Policy Gradient Play with Networked Agents in Markov Potential Games

Sarper Aydın SARPER.AYDIN@TAMU.EDU
Texas A&M University, College Station, TX, USA

Ceyhun Eksin EKSINC@TAMU.EDU
Texas A&M University, College Station, TX, USA

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

We introduce a distributed policy gradient play algorithm with networked agents playing Markov potential games. Agents have rewards at each stage of the game, that depend on the joint actions of agents given a common dynamic state. Agents implement parameterized and differentiable policies to take actions against each other. Markov potential games assume the existence of potential value functions. In a differentiable Markov potential game, partial gradients of a potential function are equal to the local gradients with respect to the individual parameters. In this work, agents receive information on other agents' parameters via a communication network in addition to rewards. Agents then use stochastic gradients with respect to local estimates of joint policy parameters to update their policy parameters. We show that agents' joint policy converges to a first-order stationary point of Markov potential value function with any type of function approximation, state and action spaces. Numerical experiments confirm the convergence result in the lake game, a Markov potential game.

Keywords: Game theory, reinforcement learning, distributed algorithms

1. Introduction

Multiple agents learn and adapt their actions in dynamic states to optimize their utilities using multi-agent reinforcement learning (MARL) algorithms without explicitly knowing the analytical structure of their rewards and dynamic state transitions (Zhang et al. (2021a)). Agents learn how to take actions in networked MARL, using the information gathered through communication (Zhu et al. (2022)). Many real-life applications such as autonomous driving (Shalev-Shwartz et al. (2016)), electric vehicles (Qiu et al. (2022)), and power grids (Hu et al. (2022)) possess a competitive multi-agent nature where agents obtain rewards and transition to next state as the result of joint actions taken. Markov games represent the framework for the competitive (MARL) algorithms where agents select their actions to gain more rewards while their rewards and state changes are determined by joint actions taken (Littman (1994)). In this study, we address and propose a new algorithm to solve Markov potential games as a subclass of Markov games. They admit a potential value function mirroring individual value function changes by one-sided policy updates against fixed policies of other agents.

Our algorithm is built upon single-agent policy gradient algorithms (Williams (1992); Sutton et al. (1999)). Agents iteratively play through episodes to estimate gradients given their parametrized policies. They update their parameters with stochastic gradients of their value functions. We define and introduce a novel version of policy gradient play where agents implement policy functions con-

sidering other agents' parameters. More specifically, agents assign parameters to others' policies in addition to state variables. Agents sample actions from their policies and observe their rewards together with the next state during two different episodes with randomly generated horizon lengths. They need to retrieve the parameters of all other agents, which may not be possible in reality. They alternatively store their local estimates and update them with signals coming from their neighbors. We prove that the joint policy of agents converges to a first-order stationary point of the potential function (Theorem 8). This result relies on the properties that the estimation procedure with random horizons provides unbiased policy gradient estimates (Lemma 5), and local estimates on others' parameters converge to the correct values of others' parameters (Lemma 6).

Early studies on policy gradient play in Markov potential games consider continuous state and action spaces, assuming that rewards and state transition probabilities are known (González-Sánchez and Hernández-Lerma (2013); Macua et al. (2018)). More recent works study direct or softmax parameterization of finite state and action problems. They design independent policy gradients among agents with several variants of first-order methods such as projected and natural gradients in addition to the standard version of gradient ascent (Zhang et al. (2021b); Leonardos et al. (2021); Ding et al. (2022); Mguni et al. (2021); Giannou et al. (2022); Chen et al. (2022); Fox et al. (2022); Mao et al. (2022)). Our analysis generalizes the setting of these recent studies to continuous state and action pairs given unknown rewards and state transition dynamics. Policies that incorporate other agents' policy parameters and local exchange information are additional features of the proposed algorithm that distinguish it from existing MARL algorithms. Numerical experiments on the lake game (a Markov potential game Dechert and O'Donnell (2006)) demonstrate the convergence of the proposed networked policy gradient play.

2. Markov Potential Games

Agents defined by the set $\mathcal{N} := \{1, \dots, N\}$ play a Markov game (Shapley (1953)), and each agent $i \in \mathcal{N}$ takes an action $a_i \in \mathcal{A}_i$ at a common state $s \in \mathcal{S}$. The joint action profile is formed by individual actions, $a = (a_1, a_2, \cdots, a_N) \in \mathcal{A}^N := \times_{i\in\mathcal{N}} \mathcal{A}_i$. Note that the sets of actions and states are not necessarily finite. The joint actions and the state generate a conditional probability transition to a next state, $\mathcal{P}^a_{s'',s'} = \mathbb{P}(s''|s', a)$, whereas initial state s_0 is also distributed with $\rho : \mathcal{S} \to [0, 1]$. Each agent obtains a reward $r_{i,t} : \mathcal{S} \times \mathcal{A}^N \to \mathbb{R}$ as the result of the joint actions and the state $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}^N$ at time $t \in \mathbb{N}$. Rewards are accumulated over an infinite horizon by the discount rate $\gamma \in (0, 1)$. A Markov game is formally defined by the tuple $\Gamma := (\mathcal{N}, \mathcal{A}^N, \mathcal{S}, \{r_i\}_{i\in\mathcal{N}}, \mathcal{P}, \gamma, \rho)$.

