Distributed off-Policy Actor-Critic Reinforcement Learning with Policy Consensus

Yan Zhang and Michael M. Zavlanos

Abstract— **In this paper, we propose a distributed off-policy actor critic method to solve multi-agent reinforcement learning problems. Specifically, we assume that all agents keep local estimates of the global optimal policy parameter and update their local value function estimates independently. Then, we introduce an additional consensus step to let all the agents asymptotically achieve agreement on the global optimal policy function. The convergence analysis of the proposed algorithm is provided and the effectiveness of the proposed algorithm is validated using a distributed resource allocation example. Compared to relevant distributed actor critic methods, here the agents do not share information about their local tasks, but instead they coordinate to estimate the global policy function.**

I. INTRODUCTION

Reinforcement learning (RL) algorithms have been widely used to solve decision making problems in unknown and stochastic environments [1,2]. Existing RL algorithms fall in two main categories, tabular-based methods and methods that use function approximation. Tabular-based methods are generally easier to analyze [3], however, they require the state and action spaces to be discrete and finite. On the other hand, using function approximation, such as Neural Networks [2], allows to solve RL problems in continuous state and action spaces. However, methods that rely on function approximation can be sensitive to approximation errors and can diverge in some cases [4]. Understanding convergence of RL algorithms with function approximation is an active area of research that is also the focus of this work.

Existing RL algorithms can also be classified as value-based methods [5,6] or policy gradient methods [7,8]. Value-based methods learn the value functions of states or state-action pairs and can only be used for control when the state-action space is discrete and of small size. Instead, policy gradient methods directly learn the policy function and are suitable for continuous action spaces. One of the most popular policy gradient methods is the Actor-Critic (AC) method [9], where the Critic evaluates the current policy and the Actor uses the feedback from the Critic to improve the policy function. In this paper, we are interested in distributed Actor-Critic methods. Specifically, we consider networks of agents that have their own tasks and their state dynamics are coupled with other agents' actions. The goal is to let the agents collaborate to learn a global optimal policy that maximizes the aggregate accumulated rewards over the network. Centralized AC methods [10,11] require the Critic to collect local rewards from all the agents, causing significant communication overhead and privacy issues. The works in [12–16] employ distributed optimization methods to evaluate a given fixed policy in multi-agents systems. However, these methods do not improve the policy parameters. The method in [17] lets homogeneous agents learn the same task and collect data independently. Instead, here we assume that the agents can have different tasks, different state and action spaces, and their behavior can affect each other.

Perhaps the most relevant work to the method proposed here is [18,19]. The key idea in these works is to let the agents estimate the global value function in a distributed way through a consensus mechanism. Compared to [18,19], here the agents keeps their own local value function estimates associated with their own tasks that they never share with their neighbors. Therefore, information about the local tasks is not revealed to other agents, as in [18,19]. Instead, the agents keep local estimates of the global policy function and a consensus step is introduced so that all agents agree on the optimal policy.

The rest of the paper is organized as follows. In Section II, we introduce the distributed reinforcment learning problem under consideration as well as some preliminary results. In Section III, we formulate the decentralized off-policy reinforcement learning problem and formally present our proposed distributed Actor-Critic algorithm. In Section IV, we analyze the convergence of the proposed algorithm. In Section V, we present a numerical example to validate our analysis. In Section VI, we conclude the paper.

II. PRELIMINARIES

A. The Reinforcement Learning Problem

Consider network of N agents. We define the state of the system $s = [s_1, s_2, \dots, s_N] \in \mathcal{S}$, where $s_i \in \mathcal{S}_i$ denotes the state of agent i. Moreover, we define the action of the system $a = [a_1, a_2, \dots, a_N] \in \mathcal{A}$, where $a_i \in \mathcal{A}_i$ denotes the action of agent i. The state space $\mathcal{S}_i$ and action space $\mathcal{A}_i$ are continuous. We denote the state transition function by $s(t+1) = T(s(t), a(t), \omega(t))$, where $\omega(t)$ represents the noise in the state dynamics at time t. We assume that the global state and action can be observed by all the agents. This is a common assumption in current RL literature, [10,11,18,19]. Let $r_i(s(t), a(t))$ denote the local reward received by agent i at time t. Define also the global deterministic policy function, $\pi(s) : \mathcal{S} \to \mathcal{A}$. Moreover, denote by $V^\pi(s) = \mathbb{E}_{\rho^\pi}[\sum_{t=0}^{\infty} \gamma^t \sum_{i=1}^{N} r_i(s(t), a(t)) | s(0) =$

Yan Zhang and Michael M. Zavlanos are with the Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA. {yan.zhang2,michael.zavlanos}@duke.edu This work is supported in part by AFOSR under award #FA9550-19-1-0169.

$s, \pi]$ the state value function and by $Q^\pi(s,a) = \mathbb{E}_{\rho^\pi}[\sum_{t=0}^\infty \gamma^t \sum_{i=1}^N r_i(s(t),a(t))|s(0) = s, a(0) = a, \pi]$ the state-action value function, where $\gamma \in (0,1)$ is the discounted factor. The goal is to find an optimal policy function π^* to maximize the infinite-time discounted-reward value function $J(\pi) = \mathbb{E}_{\rho^\pi}[V^\pi(s(0))]$, where ρ^π is the stationary state distribution under policy π.

