This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2024.3409676



# Resilient Multi-agent Reinforcement Learning with Function Approximation

Lintao Ye, Martin Figura, Yixuan Lin, Mainak Pal, Pranoy Das, Ji Liu, and Vijay Gupta

***Abstract*—Adversarial attacks during training can strongly influence the performance of multi-agent reinforcement learning algorithms. It is, thus, highly desirable to augment existing algorithms such that the impact of adversarial attacks on cooperative networks is at least bounded. We consider a fully decentralized network, where each agent receives a local reward and observes the global state and action. We propose a resilient consensus-based actor-critic algorithm, whereby each agent estimates the team-average reward and value function, and communicates the associated parameter vectors to its immediate neighbors. We show that in the presence of Byzantine agents, whose estimation and communication strategies are completely arbitrary, the estimates of the cooperative agents converge to a bounded consensus value with probability one, provided that there are at most $H$ Byzantine agents in the network that is $(2H+1)$-robust. Furthermore, we prove that the policy of the cooperative agents converges with probability one to a bounded neighborhood around a stationary point of their team-average objective function under the assumption that the policies of the adversarial agents asymptotically become stationary.**

***Index Terms*—Cooperative multi-agent reinforcement learning, Byzantine-resilient learning, adversarial attacks, consensus**

## I. Introduction

In multi-agent reinforcement learning (MARL), agents interact with each other and a common environment to learn policies that maximize their objective functions. Cooperative MARL, in which agents wish to maximize a team objective function, has emerged as an exciting method to solve dynamic programming approximately for teams of agents with aligned objectives in numerous potential applications [1]–[5].

In cooperative MARL, there is a long line of literature that assumes that the agents first participate in centralized training and then execute their decentralized policy at test time [6]–[8]. These methods assume that the agents share their local rewards, which they may wish to keep private in certain applications. Recently, this assumption was removed by establishing methods in which the agents receive only local rewards and communicate local information about the team performance to their neighbors according to a graph [9]–[13].

In this paper, we focus on decentralized learning using consensus-based actor-critic (AC) MARL methods [9], [14] that have been shown to scale well with the size of the multi-agent Markov decision processes (MMDP) and to ensure sufficient exploration through the implementation of stochastic policies. These algorithms are based on parameter sharing, i.e., the agents locally update parameters of the team-average value function surrogate using their local reward signal and communicate the updated parameters to their immediate neighbors. This idea, which originated in decentralized Q-learning [15], was adopted in two consensus-based AC MARL algorithms with linear function approximation proposed in [9] and the consensus-based AC MARL algorithm with nonlinear function approximation proposed in [14].

An important aspect of multi-agent systems is its resilience, i.e., the ability to preserve performance when a subset of agents are compromised by potential adversarial attacks [16]. While there is much to be admired about the state-of-the-art consensus-based AC MARL algorithms, there are some question marks about their resilience. In [17], it was shown that the consensus-based AC MARL algorithm in [9] is susceptible to a simple adversarial attack. Specifically, a single self-interested agent can mislead all the other (cooperative) agents to learn policies that maximize the objective function of the self-interested agent, even though these policies may be arbitrarily poor for the team objective. This has motivated work on designing a consensus-based AC MARL algorithm that is provably resilient to adversarial attacks. Formally, the cooperative agents aim to learn optimal policies to a team-average objective function among them by communicating with their neighbors in an environment influenced by the adversarial (i.e., Byzantine) agents. Throughout this paper, we assume that a Byzantine agent can communicate arbitrary and distinct information to each neighboring agent in the environment and enact an arbitrary policy. Importantly, the Byzantine agents impact the other agents both due to the information they communicate to them as well as through implementation of adversarial policies that affect the evolution of the state of the environment; further, we do not assume that the agents are aware of the policies of one another.

Unfortunately, resilient multi-agent learning algorithms that *eliminate* the effects of Byzantine attacks entirely quickly run into the curse of dimensionality. For instance, even in the simple context of agents trying to learn the mean of a static vector value that they hold when some agents are Byzantine –

L. Ye is with the Key Laboratory of Image Processing and Intelligent Control, Ministry of Education and the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. yelintao93@hust.edu.cn. M. Figura is with Fortna, Denver, CO. figuramartin1@gmail.com. Y. Lin is currently with Meta and was previously affiliated with the Department of Applied Mathematics and Statistics at Stony Brook University, NY. yixuan.lin.1@stonybrook.edu. J. Liu is with the Department of Electrical and Computer Engineering at Stony Brook University, NY. ji.liu@stonybrook.edu. M. Pal, P. Das, V. Gupta are with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN {pal42,das211,gupta869}@purdue.edu. The work was supported in part by the National Science Foundation under grants 2230101 and 2300355, the National Natural Science Foundation of China under grant 62203179, ARO under grants W911NF2310111, W911NF-23-1-0316 and W911NF2310266, AFOSR under grant FA9550-21-1-0231, and ONR under grants 13001364 and F.10052139.02.012.

the so-called *Byzantine consensus vector* problem, the number of reliable machines must be proportional to the product of the number of Byzantine machines and the dimension of the shared vectors [18]. One field that has studied the presence of Byzantine adversarial agents in the context of learning is distributed machine learning with master-worker networks. In this setting, each worker agent applies stochastic gradient descent (SGD) to update shared function parameters and a central machine aggregates parameters received from multiple worker agents. If some agents can be adversarial, we can point to algorithms based on the concept of the geometric median [19], [20]. To reduce the computational complexity, the method of median of means was proposed in [21] and the method of entry-wise median was introduced in [22]. Another recently proposed approach, the Krum method, is based on medoids [23]. It is important to note that all these methods merely bound the deviation of the aggregated parameter vector from the desired one. A resilient MARL algorithm based on centralized training and decentralized execution was proposed in [24], where a team of protagonist agents compete against a team of antagonist agents. Their aim is for the protagonist agents to perform well even if some of the agents fail or behave erroneously during the evaluation phase.

In consensus-based MARL, a popular method to attenuate the effect of adversarial data injected by Byzantine agents is to apply the element-wise trimmed-mean [25]–[27], where the agents discard $H$ largest and $H$ smallest parameter values received from the neighbors. The hyperparameter $H$ denotes the maximum number of Byzantine agents allowed in the network. [25] proposed this method in decentralized $Q$-learning and shows that the policies of the cooperative agents converge to a neighborhood around the team-optimal policy. Matters become significantly more complicated in large MMDPs, where agents employ function approximation and execute consensus updates in the parameter space. A resilient consensus-based AC MARL algorithm with linear function approximation was proposed in [26]; however, this method involves a centralized coordinator that receives parameter vectors from all agents and provides them with an element-wise trimmed mean of each parameter. The approach was later extended to fully decentralized MARL in [27], where the element-wise trimmed mean is computed on a local level. While the element-wise truncation methods guarantee boundedness of the estimated parameters, the Byzantine agents can still design attacks that manipulate individual parameters within bounded intervals. If properly designed, these attacks lead to the overestimation of selected features, which may compound large errors in the approximated functions.

In this work, we introduce a novel resilient projection-based consensus algorithm for decentralized AC MARL, where parametric models must be used to approximate the AC networks. The algorithm (Algorithm 2) includes two important steps. In the first step, the received parameters are projected into the feature vectors that happen to be the same for all agents if the agents utilize linear function approximation. The projection maps the received parameter vectors into scalar values that approximately yield the estimation errors applied by the neighbors in the SGD updates. In the second step, the cooperative agents perform resilient aggregation of the estimation errors and apply the aggregated value in another SGD update, which ensures diffusion of local data across the network. The closest work to ours is [27], which proposes a similar rule for resilient aggregation; however, as opposed to aggregating parameter vectors that are of the same dimension as the feature vectors as in [27], our projection-based method is performed over scalars. We show numerically that this leads to better estimates of the team advantage functions and better performance of the MARL algorithm. With linear function approximation, we prove that the joint policy of the cooperative agents converges to the neighborhood of a policy that forms a stationary point of the team utility function under reasonable assumptions on the policies of the Byzantine agents. We also show via simulations that Algorithm 2 can be applied even with nonlinear approximation in a cooperative navigation task. We show through our simulations that low-dimensional aggregation makes it significantly more difficult for Byzantine agents to stage a successful attack. As a side contribution regarding non-resilient consensus AC algorithms, we introduce and analyze the convergence of the projection-based consensus AC algorithm (Algorithm 1) which is a special case of Algorithm 2 with no trimming applied in the consensus updates. Finally, while our analyses and those in [9] focus on the asymptotic convergence of the proposed decentralized AC MARL algorithms, the more recent work [28] analyzes the finite-time performance of a decentralized AC MARL algorithm (with partial policy sharing among the agents and without adversarial agents). We leave adapting the finite-time analyses in [28] to the resilient decentralized AC MARL algorithms proposed in this paper for future work.

The paper is organized as follows. In Section II, we present some background and formulate the problem. The two consensus-based AC MARL algorithms are presented in Section III. We provide a convergence analysis for Algorithm 1 and Algorithm 2 with linear approximation in Section IV and demonstrate the efficacy of Algorithm 2 with nonlinear approximation in Section V. For ease of reading, all the proofs are collected in the appendix.

## II. Background and Problem Formulation

**Notations:** Let $\mathbb{R}$ denote the set of real numbers. The spectral radius of a matrix is denoted as $\rho(\cdot)$. For a vector $y$, let $\|y\|$, $\|y\|_1$ and $\mathbf{dim}(y)$ be its $l_2$ norm, $l_1$ norm and dimension, respectively. Let $I$ be the identity matrix and $\mathbf{1}$ be an all-one vector with proper dimensions that can be inferred from the context. We write $P \succeq 0$ if $P$ is a positive semidefinite matrix; and $P \preceq Q$ if $Q - P$ is positive semidefinite. For a set $\mathcal{S}$, let $|\mathcal{S}|$ be the cardinality.

**Multi-agent Markov Decision Process:** Consider an MMDP given as a tuple $(\mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \mathcal{P}, \{\mathcal{R}^i\}_{i\in\mathcal{N}}, \mathcal{G})$, where $\mathcal{N} = \{1, \dots, N\}$ is the set of all agents, $\mathcal{S}$ is a set of (global) states, $\mathcal{P}$ is a set of transitional probabilities, $\gamma \in [0, 1)$ is a discount factor, $\mathcal{G}$ represents a set of communication graphs, and $\mathcal{A}^i$ and $\mathcal{R}^i$ are a set of actions and rewards of agent $i$, respectively. The communication graph active at time $t$ is denoted by $\mathcal{G}_t$. With a small abuse of notation, we let

$\mathcal{G}_t = (\mathcal{N}, \mathcal{E}_t)$ so that the set of vertices is also denoted by $\mathcal{N}$, with each vertex $i$ being associated with agent $i$, and a set of undirected edges $\mathcal{E}_t \subseteq \mathcal{N} \times \mathcal{N}$. Furthermore, we define sets $\mathcal{N}^i_{in,t}$ and $\mathcal{N}^i_{out,t}$ that include all agents that transmit data to and receive data from agent $i$ at time $t$, respectively. The global state is denoted by $s \in \mathcal{S}$. The global action is obtained by stacking the actions of all the agents in a vector denoted by $a$. We will use $s'$ to denote the global state at the future step. All variables with the superscript $i$ pertain to agent $i$. We let $r^i(s,a) : \mathcal{S} \times \mathcal{A} \to \mathcal{R}^i \subset \mathbb{R}$ denote the local reward of subsystem $i$, $p(s'|s,a) : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathcal{P} \subset \mathbb{R}$ the joint transitional probability, and $\pi^i(a^i|s) : \mathcal{S} \times \mathcal{A}^i \to (0,1)$ the policy of subsystem $i$. The global policy is given as $\pi(a|s) = \prod_{i\in\mathcal{N}} \pi^i(a^i|s)$. If needed, we emphasize the dependence of a signal on time by using subscript $t$, i.e., $r^i_{t+1}(s_t, a_t)$. If the dependence is clear from the context, we drop the subscript to reduce notational clutter. The rewards remain private and each agent generally receives a different reward, i.e., $r^i \neq r^j$ for $i, j \in \mathcal{N}, i \neq j$. We assume that every agent observes the global state $s$ and action $a$ at each step in training. We define the average individual reward under global policy $\pi(a|s)$ as $r^i_\pi(s) = \sum_a \pi(a|s) r^i(s,a)$, the average individual reward under global policy $\pi(a|s)$ at all states $s \in \mathcal{S}$ as $R^i_\pi = [r^i_\pi(s), s \in \mathcal{S}]^T \in \mathbb{R}^{|\mathcal{S}|}$, and the average individual reward at all state-action pairs $(s,a)$ as $R^i = [r^i(s,a), s \in \mathcal{S}, a \in \mathcal{A}]^T \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|}$. The distributions of states and state-action pairs visited by the agents under a fixed policy $\pi(a|s)$ are denoted as $d_\pi(s)$ and $d'_\pi(s,a)$, respectively.

**Objective Functions:** We divide the agents into a set of cooperative agents and a set of Byzantine agents, which we denote by $\mathcal{N}^+$ and $\mathcal{N}^-$, respectively, with $\mathcal{N}^+ \cup \mathcal{N}^- = \mathcal{N}$ and $\mathcal{N}^+ \cap \mathcal{N}^- = \emptyset$. A Byzantine agent is one that communicates arbitrary and generally distinct information to each of its neighbors in the set $\mathcal{N}^i_{out,t}$ and enacts an arbitrary policy $\pi^i(a^i|s)$. We note that the membership or cardinality of the sets $\mathcal{N}^+$ and $\mathcal{N}^-$ is not known. In other words, we do not know a priori whether an agent is cooperative or Byzantine. We let $\pi^+(a^+|s) = \prod_{i\in\mathcal{N}^+} \pi^i(a^i|s)$ denote the aggregated policy of the cooperative agents, where $a^+$ is the aggregated action of the cooperative agents. Similarly, we define the aggregated policy of the Byzantine agents as $\pi^-(a^-|s) = \prod_{i\in\mathcal{N}^-} \pi^i(a^i|s)$, where $a^-$ represents the aggregated action of the Byzantine agents. Each cooperative agent $i$, $i \in \mathcal{N}^+$, is associated with an objective function:

$$J^i(\pi) = J^i(\pi^+, \pi^-) = \mathbb{E}_{\pi,d_\pi}[\sum_{t=0}^{\infty} \gamma^t r^i_{t+1}(s_t, a_t)],$$

for a discount factor $\gamma \in (0,1)$. The cooperative agents solve the following well-defined optimization problem:

$$\pi^+_* = \arg\max_{\pi^+} J^+(\pi^+, \pi^-). \tag{1}$$

where

$$J^+(\pi^+, \pi^-) \triangleq \mathbb{E}_{\pi,d_\pi}\left[\frac{1}{N^+} \sum_{i\in\mathcal{N}^+} \sum_{t=0}^{\infty} \gamma^t r^i_{t+1}(s_t, a_t)\right],$$

and $N^+ = |\mathcal{N}^+|$. It is important to note that the cooperative agents search for a policy that is optimal when the MMDP evolution is affected by the Byzantine agents rather than a policy that is optimal when the Byzantine agents are absent. The Byzantine agents seek to maximize an arbitrary and potentially unknown objective function, denoted as $J^-(\pi^-, \pi^+)$, that may not be aligned with $J^+(\pi^+, \pi^-)$ in general. The policy $\pi^-$ is unknown to the cooperative agents. We also assume that the Byzantine agents cannot be identified in the training process.