Each agent takes an action sampled from a policy function $\pi_i : \mathcal{S} \times \pi_{-i} \to \Delta(\mathcal{A}_i)$ given a state and other agents' policies π_{-i}. $\Delta(.)$ defines any probability distribution over the given set, and $-i := \mathcal{N}\backslash\{i\}$ denotes the set of all agents except agent i. The rewards $r_{i,t}$ at each time step $t \in \mathbb{N}$, induced by the joint policy $\Pi = \times_{i\in\mathcal{N}} \pi_i$, constitute individual value functions $V_i^\Pi : \mathcal{S} \to \mathbb{R}$, as a discounted sum of rewards over infinite horizon starting from each state,

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

where the sequence of states and actions generated by the joint policy is distributed with $\mathcal{P}$ [1]. Similarly, the Q-function of agent i, $Q_i : \mathcal{S} \times \mathcal{A}^N \to \mathbb{R}$ is defined as a discounted sum of rewards

1. We remove the distribution notation from the expectations in the rest of the paper for simplicity, unless necessary.

starting given each state $s \in \mathcal{S}$, and joint action $a \in \mathcal{A}^N$ sampled from the joint policy Π as below,

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

Potential games are an important class of games with a potential function expressing the change in individual utilities based on unilateral action changes when other agents' actions are fixed (Monderer and Shapley (1996)). Markov potential games hold the same property in the setting of Markov games, and have a potential value function capturing value function changes of agents when a unilateral change in policies occurs.

Definition 1 (Markov Potential Games) *A game Γ is a Markov potential game, if there exist a $r_t \in \mathbb{R}$ and a potential value function $V^{\Pi}(s) : \Pi \times \mathcal{S} \to \mathbb{R}$ that is equal to the discounted sum of the 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 $i \in \mathcal{N}$*

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

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

We suppose that agents use parametrized joint policies by unconstrained and continuous variables $\theta = (\theta_i, \theta_{-i}) \in \mathbb{R}^M$ where individual policy parameters $\theta_i \in R^{K_i}$ are such that it holds $\sum_{i\in\mathcal{N}} K_i = M$.

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

where Π_θ is the joint policy parametrized by the parameters $\theta = (\theta_i, \theta_{-i}) \in \mathbb{R}^M$. Differentiable Markov potential games are equivalently defined 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 $u : \mathbb{R}^M \to \mathbb{R}$ such the following holds,*

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

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.

Differentiable Markov potential games are the natural extensions of the standard definition of potential games (Monderer and Shapley (1996)) with discrete actions.

3. Policy Gradient Play with Networked Agents

Each agent i's policy $\pi_{i,\theta}(a_i|s)$ is conditionally independent given the state and joint policy parameters $\theta = (\theta_i, \theta_{-i})$,

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

where $\mathcal{A}_q^N = \times_{i\in\mathcal{N}} \mathcal{A}_{i,q}$ and $\mathcal{A}_{i,q}$ are countable measurable subsets over the joint and individual set of actions respectively such that probability distributions can be defined, i.e, $\mathcal{A}_i = \bigcup_{q=1}^{\infty} \mathcal{A}_{i,q}$. Each agent aims to maximize its value function as the cumulative reward in the long term given the joint actions and state transition dynamics.

The gradient of agent i's value function is defined in terms of the Q-function and sum of log-policies—see Section 7 for proof.

Lemma 3 *Given the parameterized value functions $u_i : \mathbb{R}^M \to \mathbb{R}$, the gradient of each value function u_i with respect to agent i's parameters θ_i is equal to,*

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

For this aim, agents are assumed to use their gradient information iteratively against each other, named policy gradient play. Each agent computes 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 α_t is a common step size (for the sake of simplicity), and $\hat{\nabla}_i u_i(\theta_{i,t-1}, \theta_{-i,t-1})$ is the stochastic gradient computed based on estimated rewards collected on a roll-out horizon (episode).

As per the gradient definition (8), the individual policies π_i and their stochastic gradients $\hat{\nabla}_i u_i$ depend on joint policy parameters $\theta = (\theta_i, \theta_{-i})$. However, other agents' parameters θ_{-i} may not be available to agent i. In this setting, agent i keeps an estimate of other agents' policy parameters with the information received from its neighbors $\mathcal{N}_i := \{j : (i,j) \in \mathcal{E}\}$ in the communication network $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. Agent i updates its local estimate on agent j's parameters $\hat{\theta}_{j,t}^i$ with the the weights $w_{j,l}^i \geqslant 0$ that agent i gives agent l's estimate on agent j's parameters,

$$\hat{\theta}_{j,t}^i = \sum_{l\in\mathcal{N}_i \bigcup \{i\}} w_{j,l}^i \hat{\theta}_{j,t}^l, \tag{9}$$

We assume that the communication network and the weights satisfy the following properties,

Assumption 1 *The network $\mathcal{G}$ is strongly connected with weights satisfying* **a)** $w_{j,l}^i \geqslant \upsilon$ *for* $\upsilon > 0$ *only if* $l \in \mathcal{N}_i \cup \{i\}$, *otherwise* $w_{j,l}^i = 0$, **b)** $w_{i,i}^i = 1$, *and* **c)** $\sum_{l\in\mathcal{N}_i \bigcup \{i\}} w_{j,l}^i = 1$ *for all* i, j.

