DTDE: A new cooperative multi-agent reinforcement learning framework

Guanghui Wen,1,* Junjie Fu,1 Pengcheng Dai,1 and Jialing Zhou2
1Department of Systems Science, School of Mathematics, Southeast University, Nanjing 210096, China
2School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China
*Correspondence: ghwen@seu.edu.cn

Received: April 23, 2021; Accepted: August 30, 2021; Published Online: September 1, 2021; https://doi.org/10.1016/j.xinn.2021.100162

© 2021 The Authors. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Citation: Wen G., Fu J., Dai P., and Zhou J. (2021). DTDE: A new cooperative multi-agent reinforcement learning framework. The Innovation 2(4), 100162.

A significant body of work on reinforcement learning has been focused on the single-agent tasks where the agent aims to learn a policy that maximizes the cumulative reward in a dynamic environment.1 In the past decades, quite a few single-agent-based reinforcement learning algorithms have been developed in the literature.1 Yet, it is increasingly recognized that the single-agent-based reinforcement learning algorithms may fail to effectively handle large-scale optimization (decision) tasks with joint features. Within this context, cooperative multi-agent reinforcement learning (CMARL) algorithms have been proposed, where the agents aim to complete the multi-agent learning goal cooperatively through information exchange between neighboring agents. It has been witnessed in the past few years that CMARL algorithms have received increasing attention due to their broad applications in various fields, such as traffic signal control in intelligent transportation systems, energy management of smart grid, and coordination control of robot swarms.

Compared with the single-agent reinforcement learning algorithm that considers only a single agent’s state-action space, the joint state-action space of the CMARL algorithm grows exponentially as the number of agents increases.2 Therefore, CMARL encounters major challenges of algorithm complexity and scalability. Another challenge in designing an efficient CMARL algorithm is the partial observability of the environment in which each agent has to make its individual decisions based on the local observations. Within the context of CMARL, the independent Q-learning (IQL)-based algorithms, where each agent establishes a local Q-function by local state-action information, have been suggested and discussed in the literature.2 The advantages of several IQL-based CMARL algorithms compared with the independent MARL algorithms have been also examined.2 Although the IQL has good scalability, it sometimes cannot guarantee the collection of individual optimal actions of agents produced by local Q-functions equivalent to the optimal joint action, that is, the individual global max (IGM) principle may not be satisfied. Motivated partly by this observation, a new kind of CMARL paradigm based on the centralized training with decentralized execution (CTDE) mechanism has recently attracted significant attention. In CTDE, the agents’ policies are trained with access to global information in a centralized manner and executed based only upon local observation in a decentralized way. Typical CTDE-based CMARL algorithms include value decomposition networks (VDN), among others.3

The aforementioned results advanced our knowledge of how to design CTDE-based algorithms for coping with CMARL problems. However, most of the above-mentioned CTDE-based algorithms are preliminarily focused on solving the CMARL problems in the absence of constraints on agents’ actions, especially the inherent coupling (joint) constraints. However, due to the inherent complexity of large-scale CMARL problems, the feasible actions of an individual agent are generally affected by those of the other agents. For example, when designing CMARL algorithms to address the economic power dispatch problem for the smart grid in the presence of multiple generating units, the generating units are always modeled as agents whose feasible actions (corresponding to the power outputs) are dependent on each other, as the total power output should equal to the power demand.4 The capability of addressing the constraints on actions of multiple agents has become an important index of the practicality of CMARL algorithms. In the context of distributed consensus of multi-agent systems

![Figure 1: An illustration of the DTDE structure with N agents labeled from 1 to N, where } is the set of neighbors of agent i in the undirected and connected communication network, } and } are respectively the estimates of the global state and the average reward of the neighbors of agent i, } is the information coming from the communication network and will be iteratively updated until the joint action } is a feasible action, } is the local Q-function with the local Q-function parameter }, } is the Q-functions of the neighboring agents, and } is related to the local Q-functions of all agents through “mixing Q-function,” where } is the consensus value of the global state estimations and .

---

1. The Innovation
2. The Innovation
3. The Innovation
4. The Innovation
(MASs), various distributed information exchange protocols have been constructed and embedded into the agents such that the states of all agents will converge to a common value.\textsuperscript{12} During the consensus-seeking process, the agents equipped with distributed information exchange protocols will share information with their neighbors through the underlying communication network among them. Based on the ideas of CTDE and distributed consensus of MASs, a distributed reinforcement learning-based framework called distributed training with decentralized execution (DTDE) is envisioned in this paper. Specifically, the DTDE structure suggested in this work is shown in Figure 1.

Specifically, different from the method of CTDE, where each individual agent can obtain the global state $s$ and the common (joint) reward $r$ directly, each agent $i$ in the present DTDE algorithm will employ a "consensus protocol" to respectively estimate the global state $s$ and the average reward of the agents by using only local information. Certainly, the global state and the average reward are defined according to the specific MARL tasks under consideration. For example, the total power demand can be selected as the global state, while the reciprocal value of the averaged generation cost of all agents can be selected as the average reward, when using CMARL algorithms to solve the economic power dispatch problem for a smart grid in consideration. For example, the total power demand can be selected as the global state, while the reciprocal value of the averaged generation cost of all agents can be selected as the average reward, when using CMARL algorithms to solve the economic power dispatch problem for a smart grid in consideration.