B. Actor Critic Method

Parameterizing the policy function as $\pi_\theta(s) = \phi_\pi(s)^T\theta$, where $\theta \in \mathbb{R}^{n_\theta}$ is the policy parameter and $\phi_\pi(s) : \mathcal{S} \to \mathbb{R}^{n_\theta}$ is a vector of n_θ policy feature functions, the parameterized policy optimization problem

$$\max_\theta J(\theta) \tag{1}$$

can be solved using stochastic gradient descent methods. Specifically, the gradient $\nabla_\theta J$ is given in [8] by $\nabla_\theta J(\pi) = \mathbb{E}_{\rho^\pi}[\nabla_\theta \pi(s) \nabla_a Q^\pi(s,a)|_{a=\pi_\theta(s)}]$. Since the function $Q^\pi(s,a)$ under policy π_θ is unknown, policy evaluation algorithms are needed to approximate $Q^\pi(s,a)$ and furthermore the gradient $\nabla_\theta J(\pi)$. To do so, the function $Q^\pi(s,a)$ can be parameterized as $Q^\pi(s,a) = \phi_Q(s,a)^T w$, $w \in \mathbb{R}^{n_w}$, where $\phi_Q(s,a)$ is a vector of n_w feature functions. Then, to solve problem (1) we can use the following actor-critic algorithm consisting of two time scale updates

$$\begin{aligned} w(t+1) &= w(t) + \alpha_w(t)(h(w(t),\theta(t)) + M(t+1)), \\ \theta(t+1) &= \theta(t) + \alpha_\theta(t)(f(w(t),\theta(t)) + N(t+1)), \end{aligned} \tag{2}$$

where $\alpha_w(t), \alpha_\theta(t)$ are update rates at different time scales, $h(w(t),\theta(t))$ represents the Critic update to evaluate the current policy, e.g., [5,6], $f(w(t),\theta(t))$ represents the Actor update to improve the current policy, and $M(t+1)$ and $N(t+1)$ are sampling noises. Convergence of this scheme typically depends on analyzing the two time scale updates [20], which we explain in Section IV.

III. Problem formulation

Since the estimation of the function $Q^\pi(s,a)$ requires the global reward function $r(t) = \sum_{i=1}^N r_i(s(t),a(t))$, the actor-critic method in (2) is centralized. To design a decentralized algorithm, we first decompose the value and action value functions using linearity of the expectation as $V^\pi(s) = \sum_{i=1}^N V_i^\pi(s)$ and $Q^\pi(s,a) = \sum_{i=1}^N Q_i^\pi(s,a)$, where $V_i^\pi(s)$ is the local value function and $Q_i^\pi(s,a)$ is the local action value function under policy π. Then, problem (1) can be written as

$$\min_\theta \sum_{i=1}^N J_i(\theta), \tag{3}$$

where $J_i(\theta) = \mathbb{E}[V_i^\pi(s(0))|\rho(s(0)), \pi_\theta]$. A common approach to solve problem (3) in a distributed way is by introducing local estimates $\theta_i \in \mathbb{R}^{n_\theta}$ that are subject to the consensus update

$$\theta_i(t+1) = \sum_{j\in\mathcal{N}_i} W_{ij}(\theta_j(t) + \alpha_\theta(t)\nabla_\theta J_j(\theta)|_{\theta=\theta_j(t)}), \tag{4}$$

where the objective function $J_i(\theta)$ is usually nonlinear and the gradient $\nabla_\theta J_i$ is usually evaluated in a stochastic way. The convergence of (4) in this case is studied in [21]. The key idea in [21] is to evaluate the local gradient $\nabla_\theta J_j(\theta)|_{\theta=\theta_j(t)}$ in an on-policy fashion, for which all agents need to behave under policy $\pi_{\theta_j(t)}$. This suggests that to execute the update (4) at time t, every agent i needs to send its policy parameter $\theta_i(t)$ to all other agents and let all execute the policy $\pi_{\theta_i(t)}$ for multiple time steps so that agent i can collect local rewards to estimate $\nabla_\theta J_i(\theta)|_{\theta=\theta_i(t)}$. This on-policy scheme is impractical even though its convergence analysis is simpler, as seen in [21]. This motivates us to consider off-policy actor-critic methods as in [8,22]. The idea is to let all agents behave under a fixed policy, named behavorial policy $\beta(s)$, and optimize an approximate objective function, $J_\beta(\theta) = \mathbb{E}_{\rho^\beta}[V^\pi(s)]$, where the expection is taken over the stationary state distribution ρ^β instead of ρ^π as in (1). Similarly, the objective $J_\beta(\theta)$ can also be decomposed into local costs as

$$\min_\theta \sum_{i=1}^N J_{i,\beta}(\theta). \tag{5}$$

To achieve consensus on the local policy parameters θ_i, we can apply a similar update as in (4),

$$\theta_i(t+1) = \sum_{j\in\mathcal{N}_i} W_{ij}(\theta_j(t) + \alpha_\theta(t)\nabla_\theta J_{j,\beta}(\theta)|_{\theta=\theta_j(t)}). \tag{6}$$

However, as discussed in [8], the gradient $\nabla J_{j,\beta}(\theta)|_{\theta=\theta_j(t)}$ cannot be exactly estimated in this off-policy setting, therefore, it is replaced with an approximate gradient

$$\hat{\nabla} J_{j,\beta}(\theta) = \mathbb{E}_{\rho^\beta}[\nabla_\theta \pi(s) \nabla_a Q_j^\pi(s,a)|_{a=\pi_\theta(s)}]. \tag{7}$$

Convergence of the centralized Actor Critic method using the off-policy gradient in (7) is studied in [22]. However, the convergence of its decentralized counterpart (6) is unknown, which is the focus of this paper.

