Cooperation in Multi-Agent Reinforcement Learning

Karush Suri, Dian Gadjov, Lacra Pavel
Department of Electrical & Computer Engineering, University of Toronto, Canada.
karush.suri@mail.utoronto.ca

Abstract

Advancements in Multi-Agent Reinforcement Learning (MARL) are motivated by cooperation in agents arising from Game Theory (GT). Agents must collaborate in practical scenarios in order to achieve complex objectives and attain strategies which depict optimal behavior. The need for cooperation is further highlighted in the case of partially-observed settings wherein agents have restricted access to environment observations. We revisit cooperation in MARL from the viewpoint of GT and stochastic dynamics of environments. The contributions of our work are twofold. (1) We analyze and demonstrate the effectiveness of cooperative MARL in the case of complex and partially-observed tasks consisting of high-dimensional action spaces and stochastic dynamics. (2) We leverage the empirical demonstrations to construct a novel optimization objective which addresses the detrimental effects of spurious states across agents. Our large-scale experiments carried out on the StarCraft II benchmark depict the effectiveness of cooperative MARL and our novel objective for obtaining optimal strategies under stochastic dynamics.

1 Introduction

Reinforcement Learning (RL) has seen tremendous growth in applications such as arcade games **?**, board games **??**, robot control tasks **??** and lately, real-time games **?**. The rise of RL has led to an increasing interest in the study of multi-agent systems **??**, commonly known as Multi-Agent Reinforcement Learning (MARL). MARL provides significant benefits in comparison to contemporary single-agent methods **?**. The Multi-Agent framework allows the modelling of complex real-world systems which consist of dynamic and large-scale interactions between multiple agents **?**. Additionally, MARL enables the learning of diverse strategies which are essential for executing a range of different tasks by the same set of agents.

In the case of partially observable settings, MARL enables the learning of strategies from a GT perspective by utilizing cooperation across agents**?**. Agents collaborate with each other in a given environment to optimize the cumulative payoffs by means of a single utility function. Optimization of the joint utility function leads to optimal behavior **??** in the long-horizon which is characterized by each agent executing its optimal strategy irrespective of other agents. Such a framework of learning strategies with collaborators and executing behaviors independently is often referred to as centralized training with decentralized control **?**.

The regime of decentralized control is hindered by intrinsic stochasticity in the environment. Spurious states are a common phenomenon observed in the case of single-agent RL methods. In the case of model-based RL **?**, agents build a model of the environment which learns the dynamics of the environment. Such a scheme is used as an effective planning tool in the case of long-horizon tasks **?**. In the case of model-free RL methods, environment stochasticity is addressed by utilizing robust utility functions **??** and effective exploration strategies **?**. On the other hand, MARL does not account for spurious states across agents as a result of which the system remains unaware of drastic changes in the environment **?**. Thus, addressing the learning of stochastic dynamics in the case of multi-agent settings requires attention from a critical standpoint.

We revisit cooperation in MARL from the perspective of GT and stochastic dynamics in the agents' environment. Our work assesses and demonstrates collaborative schemes in MARL under partially-observed settings which pose ill-conditioned objectives for the multi-agent system. More specifically, our twofold contributions are the following-

- We analyze and demonstrate the effectiveness of cooperative MARL for complex and partially-observed tasks consisting of high-dimensional action spaces and spurious states.
- We leverage the empirical demonstrations to construct a novel optimization objective which addresses the detrimental effects of spurious states across agents.

Our large-scale experiments carried out on the StarCraft II benchmark depict the effectiveness of cooperative MARL and our novel objective for obtaining optimal strategies under stochastic dynamics.

2 Related Work

Growing advances in GT have given rise to efficient MARL algorithms and implementations in stochastic scenarios. Thsi section highlights some of the main contributions in learning of stochastic games which have paved the way for Multi-Agent learning.

2.1 Learning in Games

Learning in games is an active area of development which is motivated by the framework of repeated games **?**. Repetitions of games are also modeled as episodes which has given rise to episodic play and continuous control in the case of single-agent systems. While episodic play serves as the basis for fictitous **?** and best-response type learning **?**, learning algorithms in games are primarily motivated by developments in the reinforcement play regime**?**. Learning algorithms are coupled with fast optimization techniques **?** to iterate over complex strategy spaces and achieve optimal behavior **?**. Additionally, developments in the learning regime such as the introduction of complex function approximators **?** for optimizing higher order utility functions has played a significant role in expanding computational capabilities of game theoretic learning **?**.

