2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Policy Gradient Play over Time-Varying Networks in Markov Potential Games

Sarper Aydın and Ceyhun Eksin

Abstract— **We design a multi-agent and networked policy gradient algorithm in Markov potential games. Each agent has its own rewards and utility as functions of joint actions and a shared state among agents. The state dynamics depend on the joint actions taken. Differentiable Markov potential games are defined based on the existence of a potential (value) function having partial gradients equal to the local gradients of agents' individual value functions. Agents implement continuous parameterized policies defined over the state and other agents' parameters to maximize their utilities against each other. Agents compute their stochastic policy gradients to update their parameters with respect to their local estimates of Q-functions and joint parameters. The updated parameters are shared with neighbors over a time-varying network. We prove the convergence of joint parameters to a first-order stationary point of the potential function in probability for any type of state and action spaces. Numerical results illustrate the potential advantages of using networked policies compared to independent policies.**

I. Introduction

Networked systems are multi-agent structures in which agents share their information (including but not limited to observations, parameters, and actions) with each other through a communication network. Typical examples of the networked learning in multi-agent reinforcement learning (MARL) include stabilization of the joint behavior [1] and gaining more information about the overall system in the case of partial observability by individuals [2] (see [3] for a survey of Networked MARL). Recent success of networked MARL in mobile robotics [4], navigation [5], and traffic management [6] motivate understanding the theoretical properties networked MARL in networked systems. In the given settings, agents need to individually reason and make decisions in dynamics environments. This renders game-theoretical learning methods as natural options for solving the aforementioned (dynamic) multi-agent problems.

Markov games represent competitive multi-agent interactions in dynamic environments. In this study, we aim to develop networked policy learning in the setting of Markov potential games, a well-known class of Markov games, defined by the existence of a global potential function aligned with utility changes by unilateral policy updates. Markov potential games provide a framework to design locally implemented policies in various applications such as navigation [7], path planning [8], and electricity demand management [9].

We develop a new class of policy gradient algorithm where agents take into account other agents' parameters given the state. We propose a networked and parameterized multi-agent policy structure based on episodic policy gradient methods [10], [11].The underlying assumption here is that agents sample actions from parameterized policy functions, which define probability distributions over their action spaces. Agents update their parameters with the gradient information estimated. This requires estimations of discounted sum of rewards, and score functions of policies. In more detail, the gradients are computed through two consecutive episodes whose lengths are randomly sampled from geometric distributions. Agents take actions against each other by employing their individual policy functions during the episodes. They estimate their sum of discounted rewards at the first episode, and then they sample the state-action pair to compute the score functions at the second episode. By the definition of their policies, agents need to gather information about others' parameters to sample their actions. Agents may not have instantaneous access to others' parameters. Instead, agents keep local estimates of others' parameters and share information via a time-varying communication network.

Given the policy gradient play over a network, we prove that the joint policy parameters converge to a stationary point of the potential function in probability. This result exploits the intermediate results that the stochastic gradients generated by random horizon sampling are unbiased estimates (Lemma 3), and local beliefs on others' parameters converge to true parameter values (Lemma 4) thanks to Lipschitz (Lemma 2) and bounded gradients (Lemma 3).

Early studies on policy gradient play in Markov potential games consider continuous state and action spaces, but with the restrictive assumptions that state dynamics and rewards are known [12], [13]. A new generation of studies only concentrates on the direct or softmax parameterization for the problems with finite state and actions. They derive gradient-based update schemes e.g., projected, natural by enabling agents to take independent actions without any communication in Markov potential or general-sum games [14]–[19]. A recent paper [20] focuses on the setting of networked Markov potential games where individual rewards and state are affected only by the neighboring agents' actions and states with independent softmax policies.

Our proposed algorithm does not make any assumptions on the cardinality of the state and the action space in addition to unknown rewards and state transition dynamics. The contributions of this study are two-fold, *i)* we derive the joint networked policy and its gradient, *ii)* we define an information exchange protocol over time-varying networks. Numerical results suggest that networked policies provide numerically better results compared to independent policies

S. Aydin and C. Eksin are with the Industrial and Systems Engineering Department, Texas A&M University, College Station, TX 77843. E-mail: sarper.aydin@tamu.edu; eksinc@tamu.edu.

2023 62nd IEEE Conference on Decision and Control (CDC) | 979-8-3503-0124-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CDC49753.2023.10383556

979-8-3503-0124-3/23/$31.00 ©2023 IEEE

with respect to average accumulated rewards (Q-functions) and convergence of gradient estimations.