In practice, we compute the gradient $\nabla_a Q_j^\pi(s,a)$ in (7) using the parameterization $Q_i^\pi(s,a) \approx \sum_{p=1}^{n_w} \phi_{w_i}(s,a)^T w_i$. To ensure that this parameterization preserves the update (7) that uses the true state-action value function $Q_i^\pi(s,a)$, as discussed in [8], we make the following assumption:

Assumption III.1. (Function Compatibility) All the local value functions $Q_i^\pi(s,a)$ are parameterized as $Q_i^\pi(s,a) = (a - \pi(s))^T(\nabla_\theta \pi(s))^T w_i$. That is, the feature function for $Q_i^\pi(s,a)$ satisfies that $\phi_{w_i}(s,a) = \nabla_\theta \pi(s)(a - \pi(s))$.

Given Assumption III.1 and the expression for the off-policy gradient (7), the consensus update (6) of the local policy parameters θ_i becomes

$$\theta_i(t) = \sum_{j\in\mathcal{N}_i} W_{ij}(\theta_i(t-1) + \alpha_\theta \nabla\pi(s(t))_\theta \nabla_\theta \pi(s(t))^T w_i(t)). \tag{8}$$

To implement the update in (8), we have to compute $w_i(t)$ in an off-policy way. In this paper, we employ the gradient temporal difference (GTD) learning algorithm studied in [5,6,8] to estimate the local value function parameters w_i

Algorithm 1 Distributed Actor Critic with policy consensus

```
Require: Initial value function parameters $\{w_i(0)\}$ and policy parameters $\{\theta_i(0)\}$. Step sizes $\alpha_w(0)$ and $\alpha_\theta(0)$. Set $t = 0$. Maximum time limit $T$. Discount factor $\gamma$. Fixed behavorial policy function $\beta(s)$. Agents' inital state $s(0)$.
1: for $0 \leq t \leq T$ do
2:     All agents take actions according to $\beta(s(t))$ and observe the actions $a(t)$, next states $s(t+1)$ and the local rewards $r_i(t)$ from the environment.
3:     Every agent $i$ updates its local value function estimate $w_i$ based on the observed transition $[s(t), a(t), s(t+1)]$ and local reward $r_i(t)$ using the GTD algorithm in [6].
4:     Every agent $i$ updates its local policy parameter $\theta_i(t)$ according to equation (8).
5: end for
```

in (8), because this method is known to be stable in the off-policy setting. The proposed distributed Actor Critic method is presented in Algorithm 1.

IV. Convergence Analysis

In this section, we analyze the convergence of Algorithm 1 using the two-time scale technique in [20]. The key idea is that the Critic updates at a faster rate than the Actor so that to analyze the convergence of the Critic update in line 3 in Algorithm 1, we can assume that each local policy parameter θ_i is fixed. Then, each local Critic can independently estimate its own local value function and the convergence analysis of this local policy evaluation is the same as GTD in [6]. To analyze the convergence of the Actor, we can assume that every local Critic has already converged to the correct value function estimate. We first introduce several assumptions that are common in the reinforcement learning literature.

Assumption IV.1. We assume that the behavioral policy $\beta(s)$ is stationary, and the Markov chain that governs the state $s(t)$ under policy $\beta(s)$ is irreducible and aperiodic.

The above assumption ensures that when the agents behave under policy $\beta(s)$, the system states will reach the stationary state distribution ρ^β.

Assumption IV.2. We assume that for all i, the reward $|r_i(s, a)|$ is uniformly bounded for all s and a.

This assumption ensures that the objective function in (5) is upper bounded when the discount factor satisfies $0 < \gamma < 1$, so that the problem (5) is well defined.

Assumption IV.3. We assume that the stepsizes $\alpha_w(t)$ and $\alpha_\theta(t)$ are deterministic and satisfy that $\sum_{t=0}^{\infty} \alpha_w(t) \to \infty$, $\sum_{t=0}^{\infty} \alpha_\theta(t) \to \infty$, $\sum_{t=0}^{\infty} \alpha_w(t)^2 < \infty$ and $\sum_{t=0}^{\infty} \alpha_\theta(t)^2 < \infty$. Moreover, $\frac{\alpha_\theta(t)}{\alpha_w(t)} \to 0$.

This assumption is standard in the literature employing two-time scale analysis, [18–20]. Furthermore, let $W_t \in \mathbb{R}^{N \times N}$ be a random weight matrix of the communication graph at time t. Define the filtration $\mathcal{F}_t$ to be a $\sigma-$algebra $\sigma(\{\theta_i(0)\}, s(\tau), \{r_i(\tau)\}, W(\tau), \tau \leq t)$. Then, we have the following assumptions on W_t.

