Joint Inference of Reward Machines and Policies for Reinforcement Learning

Zhe Xu[1], Ivan Gavran[2], Yousef Ahmad[1], Rupak Majumdar[2], Daniel Neider[2], Ufuk Topcu[1], Bo Wu[1]

[1]University of Texas at Austin, Austin, TX

[2]Max Planck Institute for Software Systems, Kaiserslautern, Germany

Abstract

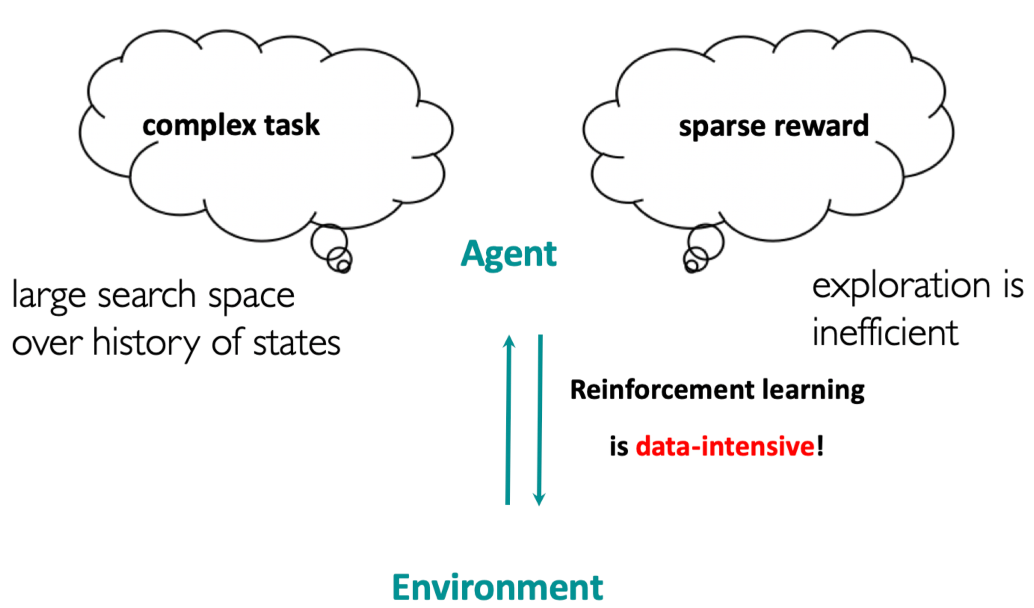
We investigate an RL problem where the high-level knowledge is in the form of reward machines, i.e., a type of Mealy machine that encodes non-Markovian reward functions.

We present an iterative algorithm that performs joint inference of reward machines and policies (JIRP) for RL (more specifically, q-learning).

