An Asynchronous Multi-Agent Actor-Critic Algorithm for Distributed Reinforcement Learning

Yixuan Lin, Yuehan Luo, Kaiqing Zhang, Zhuoran Yang, Zhaoran Wang,
Tamer Başar, Romeil Sandhu, and Ji Liu [*]

July 16, 2019

Abstract

This paper studies a distributed reinforcement learning problem in which a network of multiple agents aim to cooperatively maximize the globally averaged return through communication with only local neighbors. An asynchronous multi-agent actor-critic algorithm is proposed for possibly unidirectional communication relationships depicted by a directed graph. Each agent independently updates its variables at "event times" determined by its own clock. It is not assumed that the agents' clocks are synchronized or that the event times are evenly spaced. It is shown that the algorithm can solve the problem for any strongly connected graph in the presence of communication and computation delays.

I. Introduction

Distributed machine learning algorithms have drawn increasing attention recently, with some notable examples such as distributed multi-arm bandit [1], linear regression [2], deep learning [3], and reinforcement learning (RL) [4]. Promising applications of these algorithms are in large-scale networks without any central controller/coordinator, including online economic networks, Internet of Things, cyber-physical systems, and social platforms, primarily because in these examples, collecting all information at a single point is infeasible, due to privacy issues such that agents are not willing to share their private information, or expensive communication overhead in maintaining such big data.

Among these distributed machine learning algorithms, there has been an ever-growing interest in multi-agent reinforcement learning (MARL). In general, MARL problems are addressed in three settings, namely collaborative, competitive, and a mixture of the two. In the collaborative setting, the canonical multi-agent Markov decision process model [5, 6] appeared to be the most basic framework, where a common reward function is shared by all agents and affected by all agents' joint actions. Moreover, the team Markov game can also be used as a collaborative model, where the agents also share an identical reward function [7, 8]. Later, a more challenging but practical setting where agents can have heterogeneous reward functions, with the goal of maximizing the long-term return corresponding to the team averaged reward, was proposed in [4, 9–12]. Particularly, the focuses of these works are on a *fully-decentralized/distributed* setting,

[*] Y. Lin and Y. Luo are with the Department of Applied Mathematics and Statistics at Stony Brook University ({yixuan.lin.1,yuehan.luo}@stonybrook.edu). K. Zhang and T. Başar are with the Coordinated Science Laboratory at University of Illinois at Urbana-Champaign ({kzhang66,basar1}@illinois.edu), and their research was supported in part by the US Army Research Laboratory (ARL) Cooperative Agreement W911NF-17-2-0196. Z. Yang is with the Department of Operations Research and Financial Engineering at Princeton University (zy6@princeton.edu). Z. Wang is with the Department of Industrial Engineering and Management Sciences at Northwestern University (zhaoran.wang@northwestern.edu). R. Sandhu is with the Departments of Bioinformatics and Computer Science at Stony Brook University (romeil.sandhu@stonybrook.edu). J. Liu is with the Department of Electrical and Computer Engineering at Stony Brook University (ji.liu@stonybrook.edu).

where no central controller/decision maker exists to coordinate the agents and maximize the team averaged return. Instead, a communication network exists to connect the agents in which information exchange is allowed only between neighboring agents over the network. There is also a huge body of literature on MARL for the competitive and mixed settings [13–16], many of which are empirical works without theoretical convergence guarantees. Here, our focus is on the collaborative MARL with decentralized/distributed and networked agents, as in [4, 9, 17].

The work of [4] developed the first fully decentralized/distributed, synchronous actor-critic algorithm under the collaborative setting, in which doubly stochastic matrices were used to devise the consensus update. Such an update essentially needs the communication between each pair of neighboring agents to be bidirectional. This confines the applicability of the algorithm into scenarios with possibly unidirectional communication. More importantly, the requirement of doubly stochastic matrices further restricts its extension to the cases with communication delays and asynchronous updating, as there is no existing distributed way to devise a consensus update using a doubly stochastic matrix in the presence of communication delays or asynchronous updating.

Asynchronous RL methods [18–20] have gained great popularity recently as they can achieve successful real-world applications such as games and robotics [21, 22]. In these existing settings, a large number of RL agents collect experiences in independent environments and interact with a centralized parameter server. Compared with traditional RL algorithms, these asynchronous algorithms enjoy better exploration properties and are more tolerant of computation faults; however, they cannot be directly and easily, if not impossible, extended to fully distributed/decentralized settings in which there is no centralized parameter server. Typical application examples of such settings include robotic teams and drone fleets. It is also worth emphasizing that for a large wireless network, it is difficult and sometimes impossible to synchronize all components' clocks over the network [23]; and that is also the case with a large-scale distributed RL network.

In this paper, we propose an asynchronous, fully distributed actor-critic algorithm using the idea of push-sum [24, 25]. In our algorithm, each agent independently decides when to take actions according to its own clock. It is not assumed that the agents' clocks are synchronized. The algorithm also takes communication and computation delays into account. We show convergence of the algorithm under linear function approximation, which is validated via simulation.

II. Problem Formulation

In this section, we introduce the background and formulation of the MARL problem with networked agents. The problem was first proposed in [17] which provides two distributed algorithms for synchronous case without considering any delays.

A. Networked Multi-Agent MDP

Consider a team of N agents, denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$, operating in a common environment. There is no central controller that can either collect rewards or make the decisions for all the agents. In contrast, the agents are connected by a possibly sparse communication network depicted by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{E}$ denotes the set of communication links. A *networked multi-agent MDP* model can be defined by a tuple $(\mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, P, \{R^i\}_{i\in\mathcal{N}}, \mathcal{G}_t)$, where $\mathcal{S}$ is the state space shared by all the agents in $\mathcal{N}$, and $\mathcal{A}^i$ is the action space of agent i. For each agent i, $R^i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the local reward function, where $\mathcal{A} = \prod_{i=1}^{N} \mathcal{A}^i$ is the joint action space. $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ denotes the state transition probability of the MDP. It is assumed throughout the paper that the states are globally observable and the rewards are observed only locally. Each agent's rewards are only locally observed by itself, primarily due to privacy issues in the sense that the agents do not have motivation to share private reward information directly with others. Practical examples of this setting include cooperative navigation, motion planning of teamed robots, and dynamic operation of distributed energy resources in the smart grid; for more examples and justifications of the setting, see [4].

The networked multi-agent MDP evolves as follows. Each agent i chooses its own action a_t^i given state s_t at time t, according to a local policy, i.e., the probability of choosing action a^i at state s, $\pi^i : \mathcal{S} \times \mathcal{A}^i \to [0,1]$. Note that the joint policy of all agents, $\pi\colon \mathcal{S} \times \mathcal{A} \to [0,1]$, satisfies $\pi(s,a) = \prod_{i\in\mathcal{N}} \pi^i(s,a^i)$. Also, a reward r_{t+1}^i is received by agent i after executing the action. To make the search of the optimal joint policy tractable, we assume that the local policy is parameterized by $\pi_{\theta^i}^i$, where $\theta^i \in \Theta^i$ is the parameter, and $\Theta^i \subseteq \mathbb{R}^{m_i}$ is a compact set. The parameters are concatenated as $\theta = [(\theta^1)^\top, \cdots, (\theta^N)^\top]^\top \in \Theta$, where $\Theta = \prod_{i=1}^N \Theta^i$. The joint policy is thus given by $\pi_\theta(s,a) = \prod_{i\in\mathcal{N}} \pi_{\theta^i}^i(s,a_i)$. We first make a standard regularity assumption on the model and the policy parameterization.

Assumption 1. *For any $i \in \mathcal{N}$, $s \in \mathcal{S}$, and $a^i \in \mathcal{A}^i$, the policy function $\pi_{\theta^i}^i(s,a^i) > 0$ for any $\theta^i \in \Theta^i$. Also, $\pi_{\theta^i}^i(s,a^i)$ is continuously differentiable with respect to the parameter θ^i over Θ^i. In addition, for any $\theta \in \Theta$, let P^θ be the transition matrix of the Markov chain $\{s_t\}_{t\ge 0}$ induced by policy π_θ, that is, for any $s, s' \in \mathcal{S}$*

$$P^\theta(s' \mid s) = \sum_{a\in\mathcal{A}} \pi_\theta(s,a) \cdot P(s' \mid s,a). \tag{1}$$

The Markov chain $\{s_t\}_{t\ge 0}$ is irreducible and aperiodic under any π_θ, with the stationary distribution denoted by d_θ.

Assumption 1 has been imposed in the existing work on centralized actor-critic algorithms with function approximation [26, 27]. It implies that the Markov chain of the state-action pair $\{(s_t,a_t)\}_{t\ge 0}$ has a stationary distribution $d_\theta(s) \cdot \pi_\theta(s,a)$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

The objective of the agents is to collaboratively find a policy π_θ that maximizes the *globally* averaged long-term return over the network based solely on *local* information, namely,

$$\max_\theta \quad J(\theta) = \lim_T \ \frac{1}{T}\mathbb{E}\biggl(\sum_{t=0}^{T-1} \frac{1}{N}\sum_{i\in\mathcal{N}} r_{t+1}^i\biggr) = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} d_\theta(s)\pi_\theta(s,a) \cdot \overline{R}(s,a), \tag{2}$$

where $\overline{R}(s,a) = N^{-1} \cdot \sum_{i\in\mathcal{N}} R^i(s,a)$ is the globally averaged reward function. It is worth noting that such an averaged reward can be viewed as an example of Benthamite social welfare [28]. Let $\overline{r}_t = N^{-1} \cdot \sum_{i\in\mathcal{N}} r_t^i$; then, we have $\overline{R}(s,a) = \mathbb{E}[\overline{r}_{t+1} \mid s_t = s, a_t = a]$. Thus, the global relative action-value function under policy π_θ can be defined accordingly as

$$Q_\theta(s,a) = \sum_t \mathbb{E}\bigl[\overline{r}_{t+1} - J(\theta) \mid s_0 = s, a_0 = a, \pi_\theta\bigr],$$

and the global relative state-value function $V_\theta(s)$ is defined as $V_\theta(s) = \sum_{a\in\mathcal{A}} \pi_\theta(s,a) Q_\theta(s,a)$. For simplicity, hereafter we will refer to V_θ and Q_θ as *state-value* function and *action-value* function only. Furthermore, the *advantage function* can be defined as $A_\theta(s,a) = Q_\theta(s,a) - V_\theta(s)$.

