Critical Review for Advanced Topics in Distributed Systems

Stefanos Antaris[1]

[1]**antaris@kth.se**

ABSTRACT

This is the final report of the papers selected for the course Advanced Topics in Distributed Systems.

JUSTIFICATION

The papers selected for this course are under the domain of Multi-Agent Reinforcement Learning (MARL). Provided that this research domain attracted a lot of attention during the last decade and it is still an active research topic with plenty of research questions to be addressed, the selected papers do not address only one research question. However, the selected papers range from 2017 to 2019 and each paper complements the previous one. However, the selected papers are solving problems of MARL. The papers are the following:

- ”Deep Decentralized Multi-Task Multi-Agent Reinforcement Learning under Partial Observability”, Omidshafiei S., Pazis J., Amato C., How P. J., Vian J., ICML 2017
- ”Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents”,Zhang K., Yang Z., Liu H., Zhang T., Basar T., ICML, 2018
- ”Multi-Agent Adversarial Inverse Reinforcement Learning”, Yu L., Song J., Ermon S., ICML 2019

The justification for the selection of the above papers is the following:

> ”Distributed learning and more specifically multi-agent reinforcement learning (MARL) has received significant interest in recent years notably due to the ability to transfer the learning complexity to the agents and allow them to optimize their policies in real time avoiding the bottleneck of the centralized nodes. However, many real-world tasks involve multiple agents with partial observability and limited communication which require more sophisticated algorithms and architectures such that the agents can collaborate and behave optimally in the presence of uncertainty by interacting with the environment. The three selected papers present multi-agent reinforcement learning algorithms that allow the agents to converge to their optimal policy. Moreover, these papers take into consideration the challenge that different agents may have completely different goals and as such they require different reward functions that will allow them to capture complex behaviors in a highly dynamic environment. This challenge becomes increasingly more difficult as the number of agents increases. ”

As stated in the justification above, MARL is a research topic that considers multiple independent agents that communicate each other with the ultimate goal to learn an optimal policy. This policy will allow them to learn how to execute their task(s) in the environment that they operate. It is obvious that MARL can be considered as a distributed system where agents communicate and exchange information in order to achieve their goals. Thus, I selected to present this topic in the course.

As mentioned above, the three papers complement each other by solving problems in MARL that the previous paper didn’t consider. Therefore, in the first paper the authors address the problem of multiple agents learning more than a single tasks when they also present partial observations of the environment. However, as we will explain later, learning multiple tasks require a central authority that provides more

information to the agents. In the second paper, the authors consider the problem of having a fully decentralized methodology where no central authority is required and the agents operate independently. Finally, the third paper considers the problem of different reward functions for each agent. The authors claim that each agent should have different reward function and they present an approach to learn different reward functions by applying imitation learning and adversarial techniques.

PAPER 1. DEEP DECENTRALIZED MULTI-TASK MULTI-AGENT REINFORCEMENT LEARNING UNDER PARTIAL OBSERVABILITY

In this paper, the authors present a methodology for Multi-Task MARL (MT-MARL) so as different agents learn to operate different tasks simultaneously. A typical example of such MT-MARL can be several Autonomous Underwater Vehicles (AUVs) that learn to i) detect, ii) repair the fault of the deep-sea equipment, and iii) navigate in a dynamic environment (varying water currents, moving object, etc.). Additionally, for scalability reasons, agents present limited communication with the rest of the agents which limits the observability of the environment for every agent.

To address the above problem, the authors propose the following:

- The learning process consists of two main phases: i) the task specialization for each agent, and ii) the aggregation of different task specializations to a unified joint policy.

- For the task specialization, the agents use a variation of Distributed Q-Learning to learn the Q-values of the actions given a policy. The variation of Distributed Q-Learning is called *Decentralized Hysteretic Deep Recurrent Q-Networks* (Dec-HDRQNs) and it allows the agents avoid Q-value degradation of positive past experiences. This means that if the agent notices only positive past experiences will not assume that he found the optimal policy but it will continue exploring better policies. To achieve this, Dec-HDRQN uses different learning rates $0 < \beta < \alpha < 1$ for the Temporal Difference(TD) error, as follows:

$$Q(s,a) = \begin{cases} Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] & \text{if TD-error is positive} \\ Q(s,a) + \beta[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] & \text{if TD-error is negative} \end{cases}$$

- To avoid shadowed equilibria, the authors exploit concurrent experience replay trajectories (CERTs). More specifically, the agents report their training samples, i.e. actions, rewards, observations, states, to a memory bank that stores this information for every agent and every task that each agent operated.

- Provided that the memory bank is accessible by all agents and they have already stored several training samples for each task, the agents can exploit this bank to learn the optimal policy. Thus, the agents sample several trajectories and exploit these trajectories to learn well-performing distilled policies.

- For the second phase of the learning process, each agent collects a mini-batch of trajectories that are stored in the memory bank. The mini-batch is not for a specific task only but it can contain trajectories from different tasks. In doing so, the agent can perform regression analysis over the Q-values for the mini-batch trajectories and learn a joint-policy for all the tasks.

To evaluate the performance of the proposed methodology, the authors used the multi-agent target capture game. In this game, multiple agents cooperate so that they can capture multiple other agents. The number of tasks defined by the grid size of the environment that the agents operate and the unique target that each agent is assigned. Note that different agents have different targets. During the experimentation, they used 3×3 to 7×7 grids with maximum 3 agents and 3 targets. The authors present that their approach outperforms significantly baseline approaches where CERTs are not applied and the agents do not learn joint policies.