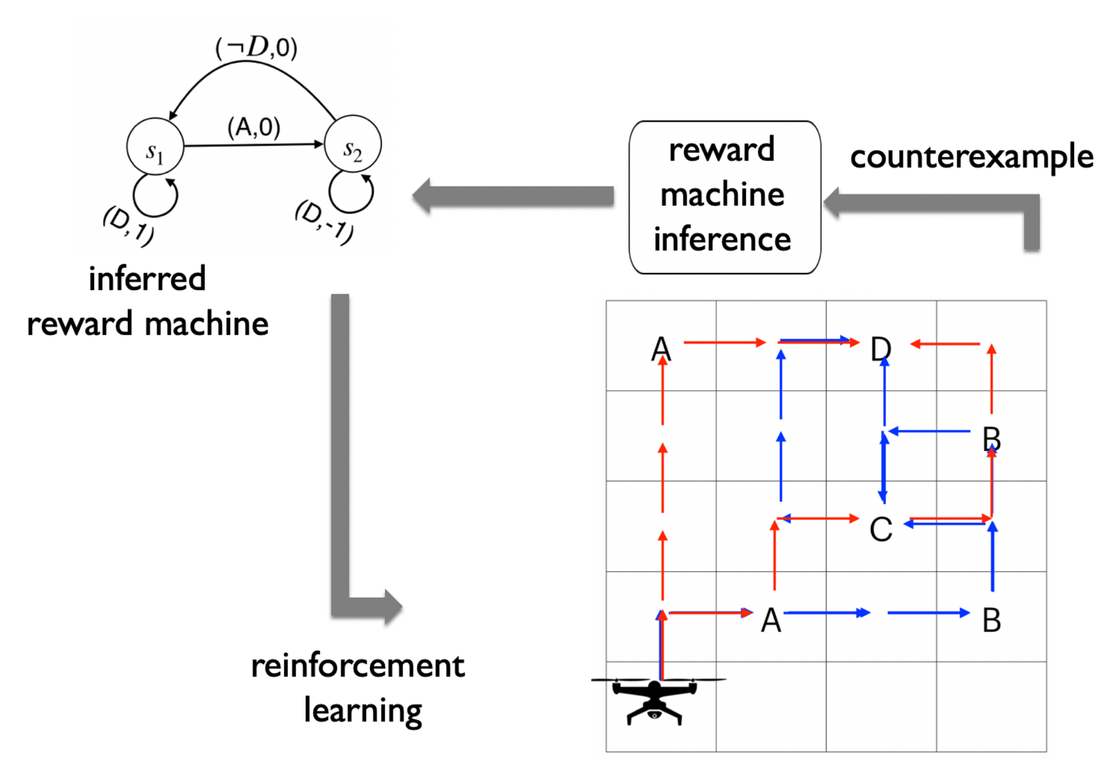
We implement the proposed JIRP approach and compare it with three baseline methods. We evaluated them in three scenarios: an autonomous vehicle scenario, an office world scenario, and a Minecraft world scenario, running a set of tasks for every scenario.

The experiments show that the proposed JIRP approach significantly outperforms the three baseline algorithms in all three scenarios.

Challenges in RL



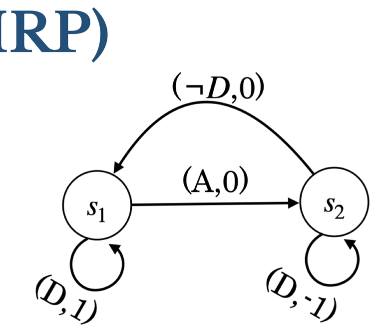
Methodology



The Algorithm (JIRP)

initialize H, Q, X;
repeat
- $(\lambda, \rho, Q) = \mathrm{QRM}(H, Q)$;
- **if** $H(\lambda) \neq \rho$ **then**
  - add (λ, ρ) to X;
  - $H = \mathrm{infer}(X)$;
  - re-initialize Q;

H:



Q: $\{q^{s_1}, q^{s_2}\}$

X: $\{$ A..B..C..D, 0..0..0..1 ; 0..-1, A..D $\}$

Theorem

Let $\mathcal{M}$ be a labeled Markov decision process and $\mathcal{A}$ be the ground truth reward machine for $\mathcal{M}$. Then, JIRP with episode length of at least $2^{|\mathcal{M}|+1}(|\mathcal{A}|+1)-1$ almost surely converges to an optimal policy in the limit.

Theorem

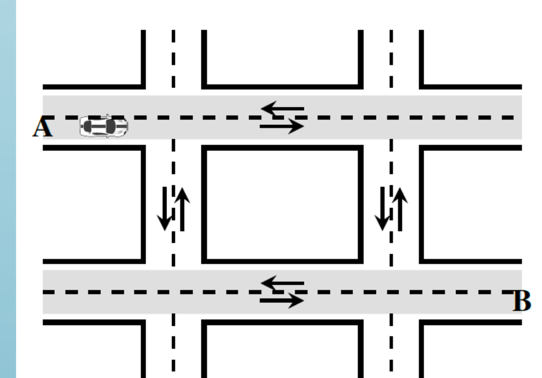
Let $\mathcal{M}$ be a labeled Markov decision process and $\mathcal{A}$ be the ground truth reward machine for $\mathcal{M}$. Then, JIRP with episode length of at least $2^{|\mathcal{M}|+1}(|\mathcal{A}|+1)-1$ almost surely learns a reward machine in the limit that is equivalent to $\mathcal{A}$ for all possible trajectories from the labeled MDP $\mathcal{M}$.

