Bayesian RL in Factored POMDPs

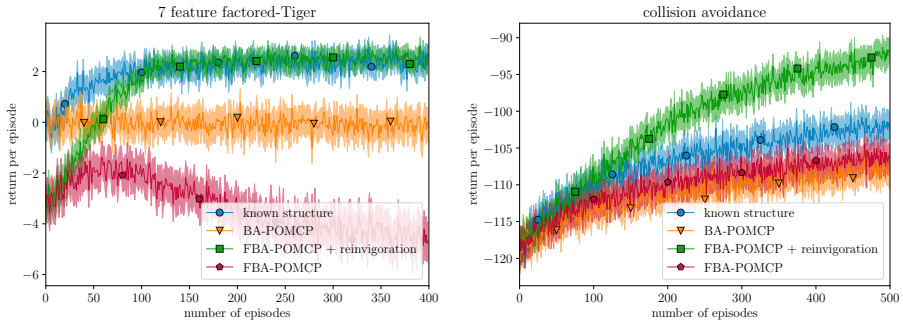
Sammie Katt[1], Frans Oliehoek[2], and Chris Amato[1]

[1] Northeastern University, USA
[2] TUDelft, Netherlands

Introduction: Robust decision-making agents in any non-trivial system must reason over uncertainty of various types such as action outcomes, the agent's current state and the dynamics of the environment. The outcome and state uncertainty are elegantly captured by the Partially Observable Markov Decision Processes (POMDP) framework [1], which enable reasoning in stochastic, partially observable environments. POMDP solution methods, however, typically assume complete access to the system dynamics, which unfortunately are often not available. When such a model is not available, model-based Bayesian Reinforcement Learning (BRL) methods explicitly maintain a posterior over the possible models of the environment, and use this knowledge to select actions that, theoretically, trade off exploration and exploitation optimally. However, few of the BRL methods are applicable to partial observable settings, and those that are, have limited scaling properties. The Bayes-Adaptive POMDP (BA-POMDP) [4], for example, models the environment in a tabular fashion, which poses a bottleneck for scalability. Here, we describe previous work [3] that proposes a method to overcome this bottleneck by representing the dynamics with Bayes Network, an approach that exploits structure in the form of independence between state and observation features.

Contribution: We introduce the Factored Bayes-Adaptive POMDP (FBA-POMDP) that allows the agent to learn and exploit structure in the environment which, if solved optimally, is guaranteed to be as sample efficient as possible. The FBA-POMDP considers the unknown dynamics as part of the hidden state of a larger *known* POMDP, effectively casting the learning problem into a planning problem. A solution to this task consists of a method for maintaining the belief and a policy that picks actions with respect to this belief. Both parts are non-trivial due to the large state space and cannot easily be addressed by of the shelve POMDP solvers. To this end we develop FBA-POMCP (inspired by BA-POMCP [2]), a Monte-Carlo Tree Search algorithm [5], that scales favorably in the size of the state space. Second, we propose a Monte-Carlo Monte-Chain reinvigoration method to tackle particle degeneracy of vanilla particle filtering methods (such as Importance Sampling, which are shown to be insufficient to track the belief). We show the favorable theoretical guarantees of this approach and demonstrate empirically that we outperform current state-of-the-art methods on three domains, one of which previous method BA-POMCP.

Fig. 1: Experimental results



Experiments: The paper contains an ablation study and a comparison on three domains of our method with current state-of-the-art method BA-POMCP and an baseline Thompson Sampling inspired planner. Here we highlight two domains: an extended version of the Tiger problem [1] and a larger Collision Avoidance problem. Our experiments show that we (green) significantly outperform BA-POMCP, which is unable to learn in the Factored Tiger problem (left). While none of the methods have converged on the collision avoidance problem (right) yet, BA-POMCP is clearly the slowest learner. The need for reinvigorating the belief is clearest in the Tiger problem, where plain FBA-POMCP occasionally converges to an incorrect belief due to approximation errors and performs poorly on average. Lastly, an interesting observation is that reinvigoration can improve on knowing the correct structure a priori (right figure).

Conclusion: This paper pushes the state of the art in model-based Bayesian reinforcement learning for partially observable settings. We defined the FBA-POMDP framework, which exploits factored representations to compactly describe the belief over the dynamics of the underlying POMDP. And in order to effectively solve the FBA-POMDP, we designed a novel particle reinvigorating algorithm to track the complicated belief and paired it with FBA-POMCP, a new Monte-Carlo Tree Search based planning algorithm. We proved that this method, in the limit of infinite samples, is guaranteed to converge to the optimal policy with respect to the initial belief. In an empirical evaluation we demonstrated that our structure-learning approach is roughly as effective as learning with given structure in two domain. In order to further scale these methods up future work can take several interesting directions. For domains too large to represent with Bayes Networks one could investigate other models to capture the dynamics. For domains that require learning over long sequences, reinvigoration methods that scale more gracefully with history length would be desirable.

References

1. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. In: Artificial intelligence. vol. 101, pp. 99–134 (1998)
2. Katt, S., Oliehoek, F.A., Amato, C.: Learning in pomdps with monte carlo tree search. In: International Conference on Machine Learning. pp. 1819–1827 (2017)
3. Katt, S., Oliehoek, F.A., Amato, C.: Bayesian reinforcement learning in factored pomdps. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. pp. 7–15 (2019)
4. Ross, S., Pineau, J., Chaib-draa, B., Kreitmann, P.: A bayesian approach for learning and planning in partially observable markov decision processes. In: The Journal of Machine Learning Research. vol. 12, pp. 1729–1770 (2011)
5. Silver, D., Veness, J.: Monte-carlo planning in large pomdps. In: Advances in Neural Information Processing Systems. pp. 2164–2172 (2010)