**Multi-agent Policy Gradient:** Since AC algorithms are gradient-based optimization methods, we first establish a general framework to evaluate the gradient of the objective function, $\nabla_{\pi^+} J^+(\pi^+, \pi^-)$. We recall the well-known policy gradient theorem [29], according to which the gradient of the objective function $J^+(\pi^+, \pi^-)$ is given as follows

$$\nabla_\pi J^+(\pi^+, \pi^-) = \mathbb{E}_{\pi,d_\pi}[Q_{\pi^+}(s,a) \nabla_\pi \log \pi(a|s)]$$
$$Q_{\pi^+}(s,a) = \mathbb{E}_\pi[\frac{1}{N^+} \sum_{i\in\mathcal{N}^+} \sum_{t=0}^{\infty} \gamma^t r^i_{t+1}(s_t, a_t) | s_0 = s, a_0 = a].$$

Thus, the policy gradient can be expressed as a sum of gradients with respect to the local policies, i.e., $\nabla_{\pi^+} J^+(\pi^+, \pi^-) = \sum_{i\in\mathcal{N}^+} \nabla_{\pi^i} J^+(\pi^+, \pi^-)$, where

$$\nabla_{\pi^i} J^+(\pi^+, \pi^-) = \mathbb{E}_{\pi,d_\pi}[Q_{\pi^+}(s,a) \nabla_{\pi^i} \log \pi^i(a^i|s)].$$

To reduce the variance of the gradient, define a critic $V_{\pi^+}(s) = \mathbb{E}_\pi[Q_{\pi^+}(s,a)]$, so that the baseline policy gradient $\nabla_{\pi^i} J^+(\pi^+, \pi^-)$ equals

$$\mathbb{E}_{\pi,d_\pi}[(Q_{\pi^+}(s,a) - V_{\pi^+}(s)) \nabla_{\pi^i} \log \pi^i(a^i|s)].$$

We consider the simple temporal difference TD(0) method, where the advantage function $Q_{\pi^+}(s,a) - V_{\pi^+}(s)$ is sampled as the team-average TD error

$$\frac{1}{N^+} \sum_{i\in\mathcal{N}^+} r^i(s,a) + \gamma V_{\pi^+}(s') - V_{\pi^+}(s). \tag{2}$$

**Consensus-based AC MARL:** The distributed AC policy gradient consists of two components: the team-average advantage function $Q_{\pi^+}(s,a) - V_{\pi^+}(s)$ and the term $\nabla_{\pi^i} \log \pi^i(a^i_t|s_t)$. Whereas the latter can be evaluated locally by each agent, the former cannot be sampled directly in decentralized networks because the agents neither observe the team-average rewards $\frac{1}{N^+} \sum_{i\in\mathcal{N}^+} r^i(s,a)$ nor have access to the centralized critic $V_{\pi^+}(s)$. However, as shown by [9], there exists a solution method that uses function approximation of the critic and team-average reward function and communication between agents, which enables the agents to approximately sample the team-average advantage function $Q_{\pi^+}(s,a) - V_{\pi^+}(s)$. Let $\bar{r}(s,a;\lambda^i)$ and $V(s;v^i)$ denote the approximation of the team-average reward function and the critic at agent $i$, where $\lambda^i$ and $v^i$ are the associated parameters. The goal of the cooperative network in the policy evaluation of the consensus-

based AC MARL algorithm is to solve constrained distributed optimization problems:

$$v_\pi = \arg\min_{v^i} \mathbb{E}_{d_\pi}\Bigg\{\mathbb{E}_{\pi,p}\Bigg(\frac{1}{N^+}\sum_{i\in\mathcal{N}^+} r^i(s,a) + \gamma V(s';v^i) - V(s;v^i)\Bigg)^2\Bigg\} \quad \text{s.t.} \quad v^i = v^j, \tag{3}$$

$$\lambda_\pi = \arg\min_{\lambda^i} \mathbb{E}_{d_\pi}\Bigg\{\mathbb{E}_{\pi}\Bigg(\frac{1}{N^+}\sum_{i\in\mathcal{N}^+} r^i(s,a) - \bar{r}(s,a;\lambda^i)\Bigg)^2\Bigg\} \quad \text{s.t.} \quad \lambda^i = \lambda^j, \tag{4}$$

where the equality constraint applies to $i, j \in \mathcal{N}^+$. The distributed optimization problems in (3) and (4) are mean squared error estimation problems with consensus constraints. [9], proposed complementing the local updates of the critic parameters $v^i$ and team-average reward function parameters $\lambda^i$ with consensus updates to ensure that the local estimates are averaged over the network, leading to the notion of team-average estimation. Denoting the local estimation errors by $\delta^i_{v,t} = r^i_{t+1}(s_t,a_t) + \gamma V(s_{t+1};v^i_t) - V(s_t;v^i_t)$ and $\delta^i_{\lambda,t} = r^i_{t+1}(s_t,a_t) - \bar{r}_{t+1}(s_t,a_t;\lambda^i_t)$, the updates of the critic and team-average reward function are given as:

$$\tilde{v}^i_t = v^i_t + \alpha_{v,t}\cdot\delta^i_{v,t}\cdot\nabla_v V(s_t;v^i_t) \tag{5}$$

$$v^i_{t+1} = \sum_{j\in\mathcal{N}} c_{v,t}(i,j)\tilde{v}^j_t \tag{6}$$

$$\tilde{\lambda}^i_t = \lambda^i_t + \alpha_{\lambda,t}\cdot\delta^i_{\lambda,t}\cdot\nabla_\lambda \bar{r}_{t+1}(s_t,a_t;\lambda^i_t) \tag{7}$$

$$\lambda^i_{t+1} = \sum_{j\in\mathcal{N}} c_{\lambda,t}(i,j)\tilde{\lambda}^j_t, \tag{8}$$

where $c_{v,t}(i,j)$ and $c_{\lambda,t}(i,j)$ are the consensus weights applied by agent $i$ to values received from agent $j$, and $\alpha_{v,t}$ and $\alpha_{\lambda,t}$ denote the step sizes in the SGD updates. The parameters $v^i$ and $\lambda^i$ are updated locally based on the most recent local reward $r^i(s,a)$ and observation of the global state $s$ and action $a$ before being transmitted to neighbors for aggregation. Combining the local updates with consensus updates ensures the MMSE estimation defined in (3) and (4).

**Resilient Consensus-based AC MARL:** The success of the consensus-based method for distributed estimation in (6) and (8) hinges on the assumption that all agents are reliable, i.e., $\mathcal{N}^- = \emptyset$. Even if a single agent deviates from the proposed updates, the distributed stochastic approximation can yield arbitrarily poor results, e.g., a single adversary can drive the cooperative network to maximize only the objective function $J^-(\pi^-,\pi^+)$ [17]. To provide resilience against adversarial attacks in consensus-based MARL, the method of trimmed means was proposed in [27]. In this method, assuming that there are at most $H$ Byzantine agents in the network, the cooperative agents rank the values received from their neighbors for each parameter and trim the $H$ largest and $H$ smallest values in the consensus update. Our intention is to design a resilient consensus method that is suitable for function approximation in MARL algorithms.

## III. Proposed Algorithms

In this section, we design novel resilient consensus-based AC MARL algorithm (Algorithm 2). In Section III-A, we introduce the projection-based consensus AC MARL algorithm with linear approximation (Algorithm 1) that allows agents to conservatively perform consensus updates. In Section III-B, we present the resilient projection-based consensus AC MARL algorithm with linear approximation (Algorithm 2) that further includes trimming that provides resilience in the team-average estimation. The presented algorithms are based on parameter-sharing, and thus the communication complexity scales with the number of parameters used in the function approximation. In Section V, we discuss the implementation of Algorithm 2 with nonlinear approximation.

### A. Projection-based Algorithm with Linear Approximation

We begin by considering linear function approximation to allow rigorous convergence analysis. Thus, we let $\bar{r}(s,a;\lambda^i) = f(s,a)^T\lambda^i$ and $V(s;v^i) = \phi(s)^T v^i$, where $f(s,a)$ and $\phi(s)$ denote the basis functions. Throughout the paper, we use shorthand $f_t$ and $\phi_t$ to denote the feature vectors $f(s_t,a_t)$ and $\phi(s_t)$ evaluated at time $t$, respectively. The assumption of linear function approximation allows us to rewrite the updates from (6) and (8) as follows:

$$\tilde{v}^i_t = v^i_t + \alpha_{v,t}(r^i_{t+1} + \gamma\phi^T_{t+1}v^i_t - \phi^T_t v^i_t)\phi_t \tag{9}$$

$$\tilde{\lambda}^i_t = \lambda^i_t + \alpha_{\lambda,t}(r^i_{t+1} - f^T_t\lambda^i_t)f_t. \tag{10}$$

It is easy to see that a single update is performed in the subspace spanned by the feature vectors and its magnitude and direction in this subspace are governed by the step size and estimation error. The agents can exploit the knowledge of the common feature vectors $\phi_t$ and $f_t$ to estimate the estimation error of their neighbors using scalar projection as

$$r^j_{t+1} + \gamma\phi^T_{t+1}v^j_t - \phi^T_t v^j_t \approx \frac{\phi^T_t(\tilde{v}^j_t - v^i_t)}{\alpha_{v,t}\|\phi_t\|^2}, \tag{11}$$

$$r^j_{t+1} - f^T_t\lambda^j_t \approx \frac{f^T_t(\tilde{\lambda}^j_t - \lambda^i_t)}{\alpha_{\lambda,t}\|f_t\|^2}. \tag{12}$$

Lemma 1 shows that the approximation in (12) becomes exact once the agents reach consensus on the parameter values.

**Lemma 1.** *Suppose that agent $i$ reaches consensus on the critic and team-average reward function parameters with its neighbors, i.e., $x^i_t = x^j_t$ for $x \in \{v,\lambda\}$ and all $j \in \mathcal{N}^i_{in,t}$. Then, the agent can exactly evaluate the estimation errors $\delta^j_{\lambda,t} = r^j_{t+1} - f^T_t\lambda^j_t$ and $\delta^j_{v,t} = r^j_{t+1} + \gamma\phi^T_{t+1}v^j_t - \phi^T_t v^j_t$.*

*Proof.* If $x^i_t = x^j_t$ for $x \in \{v,\lambda\}$, manipulating the neighbor updates in (10) and applying scalar projection into their respective feature vectors $\phi_t$ and $f_t$ yields $r^j_{t+1} + \gamma\phi^T_{t+1}v^j_t - \phi^T_t v^j_t = \frac{\phi^T_t(\tilde{v}^j_t - v^i_t)}{\alpha_{v,t}\|\phi_t\|^2}$ and $r^j_{t+1} - f^T_t\lambda^j_t = \frac{f^T_t(\tilde{\lambda}^j_t - \lambda^i_t)}{\alpha_{\lambda,t}\|f_t\|^2}$. Therefore, agent $i$ evaluates the estimation errors exactly. □

In contrast to the vector aggregation via consensus updates in (6) and (8), the aggregation in the projection-based consensus method is done over the estimated neighbors' estimation errors that take scalar values. The method is incorporated in

the parameter updates of a projection-based consensus AC algorithm with the pseudo-code in Algorithm 1.

**Algorithm 1:** Projection-based consensus AC with linear approximation

**Initialize** $s_0, \{\alpha_{v,t}\}_{t\geq 0}, , \{\alpha_{\lambda,t}\}_{t\geq 0}, \{\alpha_{\theta,t}\}_{t\geq 0}, t \leftarrow 0, \theta_0^i, \lambda_0^i, \tilde{\lambda}_0^i, v_0^i, \tilde{v}_0^i, \forall i \in \mathcal{N}$

**Repeat until convergence**

**for** $i \in \mathcal{N}$ **do**

- **Take** action $a_t^i \sim \pi^i(a_t^i|s_t;\theta_t^i)$;
- **Observe** state $s_{t+1}$, action $a_t$, and reward $r_{t+1}^i$
- **Update actor**
- $\delta_t^i \leftarrow \bar{r}(s_t,a_t;\lambda_t^i) + \gamma V(s_{t+1};v_t^i) - V(s_t;v_t^i)$
- $\theta_{t+1}^i \leftarrow \theta_t^i + \alpha_{\theta,t}\cdot\delta_t^i\cdot\nabla_{\theta^i}\log\pi^i(a_t^i|s_t^i;\theta_t^i)$
- **Update critic and team reward function**
- $\tilde{v}_t^i \leftarrow v_t^i + \alpha_{v,t}\cdot(r_{t+1}^i + \gamma V(s_{t+1};v_t^i) - V(s_t;v_t^i))\cdot \nabla_{v^i}V(s_t;v_t^i)$
- $\tilde{\lambda}_t^i \leftarrow \lambda_t^i + \alpha_{\lambda,t}\cdot(r_{t+1}^i - \bar{r}(s_t,a_t;\lambda_t^i))\cdot\nabla_{\lambda^i}\bar{r}(s_t,a_t;\lambda_t^i)$
- **Send** $\tilde{\lambda}_t^i, \tilde{v}_t^i$ to $j \in \mathcal{N}_{out,t}^i$

**end**

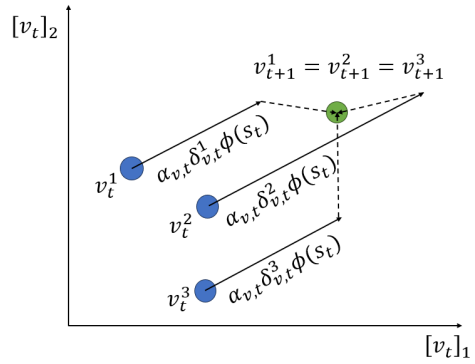
**for** $i \in \mathcal{N}$ **do**

- **Receive** $\tilde{\lambda}_t^j, \tilde{v}_t^j$ from $j \in \mathcal{N}_{in,t}^i$
- **Projection-based consensus step**
- $\epsilon_{v,t}^{ij} \leftarrow \frac{\phi_t^T(\tilde{v}_t^j - v_t^i)}{\alpha_{v,t}\|\phi_t\|^2}$, $\epsilon_{\lambda,t}^{ij} \leftarrow \frac{f_t^T(\tilde{\lambda}_t^j - \lambda_t^i)}{\alpha_{\lambda,t}\|f_t\|^2}$ for $j \in \mathcal{N}_{in,t}^i$
- $\epsilon_{v,t}^i \leftarrow \sum_{j\in\mathcal{N}_{in,t}^i} c_{v,t}(i,j)\cdot\epsilon_{v,t}^{ij}$,
- $\epsilon_{\lambda,t}^i \leftarrow \sum_{j\in\mathcal{N}_{in,t}^i} c_{\lambda,t}(i,j)\cdot\epsilon_{\lambda,t}^{ij}$
- $v_{t+1}^i \leftarrow v_t^i + \alpha_{v,t}\cdot\epsilon_{v,t}^i\cdot\nabla_{v^i}V(s_t;v_t^i)$
- $\lambda_{t+1}^i \leftarrow \lambda_t^i + \alpha_{\lambda,t}\cdot\epsilon_{\lambda,t}^i\cdot\nabla_{\lambda^i}\bar{r}(s_t,a_t;\lambda_t^i)$

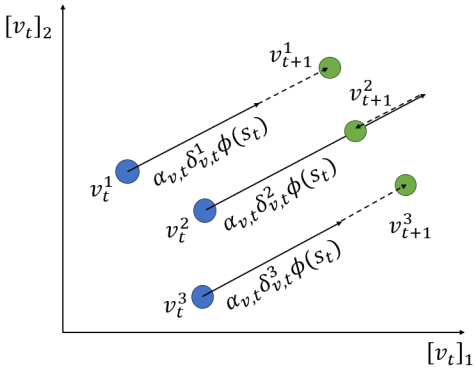
**end**

**Update** iteration counter $t \leftarrow t+1$

The agents perform a stochastic update using their local reward $r_{t+1}^i$ and exchange the updated parameters $\tilde{v}_t^i$ and $\tilde{\lambda}_t^i$ over the communication graph. Then, they estimate the average estimation errors $\epsilon_{v,t}^i$ and $\epsilon_{\lambda,t}^i$ through the projection-based consensus update, and apply the average estimation errors in the parameter updates that yield new values $v_{t+1}^i$ and $\lambda_{t+1}^i$. Figure 1 provides a comparison between the consensus-based AC updates that are given in (6) and (8), and the projection-based consensus AC updates. We note that the projection-based consensus AC algorithm performs more conservative updates than the consensus-based AC MARL algorithm. Consider any time step $t \geq 0$ in Algorithm 1. First, each agent $i \in \mathcal{N}$ needs to compute local updates for $\delta_t^i, \theta_t^i, \epsilon_{v,t}^{ij}, \epsilon_{v,t}^i \in \mathbb{R}$ for all $j \in \mathcal{N}_{in,t}^i$ and $\lambda_t^i, v_t^i$, where the dimension of $\lambda_t^i$ (resp., $v_t^i$) is equal to that of the feature vector $f_t$ (resp., $\phi_t$) used in the function approximation for the reward (resp., value function). The computation complexity of agent $i$ in time step $t$ is then given by $O((\mathbf{dim}(f_t)+\mathbf{dim}(\phi_t))|\mathcal{N}_{in,t}^i|+|\mathcal{A}^i|)$. Next, each agent $i \in \mathcal{N}$ needs to send the updated vectors $\tilde{\lambda}_t^i$ and $\tilde{v}_t^i$ to all the neighbors in $\mathcal{N}_{out,t}^i$. In summary, the time complexity and communication latency of Algorithm 1 are governed by the number of features used in the linear function approximation and the size of the action space $\mathcal{A}^i$.




(a) Consensus




(b) Projection-based consensus

Fig. 1: Critic updates of the consensus-based AC and the projection-based consensus AC algorithm in a 2-D parameter space. Local updates are depicted by solid lines and consensus updates by dashed lines. Updated parameters are in green.

### *B. Resilient Projection-based Consensus Actor-critic with Linear Approximation*

To design a defense mechanism against Byzantine agents, we adopt the basic idea of resilient consensus from W-MSR (Weighted Mean-Subsequence-Reduced) algorithms in which each agent *reduces* scalar values received from its neighbors and, *subsequently*, computes a *weighted mean* of the remaining values. By eliminating the most extreme values at every step, the final agreed value among agents is guaranteed to lie within a convex hull of non-faulty agents if the network is sufficiently robust [16]. We apply the W-MSR consensus method over the estimated neighbors' estimation errors. Assuming that there are no more than $H$ Byzantine agents, each agent forms lists of sorted values $\{\epsilon_{v,t}^{ij}\}_{j\in\mathcal{N}_{in,t}^i}$ and $\{\epsilon_{\lambda,t}^{ij}\}_{j\in\mathcal{N}_{in,t}^i}$, and removes $H$ largest values and $H$ smallest values from each set, except for values that are smaller and larger than the value of the agent, respectively. We note that the resilient projection-based consensus method does not suffer from overestimation of the approximated functions because the Byzantine agents can no longer directly manipulate individual parameters in $v_t^j$ and $\lambda_t^j$. Since the resilient aggregation in our algorithm assumes removal of $2H$ values, there are fundamental limitations on the number of Byzantine agents in the network for which the cooperative network remains resilient captured as follows [16].

