Exploring the effects of conditioning Independent Q-Learners on the Sufficient Statistic for Dec-POMDPs

Alex Mandersloot[0000−0003−1617−2934], Frans Oliehoek[1][0000−0003−4372−5055], and Aleksander Czechowski[1][0000−0002−6054−9842]

[1] Department of Intelligent Systems, Delft University of Technology, Delft, The Netherlands

Abstract. In this study, we investigate the effects of conditioning Independent Q-Learners (IQL) not solely on the individual action-observation history, but additionally on the sufficient plan-time statistic for Decentralized Partially Observable Markov Decision Processes. In doing so, we attempt to address a key shortcoming of IQL, namely that it is likely to converge to a Nash Equilibrium that can be arbitrarily poor. We identify a novel exploration strategy for IQL when it conditions on the sufficient statistic, and furthermore show that sub-optimal equilibria can be escaped consistently by sequencing the decision-making during learning. The practical limitation is the exponential complexity of both the sufficient statistic and the decision rules.

Keywords: Deep Reinforcement Learning · Multi-Agent · Partial Observability · Decentralized Execution.

Introduction: The Decentralized Partially Observable Markov Decision Process (Dec-POMDP) is a widely used framework to formally model scenarios in which multiple agents must collaborate using private information. A key difficulty of a Dec-POMDP is that to coordinate successfully, an agent should decide on actions not only using its own action-observation history, but also by reasoning about the information that might be available to the other agents.

Independent Q-Learning (IQL) [1] is an easily-scalable multi-agent Reinforcement Learning method in which each agent concurrently learns the value of individual actions based on its individual information. It is well understood that such individual action-values are insufficient to capture the inter-agent dependency, and consequently IQL is not guaranteed to converge to the optimal joint policy. Instead, it is likely to converge to a joint policy that is in Nash Equilibrium [2]. However, such equilibria can be arbitrarily poor.

Precisely the obliviousness of IQL to the presence of other learning agents is our motivation for additionally conditioning IQL on the sufficient statistic for Dec-POMDPs [3], which contains a distribution over the joint action-observation history induced by the joint policy followed thus far. As a result, each agent is then equipped with an accurate belief over the local information available to the other agents, and is able to adjust its own behavior accordingly.

Experiments: We train a Deep Q-Network for each agent that conditions on the individual action-observation history $\overline{\theta}_t^i$ and the sufficient statistic σ_t, and learns the value of individual actions $Q_t^i(\overline{\theta}_t^i, \sigma_t, a_t^i)$. Methods are evaluated in the two agent Decentralized Tiger environment, whereby a horizon of 3 is employed.

To escape poor equilibria, an exploratory action of one agent should be observable to the others. To accomplish this, our agents explore in the space of entire *decision rules*. The sufficient statistic captures such decision rules, and thus facilitates the communication of exploratory actions among the agents. Importantly, however, the sufficient statistic summarizes only the *history* of joint decision rules. For *current* exploratory decision rules to be observable to others, we therefore additionally sequence the decision-making during learning. Specifically, agent 1 acts first and agent n is last to act. Each agent i then additionally conditions on the current (possibly exploratory) decision rules $\delta_t^{1:i-1}$ of the agents that acted before it to learn $Q_t^i(\overline{\theta}_t^i, \sigma_t, \delta_t^{1:i-1}, a_t^i)$. Our learners are able to consistently escape sub-optimal equilibria and learn the optimal policy, even when we explicitly force such equilibria upon the agents initially (Fig. 1).

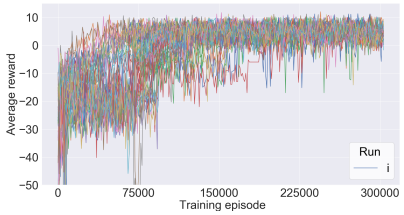

Fig. 1: All 50 learning curves.

Average Reward (std)	5.00 (0.77)
Ratio Optimal Policies	0.92

Table 1: Results across the 50 runs.

This project had received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 758824 —INFLUENCE).

References

1. Tan, M.. Multi-agent reinforcement learning: Independent vs. cooperative agents. In: Proceedings of the tenth international conference on machine learning. 1993. p. 330-337.
2. Boutilier, C.. Sequential optimality and coordination in multiagent systems. In: IJCAI. 1999. p. 478-485.
3. Oliehoek, F. A.. Sufficient plan-time statistics for decentralized POMDPs. In: Twenty-Third International Joint Conference on Artificial Intelligence. 2013.