As the basis for developing multi-agent actor-critic algorithms for distributed reinforcement learning, the following policy gradient theorem was established in [4] for MARL.

Policy Gradient Theorem for MARL [Theorem 3.1 in [4]]: For any $\theta \in \Theta$ and any agent $i \in \mathcal{N}$, define the local advantage function $A_\theta^i\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$A_\theta^i(s,a) = Q_\theta(s,a) - \tilde{V}_\theta^i(s,a^{-i}),$$

where $\tilde{V}_\theta^i(s,a^{-i}) = \sum_{a^i\in\mathcal{A}^i} \pi_{\theta^i}^i(s,a^i) \cdot Q_\theta(s,a^i,a^{-i})$, and a^{-i} denotes the actions of all agents except for agent i. Then, the gradient of $J(\theta)$ with respect to θ^i is given by

$$\nabla_{\theta^i} J(\theta) = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[\nabla_{\theta^i} \log \pi_{\theta^i}^i(s,a^i) \cdot A_\theta(s,a)\right] = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[\nabla_{\theta^i} \log \pi_{\theta^i}^i(s,a^i) \cdot A_\theta^i(s,a)\right].$$

B. The Asynchronous Algorithm

We consider a general asynchronous scenario in which each agent has its own independent clock. It is not assumed that all agents' clocks are synchronized, and thus the asynchronous system needs to be described in continuous time as follows. Each agent independently determines times at which it updates its variables. Specifically, each agent i has a strictly increasing, infinite sequence of *event times*, denoted by $t_{i0}, t_{i1}, t_{i2}, \ldots$, with the understanding that t_{i0} is the time agent i initializes its variables, and the remaining t_{ik}, $k \geq 1$, are the times at which agent i takes active actions such as transmitting information and updating variables. Between any two successive event times t_{ik} and $t_{i(k+1)}$, $k \geq 1$, agent i completes updating which may take time. Without loss of generality, we assume that all agents complete initialization before their first update, i.e., $t_{i0} < t_{j1}$ for all $i, j \in \mathcal{N}$. It is assumed that the difference between any two successive event times of each agent cannot be too large or too small. To be more precise, for any $i \in \mathcal{N}$, agent i's event times satisfy

$$\bar{T}_i \geq t_{i(k+1)} - t_{ik} \geq T_i, \quad k \geq 0, \tag{3}$$

where $\bar{T}_i$ and T_i are positive numbers such that $\bar{T}_i > T_i$. This assumption is natural as unbounded difference will make an algorithm suspend, and too frequent event times may cause an algorithm inefficient and sometimes even impossible due to hardware constraints. It is worth emphasizing that we make no assumptions about the relationships between the event times of different agents. Any two agents may have completely different unsynchronized event time sequences.

Each agent i communicates with the network at each of its event times t_{ik}, $k \geq 1$, by transmitting its current variables to its "out-neighbors". We say that an agent j is an out-neighbor of agent i if (i, j) is a directed edge in $\mathcal{G}$. Similarly, we say that an agent k is an in-neighbor of agent i if (k, i) is a directed edge in $\mathcal{G}$. An agent can send information only to its out-neighbors and receive information only from its in-neighbors. Thus, directions of the directed edges in $\mathcal{G}$ represent directions of information flow. We use $\mathcal{N}_-^i$ and $\mathcal{N}_+^i$ to denote the sets of out- and in-neighbors of agent i, respectively. For simplicity, we assume that each agent is always an out- and in-neighbor of itself, i.e., $i \in \mathcal{N}_-^i$ and $i \in \mathcal{N}_+^i$ for all $i \in \mathcal{N}$. In other words, $\mathcal{G}$ has a self-arc at each node. Thus, $|\mathcal{N}_-^i| \geq 1$ and $|\mathcal{N}_+^i| \geq 1$, where $|\mathcal{N}_-^i|$ and $|\mathcal{N}_+^i|$ denote the cardinality of $\mathcal{N}_-^i$ and $\mathcal{N}_+^i$, i.e., the number of out- and in-neighbors of agent i, respectively.

Each agent i has control over a set of variables, denoted $\mu_t^i, \omega_t^i, v_t^i, y_t^i, z_t^i, r_t^i, \theta_t^i$, whose purposes will be introduced shortly, and an additional scalar-valued variable y_t^i whose initial value $y_{t_{i0}}^i = 1$.

At each event time t_{ik}, $i \in \mathcal{N}$, $k \geq 1$, agent i sends a pair of scaled versions of its variables, $\frac{v_{t_{ik}}^i}{|\mathcal{N}_-^i|}$ and $\frac{y_{t_{ik}}^i}{|\mathcal{N}_-^i|}$, to each of its out-neighbors. Agent i's out-neighbors may receive this pair of variables at different times as the transmissions are subject to communication delays which are heterogeneous among the agents. We use $d_{t_{ik}}^{ij}$ to denote the communication delay when agent i sends information to its out-neighbor j at its event time t_{ik}. In other words, agent j will receive this information at time $t_{ik} + d_{t_{ik}}^{ij}$. Similarly, agent i receives pairs of variables from its in-neighbors from time to time which were transmitted at earlier times. We do not impose any restrictions on communication delays, except for a natural assumption that communication delays are bounded.

Each agent i computes new values of its variables based on those in-neighbors' variables received during the interval $(t_{i(k-1)}, t_{ik}]$. It is worth emphasizing that we take computation time/delays into account. It is assumed that all computations can be completed before the next event time $t_{i(k+1)}$ arrives. With this in mind, each agent i can define its next event time to be the time at/after which it finishes its last round of updating.

The asynchronous algorithm consists of two steps in each iteration, a critic step followed by an actor step. For each event time t_{ik}, $k \geq 1$, the critic step of agent i is as follows:

$$\begin{cases} \mu^i_{t_{i(k+1)}} = (1 - \beta_{\omega,t_{ik}}) \cdot \mu^i_{t_{ik}} + \beta_{\omega,t_{ik}} \cdot r^i_{t_{i(k+1)}}, & (4) \\ v^i_{t_{ik}} = \omega^i_{t_{ik}} + \beta_{\omega,t_{ik}} \cdot \delta^i_{t_{ik}} \cdot \nabla_z Q_{t_{ik}}(z^i_{t_{ik}}), & (5) \\ \omega^i_{t_{i(k+1)}} = \dfrac{v^i_{t_{ik}}}{|\mathcal{N}^i_-|} + \displaystyle\sum_{j \in \mathcal{N}^i_+} \sum_{s \geq 1} \dfrac{v^j_{t_{js}} \chi_{(t_{i(k-1)}, t_{ik}]}(t_{js} + d^{ji}_{t_{js}})}{|\mathcal{N}^j_-|}, & (6) \\ y^i_{t_{i(k+1)}} = \dfrac{y^i_{t_{ik}}}{|\mathcal{N}^i_-|} + \displaystyle\sum_{j \in \mathcal{N}^i_+} \sum_{s \geq 1} \dfrac{y^j_{t_{js}} \chi_{(t_{i(k-1)}, t_{ik}]}(t_{js} + d^{ji}_{t_{js}})}{|\mathcal{N}^j_-|}, & (7) \\ z^i_{t_{i(k+1)}} = \dfrac{\omega^i_{t_{i(k+1)}}}{y^i_{t_{i(k+1)}}}, & (8) \end{cases}$$

where $\mu^i_{t_{ik}}$ tracks the long-term average return of agent i, $\beta_{\omega,t_{ik}} > 0$ is the stepsize, $Q_{t_{ik}}(z)$ denotes $Q(s_{t_{ik}}, a_{t_{ik}}; z)$ for any z, $d^{ji}_{t_{js}}$ is the communication delay of information transmitted from agent j to agent i at time t_{js}, and $\chi_{(t_{i(k-1)}, t_{ik}]}(t_{js} + d^{ji}_{t_{js}})$ is an indicator function defined as $\chi_{(t_{i(k-1)}, t_{ik}]}(t_{js} + d^{ji}_{t_{js}}) = 1$ if $t_{i(k-1)} < t_{js} + d^{ji}_{t_{js}} \leq t_{ik}$, otherwise $\chi_{(t_{i(k-1)}, t_{ik}]}(t_{js} + d^{ji}_{t_{js}}) = 0$. It is worth noting that the second items at the right hand side of (6) and (7) take sum of all received scaled v and y variables, respectively, from agent i's in-neighbors during the interval $(t_{i(k-1)}, t_{ik}]$. The local *action-value TD-error* $\delta^i_{t_{ik}}$ in (5) is given by

$$\delta^i_{t_{ik}} = r^i_{t_{i(k+1)}} - \mu^i_{t_{ik}} + Q_{t_{i(k+1)}}(z^i_{t_{ik}}) - Q_{t_{ik}}(z^i_{t_{ik}}). \tag{9}$$

As for the actor step, agent i improves its policy via

$$\theta^i_{t_{i(k+1)}} = \theta^i_{t_{ik}} + \beta_{\theta,t_{ik}} \cdot A^i_{t_{ik}} \cdot \psi^i_{t_{ik}}, \tag{10}$$

where $\beta_{\theta,t_{ik}} > 0$ is the stepsize, A^i_t and ψ^i_t are defined as

$$A^i_t = Q_t(z^i_t) - \sum_{a^i \in \mathcal{A}^i} \pi^i_{\theta^i_t}(s_t, a^i) \cdot Q(s_t, a^i, a^{-i}_t; z^i_t), \tag{11}$$

$$\psi^i_t = \nabla_{\theta^i} \log \pi^i_{\theta^i_t}(s_t, a^i_t). \tag{12}$$

It is worth emphasizing that all above updating can be computed at agent i in a distributed manner.

We impose the following assumptions for the asynchronous actor-critic algorithm which are either mild or standard; see [17] for detailed discussions on these assumptions. In particular, we focus on convergence under linear approximation since even for centralized actor-critic algorithms, there is no convergence guarantee for nonlinear approximation.

