ABC reinforcement learning

Christos Dimitrakakis Nikolaos Tziortziotis

May 8, 2013

1 Introduction

In the *reinforcement learning problem*, an agent is acting in an unknown environment μ, according to a policy π, both of which depend on the history $h \in \mathcal{H}$. At time t, the action distribution is $\pi_t(A) \triangleq \mathbb{P}^\pi(a_t \in A \mid x^t, r^t, a^{t-1})$ and the observation distribution is $\mu_t(B) \triangleq \mathbb{P}_\mu((x_{t+1}, r_{t+1}) \in B \mid x^t, r^t, a^t)$. In the Bayesian case, the agent's goal is to maximise $\mathbb{E}^\pi_\xi U = \int_{\mathcal{M}} (\mathbb{E}^\pi_\mu U) \, \mathrm{d}\xi(\mu)$, given a prior probability ξ on $\mathcal{M}$.

A fundamental difficulty is the specification of the prior and model class. While there exist a number of non-parametric Bayesian model classes, it may not be a trivial matter to select the correct class and prior. On the other hand, it is frequently known that the process can be approximated well by a complex parametrised simulator. The question is what to do when the simulator parameters are not known.

One idea is to use employ the principles of ABC (Approximate Bayesian Computation) for performing Bayesian inference using simulation. In doing so, we extend simulation-based algorithms for reinforcement learning to the case where we do not know the correct simulator parameters.

2 ABC reinforcement learning

The main idea of ABC is to approximate samples from the posterior distribution via simulation. We produce a sequence of sample models $\mu^{(k)}$ from the prior ξ, and then generate data $h^{(k)}$ from each. If the generated data is "close" to the history h, then the k-th model is accepted as a sample from the posterior $\xi(\mu \mid h)$. More specifically, we first define an approximately sufficient statistic $f : \mathcal{H} \to \mathcal{W}$ on some normed vector space $(\mathcal{W}, \|\cdot\|)$. If $\|f(h) - f(h^{(k)})\| \leq \varepsilon$ then $\mu^{(k)}$ is accepted as a sample from the posterior.

Just as in standard ABC, if f is sufficient, then the samples will be generated from the posterior. We can generalise this result by assuming that the likelihood is smooth (Lipschitz) with respect to the statistical distance:

Assumption 2.1. *For a given policy π, for any $\mu \in \mathcal{M}$, and histories $x, h \in \mathcal{H}$, there exists $L_\pi > 0$ such that $\left|\ln \left[\mathbb{P}^\pi_\mu(h) / \mathbb{P}^\pi_\mu(x)\right]\right| \leq L_\pi \|f(h) - f(x)\|$.*

Theorem 2.1. *Under a policy π and statistic f satisfying Assumption 2.1, the approximate posterior distribution $\xi_\epsilon(\cdot \mid h)$ satisfies:*

$$D\left(\xi(\cdot \mid h) \parallel \xi_\epsilon(\cdot \mid h)\right) \leq (1 + \ln |A^h_\epsilon|) L_\pi \epsilon, \tag{2.1}$$

where $A^h_\epsilon \triangleq \{ z \in \mathcal{H} \mid \|f(z) - f(h)\| \leq \epsilon \}$ is the ϵ-ball around the observed history h with respect to the statistical distance and $|A^h_\epsilon|$ denotes its size.

Although the above assumption can be relaxed in various ways, the main practical question is what statistic to use that can give good posterior approximations with little computation.

Observation-based statistics A simple idea is to select features on which to calculate statistics. Discounted cumulative feature expectation are especially interesting, due to their connection with value functions (e.g. **?**, Sec. 6.9.2). The main drawback is that this adds yet another hyper-parameter to tune. In addition, unlike econometrics or bioinformatics, we may not be interested in model identification *per se*, but only in finding a good policy.

Utility-based statistics Quantities related to the utility may be a good match for reinforcement learning. In the simplest case, it may be sufficient to only consider unconditional moments of the utility, which is the approach followed in this paper. However, these may only trivially satisfy Ass. 2.1 for arbitrary policies. Nevertheless, as we shall see, even a very simple such statistic has a reasonably good performance.

3 Conclusion

ABC-RL appears a viable approach, even with a very simple sampling scheme, and a utility-based statistic, as our recent results show. In this talk, we will present results on more elaborate ABC schemes, using statistics that are closer to sufficient, such as discounted feature expectations and conditional utilities. We shall also discuss possible extensions to the current theory.