i|^2 \log \left( \frac{16T_j |S|^2 \sum_{i=1}^{|A_i|} |A_i|}{(1 - \gamma)^2 \alpha^2 \sigma^2} \right) + \tau
\]
with probability at least \( 1 - \delta \), we have
\[
\left| \nabla \Phi(\theta(k)) - \nabla \Phi(\theta(k)) \right| \leq \epsilon \ \forall 0 \leq k \leq T_G - 1.
\]

3) Gradient Ascent With Biased Gradient Estimation:

We first show that the total potential function \( \Phi \) is smooth.

Lemma 9: (Smoothness)
\[
\left| \nabla \Phi(\theta') - \nabla \Phi(\theta) \right| \leq \frac{\frac{1}{(1 - \gamma)^2}}{\sum_{i=1}^{|A_i|} \| \theta' - \theta \|}
\]
where \( g(\theta) = \{ \nabla_\theta, J_\theta(\theta) \} \).

Due to space limit, we refer the reader to [38, Appendix K] for the proof of the smoothness lemma. The proof resembles the smoothness lemma for centralized RL developed in [28].

We define the following variables that will be useful in this section’s analysis:
\[
\overline{G}(\theta) := \frac{1}{\eta} \left( \text{Proj}_\kappa(\theta + \eta \nabla \Phi(\theta)) - \theta \right)
\]
\[
\overline{G}_\alpha(\theta) := \frac{1}{\eta} \left( \text{Proj}_{\kappa_\alpha}(\theta + \eta \nabla \Phi(\theta)) - \theta \right).
\]
Note that \( \hat{G}^\eta(\theta) \) is the same as gradient mapping in [28, Proposition B.1] if \( \nabla \Phi(\theta) \) is replaced with the true gradient \( \nabla \Phi(\theta) \). \( \hat{G}^{\eta, \alpha}(\theta) \) only differs from \( \hat{G}^{\eta}(\theta) \) in terms of the set of projection, which is the \( \alpha \)-greedy set \( \mathcal{X}_\alpha \) instead of the true feasibility set \( \mathcal{X} \). Following similar techniques for nonconvex optimization (cf. [59]), we have the following sufficient ascent lemma in terms of \( \hat{G}^{\eta, \alpha}(\theta) \).

**Lemma 10 (Sufficient ascent):** Suppose that \( \Phi(\theta) \) is \( \beta \)-smooth. Let \( \theta^+ = \text{Proj}_{\mathcal{X}_\alpha}(\theta + \eta \hat{\nabla} \Phi(\theta)) \). Then, for \( \eta \leq \frac{1}{2\beta g} \), we have

\[
\Phi(\theta^+) - \Phi(\theta) \geq \frac{\eta}{4} \| \hat{G}^{\eta, \alpha}(\theta) \|^2 - \frac{\eta}{2} \| \hat{\nabla} \Phi(\theta) - \hat{\nabla} \Phi(\theta) \|^2.
\]

**Proof:** From the smoothness property, we have

\[
\Phi(\theta^+) - \Phi(\theta) \geq \hat{\nabla} \Phi(\theta) \Phi(\theta^+) - \frac{\eta}{2} \| \hat{\nabla} \Phi(\theta) - \hat{\nabla} \Phi(\theta) \|^2.
\]

Since \( \theta^+ = \text{Proj}_{\mathcal{X}_\alpha}(\theta + \eta \hat{\nabla} \Phi(\theta)) \), we have

\[
(\theta + \eta \hat{\nabla} \Phi(\theta^+) - \theta^+) \leq 0 \quad \forall \theta' \in \mathcal{X}_\alpha.
\]

Taking \( \theta^+ = \theta \), we get

\[
\hat{\nabla} \Phi(\theta) \Phi(\theta^+) - \frac{\eta}{2} \| \hat{\nabla} \Phi(\theta) - \hat{\nabla} \Phi(\theta) \|^2.
\]