Assumption IV.4. We assume that W_t satisfies the following conditions: (a) W_t is row stochastic and $\mathbb{E}[W_t]$ is column stochastic for all $t > 0$. That is, $W_t\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T\mathbb{E}[W_t] = \mathbf{1}^T$; (b) The spectral norm satisfies $\mathbb{E}[W_t^T(I - \frac{1}{N}\mathbf{1}\mathbf{1}^T)W_t] = \rho_W < 1$; (c) W_t and (s_t, r_t) are conditionally independent given the filtration $\mathcal{F}_{t-1}$.

The above assumptions are standard in the stochastic consensus optimization literature [21].

Assumption IV.5. We assume the vector of feature functions $\phi_\pi(s)$ is uniformly bounded for all s.

This assumption is common and essential to show stability of RL algorithms, see [5,6,18,19].

Assumption IV.6. We assume that through the whole history of the algorithm, $\{\theta_i(t)\}$ belongs to a compact set for all i and t. We also assume that this compact set contains at least one local maximum of the problem (5).

This assumption is necessary to show stability of the policy parameter updates. Moreover, boundedness of the policy parameters is commonly observed in practice when implementing RL algorithms as mentioned in [22]. Let $\chi_i(\theta)$ denote a function that maps the policy parameter θ to the optimal value function parameter by means of the policy evaluation algorithm [6]. Then, we can show the following result for the function $\chi_i(\theta)$.

Lemma IV.7. Let policy parameter θ_i at agent i be fixed, and let agent i run the GTD algorithm in [6] to evaluate this policy. Then, the local value function parameter w_i converges to $\chi_i(\theta)$ almost surely (a.s.). Moreover, the function $\chi_i(\theta)$ is Lipschitz continous.

Proof. Since each local policy evaluation is performed independently, Lemma IV.7 can be directly shown using Lemmas 4 and 5 in [22]. □

Next, we have the following result on the stability of the value function parameter w_i.

Lemma IV.8. Given Assumptions IV.2 and IV.6, the value function parameter w_i is a.s. uniformly bounded over time.

Proof. According to Lemma IV.7, we have that w_i converges to $\chi_i(\theta)$ a.s. Then, the boundness of w_i is given by combining Assumption IV.6 and the fact that the function $\chi_i(\theta)$ is $L_\chi-$Lipschitz continuous. □

In what follows, we stack all policy function parameters in a vector $\theta(t) = [\theta_i(t)^T, \ldots, \theta_N(t)^T]^T$. The update of $\theta(t)$ (8) can be compactly written as

$$\theta(t+1) = (W_t \otimes I)(\theta(t) + \alpha_\theta(t)\hat{\nabla}J(\theta(t))), \qquad (9)$$

where $\hat{\nabla}J(\theta(t)) = \begin{bmatrix} \vdots \\ \phi_\pi(s_{t+1})(\phi_\pi(s_{t+1})^T\chi_i(\theta_i(t))) \\ \vdots \end{bmatrix}$ and

I is an identity matrix. The expression of $\hat{\nabla} J(\theta(t))$ is due to Assumption III.1. Moreover, define the disagreement between local policy parameters as $\theta_\perp(t) = \theta(t) - \mathbf{1} \otimes \bar{\theta}(t)$, where $\mathbf{1}$ is a vector of N 1's, and $\bar{\theta}(t) = \frac{1}{N}\sum_i \theta_i(t)$. We have the following lemma.

Lemma IV.9. Given Assumptions IV.3, IV.4, IV.5 and IV.6, we have that $\sum_t \mathbb{E}[\|\theta_\perp(t)\|^2] < \infty$. Therefore, $\theta_\perp(t) \to 0$ a.s.

Proof. First, we establish the dynamcis of $\theta_\perp(t)$. To achieve this, we introduce an operator $J_\perp := (I - \frac{1}{N}\mathbf{1}\mathbf{1}^T) \otimes I$. Then, multiplying both sides of (9) with $J_\perp$, and replacing $\theta(t)$ with $\mathbf{1} \otimes \bar{\theta}(t) + \theta_\perp(t)$, we have that $\theta_\perp(t) = J_\perp(W_t \otimes I)(\mathbf{1} \otimes \bar{\theta}(t-1) + \theta_\perp(t-1) + \alpha_\theta(t)\hat{\nabla} J(\theta(t-1)))$. Using Assumption IV.4(a), we have that $J_\perp(W_t \otimes I)(\mathbf{1} \otimes \bar{\theta}(t)) = 0$ for all t. Therefore, we obtain $\theta_\perp(t) = J_\perp(W_t \otimes I)(\theta_\perp(t-1) + \alpha_\theta(t)\hat{\nabla} J(\theta(t-1)))$. Taking the square of the Euclidean norm on both sides of the above equation, we have that

$$\|\theta_\perp(t)\|^2 = \|\theta_\perp(t-1) + \alpha_\theta(t)\hat{\nabla} J(\theta(t-1))\|^2_{(W_t^T(I - \frac{1}{N}\mathbf{1}\mathbf{1}^T)W_t)},$$

where $\|v\|_M^2 := v^T M v$ for any vector v and matrix M. According to Assumption IV.4(b,c), taking expectation over the random matrix W_t, given the filtration $\mathcal{F}_{t-1}$ and the random sample $(s(t), r(t))$, we have that

