Proceedings of the 2025 SICE Festival with Annual Conference
September 9-12, 2025, Chiang Mai, Thailand

Decentralized Counterfactual Multi-Agent Actor-Critic Algorithms

Yiliu Jiang[1†] and Guanghui Wen[2]

[1]School of Mathematics, Southeast University, Nanjing, China
(E-mail: yiliujiang@seu.edu.cn)
[2]School of Automation, Southeast University, Nanjing, China
(E-mail: ghwen@seu.edu.cn)

Abstract: Current research indicates a great application potential of multi-agent reinforcement learning (MARL) on real-world network systems, such as power systems, traffic networks, and multi-unmanned aerial vehicle systems. There is a challenge for MARL is efficiently achieve global cooperative performance while in a decentralized learning style that is applicable to networked systems. To this end, we propose novel decentralized multi-agent actor-critic algorithms inspired by the ideas of parameter consensus learning and counterfactual baseline of multi-agent. Specifically, we reasonably consider that the reward functions of the agents are different and are available only to the corresponding agent. A consensus update is designed to approximate the joint reward function via communication over the network to ensure global goal consistency. Based on this, in the critic step, a truncated representation of joint rewards is used for respective value function learning that reduces variance. For the actor step, the truncated-based counterfactual advantage is accordingly computed to enable an efficient credit assignment for each agent's policy improvement that ensures effective joint policy learning. Our algorithms possess provable convergence when the approximation functions are within the class of linear functions and are general for both discrete and continuous spaces of tasks. Experimental results with both linear and nonlinear function approximations show the effectiveness of the proposed algorithms.

Keywords: Multi-Agent Systems, Multi-Agent Reinforcement Learning, Counterfactual Multi-Agent Actor-Critic, Decentralized Training with Decentralized Execution.

1. INTRODUCTION

Multi-agent reinforcement learning (MARL) has gained remarkable research attention nowadays due to its great potential of assisting multi-agent systems (MASs) in complex real-world applications, including multi-robot systems [1, 2], autonomous driving [3, 4], traffic signal control [5, 6], and energy network regulation [7, 8]. To realize the cooperative behavior of a collective of agents, a popular MARL framework is centralized training with decentralized execution (CTDE), which usually receives a total reward of all agents' joint actions for the global goal evaluation [9, 10]. However, CTDE becomes incapable when conducting learning in large-scale networked MASs, since the reward functions may be set in a distributed and secure way, and the classical overall evaluation approach has inefficient distinguishability for the decentralized policies' improvement. The decentralized training with decentralized execution (DTDE) framework was proposed sequentially, which mainly aims to achieve global-level objectives with local-level information communication and collaborative learning [11, 12, 15]. A critical challenge for decentralized MARL is to achieve the overall objective, i.e., the globally averaged return maximization, effectively and efficiently.

With the goal that realizing efficient global collaboration for decentralized networked MARL, in this paper, we propose two novel decentralized multi-agent actor-critic algorithms based on state value function and action value function, respectively. The main ideas of our algorithms are threefold. First, to maximize the global average return under different reward functions of agents, we design a parameterized function of the global reward for each agent. Based on the parameter consensus update with their neighbors in the network, agents are available to maintain the consistent global reward estimations locally. Second, with the approximated global reward defined on joint actions of all agents, for the critic step, we design a truncated term of the global reward represented only by the individual action of the agent for respective value function approximation. In this way, the variance induced by other agents' actions is reduced during each agent's value function learning process, while without impairing global goal consistency. Third, for the actor step, the truncated-based counterfactual advantage function is proposed for the respective policies to efficiently assign credit and improve. The version of the policy gradient theorem adapted to this setting is proved, which is applicable for both state-based and action-based value functions. We validate our two algorithms with both linear and nonlinear function approximations, in the designed numerical experiment and the benchmark Multi-Agent Particle Environment (MPE) [17], respectively.

2. RELATED WORKS

Zhang *et al.* [12] have taken the lead in studying multi-agent reinforcement learning with a networked structure. On the thought of distributed optimization, the parameter consensus updated approach is proposed for agents to estimate the global objective by local information transfer. The class of consensus method usually focus on maintaining consistency on the global value function to realize coordination of decentralized policies [12-14]. However,

† Yiliu Jiang is the presenter of this paper.

with only such consistency on the overall evaluation is incapable of efficient policy improvement and collaboration with the increased number of agents and complexity of tasks.

On the other hand, the counterfactual multi-agent (COMA) actor-critic was proposed [9] to address the challenges of multi-agent credit assignment to learn decentralized policy efficiently. On a centralized joint action value function, it designs a counterfactual baseline that marginalises out a single agent's action, while keeping the other agents' actions fixed. The proposed centralized counterfactual advantage defined on the joint action with a total reward has limited scalability, and it is applicable only for discrete action spaces.