Assumption 2. *The instantaneous reward r^i_t is uniformly bounded for any $i \in \mathcal{N}$ and $t \geq 0$.*

Assumption 3. *The stepsizes $\beta_{\omega,t}$ and $\beta_{\theta,t}$ satisfy, for all $i \in \mathcal{N}$, $\sum_{k \geq 1} \beta_{\omega,t_{ik}} = \sum_{k \geq 1} \beta_{\theta,t_{ik}} = \infty$ and $\sum_{k \geq 1}(\beta^2_{\omega,t_{ik}} + \beta^2_{\theta,t_{ik}}) < \infty$. In addition, $\beta_{\theta,t_{ik}} = o(\beta_{\omega,t_{ik}})$.*

Assumption 4. *For each agent i, the function $Q(s, a; z)$ is parametrized as $Q(s, a; z) = z^\top \phi(s, a)$, where $\phi(s, a) = [\phi_1(s, a), \cdots, \phi_K(s, a)]^\top \in \mathbb{R}^K$ is the feature associated with (s, a). The feature vector $\phi(s, a)$ is uniformly bounded for any $s \in \mathcal{S}, a \in \mathcal{A}$. Furthermore, the feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}| \cdot |\mathcal{A}| \times K}$ has full column rank, where the k-th column of Φ is $[\phi_k(s, a), s \in \mathcal{S}, a \in \mathcal{A}]^\top$ for any $k \in [K]$. Also, for any $u \in \mathbb{R}^K$, $\Phi u \neq \mathbf{1}_K$, where $\mathbf{1}_K$ denotes the K-dimensional vector whose entries all equal one.*

Assumption 5. *The update of the policy parameter θ_t^i includes a local projection operator, $\Gamma^i : \mathbb{R}^{m_i} \to \Theta^i \subset \mathbb{R}^{m_i}$, that projects any θ_t^i onto the compact set Θ^i. Also, we assume that $\Theta = \prod_{i=1}^N \Theta^i$ is large enough to include at least one local minimum of $J(\theta)$.*

For simplicity, we define $P^\theta(s', a' \mid s, a) = P(s' \mid s, a)\pi_\theta(s', a')$, $\mathrm{D}_\theta^{s,a} = \mathrm{diag}[d_\theta(s) \cdot \pi_\theta(s, a), s \in \mathcal{S}, a \in \mathcal{A}]$, and $\overline{R} = [\overline{R}(s, a), s \in \mathcal{S}, a \in \mathcal{A}]^\top \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$. Note that, with slight abuse of notation, the expression P^θ has the same form as the transition probability matrix of the Markov chain $\{s_t\}_{t\geq 0}$ under policy π_θ; see (1). These two matrices can be easily differentiated by the context.

To state our main result, we define the operator $T_\theta^Q : \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ for any action-value vector $Q \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$ as

$$T_\theta^Q(Q) = \overline{R} - J(\theta) \cdot \mathbf{1}_{|\mathcal{S}|\cdot|\mathcal{A}|} + P^\theta Q.$$

We also define the vector $\hat{\Gamma}^i(\cdot)$ as

$$\hat{\Gamma}^i[g(\theta)] = \lim_{0<\eta\to 0} \{\Gamma^i[\theta^i + \eta \cdot g(\theta)] - \theta^i\}/\eta \tag{13}$$

for any $\theta \in \Theta$ and continuous function $g : \Theta \to \mathbb{R}^{\sum_{i\in\mathcal{N}} m_i}$. In case the limit above is not unique, $\hat{\Gamma}^i[g(\theta)]$ is defined as the set of all possible limit points of (13).

With the above notation, we establish the following convergence results of the critic step (4) – (9) and actor step (10) – (12) given policy π_θ.

Theorem 1. *Suppose that Assumptions 1 – 4 hold, and that communication graph sequence $\{\bar{\mathcal{G}}_\tau\}_{\tau=1}^\infty$ is repeatedly jointly strongly connected. Then, for any given policy π_θ, with the sequences $\{\bar{\mu}_\tau^i\}$ and $\{\bar{z}_\tau^i\}$ generated from (4) and (8), we have $\lim_{k\to\infty} \sum_{i\in\mathcal{N}} \mu_{t_{ik}}^i \cdot N^{-1} = J(\theta)$ and $\lim_{k\to\infty} z_{t_{ik}}^i = \omega_\theta$ almost surely for any $i \in \mathcal{N}$, where $J(\theta)$ is the globally averaged return as defined in (2), and ω_θ is the unique solution to*

$$\Phi^\top \mathrm{D}_\theta^{s,a}\left[T_\theta^Q(\Phi\omega_\theta) - \Phi\omega_\theta\right] = 0.$$

Suppose further that Assumption 5 holds. Then, for all $i \in \mathcal{N}$, the sequence $\{\theta_{t_{ik}}^i\}$ obtained from (10) converges almost surely to a point in the set of the asymptotically stable equilibria of

$$\dot{\theta}^i = \hat{\Gamma}^i\left[\mathbb{E}_{s_t\sim d_\theta, a_t\sim\pi_\theta}\left(A_{t,\theta}^i \cdot \psi_{t,\theta}^i\right)\right].$$

III. Analysis

In this section, we use the concept of analytic synchronization [29] to derive a synchronous system whose limiting behavior is the same as the asynchronous system under consideration, which serves a critical step toward the proof of Theorem 1.

We first need a common time scale on which all n agents' update rules can be defined. For this, let $\mathcal{T}_i$ be the set of the event times of agent i which are greater than or equal to t_{i1}, and let $\mathcal{T}$ be the union of all $\mathcal{T}_i$. Relabel the times in $\mathcal{T}$ as $t_1, t_2, \dots, t_\tau, \dots$ so that $t_\tau < t_{\tau+1}$ for $\tau \geq 1$. It is easy to see that $\bar{T}_{\max} = \max\{\bar{T}_1, \bar{T}_2, \dots, \bar{T}_n\}$ uniformly bounds above the time interval between any two successive event times in $\mathcal{T}$.

For each $i \in \mathcal{N}$ and $t_\tau \in \mathcal{T}$, we define the *extended neighbor sets* for agent i as follows, which are for analysis purpose only. If $t_\tau \in \mathcal{T}_i$, the extended in-neighbor set of agent i, denoted $\bar{\mathcal{N}}_+^i(\tau)$, is defined as the set of those agents, including agent i itself, whose scaled variables are received by agent i during the time interval $(t_{i(q-1)}, t_{iq}]$ where $t_{iq} = t_\tau$. If $t_\tau \notin \mathcal{T}_i$, $\bar{\mathcal{N}}_+^i(\tau)$ is defined as a simply index i. In other words,

$$\begin{aligned} &\bar{\mathcal{N}}_+^i(\tau) = \{i\} \cup \{j \mid \exists s \geq 1 \text{ such that } t_{js} + d_{t_{js}}^{ji} \in (t_{i(q-1)}, t_{iq}]\}, && t_\tau \in \mathcal{T}_i, \\ &\bar{\mathcal{N}}_+^i(\tau) = \{i\}, && t_\tau \notin \mathcal{T}_i. \end{aligned}$$

It is clear that $\bar{\mathcal{N}}_{+}^{i}(\tau) \subseteq \mathcal{N}_{+}^{i}$ for all $\tau \geq 1$. Similarly, the extended out-neighbor set of agent i, denoted $\bar{\mathcal{N}}_{-}^{i}(\tau)$, is defined as

$$\begin{aligned} \bar{\mathcal{N}}_{-}^{i}(\tau) &= \mathcal{N}_{-}^{i}, & t_\tau &\in \mathcal{T}_i, \\ \bar{\mathcal{N}}_{-}^{i}(\tau) &= \{i\}, & t_\tau &\notin \mathcal{T}_i. \end{aligned}$$

Thus, $\bar{\mathcal{N}}_{-}^{i}(\tau)$ coincides with $\mathcal{N}_{-}^{i}$ whenever t_τ is an event time of agent i and the simple index i otherwise. We describe all defined neighbor relationships at time $\tau \in \{1,2,\ldots\}$ to be the time-varying directed graph $\bar{\mathcal{G}}_\tau$ with vertex set $\mathcal{N}$ and edge set $\bar{\mathcal{E}}_\tau \subset \mathcal{N} \times \mathcal{N}$ which satisfies the above extended in- and out-neighbor relationships. We call $\bar{\mathcal{G}}_\tau$ the *extended neighbor graph* of the asynchronous system under consideration at time τ (or equivalently, event time t_τ). It is worth noting that even though the real underlying neighbor graph $\mathcal{G}$ is time-invariant, the nominal graph $\bar{\mathcal{G}}_\tau$ defined on event times $t_\tau \in \mathcal{T}$ is time-varying due to asynchrony and time delays. It is worth noting that each agent i is always an extended in- and out-neighbor of itself, and thus $\bar{\mathcal{G}}_\tau$ has a self-arc at each node for all $\tau \geq 1$. More can be said. Since $\mathcal{G}$ is strongly connected, and each agent's time intervals between any two successive two event times are uniformly bounded above due to (3), it is easy to show the following result.

Lemma 1. *If $\mathcal{G}$ is strongly connected, $\{\bar{\mathcal{G}}_\tau\}_{\tau=1}^{\infty}$ is repeatedly jointly strongly connected.*

Here an infinite sequence of graphs $\bar{\mathcal{G}}_1, \bar{\mathcal{G}}_2, \ldots$ with the same vertex set is called *repeatedly jointly strongly connected* if for some positive integer l and each integer $k > 0$, the union of $\bar{\mathcal{G}}_{kl+1}, \bar{\mathcal{G}}_{kl+2}, \ldots, \bar{\mathcal{G}}_{(k+1)l}$ is strongly connected. It is also called "B-connected" in the literature [30].

