MCTS on model-based Bayesian Reinforcement Learning for efficient learning in Partially Observable environments

Sammie Katt
Northeastern University
katt.s@husky.neu.edu

Frans Oliehoek
Delft University of Technology
f.a.oliehoek@tudelft.nl

Christopher Amato
Northeastern University
c.amato@northeastern.edu

Introduction Impressive advances in Reinforcement Learning on fully observable domains, thanks in part to Deep Learning techniques, have caused a growing interest in solving partially observable domains due to their success on ATARI games. These domains are typically modeled as Partially Observable Markov Decision Processes (POMDPs) [6], which are well-known to be hard to solve due to uncertainty as a result of stochastic transitions, partial observability, and unknown dynamics.

While the work on Deep Learning using Recurrent Neural Nets (RNNs) on POMDPs is promising, their sample inefficiency and lack of addressing the exploration-exploitation trade-off encourages the search for complementing methods. Here we look at Bayesian model-based approaches, which promise the optimal solution to this fundamental issue of the Reinforcement Learning.

The Bayesian RL (BRL) idea is to maintain a probability distribution over the possible dynamics of the POMDP *and* current state, and devise a action picking policy with respect to that distribution, explicitly reasoning over the uncertainty of the agent. In order to do so, BRL methods assume some parametrization of the dynamics and maintain a distribution over the unknown parameters. The iPOMDP [4, 5], for example, views the model as an infinite Hidden Markov Model (iHMM) and specifies the prior with a hierarchical Dirichlet Process (HDP). Other work models the posterior of the dynamics in continuous POMDPs as a Gaussian Process Dynamical Model (GPDM) [3], assuming Gaussian stochasticity. These, and others, provide methods for maintaining distributions over models over time in a principled manner.

Few of these approaches, however, directly address the question of how to then select actions with respect to the posterior, instead relying on simple and expensive look-ahead searches consisting of full-Bellman updates of small horizons (such as [11] where the depth of the tree is 1). In our previous work, we extend the Monte-Carlo Tree Search (MCTS) family of solutions to two BRL formulations [7, 8]. Unfortunately, due to the lack of scalable alternatives, no baseline planner was available as comparison.

In this work we summarize and extend the analysis from previously published work by designing a Thompson-Sampling inspired baseline planner (TSI). In contrast to our approach, this baseline does not consider the full distribution, but plans with respect to a single sample. Our experiments show that such an approach is inferior, demonstrating the need of exploiting the powerful representation that is provided by BRL methods.

Bayes-Adaptive models Here we focus on the discrete *Bayes-Adaptive* models [8, 12, 10]. The Bayes-Adaptive models consider the dynamics of the domain (either tables or Bayes Nets) as part of the hidden state, and can be seen as POMDPs where the (hyper-) state consists of both the current state and the dynamics of the domain. A domain independent dynamics function governs the transitions from one belief to the other (given actions and observations). Effectively, the problem of learning in a POMDP has been cast as a planning problem in a bigger POMDP with the state space being a cross product of the underlying POMDP's state and model space. Here we consider both the tabular BA-POMDP and factored FBA-POMDP [8] that describes the model as a Bayes Net.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

MCTS Monte-Carlo Tree Search [1, 2] has had much publicity recently due to their successful application in solving Go [13]. It is an approach to do online planning, which attempts to pick the best action for a current situation by simulating interactions with the environment. The interactions are represented as a tree, which is grown through 'simulations' with the environment, assuming there is some black-box environment simulator available (representing the POMDP dynamics). It was successfully applied in various settings, including ones with partial observability [14].

Realizing that BA-POMDP casts the learning problem as a (bigger) POMDP planning problem, recent work has also applied MCTS to those models [7, 8], showing that one may circumvent updating the model-belief parameters during simulations by simply sampling a model at the root (*root-sampling*).

Empirical analysis We experiment on 3 different domains: a gridworld inspired problem, an extended tiger problem [6, 7] and a collision avoidance domain [9]. In addition to our method, which applies MCTS on the full posterior belief, we consider a Thompson-Sampling inspired (TSI) baseline planner. This planner first samples a single state and model from the belief, then assumes those are the true state and model, and applies MCTS. The difference in the two approaches is that our method considers the uncertainty of the belief explicitly, whereas TSI plans with respect to a single sample.

In the gridworld problem the agent's task is to navigate from the bottom-left corner of a 2-dimensional grid to a goal cell, that is observed at the start of the episode. Whereas most cells are relatively easy to move over, some 'sink-cells' cause the agent to have a low probability of successfully leaving. The problem is to identify the cells the agent should avoid given noisy observations of its current location, potentially exploiting the fact that these are independent of the goal cell. The extended tiger problem extends the traditional domain with 7 additional irrelevant binary state features, which increases the state space by a factor of 2^7 but keeps the same dynamics complexity. In the collision avoidance problem the agent pilots an airplane while flying from one side of a 2-dimensional grid and must avoid a moving obstacle. The obstacle is only perceivable with a noisy sensor, and its movement is unknown.