? demonstrates the large-scale suitability of reinforcement learning to single-agent learning by making use of Q-learning **??** which allows the agent to learn complex utility functions and generalize to different games **?** by making use of a common function approximator. Other methods in literature **????** have improved upon the Q-learning framework to provide stability **?** and diversity **?** in learning. These improvements have played a key role in yielding state-of-the-art performance **?** on real-world games **??** wherein the structure of payoff function is sparse **?** and the agent needs to explore a larger action space **?** in order to achieve optimal strategies.

2.2 Multi-Agent Learning

Most multi-agent methods are based on the paradigm of centralized training and decentralized control **???** wherein agents learn to collaborate **?** and optimize their utility function **?**. The fundamental work on MARL originates from the IQL **?** framework wherein agents learn to collaborate with independent utilities. While the IQL framework serves as a critical point for advances in MARL, the work of **?** presents the common knowledge framework wherein agents collaborate by gaining mutual information about the task and establishing a structured protocol for communication **?**. Such methods have given rise to large-scale agents capable of optimal behavior on high-dimensional control tasks **???**. Some of these methods suffer from estimation biases **??** stemming from the function approximator **?** used to maximize the utility function. Various MARL methods **?** make use of a dual function approximator approach which increases the accuracy of estimates. The Weighted Double Deep Q-Network (WDDQN) provides stability and sample efficiency for fully-observable settings. In the case of partially-observed scenarios, Weighted-QMIX (WQMIX) **?** yields a more sophisticated weighting scheme which aids in the retrieval of optimal strategy **?**.

Despite the recent success of RL **??** MARL agents suffer from spurious state spaces and encounter sudden changes in trajectories. These anomalous transitions between consecutive states are often termed as surprise **?**. Quantitatively, surprise can be inferred as a measure of deviation **??** among states encountered by the agent during its interaction with the environment. In the case of single-agent methods, surprise results in sample-inefficient learning **?**. This can be tackled by making use of rigorous exploration strategies **??**. However, such solutions do not show evidence for multiple agents consisting of individual partial observations **?**.

3 Preliminaries

3.1 Stochastic Markov Games

We revisit Stochastic Markov Games **?** which serve as the fundamental basis for MARL. A Markov Game **?** is a generalization of a Markov Decision Process (MDP) **?** which is described using the tuple $(\mathcal{S}, \mathcal{A}^1, \mathcal{A}^2 ... \mathcal{A}^n, r^1, r^2, ... r^n, N, P, \gamma)$ where $\mathcal{S}$ is the finite state space, $\mathcal{A}^a$ is the action space corresponding to agent a such that $a \in N$ where $N = \{1, 2, ...n\}$ is the set of all agents, $r^a : \mathcal{S} \times \mathcal{A}^a \rightarrow [r^a_{min}, r^a_{max}]$ is the payoff observed by agent a and bounded in $[r^a_{min}, r^a_{max}]$, $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2 \times ... \mathcal{A}^n \rightarrow [0, \infty)$ presents the unknown transition model consisting of the transition probabilities to the next state $s' \in \mathcal{S}$ given the current state $s \in \mathcal{S}$ and γ is the discount factor. Each agent a performs its own action u^a which gives rise to the joint action $u = \{u^{(1)}, u^{(2)}, ... u^{(n)}\}$. Analogously, the action space can be written as the combination of all agents' action spaces $\mathcal{A} : \mathcal{A}^1 \times \mathcal{A}^2 \times ... \mathcal{A}^n$. Markov Games wherein each agent observes its own payoffs are called General Markov Games (GMGs) **?**. On the other hand, Markov Games in which all agents observe the same payoffs $r^1 = r^2 = ... r^n = r$ such that $r : \mathcal{S} \times \mathcal{A} \rightarrow [r_{min}, r_{max}]$ are called Team Markov Games (TMGs) **?**. Thus, a TMG can be compactly defined as a tuple of the form $(\mathcal{S}, \mathcal{A}, r, N, P, \gamma)$. The general framework of cooperative multi-agent learning makes use of TMGs.