**Definition 1.** *($\zeta$-connectivity) A connected graph $\mathcal{G}$ is said to be $\zeta$-connected if it has more than $\zeta$ vertices and remains connected whenever fewer than $\zeta$ vertices are removed.*

**Definition 2.** *($\zeta$-reachable sets and $\zeta$-robustness) Given a graph $\mathcal{G}_t$ and a nonempty subset of nodes $\mathcal{Z} \subset \mathcal{N}$, we say that $\mathcal{Z}$ is an $\zeta$-reachable set if there exists $i \in \mathcal{Z}$ such that $|\mathcal{N}^i_{in,t} \backslash \mathcal{Z}| \geq \zeta$, where $\zeta \in \mathbb{Z}_{\geq 0}$. Further, a graph $\mathcal{G}_t$ on $|\mathcal{N}|$ nodes ($|\mathcal{N}| \geq 2$) is $\zeta$-robust, with $\zeta \in \mathbb{Z}_{\geq 0}$, if for every pair of nonempty, disjoint subsets of $\mathcal{N}$, at least one of the subsets is $\zeta$-reachable.*

**Lemma 2.** *(Network robustness after edge removal [16, Lemma 6]) Given a $\zeta$-robust graph $\mathcal{G}_t$, any directed graph $\mathcal{G}'_t$ produced by removing up to $k$ incoming edges to any node is $(\zeta - k)$-robust.*

**Lemma 3.** *(Connectivity of robust graphs [16, Theorem 6]) Suppose $\mathcal{G}_t$ is a $\zeta$-robust undirected graph, with $0 \leq \zeta \leq |\mathcal{N}|/2$. Then, $\mathcal{G}_t$ is at least $\zeta$-connected.*

From Lemma 2 and Lemma 3, a $(2H+1)$-robust network of agents remains connected despite each agent removing $2H$ edges in the resilient aggregation. Additionally, the trimming approach ensures that $\epsilon^i_{v,t}$ and $\epsilon^i_{\lambda,t}$ are bounded by the minimum and maximum values in the set of cooperative neighbors of agent $i$. The pseudo-code for the resilient projection-based consensus AC algorithm with linear function approximation is given in Algorithm 2. The convergence analysis of the algorithm is provided in Section IV-B Similarly to our discussions for Algorithm 1, the time complexity and communication latency are governed by the number of features used in the function approximation and $|\mathcal{A}^i|$.

**Remark 1.** *In the discussion above, we assume linear approximation of the critic $V(s; v^i)$ and team-average reward function $\bar{r}(s, a; \lambda^i)$ for analytical tractability since linear approximation allows us to identify unique optimal parameters $v^i$ and $\lambda^i$ in the policy evaluation. It is important to note that by evaluating linear combinations of the parameters of nonlinear models such as neural networks, we do not generally obtain linear combinations of their outputs. However, since the output layer of a neural network is linear, we can extend Algorithm 2 in the nonlinear setting as well. Here, the additional challenge is to train the hidden layer parameters that evaluate the basis functions $f(s, a)$ and $\phi(s)$. In Section V, we present numerical simulations where we train the hidden layers by aggregating each hidden parameter value.*

## IV. Convergence Results

In this section, we provide a convergence analysis for Algorithm 1 and Algorithm 2 by showing that the cooperative agents are indeed cooperative in the sense that they maximize the objective function $J^+(\pi^+, \pi^-)$. The proof uses a two timescale argument [9], [30], [31]. First, we analyze the critic and team-average reward function updates on the faster timescale, and then we establish the actor convergence on the slower timescale. Note that Algorithm 1 is significantly different from the AC MARL algorithm proposed in [9], due to the projection in the consensus step. Nonetheless, we will show that Algorithm 1 converges under similar assumptions made in [9]. To analyze the convergence of Algorithm 2, we need to rely on a set of assumptions introduced for resilient learning with Byzantine agents [16]. In the sequel, we first present common assumptions used in both Algorithms 1-2, and then, in Section IV-A and IV-B, we introduce further assumptions that are pertinent to the respective algorithms.

**Algorithm 2:** Resilient projection-based consensus AC with linear approximation

**Initialize** $s_0, \{\alpha_{v,t}\}_{t\geq 0}, \{\alpha_{\lambda,t}\}_{t\geq 0}\{\alpha_{\theta,t}\}_{t\geq 0}, t \leftarrow 0, \theta^i_0, \lambda^i_0, \tilde{\lambda}^i_0, v^i_0, \tilde{v}^i_0, H, \forall i \in \mathcal{N}^+$

**Repeat until convergence**

**for** $i \in \mathcal{N}^+$ **do**

- **Take** action $a^i_t \sim \pi^i(a^i_t|s_t; \theta^i_t)$;
- **Observe** state $s_{t+1}$, action $a_t$, and reward $r^i_{t+1}$
- **Update actor**
- $\delta^i_t \leftarrow \bar{r}(s_t, a_t; \lambda^i_t) + \gamma V(s_{t+1}; v^i_t) - V(s_t; v^i_t)$
- $\theta^i_{t+1} \leftarrow \theta^i_t + \alpha_{\theta,t} \cdot \delta^i_t \cdot \nabla_{\theta^i} \log \pi^i(a^i_t|s^i_t; \theta^i_t)$
- **Update critic and team reward function**
- $\tilde{v}^i_t \leftarrow v^i_t + \alpha_{v,t} \cdot (r^i_{t+1} + \gamma V(s_{t+1}; v^i_t) - V(s_t; v^i_t)) \cdot \nabla_{v^i} V(s_t; v^i_t)$
- $\tilde{\lambda}^i_t \leftarrow \lambda^i_t + \alpha_{\lambda,t} \cdot (r^i_{t+1} - \bar{r}(s_t, a_t; \lambda^i_t)) \cdot \nabla_{\lambda^i} \bar{r}(s_t, a_t; \lambda^i_t)$
- **Send** $\tilde{\lambda}^i_t, \tilde{v}^i_t$ to $j \in \mathcal{N}^i_{out,t}$

**end**

**for** $i \in \mathcal{N}^+$ **do**

- **Receive** $\tilde{\lambda}^j_t, \tilde{v}^j_t$ from $j \in \mathcal{N}^i_{in,t}$
- **Resilient projection-based consensus step**
- $\epsilon^{ij}_{v,t} \leftarrow \frac{\phi^T_t(\tilde{v}^j_t - v^i_t)}{\alpha_{v,t}\|\phi_t\|^2}$ for $j \in \mathcal{N}^i_{in,t}$, $\epsilon^{ij}_{\lambda,t} \leftarrow \frac{f^T_t(\tilde{\lambda}^j_t - \lambda^i_t)}{\alpha_{\lambda,t}\|f_t\|^2}$ for $j \in \mathcal{N}^i_{in,t}$
- $\mathcal{N}^i_{v,t} \leftarrow$ remove $H$ smallest values that are smaller than and $H$ largest values that are larger than $\epsilon^{ii}_{v,t}$ from the set $\{\epsilon^{ij}_{v,t}\}_{j\in\mathcal{N}^i_{in,t}}$, return the remaining indices
- $\mathcal{N}^i_{\lambda,t} \leftarrow$ remove $H$ smallest values that are smaller than and $H$ largest values that are larger than $\epsilon^{ii}_{\lambda,t}$ from the set $\{\epsilon^{ij}_{\lambda,t}\}_{j\in\mathcal{N}^i_{in,t}}$, return the remaining indices
- $\epsilon^i_{v,t} \leftarrow \sum_{j\in\mathcal{N}^i_{v,t}} c_{v,t}(i,j) \cdot \epsilon^{ij}_{v,t}$,
- $\epsilon^i_{\lambda,t} \leftarrow \sum_{j\in\mathcal{N}^i_{\lambda,t}} c_{\lambda,t}(i,j) \cdot \epsilon^{ij}_{\lambda,t}$
- $v^i_{t+1} \leftarrow v^i_t + \alpha_{v,t} \cdot \epsilon^i_{v,t} \cdot \nabla_{v^i} V(s_t; v^i_t)$
- $\lambda^i_{t+1} \leftarrow \lambda^i_t + \alpha_{\lambda,t} \cdot \epsilon^i_{\lambda,t} \cdot \nabla_{\lambda^i} \bar{r}(s_t, a_t; \lambda^i_t)$

**end**

**Update** iteration counter $t \leftarrow t + 1$

**Assumption 1.** *The feature vectors $f(s,a) = [f_1(s,a), \ldots, f_M(s,a)] \in \mathbb{R}^M$ and $\phi(s) = [\phi_1(s), \ldots, \phi_L(s)] \in \mathbb{R}^L$ are uniformly bounded for any $s \in \mathcal{S}$, $a \in \mathcal{A}$. Furthermore, if we define the feature matrix $F \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|\times M}$ with $[f_m(s,a), s \in \mathcal{S}, a \in \mathcal{A}]^T$ as its $m$-th column for any $m \in [M]$, and the feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}|\times L}$ with $[\phi_l(s), s \in \mathcal{S}]^T$ as its $l$-th column for any $l \in [L]$, then both $\Phi$ and $F$ have full column rank.*

**Assumption 2.** *The policy $\pi^i(a^i|s; \theta^i)$ is stochastic, i.e.,*

$\pi^i(a^i|s;\theta^i) > 0$ *for any* $i \in \mathcal{N}^+$, $\theta^i \in \Theta^i$, $s \in \mathcal{S}$, $a^i \in \mathcal{A}^i$, *and continuously differentiable in* $\theta^i$. *The Markov chain* $\{s_t\}_{t\geq 0}$ *is irreducible and aperiodic under any* $\pi(a|s;\theta)$.

**Assumption 3.** *The reward* $r^i(s,a)$ *is uniformly bounded for any* $i \in \mathcal{N}^+$.

**Assumption 4.** *The step sizes* $\alpha_{x,t}$, $x \in \{v,\lambda,\theta\}$, *are positive and satisfy* $\sum_t \alpha_{x,t} = \infty$, $\sum_t \alpha_{x,t}^2 < \infty$, $\alpha_{\theta,t} = o(\alpha_{v,t} + \alpha_{\lambda,t})$, *and* $\lim_{t\to\infty} \alpha_{x,t+1} \cdot \alpha_{x,t}^{-1} = 1$.

**Assumption 5.** *The update of the actor parameters* $\theta^i$, $i \in \mathcal{N}^+$, *includes a projection operator* $\Psi_{\Theta^i} : \mathbb{R}^{m_i} \to \Theta^i \subset \mathbb{R}^{m_i}$ *that ensures that* $\theta^i \in \Theta^i$ *where the set* $\Theta^i$ *is compact and hyperrectangular.*

These assumptions are standard in the RL literature [9]. We make the assumption about uniformly bounded rewards and feature vectors to prevent unbounded gradients in the policy evaluation. On a similar note, we assume that the policy is differentiable and strictly positive to avoid unbounded gradients in the actor updates. The assumption of the full-rank features matrices $\Phi$ and $F$ allows us to characterize a unique asymptotically stable equilibrium in the estimation of the critic $V(s;v^i)$ and the team-average reward function $\bar{r}(s,a;\lambda^i)$. By Assumption 4, we can analyze strong convergence on separate timescales as the step sizes tend to zero and the updates of the actor are slower than the updates of $V(s;v^i)$ and $\bar{r}(s,a;\lambda^i)$. This is a reasonable assumption as we typically perform multiple updates of $V(s;v^i)$ and $\bar{r}(s,a;\lambda^i)$ (policy evaluation) before an actor update (policy improvement) in practice. The inclusion of the projection operator in the actor updates, as presented in Assumption 5, is considered in the convergence analysis of RL algorithms [30] but is typically omitted in the implementation. For simplicity, we consider $\Theta^i$ to be hyperrectangular. In practice, we would select a large interval for each element in the parameter vector $\theta^i$ which would constitute a hyperrectangular set $\Theta^i$.

Before we proceed to prove the convergence under Algorithm 1 and 2, we introduce definitions that are frequently used in both proofs. First, we let $v_t = \begin{bmatrix}(v_t^1)^T & \dots & (v_t^N)^T\end{bmatrix}^T$ and $\lambda_t = \begin{bmatrix}(\lambda_t^1)^T & \dots & (\lambda_t^N)^T\end{bmatrix}^T$. Second, we let $D_\pi^s = diag([d_\pi(s), s \in \mathcal{S}]) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ and $D_\pi^{s,a} = diag([d'_\pi(s,a), s \in \mathcal{S}, a \in \mathcal{A}]) \in \mathbb{R}^{(|\mathcal{S}|\cdot|\mathcal{A}|)\times(|\mathcal{S}|\cdot|\mathcal{A}|)}$ denote matrices with a stationary distribution of states and state-action pairs, respectively. Third, we define $p_\pi(s'|s) = \sum_{a\in\mathcal{A}} p(s'|s,a)\pi(a|s;\theta)$ and let $P_\pi = \left[p_\pi(s'|s), s' \in \mathcal{S}, s \in \mathcal{S}\right] \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ denote the state transition matrix of the Markov chain $\{s_t\}_{t\geq 0}$ induced by policy $\pi(a|s;\theta)$. We also define the consensus matrices $C_{v,t} = [c_{v,t}(i,j)]_{ij}$ and $C_{\lambda,t} = [c_{\lambda,t}(i,j)]_{ij}$.

**Definition 3** (Team-average)**.** *The averaging operator* $\langle\cdot\rangle : \mathbb{R}^{NK} \to \mathbb{R}^K$ *is defined such that* $\langle x\rangle = \frac{1}{N}(\mathbf{1}^T \otimes I)x = \frac{1}{N}\sum_{i\in\mathcal{N}} x^i$, *where* $\otimes$ *denotes the Kronecker product.*

**Definition 4** (Projection into consensus subspace)**.** *The operator* $\mathcal{J} = (\frac{1}{N}\mathbf{1}\mathbf{1}^T) \otimes I$ *is defined such that* $\mathcal{J}x = \mathbf{1} \otimes \langle x\rangle$.

**Definition 5** (Projection into disagreement subspace)**.** *The operator* $\mathcal{J}_\perp = I - \mathcal{J}$ *is defined such that* $x_\perp = \mathcal{J}_\perp x = x - \mathbf{1} \otimes \langle x\rangle$.

**Definition 6** (Projection into gradient subspace)**.** *The projection matrices are given as* $\Gamma_{v,t} = \frac{\phi_t\phi_t^T}{\|\phi_t\|^2}$ *and* $\Gamma_{\lambda,t} = \frac{f_t f_t^T}{\|f_t\|^2}$.

**Definition 7** (Projection into orthogonal subspace)**.** *The orthogonal projection matrices are given as* $\hat{\Gamma}_{v,t} = I - \Gamma_{v,t}$ *and* $\hat{\Gamma}_{\lambda,t} = I - \Gamma_{\lambda,t}$.

### A. Convergence of Algorithm 1

In the analysis of Algorithm 1, we prove convergence to unique asymptotically stable equilibria of the critic and team-average reward parameters under fixed policy evaluation and convergence of the actor to the stationary point of the approximated team-average objective function. We let $A'_{v,t} = \phi_t(\gamma\phi_{t+1} - \phi_t)^T$, $b^i_{v,t} = \phi_t r^i_{t+1}$, $A'_{\lambda,t} = -f_t f_t^T$, and $b^i_{\lambda,t} = f_t r^i_{t+1}$. Furthermore, we define $A_{x,t} = I \otimes A'_{x,t}$ and $b_{x,t} = \begin{bmatrix}(b^1_{x,t})^T & \dots & (b^N_{x,t})^T\end{bmatrix}^T$ for $x \in \{v,\lambda\}$. The critic and team-average reward function updates under Algorithm 1 are compactly written as

$$\begin{aligned} v_{t+1} =& v_t + (C_{v,t} \otimes \Gamma_{v,t})(v_t + \alpha_{v,t}(A_{v,t}v_t + b_{v,t})) \\ & - (I \otimes \Gamma_{v,t})v_t \qquad (13) \\ \lambda_{t+1} =& \lambda_t + (C_{\lambda,t} \otimes \Gamma_{\lambda,t})(\lambda_t + \alpha_{\lambda,t}(A_{\lambda,t}\lambda_t + b_{\lambda,t})) \\ & - (I \otimes \Gamma_{\lambda,t})\lambda_t. \qquad (14) \end{aligned}$$

We make the following assumption about the communication graph $\mathcal{G}_t$ and the consensus matrices $C_{v,t}$ and $C_{\lambda,t}$.