II. Markov Potential Games

In a Markov game, N agents defined by the set $\mathcal{N} := \{1, \dots, N\}$, play against each other [21]. Agents decide their actions $a_i \in \mathcal{A}_i \subseteq \mathbb{R}^K$ from a subset of $K \in \mathbb{N}^+$ dimensional real space given a common state $s \in \mathcal{S}$, where the sets $\mathcal{A}_i$ and $\mathcal{S}$ are not necessarily finite, and $i \in \mathcal{N}$ denotes the individual index of agents. We define the joint action profile correspondingly as $a = (a_1, a_2, \cdots, a_N) \in \mathcal{A}^N := \bigtimes_{i \in \mathcal{N}} \mathcal{A}_i$. The transition probabilities between states are dependent on the joint action and prior state $\mathcal{P}^a_{s'',s'} = \mathbb{P}(s''|s', a)$, and the initial state s_0 comes from a prior distribution $\rho : \mathcal{S} \to [0, 1]$. Agent i collects reward $r_{i,t} : \mathcal{S} \times \mathcal{A}^N \to \mathbb{R}$ at each time $t \in \mathbb{N}^+$ determined by the action profile and the state, with the discount factor $\gamma \in (0, 1)$. We formally state the game by the tuple $\Gamma := (\mathcal{N}, \mathcal{A}^N, \mathcal{S}, \{r_{i,t}\}_{i \in \mathcal{N}}, \mathcal{P}, \gamma, \rho)$.

Each agent owns a policy function $\pi_i : \mathcal{S} \times \Delta(\mathcal{A}_{-i}) \to \Delta(\mathcal{A}_i)$ as a mapping from the joint space of states and other agents' policies to a probability distribution from which their actions are sampled, and $\Delta(.)$ indicates all probability distributions on the given set, and $-i := \mathcal{N} \setminus \{i\}$ denotes the set of all agents other than each agent i. We define the value function $V_i^{\Pi} : \mathcal{S} \to \mathbb{R}$ of each agent i for each state $s \in \mathcal{S}$, if agents implement the joint policy $\Pi = \bigtimes_{i \in \mathcal{N}} \pi_i$ as a discounted sum of rewards over infinite horizon,

$$V_i^{\Pi}(s) = \mathbb{E}_{(s,a) \sim \mathcal{P}} \Big[\sum_{t=0}^{\infty} \gamma^t r_{i,t}(s_t, a_t) | s_0 = s \Big], \qquad (1)$$

where $\mathcal{P}$ is the joint distribution of the sequence of states and actions induced by the joint policy and state transition probabilities [1]. Note that we add time sub-index $t \in \mathbb{N}^+$ into a_t, and s_t, to indicate agent i's reward is a function of joint action and common state at decision epoch t. We also define the Q-function of each agent i ($Q_i : \mathcal{S} \times \mathcal{A}^N \to \mathbb{R}$) for each state $s \in \mathcal{S}$, and joint action pair $a \in \mathcal{A}^N$ given the joint policy Π as below,

$$Q_i^{\Pi}(s, a) = \mathbb{E} \Big[\sum_{t=0}^{\infty} \gamma^t r_{i,t}(s_t, a_t) | s_0 = s, a_0 = a \Big]. \qquad (2)$$

A potential game for static (one-shot) games assumes the existence of a potential function mirroring the utility changes as a result of unilateral action changes [22]. Markov potential games suppose the existence of a potential value function that captures the changes in the individual value functions at each state $s \in \mathcal{S}$ as a result of unilateral changes in policies.

Definition 1 (Definition 2 , [13]) *A game Γ is a Markov potential game, if there exists a potential value function $V^{\Pi}(s) : \Pi \times \mathcal{S} \to \mathbb{R}$ that as the discounted sum of potential rewards $r_t \in \mathbb{R}$, i.e., $V^{\Pi}(s) = \mathbb{E} \Big[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t) | s_0 = s \Big]$, such that for all $i \in \mathcal{N}$*

$$V_i^{\hat{\Pi}}(s) - V_i^{\Pi}(s) = V^{\hat{\Pi}}(s) - V^{\Pi}(s) \quad \textit{for all } s \in \mathcal{S}, \qquad (3)$$

where $\hat{\Pi}$ and Π are two joint policies that differ only by the policy of agent $i \in \mathcal{N}$ only, i.e., $\hat{\Pi} = (\hat{\pi}_i, \pi_{-i})$ and $\Pi = (\pi_i, \pi_{-i})$.

We assume agents use parametrized policies defined by unconstrained and continuous variables $\theta = (\theta_i, \theta_{-i}) \in \mathbb{R}^M$ where individual policy parameters $\theta_i \in \mathbb{R}^{M_i}$ are such that the following holds $\sum_{i \in \mathcal{N}} M_i = M$, where $M_i \in \mathbb{N}^+$. Given the parametrized policies $\Pi_\theta : \mathbb{R}^M \times \mathcal{S} \to \Delta(\mathcal{A}^N)$, we also suppose that agents have differentiable and parametrized value functions $u_i : \mathbb{R}^M \to \mathbb{R}$ defined as follows with respect to (1),

$$u_i(\theta_i, \theta_{-i}) := V_i^{\Pi_\theta}(s) = \mathbb{E}_{\Pi_\theta} \Big[\sum_{t=0}^{\infty} \gamma^t r_{i,t}(s_t, a_t) | s_0 = s \Big]. \qquad (4)$$

We further define differentiable Markov potential games as follows.