In the exploitation step of the agents, the agents choose actions under the condition that the underlying communication network is undirected and connected. In addition, each agent $i$ may utilize the information of the estimated local state $\hat{s}_i$ and the local action $a_i$ based on the technique of distributed optimization, the values of $\arg\max_a Q_{tot}(\hat{s}_i, a; \theta)$ and $\max_a Q_{tot}(\hat{s}_i, a'; \theta^*)$ under action constraints can be calculated in a distributed way through the underlying communication network where $\hat{s}_i$ is the consensus value of the next global state estimation produced by $s$ and $a$, $\theta^* = (\theta_1, ..., \theta_N)$ are the parameters of a target network and periodically copied from $\theta$ as in a deep Q-network. The key steps of DTDE are provided as follows:

Step 1. Each agent $i$ obtains the observation $o_i$ from the environment and employs the local information $\hat{s}_i$ and $\hat{r}_i$, which is obtained from the communication network to estimate the global state $s$ through the "consensus protocol" in a distributed way. By the theory of distributed consensus, $\hat{s}_i$ could generally reach a consistent value $\hat{s}$, which equals to the global state $s$ under the condition that the underlying communication network is undirected and connected. In addition, each agent $i$ utilizes the information of $\hat{r}_i$, which comes from the neighbors of agent $i$ to estimate the average reward.

Step 2. Agents employ the $\epsilon$-greedy policy in the training and execution process. In the exploration step of the agents, each agent $i$ utilizes the local information $\hat{s}_i$ and $\hat{a}_{i}'$, which is obtained from the communication network to find a feasible joint action through "distributed exploration." It should be noted that $\hat{a}_{i}'$ is the information coming from the communication network, which is iteratively updated until the joint action $a = (a_1, \ldots, a_N)$ is a feasible action. In the exploitation step of the agents, the agents choose $\arg\max_a Q_{tot}(\hat{s}, a; \theta)$ to execute.

Step 3. Each agent $i$ updates the local parameter $\theta_i$ by minimizing the temporal difference (TD) loss $J_{tot}(\theta)$, which is denoted as $J_{tot}(\theta) = \left[ r + \gamma \max_a Q_{tot}(\hat{s}', a'; \theta^*) - Q_{tot}(\hat{s}, a; \theta) \right]$, where $\theta^* = (\theta_1, \ldots, \theta_N)$ are the parameters of a target network and periodically copied from $\theta$, as in a deep Q-network. It should be noted that the value of $\max_a Q_{tot}(\hat{s}', a'; \theta^*)$ in the presence of action constraints can be calculated by the technique of "distributed optimization."

Compared with the CTDE-based CMARL algorithms, the potential advantages of the DTDE-based CMARL algorithms include:

1. As the agents can estimate the global state through distributed algorithms based on the local observations, the common assumption made in executing the CTDE-based CMARL algorithms that the agents know the global information of state can be successfully removed from the DTDE-based framework. Furthermore, the distributed training structure in the present DTDE-based framework could improve the robustness of the system. The requirement of privacy preservation may also be ensured by designing distributed privacy-preserving information exchange protocols during the practical implementation.

2. Different from the CTDE-based algorithms that are commonly utilized to solve the CMARL tasks in the absence of action constraint, the DTDE-based algorithms can deal with the CMARL tasks in the presence of constraints on agents’ actions. Moreover, specifically, the DTDE-based algorithms can calculate the value of $\max_a Q_{tot}(\hat{s}', a'; \theta^*)$ in a distributed way under the joint action constraints (e.g., using the distributed optimization technique for MASs).

3. Different from decentralized execution in the CTDE-based CMARL algorithms based on the ICM principle, the decentralized execution in the DTDE-based CMARL algorithms is realized by the distributed optimization technique. Therefore, the optimality of the learned policy can be generally preserved.

To harvest the advantages of DTDE-based CMARL algorithms, the key is to handle the partial observability constraints using the distributed interaction. Individual agents should obtain a common estimation of the global state that reflects information about the whole environment and then make the best decision during each time step. Theoretical guarantees of several typical DTDE-based CMARL algorithms can be established in the future to solidify the proposed CMARL framework.

REFERENCES

1. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction (The MIT Press).

2. Tan, M. (1993). Multi-agent reinforcement learning: independent vs. cooperative agents. In: Proc. 10th Int. Conf. Mach. Learn. 330-337.

3. Sunehag, P., Lever, G., Gruslys, A., et al. (2018). Value-decomposition networks for cooperative multi-agent learning based on team reward. In: Proc. 17th Int. Conf. Auton. Agents MultiAgent Syst. 2085-2087.

4. Dai, P., Yu, W., Wen, G., et al. (2020). Distributed reinforcement learning algorithm for MASs. In: Proc. 20th Int. Conf. Auton. Agents MultiAgent Syst. 2258-2267.

5. Ren, W., and Beard, R.W. (2008). Distributed Consensus in Multi-Vehicle Cooperative Control (Springer-Verlag).

DECLARATION OF INTERESTS

The authors declare no competing interests.