We implement the estimation of the gradient $\nabla_i u_i$ in (8) by the adaptation of the random horizon sampling method as outlined in (Zhang et al. (2020)). Agents play together during two episodes whose lengths are randomly sampled from geometric distributions where the Q-values Q_i and gradient of log-policy $\nabla_i \log \pi_\theta$, are estimated. The sequential actions and states create a bias for the estimation of gradients in deterministic episode lengths for discounted rewards over infinite horizons. The episode lengths $\mathcal{T}_1$ and $\mathcal{T}_2$ are generated from a 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 obtain unbiased estimates (see Lemma 5). The random horizons at each time step require coordination among agents compared to the episodes with constant lengths. This issue can be solved via preset random seeds to ensure that agents use the same samples over time. The steps of the networked policy gradient play with the episodes and updates are provided in Algorithm 1.

Algorithm 1 Networked Policy Gradient

Input: Local estimates $\hat{\theta}^i_{-i,0}$ and $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, initial state s_0 and initial policy $\Pi_{\theta,0}$, and discount factor γ.
for $t = 1, 2, \cdots$ **do**
  Draw $\mathcal{T}_1 \sim Geom(1-\gamma^{0.5})$ and reset s_0, for all $i \in \mathcal{N}$..
  Each agent sample actions $a_{i,0} \sim \pi_{i,\hat{\theta}_{t-1}}(.|s_0)$
  for $\tau = 1, 2 \cdots, \mathcal{T}_1$ **do**
    Each agent i reaches state $s_\tau \sim \mathcal{P}^{a_\tau}_{s_\tau, s_{\tau-1}}$.
    Each agent i samples and takes actions, $a_{i,\tau+1} \sim \pi_{i,\hat{\theta}_{t-1}}(.|s_{\tau+1})$
  end for
  Each agent i computes $\nabla_i \log \pi_{n,\theta}(a_{\mathcal{T}_1+1}|s_{\mathcal{T}_1+1})$ for all $i \in \mathcal{N}$.
  Draw $\mathcal{T}_2 \sim Geom(1-\gamma^{0.5})$ and set $\hat{Q}_i = 0$, for all $i \in \mathcal{N}$.
  for $\tau = 1, 2, \cdots, \mathcal{T}_2$ **do**
    Each agent receives rewards $r_{i,\tau+\mathcal{T}_1}$ for all $i \in \mathcal{N}$.
    Each agent collects rewards $\hat{Q}_i = \hat{Q}_i + \gamma^{\tau/2} r_{i,\tau+\mathcal{T}_1}$ for $i \in \mathcal{N}$.
    Each agent i reaches state $s_{\tau+\mathcal{T}_1+1} \sim \mathcal{P}^{a_\tau}_{s_{\tau+\mathcal{T}_1+1}, s_{\tau+\mathcal{T}_1}}$.
    Each agent i samples and takes actions $a_{i,\tau+\mathcal{T}_1+1} \sim \pi_{i,\hat{\theta}_{t-1}}(.|s_{\tau+\mathcal{T}_1})$
  end for
  Each agent i computes $\hat{Q}_i = \hat{Q}_i + \gamma^{\tau/2} r_{i,\mathcal{T}_1+\mathcal{T}_2+1}$ for $i \in \mathcal{N}$
  Each agent i computes stochastic gradients by replacing Q_i with $\hat{Q}_i$ and computing $\nabla_i \log \pi_{n,\theta}$ at the joint state-action pair $(a_{\mathcal{T}_1+1}, s_{\mathcal{T}_1+1})$ for the corresponding terms in (7).
  Each agent i updates parameters (8) with $\theta_{-i,t-1}$ replaced by $\hat{\theta}_{-i,t-1}$.
  Each agent i updates local estimates $\hat{\theta}^i_{j,t}$ using (9).
end for

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

We state the assumption on the gradient step size.

Assumption 2 *The step size α_t satisfies $\alpha = O(1/t)$.*

This assumption is standard for the analysis of stochastic gradient algorithms. It assures the square summable but not summable step-sizes. We also have the following assumptions on rewards and policies.