Definition 2 (Differentiable Markov Potential Games) *A game Γ is a Markov potential game with differentiable individual value functions u_i, if there exists a potential value function $u : \mathbb{R}^M \to \mathbb{R}$ having equivalent partial derivatives of agents' utilities as follows,*

$$\nabla_i u_i(\theta_i, \theta_{-i}) = \nabla_i u(\theta) \quad \textit{for all } \theta \in \mathbb{R}^M \qquad (5)$$

where $\nabla_i(.) = \frac{\partial(.)}{\partial \theta_i}$ denotes the partial derivative of a given function with respect to the agent i's parameters θ_i.

Note that the differentiable policies can be also implemented in the case of finite actions, (e.g. softmax policies) implying that they can provide a solution framework for both continuous and finite action spaces. In the rest of the paper, we use the parametrized value functions $u_i : \mathbb{R}^M \to \mathbb{R}$ for the analysis and refer to them as value functions.

III. Policy Gradient Play with Networked Agents

We define the joint parameterized policy, $\Pi_\theta : \mathbb{R}^M \times \mathcal{S} \to \Delta(\mathcal{A}^N)$, agent i's policy $\pi_{i,\theta}(a_i|s) := \pi_i(a_i|s, \theta)$ as conditionally independent individual policies given the state and joint policy parameters, i.e.,

$$\Pi_\theta(a \in \mathcal{A}_q^N | s) = \prod_{i \in \mathcal{N}} \pi_{i,\theta}(a_i \in \mathcal{A}_{i,q} | s) \qquad (6)$$

where $\mathcal{A}_q^N = \bigtimes_{i \in \mathcal{N}} \mathcal{A}_{i,q}$ and $\mathcal{A}_{i,q}$ are countable measurable partitions over the joint and individual set of actions in order. Note that the policy functions satisfy the axioms of probability measures i.e, countable additivity, non-negativity, and probabilities of empty sets and full spaces.

Note that each agent would like to maximize its cumulative rewards against other agents' policies given the joint action-dependent state dynamics.

Next, we also define of the gradient of agent i's value function in terms of the Q-function and sum of log-policies—see [23] for proof.

Lemma 1 (Lemma 1, [23]) *Given the value functions $u_i : \mathbb{R}^M \to \mathbb{R}$ in (4) and the joint policy function $\Pi_\theta : \mathbb{R}^M \times \mathcal{S} \to$*

[1] For brevity, we remove subscript from the expectation in the rest of the paper, unless clarity necessitates it.

$\Delta(\mathcal{A}^N)$ in (6) *respectively, the policy gradient of each value function u_i with respect to agent i's parameters θ_i is defined as,*

$$\nabla_i u_i(\theta_i, \theta_{-i}) = \frac{1}{(1-\gamma)} \mathbb{E}\big[Q_i^{\Pi_\theta}(s,a) \sum_{n \in \mathcal{N}} \nabla_i \log \pi_{n,\theta}(a_n|s)\big]. \tag{7}$$

Since agents want to maximize their cumulative rewards against other agents' actions, in policy gradient play, each agent uses stochastic gradients to update its policy parameters,

$$\theta_{i,t} = \theta_{i,t-1} + \alpha_t \hat{\nabla}_i u_i(\theta_{i,t-1}, \theta_{-i,t-1}), \tag{8}$$

where we assume that α_t is a (common) step size, and $\hat{\nabla}_i u_i(\theta_{i,t-1}, \theta_{-i,t-1})$ is the stochastic gradient computed based on rewards collected through episodes. From a game theoretic perspective, the gradient update in (8) can be interpreted as a better-reply process on expectation given that the gradient estimation is unbiased.

A. Policy Gradient Estimation

The definition of policy gradients $\nabla_i u_i$ depends on Q-values Q_i and the gradient of log policies $\nabla_i \log \pi_{n,\theta}$ as per Lemma 1. The sequential nature of decision-making may create bias in the estimation of true gradient direction with standard approaches using fixed finite time horizons. We follow and adapt the approach proposed in [24], to obtain unbiased estimations $\hat{Q}_i$ and $\hat{\nabla}_i \log \pi_{n,\theta}$ that respectively replace Q_i and $\nabla_i \log \pi_{n,\theta}$ in (7). In particular, we generate two episodes with random horizon lengths from the geometric distribution $Geom(1-\gamma^{0.5})$ such that $\mathbb{P}(\mathcal{T}_k = \tau) = (1-\gamma^{0.5})\gamma^{0.5\times\tau}$ for $k \in \{1,2\}$ in order to construct estimates for Q_i and $\nabla_i \log \pi_{n,\theta}$. The steps of the proposed sampling are provided in Algorithm 1.