We next rewrite (6) in a form which is convenient for analysis. Toward this end, fix $k \geq 1$ and $j \in \mathcal{N}_{+}^{i}$. Suppose that agent j transmits its scaled variable $\frac{v_{t_{js}}^{j}}{\mathcal{N}_{-}^{j}}$ to agent i at its event time t_{js} and agent i receives the variable at time $t \in (t_{i(k-1)}, t_{ik}]$. Agent i then holds this variable until time t_{ik} at which it is used in the computation of $\omega_{t_{i(k+1)}}^{i}$ via (6). The transmission time for this event is $t - t_{js}$ whereas the hold time is $t_{ik} - t$. Note that the hold time $t_{ik} - t$ is bounded above by $\bar{T}_i$ because of (3). We have assumed that the transmission time $t - t_{js}$ is bounded above as well. Thus, there exists a nonnegative integer $\bar{d}_{t_{ik}}^{ji}$ such that $t_{js} = t_{ik} - \bar{d}_{t_{ik}}^{ji}$. Note that t_{ik} and t_{js} are two different event times in $\mathcal{T}$. Set $t_{ik} = t_\tau$ and $t_{js} = t_\sigma$ where $\sigma, \tau \in \{1,2,\ldots\}$ and $\sigma < \tau$. We write $\bar{d}_{\tau}^{ij} = \tau - \sigma$ for the number of distinct event times in $\mathcal{T}$ during the time interval $(t_\sigma, t_\tau]$. As a consequence of (3), there must exist a bounded integer d such that $\bar{d}_{\tau}^{ij} < \bar{d}$ for all $i, j \in \mathcal{N}$ and $t_\tau \in \mathcal{T}$. Then, $v_{t_{js}}^{j} = v_{t_\tau - d_\tau^{ij}}^{j}$ and $d_\tau^{ij} \in \{0, 1, \ldots, \bar{d} - 1\}$. Since each agent i can always access the latest value of its own variables, $d_\tau^{ii} = 0$ for all $i \in \mathcal{N}$ and $t_\tau \in \mathcal{T}$. Similar arguments apply to (7).

To proceed, for each $i \in \mathcal{N}$ and $t_q \in \mathcal{T}_i$, define

$$\mu_{t_\tau}^{i} = \mu_{t_{q'}}^{i}, \quad v_{t_\tau}^{i} = v_{t_{q'}}^{i}, \quad \delta_{t_\tau}^{i} = \delta_{t_{q'}}^{i}, \quad \omega_{t_\tau}^{i} = \omega_{t_{q'}}^{i}, \quad y_{t_\tau}^{i} = y_{t_{q'}}^{i}, \quad \theta_{t_\tau}^{i} = \theta_{t_{q'}}^{i}, \quad q < \tau \leq q',$$

where $t_{q'}$ is the first event time of agent i after t_q. Note that for any $t_q \in \mathcal{T}_i$, there always exists such a q' because of (3). Then, each agent's variables are well defined at any other agent's event times, so at any event time in $\mathcal{T}$.

Now we define a new set of variables to conveniently describe a synchronous system, which is equivalent to the asynchronous system under consideration, as follows:

$$\bar{\mu}_{\tau}^{i} = \mu_{t_\tau}^{i}, \quad \bar{v}_{\tau}^{i} = v_{t_\tau}^{i}, \quad \bar{\delta}_{\tau}^{i} = \delta_{t_\tau}^{i}, \quad \bar{\omega}_{\tau}^{i} = \omega_{t_\tau}^{i}, \quad \bar{y}_{\tau}^{i} = y_{t_\tau}^{i}, \quad \bar{z}_{\tau}^{i} = z_{t_\tau}^{i}, \quad \bar{\theta}_{\tau}^{i} = \theta_{t_\tau}^{i}.$$

We also need define each agent's stepsizes for the new variables at its own and other agents' event times. Specifically, for all $i \in \mathcal{N}$ and $t_\tau \in \mathcal{T}$, define

$$\begin{aligned} \bar{\beta}_{\bar{\omega},\tau} &= \beta_{\omega,t_\tau}, & \bar{\beta}_{\bar{\theta},\tau} &= \beta_{\theta,t_\tau}, & t_\tau &\in \mathcal{T}_i, \\ \bar{\beta}_{\bar{\omega},\tau} &= 0, & \bar{\beta}_{\bar{\theta},\tau} &= 0, & t_\tau &\notin \mathcal{T}_i. \end{aligned}$$

It is easy to verify that the stepsizes $\bar{\beta}_{\bar{\omega},\tau}$ and $\bar{\beta}_{\bar{\theta},\tau}$ satisfy Assumption 3, stated as follows.

Lemma 2. *Suppose that Assumption 3 holds. Then, the stepsizes $\bar{\beta}_{\bar{\omega},\tau}$ and $\bar{\beta}_{\bar{\theta},\tau}$ satisfy, for all $i \in \mathcal{N}$, $\sum_{\tau \geq 1} \bar{\beta}_{\bar{\omega},\tau} = \sum_{\tau \geq 1} \bar{\beta}_{\bar{\theta},\tau} = \infty$ and $\sum_{\tau \geq 1}(\bar{\beta}^2_{\bar{\omega},\tau} + \bar{\beta}^2_{\bar{\theta},\tau}) < \infty$. In addition, $\bar{\beta}_{\bar{\theta},\tau} = o(\bar{\beta}_{\bar{\omega},\tau})$.*

The preceding discussion enables us to extend the domain of applicability of the asynchronous algorithm under consideration from $\mathcal{T}_i$ to all of $\mathcal{T}$, which leads to a synchronous system. The synchronous algorithm also consists of two steps in each iteration, a critic step followed by an actor step. For each $\tau \geq 1$, the critic step of agent i is as follows:

$$\begin{cases} \bar{\mu}^i_{\tau+1} = (1 - \bar{\beta}_{\bar{\omega},\tau}) \cdot \bar{\mu}^i_\tau + \bar{\beta}_{\bar{\omega},\tau} \cdot r^i_{\tau+1}, & (14) \\ \bar{v}^i_\tau = \bar{\omega}^i_\tau + \bar{\beta}_{\bar{\omega},\tau} \cdot \bar{\delta}^i_\tau \cdot \nabla_z Q_\tau(z^i_\tau), & (15) \\ \bar{\omega}^i_{\tau+1} = \dfrac{\bar{v}^i_\tau}{|\mathcal{N}^i_-|} + \displaystyle\sum_{j \in \mathcal{N}^i_+(\tau)} \sum_{s \geq 1,\, t_s \in \mathcal{T}_j} \frac{\bar{v}^j_s \chi_{(\sigma,\tau]}(s + \bar{d}^{ji}_s)}{|\mathcal{N}^j_-|}, & (16) \\ \bar{y}^i_{\tau+1} = \dfrac{\bar{y}^i_\tau}{|\mathcal{N}^i_-|} + \displaystyle\sum_{j \in \mathcal{N}^i_+(\tau)} \sum_{s \geq 1,\, t_s \in \mathcal{T}_j} \frac{\bar{y}^j_s \chi_{(\sigma,\tau]}(s + \bar{d}^{ji}_s)}{|\mathcal{N}^j_-|}, & (17) \\ \bar{z}^i_{\tau+1} = \dfrac{\bar{\omega}^i_{\tau+1}}{\bar{y}^i_{\tau+1}}, & (18) \end{cases}$$

where σ is the largest integer such that $t_\sigma \in \mathcal{T}$ is an event time in $\mathcal{T}_i$ before t_τ, $\chi_{(\sigma,\tau]}(s + \bar{d}^{ji}_s)$ is an indicator function defined as $\chi_{(\sigma,\tau]}(s + \bar{d}^{ji}_s) = 1$ if $\sigma < s + \bar{d}^{ji}_s \leq \tau$, and otherwise $\chi_{(\sigma,\tau]}(s + \bar{d}^{ji}_s) = 0$. The local action-value TD-error $\bar{\delta}^i_\tau$ in (15) is given by

$$\bar{\delta}^i_\tau = r^i_{\tau+1} - \bar{\mu}^i_\tau + Q_{\tau+1}(\bar{z}^i_\tau) - Q_\tau(\bar{z}^i_\tau). \tag{19}$$

The actor step of agent i is as follows:

$$\bar{\theta}^i_{\tau+1} = \bar{\theta}^i_\tau + \bar{\beta}_{\bar{\theta},\tau} \cdot A^i_\tau \cdot \psi^i_\tau, \tag{20}$$

where A^i_t and ψ^i_t are defined in (11) and (12), respectively.

Since the asynchronous algorithm under consideration has the same limiting behavior as the synchronous algorithm just described, Theorem 1 is an immediate consequence of the following result.

Proposition 1. *Suppose that Assumptions 1 – 4 hold, and that communication graph $\mathcal{G}$ is strongly connected. Then, for any given policy π_θ, with the sequences $\{\bar{\mu}^i_\tau\}$ and $\{\bar{z}^i_\tau\}$ generated from (14) and (18), we have $\lim_{\tau \to \infty} \sum_{i \in \mathcal{N}} \bar{\mu}^i_\tau \cdot N^{-1} = J(\theta)$ and $\lim_{\tau \to \infty} \bar{z}^i_\tau = \omega_\theta$ almost surely for any $i \in \mathcal{N}$, where $J(\theta)$ is the globally averaged return as defined in (2), and ω_θ is the unique solution to*

$$\Phi^\top \mathrm{D}^{s,a}_\theta \big[T^Q_\theta(\Phi\omega_\theta) - \Phi\omega_\theta\big] = 0.$$

Suppose further that Assumption 5 holds. Then, for all $i \in \mathcal{N}$, the sequence $\{\bar{\theta}^i_\tau\}$ obtained from (20) converges almost surely to a point in the set of the asymptotically stable equilibria of

$$\dot{\theta}^i = \hat{\Gamma}^i\big[\mathbb{E}_{s_t \sim d_\theta, a_t \sim \pi_\theta}\big(A^i_{t,\theta} \cdot \psi^i_{t,\theta}\big)\big].$$

The proof of this proposition can be found in the Appendix.