$$\begin{aligned}
&\mathbb{E}[\|\theta_\perp(t)\|^2 | \mathcal{F}_{t-1}, s(t), r(t)] \leq \\
&\qquad \rho_W \|\theta_\perp(t-1) + \alpha_\theta(t)\hat{\nabla} J(\theta(t-1))\|^2 \\
&\leq \rho_W(\|\theta_\perp(t-1)\|^2 + 2\alpha_\theta(t)\|\theta_\perp(t-1)\|\|\hat{\nabla} J(\theta(t-1)\| \\
&\qquad + \alpha_\theta(t)^2\|\hat{\nabla} J(\theta(t-1))\|^2),
\end{aligned}$$

where the second inequality above is by expanding the two norm and using the Cauchy-Swartz inequality. Taking the expectation of both sides of above inequality and using Jensen's inequality, we have that

$$\begin{aligned}
&\mathbb{E}[\|\theta_\perp(t)\|^2] \leq \rho_W \mathbb{E}[\|\theta_\perp(t-1)\|^2] \\
&\quad + 2\rho_W \alpha_\theta(t)\sqrt{\mathbb{E}[\|\theta_\perp(t-1)\|^2 \|\hat{\nabla} J(\theta(t-1)\|^2]} \\
&\quad + \rho_W \alpha_\theta(t)^2 \mathbb{E}[\hat{\nabla} J(\theta(t-1)\|^2].
\end{aligned} \tag{10}$$

Recalling the expression for $\hat{\nabla} J(\theta(t-1)$ under (9), and using Assumption IV.5 and Lemma IV.8, we have that $\|\hat{\nabla} J(\theta(t-1)\|$ can be bounded by a constant K_1 for all t. Denote $v(t) = \mathbb{E}[\|\theta_\perp(t)\|^2]$. Then, recalling that $\rho_W < 1$ due to Assumption IV.4(b), (10) can be written as

$$v(t) \leq \rho_W v(t-1) + 2\alpha_\theta(t)K_1 v(t-1) + \alpha_\theta(t)^2 K_1^2.$$

The above inequality is the same as (17) in [21]. Since Assumptions IV.4 and IV.3 imply that Assumptions 1 and 2 in [21] are also satisfied, we can use the same proof as in Lemma 1 in [21] to show that $\mathbb{E}[\|\theta_\perp(t)\|^2]$ satisfies that $\mathbb{E}[\|\theta_\perp(t)\|^2] < \infty$ and $\theta_\perp(t) \to 0$ a.s.. □

Since $\theta_\perp(t) \to 0$ a.s., it is sufficient to study the dynamics of $\bar{\theta}(t)$. In what followss, we show that with the policy update in (9), $\bar{\theta}(t)$ asymptotically approaches the following ODE dynamics

$$\dot{\bar{\theta}} = F(\bar{\theta}), \tag{11}$$

where

$$F(\bar{\theta}) = \mathbb{E}[\frac{1}{N}\phi_\pi(s)\phi_\pi(s)^T \sum_i \chi_i(\bar{\theta})]. \tag{12}$$

Note that it is standard to study the discrete-time dynamics (8) by relating them to the behavior of the ODE in (11); see, e.g., relevant literature on RL [18,19,22] and stochastic optimization [21], as well as Chapter 2 in [20] that establishes conditions on the step sizes and noise terms in the discrete-time dynamics so that they asymptotically approach their continuous-time ODE counterpart.

To show that the discrete-time dynamics of $\bar{\theta}(t)$ asymptotically approach the ODE (11), we first multiply both sides of (9) with $\frac{1}{N}\mathbf{1}^T \otimes I$ on the left to obtain

$$\begin{aligned}
\bar{\theta}(t+1) &= \frac{1}{N}(\mathbf{1}^T \otimes I)(\theta(t) + \alpha_\theta(t+1)\hat{\nabla} J(\theta(t))) \\
&= \bar{\theta}(t) + \alpha_\theta(t+1)\frac{1}{N}(\phi_\pi(s_{t+1})\phi_\pi(s_{t+1})^T)\sum_i \chi_i(\theta_i(t))
\end{aligned}$$

Since $\sum_i \chi_i(\theta_i(t)) = \sum_i(\chi_i(\theta_i(t)) - \chi_i(\bar{\theta}_i(t)) + \chi_i(\bar{\theta}_i(t)))$, the above update of $\bar{\theta}(t)$ can be written in the following form

$$\begin{aligned}
\bar{\theta}(t+1) = \bar{\theta}(t) &+ \alpha_\theta(t+1)F(\bar{\theta}(t)) \\
&+ \alpha_\theta(t+1)\xi(t) + \alpha_\theta(t+1)r(t),
\end{aligned} \tag{13}$$

where we have

$$F(\bar{\theta}_t) = \mathbb{E}[\frac{1}{N}\phi_\pi(s_{t+1})\phi_\pi(s_{t+1})^T \sum_i \chi_i(\bar{\theta}(t))], \tag{14a}$$

$$\begin{aligned}
\xi(t) = &\frac{1}{N}\phi_\pi(s_{t+1})\phi_\pi(s_{t+1})^T \sum_i \chi_i(\bar{\theta}(t)) \\
&- \mathbb{E}[\frac{1}{N}\phi_\pi(s_{t+1})\phi_\pi(s_{t+1})^T \sum_i \chi_i(\bar{\theta}(t))].
\end{aligned} \tag{14b}$$

$$r(t) = \frac{1}{N}\phi_\pi(s_{t+1})\phi_\pi(s_{t+1})^T \sum_i(\chi_i(\theta_i(t)) - \chi_i(\bar{\theta}(t))). \tag{14c}$$