3.2 Multi-Agent Learning

We review the MARL setup. The problem is modeled as a Partially Observable and Stochastic TMG **?** defined by the tuple $(\mathcal{S}, \mathcal{A}, r, N, P, Z, O, \gamma)$ where the state space $\mathcal{S}$ and action space $\mathcal{A}$ are discrete, $r : \mathcal{S} \times \mathcal{A} \rightarrow [r_{min}, r_{max}]$ presents the payoff observed by agents $a \in N$ bounded in the interval $[r_{min}, r_{max}]$ where N is the set of all agents, $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0, \infty)$ presents the unknown transition model consisting of the transition probability to the next state $s' \in \mathcal{S}$ given the current state $s \in \mathcal{S}$ and joint action $u \in \mathcal{A}$ where $u = \{u^{(1)}_t, u^{(2)}_t ... u^{(n)}_t\}$ at time step t and γ is the discount factor. We consider a partially observable setting in which each agent a draws individual observations $z \in Z$ according to the observation function $O(s, u) : \mathcal{S} \times \mathcal{A} \rightarrow Z$. We consider a joint policy $\pi_\theta(u|s)$ which quantifies the probability of taking action u in state s as a function of the multi-agent model with its control parameters as θ. Standard RL defines the agent's objective to maximize the expected discounted payoff $\mathbb{E}_{\pi_\theta}[\sum_{t=0}^{T} \gamma^t r(s_t, u_t)]$ as a function of the parameters θ.

3.3 Q-Learning

We review the Q-learning setup in MARL. The action-value function, which is the expected sum of payoffs obtained in state s upon performing action u by following the policy π_θ, for an agent is represented in **??**.

$$Q(u, s; \theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=1}^{T} \gamma^t r(s, u) | s = s_t, u = u_t] \tag{1}$$

We denote the optimal policy π^*_θ such that $Q(u, s; \theta^*) \geq Q(u, s; \theta) \forall s \in S, u \in A$. In the case of multiple agents, the joint optimal policy can be expressed as the Nash Equilibrium **?** of the Stochastic TMG as expressed in **??**.

$$\begin{aligned} \pi^* = (\pi^{1,*}, \pi^{2,*}, ... \pi^{N,*}) \\ s.t. \quad Q(u^a, s; \theta^*) \geq Q(u^a, s; \theta) \\ \forall s \in S, u \in A, a \in N \end{aligned} \tag{2}$$

Q-Learning is an off-policy, model-free algorithm suitable for continuous and episodic tasks. The algorithm uses semi-gradient descent to minimize the Temporal Difference (TD) error expressed in **??**.

$$\mathbb{L}(\theta) = \mathop{\mathbb{E}}_{b \sim R}[(y - Q(u, s; \theta))^2] \tag{3}$$

Here $y = r + \gamma \max_{u' \in A} Q(u', s'; \theta^-)$ is the TD target consisting of θ^- as the target parameters and b is the batch of tuples (s, u, r, s') sampled from memory R.

4 Cooperation in Multi-Agent Learning

We assess and lay out the framework for cooperative multi-agent learning in this section. The setting of partially-observable states consisting of stochastic dynamics is discussed from an intuitive viewpoint followed by detailed learning mechanisms in state-of-the-art MARL algorithms.

4.1 The Partial Observability Setting

Assessing an agent's actions with unknown and spurious dynamics requires a partially-observable setting. In the case of a partially-observable TMG, the multi-agent system observes a common state s with each of its agents observing individual observations $z \in Z$. These individual observations serve as the agent's basis for selecting action a and optimizing the policy distribution $\pi(a_t|s_t)$.

The agent optimizes over its policy by maintaining a belief b over its actions and partial observations. Since the environment is Markovian, the belief of the agent b_{t+1} in state s_{t+1} depends on its belief b_t in previous state s_t. We select StarCraft II scenarios particularly for two reasons. Firstly, micromanagement scenarios consist of a larger number of agents with different action spaces. This requires a greater deal of coordination. Lastly, micromanagement scenarios in StarCraft II consist of multiple opponents which introduce a greater degree of surprise within consecutive states. Irrespective of the time evolution of an episode, environment dynamics of each scenario change rapidly as the agents need to respond to enemy's behavior.