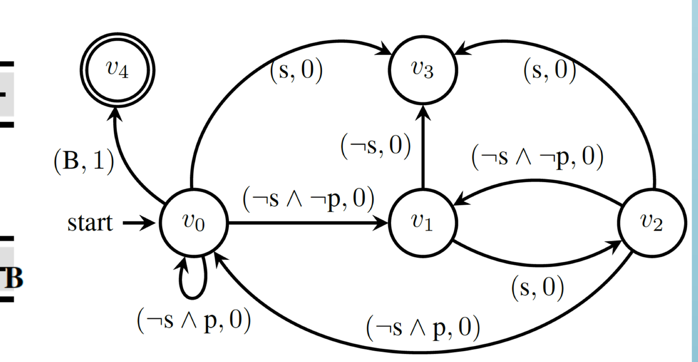
Optimizations

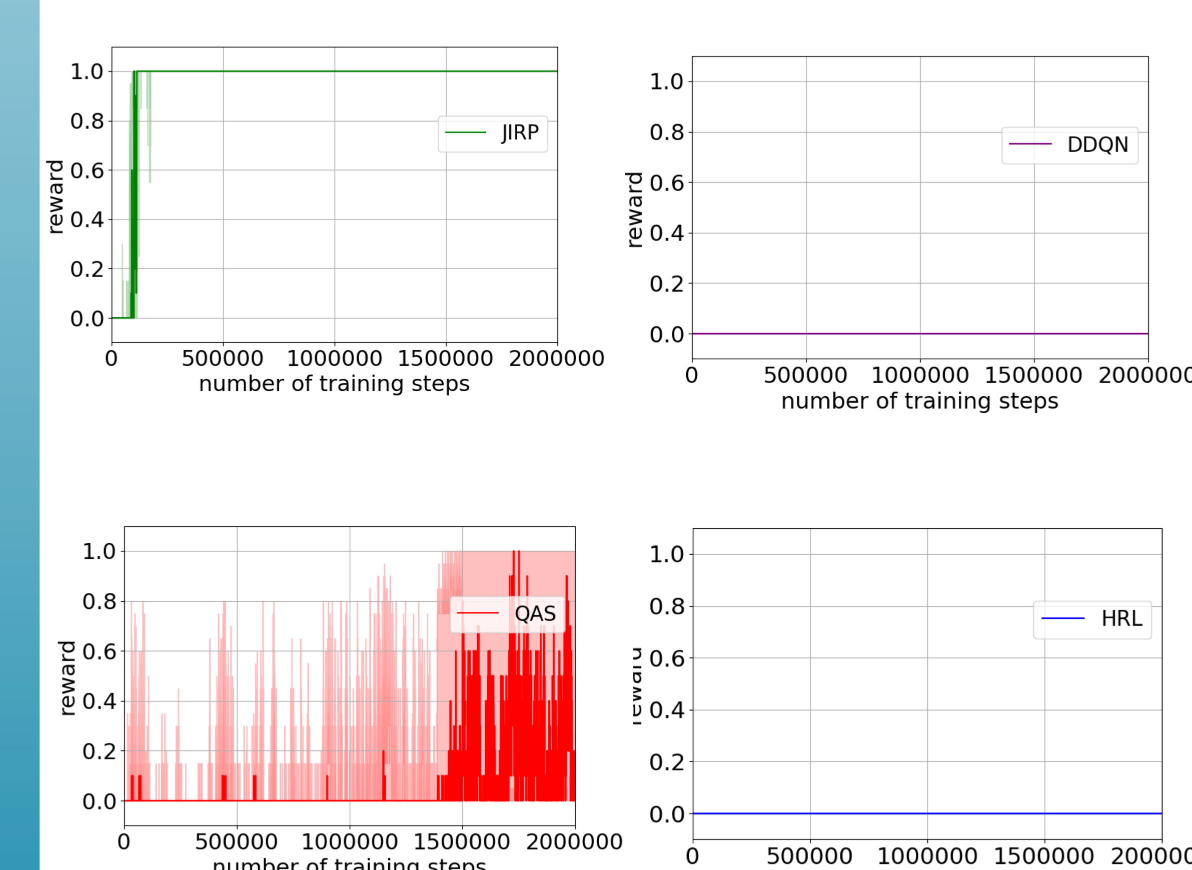
1. Batching of counterexamples
2. Transfer of q-functions

