Inducing Partially Observable Markov Decision Processes

Michael L. Littman MLITTMAN@CS.BROWN.EDU
Department of Computer Science, Brown University, Providence, RI

Editors: Jeffrey Heinz, Colin de la Higuera and Tim Oates

In the field of reinforcement learning (Sutton and Barto, 1998; Kaelbling et al., 1996), agents interact with an environment to learn how to act to maximize reward. Two different kinds of environment models dominate the literature—Markov Decision Processes (Puterman, 1994; Littman et al., 1995), or MDPs, and POMDPs, their Partially Observable counterpart (White, 1991; Kaelbling et al., 1998). Both consist of a Markovian state space in which state transitions and immediate rewards are influenced by the action choices of the agent. The difference between the two is that the state is directly observed by the agent in MDPs whereas agents in POMDP environments are only given indirect access to the state via "observations".

This small change to the definition of the model makes a huge difference for the difficulty of the problems of learning and planning. Whereas computing a plan that maximizes reward takes polynomial time in the size of the state space in MDPs (Papadimitriou and Tsitsiklis, 1987), determining the optimal first action to take in a POMDP is undecidable (Madani et al., 2003). The learning problem is not as well studied, but algorithms for learning to approximately optimize an MDP with a polynomial amount of experience have been created (Kearns and Singh, 2002; Strehl et al., 2009), whereas similar results for POMDPs remain elusive.

A key observation for learning to obtain near optimal reward in an MDP is that inducing a highly accurate model of an MDP from experience can be a simple matter of counting observed transitions between states under the influence of the selected actions. The critical quantities are all directly observed and simple statistics are enough to reveal their relationships. Learning in more complex MDPs is a matter of properly generalizing the observed experience to novel states (Atkeson et al., 1997) and can often be done provably efficiently (Li et al., 2011).

Inducing a POMDP, however, appears to involve a difficult "chicken-and-egg" problem. If a POMDP's structure is known, it is possible to keep track of the likelihood of occupying each Markovian state at each moment of time while selecting actions and making observations, thus enabling the POMDP's structure to be learned. But, if the POMDP's structure is not known in advance, this information is not available, making it unclear how to collect the necessary statistics. Thus, in many ways, the POMDP induction problem has elements in common with grammatical induction. The hidden states, like non-terminals, are important for explaining the structure of observed sequences, but cannot be directly detected.

Several different strategies have been used by researchers attempting to induce POMDP models in the context of reinforcement learning. The first work that explicitly introduced the POMDP model and algorithms for learning them came from Chrisman and McCallum in the 1990s (Chrisman, 1992; McCallum, 1993, 1994, 1995). They explored a variety of approaches beginning with Expectation-Maximization (EM). EM is an extremely natural approach to this problem as a POMDP is essentially a controlled Hidden Markov Model (HMM) and EM is one of the best studied approaches to learning in HMMs (Rabiner, 1989). However, the tendency of EM to get trapped in local optima led to considering instance-based approaches as an alternative. While showing some promise, none of these approaches has been broadly deployed.

A series of papers in the 2000s attempted to improve on these earlier results by changing the nature of the model being learned. Sutton and colleagues hypothesized that a state representation based solely on observables could avoid the problem of learning parameters on purely unobservable quantities. They introduced predictive state representations, or PSRs (Littman et al., 2002; Singh et al., 2003, 2004), showed they could capture all of the environments representable by POMDPs and evaluated several local search algorithms for inducing PSRs from data. While PSR learning algorithms seem to be more consistent than algorithms based on EM, the set of learnable environments in practice appears to be roughly the same.

Some more recent work observed that different POMDPs exhibit different size representations. A POMDP that is very complicated to write down in the learner's target representation is unlikely to be learned effectively. On the other hand, systematic enumeration and testing of models in order of increasing complexity of representation (Walsh and Littman, 2007; Zhang et al., 2012) allows POMDPs to be induced in time exponential in the size of its representation. This class of algorithms is an advance over earlier work because it provides a global search method that cannot become stuck in local optima. However, the exponential running time guarantees that the method cannot be applied beyond the tiniest problems.

In spite of the many insights of this community of researchers, POMDP induction remains a very challenging problem.

References

Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally weighted learning for control. *Artificial Intelligence Review*, 11:75–113, 1997.

Lonnie Chrisman. Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 183–188, San Jose, California, 1992. AAAI Press.

Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Michael J. Kearns and Satinder P. Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002.

Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: A framework for self-aware learning. *Machine Learning*, 82(3):399–443, 2011.

Michael L. Littman, Thomas L. Dean, and Leslie Pack Kaelbling. On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence (UAI–95)*, pages 394–402, Montreal, Québec, Canada, 1995.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems 14*, pages 1555–1561, 2002.

Omid Madani, Steve Hanks, and Anne Condon. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, 147(1–2): 5–34, 2003.

R. Andrew McCallum. Overcoming incomplete perception with utile distinction memory. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 190–196, Amherst, Massachusetts, 1993. Morgan Kaufmann.

R. Andrew McCallum. Instance-based state identification for reinforcement learning. In *Neural Information Processing Systems 6 (NIPS)*, pages 377–384, 1994.

R. Andrew McCallum. Instance-based utile distinctions for reinforcement learning with hidden state. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 387–395, San Francisco, CA, 1995. Morgan Kaufmann.

Christos H. Papadimitriou and John N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, August 1987.

Martin L. Puterman. *Markov Decision Processes—Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., New York, NY, 1994.

Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

Satinder Singh, Michael L. Littman, Nicholas K. Jong, David Pardoe, and Peter Stone. Learning predictive state representations. In *The Twentieth International Conference on Machine Learning (ICML-2003)*, pages 712–719, 2003.

Satinder Singh, Michael R. James, and Matthew R. Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI)*, pages 512–519, 2004.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* The MIT Press, 1998.

Thomas J. Walsh and Michael L. Littman. A multiple representation approach to learning dynamical systems. In *Computational Approaches to Representation Change During Learning and Development: AAAI Fall Symposium*, 2007.

Chelsea C. White, III. Partially observed Markov decision processes: A survey. *Annals of Operations Research*, 32, 1991.

Zongzhang Zhang, XiaoPing Chen, and Michael Littman. Covering number as a complexity measure for POMDP planning and learning. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1853–1859, 2012.