Algorithm 1 Gradient Estimation for Agent $i \in \mathcal{N}$

```
Input: The parameters $\theta$ initial state $s_0$ and discount factor $\gamma$.
Draw $\mathcal{T}_1 \sim Geom(1-\gamma^{0.5})$ and reset $s_0$.
Sample actions $a_{i,0} \sim \pi_{i,\theta}(.|s_0)$
for $\tau = 1, 2, \cdots, \mathcal{T}_1$ do
    Reach state $s_\tau \sim \mathcal{P}^{a_{\tau-1}}_{s_\tau, s_{\tau-1}}$
    Sample and take actions, $a_{i,\tau} \sim \pi_{i,\theta}(.|s_\tau)$
end for
Compute $\nabla_i \log \pi_\theta(a_{\mathcal{T}_1}|s_{\mathcal{T}_1})$
Draw $\mathcal{T}_2 \sim Geom(1-\gamma^{0.5})$ and set $\hat{Q}_i = 0$.
for $\tau = 1, 2, \cdots, \mathcal{T}_2$ do
    Receive rewards $r_{i,\tau+\mathcal{T}_1}$
    Collect rewards $\hat{Q}_i = \hat{Q}_i + \gamma^{\tau/2} r_{i,\tau+\mathcal{T}_1}$
    Reach state $s_{\tau+\mathcal{T}_1+1} \sim \mathcal{P}^{a_{i,\tau}}_{s_{\tau+\mathcal{T}_1+1}, s_{\tau+\mathcal{T}_1}}$.
    Sample and take actions $a_{i,\tau+\mathcal{T}_1+1} \sim \pi_{i,\theta}(.|s_{\tau+\mathcal{T}_1})$
end for
Compute $\hat{Q}_i = \hat{Q}_i + \gamma^{\tau/2} r_{i,\mathcal{T}_1+\mathcal{T}_2+1}$
Return $\hat{\nabla}_i u(\theta_i, \theta_{-i}) = (1/\gamma)\hat{Q}_i \nabla_i \log \pi_\theta(a_{\mathcal{T}_1}|s_{\mathcal{T}_1})$
```

Remark 1 *The existence of random horizon sampling requires a higher degree of coordination among agents since it needs agreement over the length of horizons during the play. This issue can be solved using random number generation with pre-determined seeds in practice.*

B. Belief Exchange and Communication

Based on Eq. (6), agents need to access other agents' policy parameters to compute their gradients $\hat{\nabla}_i u_i$. In the case that this information is not perfectly available, each agent keeps estimates $\hat{\theta}^i_{-i,t}$ of other agents' parameters $\theta_{-i,t}$ by communicating with their time-varying neighbors $\mathcal{N}_{i,t} := \{j : (i,j) \in \mathcal{E}_t\}$ at time step t created by the communication networks $\mathcal{G}_t = (\mathcal{N}, \mathcal{E}_t)$, where $\mathcal{E}_t$ represents the communication links with time-varying neighbors. Agent i updates its estimate about agent j's policy parameters $\hat{\theta}^i_{j,t}$ locally as follows,

$$\hat{\theta}^i_{j,t} = \sum_{l \in \mathcal{N}_{i,t} \bigcup \{i\}} w^i_{jl,t} \hat{\theta}^l_{j,t}, \tag{9}$$

where $w_{jl,t} \geq 0$ is the weight that agent i puts on agent l's estimate of agent j's parameters at time t. We have the following assumptions on the structure of the communication network between agents.

Assumption 1 *The network $\mathcal{G} = (\mathcal{N}, \mathcal{E}_\infty)$ is connected, where $\mathcal{E}_\infty = \{(i,j)|(i,j) \in \mathcal{E}_t,$ for infinitely many $t \in \mathbb{N}\}$.*

This assumption implies the connectivity between any pair of agents after some finite t.

Assumption 2 *There exists a time step $T_B > 0$, such that for any edge $(i,j) \in \mathcal{E}_\infty$ and $t \geq 1$, it holds $(i,j) \in \bigcup_{\tau=0}^{T_B-1} \mathcal{E}_{t+\tau}$.*

This assumption ensures the any edge $(i,j) \in \mathcal{E}_\infty$ also exists in a bounded time interval T_B. Assumptions 1 and 2 are standard and named as *connectivity* and *bounded communication interval*, respectively in [25].

Assumption 3 *There exists a scalar $h \in (0,1)$, such that the following statements hold for all $i \in \mathcal{N}$, $j \in \mathcal{N}$ and $t \in \mathbb{N}^+$,*

(i) If $l \in \mathcal{N}_{i,t} \cup \{i\}$, then $w^i_{jl,t} \geq h$, else $w^i_{jl,t} = 0$,
(ii) $w^i_{ii,t} = 1$,
(iii) $\sum_{l \in \mathcal{N}_{i,t} \cup \{i\}} w^i_{jl,t} = 1$.

Assumption 3*(i)* reflects that agents only assign positive weights on their current neighbors' estimates at each time in (9). Assumption 3*(ii)* ensures that agents do not use any information from other agents on their own parameters $\hat{\theta}^i_{i,t} = \theta_{i,t}$ for all $t > 0$. Assumption 3*(iii)* implies that the construction of $N \times N$ and a row stochastic matrix weights matrix $W_{j,t}$ at any time t, associated with the updates of local estimates on agent j's parameters where $[W_{j,t}]_{i,l} = w^i_{jl,t}$.