Qu *et al.* [15, 16] investigate the exponential decay condition for "truncated" value function, which is defined on the local state-action instead of the global one, that achieves scalability for general networked MARL. Nevertheless, the local reward-based learning is unable to ensure a global objective.

Based on the inspirations and limitations of the related works, for the networked MARL, we investigate decentralized counterfactual multi-agent actor-critics to realize efficient collaborative learning with the global objective.

3. BACKGROUND

A Markov decision processes involved in N agents is characterized by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^1, \cdots, \mathcal{A}^N, r^1, \cdots, r^N, P \rangle$. $\mathcal{S}$ is the state space, $\mathcal{A}^1$ is the action space of agent $i, i = 1, ...N$, $r^i : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \to \mathbb{R}$ is the reward function for agent i, and $P : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \to [0, 1]$ is the transition probability map, denotes a state transition probability from state s to s' determined by all agents' actions. The policy of agent i is a mapping $\pi^i : \mathcal{S} \times \mathcal{A}^i \to [0, 1]$, representing the probability of choosing action a^i at state s. The joint policy of all agents is denoted as $\boldsymbol{\pi} = \left(\pi^1, \cdots, \pi^N\right)$, and the probability value of it is $\pi(s, \boldsymbol{a}) = \prod_{i=1}^{N} \pi^i \left(s, a^i\right)$, where $\boldsymbol{a} = \left(a^1, \cdots a^N\right)$ is the joint action of all agents. At each time step t, given current state s_t, each agent i choose action a^i according to individual policy $\pi^i\left(s, a^i\right)$. All agent form a joint action $\boldsymbol{a}_t$ and each agent receives a reward $r^i(s_t, \boldsymbol{a}_t)$. The state then transits to the next s_{t+1} according to the transition probability P.

Networking is a common framework for studying MASs in their entirety. Agents in a networked system are denoted as a set $\mathcal{N} = \{1, \cdots, N\}$. The system can be modeled as an undirected graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$, where each agent i serves as vertex i and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of all edges. The edge $e_{ij} = (i, j) \in \mathcal{E}$ indicates the connection relationship of two agents $i, j \in \mathcal{N}$. Two agents associated with an edge are neighbors, agent i and its all neighbors in the graph together is denoted as a set $\mathcal{N}^i$.

The global objective of the MAS is defined as maximizing the overall average reward per time step:

$$\begin{aligned} \underset{\boldsymbol{\pi}}{\text{maximize}}\ J &= \lim_{T\to\infty} \frac{1}{T} \mathop{\mathbb{E}}_{s\sim P, \boldsymbol{a}\sim\boldsymbol{\pi}} \left(\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i\in\mathcal{N}} r^i(s, \boldsymbol{a}) \right) \\ &= \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(s, \boldsymbol{a}) \cdot \bar{r}(s, \boldsymbol{a}), \end{aligned} \tag{1}$$

where $d_\pi(s) = \lim_{t\to\infty} P(s_t = s|\pi)$ is the stationary distribution of the Markov chain under policy π, and $\bar{r}(\cdot) = \frac{1}{N} \sum_{i\in\mathcal{N}} r^i(s, \boldsymbol{a})$ is the average function of all agents' rewards.

Given any policy π, the relative action-value function associated with a state-action pair $(s, \boldsymbol{a})$ is defined as

$$Q_\pi(s, \boldsymbol{a}) = \sum_{t=0}^{\infty} \mathop{\mathbb{E}}_{s\sim P, \boldsymbol{a}\sim\boldsymbol{\pi}} [\bar{r}(s_t, \boldsymbol{a}_t) - \bar{\mu}_\pi | s_0 = s, \boldsymbol{a}_0 = \boldsymbol{a}], \tag{2}$$

where $\bar{\mu}_\pi = \frac{1}{N}\mu_\pi^i$, μ_π^i is the average reward per time step of agent i defined as $\mu_\pi^i = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathop{\mathbb{E}}_{s\sim P, \boldsymbol{a}\sim\boldsymbol{\pi}} [r^i(s_t, \boldsymbol{a}_t)]$ $= \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(s, \boldsymbol{a}) \cdot r^i(s, \boldsymbol{a})$. Accordingly, the relative state-value associated with state s under policy π can be defined as $v_\pi(s) = \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(s, \boldsymbol{a}) \cdot Q_\pi(s, \boldsymbol{a})$. The advantage function is denoted as

$$A_\pi(s, \boldsymbol{a}) = Q_\pi(s, \boldsymbol{a}) - V_\pi(s). \tag{3}$$

4. DECENTRALIZED COUNTERFACTUAL MULTI-AGENT ACTOR-CRITIC

In this section, we present the proposed decentralized actor-critic algorithms for the networked agents. We first establish a policy gradient theorem for our MARL setting. On this basis, two collaborative learning algorithms through parameter consensus update are designed with the Q-function and the V-function, respectively.