Results

autonomous vehicle scenario



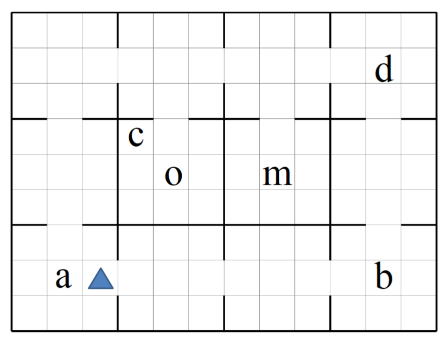


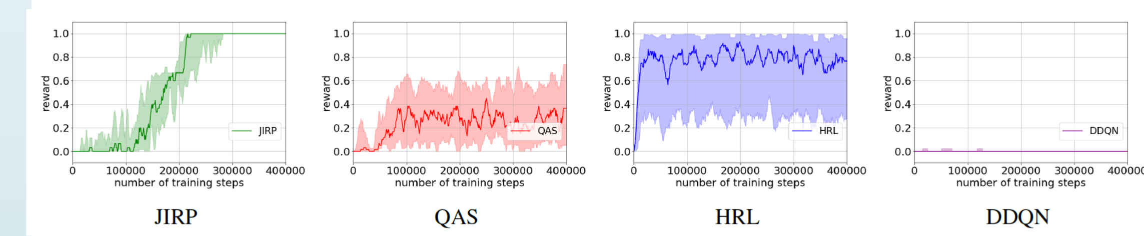


Sampling efficiency can be improved by at least 20 times!

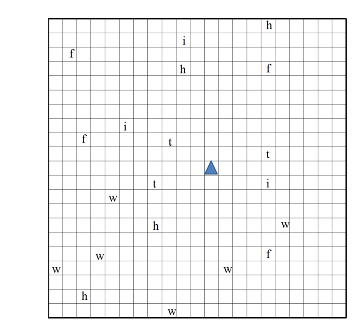
office world scenario

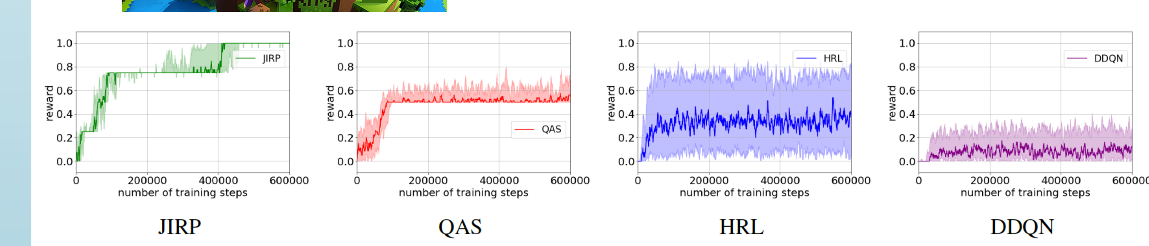
Task: get mail at m, then deliver the mail to the office o, finally go to the front desk d





Minecraft world scenario





Related Work

Icarte et al., NeurIPS 2019	Partially observable MDPs (non-Markovian dynamics)	Reward machines	Constrained optimization of the likelihood of seen observations
Gaon & Brafman, AAAI 2020	Non-Markovian rewards	Deterministic finite-state automata (DFA)	Both active and passive automata-learning methods
Furelos-Blanco et al., AAAI 2020	Learning subgoals	Subgoal automata (DFA)	Inductive logic programming

Conclusion

We proposed an iterative approach that alternates between reward machine inference and RL for the inferred reward machine.

Future work:
First, we are interested in exploring a scenario in which the reward machine is not known, but a number of hints (properties) about the reward are given.

Second, we will explore methods that can infer the reward machines incrementally.

Finally, the method to transfer the q-functions between equivalent states of reward machines can be also used for transfer learning between different tasks where the reward functions are encoded by reward machines.

Acknowledgements

This work is supported in part by grants DFG 389792660-TRR 248, DFG 434592664, ERC 610150, DARPA D19AP00004, ONR N000141712623, and NASA 80NSSC19K0209.