Reinforcement Learning and the Bayesian Control Rule*

Pedro A. Ortega, Daniel A. Braun, and Simon Godsill

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK
{pao32,dab54,sjg}@cam.ac.uk

Abstract. We present an actor-critic scheme for reinforcement learning in complex domains. The main contribution is to show that planning and I/O dynamics can be separated such that an intractable planning problem reduces to a simple multi-armed bandit problem, where each lever stands for a potentially arbitrarily complex policy. Furthermore, we use the Bayesian control rule to construct an adaptive bandit player that is universal with respect to a given class of optimal bandit players, thus indirectly constructing an adaptive agent that is universal with respect to a given class of policies.

Keywords: Reinforcement learning, actor-critic, Bayesian control rule.

1 Introduction

Actor-critic (AC) methods [1] are reinforcement learning (RL) algorithms [9] whose implementation can be conceptually subdivided into two modules: the *actor*, responsible for interacting with the environment; and the *critic*, responsible for evaluating the performance of the actor. In this paper we present an AC method that conceptualizes learning a complex policy as a *multi-armed bandit problem* [2, 6] where pulling one lever corresponds to executing one iteration of a policy. The critic, who plays the role of the multi-armed bandit player, is implemented using the recently introduced *Bayesian control rule* (BCR) [8, 7]. This has the advantage of bypassing the computational costs of calculating the optimal policy by translating adaptive control into a probabilistic inference problem. The actor is implemented as a pool of parameterized policies. The scheme that we put forward here significantly simplifies the design of RL agents capable of learning complex I/O dynamics. Furthermore, we argue that this scheme offers important advantages over current RL approaches as a basis for general adaptive agents in real-world applications.

*This research was supported by the European Commission FP7-ICT, "GUIDE—Gentle User Interfaces for Disabled and Elderly citizens".

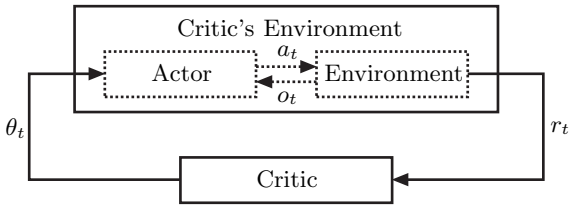

Fig. 1. The critic is an "agent" interacting with actor-environment system.

2 Setup

The interaction between the agent and the environment proceeds in cycles $t = 1, 2, \ldots$ where at cycle t, the agent produces an action $a_t \in \mathcal{A}$ that is gathered by the environment, which in turn responds with an observation $o_t \in \mathcal{O}$ and a reinforcement signal $r_t \in \mathbb{R}$ that are collected by the agent. To implement the actor-critic architecture, we introduce a signal $\theta_t \in \Theta$ generated by the critic at the beginning of each cycle, i.e. immediately before a_t, o_t and r_t are produced.

2.1 Critic

The critic is modeled as a multi-armed bandit player with a possibly (uncountably) infinite number of levers to choose from. More precisely, the critic iteratively tries out levers $\theta_1, \theta_2, \theta_3, \ldots$ so as to maximize the sum of the reinforcements $r_1, r_2, r_3, \ldots$ In this sense, the θ_t and the r_t are the critic's actions and observations respectively, not to be confused with the actions a_t and observations o_t of the actor.

According to the BCR, the critic has to sample the lever θ_t from the distribution [8]

$$P(\theta_t|\hat{\theta}_{1:t-1}, r_{1:t-1}) = \int_{\Phi} P(\theta_t|\phi, \theta_{1:t-1}, r_{1:t-1}) P(\phi|\hat{\theta}_{1:t-1}, r_{1:t-1})\, d\phi, \quad (1)$$

where the "hat"-notation $\hat{\theta}_{1:t-1}$ denotes causal intervention rather than probabilistic conditioning. The expression (1) corresponds to a weighted mixture of policies $P(\theta_t|\phi, \theta_{1:t-1}, r_{1:t-1})$ parameterized by $\phi \in \Phi$ with weights given by the posterior $P(\phi|\hat{\theta}_{1:t-1}, r_{1:t-1})$. The posterior can in turn be expressed as

$$P(\phi|\hat{\theta}_{1:t-1}, r_{1:t-1}) = \frac{P(\phi) \prod_{\tau=1}^{t} P(r_\tau|\phi, \theta_{1:\tau}, r_{1:\tau-1})}{\int_{\Phi} P(\phi') \prod_{\tau=1}^{t} P(r_\tau|\phi', \theta_{1:\tau}, r_{1:\tau-1})\, d\phi'}, \quad (2)$$

where the $P(r_t|\phi, \theta_{1:t-1}, r_{1:t-1})$ are the likelihoods of the reinforcements under the hypothesis ϕ, and where $P(\phi)$ is the prior over Φ. Note that there are no interventions on the right hand side of this equation.

