Model-Free Monte Carlo–like Policy Evaluation

Raphael Fonteneau
Department of EECS
University of Liège
BELGIUM

Susan Murphy
Department of Statistics
University of Michigan
USA

Louis Wehenkel
Department of EECS
University of Liège
BELGIUM

Damien Ernst
Department of EECS
University of Liège
BELGIUM

1 Introduction

We propose an algorithm for estimating the finite-horizon return of a closed loop control policy from an a priori given (off-policy) sample of one-step transitions [1]. It averages cumulated rewards along a set of "broken trajectories" made of one-step transitions selected from the sample on the basis of the control policy. Under some Lipschitz continuity assumptions on the system dynamics, reward function and control policy, we provide bounds on the bias and variance of the estimator that depend only on the Lipschitz constants, on the number of broken trajectories used in the estimator, and on the sparsity of the sample of one-step transitions.

2 Monte Carlo policy evaluation

Discrete-time stochastic optimal control problems arise in many fields such as finance [2], medicine [3], engineering [4] as well as artificial intelligence [5]. Many techniques for solving such problems use an oracle that evaluates the performance of any given policy in order to navigate rapidly in the space of candidate optimal policies to a (near-)optimal one. When the considered system is accessible to experimentation at low cost, such an oracle can be based on a Monte Carlo (MC) approach. With such an approach, several "on-policy" trajectories are generated by collecting information from the system when controlled by the given policy, and the cumulated rewards observed along these trajectories are averaged to get an unbiased estimate of the performance of that policy. However if obtaining trajectories under a given policy is very costly, time consuming or otherwise difficult, e.g. in medicine or in safety critical problems, the above approach is not feasible.

3 Model-free Monte Carlo policy evaluation

In this paper, we propose a policy evaluation oracle in a *model-free* setting. In our setting, the only information available on the optimal control problem is contained in a sample of one-step transitions of the system, that have been gathered by some arbitrary experimental protocol, i.e. independently of the policy that has to be evaluated. Our estimator is inspired by the MC approach. Similarly to the MC estimator, it evaluates the performance of a policy by the average of the cumulated rewards along some trajectories. However, rather than "real" on-policy trajectories of the system generated by fresh experiments, it uses a set of "broken trajectories" that are rebuilt from the given sample and from the policy that is being evaluated.

4 Preliminary results

Under some Lipschitz continuity assumptions on the system dynamics, reward function and policy, we provide bounds on the bias and variance of our model-free policy evaluator, and show that it behaves like the standard MC estimator when the sample sparsity decreases towards zero. These theoretical properties are illustrated with some promising simulations results.

References

[1] R. Fonteneau, S. Murphy, L. Wehenkel, and D. Ernst, "Model-free Monte-Carlo like policy evaluation," *Submitted*, 2010.

[2] J. Ingersoll, *Theory of Financial Decision Making*. Rowman and Littlefield Publishers, Inc., 1987.

[3] S. Murphy, "Optimal dynamic treatment regimes," *Journal of the Royal Statistical Society, Series B*, vol. 65(2), pp. 331–366, 2003.

[4] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[5] R. Sutton and A. Barto, *Reinforcement Learning*. MIT Press, 1998.