Model-based Bayesian Reinforcement Learning with Adaptive State Aggregation

Cosmin Paduraru, Arthur Guez, Doina Precup and Joelle Pineau
McGill University
Montreal, Quebec, Canada

Model-based Bayesian reinforcement learning provides an elegant way of incorporating model uncertainty for trading off between exploration and exploitation. We propose an extension of model-based Bayesian RL to continuous state spaces. The key feature of our approach is its search through the space of model structures, thus adapting not only the model parameters but also the structure itself to the problem at hand. We currently present algorithms and results for structures that are discretizations of the state space, but we hope to extend this to more powerful representations.

The class of models we are working with are composed of a partitioning (also referred to as an aggregation or discretization) of the state space, and a transition matrix which defines the probabilities of transitioning between partitions. For simplicity, we assume a small set A of discrete actions (although an extension to the continuous action case is also in the works), and we maintain a separate transition matrix for each action. We denote the partitioning by Ω, and the transition matrix for action a by Θ_a. We also assume that the transition distribution is uniform within the next state partition. Thus, we approximate the transition probabilities from state s to state s' by

$$P(s'|s,a) = \frac{1}{Vol(\omega(s'))}\Theta_a(\omega(s),\omega(s'))$$

where $Vol(\omega(s'))$ is the volume of $\omega(s')$, the partition containing s'. The partitions in Ω are created by splitting existing partitions, making a tree structure the most appropriate way of describing Ω.

State aggregation models come with several caveats. First, better discrimination between states always comes at the cost of decreased generalization, and vice versa. Second, this class of models will in most cases not contain the true transition model. Third, the probabilities of transitioning between partitions are not in fact multinomial. Indeed, these probabilities depend on the distribution of states inside a partition, which makes them have a non-tractable dependency on the history and the action selection mechanism.

Bayesian learning actually enables us to get around some of these problems. The main reason for this is that our model-based Bayesian RL algorithm causes the distribution over partitionings to change over time in response to the observed data (this is also done heuristically in other state aggregation work, such as Munos and Moore (1999)). Thus, the first issue could be handled by having good generalization (large partitions) in the beginning, and then discretizing more as sufficient data is available. The second and third issue should also be less problematic as the discretization becomes finer.

For the technical implementation of these ideas, we draw inspiration from two existing papers: Ross and Pineau (2008) and Chipman, George and McCulloch (1998). Ross and Pineau proposed a Bayesian RL algorithm for factored MDPs, which allows them to handle finite MDPs with large state spaces. Chipman, George and McCulloch describe a Bayesian approach to supervised learning that uses tree-based state aggregation. In this paper, we only present an overview of our implementation, and leave most of the details for a future lengthier publication.

Similarly to previous work, we break up the probability of a model as $P(\Omega,\Theta) = P(\Omega)P(\Theta|\Omega)$, where $\Theta = \{\Theta_a | a \in A\}$. For the posterior, we use the same factorization, but additionally condition on the history.

Since there is an infinite number of possible discretizations, the distribution $P(\Omega)$ over the discretizing trees cannot be maintained explicitly, so an approximate, particle filter style approach is taken. A set of trees is initially sampled from the prior (which gives more weight to smaller trees), and the probabilities of these structures is maintained. When the likelihood of the maintained set of trees falls below some threshold, a new set of trees is sampled using the well-known Metropolis-Hastings algorithm, as described by Chipman et al. (1998). This requires that previous transitions be stored, because we need to compute the likelihood of the data under different models.

The distribution over partition-to-partition transition models $P(\Theta|\Omega)$ is represented as a Dirichlet and maintained by updating counts. At several points in the algorithm (updating the probabilities of the structures, computing the Metropolis-Hastings ratio) we need to marginalize over Θ. Fortunately, this can be done in closed form.

To find the optimal action with respect to the posterior uncertainty in the model, we have to find an approximation to the optimal policy in the resulting Bayes-adaptive MDP. One possibility for approximating the optimal BAMDP solution that has shown good results is sample-based online planning, as used for instance by Ross and Pineau (2008). They used a very simple planning strategy, where the actions are sampled uniformly, but other more informed action selection methods have also been proposed (e.g. Castro, 2007; Wang et al., 2005). Online methods, however, seemed to have prohibitively large variance in our preliminary experiments, so we are also considering myopic heuristics such as value of perfect information (Dearden et al., 1999) .

There are several existing Bayesian RL algorithms designed for continuous state spaces. Ross et al. (2008) present a model-based Bayesian RL method for continuous-state, continuous-action POMDPs; it requires, however, the transition model to be Gaussian. The Gaussian process temporal difference work of Engel et al. (2005) can handle continuous state spaces by representing the value function as a kernel-based Gaussian. Several extensions of non-bayesian exploration methods to continuous state spaces have also been proposed (Kakade et al., 2003; Nouri and Littman, 2008).

We are currently empirically evaluating our method on continuous-state reinforcement learning problems; in the near future, we intend to compare our on-line learning performance against some of the methods mentioned in the previous paragraph.

References

Castro, P. S. (2007). Bayesian learning in Markov decision processes. Master's thesis, McGill University, Montreal, Canada.

Chipman, H., George, E., & McCulloch, R. (1998). Bayesian CART Model Search . *Journal of the American Statistical Association*, 935–960.

Dearden, R., Friedman, N., & Andre, D. (1999). Model based bayesian exploration. *In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (pp. 150–159).

Engel, Y., Mannor, S., & Meir, R. (2005). Reinforcement learning with gaussian processes. *Proceedings of the International Conference on Machine Learning (ICML)*.

Munos, R., & Moore, A. (1999). Variable resolution discretization for high-accuracy solutions of optimal control problems. *International Joint Conference on Artificial Intelligence*.

Nouri, A., & Littman, M. L. (2008). Multi-resolution exploration in continuous spaces. *NIPS*.

Ross, S., Chaib-draa, B., & Pineau, J. (2008). Bayesian reinforcement learning in continuous pomdps with application to robot navigation. *Proceedings of the IEEE International Conference on Robotics and Automation*.

Ross, S., & Pineau, J. (2008). Model-based bayesian reinforcement learning in large structured domains. *Proceedings of the 24th UAI*.

Wang, T., Lizotte, D., Bowling, M., & Schuurmans, D. (2005). Bayesian sparse sampling for on-line reward optimization. *Proceedings of the Twenty-second International Conference on Machine Learning*.