Furthermore, we assume that each parameter $\phi \in \Phi$ fully determines the likelihood function $P(r_t|\phi, \theta_{1:t}, r_{1:t-1})$ representing the probability of observing the reinforcement r_t given that (an arbitrary) lever θ_t was pulled. The terms $\theta_{1:t-1}, r_{1:t-1}$ can be used to model the internal state of the bandit at cycle t. We assume that each bandit has a unique lever $\theta^\phi \in \Theta$ that maximizes the expected sum of rewards, and that the optimal strategy consists in pulling it in every time step:

$$P(\theta_t|\phi, \theta_{1:t-1}, r_{1:t-1}) = P(\theta_t|\phi) = \begin{cases} 1 & \text{if } \theta_t = \theta^\phi, \\ 0 & \text{if } \theta_t \neq \theta^\phi. \end{cases}$$

Finally, we assume a prior $P(\phi)$ over the set of operation modes Φ. This completes the specification of the critic. We will give a concrete example in Sec. 3.

2.2 Actor

The aim of the actor is to offer an rich pool of I/O dynamics parameterized by Θ that the critic can choose from. More precisely, from Fig. 1 it is seen that the actor implements the stream over the actions, i.e.

$$P(a_t|\theta_{1:t}, a_{1:t-1}, o_{1:t-1}) = P(a_t|\theta_t, a_{1:t-1}, o_{1:t-1}),$$

where we have assumed that this distribution is independent of the previously chosen parameters $\theta_{1:t-1}$. For implementation purposes it is convenient to summarize the experience $a_{1:t}, o_{1:t}$ as a sufficient statistics s^θ_{t+1} representing an internal state of the I/O dynamics $\theta \in \Phi$ at time $t+1$. States are then updated recursively as

$$s^\theta_{t+1} = f_\theta(s^\theta_t, a_t, o_t),$$

where f_θ maps the old state s^θ_t and the interaction (a_t, o_t) into the new state s^θ_{t+1}. This scheme facilitates running the different I/O dynamics in parallel. The behavior of our proposed actor-critic scheme is described in the pseudo-code listed in Alg. 1.

3 Experimental Results

We have applied the proposed scheme to a toy problem containing elements that are usually regarded as challenging in the literature: non-linear & high-dimensional dynamics and only partially observable state. The I/O domains are $\mathcal{A} = \mathcal{O} = [-1, 1]^{10}$ with reinforcements in $\mathbb{R}$.

Algorithm 1: Actor-Critic BCR

foreach $\phi \in \Phi$ **do** Set $P_1(\phi) \leftarrow P(\phi)$
foreach $\theta \in \Theta$ **do** Initialize states s_0^θ
for $t \leftarrow 1, 2, 3, \ldots$ **do**
  Sample $\phi_t \sim P_t(\phi)$
  Set $\theta_t \leftarrow \theta^{\phi_t}$
  Sample $a_t \sim P(a_t|\theta_t, s_t^{\theta_t})$
  Issue a_t and collect o_t and r_t
  foreach $\phi \in \Phi$ **do** Set $P_{t+1}(\phi) \propto P_t(\phi)P(r_t|\phi, \theta_{1:t}, r_{1:t-1})$
  foreach $\theta \in \Theta$ **do** Set $s_{t+1}^\theta \leftarrow f_\theta(s_t^\theta, a_t, o_t)$

The environment produces observations following the equation

$$[o_t, q_t]^T = f(\mu \cdot [a_t, o_{t-1}, q_{t-1}]^T),$$

where $a_t, o_t, q_t \in [-1, 1]^{10}$ are the 10-dimensional action, observation and internal state vectors respectively, μ is a 20×30 parameter matrix, and $f(\cdot)$ is a sigmoid mapping each component x into $2/(1 + e^{-x}) - 1$. Rewards are issued as $r_t = h(\theta) + \nu_t$ where h is an unknown reward mean function and ν_t is Gaussian noise with variance σ^2. Analogously, the actor implements a family of 300 different policies, where each policy is of the form

$$[a_t, s_t]^T = f(\theta \cdot [a_{t-1}, o_{t-1}, s_{t-1}]^T),$$

where $s_t \in [-1, 1]^{10}$ is the internal state vector and $\theta \in \Theta$ is the parameter matrix of the policy. These 300 matrices were sampled randomly.