Thus, from smoothness, we have

\[
\Phi(\theta^+) - \Phi(\theta) \geq \left( \frac{1}{2\eta g} \beta \right) \| \theta^+ - \theta \|^2 - \frac{\eta}{2} \| \hat{\nabla} \Phi(\theta) - \hat{\nabla} \Phi(\theta) \|^2
\]

\[
\geq \frac{1}{4\eta g} \| \theta^+ - \theta \|^2 - \frac{\eta}{2} \| \hat{\nabla} \Phi(\theta) - \hat{\nabla} \Phi(\theta) \|^2
\]

\[
= \frac{\eta}{4} \| \hat{G}^{\eta, \alpha}(\theta) \|^2 - \frac{\eta}{2} \| \hat{\nabla} \Phi(\theta) - \hat{\nabla} \Phi(\theta) \|^2,
\]

which completes the proof.

Lemma 10 immediately results in the following corollary.

**Corollary 2 (of Lemma 10):** In Algorithm 1, suppose that \( \| \hat{\nabla} \Phi(\theta^k) - \hat{\nabla} \Phi(\theta^k) \| \) holds for every \( 0 \leq k \leq T_G - 1 \); then, running Algorithm 1 will guarantee that

\[
\frac{1}{T_G} \sum_{k=0}^{T_G-1} \| \hat{G}^{\eta, \alpha}(\theta^k) \|^2 \leq \frac{4(\Phi_{\max} - \Phi_{\min})}{\eta T_G} + 2g^\alpha.
\]

**Proof:** From Lemma 10, we have

\[
\Phi(\theta^{(k+1)}) - \Phi(\theta^k) \geq \frac{\eta}{4} \| \hat{G}^{\eta, \alpha}(\theta^k) \|^2 - \frac{\eta}{2} \| \hat{\nabla} \Phi(\theta^k) - \hat{\nabla} \Phi(\theta^k) \|^2
\]

\[
\geq \frac{\eta}{4} \| \hat{G}^{\eta, \alpha}(\theta^k) \|^2 - \frac{\eta}{2} 2g^\alpha.
\]

Thus, by telescoping

\[
\frac{1}{T_G} \sum_{k=0}^{T_G-1} \| \hat{G}^{\eta, \alpha}(\theta^k) \|^2 \leq \frac{4(\Phi(\theta^0) - \Phi(T_G))}{\eta T_G} + 2g^\alpha
\]

\[
\leq \frac{4(\Phi_{\max} - \Phi_{\min})}{\eta T_G} + 2g^\alpha.
\]

4) **Proof of the Main Theorem (see Theorem 5):** We first link the norm of \( \| \hat{G}^{\eta, \alpha}(\theta) \| \) with the first-order stationarity condition via the following lemma.

**Lemma 11 (First-order stationarity and \( \| \hat{G}^{\eta, \alpha}(\theta) \| \):** Suppose that \( \Phi(\theta) \) is \( \beta \)-smooth. Let \( \theta^+ = \text{Proj}_{\mathcal{X}_\alpha}(\theta + \eta \hat{\nabla} \Phi(\theta)) \).

Then

\[
\nabla_\theta \Phi(\theta^+) \nabla_\theta \Phi(\theta) + \| \hat{\nabla} \Phi(\theta) - \hat{\nabla} \Phi(\theta) \| \| \theta^+ - \theta \|^2 \quad \forall \theta' \in \mathcal{X}_\alpha.
\]

Furthermore

\[
\max_{\theta' \in \mathcal{X}_\alpha} \nabla_\theta \Phi(\theta^+) \nabla_\theta \Phi(\theta) + \| \hat{\nabla} \Phi(\theta) - \hat{\nabla} \Phi(\theta) \| + \frac{2\alpha M}{1 - \gamma}.
\]

Proof of Lemma 11 is in Appendix E3. Theorem 5 can be completed by combining all the above results.

**Proof:** (of Theorem 5) Recall that \( \Phi(\theta) \) is \( \beta \)-smooth with \( \beta = \frac{1}{(1 - \gamma)^3 \sum_{i=1}^{T_G} |A_i|} \). The step size \( \eta \) in Theorem 5 satisfies \( \eta \leq \frac{1}{T_G} \frac{1}{(1 - \gamma)^3} = \frac{1}{2\beta g} \).