Assumption 3 *The absolute value of rewards for any agent i at any state and joint action $(s, a) \in \mathcal{S} \times \mathcal{A}^N$ is bounded, $|r_{i,t}(s, a)| \leqslant R$ and where $R > 0$.*

Assumption 4 *The gradient of log-policy of agent $i \in \mathcal{N}$, $\nabla_i \log \pi_{i,\theta}$ exists and its norm is bounded, $||\nabla_i \log \pi_{i,\theta}|| \leqslant B$ for any θ, state $s \in \mathcal{S}$ and action $a_i \in \mathcal{A}_i$, where $B \geqslant 0$. Furthermore, it is Lipschitz continuous, i.e., $||\nabla_i \log \pi_{i,\theta_1} - \nabla_i \log \pi_{i,\theta_2}|| \leqslant \mathcal{L}||\theta^1 - \theta^2||$ for any $\theta_1, \theta_2 \in \mathbb{R}^M$, where $\mathcal{L} > 0$.*

These assumptions are standard to show that (stochastic) gradients are bounded and Lipschitz continuous.

Lemma 4 (Lipschitz-Continuity of Networked Policy Gradient) *Suppose Assumptions 3-4 hold. The policy gradient of any agent $i \in \mathcal{N}$, $\nabla_i u_i(\theta_i, \theta_{-i})$ is Lipschitz continuous with some constant $L > 0$, i.e., for any $\theta_i^1, \theta_i^2 \in \mathbb{R}^d$*

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

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

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

The proof relies on the exchange of the order of summations and integrals by the Fubini Theorem, and the Lipschitz continuity follows after using the Taylor expansion.

Lemma 5 (Unbiased and Bounded Stochastic Gradient of Agents) *The stochastic gradient $\hat{\nabla}_i u_i(\theta_i, \theta_{-i})$ to estimate the policy gradient in (7) is unbiased and its norm is bounded for all $i \in \mathcal{N}$ and for any $\theta \in \mathbb{R}^M$, i.e. $\mathbb{E}(\hat{\nabla}_i u_i(\theta_i, \theta_{-i})) = \nabla_i u_i(\theta_i, \theta_{-i})$ and $||\hat{\nabla}_i u_i(\theta_i, \theta_{-i})|| \leqslant \hat{\lambda}$ where $\hat{\lambda} > 0$.*

The result follows from Zhang et al. (2020) by the fact that the rewards are collected with special discount rates $\gamma^{\tau/2}$. This assures that agents have unbiased estimates of their Q-values. Using two independent identically sampled random horizon lengths $\mathcal{T}_1$ and $\mathcal{T}_2$ (steps 13 and 17 in Algorithm 1), it can be shown that the term $\sum_{n\in\mathcal{N}} \nabla_i \log \pi_{n,\theta}(a_n|s)$ is unbiasedly estimated, thus giving the unbiased stochastic gradients. Moreover, the estimates $\hat{Q}_i$ of Q_i-functions for any $i \in \mathcal{N}$ as the rewards collected with the discount rate $\gamma^{\tau/2}$ over a random horizon is still bounded by the fact $\gamma^{\tau/2} \in (0,1)$ given $\gamma \in (0,1)$ for $\tau \in \mathbb{N}$ and Assumption 3. Again by Assumption 4, the gradient of log gradient is also bounded, which certifies the boundedness of stochastic gradients. This lemma assures that agents update their parameters with the correct gradient direction in expectation. Together with Lemma 4, this leads to the exploitation of standard gradient ascent analysis.

Lemma 6 (Consensus on Parameters) *Suppose Assumptions 1-3 hold. If $\hat{\theta}_{j0}^i = \theta_{j0}$ is satisfied for any pair of agents $(i,j) \in \mathcal{N} \times \mathcal{N}\backslash\{i\}$, then local copies $\hat{\theta}_{j,t}^i$ converges to $\theta_{j,t}$ with the rate $O(\log t/t)$, i.e. $||\hat{\theta}_{j,t}^i - \theta_{j,t}|| = O(\log t/t)$.*

The proof is provided in Section 7. The result follows by showing that change in parameters is bounded given bounded stochastic gradients (Lemma 5). When the step size is such that $\alpha_t = O(1/t)$, the change in parameters is slowed down while agents' estimates about others' policy parameters continue to be updated according to (9) given weights that form a row stochastic weights.

Lemma 7 *The potential function $u : \mathbb{R}^M \to \mathbb{R}$ of the Markov game has the following relation between any consecutive time steps t and $t+1$ during Algorithm 1,*

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

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 lenghts of random horizons generated at time step t, given the parameters θ_t at time t.