IV. Simulation

We evaluate a setting in which linear function approximation is adopted. Consider in total $N = 20$ agents, each having a binary-valued action space, i.e., $\mathcal{A}^i = \{0, 1\}$, for all $i \in \mathcal{N}$. Thus, the cardinality of the set of actions $\mathcal{A}$ is 2^{20}. In addition, there are in total $|\mathcal{S}| = 20$ states. The elements in the transition probability matrix P are uniformly sampled from the interval $[0,1]$ and normalized to be a stochastic matrix. We also add a small constant 10^{-5} onto each element in the matrix to ensure ergodicity of the MDP such that Assumption 1 is satisfied. For each agent i and each state-action pair (s,a), the mean reward $R^i(s,a)$ is sampled uniformly from $[0,4]$, which varies among agents. The instantaneous rewards r_t^i are sampled from the uniform distribution $[R^i(s,a)-0.5, R^i(s,a)+0.5]$. The policy $\pi^i_{\theta^i}(s,a^i)$ is parameterized following the Boltzman policies, i.e.,

$$\pi^i_{\theta^i}(s,a^i) = \frac{\exp\left(q_{s,b^i}^{\intercal}\theta^i\right)}{\sum_{b^i\in\mathcal{A}^i}\exp\left(q_{s,b^i}^{\intercal}\theta^i\right)}$$

where $q_{s,b^i}\in\mathbb{R}^{m_i}$ is the feature vector with the same dimension as θ^i, for any $s\in\mathcal{S}$ and $i\in\mathcal{N}$. Here we set $m_1 = m_2 = \cdots = m_N = 5$. The elements of q_{s,b^i} are also uniformly sampled from $[0,1]$. In particular, the gradient of the score function thus has the form

$$\nabla_{\theta^i}\log\pi^i_{\theta^i}(s,a^i) = q_{s,a^i} - \sum_{b^i\in\mathcal{A}^i}\pi^i_{\theta^i}(s,a^i)q_{s,b^i}.$$

The feature vectors $\varphi\in\mathbb{R}^K$ for the action-value function $Q(\cdot,\cdot;\omega)$ are all uniformly sampled from $[0,1]$, of dimensions $K = 5 \ll |\mathcal{S}|\cdot|\mathcal{A}|$.

The communication graph $\mathcal{G}$ is fixed and strongly connected. The stepsizes are selected as $\beta_{\omega,t_{ik}} = 1/k^{0.65}$ and $\beta_{\theta,t_{ik}} = 1/k^{0.85}$, which satisfy Assumption 2. For each agent, the time between two consecutive events Δt_{ik} are uniformly sampled from $[0.5,1.5]$, so that $\mathbb{E}[\Delta t_{ik}] = 1$. The delay time $d_{t_{ik}}$ is uniformly sampled from $[0,2]$. Figure 1 shows the convergence of relative Q-value functions of the asynchronous algorithm under linear function approximation.

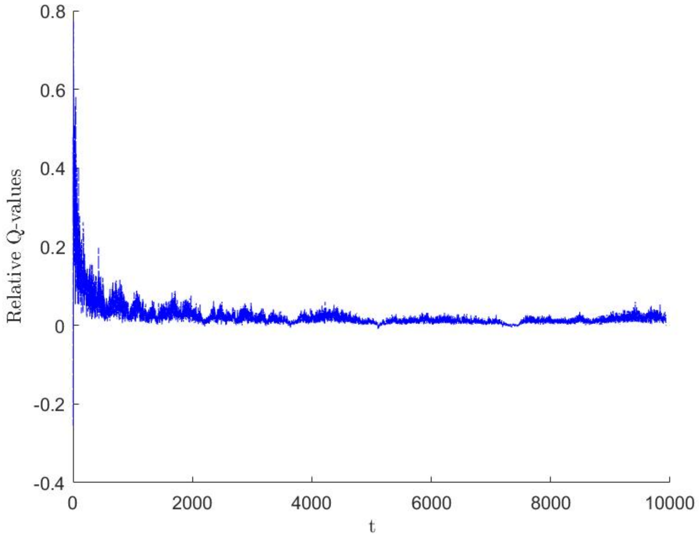

Figure 1: Convergence of relative Q-value functions

V. Conclusion

In this paper, we have proposed an asynchronous distributed actor-critic algorithm for solving a networked reinforcement learning problem. We have shown that the algorithm converges under linear function approximation. One extension would be to relax the uniform boundedness of consecutive times of action of an agent to a probabilistic one, as in [31]. For future work, we will consider other communication issues such as packet drops and other reinforcement learning algorithms such as off-policy actor-critic algorithms.

VI. Appendix

This appendix material provides a complete proof of Proposition 1.

We first construct a new graph, denoted $\bar{\mathcal{G}}_\tau$, based on $\mathcal{G}_\tau$ for each τ for analysis purpose as follows. We regard each agent i as node $n^{i(0)}$ in $\bar{\mathcal{G}}_\tau$ and introduce virtual nodes $n^{i(1)}, \ldots, n^{i(\bar{d}-1)}$ for each agent i, where $\bar{d}-1$ is the maximum delay time among all agents. At each time τ, virtual node $n^{i(k)}$ holds the sum of the values that will be received by node $n^{i(0)}$ in k time steps. Besides the directed edges in $\mathcal{G}_\tau$ including self-arcs at all nodes $n^{i(0)}$, $i \in \mathcal{N}$, we add the following directed edges: for each directed edge $(n^{j(0)}, n^{i(0)})$ in $\mathcal{G}_\tau$, we add directed edges $(n^{j(1)}, n^{i(0)}), (n^{j(2)}, n^{i(0)}), \ldots, (n^{j(\bar{d}-1)}, n^{i(0)})$ and $(n^{j(0)}, n^{j(1)}), (n^{j(1)}, n^{j(2)}), \ldots, (n^{j(\bar{d}-2)}, n^{j(\bar{d}-1)})$ for all $i, j \in \mathcal{N}$. Thus, each $\bar{\mathcal{G}}_\tau$ has $N\bar{d}$ nodes, and not all $N\bar{d}$ nodes have a self-arc.

Let $\bar{\omega}_\tau^{i(0)}, v_\tau^{i(0)}$ and $\bar{y}_\tau^{i(0)}$ be the values of $\bar{\omega}_\tau^i$, $\bar{v}_\tau^i$ and $\bar{y}_\tau^i$ respectively for each agent i, which are given in equations (15)-(17). Let

$$\begin{aligned}
W_\tau &= [(\bar{\omega}_\tau^{1(0)})^\top, \cdots, (\bar{\omega}_\tau^{N(0)})^\top, (\bar{\omega}_\tau^{1(1)})^\top, \cdots, (\bar{\omega}_\tau^{N(1)})^\top, \cdots, (\bar{\omega}_\tau^{1(\bar{d}-1)})^\top, \cdots, (\bar{\omega}_\tau^{N(\bar{d}-1)})^\top]^\top, \\
\tilde{W}_\tau &= [(\bar{v}_\tau^{1(0)})^\top, \cdots, (\bar{v}_\tau^{N(0)})^\top, (\bar{v}_\tau^{1(1)})^\top, \cdots, (\bar{v}_\tau^{N(1)})^\top, \cdots, (\bar{v}_\tau^{1(\bar{d}-1)})^\top, \cdots, (\bar{v}_\tau^{N(\bar{d}-1)})^\top]^\top, \\
Y_\tau &= [\bar{y}_\tau^{1(0)}, \cdots, \bar{y}_\tau^{N(0)}, \bar{y}_\tau^{1(1)}, \cdots, \bar{y}_\tau^{N(1)}, \cdots, \bar{y}_\tau^{1(\bar{d}-1)}, \cdots, \bar{y}_\tau^{N(\bar{d}-1)}]^\top, \\
\bar{z}_\tau &= [(\bar{z}_\tau^1))^\top, \cdots, (\bar{z}_\tau^N)^\top]^\top, \\
\tilde{U}_\tau &= [(\bar{\delta}_\tau^1 \cdot \nabla_z Q(\bar{z}_\tau^1))^\top, \cdots, (\bar{\delta}_\tau^N \cdot \nabla_z Q(\bar{z}_\tau^N))^\top, 0_{NK(\bar{d}-1)\times 1}^\top]^\top,
\end{aligned}$$

where $\bar{\omega}_\tau^{i(s)}$, $\bar{v}_\tau^{i(s)}$ and $\bar{y}_\tau^{i(s)}$ denote the variables of nodes $n^{i(s)}$ at time τ, respectively, $\bar{z}_\tau^i = [\bar{z}_\tau^i(1), \cdots, \bar{z}_\tau^i(K)]^\top$, and $\bar{z}_\tau^i(k) = \bar{\omega}_\tau^{i(0)}(k)/\bar{y}_\tau^{i(0)}$, where $\bar{\omega}_\tau^{i(0)}(k)$ denotes the kth entry of vector $\bar{\omega}_\tau^{i(0)}$, $\forall i = 1, \cdots, N, k = 1, \ldots, K$.

Let $\bar{\chi}_\tau^{ji}(d)$ be an indicator function defined as $\bar{\chi}_\tau^{ji}(d) = 1$ if agent j sends information to agent i at time $t_{\tau-d}$ (agent i updates at time t_τ), otherwise $\bar{\chi}_\tau^{ji}(d) = 0$. Then, we can rewrite the equations (16) and (17) as (21) and (22) in the following update:

$$\begin{cases}
\bar{\mu}_{\tau+1}^i = (1 - \bar{\beta}_{\bar{\omega},\tau}) \cdot \bar{\mu}_\tau^i + \bar{\beta}_{\bar{\omega},\tau} \cdot r_{\tau+1}^i, \\
\bar{v}_\tau^i = \bar{\omega}_\tau^i + \bar{\beta}_{\bar{\omega},\tau} \cdot \bar{\delta}_\tau^i \cdot \nabla_z Q_\tau(z_\tau^i), \\
\bar{\omega}_{\tau+1}^i = \dfrac{\bar{v}_\tau^i}{|\mathcal{N}_-^i|} + \displaystyle\sum_{j\in\mathcal{N}} \sum_{d=0}^{\bar{d}-1} \frac{\bar{v}_{\tau-d}^j \chi_\tau^{ji}(d)}{|\mathcal{N}_-^j|}, & (21) \\
\bar{y}_{\tau+1}^i = \dfrac{\bar{y}_\tau^i}{|\mathcal{N}_-^i|} + \displaystyle\sum_{j\in\mathcal{N}} \sum_{d=1}^{\bar{d}-1} \frac{\bar{y}_{\tau-d}^j \chi_\tau^{ji}(d)}{|\mathcal{N}_-^j|}, & (22) \\
\bar{z}_{\tau+1}^i = \dfrac{\bar{\omega}_{\tau+1}^i}{\bar{y}_{\tau+1}^i}.
\end{cases}$$

