

# Scalable Multi-Agent Reinforcement Learning with General Utilities

Donghao Ying<sup>1</sup>, Yuhao Ding<sup>1</sup>, Alec Koppel<sup>2</sup> and Javad Lavaei<sup>1</sup>

**Abstract**— We study the scalable multi-agent reinforcement learning (MARL) with general utilities, defined as nonlinear functions of the team’s long-term state-action occupancy measure. The objective is to find a localized policy that maximizes the average of the team’s local utility functions without the full observability of each agent in the team. By exploiting the spatial correlation decay property of the network structure, we propose a scalable distributed policy gradient algorithm with shadow reward and localized policy that consists of three steps: (1) shadow reward estimation, (2) truncated shadow Q-function estimation, and (3) truncated policy gradient estimation and policy update. Our algorithm converges, with high probability, to  $\epsilon$ -stationarity with  $\tilde{\mathcal{O}}(\epsilon^{-2})$  samples up to some approximation error that decreases exponentially in the communication radius. This is the first result in the literature on multi-agent RL with general utilities that does not require the full observability.

## I. INTRODUCTION

Many decision-making problems take a form beyond the classic cumulative reward, such as apprenticeship learning [1], diverse skill discovery [2], pure exploration [3], and state marginal matching [4], among others. Such problems can be abstracted as *reinforcement Learning (RL) with general utilities* [5], [6], which focus on finding a policy to maximize a nonlinear function of the induced state-action occupancy measure. It generalizes the standard RL in which the objective is only an inner product between the state-action occupancy measure induced by the policy and a policy-independent reward for each state-action pair.

Beyond the single agent RL, consider the multi-agent problem where different agents need to interact to obtain a favorable outcome by finding a decision policy that maximizes the global accumulation of all agent’s general utility. This setting captures a wide range of applications, e.g. epidemics [7], social networks [8], finance [9], intelligent transportation [10] and wireless communication networks [11]. Recently, [12] proposed a new mechanism for cooperation that allows agents to incorporate general utilities for multi-agent RL (MARL) with common payoffs among agents. To enable the decentralization of agents’ policies under general utilities, [12] defines local occupancy measure of each agent as a marginalization of the global occupancy measure, and it defines the local general utility of the agent as an arbitrary function of its local occupancy measure. Based on these definitions, [12] derives a policy gradient-based algorithm, namely Decentralized Shadow Reward Actor-Critic, where

each agent estimates its policy gradient based on local information and communications with its neighbors.

However, their approach assumes the full observability, i.e., each agent has access to the global states and actions. Such assumption has two limitations. First, it is expensive and sometimes impossible to communicate with all agents in the team when the size of the team is large. In addition, full observability also implies that the policy and critics in this approach depend on the global states and actions, which may be a barrier to effective distributed implementation in practice. Moreover, even if individual state and action spaces are often small, the size of global state and action spaces can be exponentially large in the number of agents, which can be fundamentally intractable for large numbers of agents [13].

To address these issues, we aim to develop a scalable algorithm for multi-agent RL with general utilities without the full observability assumption. Inspired by the localization idea proposed in [14], our work makes the following contributions:

- We derive a truncated policy gradient estimator using the shadow reward and the localized policy for MARL with general utilities. We further establish the approximation error of the proposed truncated policy gradient estimator based on the spatial correlation decay assumptions;
- We propose a distributed policy gradient algorithm with shadow reward and localized policy that consists of three pieces: (1) shadow reward estimation, (2) truncated shadow Q-function estimation, and (3) truncated policy gradient estimation and policy update.
- We establish that, with high probability, our algorithm requires  $\tilde{\mathcal{O}}(\epsilon^{-2})$  samples to achieve  $\epsilon$ -stationarity with the error term  $\mathcal{O}(n\phi_0^{2\kappa})$ , where  $\phi_0 \in (0, 1)$ ,  $n$  is the number of agents,  $\mathcal{N}$  is the set of agents.

It is critical to note that the operating hypotheses we require are related to, but distinct from [14]: we assume the transition dynamics and policies of all agents are globally correlated and the correlation satisfies a spatial decay property. In contrast, the agents are considered to act on their own with their transitions only affected by the nearest neighbors in [14].

## A. Notations

For a set  $\mathcal{S}$ , let  $|\mathcal{S}|$  denote its cardinality and let  $\text{TV}(\mu, \mu') := \sup_{A \subset \mathcal{S}} |\mu(A) - \mu'(A)|$  be the total variation distance between two distributions  $\mu$  and  $\mu'$  on  $\mathcal{S}$ . When the variable  $s$  follows the distribution  $\xi$ , we write it as  $s \sim \xi$ . Let  $\mathbb{E}[\cdot]$  and  $\mathbb{E}[\cdot | \cdot]$ , respectively, denote the expectation and conditional expectation. For vectors  $x$  and  $y$ , we use  $x^\top$  to denote the transpose of  $x$  and use  $\langle x, y \rangle$  to denote the inner product  $x^\top y$ . We use the convention that  $\|x\|_1 = \sum_i |x_i|$ ,  $\|x\| := \|x\|_2 = \sqrt{\sum_i x_i^2}$ , and  $\|x\|_\infty = \max_i |x_i|$ .