An outline of Networked Policy Gradient Play is given in Algorithm 2.

Algorithm 2 Networked Policy Gradient Play

```
Input: Local estimates $\hat{\theta}^i_{-i,0}$ and $\mathcal{G} = (\mathcal{N}, \mathcal{E}_t)$, initial state $s_0$ and initial policy $\Pi_{\theta,0}$, and discount factor $\gamma$.
for $t = 1, 2\cdots$, do
    Run Algorithm 1 with local beliefs $\hat{\theta}^i_{-i,t}$ for $i \in \mathcal{N}$
    Update parameters (8) for $i \in \mathcal{N}$
    Update local copies $\hat{\theta}^i_{j,t}$ using (9) for $j \in -i$ and $i \in \mathcal{N}$.
end for
```

IV. Convergence of Networked Policy Gradient Play in Markov Potential Games

We introduce the assumption on the stepsize on gradient updates.

Assumption 4 (Decaying Stepsizes) *The step size α_t satisfies $\alpha_t = O(1/t)$.*

It is a common assumption in optimization literature for the analysis of the stochastic first-order methods [26]. The order of step-size satisfies divergent infinite sum of step-sizes, whereas the infinite sum of its square becomes convergent. The next set of assumptions enforces the following regularity conditions on rewards and policy functions.

Assumption 5 (Bounded Rewards) *The absolute value of rewards for any agent i at any state and joint action $(s, a) \in \mathcal{S} \times \mathcal{A}^N$ at any time $t \in \mathbb{N}^+$ is bounded, $|r_{i,t}(s, a)| \le R$ where $R > 0$.*

Assumption 6 *The gradient of log-policy of agent $n \in \mathcal{N}$, $\nabla_i \log \pi_{n,\theta}$ with respect to agent i' parameters exists and is bounded, $||\nabla_i \log \pi_{n,\theta}(a_i|s)|| \le B$ for any $\theta \in \mathbb{R}^M$, state $s \in \mathcal{S}$ and action $a_i \in \mathcal{A}_i$, where $B \ge 0$. Furthermore, it is Lipschitz continuous, i.e., $||\nabla_i \log \pi_{n,\theta_1}(a_i|s) - \nabla_i \log \pi_{n,\theta^2}(a_i|s)|| \le \mathcal{L}||\theta^1 - \theta^2||$ for any $n \in \mathcal{N}$ and $\theta^1, \theta^2 \in \mathbb{R}$, where $\mathcal{L} \ge 0$.*

These assumptions are commonly used to show that value functions and their gradients are bounded and Lipschitz continuous.

Lemma 2 (Lipschitz-Continuity of Policy Gradients) *Suppose Assumptions 5-6 hold. The policy gradient of any agent $i \in \mathcal{N}$, $\nabla_i u_i(\theta_i, \theta_{-i})$ is Lipschitz continuous with the constant $L > 0$, i.e., for any $\theta_i^1, \theta_i^2 \in \mathbb{R}^M$, defined as below,*

$$||\nabla_i u_i(\theta_i^1, \theta_{-i}^1) - \nabla_i u_i(\theta_i^2, \theta_{-i}^2)|| \le L||\theta^1 - \theta^2||, \quad (10)$$

where the value of the Lipschitz constant L is defined as,

$$L := NR\bigg(\frac{1}{(1-\gamma^2)}\mathcal{L} + \frac{(1+\gamma)B^2}{(1-\gamma)^3}\bigg). \quad (11)$$

With Assumptions 5-6, the proof depends on the change of orders between expectation operators and the sum of rewards by Fubini's Theorem. Then, the proof utilizes Taylor expansion of the difference between two discounted state-action probability distributions defined by two arbitrary parameter vectors $\theta^1 \in \mathbb{R}^M$ and $\theta^2 \in \mathbb{R}^M$. The result implies that the local gradient estimations with local beliefs are close to the unbiased estimation as long as local beliefs are close enough to their true values, along with the following result.

Lemma 3 (Unbiased and Bounded Stochastic Gradients) *Suppose Assumptions 5-6 hold. The stochastic estimate $\hat{\nabla}_i u_i(\theta_i, \theta_{-i})$ of policy gradient $\nabla u_i(\theta_i, \theta_{-i})$ of any agent $i \in \mathcal{N}$ is unbiased and bounded. $\mathbb{E}_{\mathcal{T}_1,\mathcal{T}_2}[\hat{\nabla}_i u_i(\theta_i, \theta_{-i})|\theta] = \nabla_i u_i(\theta_i, \theta_{-i})$ and $||\hat{\nabla}_i u_i(\theta_i, \theta_{-i})|| \le \hat{l}$ where $\hat{l} := \frac{NBR}{(1-\gamma)(1-\gamma)^{1/2}}$, where $\mathcal{T}_1$ and $\mathcal{T}_2$ are the random horizon lengths generated in Algorithm 1.*

The proof uses the fact that the policy gradient is the product of the Q_i-function and the gradient of log-policy function for each agent $i \in \mathcal{N}$ (Lemma 1). The proof concludes the result by showing that the estimations of each part computed as per Algorithm 1 is unbiased. This lemma ensures that agents update their parameters with the true ascent direction in expectation, together with the next result that the local beliefs converge to true parameters.