The critic is modeled as a bandit player with $|\Theta| = 300$ levers to choose from, where each lever is a parameter matrix $\theta \in \Theta$. We assume that pulling lever θ produces a normally distributed reward $r \sim N(\phi_\theta, 1/\lambda)$, where $\phi_\theta \in \mathbb{R}$ is the mean specific to lever θ and where $\lambda > 0$ is a known precision term that is common to all levers and all bandits. Thus, a bandit is fully specified by the vector $\phi = [\phi_\theta]_{\theta \in \Theta}$ of all its means. To include all possible bandits we use $\Phi = \mathbb{R}^{|\Theta|}$. For each $\phi \in \Phi$, the likelihood model is

$$P(r_t|\phi, \theta_{1:t}, r_{1:t-1}) = P(r_t|\phi, \theta_t) = N(r_t; \phi_{\theta_t}, 1/\lambda),$$

and the policy is $P(\theta_t|\phi) = 1$ if $\theta_t = \arg\max_\theta \{\phi_\theta\}$ and zero otherwise. Because the likelihood is normal we place a conjugate prior $P(\phi_\theta) = N(\phi_\theta; m_\theta, 1/p_\theta)$ over Φ, where m_θ and p_θ are the mean and precision hyperparameters. This allows an easy update of the posterior after obtaining a reward [4]. To assess the performance of our algorithm, we have averaged

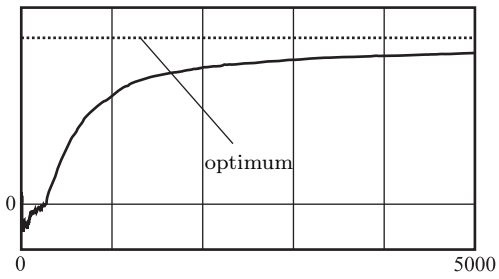

Fig. 2. Time-averaged reward of the adaptive system versus optimum performance.

a total of 100 runs with 5000 time steps. Fig. 2 shows the performance curve. It can be seen that the interaction moves from an exploratory to an exploitative phase, converging towards the optimal performance.

4 Discussion and Conclusion

The main contribution of this paper is to show how to separate the planning problem from the underlying I/O dynamics into the critic and the actor respectively, reducing reducing a complex planning problem to a simple multi-armed bandit problem. The critic is a bandit player based on the Bayesian control rule. The actor is treated as a black box, possibly implementing arbitrary complex policies.

There are important differences between our approach and other actor-critic methods. First, current actor-critic algorithms critically depend on the state-space view of the environment—see for instance [3, 9, 5]. In our opinion, this view leads to an entanglement of planning and dynamics that renders the RL problem far more difficult than necessary. Rather, we argue that this separation allows tackling domains that are intractable otherwise. Second, current reinforcement learning algorithms rely on constructing a point-estimate of the optimal policy, which is intractable when done accurately, and very costly even when only approximated. In contrast, we use the Bayesian control rule to maintain a distribution over optimal policies that is refined on-line as more observations become available. Additional experimental work is required to investigate the scalability of our actor-critic scheme to larger and more realistic domains.

References

1. A. Barto, R. Sutton, and C. Anderson. Neuron like elements that can solve difficult learning control problems. *IEEE Trans. on Systems, Man and Cybernetics*, 13, 1983.
2. D. A. Berry and B. Fristedt. *Bandit problems: Sequential allocation of experiments*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1985.
3. D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
4. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
5. M. Ghavamzadeh and Y. Engel. Bayesian actor-critic algorithms. In *Proc. of the 24th International Conference on Machine Learning*, 2007.
6. J. C. Gittins. *Multi-armed bandit allocation indices*. Wiley-Interscience Series in Systems and Optimization. John Wiley & Sons, Ltd., Chichester, 1989.
7. P. A. Ortega and D. A. Braun. A bayesian rule for adaptive control based on causal interventions. In *The third conference on artificial general intelligence*, pages 121–126, Paris, 2010. Atlantis Press.
8. P. A. Ortega and D. A. Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.
9. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.