Contribution ID: **39** Type: **not specified**

Exploration in reward machines with near-optimal regret

Monday, June 17, 2024 2:00 PM (30 minutes)

We study reinforcement learning for decision processes with Markovian dynamics but non-Markovian rewards, in which high-level knowledge in the form of a finite-state automaton is available to the learner. Such an automaton, often called Reward Machine (RM) (Toro Icarte et al., 2018), generates rewards based on its internal state as well as events that are detected at various states in the environment. The associated decision processes is called an MDPRM, and we focus on average-reward MDPRMs in the regret setting. For a given MDPRM, there is an equivalent cross-product MDP, to which one can apply provably efficient off-the-shelf algorithms obliviously to the structure induced by the MDPRM. However, this would lead to a large regret in view of the large state-space of the cross-product MDP. We establish a first regret lower bound for MDPRMs and present a model-based algorithm that efficiently exploits the structure in MDPRMs, and analyze its regret non-asymptotically. Like the lower bound, our bound is independent of Q, the number of RM states. Further, it improves over regret bound of the existing baselines (e.g., UCRL2 (Jaksch et al., 2010) applied to the cross-product MDP) by up to a factor of Q^{3/2}. Our regret bound makes appear a notion of diameter in MDPRMs, where we show that it can be smaller by a factor of Q than conventional diameter thereof. Finally, we report numerical experiments that demonstrate the superiority of the proposed algorithm over existing baselines in practice.

Primary author: Dr TALEBI, Mohammad Sadegh (University of Copenhagen)

Presenter: Dr TALEBI, Mohammad Sadegh (University of Copenhagen)

Session Classification: Parallel session: Challenges and progress in statistical reinforcement learning