<sup>1</sup>Donghao Ying, Yuhao Ding and Javad Lavaei are with the Department of Industrial Engineering and Operations Research, University of California at Berkeley. Email: {donghaoy, yuhao\_ding, lavaei}@berkeley.edu

<sup>2</sup>Alec Koppel is with J.P. Morgan AI Research. Email: alec.koppel@jpmchase.com

## II. PROBLEM FORMULATION

Consider a Markov Decision Process (MDP) over a finite state space  $\mathcal{S}$  and a finite action space  $\mathcal{A}$  with a discount factor  $\gamma \in [0, 1)$ . Let  $\xi$  be the initial distribution. A policy  $\pi$  is a function that specifies the decision rule of the agent, i.e., the agent takes action  $a \in \mathcal{A}$  with probability  $\pi(a|s)$  in state  $s \in \mathcal{S}$ . When action  $a$  is taken, the transition to the next state  $s'$  from state  $s$  follows the probability distribution  $s' \sim \mathbb{P}(\cdot|s, a)$ . In standard RL, the objective is to maximize the expected (discounted) cumulative reward, i.e.,

$$\max_{\pi} V^{\pi}(r) := \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k r(s^k, a^k) \middle| a^k \sim \pi(\cdot|s^k), s^0 \sim \xi \right], \quad (1)$$

where  $r(\cdot, \cdot)$  denotes the reward function and the expectation is taken over all possible trajectories. The value function can also be written as  $V^{\pi}(r) = \langle r, \lambda^{\pi} \rangle$ , where  $\lambda^{\pi}$  is the *discounted state-action occupancy measure* defined as

$$\lambda^{\pi}(s, a) = \sum_{k=0}^{\infty} \gamma^k \mathbb{P}(s^k = s, a^k = a | \pi, s^0 \sim \xi), \quad \forall (s, a). \quad (2)$$

We consider a more general problem where the objective is to maximize a function of  $\lambda^{\pi}$ , namely

$$\max_{\pi} f(\lambda^{\pi}), \quad (3)$$

where  $f: \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|} \rightarrow \mathbb{R}$  can be a possibly nonlinear function. Such an objective arises in many problems. For instance, in apprenticeship learning [1], the objective is  $f(\lambda^{\pi}) = -\text{dist}(\lambda^{\pi}, \lambda_e)$ , where  $\lambda_e$  corresponds to the expert demonstration and  $\text{dist}(\cdot, \cdot)$  is a distance function. In maximum entropy exploration [3],  $f(\cdot)$  refers to the entropy function such that  $f(\lambda^{\pi}) = -\sum_s d^{\pi}(s) \log d^{\pi}(s)$ , where  $d^{\pi}(s) = (1-\gamma) \sum_a \lambda^{\pi}(s, a)$  is the discounted state occupancy measure.

In this work, we study the decentralized version of (3), where the system is decentralized among a network of agents associated with a graph  $\mathcal{G} = (\mathcal{N}, \mathcal{E})$  (not densely connected). The vertex set  $\mathcal{N} = \{1, 2, \dots, n\}$  denotes the set of  $n$  agents and the edge set  $\mathcal{E}$  prescribes the communication links among agents. Let  $d(i, j)$  be the distance between agents  $i$  and  $j$  on  $\mathcal{G}$ , defined as the length of the shortest path between them. For  $\kappa \geq 0$ , we define  $\mathcal{N}_i^{\kappa} = \{j \in \mathcal{N} | d(i, j) \leq \kappa\}$  as the set of agents in the neighborhood of radius  $\kappa$  of agent  $i$ , with the shorthand notations  $\mathcal{N}_{-i}^{\kappa} := \mathcal{N} \setminus \mathcal{N}_i^{\kappa}$  and  $-i := \mathcal{N} \setminus \mathcal{N}_i^0 = \mathcal{N} \setminus \{i\}$ . The details of the decentralization are as follows:

a) *Space Decomposition:* The global state and action spaces are the product of local spaces, i.e.,  $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \dots \times \mathcal{S}_n$ ,  $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \dots \times \mathcal{A}_n$ , meaning that for every  $s \in \mathcal{S}$  and  $a \in \mathcal{A}$ , we can write  $s = (s_1, s_2, \dots, s_n)$  and  $a = (a_1, a_2, \dots, a_n)$ . For each subset  $\mathcal{N}' \subset \mathcal{N}$ , we use  $(s_{\mathcal{N}'}, a_{\mathcal{N}'})$  to denote the state-action pair for agents in  $\mathcal{N}'$ . We assume that each agent has direct access to its own states and actions while accessing other agents' information requires communications.

b) *Transition Decomposition:* Given the current global state  $s$  and action  $a$ , the local states in the next period are independently generated, i.e.,  $\mathbb{P}(s'|s, a) = \prod_{i \in \mathcal{N}} \mathbb{P}_i(s'_i|s, a)$ ,  $\forall s' \in \mathcal{S}$ , where  $\mathbb{P}_i$  denotes the local transition probability.

c) *Policy Factorization:* The global policy can be decomposed as  $\pi(a|s) = \prod_{i \in \mathcal{N}} \pi^i(a_i|s_{\mathcal{N}_i^{\kappa}})$ ,  $\forall (s, a)$ , i.e., given global state  $s$ , each agent  $i$  acts independently according to its local policy  $\pi^i$ , which depends on the state of agents in  $\mathcal{N}_i^{\kappa}$ . For the policy parameterization, we assume that the local policy of agent  $i$  is parameterized by  $\theta_i$ , and therefore one can write  $\pi(a|s) = \pi_{\theta}(a|s) = \prod_{i \in \mathcal{N}} \pi_{\theta_i}^i(a_i|s_{\mathcal{N}_i^{\kappa}})$ , where  $\theta = (\theta_1, \theta_2, \dots, \theta_n) \in \Theta$  is the global parameter.

d) *Local Utility:* For each agent  $i$ , define its *local discounted state-action occupancy measure* as

$$\lambda_i^{\pi}(s_i, a_i) = \sum_{k=0}^{\infty} \gamma^k \mathbb{P}(s_i^k = s_i, a_i^k = a_i | \pi, s^0 \sim \xi), \quad \forall (s_i, a_i), \quad (4)$$

which can be viewed as the marginalization of the global occupancy measure, i.e.,  $\lambda_i^{\pi}(\hat{s}_i, \hat{a}_i) = \sum_{s_i = \hat{s}_i, a_i = \hat{a}_i} \lambda^{\pi}(s, a)$ . Then, the global utility function  $f(\cdot)$  can be written as the average of local utilities, i.e.,  $f(\lambda^{\pi}) = 1/n \times \sum_{i \in \mathcal{N}} f_i(\lambda_i^{\pi})$ , where  $f_i: \mathbb{R}^{|\mathcal{S}_i| \times |\mathcal{A}_i|} \rightarrow \mathbb{R}$  is a function of the local occupancy measure  $\lambda_i^{\pi}$  and is private to agent  $i$ . Thus, under the parameterization  $\pi_{\theta}$ , (3) can be rewritten as

$$\max_{\theta \in \Theta} F(\theta), \quad \text{where } F(\theta) := f(\lambda^{\pi_{\theta}}) = \frac{1}{n} \cdot \sum_{i \in \mathcal{N}} f_i(\lambda_i^{\pi_{\theta}}). \quad (5)$$

Finally, we remark that, by choosing all  $f_i(\cdot)$  to be linear, (5) reduces to standard MARL, where each agent  $i$  is associated with a local reward function  $r_i: \mathcal{S}_i \times \mathcal{A}_i \rightarrow \mathbb{R}$  and the global reward is defined as  $r(s, a) := 1/n \times \sum_{i \in \mathcal{N}} r_i(s_i, a_i)$ .