Recall from the gradient domination property that

\[
\nabla_\theta \Phi(\theta^k) - \hat{\nabla} \Phi(\theta^k) \| \leq \frac{M \max_{\theta' \in \mathcal{X}_\alpha} \| \nabla_\theta \Phi(\theta') \|}{\eta (1 - \gamma) \sum_{i=1}^{T_G} |A_i|}.
\]

Suppose that \( \| \hat{\nabla} \Phi(\theta^k) - \hat{\nabla} \Phi(\theta^k) \| \leq \epsilon \eta, \forall 0 \leq k \leq T_G - 1 \), we can check from Lemma 11 that

\[
\max_{\theta' \in \mathcal{X}_\alpha} \nabla_\theta \Phi(\theta^k) \nabla_\theta \Phi(\theta) + \| \hat{\nabla} \Phi(\theta) - \hat{\nabla} \Phi(\theta) \| \leq M \max_{\theta' \in \mathcal{X}_\alpha} \| \nabla_\theta \Phi(\theta') \| + \frac{2\alpha M}{1 - \gamma}.
\]

Thus

\[
\frac{1}{T_G} \sum_{k=0}^{T_G-1} \| \hat{G}^{\eta, \alpha}(\theta^k) \|^2 \leq \frac{3}{T_G} \sum_{k=0}^{T_G-1} 4M^2|S|((1 + \eta \beta)^2 \| \hat{G}^{\eta, \alpha}(\theta^k) \|)^2 + 4M^2|S|\epsilon^2 + \frac{4\alpha^2 M^2}{(1 - \gamma)^2}
\]

\[
= 12M^2|S| \left( \epsilon^2 + \frac{\alpha^2 M}{1 - \gamma} + 4M^2 + \frac{1}{(1 - \gamma)^2} \right).
\]

From Corollary 2, we have

\[
\frac{1}{T_G} \sum_{k=0}^{T_G-1} \| \hat{G}^{\eta, \alpha}(\theta^k) \|^2 \leq 12M^2|S| \epsilon^2 + 12\alpha^2 M^2 + \frac{108M^2|S|\Phi_{\max}^2}{\eta T_G}.
\]

Substituting

\[
\alpha = \frac{(1 - \gamma) \epsilon}{6M}, \quad \epsilon = \frac{\epsilon}{2\sqrt{33|S|M}}, \quad T_G \geq \frac{64 \lambda M^2 \Phi_{\max}^2}{\eta \epsilon^2}.
\]

\[
\]
into the above inequality, we obtain
\[
\frac{1}{T_G} \sum_{k=0}^{T_G-1} \text{NE} - \text{gap}(\theta^{(k+1)})^2 \leq \frac{\epsilon^2}{2} + \frac{\epsilon^2}{3} + \frac{\epsilon^2}{6} = \epsilon^2.
\]
Substituting the value of \(\alpha\) and \(\epsilon_k\) in (32) into Lemma 8 will give us
\[
T_j \geq \frac{206997\tau \eta M^4 |s^3| \max_i |A_i|^3}{(1-\gamma)^3 \epsilon^2 \sqrt{2}} \log \left( \frac{16\tau T_G |s|^2 \sum_i |A_i|}{\delta} \right) + \tau
\]
which completes the proof. \(\square\)

**E. Proof of Lemmas in Appendix D**

1) **Bound the Estimation Error of the “Averaged” MDP:**
Here, for compactness, we only provide proof for Lemma 7; the proof of Corollary 1 can be found in [38, Appendices I.1 and I.2].

**Proof:** (of Lemma 7) To avoid repetition, here, we only provide the proof for the first inequality on \(\hat{P}_I^\theta(s'|s, a_i) - \bar{P}_I^\theta(s'|s, a_i)\). The proof for \(\hat{P}_I^\theta(s'|s, a_i) - \bar{P}_I^\theta(s'|s, a_i)\) uses same techniques by leveraging the Azuma–Hoeffding inequality.