Lemma 4 (Consensus on Parameters) *Suppose Assumptions 1-6 hold. If $\hat{\theta}^i_{j0} = \theta_{j0}$ is satisfied for any pair of agents $(i, j) \in \mathcal{N} \times \mathcal{N} \setminus \{i\}$, then local copy $\hat{\theta}^i_{j,t}$ converges to $\theta_{j,t}$ with the rate $O(\log t/t)$ on expectation, i.e. $\mathbb{E}(||\hat{\theta}^i_{j,t} - \theta_{j,t}||) = O(\log t/t)$.*

The result follows from the analysis of [27] using the fact that stochastic gradients are bounded. Provided the results of Lemma 2-4, we now state the lower bound on the change in potential values of the game in two consecutive time steps.

Lemma 5 ([28], Lemma 7) *Suppose Assumptions 1-6 hold. The potential function $u : \mathbb{R}^M \to \mathbb{R}$ has the following relation between any consecutive time steps t and $t + 1$,*

$$\mathbb{E}_{\mathcal{T}_{1,t},\mathcal{T}_{2,t}}[u(\theta_{t+1})|\theta_t,] - u(\theta_t) \ge \alpha_t||\nabla u(\theta_t)||^2 - O(\log t/t^2), \quad (12)$$

where $\mathbb{E}_{\mathcal{T}_{1,t},\mathcal{T}_{2,t}}[.|\theta_t]$ is the expectation over the variables $\mathcal{T}_{1,t}, \mathcal{T}_{2,t}$ that are the lengths of random horizons generated at time step t, given the parameters θ_t, and local beliefs $\hat{\theta}_{-i}$ for each agent i at time t.

Lemma 5 relies on the unbiasedness of estimations and consensus on parameters together with Lipschitz continuity of the gradients. We use the bound on the potential change to prove the asymptotic convergence of the gradients.

Theorem 1 (Convergence of Gradients in Probability) *Suppose Assumptions 1-6 hold. For any $T \in \mathbb{N}^+$, let $\mathbb{T} = \{1, \cdots, t, \cdots, T\}$. If a time index $t \in \mathbb{N}$ is randomly chosen from the set $\mathbb{T}$ with the probabilities $\mathbb{P}(t = t') \propto \alpha_{t'}$, then the norm of the gradient $||\nabla u(\theta_{t'})||$ converges to 0 in probability as $T \to \infty$, i.e. the parameters $\{\theta_t\}_{t\ge 0}$ of networked policies converge to a stationary of the potential function in probability.*

Proof: Summing both sides (12) for the iterations $\{1, \cdots, T\}$ and using the fact that rewards are bounded by u_{sup} (Assumption 5), we have a supremum value u_{sup}

that bounds any value of the potential function satisfies the following inequality,

$$u_{sup} - \mathbb{E}[u(\theta_1)] \geq \mathbb{E}[u(\theta_{T+1})] - \mathbb{E}[u(\theta_1)] \tag{13}$$

$$\geq \sum_{t=1}^{T} (\alpha_t ||\nabla_{\theta_t} u(\theta_t)||^2 - O(\log t/t^2)). \tag{14}$$

Then, it yields by rearranging,

$$\sum_{t=1}^{T} \alpha_t ||\nabla_{\theta_t} u(\theta_t)||^2 \leq u_{sup} - \mathbb{E}[u(\theta_1)] + \sum_{t=1}^{T} O(\log t/t^2). \tag{15}$$

As $T \to \infty$ the sum on the right-hand side converges to a finite value and, the difference is $u_{sup} - \mathbb{E}[u(\theta_1)]$ is also bounded as being the difference between geometric sums of bounded rewards (Assumption 5). Therefore, the left-hand side of the inequality is bounded from the upper side. Since $\alpha_t = O(1/t)$, as $T \to \infty$, the sum $A_T = \sum_{t=1}^{T} \alpha_t$ diverges, i.e. $A_T \to \infty$, and the following result holds below,

$$\lim_{T \to \infty} \mathbb{E}[\frac{1}{A_T} \sum_{t=1}^{T} \alpha_t ||\nabla u(\theta_t)||^2] = 0. \tag{16}$$

For any $\epsilon > 0$, it holds, by Markov inequality,

$$\lim_{T \to \infty} \mathbb{P}(||\nabla u(\theta_{t'})|| \geq \epsilon) = \lim_{T \to \infty} \mathbb{P}(||\nabla u(\theta_{t'})||^2 \geq \epsilon^2) \tag{17}$$

$$\leq \lim_{T \to \infty} \epsilon^{-2} \mathbb{E}[\mathbb{E}_t[||\nabla u(\theta_t)||^2]] \tag{18}$$

$$= \lim_{T \to \infty} \epsilon^{-2} \mathbb{E}[\frac{1}{A_t} \sum_{t=1}^{T} \alpha_t ||\nabla u(\theta_t)||^2] = 0 \tag{19}$$

■

The stationary points of the potential value function are approximate-NE when no assumptions are made on the structure of the potential function or the policy structure. If the potential value function is convex and/or the policy functions of agents are direct/softmax parametrization in a tabular form with a finite number of state-action pairs, then this result is equivalent to the convergence to NE in probability.