**Assumption 6.** *The sequence of time-varying communication graphs* $\{\mathcal{G}_t\}_{t\geq 0}$ *and associated consensus matrices* $\{C_{x,t}\}_{t\geq 0} \in \mathbb{R}^{N\times N}$, *for* $x \in \{v,\lambda\}$, *satisfy:*

*(1)* $\mathcal{G}_t$ *is independent of all random variables and connected in the mean.*

*(2)* $C_{x,t}$ *respects* $\mathcal{G}_t$. *That is,* $c_{x,t}(i,j) = 0$ *if* $(i,j) \notin \mathcal{G}_t$ *and* $c_{x,t}(i,j) \geq \nu$ *if* $(i,j) \in \mathcal{G}_t$ *for some* $\nu > 0$.

*(3) Given* $\mathcal{G}_t$, $C_{x,t}$ *is conditionally independent of all other random variables.*

*(4)* $C_{x,t}$ *is row stochastic and* $\mathbb{E}[C_{x,t}]$ *is column stochastic. That is,* $C_{x,t}\mathbf{1} = \mathbf{1}$ *and* $\mathbf{1}^T\mathbb{E}[C_{x,t}] = \mathbf{1}^T$.

Assumption 6 states that the communication graph $\mathcal{G}_t$ is connected in the mean and the consensus matrix $C_{x,t}$ has a well-defined mean value that ensures balanced updates in any visited state-action pair. Along with the application of nonzero consensus weights, these assumptions guarantee contraction of the estimated function parameters to a consensus value in our analysis. We note that our assumptions are slightly different than in [9] but they essentially describe the same conditions needed to ensure balanced distributed estimation of the critic and team-average reward function. We are now ready to establish convergence in the policy evaluation.

**Lemma 4.** *Suppose there are no adversarial agents in the network. Under Assumption 1-4 and 6, the critic parameters satisfy* $\sup_t\|v_t\| < \infty$ *with probability one. Furthermore, they asymptotically converge with probability one, i.e.,* $\lim_{t\to\infty} v_t^i = v_\pi$ *for* $i \in \mathcal{N}$. *The limit* $v_\pi$ *is a unique solution to* $\Phi^T D_\pi^s(\frac{1}{N}\sum_{i\in\mathcal{N}} R_\pi^i + \gamma P_\pi \Phi v_\pi - \Phi v_\pi) = 0$.

**Lemma 5.** *Suppose there are no adversarial agents in the network. Under Assumption 1-4 and 6, the team-average reward function parameters satisfy* $\sup_t\|\lambda_t\|< \infty$ *with probability one. Furthermore, they converge with probability one, i.e.,* $\lim_{t\to\infty}\lambda_t^i = \lambda_\pi$ *for* $i \in \mathcal{N}$. *The limit* $\lambda_\pi$ *is a unique solution to* $F^T D_\pi^{s,a}(\frac{1}{N}\sum_{i\in\mathcal{N}} R^i - F\lambda_\pi) = 0$.

We note that the convergence point $\lambda_\pi$ corresponds to the minimum mean squared error (MMSE) estimate of the true team-average reward weighted over the distribution of state-action pairs $(s,a)$, $d'_\pi(s,a)$, and the convergence point $v_\pi$ corresponds to the minimum mean squared projected Bellman error estimate of the true team-average critic weighted over the state distribution $d_\pi(s)$. Lemma 4 and Lemma 5 indicate that the cooperative agents estimate the critic and the team-average reward function in the MMSE sense and hence the agents can approximately evaluate the team-average advantage function established in (2). We now establish the actor convergence by showing that the agents' policies converge to a stationary point of the objective function $J^+(\pi^+)$ defined in (1). Due to the absence of adversarial agents, the policy $\pi^-$ is omitted in the analysis.

**Theorem 6.** *Suppose there are no adversarial agents in the network. Under Assumption 1-6, the policy parameter* $\theta_t^i$, $i \in \mathcal{N}$, *converges with probability one to a point in the set of locally asymptotically stable equilibria of the ODE*

$$\dot{\theta}^i = \Psi_{\Theta^i}[\mathbb{E}_{\pi,d_\pi,p}\{(\bar{r}(s,a;\lambda_\pi) + \gamma V(s';v_\pi) - V(s;v_\pi)) \cdot \nabla_{\theta^i}\log\pi^i(a^i|s;\theta^i)\}],$$

*where the parameters* $\lambda_\pi$ *and* $v_\pi$ *are the globally asymptotically stable equilibria under policy* $\pi(a|s;\theta)$.

### *B. Convergence of Algorithm 2*

Analysis of Algorithm 2 is complicated by the fact that Assumption 6 is no longer valid because the consensus matrices $C_{v,t}$ and $C_{\lambda,t}$ take values based on the estimated estimation errors $\epsilon_{v,t}^{ij}$ and $\epsilon_{\lambda,t}^{ij}$, and thus they are not conditionally independent of other signals in the algorithm updates. We introduce a new set of assumptions about the network robustness and the behavior of the Byzantine agents following, e.g., [16]. Define a new communication graph $\mathcal{G}'_t$ that is generated by removing $2H$ incoming edges at each node in $\mathcal{G}_t$.

**Assumption 7.** *The sequence of time-varying communication graphs* $\{\mathcal{G}_t\}_{t\geq 0}$ *and associated consensus matrices* $\{C_{x,t}\}_{t\geq 0} \in \mathbb{R}^{N\times N}$, *for* $x \in \{v,\lambda\}$, *satisfy:*

*(1)* $\mathcal{G}_t$ *includes up to* $H$ *Byzantine agents.*
*(2)* $\mathcal{G}_t$ *is* $(2H+1)$*-robust.*
*(3)* $C_{x,t}$ *respects* $\mathcal{G}'_t$. *That is,* $c_{x,t}(i,j) = 0$ *if* $(i,j) \notin \mathcal{G}'_t$. *and* $c_{x,t}(i,j) \geq \nu$ *if* $(i,j) \in \mathcal{G}'_t$ *for some* $\nu > 0$.
*(4)* $C_{x,t}$ *is row stochastic. That is,* $C_{x,t}\mathbf{1} = \mathbf{1}$.

A wide range of graphs have been proven to be robust [16], [32]. For example, for an Erdős-Rényi random graph $\mathcal{G}_{n,p}$, if the probability $p$ is above (resp., below) a threshold $\frac{\ln n+(\zeta-1)\ln\ln n}{n}$, then $\mathcal{G}_{n,p}$ is $\zeta$-robust with probability one (resp., zero) as $n\to\infty$ [32]. In fact, the $\zeta$-connectivity of $\mathcal{G}_{n,p}$ is characterized by the same threshold $\frac{\ln n+(\zeta-1)\ln\ln n}{n}$ on $p$ [33]. Thus, robustness is not a stronger assumption than connectivity on the Erdős-Rényi random graphs.

**Assumption 8.** *The policy of every Byzantine agent converges to a stationary policy, i.e.,* $\lim_{t\to\infty}\pi_t^i \to \pi_*^i$ *for* $i \in \mathcal{N}^-$. *Furthermore,* $|\pi_{t+1}^i - \pi_t^i| = O(\alpha_{\theta,t})$.

The assumption that the policy of Byzantine agents converges to a fixed policy ensures that the uncontrollable parts of the environment, e.g., the Byzantine agents, eventually induce a stationary MMDP, where the objective function $J^+(\pi^+,\pi^-)$ can be maximized over $\pi^+$. We note that we do not assume a stationary behavior of the adversarial agents in the numerical simulations presented in Section V.

To distinguish between the parameters of cooperative and Byzantine agents, we make slight changes in the notation. Without loss of generality, we assume that the agents' indices are ordered such that $\mathcal{N}^+ = \{1,\dots,N^+\}$ and $\mathcal{N}^- = \{N-N^-+1,\dots,N\}$. We use superscripts $+$ and $-$ to denote signals of all cooperative and Byzantine agents, respectively. For example, the cooperative agents' rewards are given as $r_{t+1}^+ = [(r_{t+1}^1)^T \dots (r_{t+1}^{N^+})^T]^T$. We have the following result on the consensus updates.

**Proposition 7.** *[34, Prop. 5.1] Under Assumption 7, the resilient consensus update for each* $i \in \mathcal{N}^+$ *with weights* $c_{x,t}(i,j)$ *is mathematically equivalent to* $\epsilon_{x,t}^i = \sum_{j\in\mathcal{N}_{in,t}^i\cap\mathcal{N}^+} c_{x,t}^+(i,j)\epsilon_{x,t}^{ij}$, *where* $c_{x,t}^+(i,j)$ *are consensus weights that satisfy* $\sum_{j\in\mathcal{N}_{in,t}^i\cap\mathcal{N}^+} c_{x,t}^+(i,j) = 1$. *Moreover, it holds that* $c_{x,t}^+(i,i) \geq \nu$ *and* $c_{x,t}^+(i,j) \geq \nu/2$ *for some* $\nu > 0$ *and* $j \in \mathcal{N}_{in,t}^i \cap \mathcal{N}^+$.

Proposition 7 ensures that the consensus updates of the cooperative agents can be expressed purely in terms of the cooperative agents' values. We let $C_{x,t}^+ = [c_{x,t}^+(i,j)]_{ij}$, $x \in \{v,\lambda\}$, denote the equivalent consensus matrix and define $A_{x,t}^+ = (I \otimes A'_{x,t})$ and $b_{x,t}^+ = [(b_{x,t}^1)^T \dots b_{x,t}^{N^+})^T]^T$. Applying Proposition 7, we write the cooperative agents' updates in Algorithm 2 as follows

$$\begin{aligned} v_{t+1}^+ =& v_t^+ + (C_{v,t}^+ \otimes \Gamma_{v,t})(v_t^+ + \alpha_{v,t}(A_{v,t}^+ v_t^+ + b_{v,t}^+)) \\ & - (I \otimes \Gamma_{v,t})v_t^+ \qquad (15) \\ \lambda_{t+1}^+ =& \lambda_t^+ + (C_{\lambda,t}^+ \otimes \Gamma_{\lambda,t})(\lambda_t^+ + \alpha_{\lambda,t}(A_{\lambda,t}^+ \lambda_t^+ + b_{\lambda,t}^+)) \\ & - (I \otimes \Gamma_{\lambda,t})\lambda_t^+. \qquad (16) \end{aligned}$$

While these updates are similar to those in Algorithm 1, the consensus matrices $C_{v,t}^+$ and $C_{\lambda,t}^+$ are influenced by the Byzantine agents and are not unique in general. For $x \in \{x,\lambda\}$, we let $\xi_{x,t} = (r_t, s_t, a_t, C_{x,t-1}^+)$ denote a collection of random variables and $W_x$ all possible realizations of the Markov chain $\{\xi_{x,t}\}_{t\geq 0}$. The main convergence results regarding the policy evaluation are now given.

**Lemma 8.** *Under Assumption* 1-4, 7, *and* 8, *the critic parameters* $v_t^i$, $i \in \mathcal{N}^+$, *are uniformly bounded and converge to a consensus value* $\langle v_t^+ \rangle$ *with probability one. The consensus value* $\langle v_t^+ \rangle$ *converges with probability one to a bounded*

*neighborhood around a fixed point $v_\pi^+$ that satisfies*

$$\Phi^T D_\pi^s \left( \frac{1}{N^+} \sum_{i \in \mathcal{N}^+} R_\pi^i + \gamma P_\pi \Phi v_\pi^+ - \Phi v_\pi^+ \right) = 0.$$

*The limiting sequence of the team-average critic parameter, $\langle v^+ \rangle$, satisfies $\|\Phi^T D_\pi^s (\gamma P_\pi - I) \Phi (\langle v^+ \rangle - v_\pi^+)\| \leq \|\Delta_v\|$, where*

$$\|\Delta_v\| \leq \lim_{t,m} \sup_{\xi_t \in W} \left\| \frac{1}{m} \sum_{k=t}^{t+m-1} \frac{1}{N^+} \left( (\mathbf{1}^T C_{v,k}^+ r_{k+1}^+ \otimes \phi_k) + (\mathbf{1}^T C_{v,k}^+ \otimes \frac{\phi_k \phi_k^T}{\|\phi_k\|^2}) \alpha_{v,k}^{-1} v_{\perp,k}^+ \right) - \frac{1}{N^+} \sum_{i \in \mathcal{N}^+} \Phi^T D_\pi^s R_\pi^i \right\|. \tag{17}$$

**Lemma 9.** *Under Assumption 1-4, 7, and 8, the team-average reward function parameters $\lambda_t^i$ are uniformly bounded and converge to a consensus value $\langle \lambda_t^+ \rangle$ with probability one. The consensus value $\langle \lambda_t^+ \rangle$ converges with probability one to a bounded neighborhood around a fixed point $\lambda_\pi^+$ that satisfies*

$$F^T D_\pi^{s,a} \left( \frac{1}{N^+} \sum_{i \in \mathcal{N}^+} R^i - F \lambda_\pi^+ \right) = 0.$$

*The limiting sequence of the team-average update, $\langle \lambda^+ \rangle$, satisfies $\|F^T D_\pi^{s,a} F (\langle \lambda^+ \rangle - \lambda_\pi^+)\| \leq \|\Delta_\lambda\|$, where*

$$\|\Delta_\lambda\| = \lim_{t,m \to \infty} \sup_{\xi_t \in W} \left\| \frac{1}{m} \sum_{k=t}^{t+m-1} \frac{1}{N^+} \left( (\mathbf{1}^T C_{\lambda,k}^+ r_{k+1}^+ \otimes f_k) + (\mathbf{1}^T C_{\lambda,k}^+ \otimes \frac{f_k f_k^T}{\|f_k\|^2}) \alpha_{\lambda,t}^{-1} \lambda_{\perp,k}^+ \right) - \frac{1}{N^+} \sum_{i \in \mathcal{N}^+} F^T D_\pi^{s,a} R^i \right\|. \tag{18}$$

**Remark 2.** *The critic and team-average reward function parameters converge to a bounded neighborhood around the optimal MMSE parameters. The size of the neighborhood depends primarily on the behavior of the Byzantine agents. We note that the equivalent consensus matrices $C_{v,t}^+$ and $C_{\lambda,t}^+$ are not column stochastic in general since they are influenced by the parameter values communicated by the Byzantine agents. Thus, the parameters estimated by the cooperative agents generally cannot converge to the optimal MMSE parameters of the critic and team-average reward function, $v_\pi$ and $\lambda_\pi$. A special case that permits this to happen includes when all cooperative agents receive the same reward, i.e., $r^i = r^j$ for $i, j \in \mathcal{N}^+$ and $t \geq 0$.*

**Theorem 10.** *Under Assumption 1-5, 7, and 8, the policy parameter $\theta_t^i$, $i \in \mathcal{N}^+$, converges with probability one to a neighborhood of a locally asymptotically stable equilibrium of the ODE*

$$\dot{\theta}^i = \Psi_\Theta^i [\mathbb{E}_{\pi, d_\pi, p} \{ (\bar{r}(s,a;\lambda_\pi^+) + \gamma V(s';v_\pi^+) - V(s;v_\pi^+)) \cdot \nabla \log \pi^i (a^i | s; \theta^i) \}].$$

Due to the ever-present disturbance in the approximation of the critic and team-average reward function, the cooperative policy converges to a compact set of policies that are at best close to a policy that forms a stationary point of the cooperative team's objective function $J^+(\pi^+, \pi^-)$. The size of the limit set is determined by the size of the deviation from the optimal MMSE parameters in the distributed estimation, that is $\bar{r}(s,a;\langle \lambda^+ \rangle - \lambda_\pi^+)$ and $V(s';\langle v^+ \rangle - v_\pi^+)$. In the worst case scenario, the limit set is equal to the constraint set $\Theta^i$. However, if the deviation from the optimal MMSE parameters is small, then the size of the limit set is small as well.

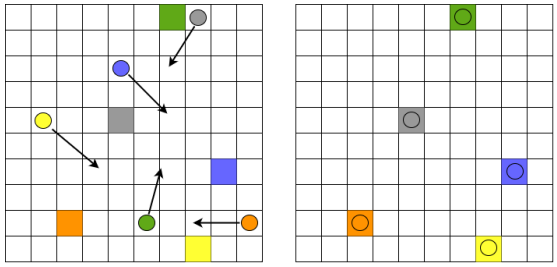


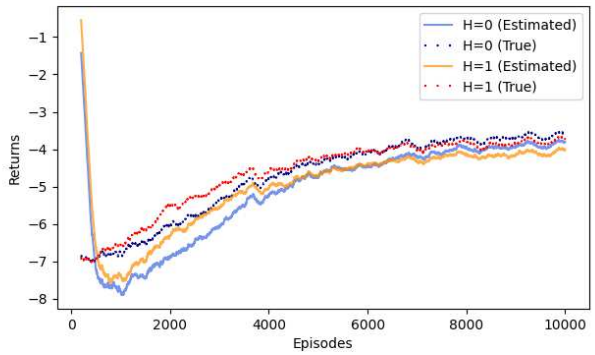
Fig. 2: Cooperative navigation task in the grid world environment.