According to the definition of \(\hat{P}_I^\theta\), we have
\[
\left\{ \begin{array}{l}
\hat{P}_I^\theta(s'|s, a_i) - \bar{P}_I^\theta(s'|s, a_i) \geq \epsilon \\
\end{array} \right. 
\]
\[
\leq \left\{ \begin{array}{l}
T-1 \sum_{t=0}^{T-1} \left( \begin{array}{l}
\{ s_{t+1} = s', s_t = s, a_{i,t} = a_i \\
-\bar{P}_I^\theta(s'|s, a_i) + \epsilon \} \{ s_t = s, a_{i,t} = a_i \} \geq 0 \\
\end{array} \right) \\
\cup \left\{ \begin{array}{l}
T-1 \sum_{t=0}^{T-1} \{ s_t = s, a_{i,t} = a_i \} = 0 \\
\end{array} \right) \\
\leq \left\{ \begin{array}{l}
T-1 \sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} \\
-\bar{P}_I^\theta(s'|s, a_i) + \epsilon \} \{ s_k = s, a_{i,k+1} = a_i \} \geq 0 \\
\right. \\
\leq \left\{ \begin{array}{l}
\sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} \\
-\bar{P}_I^\theta(s'|s, a_i) + \epsilon \} \{ s_k = s, a_{i,k+1} = a_i \} = 0 \\
\right. \\
\right.
\]

Let \(A_m\)
\[
:= \left\{ \begin{array}{l}
\sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} \\
-\bar{P}_I^\theta(s'|s, a_i) + \epsilon \} \{ s_k = s, a_{i,k+1} = a_i \} \geq 0 \\
\right. \\
\right.
\]
\[
A_m':= \left\{ \begin{array}{l}
\sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} = 0 \\
\right. \\
\right.
\]
\[
X_{m,k} := \left\{ \begin{array}{l}
\sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} \\
-\bar{P}_I^\theta(s'|s, a_i) + \epsilon \} \{ s_k = s, a_{i,k+1} = a_i \} \leq 0 \\
\right. \\
\right.
\]
\[
Y_{m,k} := \left\{ \begin{array}{l}
\sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} \geq 0 \\
\right. \\
\right.
\]
\[
Y_{m,k}' := \left\{ \begin{array}{l}
\sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} \leq 0 \\
\right. \\
\right.
\]
\[
X_{m,k} := \left\{ \begin{array}{l}
\sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} \\
-\bar{P}_I^\theta(s'|s, a_i) + \epsilon \} \{ s_k = s, a_{i,k+1} = a_i \} \leq 0 \\
\right. \\
\right.
\]
\[
Y_{m,k} := \left\{ \begin{array}{l}
\sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} \geq 0 \\
\right. \\
\right.
\]
\[
Y_{m,k}' := \left\{ \begin{array}{l}
\sum_{k=0}^{T-1} \{ s_{k+1} = s', s_k = s, a_{i,k+1} = a_i \} \leq 0 \\
\right. \\
\right.
\]
\[
\text{Then, } \{X_{m,k} \}_{k=0}^{T-1} \text{ is a martingale difference sequence. Because } \epsilon \leq 1, \text{ it is easy to verify that } |X_{m,k}| \leq 2, |X_{m,k}'| \leq 1. \text{ Then, we have}
\]
\[
\left\{ \begin{array}{l}
|X_{m,k}| \leq |X_{m,k}| + E[X_{m,k} | F_{(k-1) \tau+m}] \leq 4 \\
|Y_{m,k}'| \leq |X_{m,k}'| + E[X_{m,k}' | F_{(k-1) \tau+m}] \leq 2.
\right. \\
\right.
\]
\[
\text{Furthermore}
\]
\[
E[X_{m,k} | F_{(k-1) \tau+m}] = E[1 \{ s_{k+1} = s', s_{k+1} = s, a_{i,k+1} = a_i \} | F_{(k-1) \tau+m}] \geq \sigma
\]
\[

To move from (33) to (34), we used the fact that
\[
E[1 \{ s_{k+1} = s', s_t = s, a_{i,t+1} = a_i \} | F_{t-1}] = P(s_{k+1} = s', s_t = s, a_{i,t+1} = a_i) \leq P(s_{k+1} = s', s_t = s, a_{i,t} = a_i) \leq P(s_{k+1} = s', s_t = s, a_{i,t+1} = a_i) \leq P(s_{k+1} = s', s_t = s, a_{i,t+1} = a_i) | F_{t-1}
\]
and the inequality in (34) is derived directly from Definition 5.