Let

$$H_\tau^s = \begin{bmatrix} \frac{\chi_\tau^{11}(s)}{|\mathcal{N}_-^1|} & \frac{\chi_\tau^{21}(s)}{|\mathcal{N}_-^2|} & \cdots & \frac{\chi_\tau^{N1}(s)}{|\mathcal{N}_-^N|} \\ \frac{\chi_\tau^{12}(s)}{|\mathcal{N}_-^1|} & \frac{\chi_\tau^{22}(s)}{|\mathcal{N}_-^2|} & \cdots & \frac{\chi_\tau^{N2}(s)}{|\mathcal{N}_-^N|} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\chi_\tau^{1N}(s)}{|\mathcal{N}_-^1|} & \frac{\chi_\tau^{2N}(s)}{|\mathcal{N}_-^2|} & \cdots & \frac{\chi_\tau^{NN}(s)}{|\mathcal{N}_-^N|} \end{bmatrix}, \quad \forall s = 0, 1, \cdots, \bar{d} - 1,$$

$$H_\tau = \begin{bmatrix} H_\tau^0 & I_N & 0_N & \cdots & 0_N \\ H_\tau^1 & 0_N & I_N & \cdots & 0_N \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ H_\tau^{\bar{d}-2} & 0_N & 0_N & \cdots & I_N \\ H_\tau^{\bar{d}-1} & 0_N & 0_N & \cdots & 0_N \end{bmatrix},$$

where I_N and 0_N denote the $N \times N$ identity matrix and zero matrix, repectively. Then, the above update can be written in a compact state form:

$$\begin{cases} \tilde{W}_\tau = W_\tau + \bar{\beta}_{\bar{\varpi},\tau} \cdot \tilde{U}_\tau, \\ W_{\tau+1} = H_\tau \otimes I_K \cdot \tilde{W}_\tau, \\ Y_{\tau+1} = H_\tau \cdot Y_\tau. \end{cases} \tag{23}$$

Define the operator $\langle \cdot \rangle_1 : \mathbb{R}^{KN} \to \mathbb{R}^K$ and $\langle \cdot \rangle_2 : \mathbb{R}^{KN\bar{d}} \to \mathbb{R}^K$ as

$$\langle x \rangle_1 = \frac{1}{N}(\mathbf{1}_N^\top \otimes I_K)x = \frac{1}{N}\sum_{i\in\mathcal{N}} x^i,$$

$$\langle y \rangle_2 = \frac{1}{N}(\mathbf{1}_{N\bar{d}}^\top \otimes I_K)y = \frac{1}{N}\sum_{i\in\mathcal{N}}\sum_{s=0}^{\bar{d}-1} y^{i(s)},$$

for any $x = [(x^1)^\top, \ldots, (x^N)^\top]^\top \in \mathbb{R}^{KN}$ with $x^i \in \mathbb{R}^K$, and
$y = [(y^{1(0)})^\top, \ldots, (y^{N(0)})^\top, \cdots, (y^{1(\bar{d}-1)})^\top, \ldots, (y^{N(\bar{d}-1)})^\top]^\top \in \mathbb{R}^{KN\bar{d}}$ with $y^{i(s)} \in \mathbb{R}^K$ for all $i \in \mathcal{N}$, $s = 0, \cdots, \bar{d} - 1$.

Lemma 3. *There exists a constant $\alpha > 0$ such that $\alpha \leq \bar{y}_\tau^{i(0)} \leq N$ for any i and τ almost surely.*

Proof: From Proposition 1 in [32], when $\tau > N\bar{d}$, there exists a positive constant $c_{\min}$ such that the first N rows of matrix product $\Pi_{s=0}^\tau H_s$ are strictly positive with minimum entry greater than or equal to $c_{\min}$. Moreover, we have the update $Y_\tau = \Pi_{s=0}^\tau H_s \cdot Y_0$. Then, when $\tau > N\bar{d}$, there exists a positive constant α for which $0 \leq \alpha \leq N \cdot c_{\min}$ and the first N entries of Y_τ are greater than or equal to α. Thus, we know that $\alpha \leq \bar{y}_\tau^{i(0)} \leq N$ for all i and τ. ■

To proceed, let

$$W_\tau^k = [\bar{\varpi}_\tau^{1(0)}(k), \cdots, \bar{\varpi}_\tau^{N(0)}(k), \cdots, \bar{\varpi}_\tau^{1(\bar{d}-1)}(k), \cdots, \bar{\varpi}_\tau^{N(\bar{d}-1)}(k)]^\top,$$
$$\tilde{W}_\tau^k = [\bar{v}_\tau^{1(0)}(k), \cdots, \bar{v}_\tau^{N(0)}(k), \cdots, \bar{v}_\tau^{1(\bar{d}-1)}(k), \cdots, \bar{v}_\tau^{N(\bar{d}-1)}(k)]^\top,$$
$$\tilde{U}_\tau^k = [\tilde{u}_\tau^{1(0)}(k), \cdots, \tilde{u}_\tau^{N(0)}(k), \cdots, \tilde{u}_\tau^{1(\bar{d}-1)}(k), \cdots, \tilde{u}_\tau^{N(\bar{d}-1)}(k)]^\top.$$

Then, we have

$$\begin{cases} \tilde{W}_\tau^k = W_\tau^k + \bar{\beta}_{\bar{\varpi},\tau}\tilde{U}_\tau^k, \\ W_{\tau+1}^k = H_\tau \tilde{W}_\tau^k, \\ Y_{\tau+1} = H_\tau Y_\tau, \\ \bar{z}_{\tau+1}^i(k) = \dfrac{W_{\tau+1}^k(i)}{Y_{\tau+1}(i)}, \quad \forall i = 1, \cdots, N, \end{cases}$$

where $W_{\tau+1}^k(i)$ and $Y_{\tau+1}(i)$ are the ith entry of vector $W_{\tau+1}^k$ and $Y_{\tau+1}$, respectively.

Lemma 4. *For all $i = 1, \cdots, N$, and $k = 1, \cdots, K$, $\lim_{\tau\to\infty} \bar{z}_\tau^{ik} = \lim_{\tau\to\infty} \mathbf{1}_{N\bar{d}}^\top W_\tau^k / N$.*

Proof: Let $H(\tau : s) = \Pi_{k=s}^\tau H_k$. There exists a vector l_τ such that $|[H(t : s)]_{i,j} - l_\tau^i| \leq C\lambda^{\tau-s}, \forall \tau \geq s \geq 0$, where C is a constant and $\lambda \in (0,1)$. Let $D(\tau, s) = H(\tau, s) - l_\tau \mathbf{1}^\top$. Then, we have

$$\tilde{W}_{\tau+1}^k = H(\tau : 0)\tilde{W}_0^k + \sum_{s=1}^{\tau} H(\tau : s)\tilde{U}_s^k \bar{\beta}_{\bar{\omega},s} + \tilde{U}_{\tau+1}^k \bar{\beta}_{\bar{\omega},\tau+1},$$

$$\mathbf{1}^\top \tilde{W}_{\tau+1}^k = \mathbf{1}^\top \tilde{W}_0^k + \sum_{s=1}^{\tau+1} \mathbf{1}^\top \tilde{U}_s^k \bar{\beta}_{\bar{\omega},s},$$

$$H_{\tau+1}\tilde{W}_{\tau+1}^k - l_{\tau+1}\mathbf{1}^\top \tilde{W}_{\tau+1}^k = (H(\tau+1 : 0) - l_{t+1}\mathbf{1}^\top)\tilde{W}_0^k + \sum_{s=1}^{\tau+1}(H(\tau+1 : s) - l_{\tau+1}\mathbf{1}^\top)\tilde{U}_s^k \bar{\beta}_{\bar{\omega},s}.$$

From this, we have, for $\tau \geq 1$,

$$W_{\tau+1}^k = H_\tau \tilde{W}_\tau^k = l_\tau \mathbf{1}^\top \tilde{W}_\tau^k + D(\tau, 0)\tilde{W}_0^k + \sum_{s=1}^{\tau} D(\tau, s)\tilde{U}_s^k \bar{\beta}_{\bar{\omega},s},$$

$$Y_{\tau+1} = H(\tau : 0)Y_0 = l_\tau \mathbf{1}^\top Y_0 + D(\tau, 0)Y_0 = Nl_\tau + D(\tau, 0)Y_0.$$

Thus, for every $\tau \geq 1$ and for all i,

$$\begin{aligned} \bar{z}_{\tau+1}^i(k) &= \frac{W_{\tau+1}^k(i)}{Y_{\tau+1}(i)} \\ &= \frac{l_\tau(i)\mathbf{1}^\top \tilde{W}_\tau^k + [D(\tau, 0)\tilde{W}_0^k](i) + \sum_{s=1}^\tau [D(\tau, s)\tilde{U}_s^k \bar{\beta}_{\bar{\omega},s}](i)}{Nl_\tau(i) + [D(\tau, 0)Y_0](i)}. \end{aligned}$$

Therefore,

$$\begin{aligned} \bar{z}_{\tau+1}^i(k) - \frac{\mathbf{1}^\top \tilde{W}_\tau^k}{N} &= \frac{l_\tau(i)\mathbf{1}^\top \tilde{W}_\tau^k + [D(\tau, 0)\tilde{W}_0^k](i) + \sum_{s=1}^\tau [D(\tau, s)\tilde{U}_s^k \bar{\beta}_{\bar{\omega},s}](i)}{Nl_\tau(i) + [D(\tau, 0)Y_0](i)} - \frac{\mathbf{1}^\top \tilde{W}_\tau^k}{N} \\ &= \frac{[D(\tau, 0)\tilde{W}_0^k](i) + \sum_{s=1}^\tau \tau[D(\tau, s)\tilde{U}_s^k \bar{\beta}_{\bar{\omega},s}](i)}{Nl_\tau(i) + [D(\tau, 0)Y_0](i)} - \frac{\mathbf{1}^\top \tilde{W}_\tau^k [D(\tau : 0)Y_0](i)}{N[Nl_\tau(i) + [D(\tau : 0)Y_0](i)]}. \end{aligned}$$