## III. TRUNCATED POLICY GRADIENT ALGORITHM WITH SHADOW REWARD

In RL with cumulative reward, the *policy gradient theorem* [15] applies to computing the gradient of the value function:

$$\begin{aligned} \nabla_{\theta} V^{\pi_{\theta}}(r) &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_{\theta}}, a \sim \pi_{\theta}(\cdot|s)} [\psi_{\theta}(a|s) \cdot Q^{\pi_{\theta}}(r; s, a)], \\ &= \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k \psi_{\theta}(a^k|s^k) \cdot Q^{\pi_{\theta}}(r; s^k, a^k) \middle| \pi, s^0 \sim \xi \right], \end{aligned} \quad (6)$$

where  $\psi_{\theta}(\cdot|\cdot) := \nabla_{\theta} \log \pi_{\theta}(\cdot|\cdot)$  denotes the score function, and  $Q^{\pi}(r; s, a)$  is the state-action value function (Q-function) under reward  $r(\cdot, \cdot)$ , defined as

$$Q^{\pi}(r; s, a) = \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k r(s^k, a^k) \middle| \pi, s^0 = s, a^0 = a \right]. \quad (7)$$

However, for objective (5) with general utilities, this elegant result no longer holds. Instead, we have the following lemma.

**Lemma 1.** *For every policy  $\pi_{\theta}$ , it holds that*

$$\nabla_{\theta} F(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_{\theta}}, a \sim \pi_{\theta}(\cdot|s)} [\psi_{\theta}(a|s) \cdot Q_f^{\pi_{\theta}}(s, a)], \quad (8)$$

where  $Q_f^{\pi_{\theta}}(\cdot, \cdot) := Q^{\pi_{\theta}}(r^{\pi_{\theta}}; \cdot, \cdot)$  is the shadow Q-function and  $r^{\pi_{\theta}} := \nabla_{\lambda} f(\lambda^{\pi_{\theta}}) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$  is the shadow reward associated with policy  $\pi_{\theta}$ .

*Proof.* For value functions with cumulative reward, we observe the relation  $\nabla_{\theta} V^{\pi_{\theta}}(r) = \nabla_{\theta} \langle r, \lambda^{\pi_{\theta}} \rangle = \langle r, \nabla_{\theta} \lambda^{\pi_{\theta}} \rangle$ . Thus, by the chain rule, we have that

$$\nabla_{\theta} F(\theta) = \nabla_{\theta} f(\lambda^{\pi_{\theta}}) = \langle \nabla_{\lambda} f(\lambda^{\pi_{\theta}}), \nabla_{\theta} \lambda^{\pi_{\theta}} \rangle = \nabla_{\theta} V^{\pi_{\theta}}(r^{\pi_{\theta}}), \quad (9)$$

which completes the proof by the policy gradient theorem.  $\square$

In the decentralized formulation (5), for each agent  $i$ , let  $r_i^{\pi_\theta} := \nabla_{\lambda_i} f_i(\lambda_i^{\pi_\theta}) \in \mathbb{R}^{|\mathcal{S}_i| \times |\mathcal{A}_i|}$  be the local shadow reward, which only depends on the local state and action for a given policy  $\pi_\theta$ , and we define the local shadow Q-function as  $Q_i^{\pi_\theta}(s, a) := Q^{\pi_\theta}(r_i^{\pi_\theta}, s, a)$ . Then, it is clear that  $r^{\pi_\theta} = 1/n \times \sum_{i \in \mathcal{N}} r_i^{\pi_\theta}$  and  $Q_f^{\pi_\theta}(s, a) = 1/n \times \sum_{i \in \mathcal{N}} Q_i^{\pi_\theta}(s, a)$ , and the gradient of  $F(\theta)$  with respect to agent  $i$ 's local parameter  $\theta_i$  can be written as

$$\nabla_{\theta_i} F(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim a^{\pi_\theta}} \left[ \psi_{\theta_i}(a_i | s_{\mathcal{N}_i^\kappa}) \cdot \frac{1}{n} \sum_{j \in \mathcal{N}} Q_j^{\pi_\theta}(s, a) \right], \quad (10)$$

where we use the policy factorization to derive that  $\nabla_{\theta_i} \log \pi_\theta(a|s) = \nabla_{\theta_i} \log \pi_{\theta_i}^i(a_i | s_{\mathcal{N}_i^\kappa}) =: \psi_{\theta_i}(a_i | s)$ , and we refer to  $\psi_{\theta_i}(\cdot)$  as the local score function. Thus, updating the local parameter  $\theta_i$  with the gradient (10) requires knowing the global state and action as well as the shadow Q-functions of all agents, which can be inefficient in large networks due to the communication cost. In the remainder of the section, we show that an accurate gradient estimator can be designed for all agents while only local communications with neighbors are required under some correlation decay assumptions.

#### A. Spatial Correlation Decay Assumption

Following [16], we assume that a form of correlation decay property holds for the transition probability [17], [18].

**Assumption 1.** For a matrix  $M \in \mathbb{R}^{n \times n}$  whose  $(i, j)$  entry is defined as

$$M_{ij} = \sup \text{TV} \left( \mathbb{P}_i(\cdot | s_j, s_{-j}, a_j, a_{-j}), \mathbb{P}_i(\cdot | s'_j, s_{-j}, a'_j, a_{-j}) \right), \quad (11)$$

$s_j, a_j, s'_j, a'_j, s_{-j}, a_{-j}$

assume that there exists  $\beta \geq 0$  such that

$$\max_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} e^{\beta d(i,j)} M_{ij} \leq \rho, \quad (12)$$

with  $\rho < 1/\gamma$ , where  $\gamma$  is the discount factor.