Then, to show that the discrete-time trajectory in (13) approaches the continuous trajectory of (11), we need to define the following functions generated by these trajectories; cf. Chapter 2.1 in [20]. First, let $\bar{x}(n)$ denote a continuous piecewise linear function that passes through the discrete-time updates in (13), so that $\bar{x}(n(t)) = \bar{\theta}(t)$ for $t \geq 0$ and $\bar{x}(n) = \bar{x}(n(t)) + \frac{\bar{x}(n(t+1)) - \bar{x}(n(t))}{n(t+1) - n(t)}(n - n(t))$ for $n(t) < n < n(t+1)$, where $n(0) = 0$, $n(t) = \sum_{m=0}^{t-1}\alpha_\theta(m)$ and n denotes the continuous time index. Moreover, define the function $x^s(n)$ that is the unique solution of the dynamical equation (11) for $n \geq s$ with initial condition $x^s(s) = \bar{\theta}(s)$, and the function $x_s(n)$ that is the unique solution of (11) for $n \leq s$ with the ending condition $x_s(s) = \bar{\theta}(s)$. Then, we can show the following result.

Lemma IV.10. Given Assumptions IV.3, IV.4, IV.5 and IV.6, we have that for any $T > 0$,

$$\begin{aligned}
&\lim_{s \to \infty} \sup_{n \in [s, s+T]} \|\bar{x}(n) - x^s(n)\| = 0, \text{ a.s.} \\
&\lim_{s \to \infty} \sup_{n \in [s, s-T]} \|\bar{x}(n) - x_s(n)\| = 0, \text{ a.s..}
\end{aligned}$$

Proof. According to Lemma 1 in Chapter 2 [20], it is sufficient to show that the following conditions are satisfied: (i) the function $F(\bar{\theta}_t)$ in (14a) is Lipschtiz continuous, (ii) $\xi(t)$ in (14b) satisfies that $\mathbb{E}[\xi(t)|\mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\|\xi(t)\|^2|\mathcal{F}_{t-1}] \leq K_2(1+\|\bar{\theta}_t\|^2)$ a.s. for some constant $K_2 > 0$, and (iii) that $\|r(t)\| \to 0$ a.s.. These conditions can be found in Lemma 1 and its extension in [20].

Combining Assumption IV.5 and Lemma IV.7, and recalling the definition in (14a), condition (i) is satisfied. In addition, from the construction of $\xi(t)$, it is simple to see that $\mathbb{E}[\xi(t)|\mathcal{F}_{t-1}] = 0$. Particularly, since $\phi_\pi(s)$ is bounded for all s and $\chi_i(\bar{\theta}(t))$ is also bounded according to Assumption IV.5 and Lemma IV.7, we have that $\mathbb{E}[\|\xi(t)\|^2|\mathcal{F}_{t-1}]$ is uniformly bounded for all t. Therefore, the constant K_2 in condition (ii) must exist and this condition is satisfied. Finally, by the Lipschitz property of the function $\chi_i(\theta)$ shown in Lemma IV.7, we have that $\|r(t)\| \leq \frac{L_\chi}{\sqrt{N}}\|\phi_\pi(s_{t+1})\phi_\pi(s_{t+1})^T\|\|\theta_\perp(t)\|$. Due to boundness of $\phi_\pi(s)$ and Lemma IV.9, we have that $\|r(t)\| \to 0$ a.s.. Therefore, condition (iii) is also satisfied. By conditions (i-iii), Assumption IV.3 and Lemma 1 in [20], the proof is complete. □

Before we state our main result, define the set $\Lambda = \{\bar{\theta} : (\mathbf{1}^T \otimes I)\hat{\nabla} J(\mathbf{1}\otimes\bar{\theta}) = 0\}$ and make the following assumption.

Assumption IV.11. We assume that set Λ is compact. Meanwhile, the set $\sum_{i=1}^N J_{i,\beta}(\Lambda)$ has an empty interior.

This assumption is satisfied when the objective function $J_\beta(\theta)$ is smooth, according to Sard's theorem. It is a common assumption in the stochastic approximation and optimization literature, e.g., [21,23].

Theorem IV.12. Given Assumptions III.1 and from IV.1 to IV.11, $\theta_i(t)$ converges to the set Λ a.s. for all i.

Proof. Using Lemma IV.9, we need to show that $\bar{\theta}(t)$ given by (13) converges to the set Λ. Moreover, using Lemma IV.10, we need to show that the dynamics (11) converge to the set Λ. To achieve this, define function $f(\bar{\theta}) = -J_\beta(\bar{\theta})$. We shall prove that the function $f(\bar{\theta})$ can serve as a Lyapunov function to show stability of the set Λ under dynamics (11). For this, we need to show that $\dot{f}(\bar{\theta}(n)) \leq 0$ for any solution $\bar{\theta}(n)$ of the ODE in (11), where n is the continuous time index, and that the inequality is strict for any $\bar{\theta} \notin \Lambda$.