According to the Azuma–Hoeffding inequality, we have
\[
\Pr(A_m) = \Pr \left( \sum_{k=0}^{T-1} X_{m,k} \geq 0 \right)
\]
\[
= \Pr \left( \sum_{k=0}^{T-1} Y_{m,k} \geq -\sum_{k=0}^{T-1} E[X_{m,k} | F_{(k-1) \tau+m}] \right)
\]
\[
\leq \Pr \left( \sum_{k=0}^{T-1} Y_{m,k} \geq \left( \frac{T-1-m+\tau}{\tau} \right) \sigma \right)
\]
\[
\leq \exp \left( -\frac{\sigma^2 \left( \frac{T^2}{32} \right)}{8} \right).
\]
Similarly, from the Azuma–Hoeffding inequality, we have
\[
\Pr(A'_m) = \Pr \left( \sum_{k=0}^{T-1} X_{m,k}' \geq 0 \right)
\]
\[
= \Pr \left( \sum_{k=0}^{T-1} Y_{m,k}' \geq -\sum_{k=0}^{T-1} E[X_{m,k}' | F_{(k-1) \tau+m}] \right)
\]
\[
\leq \Pr \left( \sum_{k=0}^{T-1} Y_{m,k}' \geq \left( \frac{T-1-m+\tau}{\tau} \right) \sigma \right)
\]
\[
\leq \exp \left( -\frac{\sigma^2 \left( \frac{T^2}{8} \right)}{8} \right).
\]

Thus
\[
\Pr \left( \bar{P}_I^\theta(s'|s, a_i) - \hat{P}_I^\theta(s'|s, a_i) \leq \epsilon \right)
\]
\[
\begin{align*}
\sum_{m=0}^{\tau-1} \Pr(A_m) + \Pr(A'_m) & \leq 2\tau \exp \left(-\frac{\epsilon^2 \sigma^2}{32} \frac{T}{\tau} \right).
\end{align*}
\]

Similarly
\[
\Pr \left( |F_i(s', s, a_i) - F_i(s, s, a_i)| \leq -\epsilon \right) \leq 2\tau \exp \left(-\frac{\epsilon^2 \sigma^2}{32} \frac{T}{\tau} \right).
\]

Thus
\[
\Pr \left( |F_i(s', s, a_i) - F_i(s, s, a_i)| \geq \epsilon \right) \leq 4\tau \exp \left(-\frac{\epsilon^2 \sigma^2}{32} \frac{T}{\tau} \right)
\]

which completes the proof. \(\square\)

2) **Bound the Gradient Estimation Error:** Proof: (Of Lemma 8) Since the SG satisfies \((\tau, \sigma S, \delta)-\text{sufficient exploration on states, then for any } \theta \in \mathcal{X}^\alpha, \text{ we know that it satisfies}

\((\tau, \sigma S, \delta)-\text{sufficient exploration. Substituting this into Corollary 1, we have that for}

\[
T_J \geq \frac{32(1+\alpha)\sqrt{\sum |A_i| \max |A_i|^2 \log (16\tau G^2 |S| \sum |A_i|) + \tau}}{(1-\gamma)^2 \alpha \sigma^2}
\]

with probability at least 1 - \(\frac{\delta}{2TG}\), we have

\[
\|Q^{(k+1)} - Q^{(k)}\| \leq \frac{(1-\gamma)\epsilon_g}{\sqrt{|S| \sum |A_i|}}
\]

\[
\|d_{(k+1)} - d_{(k)}\| \leq \frac{(1-\gamma)^2 \epsilon_g}{\sqrt{|S| \sum |A_i|}}
\]