By definition of α, $\tilde{W}_\tau^k(i)N + [D(\tau, 0)\mathbf{1}](i) = [H(\tau : 0) \cdot [\mathbf{1}_N^\top, 0_{N(\bar{d}-1)}^\top]^\top](i) \geq \alpha, i = 1, \cdots, N$. Thus, for all $i = 1, \cdots, N$ and $\tau \geq 1$, we have

$$\begin{aligned} |\bar{z}_{\tau+1}^{ik} - \frac{\mathbf{1}^\top \bar{W}_\tau^k}{N}| &\leq \frac{|[D(\tau, 0)\tilde{W}_0^k](i) + \sum_{s=1}^\tau [D(\tau, s)\tilde{U}_s^k \bar{\beta}_{\bar{\omega},s}](i)|}{Nl_\tau(i) + [D(\tau, 0)Y_0](i)} + \frac{|\mathbf{1}^\top \tilde{W}_\tau^k [D(\tau, 0)Y_0](i)|}{N[Nh_\tau(i) + [D(\tau, 0)Y_0](i)]} \\ &\leq \frac{1}{\alpha}[\max_j |[D(\tau, 0)](i, j)| \cdot \|\tilde{W}_0^k\|_1 + \sum_{s=1}^{\tau} \max_j |[D(\tau, s)](i, j)| \cdot \|\tilde{U}_s^k\|_1 \bar{\beta}_{\bar{\omega},s}] \\ &\quad + \frac{1}{\alpha}|\mathbf{1}^\top \tilde{W}_\tau^k| \cdot \max_j |[D(\tau, 0)](i, j)| \\ &\leq \frac{C\lambda^\tau}{\alpha}\|\tilde{W}_0^k\|_1 + \frac{C}{\alpha}\sum_{s=1}^{\tau} \lambda^{\tau-s}\|\tilde{U}_s^k\|_1 \bar{\beta}_{\bar{\omega},s} + \frac{C\lambda^\tau}{\alpha}|\mathbf{1}^\top \tilde{W}_\tau^k| \\ &= \frac{C\lambda^\tau}{\alpha}\|\tilde{W}_0^k\|_1 + \frac{C}{\alpha}\sum_{s=1}^{\tau} \lambda^{\tau-s}\|\tilde{U}_s^k\|_1 \bar{\beta}_{\bar{\omega},s} + \frac{C\lambda^\tau}{\alpha}|\mathbf{1}^\top \tilde{W}_0^k + \sum_{s=1}^{\tau+1} \mathbf{1}^\top \tilde{U}_s^k \bar{\beta}_{\bar{\omega},s}| \\ &\leq \frac{C\lambda^\tau}{\alpha}\|\tilde{W}_0^k\|_1 + \frac{C}{\alpha}\sum_{s=1}^{\tau} \lambda^{\tau-s}\|\tilde{U}_s^k\|_1 \bar{\beta}_{\bar{\omega},s} + \frac{C\lambda^\tau}{\alpha}\|\tilde{W}_0^k\|_1 + \frac{C\lambda^\tau}{\alpha}\sum_{s=1}^{\tau+1} \|\tilde{U}_s^k\|_1 \bar{\beta}_{\bar{\omega},s} \\ &\leq \frac{2C}{\alpha}[\lambda^\tau \cdot \|\tilde{W}_0^k\|_1 + \sum_{s=1}^{\tau} \lambda^{\tau-s}\|\tilde{U}_s^k\|_1 \bar{\beta}_{\bar{\omega},s}], \end{aligned}$$

where $[D(\tau,0)](i,j)$ is the entry at ith row and jth column of matrix $D(\tau,0)$. Since $\lambda \in (0,1)$, for all agent i,

$$\begin{aligned}\lim_{\tau\to\infty} |\bar{z}^i_{\tau+1}(k) - \frac{\mathbf{1}^\top \tilde{W}^k_\tau}{N}| &\le \lim_{\tau\to\infty} \frac{2C}{\alpha}[\lambda^\tau \cdot \|\tilde{W}^k_0\|_1 + \sum_{s=1}^{\tau} \lambda^{\tau-s}\|\tilde{U}^k_s\|_1 \bar{\beta}_{\bar{\omega},s}] \\ &\le \lim_{\tau\to\infty} \frac{2C}{\alpha}\sum_{s=1}^{\tau} \lambda^{\tau-s}\|\tilde{U}^k_s\|_1 \bar{\beta}_{\bar{\omega},s}].\end{aligned}$$

Since $\tilde{U}^k_s(i) \to 0$ for every agent i, then $\|\tilde{U}^k_s\|_1 \to 0$, and from the Lemma 5(a) in [30], we have

$$\lim_{\tau\to\infty} |\bar{z}^i_{\tau+1}(k) - \frac{\mathbf{1}^\top \tilde{W}^k_\tau}{N}| \le \lim_{\tau\to\infty} \frac{2C}{\alpha}\sum_{s=1}^{\tau} \lambda^{\tau-s}\|\tilde{U}^k_s\|_1 \bar{\beta}_{\bar{\omega},s} = 0.$$

This completes the proof. ■

From Lemma 4, since it is easy to show the following result, the proof is omitted.

Lemma 5. *For all $i \in \mathcal{N}$, $\lim_{\tau\to\infty} \bar{z}^i_\tau = \lim_{\tau\to\infty}\langle W_\tau\rangle_2 = \lim_{\tau\to\infty}\langle \bar{z}_\tau\rangle_1$.*

Lemma 6. *Under Assumption 1 and Lemma 2, the sequence $\{\bar{\mu}^i_\tau\}$ generated as in* (16) *is bounded almost surely.*

Proof: The proof of the lemma is the same as that of Lemma 5.2 in [4]. ■

Lemma 7. *Under Assumptions 1, 2, 4 and Lemma 2, the sequence $\{\bar{\omega}^{i(k)}_\tau\}$ is bounded almost surely, i.e., $\sup_\tau \|W_\tau\| < \infty$.*

Proof: Recall that the update of W is $W_{\tau+1} = H_\tau \otimes I_K \cdot (W_\tau + \beta_{\bar{\omega},t} \cdot \tilde{U}_\tau)$ given in (23). Let $\tilde{U}_\tau = [\tilde{u}^{1(0)}_\tau, \cdots, \tilde{u}^{N(0)}_\tau, \cdots, \tilde{u}^{1(\bar{d}-1)}_\tau, \cdots, \tilde{u}^{N(\bar{d}-1)}_\tau]$. From the definition of $\tilde{U}_\tau$, for all agent i,

$$\begin{cases} \tilde{u}^{i(0)}_\tau = (r^i_{\tau+1} - \bar{\mu}^i_\tau + (\phi_{\tau+1} - \phi_\tau)\bar{\omega}^{i(0)}_\tau / \bar{y}^{i(0)}_\tau)\phi_\tau, \\ \tilde{u}^{i(s)}_\tau = 0, \quad \forall s > 0. \end{cases}$$

Moreover, we have

$$\bar{\omega}^{i(s)}_{\tau+1} = \sum_{l=0}^{\bar{d}-1}\sum_{j=1}^{N} H_\tau(i+sN, j+lN)(\bar{\omega}^{j(l)}_\tau + \bar{\beta}_{\bar{\omega},t}\tilde{u}^{j(l)}_\tau).$$

Let $\{\mathcal{F}_{\tau,1}\}$ be the filtration with $\mathcal{F}_{\tau,1} = \sigma(r_l, \bar{\mu}_l, W_l, \bar{z}_l, Y_l, s_l, a_l, B_{l-1}, l < \tau)$, and

$$h^{i(k)}(\bar{\omega}^{i(k)}_\tau, \bar{\mu}^i_\tau, \bar{y}^{i(k)}_\tau, s_\tau, a_\tau) = \mathbb{E}(\tilde{u}^{i(k)}_\tau | \mathcal{F}_{\tau,1}), M^{i(k)}_{\tau+1} = \tilde{u}^{i(k)}_\tau - \mathbb{E}(\tilde{u}^{i(k)}_\tau | \mathcal{F}_{\tau,1}).$$

Since the Markov chain $\{(s_\tau, a_\tau)\}_{\tau\ge 0}$ is irreducible and aperiodic given policy π_θ, we have that when $s = 0$, $\bar{h}^{i(0)}(\bar{\omega}^{i(0)}_\tau, \bar{\mu}^i_\tau, \bar{y}^{i(0)}_\tau) = \mathbb{E}_{s_\tau\sim d_\theta, a_\tau\sim\pi_\theta}[h^{i(0)}(\bar{\omega}^{i(0)}_\tau, \bar{\mu}^i_t, \bar{y}^{i(0)}_\tau, s_\tau, a_\tau)] = \Phi^\top D^{s,a}_\theta[R^i - \bar{\mu}^i_\tau \mathbf{1}_K + \frac{1}{\bar{y}^{i(0)}_\tau}(P^\theta\Phi - \Phi)\bar{\omega}^{i(0)}_\tau]$ and for $k > 0$, $\bar{h}^{i(k)}(\bar{\omega}^{i(k)}_\tau, \bar{\mu}^i_\tau, \bar{y}^{i(k)}_\tau) = \mathbb{E}_{s_\tau\sim d_\theta, a_\tau\sim\pi_\theta}[h^{i(k)}(\bar{\omega}^{i(k)}_\tau, \bar{\mu}^i_\tau, \bar{y}^{i(k)}_\tau, s_\tau, a_\tau)] = 0$. From Assumptions 2 and 4, and Lemmas 3 and 6, we know that $\exists K_1, K_2 > 0$, s.t. $\|\frac{\phi^k_\tau}{\bar{y}^{i(0)}_\tau}\|_\infty \le K_1$, and $\|r^i_{\tau+1} - \bar{\mu}^i_\tau\| \le K_2, \forall k, i$. Thus, $\exists K_3 > 0$ such that $\|\bar{h}^{i(k)}(\bar{\omega}^{i(k)}_\tau, \bar{\mu}^i_\tau, \bar{y}^{i(k)}_\tau) - h^{i(k)}(\bar{\omega}^{i(k)}_\tau, \bar{\mu}^i_\tau, \bar{y}^{i(k)}_\tau, s_\tau, a_\tau)\|^2 \le K_3 \cdot (1 + \|W_\tau\|^2)$. Moreover, we know $h^{i(k)}(\bar{\omega}^{i(k)}_\tau, \bar{\mu}^i_\tau, \bar{y}^{i(k)}_\tau, s_\tau, a_\tau)$ is Lipschitz continuous in W^i_τ, and $M^i_{\tau+1}$ is a martingale difference sequence. Since H_τ is a column stochastic matrix, it has bounded norm. Thus, by Theorem A.2 in [4], W^i_τ is bounded almost surely. ■

Lemma 8. *Under Assumptions 1, 2, 4 and Lemma 2, the sequence $\{\bar{z}^i_\tau\}$ is bounded almost surely, i.e., $\sup_\tau \|\bar{z}^i_\tau\| < \infty$, $\forall i = 1, \cdots, N$.*

Proof: From (16), we know that for each entry k in $\bar{z}^i_\tau$, $\bar{z}^i_\tau(k) = \bar{\omega}^{i(0)}_\tau(k)/\bar{y}^{i(0)}_\tau$, $k \in \{1, \ldots, K\}$. Moreover, from Lemmas 3 and 7, $\bar{\omega}^{i(0)}_\tau$ and $\bar{y}^{i(0)}_\tau$ are bounded almost surely. Therefore, it is easy to show that $\bar{z}_\tau$ is also bounded almost surely. ■

We are now in a position to prove Proposition 1.

