PAC-MDP Learning with Knowledge-based Admissible Models

Marek Grześ and Daniel Kudenko
Department of Computer Science, University of York, United Kingdom

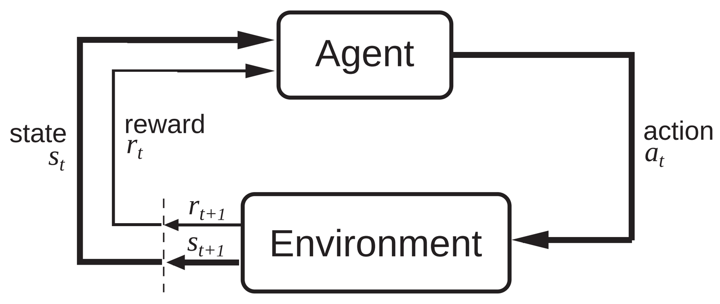
THE UNIVERSITY of York
DEPARTMENT OF COMPUTER SCIENCE

Introduction

- Reinforcement learning suffers scalability problems due to state space explosion and the temporal credit assignment problem
  - a knowledge-based approach to reinforcement learning is desirable
  - relation between *uninformed search* ⟶ *informed search* is like relation between *basic RL* ⟶ *RL with knowledge*
- In this work:
  - we are looking for new ways of incorporating domain knowledge (heuristics) into reinforcement learning algorithms
  - knowledge and methods of using knowledge, which preserve theoretical properties of PAC-MDP learning are sought

Reinforcement Learning

- The loop of interaction:
  - Agent can see the current state of the environment
  - Agent chooses an action
  - State of the environment changes, agent receives reward or punishment



- The goal of learning: quickly learn the policy that maximises the long-term expected reward

Exploration-Exploitation Trade-off

- We have found a reward of 100. Is it the best reward which can be achieved?
- **Exploitation**: should I stick to the best reward which was found? But, there may still be a high reward undiscovered
- **Exploration:** should I try more new actions to find a region with a higher reward? But, a lot of negative reward may be collected while exploring unknown actions

PAC-MDP Learning

- While learning the policy, also learn the model of the environment
- Assume that all unknown actions lead to a state with a highest possible reward, R-max
- This approach has been proven to be PAC, i.e., the number of suboptimal decisions is bounded polynomially by relevant parameters

Problem Formulation

- PAC-MDP learning vs. heuristic search
  - Heuristic search is efficient when used with good informative heuristics (knowledge)
  - It is useful and desirable to transfer this idea to reinforcement learning
- Existing literature shows how admissible heuristics can improve PAC-MDP learning via reward shaping (Asmuth, Littman & Zinkov 2008)
- In this work, we are looking for alternative ways of incorporating knowledge (heuristics) into reinforcement learning algorithms
  - Different knowledge (global admissible heuristics may not be available)
  - Different ways of using knowledge (potentially more efficient than reward shaping)
  - We want to guarantee that the algorithm remains PAC-MDP

Determinisation in Symbolic Planning

- Action representation: Probabilistic Planning Domain Description Language (PPDDL)

$$(\mathbf{a}\ \mathbf{p_1}\ \mathbf{e_1}\ \ldots\ \mathbf{p_n}\ \mathbf{e_n})$$

- Determinisation (probabilities known but ignored), e.g., FF-Replan, P-Graphplan
- In reinforcement learning probabilities are not known anyway

(1) All-outcomes (AO) Determinisation

- Available knowledge: all outcomes $\mathbf{e_i}$ of each action, $\mathbf{a}$

$$(\mathbf{a}\ \mathbf{p_1}\ \mathbf{e_1}\ \ldots\ \mathbf{p_n}\ \mathbf{e_n})$$

- Create a new MDP $\hat{\mathbf{M}}$ in which there is a deterministic action $\mathbf{a_d}$ for each possible effect, $\mathbf{e_i}$, of a given action $\mathbf{a}$
- For any state $\mathbf{s}$ and action $\mathbf{a}$, the condition $\hat{\mathbf{Q}}(\mathbf{s},\mathbf{a}) \geq \mathbf{Q}^*(\mathbf{s},\mathbf{a})$ is satisfied after value iteration on the MDP $\hat{\mathbf{M}}$ which is obtained from all-outcomes determinisation

(2) Free Space Assumption (FSA)