V. Numerical Experiments

We use the Lake game which is shown to be a Markov potential game with open-loop policies [29]. Each agent $i \in \mathcal{N}$ decides on phosphorus rate $a_{i,t} \in (0, 1)$, around a lake, with the given state transition dynamics,

$$s_t = b s_{t-1} + \frac{s_{t-1}^c}{s_{t-1}^c + 1} + \sum_{i \in \mathcal{N}} a_{i,t-1}, \tag{20}$$

where b and c are positive constants. The reward of each agent i increases with the logarithmic rate of phosphorous usage and observes a quadratic rate of decrease in the phosphorus level,

$$r_{i,t} = c_r(\log(d a_{i,t}) - s_t^2), \tag{21}$$

where we also scaled the reward values by the constant $c_r = 10^{-4}$ to obtain better numerical stability, and use $d = 10^2$

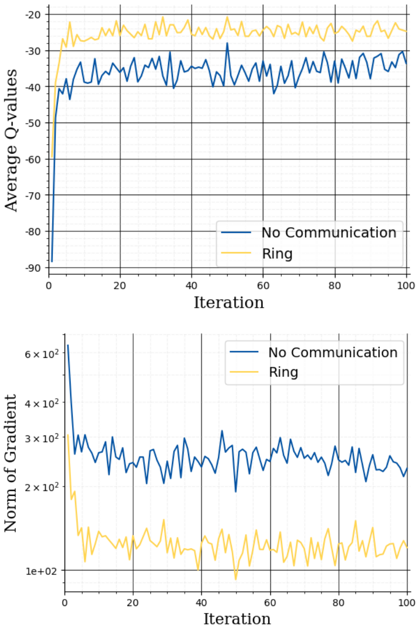

Fig. 1. Networked vs. independent policy gradient in lake game over 100 runs. (Top) Average Q-values of agents $\frac{1}{N} \sum_{i \in \mathcal{N}} \hat{Q}_{i,t}$, (Bottom) Average local gradients $||\nabla_i u_i(.)||$

as the action coefficient. We experiment with $N = 5$ agents, and game parameters $b = 0.4$, $c = 2$.

We utilize logit-normal distribution to let agents map the unconstrained parameters to a bounded interval between 0 and 1. Logit-normal distribution is a continuous probability distribution where the sampled values from a normal distribution are transformed via the sigmoid function. Note that moments of logit-normal distribution exist while they are not analytically computable. Hence, agents use logit-normal distribution Logit-Normal$(\mu_{i,\theta}, I)$ where I is an identity covariance matrix and μ_i is the parametrized mean of the corresponding normal distribution. We first derived the policy function for a no-communication scheme where agents only learn from their rewards values and the independent policy function has the following form,

$$\mu_{i,\theta} = \theta_i s - \lambda_1 \theta_i + c_\mu, \tag{22}$$

where c_μ is a constant set to 10^{-6} together with the penalty term $-\lambda_1 \theta_i$ given $\lambda_1 = 1$. We employ these penalty and constant terms to have stabilized behavior in addition to the scaled rewards. Similarly, we can define the networked policies as below,

$$\mu_{i,\theta} = \theta_i s - \lambda_2 \max(0, \theta_i - \sum_{j \in \mathcal{N} \setminus \{i\}} \hat{\theta}_j) + c_\mu, \tag{23}$$

where $\lambda_2 = 10$ and $c_\mu = 10^{-6}$. The main difference between the two policies is that the networked policy (23) includes others' parameters and penalizes the individual policies when an individual policy parameter value θ_i is higher than the average others' policy parameters.

We initialize policy parameters with uniformly sampled random values between $[0, 0.5]$ and we set the initial state variable $s_0 = 1$. We chose the discount factor and stepsize as $\gamma = 0.99$ and $\alpha_t = (10^{-3}/t)$ in order for both sets of experiments. Agents communicate over a ring network given weights on local beliefs $w^i_{i,l} = 0.30$, and remaining weights on information received from neighbors equally distributed, i.e., $w^i_{j,l} = 0.70/|\mathcal{N}_i|$ for all $j \in \mathcal{N}_i$.

Over 100 runs with 100 iterations, we eliminated the instances in which agents gain accumulated rewards (Q-functions sampled as in Algorithm 1) less than -50 on average in 10 last iterations. In these instances, the parameters did not converge to an acceptable region of parameters due to the extreme behavior of at least one agent. These extreme behaviors may stem from different factors. First, unbiased estimation of policy gradients seems to lead to higher variance due to the fact that the score function is computed at a randomly chosen point, and also the episode lengths are randomly distributed. The number of removed cases for independent policies is 42, whereas this number is reduced only to 9 cases for networked policies.

