PAC-MDP Learning with Knowledge-based Admissible Models

Marek Grześ and Daniel Kudenko

Department of Computer Science

The University of York

United Kingdom

AAMAS 2010

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
- 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
  - Default R-max ‘is like’ best-first search (i.e., A*) with a trivial heuristic h(s)=0
  - Heuristic search is efficient when used with good informative heuristics
  - It is useful and desirable to transfer this idea to reinforcement learning

Problem Formulation ctd

- 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 (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)

$$(a\ p_1\ e_1\ \dots\ p_n\ e_n)$$

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

All-outcomes (AO) Determinisation

- Available knowledge: all outcomes e_i of each action, a.

$$(a\ p_1\ e_1\ \ldots\ p_n\ e_n)$$

- Create a new MDP $\hat{M}$ in which there is a deterministic action a_d for each possible effect, e_i, of a given action a.
- The value function of a new MDP, $\hat{M}$, is admissible, i.e., $\hat{V}(s) \geq V^*(s)$

Free Space Assumption (FSA)

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

$$(a\ p_1\ e_1\ \ldots\ p_n\ e_n)$$

- Create a new MDP $\hat{M}$ in which each action, a, is replaced by its intended outcome.
- The value function of a new MDP, $\hat{M}$, is admissible, i.e., $\hat{V}(s) \geq V^*(s)$

PAC-MDP Learning with Admissible Models

- Rmax
  - If (s,a) not known (i.e., $n(s, a) < m$): use Rmax
  - if (s,a) known (i.e., $n(s, a) \geq m$): use estimated model

PAC-MDP Learning with Admissible Models

- Rmax
  - If (s,a) not known (i.e., $n(s,a) < m$): use Rmax
  - if (s,a) known (i.e., $n(s,a) \geq m$): use estimated model

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

Results

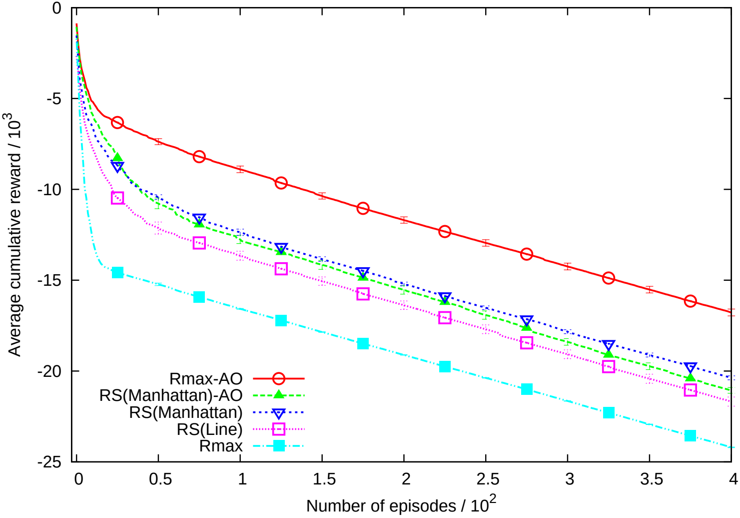

Figure: Results on a 25 × 25 maze domain. AO knowledge.

Results

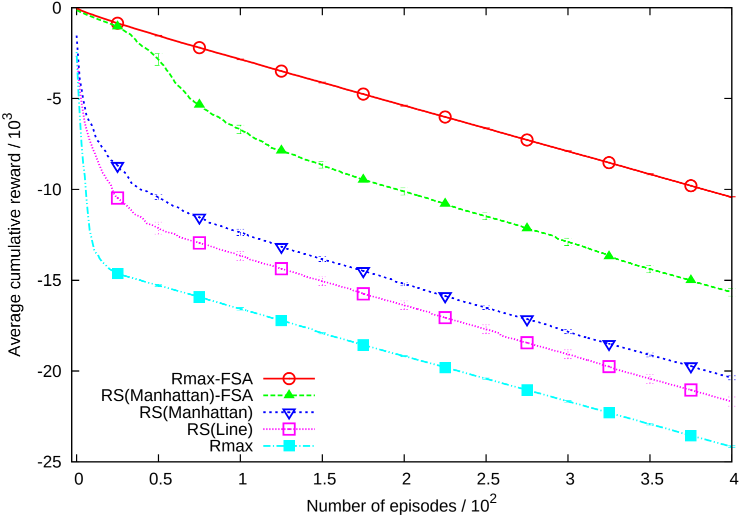

Figure: Results on a 25×25 maze domain. FSA knowledge.

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 (see our paper).

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).