Proof By Taylor Expansion and Lemma 4, we obtain,

$$u(\theta_{t+1}) - u(\theta_t) \geqslant (\theta_{t+1} - \theta_t)^T \nabla_{\theta_t} u(\theta_t) - \frac{1}{2}L\|\theta_{t+1} - \theta_t||^2, \tag{13}$$

$$\geqslant \alpha_t g(\theta_t)^T \nabla_{\theta_t} u(\theta_t) - \frac{1}{2}L\alpha_t^2\hat{\lambda}^2 \tag{14}$$

where $g(\theta_t) = [\hat{\nabla}_1 u_1(\theta_{1,t}, \hat{\theta}_{-1,t}), \cdots, \hat{\nabla}_N u_N(\theta_{N,t}, \hat{\theta}_{-N,t})]$ is the vector of stochastic policy gradient of each agent $i \in \mathcal{N}$ with respect to $\theta_{i,t}$ against the local copies $\hat{\theta}_{-i,t}$. By Lemma 5 and Definition 2, it holds, $\mathbb{E}[\hat{\nabla}_{\theta_{i,t}} u_i(\theta_{i,t}, \hat{\theta}_{-i,t})] = \nabla_{\theta_{i,t}} u_i(\theta_{i,t}, \hat{\theta}_{-i,t}) = \nabla_{\theta_{i,t}} u(\theta_{i,k}, \hat{\theta}_{-i,t})$,

$$\mathbb{E}_{\mathcal{T}_{1,t}, \mathcal{T}_{2,t}}[u(\theta_{t+1})|\theta_t] - u(\theta_t) \geqslant \alpha_t \mathbb{E}(g(\theta_t))^T \nabla_{\theta_t} u(\theta_t) - \frac{1}{2} L \alpha_t^2 \hat{\lambda}^2 \tag{15}$$

$$\geqslant \alpha_t \nabla_{\theta_t} u(\hat{\theta}_t)^T \nabla_{\theta_t} u(\theta_t) - \frac{1}{2} L \alpha_t^2 \hat{\lambda}^2 \tag{16}$$

$$\geqslant \alpha_t ||\nabla_{\theta_t} u(\theta_t)||^2 - \alpha_t O(\log t/t) - \frac{1}{2} L \alpha_t^2 \hat{\lambda}^2 \tag{17}$$

$$\geqslant \alpha_t ||\nabla_{\theta_t} u(\theta_t)||^2 - O(\log t/t^2). \tag{18}$$

where $\nabla_{\theta_t} u(\hat{\theta}_t) = [\nabla_1 u(\theta_{1,t}, \hat{\theta}_{-1,t}), \cdots, \nabla_N u(\theta_{N,k}, \hat{\theta}_{-N,k})]$. Then by Lemmas 4-6, it follows that $||\nabla_i u(\theta_{i,t}, \hat{\theta}_{-i,t}) - \nabla_i u(\theta_{i,t}, \theta_{-i,t})|| = O(\log t/t)$. Hence, it also holds, $||u(\theta_t) - \nabla_{\theta_t} u(\theta_t)|| = O(\log t/t)$. Thus, the result follows by the fact that $\alpha_t = O(1/t)$ and Lemma 5. ■

We use the obtained lower bound on the potential change in expectation to show convergence of the gradients.

Theorem 8 *Let $\{\theta_t\}_{t \geqslant 1}$ be the sequence of policy parameters generated by Algorithm 1. Then, the policy parameters $\{\theta_t\}_{t \geqslant 1}$ converge to a first-order stationary point of the potential function in expectation,*

$$\lim_{t \to \infty} \mathbb{E}(||\nabla_{\theta_t} u(\theta_t)||^2) = 0. \tag{19}$$

Proof is omitted. The proof is mainly based on the fact that the limit sum over the product of step sizes and norm of gradients over iterations is finite implying that the gradients goes to zero, since the infinite sum of step sizes are divergent by Assumption 2. This result suggests that agents converge to a Nash equilibrium (NE), i.e., optimal behavior, for convex potential value functions. For non-convex potential value functions, a stationary point is not necessarily a NE, indicating that a stationary point is an approximate NE.

5. Numerical Experiments

We use the Lake game as an example of Markov potential games (Dechert and O'Donnell (2006))[2]. Each agent $i \in \mathcal{N}$ decides on its usage of phosphorus rate $a_{i,t} \in [0, 1]$ in a dynamic state of phosphorus level $s_t \in \mathbb{R}^+$ around a lake, with the given 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,

$$r_{i,t} = \log(d a_{i,t} + 1) - s_t^2, \tag{21}$$

where $d > 0$ and, we bound the rewards in the range of $[-4, 4]$ to satisfy Assumption 3. We experiment with $N = 5$ agents, and game parameters $b = 0.45$, $c = 2$, and $d = 10^4$. Agents use