If the function $F(\bar{\theta})$ is the gradient of the function $f(\bar{\theta})$, then we can directly use Proposition 4 in [21] to get the desired result. However, due to the proposed off-policy framework as we discussed in Section III, $F(\bar{\theta}_t)$ is only an approximation of the exact gradient. In what follows, we show that $F(\bar{\theta})$ behaves in a similar way as the exact gradient. Recalling the off-policy policy gradient in (7) and the parameterization scheme in Assumption III.1, it is simple to check that $F(\bar{\theta})$ in (12) equals the off-policy gradient $\hat{\nabla} J_\beta(\bar{\theta})$. According to the policy improvement theorem (Theorem 1) in [22], we have that there exists $\epsilon > 0$, such that for all positive $\alpha_\theta < \epsilon$ and $\bar{\theta}' = \bar{\theta} + \alpha_\theta \hat{\nabla} J_\beta(\bar{\theta})$, we have that $J_\beta(\bar{\theta}') \geq J_\beta(\bar{\theta})$. And for all $\bar{\theta} \notin \Lambda$, the above inequality is strict. Considering the first order Taylor expansion of the value function $J_\beta(\bar{\theta}') = J_\beta(\bar{\theta}) + \langle \nabla J_\beta(\bar{\theta}), \bar{\theta}' - \bar{\theta}\rangle + o(\|\bar{\theta}' - \bar{\theta}\|)$, when α_θ goes to 0, the term $\langle \nabla J_\beta(\bar{\theta}), \bar{\theta}' - \bar{\theta}\rangle$ dominates $o(\|\bar{\theta}' - \bar{\theta}\|)$. Therefore, we have that $\langle \nabla J_\beta(\bar{\theta}), \hat{\nabla} J_\beta(\bar{\theta})\rangle \geq 0$ for all $\bar{\theta}$. And if $\bar{\theta} \notin \Lambda$, we have that $\langle \nabla J_\beta(\bar{\theta}), \hat{\nabla} J_\beta(\bar{\theta})\rangle > 0$.

Since $\dot{f}(\bar{\theta}) = \langle -\nabla J_\beta(\bar{\theta}), \dot{\bar{\theta}}\rangle$ and $\dot{\bar{\theta}} = \hat{\nabla}_\beta J(\bar{\theta})$, we have that $\dot{f}(\bar{\theta}) = -\langle \nabla J_\beta(\bar{\theta}), \hat{\nabla} J_\beta(\bar{\theta})\rangle < 0$ when $\bar{\theta} \notin \Lambda$. We conclude that $f(\bar{\theta})$ is a valid Lyapunov function that can be used to show stability of the set Λ under dynamics (11). Finally, given Assumption IV.11, combining Lemma IV.10 and Theorem 2 in [21], we have that $\bar{\theta}(t)$ converges to the set Λ a.s.. Recalling that $\|\theta_\perp\| \to 0$ a.s., we obtain the desired result. The proof is complete. □

To conclude, we make the following remark on the two-time scale analysis we have employed in this paper that is described at the end of Chapter 6.1 [20]. Specifically, in this scheme, the local critic updates $w_i(t)$ in the fast time scale and the local actor keeps its local policy estimate fixed until the local critic has taken N_t update steps. Then, according to Lemma IV.7, when N_t is chosen large enough, the local critic converges. Therefore, the convergence of the local actors can be analyzed using the two-time scale approach we have employed in this section.

V. Numerical Simulation

In this section, we illustrate our proposed algorithm using a distributed resource allocation example. Specifically, we consider 6 resource dispatch centers in an area of interest. These centers make decisions as to how to allocate available resources amongst each other. We define the state of each center i at time t by $s_i(t)$ that captures the quantity of available resources $m_i(t)$ and the local demands $d_i(t)$. We also define the local action as $a_i(t) = \{a_{ij}(t)\}_{j\in\mathcal{N}_i}$, which denotes the amount of resources sent from center i to its neighbor j at time t. In this simulation, we assume the 6 centers are located in a 2×3 grid. Each center communicates with its $1-$hop neighboring centers located at its up/down/left/right direction. We assume the demand at each agent is given by $d_i(t) = A_i \sin(\omega_i t + \phi_i) + w_i(t)$, where $\{A_i, \omega_i, \phi_i\}$ are randomly generated. We denote by $T_i = \frac{2\pi}{\omega_i}$ the periodicity of the demand at agent i. The noise $w_i(t)$ is subject to a zero-mean Gaussian distribution $\mathcal{N}(0, \sigma_i^2)$, and we set σ_i to be 10% of A_i. Given this demand model, we let $s_i(t) = [m_i(t), \bar{t}_i(t)]^T$, where $\bar{t}_i(t)$ denotes the phase of the local demand. Moreover, we define the state transition function $T(s, a, w_i)$ as $m_i(t+1) = m_i(t) + \sum_{j\in\mathcal{N}_i} a_{ji} - \sum_{j\in\mathcal{N}_i} a_{ij} - d_i(t)$, and $\bar{t}_i(t+1) = \bar{t}_i(t) + \delta$, if $\bar{t}_i(t) + \delta t < \frac{T_i}{2}$, or $\bar{t}_i(t+1) = \bar{t}_i(t) + \delta t - T_i$, if $\bar{t}_i(t) + \delta t > \frac{T_i}{2}$. where δt is the sampling interval. The local reward function is designed as

$$r_i(s_i(t)) = \begin{cases} 0 & \text{if } m_i(t) > 0, \\ -(-m_i(t))^3 & \text{if } m_i(t) < 0. \end{cases} \tag{15}$$

This reward function penalizes agents for having negative resources but also does not reward them for accumulating too many resources.