Networked policies yield a better solution on average by providing an opportunity for coordination among agents.Fig. 1 (Top) indicates the average of estimated $\hat{Q}_i$ over 100 runs. We see that the average accumulated reward of the system is better with networked policies. Fig. 1 (Bottom) shows that the convergence of local gradients. We see that the average gradient values converge to smaller values for networked policies.

VI. Conclusion

We develop a policy gradient algorithm for Markov potential games with a novel policy function structure where the policies depend on the joint parameters. The proposed algorithm has novel joint policy structure structure and novel parameter exchange protocol over time-varying networks. We show the convergence to a stationary point of potential function with respect to joint policy parameters using the results of unbiased gradients and consensus on local beliefs. Numerical results show there is a performance gain when networked policies are considered compared to independent policies, in terms of average obtained rewards and norms of estimated gradients.

References

[1] J. Jiang, C. Dun, T. Huang, and Z. Lu, "Graph convolutional reinforcement learning," *arXiv preprint arXiv:1810.09202*, 2018.

[2] C. Zhang and V. Lesser, "Coordinated multi-agent reinforcement learning in networked distributed pomdps," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, 2011, pp. 764–770.

[3] C. Zhu, M. Dastani, and S. Wang, "A survey of multi-agent reinforcement learning with communication," *arXiv preprint arXiv:2203.08975*, 2022.

[4] F. Mason, F. Chiariotti, A. Zanella, and P. Popovski, "Multi-agent reinforcement learning for pragmatic communication and control," *arXiv preprint arXiv:2302.14399*, 2023.

[5] D. Kim, S. Moon, D. Hostallero, W. J. Kang, T. Lee, K. Son, and Y. Yi, "Learning to schedule communication in multi-agent reinforcement learning," *arXiv preprint arXiv:1902.01554*, 2019.

[6] S. Gupta, R. Hazra, and A. Dukkipati, "Networked multi-agent reinforcement learning with emergent communication," *arXiv preprint arXiv:2004.02780*, 2020.

[7] Y. Jia, M. Bhatt, and N. Mehr, "Rapid: Autonomous multi-agent racing using constrained potential dynamic games," *arXiv preprint arXiv:2305.00579*, 2023.

[8] A. Muralidharan and Y. Mostofi, "Path planning for minimizing the expected cost until success," *IEEE Transactions on Robotics*, vol. 35, no. 2, pp. 466–481, 2019.

[9] D. Narasimha, K. Lee, D. Kalathil, and S. Shakkottai, "Multi-agent learning via markov potential games in marketplaces for distributed energy resources," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 6350–6357.

[10] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.

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

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

[13] S. V. Macua, J. Zazo, and S. Zazo, "Learning parametric closed-loop policies for markov potential games," *arXiv preprint arXiv:1802.00899*, 2018.

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

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

[16] D. Ding, C.-Y. Wei, K. Zhang, and M. Jovanovic, "Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence," in *International Conference on Machine Learning*. PMLR, 2022, pp. 5166–5220.

[17] A. Giannou, K. Lotidis, P. Mertikopoulos, and E.-V. Vlatakis-Gkaragkounis, "On the convergence of policy gradient methods to nash equilibria in general stochastic games," *arXiv preprint arXiv:2210.08857*, 2022.

[18] R. Fox, S. M. Mcaleer, W. Overman, and I. Panageas, "Independent natural policy gradient always converges in markov potential games," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 4414–4425.

[19] W. Mao, L. Yang, K. Zhang, and T. Basar, "On improving model-free algorithms for decentralized multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 15 007–15 049.

[20] Z. Zhou, Z. Chen, Y. Lin, and A. Wierman, "Convergence rates for localized actor-critic in networked markov potential games," *arXiv preprint arXiv:2303.04865*, 2023.

[21] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[22] D. Monderer and L. S. Shapley, "Potential games," *Games and economic behavior*, vol. 14, no. 1, pp. 124–143, 1996.

[23] S. Aydın and C. Eksin, "Networked policy gradient play in markov potential games," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] K. Zhang, A. Koppel, H. Zhu, and T. Basar, "Global convergence of policy gradient methods to (almost) locally optimal policies," *SIAM Journal on Control and Optimization*, vol. 58, no. 6, pp. 3586–3612, 2020.

[25] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multiagent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, 2009.

[26] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

[27] S. Arefizadeh and C. Eksin, "Distributed fictitious play in potential games with time-varying communication networks," *arXiv preprint arXiv:1912.03592*, 2019.

[28] S. Aydin and C. Eksin, "Policy gradient play with networked agents in markov potential games," in *Learning for Dynamics and Control Conference*. PMLR, 2023, pp. 184–195.

[29] W. D. Dechert and S. O'Donnell, "The stochastic lake game: A numerical solution," *Journal of Economic Dynamics and Control*, vol. 30, no. 9-10, pp. 1569–1587, 2006.