Inferring Probabilistic Reward Machines from Non-Markovian Reward Signals for Reinforcement Learning

T. Dohmen[1], N. Topper[2], G. Atia[2] , A. Beckus[3], A. Trivedi[1], A. Velasquez[3]

[1] University of Colorado – Boulder, [2] University of Central Florida, [3] Air Force Research Laboratory

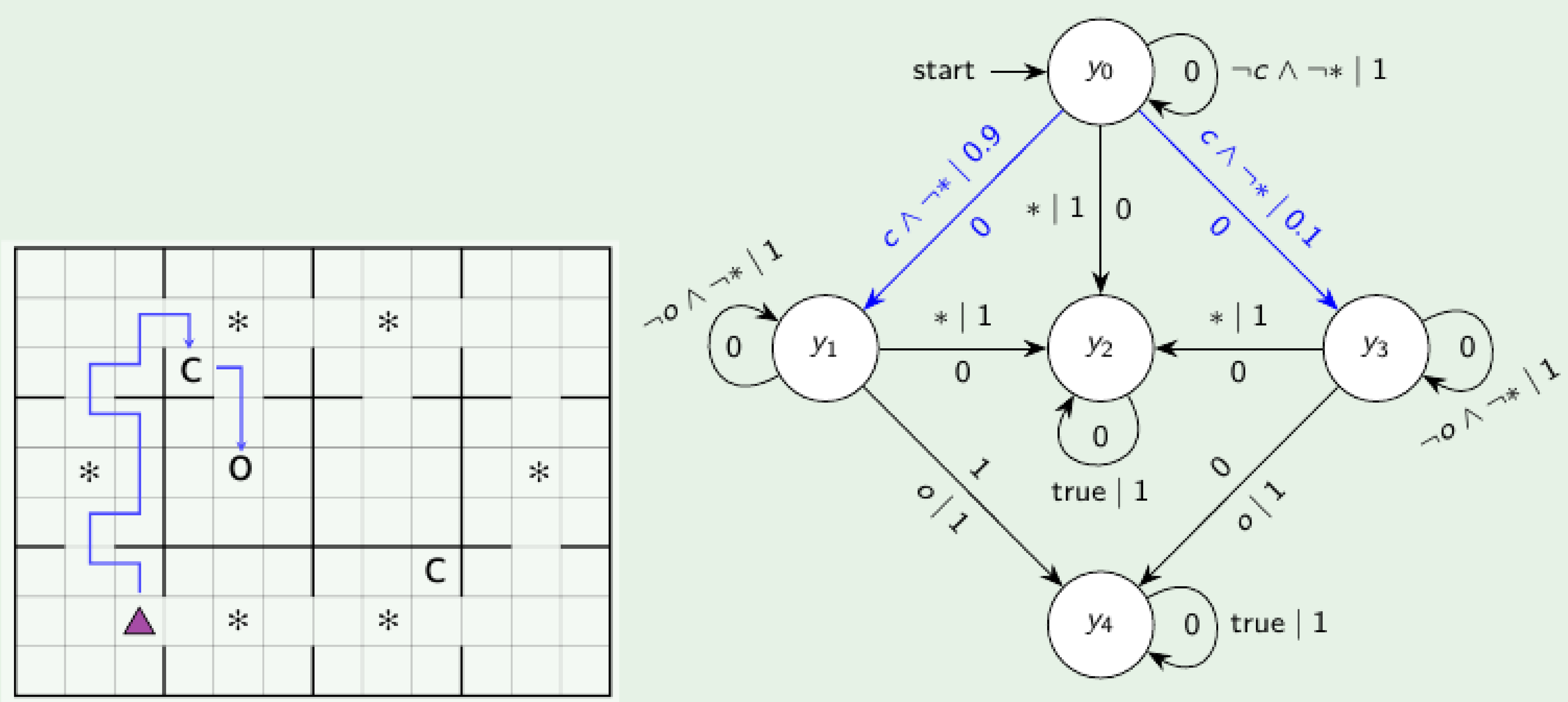
Summary

- Reinforcement learning (RL) is typically predicated on the assumption that the reward signal is Markovian, i.e. depends only on the current state and action.
- Reward machines have emerged as a structured representation based on the theory of finite automata of non-Markovian reward signals.
- We introduce and study probabilistic reward machines (PRMs) as representations of non-Markovian stochastic reward signals.

Results

- We prove that the product of a decision process and a PRM is a decision process with Markovian reward (MDP).
- We formulate an inference procedure for learning PRMs that combines a sampling based variant of the L* algorithm with an RL-driven sampling tehcnique.
- We show that the algorithm coverges to a PRM encoding of the target reward function, assuming the reward function is sufficiently regular.
- We prove an exponential upper bound on the state-space blowup for simulating a product MDP T x H of a decision process T and a PRM H by a product MDP U x J where the randomness of H is embedded in U and the non-Markovian dynamics of H preserved in J a deterministic reward machine.

Example (Office gridworld and probabilistic reward machine)



Combining $L^\star$ and RL for Learning PRMs