4.1. Truncated Counterfactual Multi-Agent Policy Gradients

Inspired by the scalable value function definition in [15], which effectively approximates the joint-action value function in a truncated way, and set each agent localized policy $\pi_{\theta^i}^i$ parameterized by θ^i, we define the "truncated" value function for our setting as follows.

Definition 1. In multi-agent MDP $\mathcal{M}$, for all $s \in \mathcal{S}$, $a \in \mathcal{A}$ under policy θ, $\theta = \left(\theta^1, \cdots, \theta^N\right)$, the truncated action-value function for agent $i, i \in \mathcal{N}$, is defined as

$$\tilde{Q}_\theta^i\left(s, a^i\right) = \sum_{t=0}^{\infty} \mathop{\mathbb{E}}_{{}_{-i}\sim\boldsymbol{\pi}^{-i}, s\sim P} \left[\bar{r}(s_t, \boldsymbol{a}_t) - \bar{\mu}_\theta | s_0 = s, a_0^i = a^i\right], \tag{4}$$

where $-i = \{1, ..., i-1, i+1, ..., N\}$ denotes the set of all agents except i, $\bar{r}$ is the the averaged reward of all

agents defined in Eq. (1), and $\bar{\mu}_\theta$ is the averaged value of the average rewards of all agents under the stationary distribution $d_\theta(s)$.

In Definition 1, we define the global value function of each agent explicitly on the respective action a^i, the performance of other agents' actions is reflected by $\bar{r}$ implicitly, to average out the disturbance of others and reduce variance during the value function learning process. On this basis, considering that the averaged reward value $\bar{r}$ is inefficient to cope with the credit assignment with an increased number of agents [9], we take advantage of its idea of counterfactual baseline for policy gradient. The following Lemma of policy gradient is proposed, which generalizes from Theorem 3.1 in [12] and characterizes the gradient of the global objective J in Eq. (1).

Lemma 1. For any $\theta = \left(\theta^1, \cdots, \theta^N\right)$, let $\pi_\theta : \mathcal{S} \times \mathcal{A} \to [0, 1]$ be a policy and let $J(\theta)$ be the globally long-term averaged return defined in Eq. (1). In addition, let A_θ be the parameterized advantage function in Eq. (3). Moreover, for any $i \in \mathcal{N}$, we define the truncated counterfactual local advantage function $\tilde{A}^i_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$\tilde{A}^i_\theta\left(s, a^i\right) = \tilde{Q}^i_\theta\left(s, a^i\right) - \sum_{a^i \in \mathcal{A}^i} \pi^i_{\theta^i}\left(s, a^i\right) \cdot \tilde{Q}^i_\theta\left(s, a^i\right). \quad (5)$$

Then the gradient of $J(\theta)$ with respect to θ^i is given by

$$\begin{aligned} \nabla_{\theta^i} J(\theta) &= \mathop{\mathbb{E}}_{s \sim d_\theta, \boldsymbol{a} \sim \boldsymbol{\pi}_\theta} \left[\nabla_{\theta^i} \log \pi^i_{\theta^i}\left(s, a^i\right) \cdot A_\theta(s, \boldsymbol{a})\right] \\ &= \mathop{\mathbb{E}}_{s \sim d_\theta, \boldsymbol{a} \sim \boldsymbol{\pi}_\theta} \left[\nabla_{\theta^i} \log \pi^i_{\theta^i}\left(s, a^i\right) \cdot \tilde{A}^i_\theta\left(s, a^i\right)\right]. \end{aligned} \quad (6)$$

Proof. The proof of this theorem follows the proof of the policy gradient theorem for MARL (Theorem 3.1) in [12], from which we have

$$\begin{aligned} \nabla_{\theta^i} J(\theta) &= \mathop{\mathbb{E}}_{s \sim d_\theta, \boldsymbol{a} \sim \boldsymbol{\pi}_\theta} \left[\nabla_{\theta^i} \log \pi^i_{\theta^i}\left(s, a^i\right) \cdot Q_\theta(s, \boldsymbol{a})\right] \\ &= \mathop{\mathbb{E}}_{s \sim d_\theta, \boldsymbol{a} \sim \boldsymbol{\pi}_\theta} \left[\nabla_{\theta^i} \log \pi^i_{\theta^i}\left(s, a^i\right) \cdot A_\theta(s, \boldsymbol{a})\right], \end{aligned} \quad (7)$$

where $Q_\theta(s, \boldsymbol{a})$ is the action-value function in Eq. (2) parameterized by θ. We write the difference term as

$$\begin{aligned} g_d = &\mathop{\mathbb{E}}_{s \sim d_\theta, \boldsymbol{a} \sim \boldsymbol{\pi}_\theta} \left[\nabla_{\theta^i} \log \pi^i_{\theta^i}\left(s, a^i\right) \cdot A_\theta(s, \boldsymbol{a})\right] \\ &- \mathop{\mathbb{E}}_{s \sim d_\theta, \boldsymbol{a} \sim \boldsymbol{\pi}_\theta} \left[\nabla_{\theta^i} \log \pi^i_{\theta^i}\left(s, a^i\right) \cdot \tilde{A}^i_\theta\left(s, a^i\right)\right]. \end{aligned}$$