Proof of Proposition 1: The iteration of $\langle W_\tau \rangle_2$ has the following form:

$$\begin{aligned}
\langle W_{\tau+1} \rangle_2 &= \frac{1}{N}(\mathbf{1}^\top_{N\bar{d}} \otimes I_K) H_\tau \otimes I_K (W_\tau + \bar{\beta}_{\bar{\omega},\tau} \tilde{U}_{t+1}) \\
&= \frac{1}{N}(\mathbf{1}^\top_{N\bar{d}} \otimes I_K)(W_\tau + \bar{\beta}_{\bar{\omega},\tau} \tilde{U}_{\tau+1}) \\
&= \langle W_\tau \rangle_2 + \bar{\beta}_{\bar{\omega},\tau} \langle \tilde{U}_{\tau+1} \rangle_2 \\
&= \langle W_\tau \rangle_2 + \bar{\beta}_{\bar{\omega},\tau} \langle \tilde{\delta}_\tau \rangle_1 \cdot \phi_\tau.
\end{aligned}$$

Hence, the updates for $\langle W_\tau \rangle_2$ and $\langle \bar{\mu}_\tau \rangle_1$ are

$$\langle \bar{\mu}_{\tau+1} \rangle_1 = \langle \bar{\mu}_\tau \rangle_1 + \bar{\beta}_{\bar{\omega},\tau} \cdot \mathbb{E}(\bar{r}_{\tau+1} - \langle \bar{\mu}_\tau \rangle_1 | \mathcal{F}_{\tau,1}) + \bar{\beta}_{\bar{\omega},\tau} \cdot \xi_{\tau+1,1}, \tag{24}$$

$$\langle W_{\tau+1} \rangle_2 = \langle W_\tau \rangle_2 + \bar{\beta}_{\bar{\omega},\tau} \cdot \mathbb{E}(\hat{\delta}_{\tau+1}\phi_\tau | \mathcal{F}_{\tau,1}) + \bar{\beta}_{\bar{\omega},\tau} \cdot \xi_{\tau+1,2} + \bar{\beta}_{\bar{\omega},\tau} \cdot \gamma_{\tau+1}, \tag{25}$$

where $\hat{\delta}_{\tau+1} = \langle r_{\tau+1} - \bar{\mu}_\tau \rangle_1 + (\Phi_{\tau+1} - \Phi_\tau)\langle W_\tau \rangle_2$, $\xi_{\tau+1,1} = r_{\tau+1} - \mathbb{E}(r_{\tau+1} - \langle \bar{\mu}_\tau \rangle | \mathcal{F}_{\tau,1})$, $\xi_{\tau+1,2} = \hat{\delta}_{\tau+1}\phi_\tau - \mathbb{E}(\hat{\delta}_{\tau+1}\phi_\tau | \mathcal{F}_{\tau,1})$, and $\gamma_{\tau+1} = \langle \tilde{\delta}_{\tau+1} \rangle \phi_\tau - \hat{\delta}_{\tau+1}\phi_\tau$.

Note that $\mathbb{E}(r_{\tau+1} - \langle \bar{\mu}_\tau \rangle_1 | \mathcal{F}_{\tau,1})$ is Lipschitz continuous in $\langle \bar{\mu}_\tau \rangle_1$, and that $\mathbb{E}(\hat{\delta}_{\tau+1}\phi_\tau | \mathcal{F}_{\tau,1})$ is Lipschitz continuous in both $\langle W_\tau \rangle_2$ and $\langle \bar{\mu}_\tau \rangle_1$. Moreover, $\xi_{\tau+1,1}$ and $\xi_{\tau+1,2}$ are martingale differences sequences. From Lemmas 3 and 7, $\{\gamma_\tau\}$ is a bounded random sequence with $\gamma_\tau \to 0$ as $\tau \to \infty$ almost surely.

From Theorem B.2 in [4], the following ODE captures the asymptotic behavior of (24) and (25):

$$\begin{bmatrix} \langle \dot{\bar{\mu}} \rangle_1 \\ \langle \dot{W} \rangle_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -\Phi^\top D^{s,a}_\theta \mathbf{1}_{NK} & \Phi^\top D^{s,a}_\theta (P^\theta - I_{NK})\Phi \end{bmatrix} \begin{bmatrix} \langle \bar{\mu} \rangle_1 \\ \langle W \rangle_2 \end{bmatrix} + \begin{bmatrix} J(\theta) \\ \Phi^\top D^{s,a}_\theta \bar{R} \end{bmatrix} \tag{26}$$

From the proof of Theorem 4.6 in [4], the ODE (26) is globally asymptotically stable and has its equilibrium satisfying

$$\begin{cases} \langle \bar{\mu} \rangle_1 = J(\theta), \\ \Phi^\top D^{s,a}_\theta [\bar{R} - \langle \bar{\mu} \rangle_1 \mathbf{1}_{NK} + P^\theta \Phi \langle W \rangle_2 - \Phi \langle W \rangle_2] = 0. \end{cases}$$

Note that the solution for $\langle \bar{\mu} \rangle_1$ at equilibrium is $J(\theta)$, and the solution for $\langle W \rangle_2$ has the form $\omega_\theta + lv$ with any $l \in \mathbb{R}$ and $v \in \mathbb{R}^K$ such that $\phi v = \mathbf{1}_K$, where ω_θ follows that $\Phi^\top \mathrm{D}^{s,a}_\theta \big[T^Q_\theta(\Phi\omega_\theta) - \Phi\omega_\theta\big] = 0$. Moreover, $\phi v \neq \mathbf{1}_K$ by Assumption 4, so ω_θ is the unique solution, which implies that $\lim_\tau \langle W_\tau \rangle_2 = \omega_\theta$. Combining the above facts with Lemma 5, we conclude that $\lim_\tau \bar{z}^i_\tau = \omega_\theta$. ■

References

[1] R.K. Kolla, K. Jagannathan, and A. Gopalan. Collaborative learning of stochastic bandits over a social network. *IEEE Transactions on Networking*, 26(4):1782–1795, 2018.

[2] G. Mateos, J.A. Bazerque, and G.B. Giannakis. Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 52(10):5262–5276, 2010.

[3] Z. Jiang, A. Balu, C. Hegde, and S. Sarkar. Collaborative deep learning in fixed topology networks. In *Advances in Neural Information Processing Systems*, pages 5904–5914, 2017.

[4] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pages 5872–5881, 2018.

[5] C. Boutilier. Planning, learning and coordination in multi-agent decision processes. In *Conference on Theoretical Aspects of Rationality and Knowledge*, pages 195–210, 1996.

[6] M. Lauer and M. Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *International Conference on Machine Learning*, pages 535–542, 2000.

[7] M.L. Littman. Value-function reinforcement learning in Markov games. *Cognitive Systems Research*, 2(1):55–66, 2001.

[8] X. Wang and T. Sandholm. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. In *Advances in Neural Information Processing Systems*, pages 1603–1610, 2003.

[9] S. Kar, J.M. Moura, and H.V. Poor. QD-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013.

[10] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar. Finite-sample analyses for fully decentralized multi-agent reinforcement learning. *arXiv preprint arXiv:1812.02783*, 2018.

[11] D. Lee, H. Yoon, V. Cichella, and N. Hovakimyan. Stochastic primal-dual algorithm for distributed gradient temporal difference learning. *arXiv preprint arXiv:1805.07918*, 2018.

[12] T.T. Doan, S.T. Maguluri, and J. Romberg. Convergence rates of distributed TD (0) with linear function approximation for multi-agent reinforcement learning. *arXiv preprint arXiv:1902.07393*, 2019.

[13] J. Hu and M.P. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4(11):1039–1069, 2003.

[14] J. Foerster, Y.M. Assael, N. Freitas, and S. Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2137–2145, 2016.

[15] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017.

[16] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690, 2017.

[17] K. Zhang, Z. Yang, and T. Başar. Networked multi-agent reinforcement learning in continuous spaces. In *Proceedings of IEEE Conference on Decision and Control*, pages 2771–2776. IEEE, 2018.

[18] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[19] Mohammad Babaeizadeh, Iuri Frosio, Stephen Tyree, Jason Clemons, and Jan Kautz. Reinforcement learning through asynchronous advantage actor-critic on a gpu. *arXiv preprint arXiv:1611.06256*, 2016.

[20] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deeprl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

[21] Ali Yahya, Adrian Li, Mrinal Kalakrishnan, Yevgen Chebotar, and Sergey Levine. Collective robot reinforcement learning with distributed asynchronous guided policy search. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 79–86. IEEE, 2017.

[22] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[23] N. M. Freris, S. R. Graham, and P. R. Kumar. Fundamental limits on synchronizing clocks over networks. *IEEE Transactions on Automatic Control*, 56(6):1352–1364, 2011.

[24] D. Kempe, A. Dobra, and J. Gehrke. Gossip-based computation of aggregate information. In *Proceedings of the 44th IEEE Symposium on Foundations of Computer Science*, pages 482–491, 2003.

[25] C.N. Hadjicostis and T. Charalambous. Average consensus in the presence of delays in directed graph topologies. *IEEE Transactions on Automatic Control*, 59(3):763–768, 2014.

[26] V.R. Konda and J.N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

[27] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

[28] Yew-Kwang Ng. Bentham or bergson? finite sensibility, utility functions and social welfare functions. *The Review of Economic Studies*, 42(4):545–569, 1975.

[29] D.P. Bertsekas and J.N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall, 1989.

[30] A. Nedić and A. Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015.

[31] S. Li and T. Başar. Asymptotic agreement and convergence of asynchronous stochastic algorithms. *IEEE Transactions on Automatic Control*, 32(7):612–618, 1987.

[32] Christoforos N. Hadjicostis and Themistoklis Charalambous. Average consensus in the presence of delays in directed graph topologies. *IEEE Transactions on Automatic Control, vol.59, no.3*, 2014.