By definition, the element  $M_{ij}$  characterizes the maximum level of impact of agent  $j$ 's state and action on the local transition probability of agent  $i$ . Then, Assumption 1 mainly requires that such impacts decrease exponentially with respect to the distance between agents. Such a decay is usually typical in engineered systems with large networks, e.g., in wireless communication where the strength of signals decreases exponentially with the distance [19], [20].

#### B. Truncated Shadow Q-function

We first introduce the notion of *exponential decay* for Q-functions [14], which is a form of correlation decay property.

**Definition 1.** For  $c \geq 0$  and  $\phi \in (0, 1)$ , the  $(c, \phi)$ -exponential decay property holds if, for every policy  $\pi_\theta$ , agent  $i$ , and state-action pairs  $(s, a), (s', a') \in \mathcal{S} \times \mathcal{A}$  with  $s_{\mathcal{N}_i^\kappa} = s'_{\mathcal{N}_i^\kappa}$ ,  $a_{\mathcal{N}_i^\kappa} = a'_{\mathcal{N}_i^\kappa}$ , the local shadow Q-function satisfies

$$|Q_i^{\pi_\theta}(s, a) - Q_i^{\pi_\theta}(s', a')| \leq c\phi^\kappa. \quad (13)$$

The exponential decay property holds when the dependency of each agent's local shadow Q-function on other agents' states and actions exponentially decreases with respect to their distances. Motivated by [14] and [18], for every  $i$ , we define  $\widehat{Q}_i^{\pi_\theta} : \mathcal{S}_{\mathcal{N}_i^\kappa} \times \mathcal{A}_{\mathcal{N}_i^\kappa} \rightarrow \mathbb{R}$  to be agent  $i$ 's truncated shadow Q-function, depending only on the states and actions of agents in the neighborhood  $\mathcal{N}_i^\kappa$ :

$$\widehat{Q}_i^{\pi_\theta}(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) := Q_i^{\pi_\theta}(s_{\mathcal{N}_i^\kappa}, \bar{s}_{\mathcal{N}_{-i}^\kappa}, a_{\mathcal{N}_i^\kappa}, \bar{a}_{\mathcal{N}_{-i}^\kappa}), \quad (14)$$

for every  $(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) \in \mathcal{S}_{\mathcal{N}_i^\kappa} \times \mathcal{A}_{\mathcal{N}_i^\kappa}$ , where  $(\bar{s}_{\mathcal{N}_{-i}^\kappa}, \bar{a}_{\mathcal{N}_{-i}^\kappa})$  is any fixed state-action pair for the agents in  $\mathcal{N}_{-i}^\kappa$ . That is, the estimator  $\widehat{Q}_i^{\pi_\theta}(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa})$  can be viewed as an approximate of the true shadow Q-function  $Q_i^{\pi_\theta}(s, a)$  by taking arbitrary values for  $(s_{\mathcal{N}_{-i}^\kappa}, a_{\mathcal{N}_{-i}^\kappa})$ . Compared with  $Q_i^{\pi_\theta}$ , the estimator  $\widehat{Q}_i^{\pi_\theta}$  depends on much smaller state and action spaces, and it is thus easy to estimate and store.

When the  $(c, \phi)$ -exponential decay property holds for Q-functions, it can be intuitively understood that the accuracy of this approximation has the order  $\mathcal{O}(\phi^\kappa)$ . The following lemma shows that, when Assumption 1 holds and the shadow reward is universally bounded, the exponential decay property is satisfied. We are thus capable of proving that  $\widehat{Q}_i^{\pi_\theta}$  is a satisfactory approximation of  $Q_i^{\pi_\theta}$ .

**Lemma 2.** Suppose that Assumption 1 holds and there exists  $M_f > 0$  such that  $\|\nabla_{\lambda_i} f_i(\lambda_i^{\pi_\theta})\|_\infty \leq M_f$ ,  $\forall i \in \mathcal{V}, \theta \in \Theta$ . Then, **(I)** the  $(c_0, \phi_0)$ -exponential decay property holds with  $(c_0, \phi_0) = \left( \frac{2\gamma\rho M_f}{1-\gamma\rho}, e^{-\beta} \right)$ , **(II)** the truncated shadow Q-function satisfies  $\sup_{s, a} |\widehat{Q}_i^{\pi_\theta}(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) - Q_i^{\pi_\theta}(s, a)| \leq c_0\phi_0^\kappa$ .

Under the bounded gradient assumption, we can treat the shadow Q-functions as standard Q-functions with bounded reward functions. We refer the reader to [16] for the proof of part **(I)** in Lemma 2. Then, part **(II)** follows directly from the definition of the exponential decay property. We note that the set of all possible state-action occupancy measures forms a convex polytope in  $\mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$  and is therefore a compact set. Thus, requiring the existence of  $M_f > 0$  in Lemma 2 is not a restrictive assumption and it naturally holds if the gradient  $\nabla_{\lambda} f(\lambda)$  is a continuous mapping on the set of occupancy measures. We additionally remark that a faster rate of the exponential decay property may be proved under extra assumptions, e.g., mixing properties of the underlying Markov chain [14].

#### C. Truncated Policy Gradient Estimator

In this section, we introduce how the exponential decay property can help design scalable algorithms.

