Action Selection in Bayesian Reinforcement Learning

Tao Wang
University of Alberta
Edmonton, AB
Canada T6G 2E8

Abstract

My research attempts to address on-line action selection in reinforcement learning from a Bayesian perspective. The idea is to develop more effective action selection techniques by exploiting information in a Bayesian posterior, while also selecting actions by growing an adaptive, sparse lookahead tree. I further augment the approach by considering a new value function approximation strategy for the belief-state Markov decision processes induced by Bayesian learning.

Bayesian Reinforcement Learning

Imagine a mobile vendor robot ("vendorbot") loaded with snacks and bustling around a building, learning where to visit to optimize its profit. The robot must choose wisely between selling snacks somewhere far away from its home or going back to its charger before its battery dies. How could a robot effectively learn to behave from its experience (previous sensations and actions) in such an environment?

Reinforcement mechanisms provide a robot an opportunity to improve its decision making via interacting with the world and receiving evaluative feedback on its actions. The goal of the robot is to maximize the total reward it obtains by taking actions in the environment. Normally, the environment is uncertain; therefore, the outcome of an action is non-deterministic. At each decision-making point the robot must choose its action, even while learning. Here there is a fundamental tradeoff: it could exploit its current knowledge to gain reward by taking actions known to give relatively high reward, or explore the environment to gain information by trying actions whose value is uncertain. This is the well-known problem of balancing exploitation with exploration in reinforcement learning, or more generally, the problem of action selection during reinforcement learning.

Interestingly, there remains little convergence on the fundamental question of on-line action selection in reinforcement learning. Beyond the standard ϵ-greedy and Boltzmann selection strategies, few techniques have been adopted beyond the papers that originally proposed them. Nevertheless, there remains a persistent belief that more complicated selection strategies can yield improved results (Dearden, Friedman, & Andre 1999; Strens 2000; Wyatt 2001). A possible reason for limited use of sophisticated methods might be the complexity of the proposals, or the assumption that the degree of improvement might not be dramatic.

In my work, I have been investigating a Bayesian approach to action selection in reinforcement learning. The Bayesian approach still appears to be under-researched given the important role it has played in other areas of machine learning. However, flexible Bayesian tools, such as Gaussian process regression, have had a significant impact on other areas of machine learning research but have only just recently been introduced to reinforcement learning (Engel, Mannor, & Meir 2003). Moreover, Bayesian approaches seem ideally suited to reinforcement learning as they offer an explicit representation of uncertainty, which is essential for reasoning about the exploration versus exploitation tradeoff. In fact, Bayesian approaches offer the prospect of *optimal* action selection. Bayesian decision theory solves the exploration versus exploitation tradeoff directly (but implicitly) by asserting that the optimal action is one which, over the entire time horizon being considered, maximizes the total expected reward averaged over possible world models. Therefore, any gain in reducing uncertainty is not valued for its own sake, but measured instead in terms of the gain in future reward it offers. In this way, explicit reasoning about exploration versus exploitation is subsumed by direct reasoning about rewards obtained over the long term.

Despite the apparent elegance of the Bayesian approach, there remain serious barriers to its application. The most obvious drawback is the computational challenge posed by optimal Bayesian decision making, which is known to be intractable in all but trivial decision making contexts. This means that with a Bayesian approach one is forced to consider heuristic approximations. In response, a small body of research has developed on on-line approximations of optimal Bayesian action selection. The number of proposals remains relatively small and no widely adopted approximation strategy has emerged. However, the potential power of Bayesian modeling for approximating optimal action selection makes this approach worth investigating.

Current Results

I have been working on the problem of on-line action selection during reinforcement learning from a Bayesian perspective. In particular, I am interested in developing practical, relatively straightforward action selection strategies for a reinforcement learner. The idea I have been exploring is to exploit a Bayesian posterior to make intelligent action selection decisions by constructing and searching a sparse lookahead tree, originally inspired by the idea of sparse sampling (Kearns, Mansour, & Ng 2002). The main extensions I have investigated are to grow the sparse lookahead tree adaptively, by sampling actions according to their relative importance (rather than enumerating actions), and growing an imbalanced tree that is targeted toward improving the value estimates of the most promising branches only. One of the key observations is that in a Bayesian setting, actions do not need to be completely enumerated to identify an approximately optimal decision with high probability. The outcome is a flexible, practical technique—"Bayesian sparse sampling"— for improving action selection in episodic reinforcement learning problems (Wang *et al.* 2005). In this work, I have demonstrated that the proposed technique works well in several abstract domains, including Bernoulli bandits, Gaussian bandits, discrete (multivariate) Gaussian process bandits, and multi-dimensional continuous action Gaussian process bandits. I have subsequently studied more interesting abstract domains with correlated actions: correlated Bernoulli bandits and correlated Gaussian bandits.