## V. Simulation Results

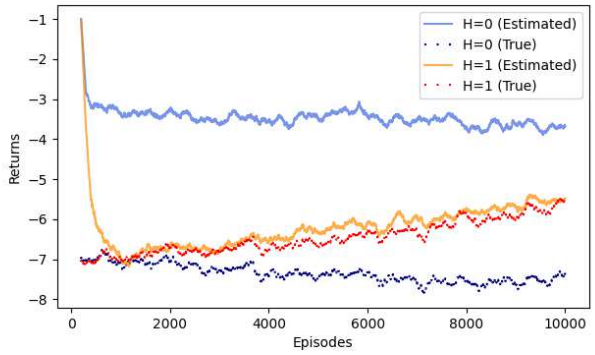
For a numerical illustration, we consider a multi-agent grid world environment of size $10 \times 10$ (Figure 2), where five agents learn to solve the cooperative navigation task by communicating in an all-to-all fashion. The cooperative goal of the agents is to follow a path to their respective targets that yields the highest team-average returns. We let $s^i$ denote the 2D positional coordinates of agent $i$ and $d^i$ the 2D positional coordinates of its desired position. Each agent chooses from a set of five actions that correspond to the cardinal direction of its next state transition and staying put in the same state. A transition to an adjacent cell occurs only if the cell is not occupied by another agent. The dynamics are deterministic and the agents are rewarded by $-\|s'^i - d^i\|_1$, and an additional $-1$ if they attempt to move into an occupied cell or leave the boundaries of the grid world. Each agent approximates the actor, critic, and team-average reward with neural networks that have two dense hidden layers with 30 units and leaky ReLU activation function with $\alpha = 0.1$. The actor includes an output layer with a softmax activation.

We define an episode as a single run of the simulation from start to finish, consisting of a sequence of time steps in the agents' dynamics described above, where the initial and target positions of each agent are chosen randomly from the $10 \times 10$ grid. We train the neural networks for 10000 episodes with each episode lasting 20 steps. Within an episode, agents interact with the environment to collect experiences, which are tuples of (state, action, reward, next state). The team-average reward function $\bar{r}(s,a;\lambda^i)$ and the critic $V(s;v^i)$ are evaluated under a fixed policy $\pi(a|s;\theta)$ every 100 episodes. The agents perform local updates and resilient projection-based consensus updates in Algorithm 2 every 20 episodes based on the most recent 1000 experiences obtained under the current policy as well as further 2000 experiences from a replay buffer. The hidden layer parameters are aggregated using the method of trimmed means. The actor is updated based on the 1000 most recent experiences. We select the discount factor $\gamma = 0.9$ and the learning rates $\alpha_{\theta,t} = 0.0005$ and $\alpha_{v,t} = \alpha_{\lambda,t} = 0.05$. Each

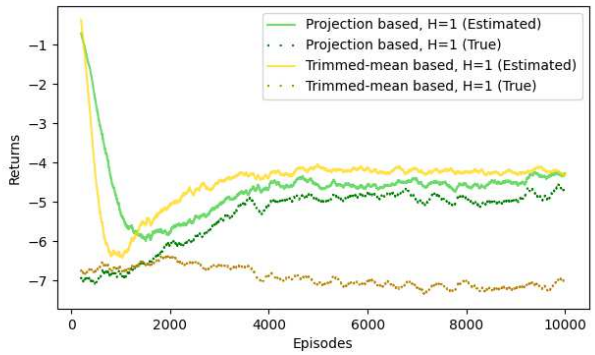
episode takes around 7.9 seconds to complete.[1]




(a) All-cooperative, $H = 0, 1$




(b) Greedy, $H = 0, 1$




(c) Strategic, $H = 1$, Projection based and Trimmed-mean based

Fig. 3: True and estimated team-averaged rewards among the cooperative agents.

We consider three scenarios. The results presented in Figure 3 correspond to a rolling average of Returns over 200 episodes, where the Return value is the (true or estimated) cumulative team-averaged reward (among the cooperative agents) in a single episode. In the first scenario, we assume that all five agents are cooperative. From Fig. 3a, we see that with the trimming parameter in the projection step in Algorithm 2 at either $H = 0$ or $H = 1$, the true and estimated team-averaged rewards are close to each other. In the second scenario, we assume that one agent is greedy and sends its parameter values to the other agents but does not utilize any values from them. From Fig. 3b, we see that when $H = 0$, the cooperative agents overestimate the team-averaged reward and end up maximizing the greedy agent's objective. In contrast, when $H = 1$, the cooperative agents are resilient to the greedy agent, and better estimate and maximize the team-average objective function among the cooperative agents. In the last scenario, we assume that one strategic agent attempts to minimize the cooperative agents' objective and maximize its own objective. Here, we compare between the performances of Algorithm 2 and the algorithm proposed in [27]. As we described in Section II, the algorithm proposed in [27] is based on trimmed means. To obtain the results in Fig. 3c, we set $H = 1$ in both the trimmed-mean based algorithm in [27] and our projection-based algorithm. From Fig. 3c, we see that the trimmed-mean based algorithm tends to overestimate the team-average reward so that the cooperative agents cannot learn an optimal policy to maximize the team-averaged reward. In contrast, Algorithm 2 is resilient to this strategic agent, and again better estimates and maximizes the team-averaged reward.

[1]All the experiments are performed on a computer cluster with 1 NVIDIA A30 GPU and 192GB RAM. Code available at `https://github.com/mainakpal08/Resilient-consensus-based-MARL`.

## VI. Conclusion

We introduced novel resilient projection-based consensus actor-critic MARL algorithms that ensure Byzantine-resilient learning of cooperative agents in environments influenced by Byzantine agents. We provided a convergence analysis of the algorithm that uses linear approximation. In simulations, we implemented the resilient algorithm that employs nonlinear approximation and demonstrated its functionality.

## Appendix A
## Stochastic Approximation

For completeness, we state here standard results on stochastic approximation.

*1) Unconstrained Stochastic Approximation with Correlated Noise:* We let $\theta_n$, $Y_n$ and $\xi_n$ denote the estimated parameter, observation, and state of a Markov chain, respectively. We define the filtration $\mathcal{F}_n = \sigma(\theta_0, Y_{i-1}, \xi_i, i \leq n)$. The unconstrained stochastic updates are given as follows

$$\theta_{n+1} = \theta_n + \epsilon_n[\mathbb{E}(Y_n|\mathcal{F}_n) + \delta M_n] \tag{19}$$
$$= \theta_n + \epsilon_n[g_n(\theta_n, \xi_n) + \delta M_n + \beta_n] \tag{20}$$

where $\epsilon_n > 0$ and $\delta M_n = Y_n - \mathbb{E}(Y_n|\mathcal{F}_n)$ is a martingale difference.

**Assumption 9.** *Consider the following assumptions:*

*(1) The function $g_n(\theta_n, \xi_n)$ is Lipschitz continuous in the first argument.*

*(2) The step size sequence $\{\epsilon_n\}_{n\geq 0}$ satisfies $\sum_n \epsilon_n = \infty$ and $\sum_n \epsilon_n^2 < \infty$, for $n \geq 0$.*

*(3) The martingale difference sequence $\{\delta M_n\}_{n\geq 0}$ satisfies $\mathbb{E}(\|\delta M_{n+1}\|^2|\mathcal{F}_n) \leq K \cdot (1 + \|\theta_n\|^2)$ for all $n \geq 0$ and some $K > 0$.*

*(4) The random sequence $\{\beta_n\}_{n\geq 0}$ is bounded and satisfies $\beta_n \to 0$ with probability one.*

*(5) $\{\xi_n\}_{n\geq 0}$ is an irreducible Markov chain with stationary distribution $\eta$.*
*(6) The Markov chain $\{\xi_n\}_{n\geq 0}$ is uniformly bounded and has a set of occupation measures $\mathcal{D}(\theta)$ for any $\theta$.*

**Theorem 11.** *[31, Chapter 6] Under Assumption 9.(1)-(5), the asymptotic behavior of the algorithm (20) is described by the ODE*

$$\dot{\theta} = \bar{g}(\theta) := \mathbb{E}_{i\in\eta}[g(\theta, i)]. \tag{21}$$

**Theorem 12.** *[31, Chapter 6] Under Assumption 9.(1)-(5), suppose that $\lim_{c\to\infty} \bar{g}(c\theta)\cdot c^{-1} = g_\infty(\theta)$ exists uniformly on compact sets for some $g_\infty \in C(\mathbb{R}^n)$. If the ODE $\dot{\theta} = g_\infty(\theta)$ has the origin as the unique globally asymptotically stable equilibrium, then $\sup_n\|\theta_n\| < \infty$ with probability one.*

**Theorem 13.** *[31, Chapter 6] If the ODE (21) has a unique globally asymptotically stable equilibrium $\theta^*$ and $\sup_n\|\theta_n\| < \infty$ with probability one, then $\theta_n \to \theta^*$ as $n \to \infty$ with probability one.*

**Theorem 14.** *[30, Chapter 6] Under Assumption 9.(1)-(4) and (6), the asymptotic behavior of the algorithm (20) is described by the differential inclusion*

$$\dot{\theta} \in G(\theta) := \left\{ \lim_{n,m\to\infty} \frac{1}{m} \sum_{i=n}^{n+m-1} g_i(\theta, j),\ j \in \mathcal{D} \right\}. \tag{22}$$

**Theorem 15.** *[35, Theorem 2] Under Assumption 9.(1)-(4) and (6), suppose that $\lim_{c\to\infty} G(c\theta)\cdot c^{-1} = g_\infty(\theta)$ exists uniformly on compact sets for all $i \in \mathcal{D}$ and some $g_\infty \in C(\mathbb{R}^n)$. If the ODE $\dot{\theta} = g_\infty(\theta, i)$ has the origin as the unique globally asymptotically stable equilibrium, then $\sup_n\|\theta_n\| < \infty$ with probability one.*

**Theorem 16.** *[30, Chapter 6] If $\sup_n\|\theta_n\| < \infty$, then the trajectories converge to the limit set of the differential inclusion $\dot{\theta} \in G(\theta)$.*

*2) Constrained Stochastic Approximation with Martingale Difference Noise:* We let $\theta_n$ and $Y_n$ denote the estimated parameter and observation, respectively. We define the filtration $\mathcal{F}_n = \sigma(\theta_0, Y_{i-1}, i \leq n)$. The constrained stochastic updates are given as follows

$$\theta_{n+1} = \Psi_\Theta(\theta_n + \epsilon_n[\mathbb{E}(Y_n|\mathcal{F}_n) + \delta M_n]) \tag{23}$$
$$= \Psi_\Theta(\theta_n + \epsilon_n[g_n(\theta_n) + \delta M_n + \beta_n]), \tag{24}$$

where $\Psi_\Theta(\cdot)$ is a projection operator that maps the stochastic updates into a compact admissible set $\Theta$, and $\delta M_n = Y_n - \mathbb{E}(Y_n|\mathcal{F}_n)$ is a martingale difference. We introduce assumptions for the algorithm updates.

**Assumption 10.** *Consider the following assumptions:*
*(1) $\sup_n \mathbb{E}(\|Y_n\||\mathcal{F}_n) < \infty$.*
*(2) The step size sequence $\epsilon_n$ satisfies $\sum_n \epsilon_n^2 < \infty$ and $\lim_{n\to\infty} \frac{\epsilon_{n+1}}{\epsilon_n} = 1$.*
*(3) The random sequence $\{\beta_n\}_{n\geq 0}$ satisfies $\beta_n \to 0$ with probability one.*
*(4) The admissible set $\Theta$ is a hyperrectangle, i.e., there exist $a$ and $b$ such that $a < b$ and $\Theta = \{\theta_n : a \leq \theta_n \leq b\}$.*
*(5) The function $g_n(\cdot)$ is continuous uniformly in $n$. Furthermore, there exists a function $\bar{g}(\theta)$ such that for all $m > 0$, we have $\lim_{n\to\infty} \|\sum_{i=n}^{n+m-1} \epsilon_i[g_i(\theta) - \bar{g}(\theta)]\| = 0$.*
*(6) The function $g_n(\cdot)$ is continuous uniformly in $n$. Moreover, there exists an upper semicontinuous set-valued function $G(\theta)$ such that $\lim_{n,m\to\infty} \frac{1}{m}\sum_{i=n}^{n+m-1} g_i(\theta) \in G(\theta)$.*

**Theorem 17.** *[36, Chapter 5] Under Assumption 10.(1)-(5), the asymptotic behavior of the algorithm (24) is described by the ODE $\dot{\theta} = \Psi_\Theta[\bar{g}(\theta)]$.*

Theorem 18 is a direct consequence of Theorem 17.

**Theorem 18.** *Under Assumption 10.(1)-(6), if there exists a continuously differentiable function $\omega(\theta)$ such that $\bar{g}(\theta) = \frac{d\omega}{d\theta}$ and $\omega(\theta)$ is constant on disjoint compact sets $\mathcal{L}_i$, $i = 1, \ldots, M$, then the parameters $\theta_n$ converge with probability one to $\mathcal{L}_i$ for some $i \in \{1, \ldots, M\}$ as $n \to \infty$.*

**Theorem 19.** *[36, Chapter 5] Under Assumption 10.(1)-(4) and (6), the limit points are contained in an invariant set of the differential inclusion $\dot{\theta} \in \Psi_\Theta[G(\theta)]$.*

## Appendix B
## Proofs of Theoretical Results in Section IV

We repeatedly use shorthand $\pi_t = \pi(a|s;\theta_t)$ to describe the dependence of the policy on the time-varying parameters $\theta_t$.

*1) Algorithm 1:* To complete the proofs, we need to establish several technical lemmas. First, we analyze the spectral radius of the mean consensus update in the disagreement subspace. We let $\mathcal{F}_t^x = \sigma(x_0, Y_{t-1}, \xi_\tau, \tau \leq t)$ denote a filtration of a random variable $x \in \{v, \lambda\}$, where $Y_t$ are the incremental changes in parameters $v$ and $\lambda$ due to the updates in (13) or (14), and $\xi_\tau = (r_\tau, s_\tau, a_\tau, C_{x,\tau-1})$ is a collection of random variables. The filtration captures the evolution of the random variable $x$ based on its initial value $x_0$ and updates $Y_t$ that occur along the trajectory of the Markov chain $\xi_\tau$.

**Lemma 20.** *Under Assumption 6, the spectral radius $\rho(\mathbb{E}[C_{x,t}^T(I - \mathbf{1}\mathbf{1}^T/N)C_{x,t}|\mathcal{F}_t^x]) < 1$, where $x \in \{v, \lambda\}$.*

*Proof.* We begin by showing some properties of the matrix $I - \mathbf{1}\mathbf{1}^T/N$. First, since $I - \mathbf{1}\mathbf{1}^T/N$ is diagonal dominant and symmetric, we know that $I - \mathbf{1}\mathbf{1}^T/N \succeq 0$ [37, Chapter 7]. Considering any $y \in \mathbb{R}^N$, we have $y^T(I - \mathbf{1}\mathbf{1}^T/N)y \leq y^T y$, which implies that $I - \mathbf{1}\mathbf{1}^T/N \preceq I$. Using the above properties of $I - \mathbf{1}\mathbf{1}^T/N$, we obtain $y^T C_{x,t}^T(I - \mathbf{1}\mathbf{1}^T/N)C_{x,t}y \leq y^T C_{x,t}^T C_{x,t} y = \|C_{x,t}y\|^2$. Since $C_{x,t}$ is row stochastic and $\mathcal{G}_t$ is connected in the mean by Assumption 6, we know that $\rho(\mathbb{E}[C_{x,t}|\mathcal{F}_t^x]) = 1$ and $\|\mathbb{E}[C_{x,t}|\mathcal{F}_t^x]y\|^2 \leq \rho(\mathbb{E}[C_{x,t}|\mathcal{F}_t^x])^2\|y\|^2$ for all $y \in \mathbb{R}^N$ and the inequality holds when $y = \mathbf{1}$ [37, Chapter 8]. Note also that $\mathbf{1}^T C_{x,t}^T(I - \mathbf{1}\mathbf{1}^T/N)C_{x,t}\mathbf{1} = 0$. Combining the above arguments, we have $y^T\mathbb{E}[C_{x,t}^T(I - \mathbf{1}\mathbf{1}^T/N)C_{x,t}|\mathcal{F}_t^x]y < \|y\|^2$ for all $y \in \mathbb{R}^N$. Finally, since $C_{x,t}^T(I - \mathbf{1}\mathbf{1}^T/N)C_{x,t} \succeq 0$, we know that $\rho(\mathbb{E}[C_{x,t}^T(I - \mathbf{1}\mathbf{1}^T/N)C_{x,t}|\mathcal{F}_t^x]) = \max_{y\in\mathbb{R}^N, \|y\|=1} y^T\mathbb{E}[C_{x,t}^T(I - \mathbf{1}\mathbf{1}^T/N)C_{x,t}|\mathcal{F}_t^x]y < 1$ [37, Chapter 4]. □

**Lemma 21.** *Suppose there are no adversarial agents in the network. Under Assumption 1-4 and 6, the agents reach consensus on the critic parameters with probability one, i.e.,* $\lim_{t\to\infty} v_{\perp,t} = 0$. *Furthermore, the term* $\mathbb{E}(\|(I \otimes \Gamma_{v,t})\alpha_{v,t}^{-1} v_{\perp,t}\|^2 | \mathcal{F}_t^v)$ *is uniformly bounded.*