As mentioned earlier, the major challenge in employing the exact policy gradient (10) comes from obtaining the global state-action pairs and the local shadow Q-functions of all agents, which may incur high costs in large networks. Instead, we consider the following truncated policy gradient estimator:

$$\widehat{g}_i(\theta) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim a^{\pi_\theta}} \left[ \psi_{\theta_i}(a_i | s_{\mathcal{N}_i^\kappa}) \cdot \frac{1}{n} \sum_{j \in \mathcal{N}_i^\kappa} \widehat{Q}_j^{\pi_\theta}(s_{\mathcal{N}_j^\kappa}, a_{\mathcal{N}_j^\kappa}) \right], \quad (15)$$

---

**Algorithm 1** Distributed Policy Gradient Algorithm With Shadow Reward and Localized Policy
 

---

- 1: **Input:** Initial policy  $\theta^0$ ; initial distribution  $\xi$ ; communication radius  $\kappa$ ; step-sizes  $\{\eta_\theta^t\}$ ; batch size  $B$ ; episode length  $H$ .
- 2: **for** iteration  $t = 0, 1, 2, \dots$  **do**
- 3: Sample  $B$  trajectories  $\tau = \{(s^0, a^0), \dots, (s^{H-1}, a^{H-1})\}$  with length  $H$ , under policy  $\pi_{\theta^t}$ , initial distribution  $\xi$ . Collect them as batch  $\mathcal{B}_t$ .
- 4: Each agent  $i$  estimates its local occupancy measure  $\lambda_i^{\pi_{\theta^t}}$  through

$$\tilde{\lambda}_i^t = \frac{1}{B} \sum_{\tau \in \mathcal{B}_t} \sum_{k=0}^{H-1} \gamma^k \cdot \mathbf{e}_i(s_i^k, a_i^k) \in \mathbb{R}^{|\mathcal{S}_i| \times |\mathcal{A}_i|}, \quad (17)$$

and computes the empirical shadow reward  $\tilde{r}_i^t = \nabla_{\lambda_i} f_i(\tilde{\lambda}_i^t)$ .

- 5: Each agent  $i$  communicates with its neighborhood  $\mathcal{N}_i^\kappa$  and estimate the truncated Q-function under  $\tilde{r}_i^t$ , denoted as  $\tilde{Q}_i^t$ .
- 6: Each agent  $i$  shares  $\tilde{Q}_i^t$  with its neighborhood  $\mathcal{N}_i^\kappa$  and estimates the truncated policy gradient through

$$\tilde{g}_i^t = \frac{1}{B} \sum_{\tau \in \mathcal{B}_t} \left[ \sum_{k=0}^{H-1} \gamma^k \psi_{\theta_i^t}(a_i^k | s_{\mathcal{N}_i^\kappa}^k) \cdot \frac{1}{n_{j \in \mathcal{N}_i^\kappa}} \sum_{j \in \mathcal{N}_i^\kappa} \tilde{Q}_j^t(s_{\mathcal{N}_j^\kappa}^k, a_{\mathcal{N}_j^\kappa}^k) \right]. \quad (18)$$

- 7: Each agent  $i$  updates the policy through

$$\theta_i^{t+1} = \theta_i^t + \eta_\theta^t \cdot \tilde{g}_i^t. \quad (19)$$

- 8: **end for**
- 

Compared to the true policy gradient (10), the estimator  $\hat{g}_i(\theta)$  replaces the shadow Q-functions with their truncated estimators. Furthermore, it only uses the truncated Q-functions of agents in  $\mathcal{N}_i^\kappa$ . In the next proposition, we evaluate the approximation error of  $\hat{g}_i(\theta)$ .

**Proposition 1.** *Let Assumption 1 hold. Suppose that there exist  $M_f, M_\psi > 0$  such that  $\|\nabla_{\lambda_i} f_i(\lambda_i^{\pi_\theta})\|_\infty \leq M_f$  and  $\|\psi_{\theta_i}(a_i | s_{\mathcal{N}_i^\kappa})\| \leq M_\psi, \forall i \in \mathcal{N}, (s, a) \in \mathcal{S} \times \mathcal{A}, \theta \in \Theta$ . Then, for all  $i \in \mathcal{N}, \theta \in \Theta$ , we have that*

$$\|\hat{g}_i(\theta) - \nabla_{\theta_i} F(\theta)\| \leq \frac{c_0 \phi_0^\kappa M_\psi}{1 - \gamma}. \quad (16)$$

Proposition 1 shows that, the accuracy of the truncated gradient estimator has the order  $\mathcal{O}(\phi_0^\kappa)$ , which decreases along with the communication radius  $\kappa$ . Thus, it indicates a feasible direction to reduce the communication of agents to their  $\kappa$ -neighborhoods. The proof of Proposition 1 can be found in the online appendix [21].

#### D. Algorithm Design

In this section, we present our method, Distributed Policy Gradient Algorithm with Shadow Reward, for solving problem (5). The algorithm, summarized in Algorithm 1, consists of the following elements:

a) *Shadow Reward Estimation (lines 3-4):* In the beginning of each iteration  $t$ , the current policy is simulated to generate a batch of  $B$  trajectories with length  $H$ . Since the local policy  $\pi_{\theta_i}(\cdot | s_{\mathcal{N}_i^\kappa})$  of each agent  $i$  only depends on the states of  $\mathcal{N}_i^\kappa$ , the process of trajectory sampling is comply with the communication requirement. Then, using local state-action information, each agent  $i$  forms an estimation  $\tilde{\lambda}_i^t$  for its local occupancy measure through (17), where we define  $\mathbf{e}_i(s_i, a_i) \in \mathbb{R}^{|\mathcal{S}_i| \times |\mathcal{A}_i|}$  as a vector with its  $(s_i, a_i)$ -th entry equal to one and other entries equal to zero. Finally, the empirical shadow reward is computed via  $\tilde{r}_i^t = \nabla_{\lambda_i} f_i(\tilde{\lambda}_i^t)$ .