From the definitions of $A_\theta(s, \boldsymbol{a})$ and $\tilde{A}^i_\theta\left(s, a^i\right)$ respectively in Eq. (3) and Eq. (5), one can easily obtain $g_d = 0$. □

Lemma 1 indicates that, based on the respective truncated counterfactual advantage $\tilde{A}^i_\theta\left(s, a^i\right)$, each agent i enables an unbiased estimate of the global objective $J(\theta)$ through the corresponding gradient $\nabla_{\theta^i} \log \pi^i_{\theta^i}$ locally. Furthermore, to realize fully decentralized MARL with only individual reward available for each agent, the total average reward $\bar{r}(\cdot)$ required in $\tilde{A}^i_\theta\left(s, a^i\right)$ and $\tilde{Q}^i_\theta\left(s, a^i\right)$ is set as a parameterized function, and we design two parameter consensus update algorithms to realize the approximation of $\bar{r}(\cdot)$.

Algorithm 1 Decentralized counterfactual multi-agent actor-critic based on action-value function

Input: Initial value of parameters λ^i_0, $\tilde{\lambda}^i_0$, μ^i_0, $\tilde{\mu}^i_0$, ω^i_0, θ^i_0, $\forall i \in \mathcal{N}$, the initial state s_0, and step-sizes $\{\alpha_t\}_{t \geq 0}$ and $\{\beta_t\}_{t \geq 0}$.
Initialize the iteration counter $t \leftarrow 0$.
repeat
for all $i \in \mathcal{N}$ **do**
Each agent i executes action $a^i_t \sim \pi^i_{\theta^i_t}(s_t, \cdot)$ and observes joint actions $a_t = \left(a^1_t, ..., a^N_t\right)$.
Observe state s_{t+1}, and reward $r^i_t = r^i(s_t, a_t)$.
Receive r^j_t from neighbors $j \in \mathcal{N}^i$.
Update $\tilde{\mu}^i_t \leftarrow \mu^i_t + \alpha_t \cdot \left(r^i_t - \mu^i_t\right)$
Update $\tilde{\lambda}^i_t \leftarrow \lambda^i_t + \alpha_t \cdot \left(r^i_{t+1} - \bar{R}^i_{\lambda_t}(s_t, a_t)\right) \cdot \nabla_\lambda \bar{R}^i_{\lambda_t}(s_t, a_t)$
Update $\delta^i_t \leftarrow \bar{R}^i_{\lambda_t}(s_t, a_t) - \mu^i_t + Q^i_{\omega^i_t}\left(s_{t+1}, a^i_{t+1}\right) - Q^i_{\omega^i_t}\left(s_t, a^i_t\right)$, $A^i_t \leftarrow Q^i_{\omega^i_t}\left(s_t, a^i_t\right) - \sum_{a^i \in \mathcal{A}^i} \pi^i_{\theta^i}\left(s_t, a^i\right) \cdot Q^i_{\omega^i_t}\left(s_t, a^i_t\right)$
Critic step: $\omega^i_{t+1} \leftarrow \omega^i_t + \alpha_t \cdot \delta^i_t \cdot \nabla_{\omega^i} Q^i_{\omega^i_t}\left(s_t, a^i_t\right)$.
Actor step: $\theta^i_{t+1} \leftarrow \theta^i_t + \beta_t \cdot A^i_t \cdot \nabla_{\theta^i} \log \pi^i_{\theta^i_t}\left(s_t, a^i_t\right)$.
end for
for all $i \in \mathcal{N}$ **do**
Consensus step: $\mu^i_{t+1} \leftarrow \sum_{j \in \mathcal{N}^i} c_t(i, j) \cdot \tilde{\mu}^i_t$, $\lambda^i_{t+1} \leftarrow \sum_{j \in \mathcal{N}^i} c_t(i, j) \cdot \tilde{\lambda}^i_t$.
end for
Update the iteration counter $t \leftarrow t + 1$.
until Convergence

4.2. Algorithms

Zhang *et al.* [12] firstly introduced the parameter consensus update approach for MARL on the general network topology, realizing the global value function estimation through local information communication during learning. Here, we set the global averaged reward function $\bar{r}(s, \boldsymbol{a}) = \frac{1}{N} \sum_{i \in \mathcal{N}} r^i(s, \boldsymbol{a})$ as the parameterized one. Specifically, let $R(\cdot, \cdot; \lambda) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ be a class of functions parameterized by $\lambda \in \mathbb{R}^K$, where $K \ll |\mathcal{S}| \cdot |\mathcal{A}|$. Each agent i maintains its own parameter λ_i and uses $R\left(\cdot, \cdot; \lambda^i\right)$ as a local estimate of R_λ, and let agents share the local parameters $\lambda^i, i \in \mathbb{N}$ with their neighbors N^i on the network to realize a consensual estimate of R_λ. In this way, each agent is able to improve its respective policy via the counterfactual advantage-defined gradient in Eq. (6).