**Lemma 22.** *Suppose there are no adversarial agents in the network. Under Assumption 1-4 and 6, the agents reach consensus on the team-average reward function parameters with probability one, i.e.,* $\lim_{t\to\infty} \lambda_{\perp,t} = 0$. *Furthermore, the term* $\mathbb{E}(\|(I \otimes \Gamma_{\lambda,t})\alpha_{\lambda,t}^{-1} \lambda_{\perp,t}\|^2 | \mathcal{F}_t^\lambda)$ *is uniformly bounded.*

*Proof.* Since the proofs of Lemma 21 and 22 are analogous, We only present the proof for the critic parameters $v_t$. The updates of $v_{\perp,t}$ can be compactly written as follows

$$\begin{aligned}
&v_{\perp,t+1}\\
=&\mathcal{J}_\perp[v_t + (C_{v,t} \otimes \Gamma_{v,t})(v_t + \alpha_{v,t}(A_{v,t} v_t + b_{v,t}))\\
&- (I \otimes \Gamma_{v,t}) v_t]\\
=&\mathcal{J}_\perp[(I \otimes \hat{\Gamma}_{v,t}) v_t + (C_{v,t} \otimes \Gamma_{v,t})(v_t + \alpha_{v,t}(A_{v,t} v_t + b_{v,t}))]\\
=&\mathcal{J}_\perp[(C_{v,t} \otimes \Gamma_{v,t})(v_{\perp,t} + \alpha_{v,t}(A_{v,t} v_{\perp,t} + b_{v,t}))]\\
&+ (I \otimes \hat{\Gamma}_{v,t}) v_{\perp,t}, \qquad (25)
\end{aligned}$$

where we used the fact that $(C_{v,t} \otimes \Gamma_{v,t})(\mathbf{1} \otimes \langle x \rangle) = \mathbf{1} \otimes (\Gamma_{v,t} \langle x \rangle)$ and $\mathcal{J}_\perp[\mathbf{1} \otimes (\Gamma_{v,t} \langle x \rangle)] = 0$. The updates of $v_{\perp,t}$ can be viewed in two mutually orthogonal subspaces as follows $v_{\perp,t+1} = (I \otimes \hat{\Gamma}_{v,t}) v_{\perp,t+1} + (I \otimes \Gamma_{v,t}) v_{\perp,t+1}$, where

$$\begin{aligned}
(I \otimes \hat{\Gamma}_{v,t}) v_{\perp,t+1} =& (I \otimes \hat{\Gamma}_{v,t}) v_{\perp,t} \qquad (26)\\
(I \otimes \Gamma_{v,t}) v_{\perp,t+1} =& [(I - \mathbf{1}\mathbf{1}^T/N) C_{v,t} \otimes I](I \otimes \Gamma_{v,t})\\
&\cdot (v_{\perp,t} + \alpha_{v,t}(A_{v,t} v_t + b_{v,t})) \qquad (27)
\end{aligned}$$

Consider the expected values $\mathbb{E}(\|(I \otimes \hat{\Gamma}_{v,t}) v_{\perp,t+1}\|^2 | \mathcal{F}_t^v)$ and $\mathbb{E}(\|(I \otimes \Gamma_{v,t}) v_{\perp,t+1}\|^2 | \mathcal{F}_t^v)$. By (26), the first term satisfies $\mathbb{E}(\|(I \otimes \hat{\Gamma}_{v,t}) v_{\perp,t+1}\|^2 | \mathcal{F}_{v,t}^v) = \mathbb{E}(\|(I \otimes \hat{\Gamma}_{v,t}) v_{\perp,t}\|^2 | \mathcal{F}_t^v)$ and the disagreement vector $v_{\perp,t}$ remains stable in the orthogonal subspace. To analyze stability of the term in (27), we write

$$\begin{aligned}
&\mathbb{E}(\|(I \otimes \Gamma_{v,t}) v_{\perp,t+1}\|^2 | \mathcal{F}_t^v)\\
=&\mathbb{E}(\|[(I - \mathbf{1}\mathbf{1}^T/N) C_{v,t} \otimes I](I \otimes \Gamma_{v,t})\\
&\cdot ((I + \alpha_{v,t} A_{v,t}) v_{\perp,t} + \alpha_{v,t} b_{v,t}))\|^2 | \mathcal{F}_t^v)\\
\le& \rho_{v,t} \cdot \mathbb{E}(\|(I \otimes \Gamma_{v,t})((I + \alpha_{v,t} A_{v,t}) v_{\perp,t} + \alpha_{v,t} b_{v,t}))\|^2 | \mathcal{F}_t^v) \qquad (28)
\end{aligned}$$

where $\rho_{v,t} = \rho(\mathbb{E}(C_{v,t}(I - \mathbf{1}\mathbf{1}^T/N) C_{v,t} | \mathcal{F}_t^v))$. We note that the inequality holds by the conditional independence of $C_{v,t}$ stated in Assumption 6. Regarding the terms that involve $v_{\perp,t}$, we have $I + \alpha_{v,t} A_{v,t} = I \otimes (I + \alpha_{v,t} A'_{v,t})$, and hence we obtain $(I \otimes \Gamma_{v,t})(I + \alpha_{v,t} A_{v,t}) = I \otimes (\Gamma_{v,t} + \alpha_{v,t} A'_{v,t})$. The eigenspace of matrix $\Gamma_{v,t} + \alpha_{v,t} A'_{v,t}$ is spanned by vector $\Gamma_{v,t} \phi_t$. Hence, for $\nu_t$ that satisfies $\nu_t \le (1 + \alpha_{v,t} K_1)^2$, where $K_1 = \sup_t \|(\gamma \phi_{t+1} - \phi_t)^T \phi_t\| < \infty$ by Assumption 1, we have $\|(I \otimes \Gamma_{v,t})(I + \alpha_{v,t} A_{v,t}) v_{\perp,t}\|^2 \le \nu_t \cdot \|(I \otimes \Gamma_{v,t}) v_{\perp,t}\|^2$. We apply this inequality in the following lines, where we first premultiply both sides of (28) by $\alpha_{v,t+1}^{-2}$ and apply the triangle inequality. Letting $\eta_{t+1} = \mathbb{E}(\|(I \otimes \Gamma_{v,t}) \alpha_{v,t+1}^{-1} v_{\perp,t+1}\|^2 | \mathcal{F}_t^v)$, we obtain the following inequality

$$\begin{aligned}
\eta_{t+1} \le& \rho_{v,t} \cdot \nu_t \cdot \frac{\alpha_{v,t}^2}{\alpha_{v,t+1}^2} (\eta_t + 2\nu_t^{-\frac{1}{2}} \eta_t^{\frac{1}{2}} \cdot \mathbb{E}(\|b_{v,t}\| | \mathcal{F}_t^v)\\
&+ \nu_t^{-1} \cdot \mathbb{E}(\|b_{v,t}\|^2 | \mathcal{F}_t^v)).
\end{aligned}$$

Under Assumption 4, $\lim_{t\to\infty} \frac{\alpha_{v,t}^2}{\alpha_{v,t+1}^2} = 1$ and $\lim_{t\to\infty} \nu_t = 1$, and so there exists finite time $t_0$ and constant $\delta > 0$ such that $\nu_t > 0$ and $\rho_{v,t} \cdot \nu_t \cdot \frac{\alpha_{v,t}^2}{\alpha_{v,t+1}^2} \le 1 - \delta$ for all $t > t_0$. Since $b_{v,t}$ is uniformly bounded by Assumption 1 and 3, we have $\nu_t^{-1} \cdot \mathbb{E}(\|b_{v,t}\|^2 | \mathcal{F}_t^v) \le K_2$ for some $K_2 < \infty$. Therefore, for $t > t_0$ we can write

$$\begin{aligned}
\eta_{t+1} \le& (1 - \delta)(\eta_t + 2\sqrt{\eta_t} \cdot \sqrt{K_2} + K_2)\\
=& (1 - \frac{\delta}{2})\eta_t - \frac{\delta}{2}(\sqrt{\eta_t} - \frac{2}{\delta}(1 - \delta)\sqrt{K_2})^2\\
&+ \frac{2}{\delta}(1 - \delta)^2 K_2 + (1 - \delta) K_2 \le (1 - \delta/2)\eta_t + K_3,
\end{aligned}$$

where $K_3 = \frac{2}{\delta}(1 - \delta)^2 K_2 + (1 - \delta) K_2$. By induction, $\eta_t \le (1 - \frac{\delta}{2})^{t - t_0} \eta_{t_0} + \frac{2K_3}{\delta}$. Therefore, we have $\sup_t \mathbb{E}(\|(I \otimes \Gamma_{v,t}) \alpha_{v,t}^{-1} v_{\perp,t}\|^2 | \mathcal{F}_t^v) < K_4$ for some $K_4 > 0$. Since the states are visited according to the stationary distribution $d_\pi(s)$, the uniform bound holds for all $\Gamma_{v,t}$ visited in the infinite sequence. Therefore, we consider $\sum_t \mathbb{E}(\|(I \otimes \Gamma_{v,t}) v_{\perp,t}\|^2 | \mathcal{F}_t^v) < K_4 \cdot \sum_t \alpha_{v,t}^2$ and obtain $\lim_{t\to\infty} (I \otimes \Gamma_{v,t}) v_{\perp,t} = 0$ with probability one by Assumption 4. This implies that $\lim_{t\to\infty} v_{\perp,t} = 0$ with probability one. □

From Lemma 21 and 22, under a sufficiently small step size, the disagreement vector scaled by the step size is contractive and subject to a bounded "input" disturbance that is due to the heterogeneous rewards observed by individual agents. Therefore, if the trajectories of $\alpha_{v,t}^{-1} v_{\perp,t}$ and $\alpha_{\lambda,t}^{-1} \lambda_{\perp,t}$ happen to escape a compact set for $t > t_0$, the trajectories exponentially converge back to the set, which implies the boundedness in the disagreement space.

*Proof.* (**Lemma 4**) We write the iteration of $\langle v_t \rangle$ as follows

$$\begin{aligned}
&\langle v_{t+1} \rangle\\
=&\left\langle (I \otimes \hat{\Gamma}_{v,t}) v_t \right\rangle + \left\langle (C_{v,t} \otimes \Gamma_{v,t})(\mathbf{1} \otimes \langle v_t \rangle + v_{\perp,t}\right.\\
&\left. + \alpha_{v,t}[(I \otimes A'_{v,t})(\mathbf{1} \otimes \langle v_t \rangle + v_{\perp,t}) + b_{v,t}]) \right\rangle\\
=&\Gamma_{v,t} \langle v_t \rangle + \hat{\Gamma}_{v,t} \langle v_t \rangle\\
&+ \alpha_{v,t} \left\langle (C_{v,t} \otimes \Gamma_{v,t})(I \otimes A'_{v,t})(\mathbf{1} \otimes \langle v_t \rangle) \right\rangle\\
&+ \alpha_{v,t} \left\langle (C_{v,t} \otimes \Gamma_{v,t})(\alpha_{v,t}^{-1} v_{\perp,t} + (I \otimes A'_{v,t}) v_{\perp,t} + b_{v,t}) \right\rangle\\
=&\langle v_t \rangle + \alpha_{v,t} A'_{v,t} \langle v_t \rangle\\
&+ \alpha_{v,t} \left\langle (C_{v,t} \otimes \Gamma_{v,t})(\alpha_{v,t}^{-1} v_{\perp,t} + A_{v,t} v_{\perp,t} + b_{v,t}) \right\rangle\\
=&\langle v_t \rangle + \alpha_{v,t}[g_t(\langle v_t \rangle, \xi_t) + \delta M_t + \beta_t],
\end{aligned}$$

where the functions $g_t(\cdot,\cdot)$, $\delta M_t$, and $\beta_t$ are given as

$$\begin{aligned} g_t(\langle v_t\rangle, \xi_t) &= \mathbb{E}(A'_{v,t}\langle v_t\rangle + \langle (C_{v,t}\otimes\Gamma_{v,t})b_{v,t}\rangle \,|\mathcal{F}_t^v) \\ \delta M_t &= A'_{v,t}\langle v_t\rangle + \langle (C_{v,t}\otimes I)b_{v,t}\rangle \\ &\quad + \left\langle (C_{v,t}\otimes\Gamma_{v,t})(\alpha_{v,t}^{-1}v_{\perp,t} + A_{v,t}v_{\perp,t})\right\rangle \\ &\quad - \mathbb{E}\Big(A'_{v,t}\langle v_t\rangle + \langle (C_{v,t}\otimes\Gamma_{v,t})b_{v,t}\rangle \\ &\quad + \left\langle (C_{v,t}\otimes\Gamma_{v,t})(\alpha_{v,t}^{-1}v_{\perp,t} + A_{v,t}v_{\perp,t})\right\rangle |\mathcal{F}_t^v\Big) \\ \beta_t &= \mathbb{E}(\left\langle (C_{v,t}\otimes\Gamma_{v,t})(\alpha_{v,t}^{-1}v_{\perp,t} + A_{v,t}v_{\perp,t})\right\rangle |\mathcal{F}_t^v). \end{aligned}$$

We now verify the conditions in Appendix A-1.

(1) We have $\|g_t(\langle x\rangle,\xi_t) - g_t(\langle y\rangle,\xi_t)\| = \|\mathbb{E}(A_{v,t}(\langle x\rangle - \langle y\rangle)|\mathcal{F}_t^v)\| \le K_1\cdot\|\langle x\rangle - \langle y\rangle\|^2$ for some $K_1>0$ since $A'_{v,t}$ is uniformly bounded by Assumption 1. Therefore, $g_t(\langle v_t\rangle,\xi_t)$ is Lipschitz continuous in $\langle v_t\rangle$.

(2) The step size sequence $\{\alpha_{v,t}\}_{t\ge0}$ satisfies $\sum_t \alpha_{v,t}=\infty$ and $\sum_t \alpha_{v,t}^2<\infty$, for $t\ge0$.

(3) The martingale difference sequence $\delta M_t$ satisfies $\mathbb{E}(\|\delta M_t\|^2|\mathcal{F}_t^v) \le K_2\cdot(1+\|\langle v_t\rangle\|^2)$, since $A_{v,t}$, $b_{v,t}$, and $\alpha_{v,t}^{-1}v_{\perp,t}$ are uniformly bounded by Assumption 1 and 3, and Lemma 21.

(4) By Assumption 6, we can write $\beta_t = \mathbb{E}(\left\langle (C_t\otimes\Gamma_{v,t})(\alpha_{v,t}^{-1}v_{\perp,t}+A_{v,t}v_{\perp,t})\right\rangle|\mathcal{F}_t^v) = \frac{1}{N}(\mathbf{1}^T\otimes I)\mathbb{E}(C_t\otimes I|\mathcal{F}_t^v)\mathbb{E}((I\otimes\Gamma_{v,t})(\alpha_{v,t}^{-1}v_{\perp,t}+A_{v,t}v_{\perp,t})|\mathcal{F}_t^v) = \frac{1}{N}(\mathbf{1}^T\otimes I)\mathbb{E}((I\otimes\Gamma_{v,t})(\alpha_{v,t}^{-1}v_{\perp,t}+A_{v,t}v_{\perp,t})|\mathcal{F}_t^v)$. The last term is uniformly bounded by Lemma 21 and Assumption 1. This and the fact that $(\mathbf{1}^T\otimes I)(I\otimes\Gamma_{v,t})(\alpha_{v,t}^{-1}v_{\perp,t}+A_{v,t}v_{\perp,t}) = 0$ imply that $\beta_t=0$.

(5) The Markov chain $\{\xi_t\}_{t\ge0}$ is irreducible and has a stationary distribution $\eta$.