2. The code is available at Aydin (2022).

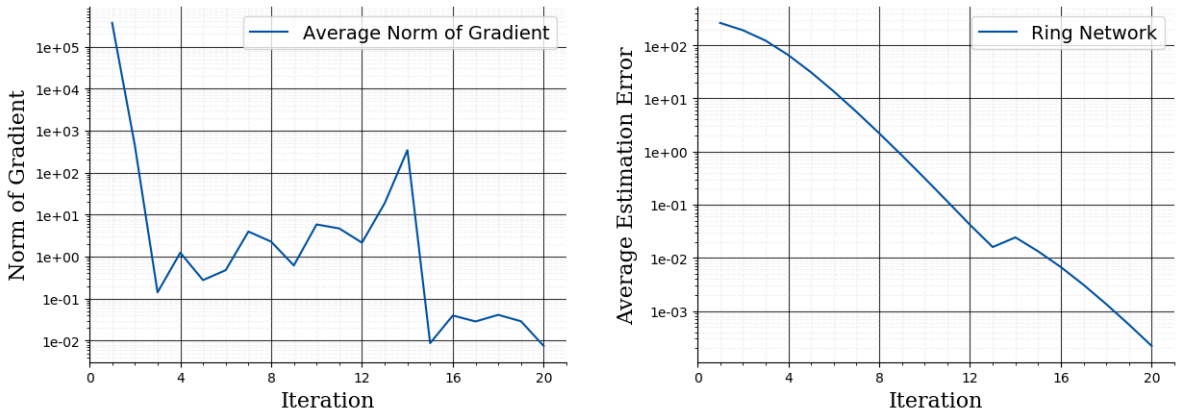

Figure 1: Networked policy gradient in lake game. (Left) Average norm of gradients of agents $\frac{1}{N}\sum_{i\in\mathcal{N}}||\nabla_i u_i(\theta_i, \hat{\theta}_{-i})||$ (Right) Average estimation error $\frac{1}{N(N-1)}\sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}\setminus\{i\}}||\theta_{i,t} - \hat{\theta}^i_{j,t}||$.

truncated normal distribution Normal$(\mu_{i,\theta}, \kappa I)$ where κI is identity covariance matrix scaled with $\kappa = 0.1$ and μ_i is the parametrized mean of the policy distribution given as,

$$\mu_{i,\theta} = \text{sigmoid}\,(\theta_{i,s}s + \theta_{i,-i}(\frac{1}{(N-1)}\sum_{j\in\mathcal{N}\setminus\{i\}}(\theta_{j,s})). \tag{22}$$

In (22), there are two policy parameters, i.e., $K_i = 2$. $\theta_{i,s} \in \mathbb{R}$ is the parameter multiplying the state and $\theta_{i,-i} \in \mathbb{R}$ multiplies the average state parameter of other agents. We note that the sigmoid function maps the unconstrained parameters $\theta \in \mathbb{R}^M$ to the range of the actions $[0, 1]$. We also utilize the unbiased estimation technique for the gradient of truncated policy distributions as outlined in (Fujita and Maeda (2018)). 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 $j \in \mathcal{N}_i$.

Fig. 1 (Left) indicates the average of individual gradients over 50 runs with random initialization converge closely to a stationary point of the value functions. Fig. 1 (Right) shows that the local beliefs on other agents' parameters converge to the true parameter values. The two observations confirm the result that the joint policies in policy gradient play converge to a stationary point.

6. Conclusion

We formulate a new class of networked policies where agents play against each other by updating their parameters with gradient information. Our algorithm has novel features in which individual policies are conditioned on others' parameters in addition to the state variables, and parameters are shared over a communication network. We prove that local beliefs on others converge to true values. The algorithm is based on random roll-out horizons which provide unbiased policy gradient estimates. We stated the convergence of the algorithm to a stationary point. We verify our results with numerical experiments.

7. Appendix

7.1. Proof of Lemma 3

We define the agents' gradients by Policy Gradient Theorem Sutton et al. (1999),

$$\nabla_i u_i(\theta_i, \theta_{-i}) = \int_{\substack{a\in\mathcal{A},\\ s\in\mathcal{S}}} Q_i^{\Pi_\theta}(s,a) d^{\Pi_\theta} \nabla_i \pi_\theta(a|s)\, da\, ds, \tag{23}$$

where $d^{\Pi_\theta} = \sum_{t=0}^{\infty} \gamma^t \rho^a_{s_0,s,t}$ is the discounted sum of density functions $\rho^a_{s_0,s,t}$ of the transition probabilities $\mathcal{P}^a_{s_0,s,t}$ from the initial state s_0 to the state s given the joint action a at t steps ahead, and similarly $\pi_\theta(a|s)$ is defined as the density function of the joint policy Π_θ. Then, we use the log-likelihood trick by dividing and multiplying the gradient of $\nabla_i \pi_\theta(a|s)$ by the density $\pi_\theta(a|s)$,

$$\nabla_i u_i(\theta_i, \theta_{-i}) = \int_{\substack{a\in\mathcal{A},\\ s\in\mathcal{S}}} Q_i^{\Pi_\theta}(s,a) d^{\Pi_\theta} \pi_\theta(a|s) \frac{\nabla_i \pi_\theta(a|s)}{\pi_\theta(a|s)}\, da\, ds \tag{24}$$

$$= \int_{\substack{a\in\mathcal{A},\\ s\in\mathcal{S}}} Q_i^{\Pi_\theta}(s,a) d^{\Pi_\theta} \pi_\theta(a|s) \nabla_i \log \pi_\theta(a|s)\, da\, ds. \tag{25}$$