- Available knowledge: intended (which is either most probable or completely blocked) outcome $\mathbf{e_i}$ of each action, $\mathbf{a}$. If the intended outcome is blocked, then all remaining outcomes, $\mathbf{e_i}$, of a given action are most probable outcomes of different actions

$$(\mathbf{a}\ \mathbf{p_1}\ \mathbf{e_1}\ \ldots\ \mathbf{p_n}\ \mathbf{e_n})$$

- Create a new MDP $\hat{\mathbf{M}}$ in which each action, $\mathbf{a}$, is replaced by its intended outcome
- For any state $\mathbf{s}$ and action $\mathbf{a}$, the condition $\hat{\mathbf{V}}(\mathbf{s}) \geq \mathbf{V}^*(\mathbf{s})$ is satisfied after value iteration on the MDP $\hat{\mathbf{M}}$ which is obtained from FSA determinisation

PAC-MDP Learning with Admissible Models: Our Approach

- Rmax
  - If (s,a) not known (i.e., $\mathbf{n(s,a)} < \mathbf{m}$): use Rmax
  - if (s,a) known (i.e., $\mathbf{n(s,a)} \geq \mathbf{m}$): use estimated model
- Our approach
  - If (s,a) not known (i.e., $\mathbf{n(s,a)} < \mathbf{m}$): use the knowledge-based admissible model
  - if (s,a) known (i.e., $\mathbf{n(s,a)} \geq \mathbf{m}$): use estimated model

Results

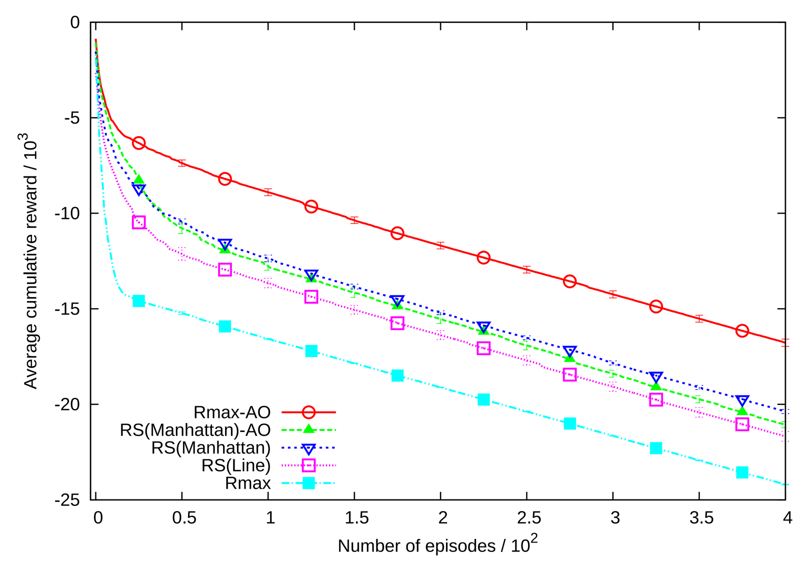


Figure: Results on a $\mathbf{25} \times \mathbf{25}$ maze domain. AO knowledge and $\varrho = \mathbf{1}$

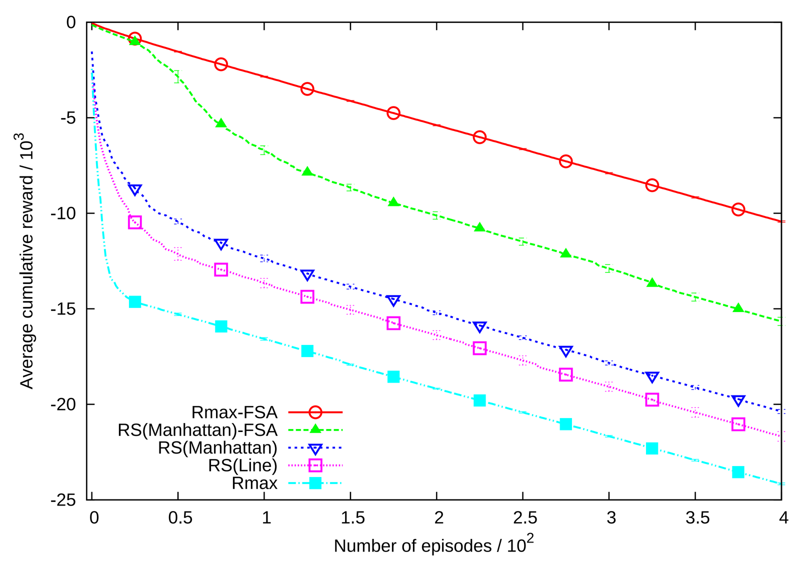


Figure: Results on a $\mathbf{25} \times \mathbf{25}$ maze domain. FSA knowledge and $\varrho = \mathbf{1}$

