Linear Fitted-Q Iteration with Multiple Reward Functions

Daniel J. Lizotte, Michael Bowling, Susan A. Murphy

Abstract

We present a general and detailed development of an algorithm for finite-horizon fitted-Q iteration with an arbitrary number of reward signals and linear value function approximation using an arbitrary number of state features (Lizotte, Murphy, and Bowling 2012). This includes a detailed treatment of the 3-reward function case using triangulation primitives from computational geometry and a method for identifying globally dominated actions. We also present an example of how our methods can be used to construct a real-world decision aid by considering symptom reduction, weight gain, and quality of life in sequential treatments for schizophrenia. Finally, we discuss future directions in which to take this work that will further enable our methods to make a positive impact on the field of evidence-based clinical decision support.

Motivation

Within the field of personalized medicine, there is increasing interest in investigating the role of sequential decision making for managing chronic disease. Reinforcement learning methods are already being used to analyze data from clinical trials wherein patients are given different treatments in multiple *stages* over time (Pineau et al. 2007; Shortreed et al. 2011). The patient data collected during each stage are very rich and commonly include several continuous variables related to symptoms, side-effects, treatment adherence, quality of life, and so on. For the ith patient in the trial, we obtain a *trajectory* of *observations* and *actions* of the form $o_1^i, a_1^i, o_2^i, a_2^i, ..., o_T^i, a_T^i, o_{T+1}^i$. Here, a_t^i represents the action (treatment) at time t, and o_t^i represents measurements made of patient i *after* action a_{t-1}^i and *before* action a_t^i. The first observations o_1^i are baseline measurements made before any actions are taken.

To analyze these data using reinforcement learning methods, we must define two functions $s_t(o_1, a_1, ..., o_t)$ and $r_t(s_t, a_t, o_{t+1})$ which map the patient's current history to a state representation and a scalar reward signal, respectively. Applying these functions to the data from the ith patient gives a trajectory $s_1^i, a_1^i, r_1^i, s_2^i, a_2^i, r_2^i, ..., s_T^i, a_T^i, r_T^i$.

These redefined data are treated as sample trajectories from a known policy. Once we have these, we will view ongoing patient care as a Markov decision process (MDP), and apply batch off-policy reinforcement learning methods (e.g. see Ernst, Geurts, and Wehenkel 2005) to learn an optimal policy that takes a patient state and indicates which action appears to be best in view of the data available. In an MDP, both the state transition dynamics and the reward distributions are assumed to have the *Markov property*. That is, given the value s_t of the current state, the distribution of next state S_{t+1} and current reward R_t is conditionally independent of s_j, a_j, r_j for all $j < t$.

The interplay between predictive power, tractability, and interpretability makes the definition of s_t a challenging problem. However, the question of how s_t should be defined can be answered at least in part by the data themselves together with expert knowledge and feature/model selection techniques analogous to those used in supervised learning settings (Keller, Mannor, and Precup 2006) *if we have an adequate definition of* r_t. However, a major difficulty with using trial data in this way is that there is often no obviously correct way to define r_t. Indeed, any definition of r_t is an attempt to answer the question "*What is the right quantity to optimize?*"—a question that is driven by the objectives and preferences of individual decision makers and *cannot be answered by the data alone.* There are many reasonable reward functions one could define, since each patient record includes a multi-dimensional measurement of that patient's overall well-being. Furthermore, these different dimensions are often optimized by different treatments, and therefore the choice of which dimension to use as the reward will affect the resulting learned policy. For example, a policy that minimizes expected symptom level will tend to choose more aggressive drugs that are very effective but that have a more severe side-effect profile. On the other hand, a policy that minimizes expected side-effect measurements will choose drugs that are less effective but that have milder side-effects.

We consider sets of MDPs that all have the same $\mathcal{S}_t$, $\mathcal{A}_t$, and state transition dynamics, but whose expected reward functions $r_t(s_t, a_t, \boldsymbol{\delta})$ have an additional parameter $\boldsymbol{\delta}$ that represents the relative importance assigned to different reward signals. One may think of $\boldsymbol{\delta}$ as a special part of state that: i) does not evolve with time, and ii) does not influence transition dynamics. Each fixed $\boldsymbol{\delta}$ identifies a single MDP by fixing a reward function, which has a corresponding optimal state-action value function. In order to mathematize the relationship between preference and $\boldsymbol{\delta} = (\delta_{[0]}, \delta_{[1]}, ..., \delta_{[D-1]})$, we define the structure of $r_t(s_t, a_t, \boldsymbol{\delta})$ to be

$$r_t(s_t, a_t, \boldsymbol{\delta}) = \delta_{[0]} r_{t[0]}(s_t, a_t) + \delta_{[1]} r_{t[1]}(s_t, a_t) + ... \\ + (1 - \sum_{d=0}^{D-2} \delta_{[d]}) r_{t[D-1]}(s_t, a_t).$$