We divide the integral by $(1-\gamma)$ to have a proper expectation satisfying the properties of probability measures,and the policy gradients become by the definition of networked policies (6),

$$= \int_{\substack{a\in\mathcal{A},\\ s\in\mathcal{S}}} Q_i^{\Pi_\theta}(s,a) d^{\Pi_\theta} \pi^\theta(a|s) \sum_{n\in\mathcal{N}} \nabla_i \log \pi_n^\theta(a_n|s)\, da\, ds \tag{26}$$

$$= \frac{1}{(1-\gamma)} \mathbb{E}_{(s,a)\sim\mathcal{P}}\big[Q_i^{\Pi_\theta}(s,a) \sum_{n\in\mathcal{N}} \nabla_i \log \pi_{n,\theta}(a_n|s)\big]. \tag{27}$$

7.2. Proof of Lemma 6

Firstly, we are going to provide the update of local copies as a recursion in relation with the values at the previous step. Let $\theta_{j,t}(m)$ and $\hat{\theta}_{j,t+1}(m) \in \mathbb{R}^N$ be m^{th} index of agent j's policy at time t and the local copies of m^{th} index of agent j's policy at time step $t+1$. $W_j \in \mathbb{R}^{N\times N}$ is communication weight matrix created by the values $W_j(i,l) = w^i_{jl}$. Note that by Assumption 1, W_j is a row stochastic matrix. Then, recursion can be written as follows,

$$\hat{\theta}_{j,t+1}(m) = W_j(\hat{\theta}_{j,t}(m) + (\theta_{j,t+1}(m) - \theta_{j,t}(m))e_j), \tag{28}$$

where e_j is the canonical vector of j^{th} base in $\mathbb{R}^N$. Subtracting the vector $\theta_{j,t+1}(m)\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^N$ is the vector in $\mathbb{R}^N$ whose all values are 1, from both sides of (28), we obtain,

$$\hat{\theta}_{j,t+1}(m) - \theta_{j,t+1}(m)\mathbf{1} = W_j(\hat{\theta}_{j,t}(m) + (\theta_{j,t+1}(m) - \theta_{j,t}(m))e_j - \theta_{j,t+1}(m)\mathbf{1}), \tag{29}$$

The term $\theta_{j,t+1}\mathbf{1}$ can go inside the matrix multiplication, since W_j is a row-stochastic matrix. By letting $y_t = \hat{\theta}_{j,t}(m) - \theta_{j,t}$, we rearrange the equation (29) as,

$$y_{t+1} = W_j(y_t + (\theta_{j,t+1}(m) - \theta_{j,t}(m))e_j - (\theta_{j,t+1}(m) - \theta_{j,t}(m))\mathbf{1}). \tag{30}$$

Next, we are going to derive an upper bound for the term $\delta_t = (\theta_{j,t+1}(m) - \theta_{j,t}(m))(e_j - \mathbf{1})$, by Lemma 5,

$$||(\theta_{j,t+1}(m) - \theta_{j,t}(m))(e_j - \mathbf{1})|| \leqslant ||((\theta_{j,t+1}(m) - \theta_{j,t}(m))||(||e_j|| + ||\mathbf{1}||) \tag{31}$$

$$\leqslant \alpha_t ||\hat{\nabla}_{j,m} u_j(\theta_{j,t}, \hat{\theta}_{-j,t})||(||e_j|| + ||\mathbf{1}||) \tag{32}$$

$$\leqslant \alpha_t ||\hat{\nabla}_{j,m} u_j(\theta_{j,t}, \hat{\theta}_{-j,t})||(N+1), \tag{33}$$

where $\hat{\nabla}_{\theta_{it,m}} u(\theta_{i.t}, \hat{\theta}_{-i,t})$ is the stochastic gradient of agent j' value function with respect to the parameter value at the index m. Since stochastic gradient of any agent is bounded by Lemma 5, and the step size $\alpha_t = O(1/t)$ by Assumption 2, it holds,

$$\delta_t = ||(\theta_{j,t+1}(m) - \theta_{j,t}(m))(e_j - \mathbf{1})|| = O(1/t). \tag{34}$$

Now we rewrite (30) as follows,

$$y_{t+1} = W_j(y_t + \delta_t) = \sum_{\zeta=0}^{t} W_j^{\zeta+1} \delta_{t-\zeta} + W_j^t y_1. \tag{35}$$

As it holds $\hat{\theta}^i_{j0} = \theta_{j0}$, then $y_1 = 0$. Hence, the norm $||y_{t+1}||$ can be expressed as follows,

$$||y_{t+1}|| = ||\sum_{\zeta=0}^{t} W_j^{\zeta+1} \delta_{t-\zeta}|| \leqslant \sum_{\zeta=0}^{t} ||W_j^{\zeta+1} \delta_{t-\zeta}||. \tag{36}$$