Such a parameter update involves a weight matrix $C_t = [c_t(i, j)]_{N \times N}$ to achieve the averaged consensus, where $c_t(i, j)$ is the weight on the message transmitted from i to j at time t through the existing communication edge in the network. We assume C_t follows the theoretical conditions (see Assumption 4.3 in [12]), thus, the

Algorithm 2 Decentralized counterfactual multi-agent actor-critic based on state-value function

Input: Initial value of parameters ζ_0^i, λ_0^i, $\tilde{\lambda}_0^i$, μ_0^i, $\tilde{\mu}_0^i$, ν_0^i, θ_0^i, $\forall i \in \mathcal{N}$, the initial state s_0, and step-sizes $\{\alpha_t\}_{t\geq 0}$ and $\{\beta_t\}_{t\geq 0}$.
Initialize the iteration counter $t \leftarrow 0$.
repeat
for all $i \in \mathcal{N}$ **do**
Each agent i executes action $a_t^i \sim \pi_{\theta_t^i}^i(s_t, \cdot)$ and observes joint actions $a_t = \left(a_t^1, ..., a_t^N\right)$.
Observe state s_{t+1}, and reward $r_t^i = r^i(s_t, a_t)$.
Receive r_t^j from neighbors $j \in \mathcal{N}^i$.
Update $\tilde{\mu}_t^i \leftarrow \mu_t^i + \alpha_t \cdot \left(r_t^i - \mu_t^i\right)$
Update
$\tilde{\lambda}_t^i \leftarrow \lambda_t^i + \alpha_t \cdot \left(r_{t+1}^i - \bar{R}_{\lambda_t}^i(s_t, a_t)\right) \cdot \nabla_\lambda \bar{R}_{\lambda_t}^i(s_t, a_t)$
$\zeta_{t+1}^i \leftarrow \zeta_t^i + \alpha_t \Big(\bar{R}_{\lambda_t}^i(s_t, a_t) - \bar{R}_{\zeta^i}^i(s_t, a_t^i)\Big) \cdot \nabla_\zeta \bar{R}_{\zeta_t}^i(s_t, a_t^i)$
Update $\delta_t^i \leftarrow \bar{R}_{\zeta^i}^i\left(s_t, a_t^i\right) - \mu_t^i + V_{\nu_t^i}^i(s_{t+1}) - V_{\nu_t^i}^i(s_t)$
Critic step: $\nu_{t+1}^i \leftarrow \nu_t^i + \alpha_t \cdot \delta_t^i \cdot \nabla_{\nu^i} V_{\nu_t^i}^i(s_t)$.
Actor step: $\theta_{t+1}^i \leftarrow \theta_t^i + \beta_t \cdot \delta_t^i \cdot \nabla_{\theta^i} \log \pi_{\theta_t^i}^i\left(s_t, a_t^i\right)$.
end for
for all $i \in \mathcal{N}$ **do**
Consensus step: $\mu_{t+1}^i \leftarrow \sum_{j \in \mathcal{N}^i} c_t(i,j) \cdot \tilde{\mu}_t^i$,
$\lambda_{t+1}^i \leftarrow \sum_{j \in \mathcal{N}^i} c_t(i,j) \cdot \tilde{\lambda}_t^i$.
end for
Update the iteration counter $t \leftarrow t+1$.
until Convergence

Q-value based actor-critic algorithm can be directly obtained under the definition Eq. (4) and Eq. (5). Algorithms 1 represents the pseudocode, where the personal Q-function and policy are parameterized by ω^i and θ^i respectively.