Since

\[
\begin{align*}
\frac{1}{1-\gamma} d_i(s) Q_i^k(s, a_i) - \frac{1}{1-\gamma} \hat{d}_i(s) \hat{Q}_i(s, a_i) & \\
& \frac{1}{1-\gamma} \hat{Q}_i(s, a_i) (d_i(s) - \hat{d}_i(s)) & \\
& \leq \frac{\epsilon_g}{(1-\gamma)^2 \sqrt{|S| \sum |A_i|}} + \frac{(1-\gamma)^2 \epsilon_g}{\sqrt{|S| \sum |A_i|}} \\
& = \frac{\epsilon_g}{\sqrt{|S| \sum |A_i|}}
\end{align*}
\]

Thus, with probability 1 - \(\delta\), we have

\[
\|\nabla \Phi(\theta^{(k)}) - \nabla \Phi(\theta)\|_2 \leq \epsilon_g \quad \forall 1 \leq k \leq T_G.
\]

3) **Proof of Lemma 11:** Proof: (of Lemma 11) Since \(\theta^+ = \text{Proj}_{\mathcal{X}^\alpha}(\theta + \eta \nabla \Phi(\theta))\), we have

\[
(\theta + \eta \nabla \Phi(\theta) - \theta^+) \leq 0 \quad \forall \theta^+ \in \mathcal{X}^\alpha
\]

\[
\Rightarrow \nabla \Phi(\theta^+) \geq (\theta - \theta^+) \nabla \Phi(\theta - \theta^+)
\]

\[
\Rightarrow \eta \nabla \Phi(\theta^+) \geq (\theta - \theta^+) \nabla \Phi(\theta - \theta^+)
\]

\[
+ \eta (\nabla \Phi(\theta) - \nabla \Phi(\theta^+)) \nabla \Phi(\theta - \theta^+)
\]

\[
+ \eta (\nabla \Phi(\theta^+) - \nabla \Phi(\theta)) \nabla \Phi(\theta - \theta^+).
\]

Thus

\[
\begin{align*}
\eta \nabla \Phi(\theta^+) \geq (\theta - \theta^+) \nabla \Phi(\theta - \theta^+)
\end{align*}
\]

which proves (30). We now prove (31). For any \(\theta^+_i \in \Delta(|A_i|)\), we know that

\[
(1-\alpha)\theta^+_i + \alpha U_{i|A_i} \in \Delta^\alpha(|A_i|).
\]

Let \(U_i := \sum_{i|A_i} U_{i|A_i}\); then, for any \(\theta^+_i \in \Delta_i, (1-\alpha)\theta^+_i + \alpha U_i \in \mathcal{X}^\alpha\).

Thus

\[
\nabla \Phi(\theta^+) \geq (\theta^+_i - \theta^+),
\]

\[
\leq \nabla \Phi(\theta^+) \geq (1-\alpha)\theta^+_i + \alpha U_i - \theta^+,
\]

\[
\nabla \Phi(\theta^+) \geq (1-\alpha)\theta^+_i - \alpha U_i,
\]

\[
\nabla \Phi(\theta^+) \geq (1+\alpha)\theta^+_i + \alpha U_i - \theta^+,
\]

\[
\nabla \Phi(\theta^+) \geq (1-\alpha)\theta^+_i + \alpha U_i - \theta^+,
\]

\[
\nabla \Phi(\theta^+) \geq (1+\alpha)\theta^+_i - \alpha U_i - \theta^+.
\]

Thus

\[
\nabla \Phi(\theta^+) \geq (\theta^+_i - \theta^+),
\]

\[
\nabla \Phi(\theta^+) \geq (1+\alpha)\theta^+_i + \alpha U_i - \theta^+,
\]

\[
\nabla \Phi(\theta^+) \geq (1-\alpha)\theta^+_i - \alpha U_i - \theta^+.
\]