Applying Theorem 11, it follows that the asymptotic behavior is described by the ODE

$$\begin{aligned} \langle\dot v\rangle &= \bar g(\langle v\rangle) = \mathbb{E}_{d_\pi}[g_t(\langle v_t\rangle,\xi_t)] \\ &= \Phi^T D_\pi^s(\gamma P_\pi - I)\Phi\langle v\rangle + \frac{1}{N}\sum_{i\in\mathcal{N}}\Phi^T D_\pi^s R_\pi^i. \end{aligned}$$

Let $\lim_{c\to\infty}\bar g(cx)\cdot c^{-1} = g_\infty(x) = \Phi^T D_\pi^s(\gamma P_\pi - I)\Phi x$. Since $\Phi$ is full column rank by Assumption 1, let $\zeta$ and $\Phi y$ denote an arbitrary eigenvalue-eigenvector pair of the matrix product $D_\pi(\gamma P_\pi - I)$. Since $D_\pi(\gamma P_\pi - I)\Phi y = \zeta\Phi y$, we can write $y^T\Phi^T(\gamma P_\pi - I)D_\pi(\gamma P_\pi - I)\Phi y = \zeta y^T\Phi^T(\gamma P_\pi - I)\Phi y$, which implies that with probability one

$$\zeta = \frac{y^T\Phi^T(\gamma P_\pi - I)^T D_\pi(\gamma P_\pi - I)\Phi y}{y^T\Phi^T(\gamma P_\pi - I)\Phi y} < 0\ ,$$

since the numerator is positive definite and the denominator is negative definite with probability one. Therefore, the system $\dot x = g_\infty(x)$ has a unique globally asymptotically stable (GAS) equilibrium. Theorem 12 yields desired boundedness of the iterates, i.e., $\sup_t\|v_t\|<\infty$ with probability one. Finally, we apply Theorem 13 to establish convergence with probability one to the GAS of the ODE $\langle\dot v\rangle = \bar g(\langle v\rangle) = \Phi^T D_\pi^s(\gamma P_\pi - I)\Phi\langle v\rangle + \frac{1}{N}\sum_{i\in\mathcal{N}}\Phi^T D_\pi^s R_\pi^i$. □

*Proof.* (**Lemma 5**) Analogous to the proof of Lemma 4. □

*Proof.* (**Theorem 6**) We define a filtration $\mathcal{F}_t^\theta = \sigma(\theta_0, Y_{\tau-1}, \tau\le t)$, where $Y_t$ are the actor updates. The recursion of agent $i$, $i\in\mathcal{N}$, is given as

$$\theta_{t+1}^i = \Psi_\Theta^i(\theta_t^i + \alpha_{\theta,t}\cdot\delta_t^i\cdot\psi_t^i) \tag{29}$$

$$= \Psi_\Theta^i(\theta_t^i + \alpha_{\theta,t}\cdot[g_t^i(\theta_t^i) + \delta M_t + \beta_t]), \tag{30}$$

where $\delta_t^i = f_t^T\lambda_t^i + \gamma\phi_{t+1}^T v_t^i - \phi_t^T v_t^i$, $\psi_t^i = \nabla_{\theta^i}\log\pi^i(a_t^i|s_t^i;\theta_t^i)$, and the functions $g_t(\cdot)$, $\delta M_t$, $\beta_t$ are given as

$$g_t^i(\theta_t^i) = \mathbb{E}_{\pi_t,d_{\pi_t},p}(\delta_{t,\pi_t}\cdot\psi_t^i|\mathcal{F}_t^\theta) \tag{31}$$

$$\delta M_t = \delta_t^i\psi_t^i - \mathbb{E}_{\pi_t,d_{\pi_t},p}(\delta_t^i\cdot\psi_t^i|\mathcal{F}_t^\theta) \tag{32}$$

$$\beta_t = \mathbb{E}_{\pi_t,d_{\pi_t},p}((\delta_t^i - \delta_{t,\pi_t})\cdot\psi_t^i|\mathcal{F}_t^\theta). \tag{33}$$

The signal $\delta_{t,\pi_t}$ is the approximated team-average TD error upon convergence of the parameters $v_t$ and $\lambda_t$ under the current network policy $\pi(a|s;\theta_t)$, i.e., $\delta_{t,\pi_t} = f_t^T\lambda_{\pi_t} + \gamma\phi_{t+1}^T v_{\pi_t} - \phi_t^T v_{\pi_t}$. To complete the convergence proof, we verify the conditions given in Appendix A-2.

(1) The function $\delta_t^i$ is bounded by Assumption 1, and Lemma 4 and 5. The function $\psi_t^i$ is bounded by Assumption 5. Therefore, we obtain $\sup_t\mathbb{E}(\|\delta_t^i\cdot\psi_t^i\||\mathcal{F}_t^\theta)<\infty$.

(2) The step size sequence $\alpha_{\theta,t}$ satisfies $\sum_t\alpha_{\theta,t}^2<\infty$ and $\lim_t\frac{\alpha_{\theta,t+1}}{\alpha_{\theta,t}}=1$.

(3) The bias term satisfies $\beta_t\to0$ with probability one since $v_t\to v_{\pi_t}$ and $\lambda_t\to\lambda_{\pi_t}$ on the faster time scale by Assumption 4.

(4) The admissible set $\Theta$ is a hyperrectangle by Assumption 5.

(5) The function $g_t^i(\cdot)$ is continuous in $\theta_t^i$ uniformly in $t$. Furthermore, $g_t^i(\cdot) := \bar g^i(\cdot)$ since it is independent of $t$.

From Theorem 17, the asymptotic behavior of the actor updates is given by the ODE $\dot\theta^i = \Psi_{\Theta^i}[\bar g^i(\theta^i)]$. Now note that $\bar g^i(\theta^i) = \nabla_{\theta^i}\tilde J(\theta)$, where $\tilde J(\theta) = \mathbb{E}_{\pi,d_\pi,p}[\bar r(s,a;\lambda_\pi) + \gamma V(s';v_\pi) - V(s;v_\pi)]$. The rate of change of $\tilde J(\theta)$ is given as $\dot{\tilde J}(\theta) = \nabla_\theta\tilde J(\theta)^T(\nabla_\theta\tilde J(\theta)+z)$, where $z$ is the reflection term that projects the actor parameters back into the admissible set $\Theta$, i.e., $z=-\nabla_\theta\tilde J(\theta)$ whenever a constraint is active and $z=0$ otherwise (elementwise). Therefore, $\dot{\tilde J}(\theta)>0$ if $\nabla_\theta\tilde J(\theta)+z\ne0$ and $\dot{\tilde J}(\theta)=0$ otherwise. By Theorem 18, the solution of the ODE $\dot\theta = \Psi_\Theta[\bar g(\theta)] = \left[\Psi_{\Theta^1}[\bar g^1(\theta^1)]^T \quad \dots \quad \Psi_{\Theta^N}[\bar g^N(\theta^N)]^T\right]^T$ converges to a set of stationary points $\nabla_\theta\tilde J(\theta)+z=0$ that correspond to the stationary points of $\tilde J(\theta)$. □

*2) Algorithm 2:*

**Lemma 23.** *Under Assumption 7, the spectral radius* $\rho_{x,t}^+(C_{x,t}^{+T}(I-\mathbf{1}\mathbf{1}^T/N^+)C_{x,t}^+)<1$, *for* $x\in\{v,\lambda\}$.

*Proof.* The proof is analogous to the proof of Lemma 20. The difference here is that the communication graph after edge removal, $\mathcal{G}_t'$, remains connected for $t>0$. Furthermore, the communication subgraph of the cooperative agents is rooted under Assumption 7 [38]. Therefore, we conclude that $C_{x,t}^+$ has only one eigenvalue equal to one by Proposition 7. Using the same reasoning about the eigenvalues as in the proof of Lemma 20, we obtain $\|(I-\mathbf{1}\mathbf{1}^T/N^+)C_{x,t}^+x\|^2\le\rho_{x,t}^+\|x\|^2$ for all $x$ and some $\rho_{x,t}^+<1$. □

**Lemma 24.** *Under Assumption 1-4 and 7, the agents reach consensus on the critic parameters with probability one, i.e.,* $\lim_{t\to\infty} v^+_{\perp,t} = 0$. *Furthermore, the term* $\|(I \otimes \Gamma_{v,t})\alpha^{-1}_{v,t} v^+_{\perp,t}\|^2$ *is uniformly bounded.*

**Lemma 25.** *Under Assumption 1-4 and 7, the agents reach consensus on the team-average reward function parameters with probability one, i.e.,* $\lim_{t\to\infty} \lambda^+_{\perp,t} = 0$. *Furthermore, the term* $\|(I \otimes \Gamma_{\lambda,t})\alpha^{-1}_{\lambda,t}\lambda^+_{\perp,t}\|^2$ *is uniformly bounded.*

*Proof.* As the proofs of Lemma 24 and 25 are analogous, we only present the proof for the critic parameters $v^+_t$. The updates of $v^+_{\perp,t}$ are given as follows

$$\begin{aligned} v^+_{\perp,t+1} =& \mathcal{J}_\perp[(C^+_{v,t}\otimes\Gamma_{v,t})(v^+_{\perp,t} + \alpha_{v,t}(A^+_{v,t}v^+_{\perp,t} + b^+_{v,t}))] \\ &+ (I\otimes\hat{\Gamma}_{v,t})v^+_{\perp,t}. \end{aligned}$$

Splitting the updates into two orthogonal subspaces yields

$$\begin{aligned} (I\otimes\hat{\Gamma}_{v,t})v^+_{\perp,t+1} =& (I\otimes\hat{\Gamma}_{v,t})v^+_{\perp,t} \\ (I\otimes\Gamma_{v,t})v^+_{\perp,t+1} =& [(I - \mathbf{1}\mathbf{1}^T/N^+)C^+_{v,t}\otimes I](I\otimes\Gamma_{v,t}) \\ &\cdot (v^+_{\perp,t} + \alpha_{v,t}(A^+_{v,t}v^+_{\perp,t} + b^+_{v,t})) \end{aligned}$$

The first term equation implies $\|(I\otimes\hat{\Gamma}_{v,t})v^+_{\perp,t+1}\|^2 = \|(I\otimes\hat{\Gamma}_{v,t})v^+_{\perp,t}\|^2$ and for the second equation we write

$$\begin{aligned} &\|(I\otimes\Gamma_{v,t})v^+_{\perp,t+1}\|^2 \\ =&\|[(I - \mathbf{1}\mathbf{1}^T/N^+)C^+_{v,t}\otimes I](I\otimes\Gamma_{v,t}) \\ &\cdot (v^+_{\perp,t} + \alpha_{v,t}(A^+_{v,t}v^+_{\perp,t} + b^+_{v,t}))\|^2 \\ \le& \rho^+_{v,t}\cdot\|(I\otimes\Gamma_{v,t})((I + \alpha_{v,t}A^+_{v,t})v^+_{\perp,t} + \alpha_{v,t}b^+_{v,t}))\|^2, \end{aligned}$$

where $\rho^+_{v,t} < 1$ by Lemma 23. Following the steps in the proof of Lemma 21, we obtain $\sup_t \|(I\otimes\Gamma_{v,t})\alpha^{-1}_{v,t}v^+_{\perp,t}\|^2 < \infty$ and $\lim_{t\to\infty} v^+_{\perp,t} = 0$ with probability one. □

Lemma 24 and 25 ensure that the updates in the disagreement subspace become contractive with a decreasing step size and are subject only to a bounded disturbance that originates in the homogeneous rewards. Therefore, the trajectories of $v^+_{\perp,t}$ and $\lambda^+_{\perp,t}$ remain in a compact set for $t > 0$.

*Proof.* (**Lemma 8**) We let $\mathcal{F}^v_t = \sigma(v^+_0, Y_{\tau-1}, \xi_\tau, \tau \le t)$ denote a filtration, where $Y_\tau$ is a critic update and $\xi_\tau = (r^+_\tau, s_\tau, a_\tau, C^+_{v,\tau-1})$ is a collection of random variables. We write the iteration of $\langle v^+_t\rangle$ as follows

$$\begin{aligned} &\langle v^+_{t+1}\rangle \\ =&\left\langle (I\otimes\hat{\Gamma}_{v,t})v^+_t + (C^+_{v,t}\otimes\Gamma_{v,t})(v^+_t + \alpha_{v,t}(A^+_{v,t}v^+_t + b^+_{v,t}))\right\rangle \\ =&\langle v^+_t\rangle + \alpha_{v,t}A'_{v,t}\langle v^+_t\rangle \\ &+ \alpha_{v,t}\left\langle (C^+_{v,t}\otimes\Gamma_{v,t})(\alpha^{-1}_{v,t}v^+_{\perp,t} + A^+_{v,t}v^+_{\perp,t} + b^+_{v,t})\right\rangle \\ =&\langle v^+_t\rangle + \alpha_{v,t}[g_t(\langle v^+_t\rangle,\xi_t) + \delta M_t + \beta_t], \end{aligned}$$

where the functions $g_t(\cdot,\cdot)$, $\delta M_t$, and $\beta_t$ are given as

$$\begin{aligned} g_t(\langle v^+_t\rangle,\xi_t) =& \mathbb{E}(A'_{v,t}\langle v^+_t\rangle|\mathcal{F}^v_t) + \left\langle (C^+_{v,t}\otimes\Gamma_{v,t})b^+_{v,t}\right\rangle \\ &+ \left\langle (C^+_{v,t}\otimes\Gamma_{v,t})\alpha^{-1}_{v,t}v_{\perp,t}\right\rangle \quad (34) \\ \delta M_t =& A'_{v,t}\langle v^+_t\rangle - \mathbb{E}(A'_{v,t}\langle v^+_t\rangle|\mathcal{F}^v_t) \quad (35) \\ \beta_t =& \left\langle (C^+_{v,t}\otimes\Gamma_{v,t})A^+_{v,t}v^+_{\perp,t}\right\rangle. \quad (36) \end{aligned}$$

We now need to verify the conditions in Appendix A-1.

(1) We have $\|g_t(\langle x\rangle,\xi_t) - g_t(\langle y\rangle,\xi_t)\| = \|\mathbb{E}(A_{v,t}(\langle x\rangle - \langle y\rangle)|\mathcal{F}^v_t)\| \le K_1\cdot\|\langle x\rangle - \langle y\rangle\|^2$ for some $K_1 > 0$ since $A'_{v,t}$ is uniformly bounded by Assumption 1. Thus, $g_t(\langle v^+_t\rangle,\xi_t)$ is Lipschitz continuous in $\langle v^+_t\rangle$.

(2) The step size sequence $\{\alpha_{v,t}\}_{t\ge0}$ satisfies $\sum_t \alpha_{v,t} = \infty$ and $\sum_t \alpha^2_{v,t} < \infty$, for $t \ge 0$.

(3) The martingale difference sequence $\delta M_t$ satisfies $\mathbb{E}(\|\delta M_t\|^2|\mathcal{F}^v_t) \le K_2\cdot(1 + \|\langle v^+_t\rangle\|^2)$, since $A'_{v,t}$ is uniformly bounded by Assumption 1.

(4) By Lemma 24 and Assumption 1, the bias term $\beta_t$ is uniformly bounded and $\beta_t \to 0$ with probability one.

(5) We let $W(\langle v\rangle)$ denote a set of all occupation measures of the Markov chain $\{\xi_t\}_{t\ge0}$ for a fixed $\langle v\rangle$. The Markov chain $\{\xi_t\}_{t\ge0}$ is uniformly bounded since $r_t$, $s_t$, $a_t$, and $C^+_{v,t}$ are uniformly bounded.

By Theorem 15, the asymptotic behavior is described by the differential inclusion $\langle\dot{v}^+\rangle \in \Phi^T D^s_\pi(\gamma P_\pi - I)\Phi\langle v^+\rangle + \frac{1}{N^+}\sum_{i\in\mathcal{N}^+}\Phi^T D^s_\pi R^i_\pi + \Delta_v$, where

$$\begin{aligned} \|\Delta_v\| \le \lim_{t,m}\sup_{\xi_t\in W}\Bigg\| &\frac{1}{m}\sum_{k=t}^{t+m-1}[(\mathbf{1}^T C^+_{v,k}r^+_{k+1}\otimes\phi_k) \\ &+ (\mathbf{1}^T C^+_{v,k}\otimes\frac{\phi_k\phi_k^T}{\|\phi_k\|^2})\alpha^{-1}_{v,k}v^+_{\perp,k}] - \frac{1}{N^+}\sum_{i\in\mathcal{N}^+}\Phi^T D^s_\pi R^i_\pi\Bigg\|. \end{aligned}$$

Since the terms $b^+_{v,t}$ and $\alpha^{-1}_{v,t}v^+_{\perp,t}$, and consequently $\Delta_v$, are uniformly bounded by Assumption 1 and 3 and Lemma 24, we can apply Theorem 12 to establish boundedness of the critic updates. Finally, we apply Theorem 16 to establish that the team-average critic value $\langle v^+\rangle$ converges with probability one to a bounded neighborhood around the cooperative-team-average true minimizer $v^+_\pi$ that satisfies $\Phi^T D^s_\pi(\gamma P_\pi - I)\Phi v^+_\pi + \frac{1}{N^+}\sum_{i\in\mathcal{N}^+}\Phi^T D^s_\pi R^i_\pi = 0$. □

*Proof.* (**Lemma 9**) The proof is nearly identical to the proof of Lemma 8. We let $\mathcal{F}^\lambda_t = \sigma(\lambda_0, \xi_\tau, \tau \le t)$ denote a filtration, where $\xi_\tau = (r_\tau, s_\tau, a_\tau, C^+_{\lambda,\tau-1})$ is a collection of random variables. We write the updates in the form

$$\begin{aligned} \langle\lambda^+_{t+1}\rangle =& \Big\langle (I\otimes\hat{\Gamma}_{\lambda,t})\lambda^+_t \\ &+ (C^+_{\lambda,t}\otimes\Gamma_{\lambda,t})(\lambda^+_t + \alpha_{\lambda,t}(A^+_{\lambda,t}\lambda^+_t + b^+_{\lambda,t}))\Big\rangle \\ =& \langle\lambda^+_t\rangle + \alpha_{\lambda,t}[g_t(\langle\lambda^+_t\rangle,\xi_t) + \delta M_t + \beta_t], \end{aligned}$$

where the functions $g_t(\cdot,\cdot)$, $\delta M_t$, and $\beta_t$ are given as

$$\begin{aligned} g_t(\langle\lambda^+_t\rangle,\xi_t) =& \mathbb{E}(A'_{\lambda,t}\langle\lambda^+_t\rangle|\mathcal{F}^\lambda_t) + \left\langle (C^+_{\lambda,t}\otimes\Gamma_{\lambda,t})b^+_{\lambda,t}\right\rangle \\ &+ \left\langle (C^+_{\lambda,t}\otimes\Gamma_{\lambda,t})\alpha^{-1}_{\lambda,t}\lambda_{\perp,t}\right\rangle \quad (37) \\ \delta M_t =& A'_{\lambda,t}\langle\lambda^+_t\rangle - \mathbb{E}(A'_{\lambda,t}\langle\lambda^+_t\rangle|\mathcal{F}^\lambda_t) \quad (38) \\ \beta_t =& \left\langle (C^+_{\lambda,t}\otimes\Gamma_{\lambda,t})A^+_{v,t}\lambda^+_{\perp,t}\right\rangle. \quad (39) \end{aligned}$$