Comparing with the Bayesian Exploration Bonus Algorithm

- Bayesian Exploration Bonus (BEB) approximates Bayesian exploration (Kolter & Ng 2009)
  - (+) It can use action knowledge (AO and FSA) via informative priors
  - (-) It is not PAC-MDP
- Our approach shows how to use this knowledge with PAC-MDP algorithms
- Comparing BEB using informative priors with our approach using knowledge-based models

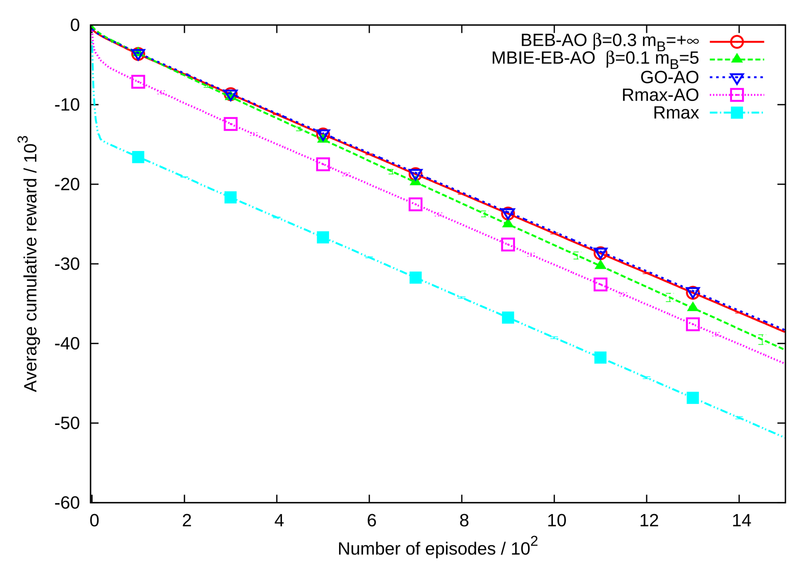


Figure: Results on a $\mathbf{25} \times \mathbf{25}$ maze domain. AO knowledge and $\varrho = \mathbf{0.8}$

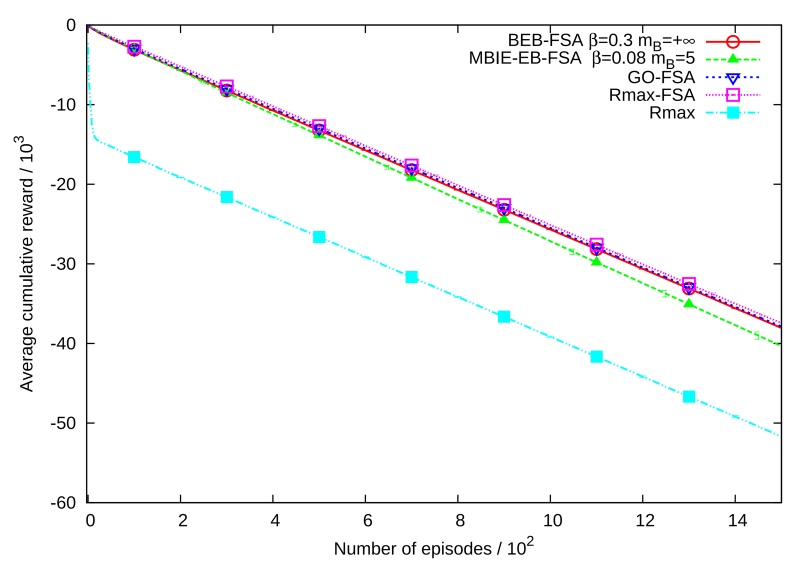


Figure: Results on a $\mathbf{25} \times \mathbf{25}$ maze domain. FSA knowledge and $\varrho = \mathbf{0.8}$

Conclusion

- The use of knowledge in RL is important
- It was shown how to use partial knowledge about actions with PAC-MDP algorithms in a theoretically correct way
- Global admissible heuristics required by reward shaping may not be available (e.g., PPDDL domains)
- Knowledge-based admissible models turned out to be more efficient than reward shaping with equivalent knowledge: in our case knowledge is used when actions are still 'unknown', whereas reward shaping helps only with known actions
- BEB can use AO and FSA knowledge via informative priors. It was shown how to use this knowledge in the PAC-MDP framework (BEB is not PAC-MDP)