b) *Truncated Shadow Q-function Estimation (line 5):* In the next stage, each agent  $i$  takes  $\tilde{r}_i^t$  as their reward function (pretending that to be the true shadow reward) and communicates with its neighborhood  $\mathcal{N}_i^\kappa$  to estimate the truncated shadow Q-function  $\tilde{Q}_i^t$ . We do not specify the estimation process and allow the use of any existing approach for Q-function evaluation as long as it satisfies the error bound required for the theoretical analysis in Section IV (see Assumption 4). For example, one can use the Temporal difference (TD) learning [22], which is a model-free method for estimating the Q-function. In TD-learning, all agents iteratively update their estimations along a common trajectory  $\tau = \{(s^0, a^0), \dots, (s^{H-1}, a^{H-1})\}$  under policy  $\pi_{\theta^t}$ . For every new global state-action pair  $(s^k, a^k)$ , the TD-learning updates the current estimation  $\tilde{Q}_i^t$  through

$$\begin{aligned} \tilde{Q}_i^t(s_{\mathcal{N}_i^\kappa}^{k-1}, a_{\mathcal{N}_i^\kappa}^{k-1}) &\leftarrow (1 - \eta_Q^{k-1}) \tilde{Q}_i^t(s_{\mathcal{N}_i^\kappa}^{k-1}, a_{\mathcal{N}_i^\kappa}^{k-1}) \\ &\quad + \eta_Q^{k-1} [\tilde{r}_i^t(s_i^{k-1}, a_i^{k-1}) + \gamma \tilde{Q}_i^t(s_{\mathcal{N}_i^\kappa}^k, a_{\mathcal{N}_i^\kappa}^k)], \end{aligned} \quad (20a)$$

$$\begin{aligned} \tilde{Q}_i^t(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) &\leftarrow \tilde{Q}_i^t(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) \\ &\quad \text{for } (s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) \neq (s_{\mathcal{N}_i^\kappa}^{k-1}, a_{\mathcal{N}_i^\kappa}^{k-1}), \end{aligned} \quad (20b)$$

where  $\{\eta_Q^k\}$  are the learning step-sizes. As shown in [14, Theorem 5], the above procedure exhibits an error rate of  $\mathcal{O}(1/\sqrt{H})$  under a local exploration assumption. Together with the error induced by the empirical shadow reward, this implies  $\|\tilde{Q}_i^t - \hat{Q}_i^t\|_\infty = \mathcal{O}(1/\sqrt{H} + \|\tilde{r}_i^t - r_i^t\|_\infty)$ . Besides the TD-learning, one can also deploy other model-free or model-based estimators depending on the sampling mechanisms, e.g., [23], [24].

c) *Truncated Policy Gradient Estimation and Policy Update (lines 6-7):* At the final stage, every agent  $i$  exchanges their estimation  $\tilde{Q}_i^t$  with the neighborhood  $\mathcal{N}_i^\kappa$  and evaluates the truncated policy gradient (15) through (18). The new policy is obtained by performing a policy gradient ascent with the estimated gradient  $\tilde{g}_i^t$ .

**Remark 1.** *In contrast to a major line of MARL research, e.g., [12], [25], full observability is not required for executing Algorithm 1, i.e., the agents do not need have access to the global information, including the global state and action. Instead, for the specified communication radius  $\kappa$ , each agent  $i$  needs to communicate with its neighborhood  $\mathcal{N}_i^\kappa$  to sample trajectories, estimate its local shadow Q-function, and estimate its truncated policy gradient.*

#### IV. CONVERGENCE ANALYSIS

In this section, we analyze the convergence behavior of Algorithm 1. The proofs of the results in this section can be found in the online appendix [21]. We first summarize the additional technical assumptions required, among which some have appeared in the previous section.

**Assumption 2.** Let  $\Lambda$  be the set of all possible occupancy measures  $\lambda$ . The utility function  $f(\cdot)$  satisfies: **(I)**  $\exists M_f > 0$  such that  $\|\nabla_{\lambda_i} f_i(\lambda_i)\|_\infty \leq M_f$ ,  $\forall i \in \mathcal{N}$  and  $\lambda \in \Lambda$ . **(II)**  $\exists L_\lambda$  such that  $\|\nabla_{\lambda_i} f_i(\lambda_i) - \nabla_{\lambda_i} f_i(\lambda'_i)\|_\infty \leq L_\lambda \|\lambda_i - \lambda'_i\|$ ,  $\forall i \in \mathcal{N}$  and  $\lambda, \lambda' \in \Lambda$ .

**Assumption 3.** The parameterized policy  $\pi_\theta$  and the associated occupancy measure  $\lambda^{\pi_\theta}$  satisfy: **(I)**  $\exists M_\psi > 0$  such that the score function  $\|\psi_{\theta_i}(a_i | s_{\mathcal{N}_i^\kappa})\| \leq M_\psi$ ,  $\forall i \in \mathcal{N}$ ,  $(s, a) \in \mathcal{S} \times \mathcal{A}$ ,  $\theta \in \Theta$ . **(II)**  $\exists L_\theta > 0$  such that the utility function  $F(\theta) = f(\lambda^{\pi_\theta})$  is  $L_\theta$ -smooth with respect to  $\theta$ .

Besides the bounded gradient and the bounded score function assumptions, we additionally assume that the utility function  $f_i(\lambda_i^{\pi_\theta})$  is smooth with respect to both the occupancy measure  $\lambda_i$  and the policy  $\theta$ . These assumptions are standard in the literature of reinforcement learning with general utilities [3], [12], [26], [27].

As discussed in Section III, we do not specify the estimation process for the truncated shadow Q-functions. Instead, we assume that an oracle is used, which produces a bounded-error approximation to the true function. Let  $\widehat{Q}_{r_i}^{\pi_\theta}(\cdot, \cdot) \in \mathbb{R}^{|\mathcal{S}_{\mathcal{N}_i^\kappa}| \times |\mathcal{A}_{\mathcal{N}_i^\kappa}|}$  be the  $\kappa$ -truncated local Q-function under reward  $r_i \in \mathbb{R}^{|\mathcal{S}_i| \times |\mathcal{A}_i|}$  for agent  $i$ .

**Assumption 4.** For every  $i \in \mathcal{N}$  and  $\theta \in \Theta$ , an approximation  $\widetilde{Q}_{r_i}^{\pi_\theta}(\cdot, \cdot)$  can be computed for  $\widehat{Q}_{r_i}^{\pi_\theta}(\cdot, \cdot)$  such that

$$\sup_{s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}} |\widetilde{Q}_{r_i}^{\pi_\theta}(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) - \widehat{Q}_{r_i}^{\pi_\theta}(s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa})| \leq \epsilon_0 \|r_i\|_\infty, \quad (21)$$

where  $\epsilon_0 > 0$  is the approximation error.