Since

\[
\nabla \Phi(\theta^+) \geq (\theta^+_i - \theta^+),
\]

\[
\nabla \Phi(\theta^+) \geq (1+\alpha)\theta^+_i + \alpha U_i - \theta^+,
\]

\[
\nabla \Phi(\theta^+) \geq (1-\alpha)\theta^+_i - \alpha U_i - \theta^+.
\]

we have

\[
\nabla \Phi(\theta^+) \geq (\theta^+_i - \theta^+),
\]

\[
\nabla \Phi(\theta^+) \geq (1+\alpha)\theta^+_i + \alpha U_i - \theta^+,
\]

\[
\nabla \Phi(\theta^+) \geq (1-\alpha)\theta^+_i - \alpha U_i - \theta^+.
\]

REFERENCES

[1] F. Daneshfar and H. Bevrani, “Load–frequency control: A GA-based multi-agent reinforcement learning,” IET Gener. Transmiss. Distrib., vol. 4, no. 1, pp. 15–23, 2010.

[2] S. Shalev-Shwartz, S. Shamma, and A. Shashua, “Safe, multi-agent, reinforcement learning for autonomous driving,” 2016, arXiv:1610.03295.

[3] T. Chen, K. Zhang, G. B. Giannakis, and T. Başar, “Communication-efficient policy gradient methods for distributed reinforcement learning,” IEEE Trans. Control Netw. Syst., vol. 9, no. 2, pp. 917–929, Jun. 2022.

[4] Y. Li, Y. Tang, R. Zhang, and N. Li, “Distributed reinforcement learning for decentralized linear quadratic control: A derivative-free policy optimization approach,” IEEE Trans. Autom. Control, vol. 67, no. 12, pp. 6429–6444, 2019.

[5] G. Qu, A. Wierman, and N. Li, “Scalable reinforcement learning of localized policies for multi-agent networked systems,” in Proc. Int. Conf. Mach. Learn. Res., 2020, pp. 256–266.

[6] L. S. Shapley, “Stochastic games,” Proc. Nat. Acad. Sci. USA, vol. 39, no. 10, pp. 1095–1100, 1953.

[7] L. Buşoniu, R. Babuška, and B. De Schutter, “Multi-agent reinforcement learning: An overview,” in Innovations in Multi-Agent Systems and Applications-1. Berlin, Germany: Springer, 2010, pp. 183–221.
A. Sidford, M. Wang, X. Wu, L. Yang, and Y. Ye, “Near-optimal time and sample complexities for solving Markov decision processes with a generative model,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2018, pp. 5192–5202.

R. Cooper, Coordination Games. Cambridge, U.K.: Cambridge Univ. Press, 1999.

Runyu (Cathy) Zhang received the B.S. degree in mathematics from Peking University, Beijing, China, in 2019. She is currently working toward the Ph.D. degree in applied mathematics with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA.

During the Summer of 2022, she was a Research Intern with Salesforce Research, working on multiagent reinforcement learning. Her research interests include online control methods, reinforcement learning, game theory and optimization, with particular focus on multiagent systems.

Ms. Runyu was the recipient of the Certificate of Distinction and Excellence in Teaching at Harvard University and the finalist of the Two Sigma Diversity PhD Fellowship in 2022.

Zhaolin Ren received the bachelor’s degree in mathematics and the master’s degree in statistics from Stanford University, Stanford, CA, USA, in 2019. He is currently working toward the Ph.D. degree, working under the supervision of Prof. Na Li, with Harvard University, Cambridge, MA, USA.

His research interests include zeroth-order optimization, control theory, and reinforcement learning.

Runyu (Cathy) Zhang

Zhaolin Ren

Na Li (Senior Member, IEEE) received the B.S. degree in mathematics and applied mathematics from Zhejiang University, Hangzhou, China, in 2007, and the Ph.D. degree in control and dynamical systems from the California Institute of Technology, Pasadena, CA, USA, in 2013.

She is currently a Winokur Family Professor with the School of Engineering and Applied Sciences, Harvard University, where she joined in 2014 as an Assistant Professor. She was a Postdoctoral Associate with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. Her research focuses on the design, analysis, optimization, and control of distributed network systems, with particular applications to power networks and systems biology/physiology.

Dr. Li was the recipient of the NSF CAREER Award, the ONR YIP Award, the Eckman Award, and the IFAC Thoma Medal.