Discussion

Despite the novelty of the proposed methodology and its significant performance against the baseline approaches, there are two main weaknesses. First, the CERTs memory bank is centralized which generates i) several scalability issues to the number of agents that can be trained, and ii) questions regarding the decentralized nature of the methodology. Replay memory has been used extensively in single agent reinforcement learning and has shown significant improvements. Therefore, it is obvious that CERTs for MARL is the main contribution of the paper. However, having a central memory bank presents limitation on the number of agents that can be trained. Thus, during the experimentation they number of agents used is limited to 3.

PAPER 2. FULLY DECENTRALIZED MULTI-AGENT REINFORCEMENT LEARNING WITH NETWORKED AGENTS

In contrast to the previous paper, where a central memory bank is required, this paper presents a fully decentralized MARL approach. In this paper, the agents rely on the communication between each other in order to learn the optimal policy. However, an all-in-all communication is unrealistic and in many real-world scenarios is unnecessary. Moreover, learning the optimal policy using a decentralized approach is a difficult problem. The main difficulty stems in the lack of convergence guarantees that the agents will learn the optimal policy without any central authority. Additionally, the communication between agents is performed using communication networks which often are unreliable and thus provide noise to the learning process.

To address the above problems, the main contributions of this paper are the following:

- The authors formulate the problem of fully decentralized MARL with network communication. Provided that the agents have limited communication, they model the networked MARL as a time-varying undirected graph, where the agents are the nodes and the connections between them are the edges.
- They propose two algorithms for decentralized MARL. Each algorithm is a variation of the centralized Actor-Critic algorithm. In the first algorithm, each agent updates the action-based policy of the Actor Critic (AC) Algorithm by observing the joint actions that each of his neighbor agents will execute in the future. Based on this information, the agent updates its Q-values and sends the updated weights to its neighbors so that they can update their own policy. Despite the efficiency of this approach, each agent requires the future action to compute the weights for the current policy, which might create bias based on the joint future actions. To avoid this problem, the authors proposed a second AC algorithm so as the agents to update the weights of their policy using the current state information that the agent and its neighbors are at each time.
- This is the first paper that theoretically prove the convergence of their proposed algorithms for the MARL domain.

To demonstrate the performance of the proposed methodology, they used the cooperative navigation problem, where N networked agents communicate each other in order to reach $|S|$ different states. Both the number of agents and number of states used for experimentation is 20. The authors report that both of their decentralized algorithms converge to the optimal policy and they resemble those learned by the relevant centralized AC algorithms.

Discussion

This paper presents the first fully decentralized MARL approach with convergence guarantees. However, while mathematically proving the convergence, the authors assumed that the network communication used to connect the different agents is always reliable. In real-world scenarios the network communication is dynamic which generates noise to the learning process. Thus, omitting this constraint from the proof is a weakness of this paper.

PAPER 3. MULTI-AGENT ADVERSARIAL INVERSE REINFORCEMENT LEARNING

In the previous selected papers, each agent optimized the learned policy using a common reward function that is designed by an expert. However, in many real-world MARL scenarios, each agent may require different reward function based on the task and the environment that it is operating. Having a well-defined common reward function for every agent creates bias to the learning process. Instead of defining multiple reward functions and manually assign one reward function to each agent, we would like the agents to learn their reward function by imitating the expert demonstrations.

To address the above problem, the main contributions of this paper are the following:

- They use imitation learning approaches to derive the optimal policy by observing the expert trajectories. More specifically, they apply the maximum entropy inverse reinforcement learning (MaxEnt IRL) which identifies the expert trajectory distribution with maximum entropy in order to match the reward expectation of the experts. However, MaxEntIRL has scalability problems, as many trajectories can match the same reward. Therefore, the authors solve the scalability problem by applying Adversarial IRL.
- To provide equilibrium between agent's optimal policy, the authors propose a new solution concept termed logistic stochastic best response equilibrium (LS-BRE). This allows the agents to better uncover the rationality of the expert demonstrations and parametrize reward functions for each agent while achieving concencus.
- Using the adversarial IRL, each agent learns a discriminator and a generator, where the discriminator classifies the expert and policy trajectories, and the generator mimics the expert trajectories. Thus, the discriminator learns the reward function that each agent requires by applying KL divergence between the experts trajectory distribution and that induced by the generator.

To evaluate the performance of the MA-AIRL methodology, the authors consider the following scenarios: i) Cooperative navigation, where multiple agents cooperate using physical actions in order to reach different landmarks, ii) cooperative communication, where multiple agents communicate in order to reach different landmarks, and iii) competitive keep-away where one agent tries to reach a landmark and a competitive agent tries to infer the landmark and prevent the first agent to reach it. In all scenarios, MA-AIRL achieve superior performance against baseline approaches.

Discussion

This paper provides a novel methodology for learning the reward function using imitation learning on MARL. As a strong point of this paper is that this approach is not well explored and they fact that they use LS-BRE instead of Nash equilibrium makes this paper novel. However, the fact that the authors exploit Adversarial Networks makes the methodology applicable to a single task. Extending the approach to multi-task domain, require different Adversarial networks which increases significantly the complexity of the methodology.