# Acting With Confidence: Quantifying Policy Uncertainty for Medical Applications

Daniel J. Lizotte, Eric Laber and Susan Murphy

April 19, 2009

In the medical field, “sequential randomized trials” are becoming an increasingly important mechanism for gathering data to aid in clinical decision making, particularly in the study of chronic illnesses. These trials produce data that have an inherent sequential structure: Each patient receives a sequence of several randomized treatments over the course of the study, and patient responses are measured that allow us to evaluate the therapeutic benefits and deficiencies of the different sequences. Reinforcement learning provides several well-studied algorithms that can learn policies from this type of data.

However, most reinforcement learning methods do not provide measures of confidence in the learned policy, which are a necessity for any real-world deployment. Reinforcement learning using clinical trial data can analyze the outcomes of hundreds or even thousands of patients and provide useful guidance in cases where, to the best of a physician’s knowledge, several treatments appear to be acceptable and equivalent alternatives. Nonetheless, no matter how comprehensive the clinical trial, the data we collect and the policy we learn cannot account for all of the possible situations that may arise in clinical practice. For example, we will never collect data about treating a patient with a drug to which they are known to be allergic—we expect a clinician who implements a policy learned from data to use his or her own expertise to avoid taking a dangerous action in this situation.

Our goal is therefore to learn a policy from a batch of patient data and then to transfer that policy to another agent—the clinician—who will make use of the policy in the future. The receiving agent in this setting will not follow the learned policy blindly, however: At each decision point, the clinician will decide what to do based on the received policy *and* based on his or her own experience and knowledge.

In order to make this policy transfer effective and useful, we must express the learned policy in a way that the receiving agent can easily interpret and understand, and we must also give the receiving agent some sense of when it should feel free to use its discretion in choosing the next action; that is, we would like to let the receiving agent know if, given a particular state, the data actually support several possible optimal actions.

Thus we not only want to learn a policy; we also want to offer a measure of confidence in the learned policy in a manner analogous to measures of confidence given in traditional clinical trials by tests for statistical significance. This is a difficult problem from an interpretability viewpoint in the case where there are more than two actions, because pairwise comparisons of action-values do not always lead to simple statements about which action or set of actions is likely to be optimal. For example, pairwise significance tests could reveal that action 1 and action 2 are not significantly different, action 2 and action 3 are not significantly different, but action 1 is significantly better than action 3. It is not clear from these results what set of actions we should consider equivalent in terms of value. Assessing confidence in policies is also a difficult problem from a technical viewpoint, because the quantities we estimate in reinforcement learning, such as Q-values for example, are non-smooth functions of the data. This is different from more traditional settings where we reason using smooth estimates like sample means and variances. The non-smoothness introduced by the max operator in Q-learning can cause commonly-used statistical procedures like the bootstrap to produce misleading results in a variety of situations.

Our work focuses on the development of useful measures of confidence in learned policies, with the additional requirement that we produce results that are easily interpretable by individuals who are not experts in reinforcement learning. We present an alternative to multiple hypothesis tests based on the "probability of replication" in which we estimate for each action the probability that we would decide the action was optimal if we drew a new training set and re-ran our learning algorithm. We estimate this quantity using an adaptive bootstrap algorithm that uses dataset re-sampling to estimate these probabilities in a way that is robust to some of the non-smoothness problems mentioned above. We also present a graphical representation of this confidence measure using examples on real-world trial data. Our visualization approach has garnered a great deal of positive feedback from physicians, who feel it provides them with the necessary policy information in a format that is at once concise and detailed.

This project is part of ongoing joint work at the University of Michigan Department of Statistics and the University of McGill School of Computer Science on the application of reinforcement learning to clinical decision making. Our group is also investigating other key problems in this area including frequentist and Bayesian inference about value functions, feature selection, and dealing with missing data. We are also working closely with psychiatrists on the analysis of real sequential randomized trial data on treatments for major depressive disorder and schizophrenia.