For the V-value based actor-critic, we rewrite the truncated counterfactual local advantage in Eq. (5) as

$$\begin{aligned} \tilde{A}_\theta^i\left(s, a^i\right) &= \tilde{Q}_\theta^i\left(s, a^i\right) - \sum_{a^i \in \mathcal{A}^i} \pi_{\theta^i}^i\left(s, a^i\right) \cdot \tilde{Q}_\theta^i\left(s, a^i\right) \\ &= \bar{r}_\theta^i\left(s, a^i\right) - \bar{\mu}_\theta + \tilde{V}_\theta^i\left(s'\right) - \tilde{V}_\theta^i(s), \end{aligned} \tag{8}$$

where $\bar{r}_\theta^i\left(s, a^i\right) = \sum_{\boldsymbol{a}^{-i} \in \mathcal{A}^{-i}} \pi_\theta^{-i}\left(s, \boldsymbol{a}^{-i}\right) \cdot \bar{r}(s, \boldsymbol{a})$, and $\tilde{V}_\theta^i(s) = \sum_{a^i \in \mathcal{A}^i} \pi_{\theta^i}^i\left(s, a^i\right) \cdot \tilde{Q}_\theta^i\left(s, a^i\right)$.

Considering the extra truncated reward term $\bar{r}_\theta^i\left(s, a^i\right)$ in Eq. (8), we set an additional function to parameterized it denoted as $R\left(\cdot, \cdot; \zeta^i\right)$, $\zeta^i \in \mathbb{R}^L$, in which $L \ll |\mathcal{S}| \left|\mathcal{A}^i\right|$. Then, the pseudocode of this case is represented as Algorithms 2, where the personal V-function is parameterized by ν^i. Based on the policy gradient in Lemma 1, the convergence of two decentralized counterfactual multi-agent actor-critic with linear function approximation can be similarly proved, through the consensus MARL theory in [12].

5. EXPERIMENTS

In this section, we evaluate the proposed decentralized counterfactual MAAC algorithms with both linear and nonlinear function approximation, in the designed numerical experiment and the benchmark Multi-Agent Particle Environment [17], respectively.

5.1. Linear Function Approximation

To test the effective global objective approximation for our decentralized algorithms, the numerical experiment setting in this case is similar to [12]. Consider in total $N = 10$ agents, each has a binary-valued action space, i.e., $A_i = \{0, 1\}$, for all $i \in \mathcal{N}$. There are total $|\mathcal{S}| = 10$ states. The elements in the transition probability matrix P are uniformly sampled from the interval [0,1] and normalized to be stochastic. For each agent i and each state-action pair (s, a), the mean reward $R^i(s, a)$ is sampled uniformly from [0,4], which varies among agents. The policy $\pi_{\theta^i}^i\left(s, a^i\right)$ is parameterized following the Boltzman policies. The feature vectors $f_{s,a^i} \in \mathbb{R}^{L_i}$ for policy function is $L_1 = ... = L_N = 6$, $\phi \in \mathbb{R}^K$ for the state-action (s, a), $\varphi \in \mathbb{R}^D$ for the state s, are all uniformly sampled from [0,1], of dimensions $K = 12 \ll |\mathcal{S}| \cdot |\mathcal{A}|$, $D = 5 < |\mathcal{S}|$. The consensus weight matrix C_t is chosen to be independent and identically distributed along time t, and to be doubly stochastic (see [12] for the details). The stepsizes for the critic and the actor are selected as $\alpha_t = 1/t^{0.85}$ and $\beta_t = 1/t^{0.95}$.

The performances of our decentralized counterfactual algorithms labeled as '-Deco' are compared with those of the centralized algorithms and the decentralized algorithms in [12], which are labeled as '-Central' and '-Decentral' respectively. For the centralized controller, the rewards r_t^i of all agents are available. As shown in Fig. 1 and Fig. 2, both decentralized counterfactual algorithms converge to the globally long-term averaged return as achieved by the two centralized algorithms. In view of the overall limited performance of the linear function approximation, moreover, we compare the V-value and Q-value of algorithms after 3000 iterations, which are shown in Fig. 3 and Fig. 4 respectively. The value of '-Deco' is computed as the average of all agents. These verify that Deco algorithms, through distributed and distinguished

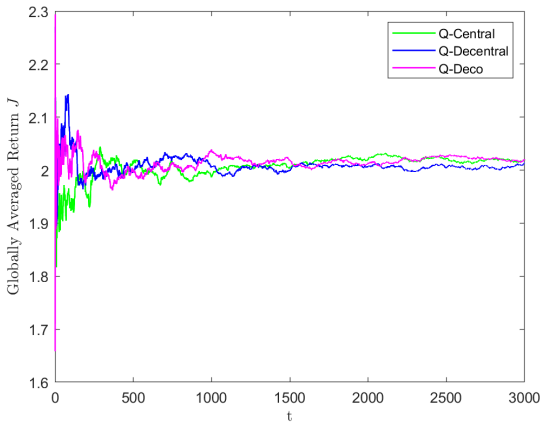

Fig. 1 The convergence of globally averaged returns of Q-algorithms.