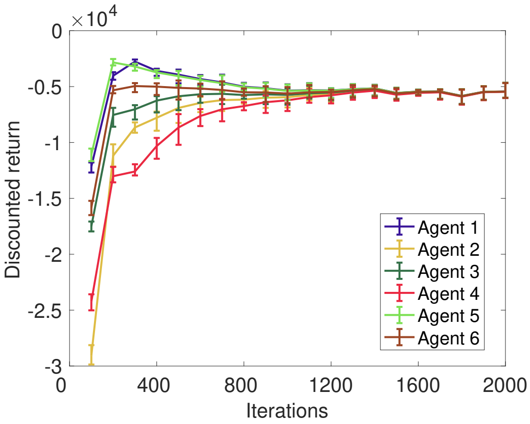

Fig. 1: Performance of Algorithm 1 on a distributed resource allocation problem. Each curve represents the policy improvement of one agent over iterations. The algorithm is run 5 times. The curves represent the mean performance over the 5 trials and the error bars represent the variance of the performance.

We apply Algorithm 1 to solve this problem. Specifically, we parmeterize the global policy function $\pi(s)$ using randomly generated radial Gaussian basis functions. As discussed at the end of Section IV, we utilize subsampling to achieve a two-time scale effect. Specifically, we let the local actors update once every 100 updates of the critics. The step sizes are chosen as $\alpha_w = \alpha_v = \alpha_u = t^{-0.35}$ and $\alpha_\theta = t^{-0.55}$. We run Algorithm 1 for 5 times with the same initialization. In Figure 1, we demonstrate that the accumulated return using each agent's policy estimate is improved. We observe in Figure 1 that by applying Algorithm 1, the local agent's policies achieve consensus and are all improved from their initial values. Agents $1, 5$ and 6 find better policies at the beginning but downgrade to achieve consensus with the other agents at the end. This is due to the nonlinear nature of the problem and that Algorithm 1 is only guaranteed to converge to local maximum solution.

VI. Conclusions

In this paper, we proposed a distributed actor critic algorithm to solve multi-agent RL problems. Specifically, we let every agent improve its local estimate of the global optimal policy function, and we introduced an additional consenesus step on these local estimates so that the agents asymptotically achieve agreement on the global optimal policy. We anlayzed the convergence of the proposed algorithm and demonstrated its effectiveness on a distributed resource allocation example. Compared to existing distributed actor critic methods for RL, using policy consensus does not require the agents to share their tasks with each other.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," in *NIPS Deep Learning Workshop*, 2013.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[3] D. P. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 1995.

[4] L. Baird, "Residual algorithms: Reinforcement learning with function approximation," in *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 30–37.

[5] R. S. Sutton, C. Szepesvári, and H. R. Maei, "A convergent o (n) algorithm for off-policy temporal-difference learning with linear function approximation," *Advances in neural information processing systems*, vol. 21, no. 21, pp. 1609–1616, 2008.

[6] H. R. Maei, C. Szepesvári, S. Bhatnagar, and R. S. Sutton, "Toward off-policy learning control with function approximation." in *International Conference on Machine Learning*, 2010, pp. 719–726.

[7] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[8] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning*, 2014.

[9] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in neural information processing systems*, 2000, pp. 1008–1014.

[10] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[11] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems*, 2017, pp. 6379–6390.

[12] S. V. Macua, J. Chen, S. Zazo, and A. H. Sayed, "Distributed policy evaluation under multiple behavior strategies," *IEEE Transactions on Automatic Control*, vol. 60, no. 5, pp. 1260–1274, 2015.

[13] M. S. Stanković and S. S. Stanković, "Multi-agent temporal-difference learning with linear function approximation: Weak convergence under time-varying network topologies," in *2016 American Control Conference (ACC)*. IEEE, 2016, pp. 167–172.

[14] K. Yuan, B. Ying, J. Liu, and A. H. Sayed, "Variance-reduced stochastic learning by networked agents under random reshuffling," *IEEE Transactions on Signal Processing*, vol. 67, no. 2, pp. 351–366, 2017.

[15] H.-T. Wai, Z. Yang, P. Z. Wang, and M. Hong, "Multi-agent reinforcement learning via double averaging primal-dual optimization," in *Advances in Neural Information Processing Systems*, 2018, pp. 9672–9683.

[16] D. Lee, H. Yoon, and N. Hovakimyan, "Primal-dual algorithm for distributed reinforcement learning: distributed gtd," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 1967–1972.

[17] P. Pennesi and I. C. Paschalidis, "A distributed actor-critic algorithm and applications to mobile sensor network coordination problems," *IEEE Transactions on Automatic Control*, vol. 55, no. 2, pp. 492–497, 2010.

[18] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*, 2018, pp. 5867–5876.

[19] K. Zhang, Z. Yang, and T. Basar, "Networked multi-agent reinforcement learning in continuous spaces," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 2771–2776.

[20] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.

[21] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 391–405, 2013.

[22] T. Degris, M. White, and R. Sutton, "Off-policy actor-critic," in *International Conference on Machine Learning*, 2012.

[23] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.