Under Assumption 4, we have that the estimator  $\widetilde{Q}_i^t$  in line 5 of Algorithm 1 satisfies  $\|\widetilde{Q}_i^t - \widehat{Q}_{r_i}^{\pi_{\theta^t}}\|_\infty \leq \epsilon_0 \|\widetilde{r}_i^t\|_\infty$ . This can be achieved, for example, with  $\mathcal{O}(1/(\epsilon_0)^2)$  samples by the TD-learning procedure (20).

Before analyzing the convergence of Algorithm 1, we first present a few auxiliary results, which evaluate the estimators  $\widetilde{\lambda}_i^t$ ,  $\widetilde{r}_i^t$ ,  $\widetilde{Q}_i^t$ , and  $\widetilde{g}_i^t$ .

**Proposition 2.** Let  $\delta_0 \in (0, 1/(2n))$  be the failure probability. Under Assumptions 2-4, it holds for every period  $t \geq 0$  that **(I)** for each agent  $i \in \mathcal{N}$ , with probability  $1 - \delta_0$

$$\|\widetilde{\lambda}_i^t - \lambda_i^{\pi_{\theta^t}}\| \leq \epsilon_1(\delta_0), \quad \|\widetilde{r}_i^t - r_i^t\|_\infty \leq L_\lambda \epsilon_1(\delta_0). \quad (22)$$

**(II)** with probability  $1 - n\delta_0$

$$\|\widetilde{Q}_i^t - \widehat{Q}_{r_i}^{\pi_{\theta^t}}\|_\infty \leq \epsilon_0 M_f + \frac{L_\lambda \epsilon_1(\delta_0)}{1 - \gamma}, \quad \forall i \in \mathcal{N}. \quad (23)$$

**(III)** with probability  $1 - 2n\delta_0$

$$\|\widetilde{g}_i^t - \widehat{g}_i(\theta^t)\| \leq \epsilon_{2,i}(\delta_0), \quad \forall i \in \mathcal{N}, \quad (24)$$

where

$$\epsilon_1(\delta_0) = \sqrt{\frac{4 + 2\gamma^{2H}B - 16 \log \delta_0}{(1 - \gamma)^2 B}}, \quad (25a)$$

$$\epsilon_{2,i}(\delta_0) = \frac{|\mathcal{N}_i^\kappa|}{n} \mathcal{O}\left(\epsilon_0 + \sqrt{\frac{\log(1/\delta_0)}{B}} + \gamma^H\right). \quad (25b)$$

Proposition 2 evaluates the accuracy of the estimation for the truncated policy gradient. Together with Proposition 1, this provides a probabilistic upper bound for the gradient estimation error  $\|\widetilde{g}_i^t - \nabla_{\theta_i} F(\theta^t)\|$ , which we will use to prove the convergence of Algorithm 1 in the following theorem.

**Theorem 1.** Suppose that Assumptions 1-4 hold and the step-sizes satisfy  $\eta_\theta^t \leq 1/(4L_\theta)$ ,  $\forall t \geq 0$ . For every  $T > 0$ , let  $\delta_0 = \delta/(2nT)$ , where  $\delta \in (0, 1)$  is the failure probability. Then, with probability  $1 - \delta$ , it holds that

$$\frac{\sum_{t=0}^{T-1} \eta_\theta^t \|\nabla_\theta F(\theta^t)\|^2}{\sum_{t=0}^{T-1} \eta_\theta^t} \leq \frac{4(F(\theta^T) - F(\theta^0))}{\sum_{t=0}^{T-1} \eta_\theta^t} + 3\Delta(\delta_0), \quad (26)$$

where

$$\Delta(\delta_0) = \mathcal{O}(n\phi_0^{2\kappa}) + \sum_{i \in \mathcal{N}} \frac{|\mathcal{N}_i^\kappa|^2}{n^2} \mathcal{O}\left(\epsilon_0^2 + \frac{\log(1/\delta_0)}{B} + \gamma^{2H}\right). \quad (27)$$

Under constant step-sizes  $\eta_\theta^t \equiv \eta_\theta$ , the bound (26) becomes

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla_\theta F(\theta^t)\|^2 \leq \frac{4(F(\theta^T) - F(\theta^0))}{\eta_\theta T} + 3\Delta(\delta_0), \quad (28)$$

which implies an  $\mathcal{O}(1/T)$  iteration complexity with the approximation error  $3\Delta(\delta_0)$ . As shown in (27), the constant  $\Delta(\delta_0)$  will be small when the rate of spatial correlation decay is fast, the computational error  $\epsilon_0$  for Q-functions is small, and enough samples are used to estimate the local occupancy measure. Notably, when the size of  $\kappa$ -neighborhood  $|\mathcal{N}_i^\kappa|$  is relatively small for all agents compared to the total number of agents  $n$ , the term  $\sum_{i \in \mathcal{N}} |\mathcal{N}_i^\kappa|^2/n^2$  approaches  $\mathcal{O}(1/n)$  and  $\Delta(\delta_0) = \mathcal{O}(n\phi_0^{2\kappa})$  approximately holds.

Suppose that an  $\mathcal{O}(1/(\epsilon_0)^2)$  oracle is used for the truncated Q-function estimation (line 5 in Algorithm 1), i.e., the approximation (21) is achieved with  $\mathcal{O}(1/(\epsilon_0)^2)$  samples. We analyze the sample complexity of Algorithm 1 to compute an  $\epsilon$ -stationary point.

**Theorem 2.** Suppose that Assumptions 1-4 hold and an  $\mathcal{O}(1/(\epsilon_0)^2)$  oracle is used for the truncated Q-function estimation. For every  $\epsilon > 0$  and  $\delta \in (0, 1)$ , let  $T = \mathcal{O}(\epsilon^{-1})$ ,  $\eta_\theta^t \equiv 1/(4L_\theta)$ ,  $\epsilon_0 = \sqrt{\epsilon}$ ,  $\delta_0 = \delta/(2nT)$ , batch size  $B = \mathcal{O}(\log(1/\delta_0)\epsilon^{-1})$ , episode length  $H = \mathcal{O}(\log(1/\epsilon))$ . Then, with probability  $1 - \delta$ , it holds that

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla_\theta F(\theta^t)\|^2 = \mathcal{O}(\epsilon + n\phi_0^{2\kappa}). \quad (29)$$

The total number of samples required is  $\widetilde{\mathcal{O}}(\epsilon^{-2})$ .