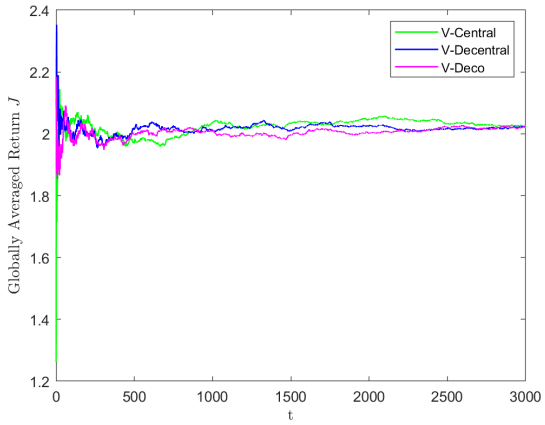

Fig. 2 The convergence of globally averaged returns of V-algorithms.

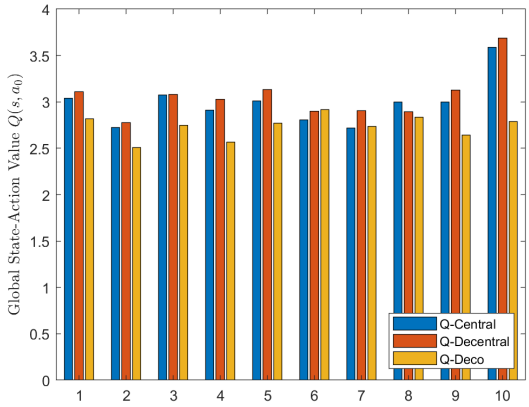

Fig. 3 The values of the Q-function after 3000 steps.

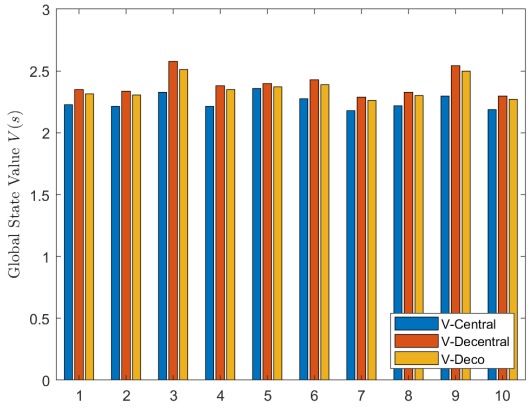

Fig. 4 The values of the V-function after 3000 steps.

counterfactual advantage function learning, indeed learn similar global value evaluations as their central counterparts.

5.2. Nonlinear Function Approximation

In the scenario of MPE, we empirically evaluate the performance of the algorithms with nonlinear function approximators, i.e., neural networks in this case. Considering the scenario of the Cooperative Spread task, in this scenario, we set 4 agents to collaboratively spread to 4 landmarks through physical movement. Agents are able to globally observe the position of the landmarks and other agents, and receive rewards based on the negative distance of each agent to a landmark that is closest. In addition, a reward of -1 is set as a collision punish for each other. The collaborative goal of the agents is then to occupy as many spread targets as possible in the episode length of 30. The illustration of the Cooperative Spread scenario is shown as Fig. 5. Both state and action spaces are continuous in this case. The neural networks in the experiment are all set to have two hidden layers with 64 neural units per layer, which all use ReLU as the activation function. The learning rate for the actor and critic steps are set as constants 1e-4 and 2e-4, respectively.

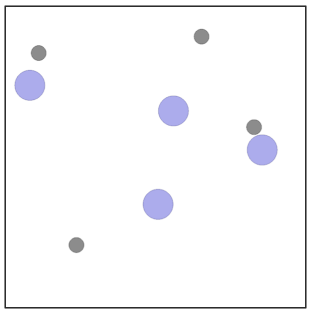

Fig. 5 The Cooperative Spread scenario in MPE.

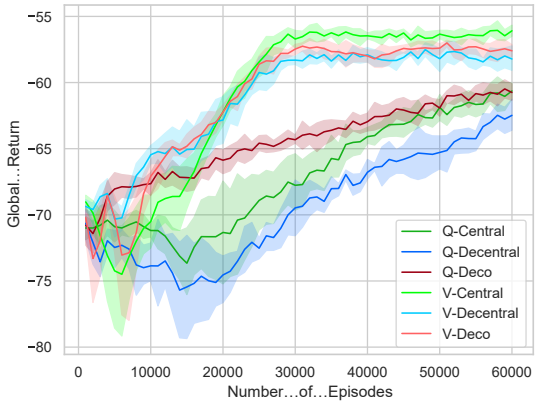

Fig. 6 The globally averaged returns for the task of Cooperative Spread.