Rather than eliciting the preference $\boldsymbol{\delta}$ and producing a policy that recommends a single action per state, our approach is to learn the optimal Q-function $Q_t(s_t, a_t, \boldsymbol{\delta})$ *for all $\boldsymbol{\delta}$ exactly and simultaneously*. This allows us for each action to answer the question, *"What range of preferences makes this action a good choice?"* This provides much richer information about the possible actions at each stage. Furthermore, even if a preference is specified, our methods allow the maker to immediately see if his or her preferences are near a "boundary"—that is, whether a small change in preference can lead to a different recommended action. In this case, two or more actions perform similarly, and therefore the final decision could be based on other considerations like dosing schedule, difference in cost, etc. Finally, we can also determine if an action is not optimal for any preference.

Approach

We show that the optimal state-action value function $Q_t(s_t, a_t, \boldsymbol{\delta})$ is piecewise-linear in the tradeoff parameter $\boldsymbol{\delta}$. Value backups for fitted Q-learning require two operations: maximization over actions, and expectation over future states. We use an exact piecewise-linear representation of the functions $Q_t(s_t, a_t, \cdot)$ which allows us to perform these operations to exactly compute value backups for all $\boldsymbol{\delta}$. We can also identify the set of dominated actions, i.e. the actions that are not optimal for any $(s_t, \boldsymbol{\delta})$ pair.

Unlike in POMDP planning, where value is always a *convex* function of belief state, we show that because our approach estimates value functions using linear regression, the Q-functions in our problem are *not* convex in $\boldsymbol{\delta}$. We therefore develop alternative methods for representing value functions based on primitives from computational geometry.

Related Work and Contributions

Early work in this direction (Barrett and Narayanan 2008) explored the problem of simultaneously computing optimal policies for a class of reward functions over a small, finite state space in a framework where the model is known. Subsequent developments were made that focussed on the infinite-horizon discounted setting and black-box function approximation techniques (Castelletti et al. 2010; Vamplew et al. 2011). Previously, we extended the approach of Barrett and Narayanan (2008) to the setting with real-valued state features and *linear* function approximation, which is a more appropriate framework for analyzing trial data (Lizotte, Bowling, and Murphy 2010). We also introduced an algorithm that is asymptotically more time- and space-efficient than the Barrett & Narayanan approach, and described how it can be directly applied to batch data. Finally, we gave an algorithm for finding the set of all non-dominated actions in the single-variable continuous state setting. This paper builds on previous work by contributing:

- A general and detailed development of finite-horizon fitted-Q iteration with an arbitrary number of reward signals and linear approximation using an arbitrary number of state features
- A detailed treatment of 3-reward function case using triangulation algorithms from computational geometry that has the same asymptotic time complexity as the 2-reward function case
- A more concise solution for identifying globally dominated actions under linear function approximation, and method for solving this problem in higher dimensions
- A real-world decision aid example that considers symptom reduction, weight gain, and quality of life when choosing treatments for schizophrenia

References

Barrett, L., and Narayanan, S. 2008. Learning all optimal policies with multiple criteria. In *Proceedings of the 25th International Conference on Machine Learning*, 41–47.

Castelletti, A.; Galelli, S.; Restelli, M.; and Soncini-Sessa, R. 2010. Tree-based reinforcement learning for optimal water reservoir operation. *Water Resources Research* 46(W06502).

Ernst, D.; Geurts, P.; and Wehenkel, L. 2005. Tree-Based Batch Mode Reinforcement Learning. *Journal of Machine Learning Research* 6:503–556.

Keller, P. W.; Mannor, S.; and Precup, D. 2006. Automatic basis function construction for approximate dynamic programming and reinforcement learning. In *Proceedings of the 23rd International Conference on Machine learning*, ICML '06, 449–456.

Lizotte, D. J.; Bowling, M.; and Murphy, S. A. 2010. Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis. In *Proceedings of the 27th International Conference on Machine Learning*, 695–702.

Lizotte, D. J.; Murphy, S. A.; and Bowling, M. 2012. Linear fitted-Q iteration with multiple reward functions. *Journal of Machine Learning Research* 13:3253–3295.

Pineau, J.; Bellemare, M. G.; Rush, A. J.; Ghizaru, A.; and Murphy, S. A. 2007. Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence* 88(Suppl 2):S52–S60.

Shortreed, S.; Laber, E. B.; Lizotte, D. J.; Stroup, T. S.; Pineau, J.; and Murphy, S. A. 2011. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine Learning* 84(1–2):109–136.

Vamplew, P.; Dazeley, R.; Berry, A.; Issabekov, R.; and Dekker, E. 2011. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning* 84:51–80.