Exploring the effects of conditioning Independent Q-Learners on the sufficient plan-time statistic for Dec-POMDPs

Alex Mandersloot, Frans A. Oliehoek, Aleksander Czechowski

TU Delft
Delft University of Technology

Introduction

Reinforcement Learning is a general technique to learn to act in different environments.

One particular set of environments, the **Decentralized Partially Observable Markov Decision Process** (Dec-POMDP), is especially hard due to the presence of multiple agents and private information amongst them, e.g. a hand of cards.

Independent Q-Learning (IQL) combats the exponential scale-up in the number of agents by letting each agent optimize its individual Q-function independently and concurrently. Each agent thus completely ignores the inter-agent dependencies.

IQL is likely to converge to a **Nash Equilibrium**, with no guarantees about the quality of such an equilibrium: it is not necessarily optimal.

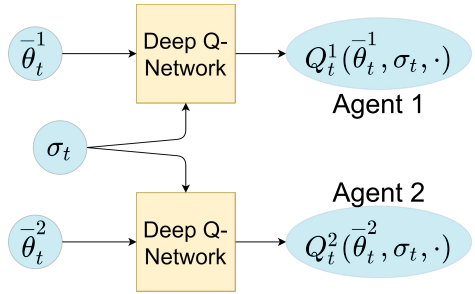
The **Sufficient Statistic** for Dec-POMDPs contains a distribution over the joint action-observation history induced by the joint policy thus far.

Methodology

Can IQL escape sub-optimal equilibria by conditioning on the sufficient statistic?

By conditioning IQL on the sufficient statistic, each agent is equipped with knowledge about the private information of others. They can then more accurately predict others' behavior and adjust their own accordingly.

Each agent learns a mapping from its individual action-observation history $\overline{\theta}_t^i$ and sufficient statistic σ_t to individual action-values.



Findings

Agents must **explore in the space of decision rules** in order for the sufficient statistic to capture exploratory actions.

The sufficient statistic summarizes the *history* of joint decision rules. To escape sub-optimal equilibria, agents must know the (possibly exploratory) individual decision rules followed *at the current timestep*.

By **sequencing the decision-making** during learning we can additionally condition each agent on the *current* decision rules of the previous agents. In doing so, an exploratory action is immediately observable to the others

Agents then consistently escape sub-optimal equilibria and learn the optimal policy.