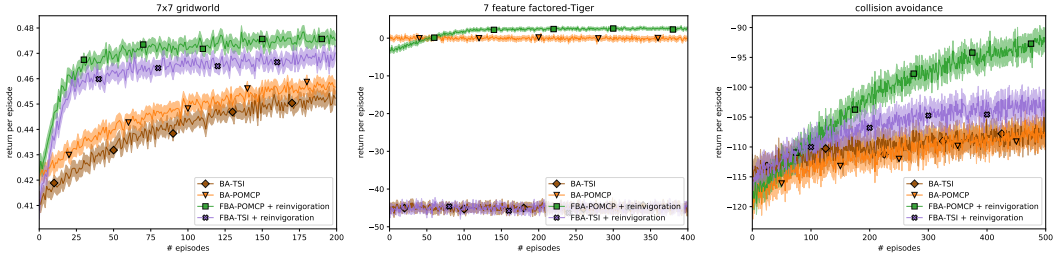

Figure 1: Average return on the gridworld (left), collision avoidance (middle) and extended Tiger (right) problem, the shaded areas indicate the the 95% confidence interval.

Our method consistently outperforms TSI on all domains, demonstrating the utility of reasoning directly over the posterior as opposed to a single sample (see results in fig. 1). The most eye-catching results are on Tiger, where TSI fails completely. This reveals the true nature the planner, as the assumption of being in a particular state leads to the policy of opening a specific door, and corresponds to a large negative reward half of the time. This domain also exhibits a large amount of independent relations, and as a result only our method on the FBA-POMDP approach, which identifies and exploit the structure, is able to learn a satisfying solution. Gridworld, on the other hand, is the least discriminating domain, most probably because the optimal policy is similar between similar states and environments, leading to a lower loss when the wrong state or model is sampled.

Future work In this work we extended empirical evaluation on two BRL approaches by designing a more naive Thompson-Sampling inspired planner. As opposed to previous work, this planner does not exploit the complete knowledge available, and performs significantly worse, demonstrating the need of exploiting the knowledge that BRL methods provide. Since essentially any BRL method for discrete spaces may be coupled with MCTS, it would be interesting to see how well that theory holds up in practice on other approaches and how it affects the tree. For example, MCTS relies on quick simulations and thus only works in practice if a step in the state-model space is efficient.

References

[1] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

[2] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer, 2006.

[3] Patrick Dallaire, Camille Besse, Stephane Ross, and Brahim Chaib-draa. Bayesian reinforcement learning in continuous pomdps with gaussian processes. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2604–2609. IEEE, 2009.

[4] Finale Doshi-Velez. The infinite partially observable markov decision process. In *Advances in neural information processing systems*, pages 477–485, 2009.

[5] Finale Doshi-Velez, David Pfau, Frank Wood, and Nicholas Roy. Bayesian nonparametric methods for partially-observable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):394–407, 2015.

[6] 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.

[7] Sammie Katt, Frans A Oliehoek, and Christopher Amato. Learning in pomdps with monte carlo tree search. In *International Conference on Machine Learning*, pages 1819–1827, 2017.

[8] Sammie Katt, Frans A Oliehoek, and Christopher Amato. Efficient exploitation of factored domains in bayesian reinforcement learning for pomdps. In *Adaptive Learning Agents*, 2018.

[9] Miao Liu, Xuejun Liao, and Lawrence Carin. Online expectation maximization for reinforcement learning in pomdps. In *IJCAI*, pages 1501–1507, 2013.

[10] Pascal Poupart and Nikos Vlassis. Model-based bayesian reinforcement learning in partially observable domains. In *Proc Int. Symp. on Artificial Intelligence and Mathematics,*, pages 1–2, 2008.

[11] Stephane Ross, Brahim Chaib-draa, and Joelle Pineau. Bayesian reinforcement learning in continuous pomdps. In *International Conference on Robotics and Automation (ICRA)*, 2008.

[12] Stéphane Ross, Joelle Pineau, Brahim Chaib-draa, and Pierre Kreitmann. A bayesian approach for learning and planning in partially observable markov decision processes. *Journal of Machine Learning Research*, 12(May):1729–1770, 2011.

[13] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[14] David Silver and Joel Veness. Monte-carlo planning in large pomdps. In *Advances in neural information processing systems*, pages 2164–2172, 2010.