Since W_j is row stochastic matrix, it holds $\lim_{t\to\infty} W_j^t = \mathbf{1}e_j^T$. Then, define the matrix $\bar{W}_j = W_j - \mathbf{1}e_j^T$. It holds $\lim_{t\to\infty} \bar{W}_j = 0$ and $\bar{W}_j \mathbf{1} = 0$. If the equation (36) is written with the relation $\bar{W}_j = W_j - \mathbf{1}e_j^T$, then we obtain the following,

$$||y_{t+1}|| \leqslant \sum_{\zeta=0}^{t} ||(\bar{W}_j + \mathbf{1}e_j^T)^{\zeta+1} \delta_{t-\zeta}||. \tag{37}$$

Since, we have $\bar{W}_j \mathbf{1} = 0$, $e_j^T \bar{W}_j = 0$, and $(\mathbf{1}e_j^T)^\zeta = \mathbf{1}e_j^T)$ for any $\zeta \geqslant 1$ by the definition of given matrices, we rewrite the upper bound in (37),

$$||y_{t+1}|| \leqslant \sum_{\zeta=0}^{t} ||\bar{W}_j^{\zeta+1} \delta_{t-\zeta}|| + |||(\mathbf{1}e_j^T)^{\zeta+1} \delta_{t-\zeta}|| \tag{38}$$

By definition, see that $\delta_t(j) = 0$, for any t, which gives $e_j^T \delta_t = 0$. Moreover, the spectral radius λ of $\bar{W}_j$ is strictly less than 1, $\lambda < 1$. Hence, we have,

$$||y_{t+1}|| \leqslant \sum_{\zeta=0}^{t} \lambda^{\zeta+1} ||\delta_{t-\zeta}||. \tag{39}$$

By Chebychev's sum inequality, it holds,

$$||y_{t+1}|| \leqslant \delta_t^{avg} \sum_{\zeta=0}^{t} \lambda^{\zeta+1}, \tag{40}$$

where $\delta_t^{avg} = (1/t) \sum_{\zeta=0}^{t} ||\delta_{t-\zeta}||$. By using the fact $\delta_t^{avg} = (1/t) \sum_{\zeta=0}^{t} ||\delta_{t-\zeta}|| = (1/t) \sum_{\zeta=0}^{t} (\frac{n+1}{t-\zeta}) = O(\log t/t)$ and hence, we have $||y_{t+1}|| = O(\log t/t)$. Thus, it assures $\mathbb{E}(||\hat{\theta}^i_{j,t} - \theta_{j,t}|| = O(\log t/t)$ and the proof is completed.

Acknowledgments

This work was supported by NSF CCF-2008855.

References

Sarper Aydin. Code for Numerical Experiments, 2022. URL https://github.com/sarperaydin/ResearchCodes.git.

Dingyang Chen, Qi Zhang, and Thinh T Doan. Convergence and price of anarchy guarantees of the softmax policy gradient in markov potential games. *arXiv preprint arXiv:2206.07642*, 2022.

W Davis Dechert and SI O'Donnell. The stochastic lake game: A numerical solution. *Journal of Economic Dynamics and Control*, 30(9-10):1569–1587, 2006.

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pages 5166–5220. PMLR, 2022.

Roy Fox, Stephen M Mcaleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pages 4414–4425. PMLR, 2022.

Yasuhiro Fujita and Shin-ichi Maeda. Clipped action policy gradient. In *International Conference on Machine Learning*, pages 1597–1606. PMLR, 2018.

Angeliki Giannou, Kyriakos Lotidis, Panayotis Mertikopoulos, and Emmanouil-Vasileios Vlatakis-Gkaragkounis. On the convergence of policy gradient methods to nash equilibria in general stochastic games. *arXiv preprint arXiv:2210.08857*, 2022.

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

Daner Hu, Zhenhui Ye, Yuanqi Gao, Zuzhao Ye, Yonggang Peng, and Nanpeng Yu. Multi-agent deep reinforcement learning for voltage control with coordinated active and reactive power optimization. *IEEE Transactions on Smart Grid*, 2022.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

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

Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for markov potential games. *arXiv preprint arXiv:1802.00899*, 2018.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Basar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049. PMLR, 2022.

David H Mguni, Yutong Wu, Yali Du, Yaodong Yang, Ziyi Wang, Minne Li, Ying Wen, Joel Jennings, and Jun Wang. Learning in nonzero-sum stochastic games with potentials. In *International Conference on Machine Learning*, pages 7688–7699. PMLR, 2021.

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

Dawei Qiu, Yi Wang, Tingqi Zhang, Mingyang Sun, and Goran Strbac. Hybrid multi-agent reinforcement learning for electric vehicle resilience control towards a low-carbon transition. *IEEE Transactions on Industrial Informatics*, 2022.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

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

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021a.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: stationary points, convergence, and sample complexity. *arXiv preprint arXiv:2106.00198*, 2021b.

Changxi Zhu, Mehdi Dastani, and Shihan Wang. A survey of multi-agent reinforcement learning with communication. *arXiv preprint arXiv:2203.08975*, 2022.