Currently I have been working on a new approach to approximate planning in partially observable Markov decision processes (POMDPs) that is based on convex quadratic function approximation (Wang *et al.* 2006). Although the approximation strategy is applicable to general POMDPs, in this work I apply it to the belief-state MDP model that arises in Bayesian reinforcement learning. The main idea behind this approximation technique is to replace the maximum over a set of α-vectors with a convex quadratic upper bound that summarizes the entire set with small information loss. Here, we achieve significant compression by reducing the set of α-vectors to a single quadratic approximation. Specifically, we approximate the optimal value function by a convex upper bound composed of a fixed number of quadratics, and optimize it at each stage by semidefinite programming. In this way, we can achieve an approximate value iteration algorithm that runs in time linear in the horizon. I have shown that this approach can achieve competitive approximation quality to current techniques while still maintaining a bounded size representation of the function approximator. Moreover, an upper bound on the optimal value function can be preserved if required. Overall, the technique requires computation time and space that is only linear in the number of iterations (horizon time).

Proposed Research

My general goal is to develop improved action selection algorithms for interesting sequential decision making problems that arise in more complicated reinforcement learning scenarios, such as robotics and games. Meanwhile I intend to analyze the theoretical properties of the action selection techniques, including establishing performance guarantees and computational bounds, but also possibly identifying limitations of the methods.

For fundamental research, I am focusing on extending the action selection strategy and Gaussian process framework developed so far. One idea is to extend the concept of a kernel between actions to a kernel between policies. The idea is that by using a kernel to measure the similarity between two policies over the same state and action space, one can learn about the value of one policy by estimating the value of another. This approach makes the Gaussian process machinery available to the POMDP action selection problem.

Two robotic applications I am pursuing are AIBO (Sony robotic dog) walking and the mobile vendorbot project (a wheeled Pioneer III robot). The goal of the AIBO walking project is to get the AIBO to learn to walk faster using techniques that require little sophisticated engineering or prior knowledge. Here the challenge is to solve a noisy high dimensional parameter optimization problem. To do so we developed a new optimization algorithm using a Gaussian Process model (Lizotte *et al.* 2005). The mobile vendorbot project, however, involves sequential decision making in a non-episodic environment—unlike the AIBO task—and therefore requires both the Bayesian sparse sampling and POMDP approximation strategies to be combined. Although we have to face challenging problems besides decision making in this case (e.g. robot navigation), the vendorbot project provides an excellent opportunity to investigate the power and flexibility of the proposed strategies.

Another direction I am pursuing is opponent modeling in games. Here I have begun to work in the domain of simplified two player poker. The difficulty of learning in a game setting is the inherent non-stationarity of the environment created by a simultaneously adapting opponent. Here I hope to develop improved strategies for this domain by employing Bayesian models of adaptive opponents, and applying the decision making techniques we have developed above.

References

Dearden, R.; Friedman, N.; and Andre, D. 1999. Model based Bayesian exploration. In *Proceedings UAI.*

Engel, Y.; Mannor, S.; and Meir, R. 2003. Bayes meets Bellman: The Gaussian process approach to temporal difference learning. In *Proceedings ICML.*

Kearns, M.; Mansour, Y.; and Ng, A. 2002. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning* 49(2-3):193–208.

Lizotte, D.; Wang, T.; Bowling, M.; and Schuurmans, D. 2005. Gaussian process regression for optimization. *NIPS Workshop on Value of Information.*

Strens, M. 2000. A Bayesian framework for reinforcement learning. In *Proceedings ICML.*

Wang, T.; Lizotte, D.; Bowling, M.; and Schuurmans, D. 2005. Bayesian sparse sampling for on-line reward optimization. In *Proceedings ICML.*

Wang, T.; Bowling, M.; Poupart, P.; and Schuurmans, D. 2006. Compact, convex upper bound iteration for approximate POMDP planning. Submitted.

Wyatt, J. 2001. Exploration control in reinforcement learning using optimistic model selection. In *Proc. ICML.*