The conditions can be verified as in the proof of Lemma 8, which leads to the convergence of $\langle\lambda^+_t\rangle$ to a limit set

of the differential inclusion $\left\langle \dot{\lambda}^+ \right\rangle \in -F^T D_\pi^{s,a} F \langle \lambda^+ \rangle + \frac{1}{N^+} \sum_{i\in\mathcal{N}^+} F^T D_\pi^{s,a} R^i + \Delta_\lambda$, where

$$\|\Delta_\lambda\| = \lim_{t,m\to\infty} \sup_{\xi_t \in W} \left\| \frac{1}{m} \sum_{k=t}^{t+m-1} \frac{1}{N^+} \Big( (\mathbf{1}^T C^+_{\lambda,k} r^+_{k+1} \otimes f_k) + (\mathbf{1}^T C^+_{\lambda,k} \otimes \frac{f_k f_k^T}{\|f_k\|^2}) \alpha^{-1}_{\lambda,t} \lambda^+_{\perp,k} \Big) - \frac{1}{N^+} \sum_{i\in\mathcal{N}^+} F^T D_\pi^{s,a} R^i \right\|.$$

Hence, the team-average value of the team-average reward function parameter, $\langle \lambda^+ \rangle$, converges with probability one to a bounded neighborhood around the desired minimizer $\lambda^+_\pi$ that satisfies $F^T D_\pi^{s,a} (\frac{1}{N^+} \sum_{i\in\mathcal{N}^+} R^i - F\lambda^+_\pi) = 0$. □

*Proof.* (**Theorem 10**) We define a filtration $\mathcal{F}^\theta_t = \sigma(\theta^i_\tau, \tau \le t)$. The actor updates of agent $i$, $i \in \mathcal{N}^+$, are given as

$$\theta^i_{t+1} = \Psi_{\Theta^i}(\theta^i_t + \alpha_{\theta,t} \cdot \delta^i_t \cdot \psi^i_t) \tag{40}$$
$$= \Psi_{\Theta^i}(\theta^i_t + \alpha_{\theta,t} \cdot [g_t(\theta^i_t) + \delta M_t]), \tag{41}$$

where $\delta^i_t = f_t^T \lambda^i_t + \gamma \phi^T_{t+1} v^i_t - \phi^T_t v^i_t$, $\psi^i_t = \nabla_{\theta^i} \log \pi^i(a^i_t | s_t; \theta^i_t)$, and the function $g_t(\cdot)$ and martingale difference $\delta M_t$ are given as

$$g_t(\theta^i_t) = \mathbb{E}_{\pi_t, d_{\pi_t}, p}(\delta_{t,\pi_t} \cdot \psi^i_t | \mathcal{F}^\theta_t) + \mathbb{E}_{\pi_t, d_{\pi_t}, p}((\delta^i_t - \delta_{t,\pi_t}) \cdot \psi^i_t | \mathcal{F}^\theta_t) \tag{42}$$
$$\delta M_t = \delta^i_t \cdot \psi^i_t - \mathbb{E}_{\pi_t, d_{\pi_t}, p}(\delta^i_t \cdot \psi^i_t | \mathcal{F}^\theta_t). \tag{43}$$

The signal $\delta_{t,\pi_t}$ is the approximated network TD error under the current network policy $\pi(a|s;\theta_t)$ evaluated at $v^+_{\pi_t}$ and $\lambda^+_{\pi_t}$, i.e., $\delta_{t,\pi_t} = f_t^T \lambda^+_{\pi_t} + \gamma \phi_{t+1} v^+_{\pi_t} - \phi^T_t v^+_{\pi_t}$. To complete the proof, we verify the conditions in Appendix A-2.

(1) The function $\delta^i_t$ is bounded by Assumption 1 and Lemma 8 and 9. The function $\psi^i_t$ is bounded by Assumption 5. Therefore, we obtain $\sup_t \mathbb{E}(\|\delta^i_t \cdot \psi^i_t\| | \mathcal{F}^\theta_t) < \infty$.
(2) The step size sequence $\alpha_{\theta,t}$ satisfies $\sum_t \alpha^2_{\theta,t} < \infty$ and $\lim_{t\to\infty} \frac{\alpha_{\theta,t+1}}{\alpha_{\theta,t}} = 1$.
(3) The bias term satisfies $\beta_t = 0$ with probability one.
(4) The admissible set $\Theta$ is a hyperrectangle by Assumption 5.
(5) The function $g^i_t(\theta^i)$ is continuous in $\theta^i$ uniformly in $t$; thus, a set-valued function $G(\theta^i) = \{\lim_{n,m\to\infty} \frac{1}{m} \sum_{t=n}^{n+m-1} g^i_t(\theta^i)\}$ is upper semicontinuous.

Applying Theorem 19, the asymptotic behavior of the actor updates is given by the differential inclusion $\dot{\theta}^i \in \Psi_{\Theta^i}[G^i(\theta^i)]$. We let $\tilde{J}^+(\theta^+, \pi^-) = \mathbb{E}_{\pi, d_\pi, p}[\bar{r}(s,a;\lambda^+_\pi) + \gamma V(s_{t+1}; v^+_\pi) - V(s_t; v^+_\pi)]$ denote the approximated team-average objective function. Since $v^+_\pi$ and $\lambda^+_\pi$ are continuously differentiable in $\theta^i$ and Assumption 2 ensures differentiability of $\nabla_{\theta^i} \log \pi^i(a^i_t | s_t; \theta^i)$, $\tilde{J}^+(\theta^+, \pi^-)$ is continuously differentiable in $\theta^i$ with the associated local AC policy gradient

$$\nabla_{\theta^i} \tilde{J}^+(\theta^+, \pi^-) = \mathbb{E}_{\pi, d_\pi, p}[(\bar{r}(s,a;\lambda^+_\pi) + \gamma V(s'; v^+_\pi) - V(s; v^+_\pi)) \cdot \nabla_{\theta^i} \log \pi^i(a^i | s; \theta^i)].$$

We note that $G(\theta^i) = \nabla_{\theta^i} \tilde{J}^+(\theta^+, \pi^-) + \varepsilon^i_t$, where $\varepsilon^i_t$ is a set-valued error due to the discrepancy $\mathbb{E}_{\pi, d_\pi, p}((\delta^i_t - \delta_{t,\pi_t}) \cdot \psi^i_t | \mathcal{F}^\theta_t)$. Using Assumption 8, the rate of change of $\tilde{J}^+(\theta^+, \pi^-)$ is given in terms of the cooperative agents

$$\dot{\tilde{J}}^+(\theta^+, \pi^-) = \sum_{i\in\mathcal{N}^+} \nabla_{\theta^i} \tilde{J}^+(\theta^+, \pi^-)^T \times (\nabla_{\theta^i} \tilde{J}^+(\theta^+, \pi^-) + \varepsilon^i_t + z^i_t).$$

Here, $z^i_t$ is the reflection term that projects the actor parameters back into the admissible set $\Theta^i$, i.e., $z^i_t = -\nabla_{\theta^i} \tilde{J}(\theta^+, \pi^-) - \varepsilon^i_t$ whenever a constraint is active and $z^i_t = 0$ otherwise (elementwise). Suppose that $\sum_{i\in\mathcal{N}^+} \|z^i_t + \varepsilon^i_t\|^2 \le \sum_{i\in\mathcal{N}^+} \|\nabla_{\theta^i} \tilde{J}(\theta^+, \pi^-)\|^2$ on a compact subset. By Cauchy-Schwartz inequality, $\dot{\tilde{J}}^+(\theta) \ge 0$ and the policies converge to a neighborhood of a stationary point of the cooperative team-average objective function provided that $\varepsilon^i$ are small. □

## REFERENCES


[1] M. Xu, J. Peng, B. Gupta, J. Kang, Z. Xiong, Z. Li, and A. A. Abd El-Latif, "Multi-agent federated reinforcement learning for secure incentive mechanism in intelligent cyber-physical systems," *IEEE Internet of Things Journal*, 2021.

[2] E. Yang and D. Gu, "Multiagent reinforcement learning for multi-robot systems: A survey," tech. rep, Tech. Rep., 2004.

[3] L. Kuyer, S. Whiteson, B. Bakker, and N. Vlassis, "Multiagent reinforcement learning for urban traffic control using coordination graphs," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 656–671.

[4] M. Hüttenrauch, A. Šošić, and G. Neumann, "Deep reinforcement learning for swarm systems," *Journal of Machine Learning Research*, vol. 20, no. 54, pp. 1–31, 2019.

[5] A. Charpentier, R. Elie, and C. Remlinger, "Reinforcement learning in economics and finance," *Computational Economics*, pp. 1–38, 2021.

[6] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[7] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2017, p. 6382–6393.

[8] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. N. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning." *Journal of Machine Learning Research*, vol. 21, pp. 178–1, 2020.

[9] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proc. International Conference on Machine Learning*, vol. 80, 2018, pp. 5872–5881.

[10] G. Qu, A. Wierman, and N. Li, "Scalable reinforcement learning of localized policies for multi-agent networked systems," *Proc. of Machine Learning Research*, vol. 1, p. 38, 2020.

[11] Y. Lin, K. Zhang, Z. Yang, Z. Wang, T. Başar, R. Sandhu, and J. Liu, "A communication-efficient multi-agent actor-critic algorithm for distributed reinforcement learning," in *IEEE 58th Conference on Decision and Control*. IEEE, 2019, pp. 5562–5567.

[12] W. Suttle, Z. Yang, K. Zhang, Z. Wang, T. Başar, and J. Liu, "A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1549–1554, 2020.

[13] D. Chen, Y. Li, and Q. Zhang, "Communication-efficient actor-critic methods for homogeneous markov games," *arXiv preprint arXiv:2202.09422*, 2022.

[14] C. Qu, S. Mannor, H. Xu, Y. Qi, L. Song, and J. Xiong, "Value propagation for decentralized networked deep multi-agent reinforcement learning," *arXiv preprint arXiv:1901.09326*, 2019.

[15] 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*, vol. 61, no. 7, pp. 1848–1862, 2013.

[16] H. J. LeBlanc, H. Zhang, X. Koutsoukos, and S. Sundaram, "Resilient asymptotic consensus in robust networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 4, pp. 766–781, 2013.

[17] M. Figura, K. C. Kosaraju, and V. Gupta, "Adversarial attacks in consensus-based multi-agent reinforcement learning," in *Proc. American Control Conference*, 2021, pp. 3050–3055.

[18] N. H. Vaidya, "Iterative byzantine vector consensus in incomplete graphs," in *Proc. International Conference on Distributed Computing and Networking*. Springer, 2014, pp. 14–28.

[19] C. Xie, O. Koyejo, and I. Gupta, "Generalized byzantine-tolerant sgd," *arXiv preprint arXiv:1802.10116*, 2018.

[20] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proc. ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, pp. 1–25, 2017.

[21] J. Tu, W. Liu, X. Mao, and X. Chen, "Variance reduced median-of-means estimator for byzantine-robust distributed inference," *Journal of Machine Learning Research*, vol. 22, no. 84, pp. 1–67, 2021.

[22] D. Alistarh, Z. Allen-Zhu, and J. Li, "Byzantine stochastic gradient descent," in *Proc. Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[23] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. International Conference on Neural Information Processing Systems*, 2017, pp. 118–128.

[24] T. Phan, L. Belzner, T. Gabor, A. Sedlmeier, F. Ritz, and C. Linnhoff-Popien, "Resilient multi-agent reinforcement learning with adversarial value decomposition," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 11 308–11 316.

[25] Y. Xie, S. Mou, and S. Sundaram, "Towards resilience for multi-agent QD-learning," *arXiv preprint arXiv:2104.03153*, 2021.

[26] Y. Lin, S. Gade, R. Sandhu, and J. Liu, "Toward resilient multi-agent actor-critic algorithms for distributed reinforcement learning," in *Proc. American Control Conference*, 2020, pp. 3953–3958.

[27] Z. Wu, H. Shen, T. Chen, and Q. Ling, "Byzantine-resilient decentralized TD learning with linear function approximation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 5040–5044.

[28] S. Zeng, T. Chen, A. Garcia, and M. Hong, "Learning to coordinate in multi-agent systems: A coordinated actor-critic algorithm and finite-time guarantees," in *Proc. Learning for Dynamics and Control Conference*, 2022, pp. 278–290.

[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[30] H. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003, vol. 35.

[31] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer, 2008, vol. 48.

[32] H. Zhang, E. Fata, and S. Sundaram, "A notion of robustness in complex networks," *IEEE Transactions on Control of Network Systems*, vol. 2, no. 3, pp. 310–320, 2015.

[33] P. Erdős and A. Rényi, "On the strength of connectedness of a random graph," *Acta Mathematica Hungarica*, vol. 12, no. 1, pp. 261–267, 1961.

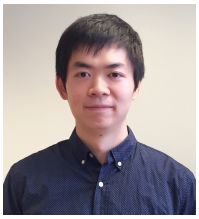
[34] S. Sundaram and B. Gharesifard, "Distributed optimization under adversarial nodes," *IEEE Transactions on Automatic Control*, vol. 64, no. 3, pp. 1063–1076, 2018.

[35] A. Ramaswamy and S. Bhatnagar, "A generalization of the borkar-meyn theorem for stochastic recursive inclusions," *Mathematics of Operations Research*, vol. 42, no. 3, pp. 648–661, 2017.

[36] H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems*. Springer Science & Business Media, 2012, vol. 26.

[37] R. A. Horn and C. R. Johnson, *Matrix analysis 2nd Edition*. Cambridge University Press, 2012.

[38] N. H. Vaidya, L. Tseng, and G. Liang, "Iterative approximate byzantine consensus in arbitrary directed graphs," in *Proc. ACM symposium on Principles of distributed computing*, 2012, pp. 365–374.

**Lintao Ye** is a Lecturer in the School of Artificial Intelligence and Automation at the Huazhong University of Science and Technology, Wuhan, China. He received his M.S. degree in Mechanical Engineering in 2017, and his Ph.D. degree in Electrical and Computer Engineering in 2020, both from Purdue University, IN, USA. He was a Postdoctoral Researcher at the University of Notre Dame, IN, USA. His research interests are in the areas of optimization algorithms, control theory, estimation theory, and network science.

**Martin Figura** received his B.S. degree in Mechanical Engineering Technology at South Carolina State University, Orangeburg, SC, USA, and his M.S. and Ph.D. in Electrical Engineering at University of Notre Dame, Notre Dame, IN, USA. His research interests include reinforcement learning, optimal control, and distributed systems.

**Yixuan Lin** received the B.S. degree in Mathematics from Fudan University, Shanghai, China, in 2017, and the M.S degree in Applied Mathematics and Statistics from Stony Brook University, Stony Brook, NY, USA, in 2020. She is currently pursuing her Ph.D. degree in Applied Mathematics and Statistics at Stony Brook University. Her research interests include reinforcement learning, optimization, and federated learning.

**Mainak Pal** received his B.E. (Hons.) in Electronics and Telecommunication Engineering from Jadavpur University, Kolkata, India. Following his graduation, he worked as a Data Scientist in the Analytics and Insights team at Tata Steel, India. He is currently pursuing his Ph.D. in Electrical and Computer Engineering at Purdue University, IN, USA. His research interests are in the areas of reinforcement learning and human-robot teaming.

**Pranoy Das** is currently pursuing his masters and PhD in the Elmore Family School of Electrical and Computer Engineering at Purdue University. He received his B. Sc (research) and M.Sc (research) degree at Indian Institute of Science, Bangalore all in Biological Sciences. His research interests include game theory, multi-agent learning in games and reinforcement learning.

**Ji Liu** received the B.S. degree in information engineering from Shanghai Jiao Tong University, Shanghai, China, in 2006, and the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, USA, in 2013. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at Stony Brook University, Stony Brook, NY, USA. His current research interests include distributed control and optimization, distributed reinforcement learning, epidemic networks, social networks, and cyber-physical systems.

**Vijay Gupta** is in the Elmore Family School of Electrical and Computer Engineering at the Purdue University. He received his B. Tech degree at Indian Institute of Technology, Delhi, and his M.S. and Ph.D. at California Institute of Technology, all in Electrical Engineering. He received the 2018 Antonio J Rubert Award from the IEEE Control Systems Society, the 2013 Donald P. Eckman Award from the American Automatic Control Council and a 2009 National Science Foundation (NSF) CAREER Award. His research interests are broadly at the interface of communication, control, distributed computation, and human decision making