Fig. 6 shows the successful convergence of the proposed algorithms. For the V-function algorithms, both '-Deco' and '-Decentral' algorithms are able to achieve globally averaged returns close to the centralized counterparts, though at a relatively slower speed. This is reasonable due to the delay of information aggregation across the network. The 'V-Deco' compared with the 'V-Decentral' does not show an obvious advantage here. We attribute this to the fact that the class of V-function-based algorithms has an inherently smaller variance, so that the credit assignment of the counterfactual advantage has limited improvement. For the Q-function al-

gorithms, the '-Deco' shows an outperforming learning efficiency compared with both the '-Central' and the '-Decentral' ones. Overall, the results validate the efficiency of the truncated counterfactual advantage for the decentralized multi-agent collaborative policy learning, especially for Q-function approximation that is affected by other agents' actions with larger variance.

6. CONCLUSIONS

In this paper, the truncated counterfactual advantage function is proposed for the decentralized MARL to achieve efficient collaborative learning on networked multi-agent systems. On this basis, we propose two multi-agent actor-critic algorithms that can generally handle both continuous and discrete spaces of tasks, with a provable global convergence property for the linear function approximation. Numerical experiments on large-scale MARL settings with numerous agents and massive state-action spaces validate the global performance of the proposed decentralized algorithms. Moreover, experiments on the benchmark Multi-Agent Particle Environment show the superior learning efficiency of the proposed truncated counterfactual MARL. The future direction of our research is to extend the algorithms in more distributed and scalable ways that reduce the reliance on global information, such as global state and joint actions.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China through Grant Nos. 62325304, U22B2046, U24A20279, and in part by the Jiangsu Provincial Scientific Research Center of Applied Mathematics under Grant No. BK20233002.

REFERENCES

[1] Z. Mou, Y. Zhang, F. Gao, H. Wang, T. Zhang, and Z. Han, "Deep reinforcement learning based three-dimensional area coverage with UAV swarm", *IEEE Journal on Selected Areas in Communications*, Vol. 39, No. 10, pp. 3160-3176, 2021.

[2] E. Marchesini and A. Farinelli, "Enhancing deep reinforcement learning approaches for multi-robot navigation via single-robot evolutionary policy search", *IEEE/International Conference on Robotics and Automation (ICRA)*, pp. 5525-5531, 2022.

[3] P. Palanisamy, "Multi-agent connected autonomous driving using deep reinforcement learning", *IEEE/International Joint Conference on Neural Networks (IJCNN)*, pp. 1-7, 2020

[4] E. Vinitsky, N. Lichtl, X. Yang, B. Amos, and J. Foerster, "Nocturne: a scalable driving benchmark for bringing multi-agent learning one step closer to the real world", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 3962-3974, 2022.

[5] J. Ault and G. Sharon, "Reinforcement learning benchmarks for traffic signal control", *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[6] W. Du, J. Ye, J. Gu, J. Li, H. Wei, and G.Wang, "Safelight: A reinforcement learning method toward collision-free traffic signal control", *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37, No. 12, pp. 14801-14810, 2023.

[7] J. Wang, W. Xu , Y. Gu, W. Song, and T. C. Green, "Multi-agent reinforcement learning for active voltage control on power distribution networks", *Advances in Neural Information Processing Systems*, Vol. 34, pp. 3271-3284, 2021.

[8] C. Ma, A. Li, Y. Du, H. Dong, and Y. Yang, "Efficient and scalable reinforcement learning for large-scale network control", *Nature Machine Intelligence*, Vol. 6, No. 9, pp. 1006-1020, 2024.

[9] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients", *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32, No. 1, 2018.

[10] J. G. Kuba, R. Chen, M. Wen, Y. Wen, F. Sun, J. Wang, and Y. Yang, "Trust region policy optimisation in multi-agent reinforcement learning", *arXiv preprint*, arXiv: 2109.11251, 2021.

[11] G. Wen, J. Fu, P. Dai, and J. Zhou, "DTDE: A new cooperative multi-agent reinforcement learning framework", *The Innovation*, Vol. 2, No. 4, 2021.

[12] K. Zhang, Z. Yang, H. Liu, T., Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents", *PMLR/International conference on machine learning*, pp. 5872-5881, 2018

[13] C. Qu, S. Mannor, H. Xu, Y. Qi, L. Song, and J. Xiong, "Value propagation for decentralized networked deep multi-agent reinforcement learning", *Advances in Neural Information Processing Systems*, pp. 32, 2019.

[14] Z. Chen, Y. Zhou, R. R. Chen, and S. Zou, "Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis", *PMLR/International Conference on Machine Learning*, pp. 3794-3834, 2022

[15] G. Qu, Y. Lin, A. Wierman, and N. Li, "Scalable multi-agent reinforcement learning for networked systems with average reward", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 2074-2086, 2020.

[16] G. Qu, A. Wierman, and N. Li, "Scalable reinforcement learning for multiagent networked systems", *Operations Research*, Vol. 70, No. 6, pp. 3601-3628, 2022.

[17] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments", *Advances in neural information processing systems*, pp. 30, 2017