Sequentially Interacting Markov Chain Monte Carlo Based Policy Iteration

Orhan Sönmez and A. Taylan Cemgil
{orhan.sonmez,taylan.cemgil}@boun.edu.tr

May 10, 2013

Introduction

The main motivation of this study is to employ state-of-the-art Monte Carlo methods to construct a sample-efficient online reinforcement learning (RL) algorithm. In this paper, we introduce a policy iteration method where policy is evaluated with sequentially interacting Markov chain Monte Carlo (SIMCMC) [1] for discrete time model-based RL with continuous state space. In the scope of this paper, we assume that the transition model and the reward function are known.

Problem and Approach

Toussaint and Storkey [2] proposed an equivalent probabilistic inference problem to RL and derive the corresponding EM algorithm. Hence, it generates a probabilistic policy iteration scheme as the following update equation,

$$\pi^{(k+1)} \leftarrow \arg\max_{\pi} \left\langle \log P(r=1, x_{0:T}, a_{0:T}; \pi) \right\rangle_{P(x_{0:T}, a_{0:T} | r=1; \pi^{(k)})} \tag{1}$$

with assuming reward $r \in \{0, 1\}$ without loss of generality where $x_{0:T}$,$a_{0:T}$ are fixed-length state, action trajectories respectively and π denotes the policy.

However, calculating the expectation in (1) at the E-step is intractable. Hence, we derived a SIMCMC method that samples from the posterior $P(x_{0:T}, a_{0:T} | r=1; \pi^{(k)})$ in order to estimate it. For this purpose, we utilized bridge functions ϕ_n annealed with an increasing $\eta(n)$ in the form of,

$$\phi_n(x_{0:T}, a_{0:T}) \propto P(x_{0:T}, a_{0:T}; \pi^{(k)}) P(r=1 | x_{0:T}, a_{0:T}; \pi^{(k)})^{\eta(n)} \quad n \in [0, N] \tag{2}$$

where ϕ_0 and ϕ_N are selected as prior and target posterior densities by explicitly choosing $\eta(0) = 0$ and $\eta(N) = 1$. This E-step actually corresponds to the policy evaluation step in a policy iteration scheme.

Fortunately, the maximization in (1) has a closed form solution due to Markov properties [2]. Yet, policies have to be somehow characterized among the continuous state space. We use a Gaussian process (GP) to approximate the policies [3] as shown in Figure 1. At the M-step, a GP with the support of samples from target posterior is one of the maximizers. Hence, instead discretizing the state space, we directly use the samples from the bridge function ϕ_N in order to approximate the policy of the next iteration.

Results and Remarks

We evaluated our method on the well-known mountain-car problem and it has converged to an suboptimal solution as shown in Figure 3 as expected.

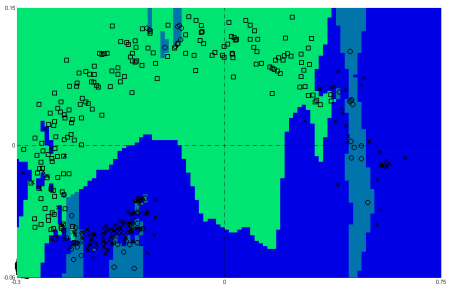

Figure 1: Policy approximation with GP by using the sampled trajectories

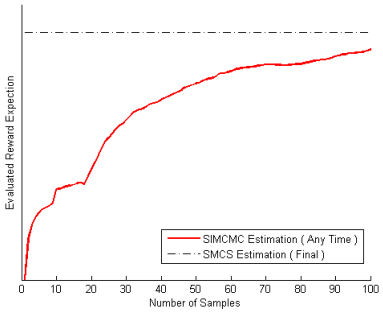

Figure 2: SIMCMC estimate at any time of the policy evaluation compared with SMCS final estimate

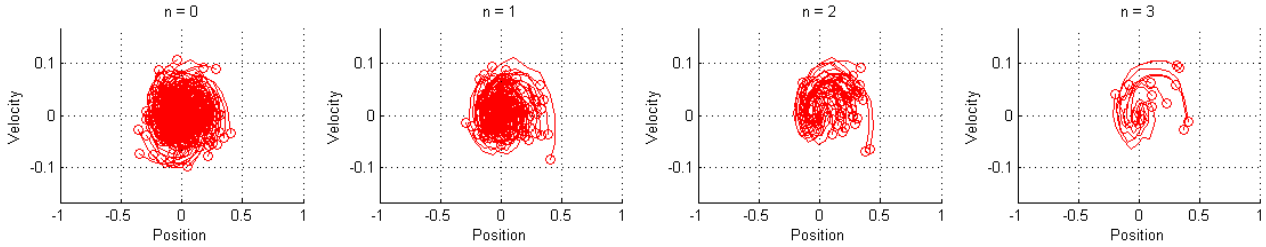

Figure 3: An example of the convergence of our method for N = 3 towards the optimal policy of the mountain-car problem at the very first iteration starting with a uniform policy.

Finally, we are also aware that SIMCMC methods are not the best choice with respect to sample efficiency compared to sequential Monte Carlo samplers (SMCS) [5] as shown in Figure 2. However, they are more appropriate for the online settings due to their estimation at any time property. Thus, we are planning to develop an online RL algoritm based on SIMCMC policy evaluation by approximating the dynamics of the model with GP [6].

References

[1] Brockwell, A., Del Moral, P., Doucet A., Sequentially Interacting Markov Chain Monte Carlo Methods. *The Annals of Statistics*, 38(6):3870–3411, December 2010.

[2] Toussaint, M., Storkey, A., Probabilistic Inference for Solving Discrete and Continuous State Markov Decision Processes. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 945–952, 2006.

[3] Deisenroth, M., Rasmussen C.E., Peters, J., Gaussian Process Dynamic Programming. *Neurocomputing*, 72(7-9):1508–1524, March 2009.

[4] Sutton, R. S., Barto, A. G., *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[5] Del Moral, P., Doucet A., Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, 68(3):411–436, June 2006.

[6] Deisenroth, M., Rasmussen C.E. PILCO: A Model-Based and Data-Efficient Approach to Policy Search. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 465–472, 2006.