As discussed in Section III, the TD-learning procedure (20) is an  $\mathcal{O}(1/(\epsilon_0)^2)$  oracle for the truncated Q-function estimation. Below, we provide two further remarks.

**Remark 2** (Global Optimality). *Suppose that the utility function  $f(\lambda)$  is concave in  $\lambda$ , which generalizes the linear objective for standard RL. If the policy parameterization satisfies [26, Assumption 5.11], then problem (5) does not have spurious local solutions. Thus, the error bound (26) implies convergence to global optimality.*

**Remark 3.** *The communication radius  $\kappa$  plays an important role in both Theorems 1 and 2. As  $\kappa$  increases, the term  $\phi_0^{2\kappa}$  decreases, yet the size of the  $\kappa$ -neighborhood  $|\mathcal{N}_i^\kappa|$  increases, making the constant  $\sum_{i \in \mathcal{N}} |\mathcal{N}_i^\kappa|^2 / n^2$  increase. Also, the increase of  $|\mathcal{N}_i^\kappa|$  will amplify the communication cost and make the estimation of truncated Q-functions less efficient. Thus, finding a good balance is important in determining  $\kappa$ .*

**Remark 4.** *In this work, we focus on the policy search in a class of localized policies, where each local policy  $\pi_{\theta_i}^i(a_i | s_{\mathcal{N}_i^\kappa})$  only depends on the states of agents in  $\mathcal{N}_i^\kappa$ . It is possible to relax this “hard” requirement to a “soft” requirement. Specifically, consider  $\pi_\theta(a | s) = \prod_{i \in \mathcal{N}} \pi_{\theta_i}^i(a_i | s)$ , where each local policy  $\pi_{\theta_i}^i$  depends on the global state  $s$ . Under a form of spatial correlation decay property on the local policy  $\pi_{\theta_i}^i(a_i | s)$  and the associated local score function  $\psi_{\theta_i}(a_i | s) = \nabla_{\theta_i} \log \pi_{\theta_i}^i(a_i | s)$ , one can show that Algorithm 1 can still be implemented without violating the observability and communication requirements. The same convergence results hold with an additional approximation error resulting from the trajectory sampling. A detailed discussion of this extension is provided in the online appendix [21].*

## V. CONCLUSIONS

In this paper, we study the scalable MARL with general utilities, defined as nonlinear functions of the team’s long-term state-action occupancy measure. We propose a scalable distributed policy gradient algorithm with shadow reward and localized policy, which has three steps: (1) shadow reward estimation, (2) truncated shadow Q-function estimation, and (3) truncated policy gradient estimation and policy update. By exploiting the spatial correlation decay property of the network structure, we rigorously establish the convergence and sample complexity of the proposed algorithm. Future work includes generalization to the safety-critical setting and considering information asymmetry among the agents.

## ACKNOWLEDGMENT

This work was supported by grants from ARO, AFOSR, ONR and NSF.

## REFERENCES

- [1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.
- [2] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.
- [3] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pages 2681–2691. PMLR, 2019.
- [4] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.
- [5] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational policy gradient method for reinforcement learning with general utilities. *Advances in Neural Information Processing Systems*, 33:4572–4583, 2020.
- [6] Tom Zahavy, Brendan O’Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdp. *Advances in Neural Information Processing Systems*, 34, 2021.
- [7] Wenjun Mei, Shadi Mohagheghi, Sandro Zampieri, and Francesco Bullo. On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control*, 44:116–128, 2017.
- [8] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.
- [9] Jae Won Lee, Byoung-Tak Zhang, et al. Stock trading system using reinforcement learning with cooperative agents. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 451–458, 2002.
- [10] Rick Zhang and Marco Pavone. Control of robotic mobility-on-demand systems: a queueing-theoretical perspective. *The International Journal of Robotics Research*, 35(1-3):186–203, 2016.
- [11] Alessandro Zocca. Temporal starvation in multi-channel csma networks: an analytical framework. *Queueing Systems*, 91(3):241–263, 2019.
- [12] Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel. Multi-agent reinforcement learning with general utilities via decentralized shadow reward actor-critic. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9031–9039, 2022.
- [13] Vincent D Blondel and John N Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, 2000.
- [14] Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning for multiagent networked systems. *Operations Research*, 2022.
- [15] 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.
- [16] Carlo Alfano and Patrick Rebeschini. Dimension-free rates for natural policy gradient in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11692*, 2021.
- [17] Hans-Otto Georgii. Gibbs measures and phase transitions. de Gruyter, 2011.
- [18] David Gamarnik. Correlation decay method for decision, optimization, and inference in large-scale networks. In *Theory Driven by Influential Applications*, pages 108–121. INFORMS, 2013.
- [19] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.
- [20] Lawrence G Roberts. Aloha packet system with and without slots and capture. *ACM SIGCOMM Computer Communication Review*, 5(2):28–42, 1975.
- [21] Donghao Ying, Yuhao Ding, Alec Koppel, and Javad Lavaei. Scalable multi-agent reinforcement learning with general utilities. *arXiv preprint arXiv:2302.07938*, 2023.
- [22] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.
- [23] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043, 2020.
- [24] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.
- [25] Siliang Zeng, Tianyi Chen, Alfredo Garcia, and Mingyi Hong. Learning to coordinate in multi-agent systems: A coordinated actor-critic algorithm and finite-time guarantees. In *Learning for Dynamics and Control Conference*, pages 278–290. PMLR, 2022.
- [26] Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.
- [27] Donghao Ying, Mengzi Guo, Yuhao Ding, Javad Lavaei, and Zuoqun Shen. Policy-based primal-dual methods for convex constrained markov decision processes. *arXiv preprint arXiv:2205.10715*, 2022.