Bayesian Reinforcement Learning + Exploration

Tor Lattimore and **Marcus Hutter**

Research School of Computer Science
Australian National University
{tor.lattimore,marcus.hutter}@anu.edu.au

July 17, 2013

1 Introduction

A reinforcement learning policy π interacts sequentially with an environment μ. In each time-step the policy π takes action $a \in \mathcal{A}$ before receiving observation $o \in \mathcal{O}$ and reward $r \in \mathcal{R}$. The goal of an agent/policy is to maximise some version of the (expected/discounted) cumulative reward. Since we are interested in the reinforcement learning problem we will assume that the true environment μ is unknown, but resides in some known set $\mathcal{M}$. The objective is to construct a single policy that performs well in some sense for all/most $\mu \in \mathcal{M}$. This challenge has been tackled for many specific $\mathcal{M}$, including bandits and factored/partially observable/regular MDPs, but comparatively few researchers have considered more general history-based environments. Here we consider arbitrary countable $\mathcal{M}$ and construct a principled Bayesian inspired algorithm that competes with the optimal policy in Cesaro average.

Notation. We assume $\mathcal{A}$, $\mathcal{O}$ and $\mathcal{R}$ are finite, but generalisations are possible. A history $x = a_1 o_1 r_2 \cdots a_t o_t r_t$ is a sequence of action/observation/reward tuples. We let $\mathcal{H}^*$ be the set of finite histories. A policy is a function $\pi : \mathcal{H}^* \to \mathcal{A}$ and an environment is a stochastic function $\mu : \mathcal{H}^* \times \mathcal{A} \rightsquigarrow \mathcal{O} \times \mathcal{R}$. Fixing a policy π and environment μ leads to a measure P^π_μ on the space of infinite history sequences. If history x is of length $t-1$, then the value function is the expected discounted reward $V^\pi_\mu(x) := \frac{1}{\Gamma_t}\mathbb{E}^\pi_\mu[\sum_{k=t}^\infty \gamma_k r_k | x]$ where $\gamma : \mathbb{N} \to [0,1]$ and $\Gamma_k := \sum_{t=k}^\infty \gamma_k$. We assume that $\sum_{t=1}^\infty \gamma_t < \infty$. The ε-effective horizon at time-step t is $H_t(\varepsilon) := \min_H \Gamma_{H+t}/\Gamma_t < \varepsilon$. The optimal policy is denoted $\pi^*_\mu := \arg\max_\pi V^\pi_\mu$, which is guaranteed to exist in this setting [LH11b].

Bayesian mixture environment. Let $w : \mathcal{M} \to (0,1]$ be a prior distribution on $\mathcal{M}$, then for fixed policy π the Bayes mixture is defined by $P^\pi_\xi := \sum_{\nu \in \mathcal{M}} w_\nu P^\pi_\nu$ and the Bayes optimal policy is denoted by π^*_ξ and studied extensively in [Hut05]. The posterior belief in environment ν having observed history x is $w_\nu(x) := w_\nu P^\pi_\nu(x)/P^\pi_\xi(x)$ where π is some policy consistent with history x.

Asymptotic optimality. We initially hoped that the Bayes optimal policy would be asymptotically optimal,

$$\forall \mu \in \mathcal{M}, \quad \lim_{t\to\infty} V^*_\mu(x_{<t}) - V^{\pi^*_\xi}_\mu(x_{<t}) = 0 \text{ with } P^\pi_\mu\text{-probability 1.}$$

But without making strong ergodicity assumptions this objective is unfortunately too strong and cannot be achieved by any policy [LH11a,Hut02]. Instead, we focus on a weaker form of optimality. A policy π is weakly asymptotically optimal if

$$\forall \mu \in \mathcal{M}, \quad \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} V_{\mu}^{*}(x_{<t}) - V_{\mu}^{\pi}(x_{<t}) = 0 \text{ with } P_{\mu}^{\pi}\text{-probability } 1.$$

It is known that π_{ξ}^{*} is not weakly asymptotically optimal [Ors13]. We modify the policy slightly by adding exploration periods based on maximising the information gain and show that the modified policy is weakly asymptotically optimal.

Information gain. Let $d \in \mathbb{N}$ and x be a finite history. Then the d-step P_{ξ}^{π}-expected information gain is defined

$$\mathbb{E}_{\xi}^{\pi}[\mathrm{IG}_d \,|x] := \sum_{y \in \mathcal{H}^d} P_{\xi}^{\pi}(y|x) \underbrace{\sum_{\nu \in \mathcal{M}} w_{\nu}(xy) \log \frac{w_{\nu}(xy)}{w_{\nu}(x)}}_{\text{information gain}} = \sum_{\nu \in \mathcal{M}} w_{\nu}(x) \sum_{y \in \mathcal{H}^d} P_{\nu}^{\pi}(y|x) \log \frac{P_{\nu}^{\pi}(y|x)}{P_{\xi}^{\pi}(y|x)}$$

Algorithm. The new algorithm (Bayes+Exp) accepts parameters $\vec{\varepsilon}_t$, $\vec{d}_t$ and operates in phases. If at time-step t there exists a policy π such that the P_{ξ}^{π}-expected d_t-step information gain is larger than ε_t, then π is followed for d_t time-steps. Otherwise the Bayes optimal policy π_{ξ}^{*} is followed.

Algorithm 1 Bayes+Exp

inputs: $\mathcal{M} = \{\nu_1, \nu_2, \cdots\}$, $\vec{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \cdots\}$ and $\vec{d} = \{d_1, d_2, \cdots\}$
loop
  $x \leftarrow$ current history and $t \leftarrow$ current time
  $\Delta \leftarrow \max_{\pi} \mathbb{E}_{\xi}^{\pi}[\mathrm{IG}_{d_t} \,|x]$ and $\pi \leftarrow \arg\max_{\pi} \mathbb{E}_{\xi}^{\pi}[\mathrm{IG}_{d_t} \,|x]$
  if $\Delta > \varepsilon_t$ **then**
    Follow π for d_t time-steps
  else
    Follow π_{ξ}^{*} for 1 time-step

2 Results

Theorem 1. *We proved the following:*

1. *If $H_t(\varepsilon) \in o(t)$ for all ε, then there exist parameters $\vec{d}$ and $\vec{\varepsilon}$ such that Bayes+Exp is weakly asymptotically optimal.*
2. *If $H_t(\varepsilon) \in \Omega(t)$ for all ε, then there exists an $\mathcal{M}$ such that no policy is weakly asymptotically optimal.*

The trick behind the proof of the first item is the fact that with probability 1 the expected cumulative information gain is bounded by some finite quantity dependent on the true unknown environment. Then, for carefully chosen $\vec{\varepsilon}$ the algorithm will explore sufficiently infrequently to guarantee that it is exploiting most of the time. On the other hand, if the expected information gain is small, then it can be shown that π_{ξ}^{*} is close to optimal. So by sending $\varepsilon_t \to 0$ sufficiently slowly we can guarantee both asymptotic optimality while exploiting combined with a small number of exploration periods. The horizon d_t is chosen to be proportional to $H_t(\varepsilon_t)$.

If the horizon grows too fast, then the idea above fails. If a policy π is to be weakly asymptotically optimal it must never stop exploring. If the horizon grows with order t, then an exploration period must be at least $\Omega(H_t(\varepsilon_t))$ time-steps long, but if $H_t(\varepsilon_t)$ grows linearly with t, then every exploration phase resets the average and ensures eternal sub-optimality. The proof in the general case requires a tricky counter-example for which we have insufficient space in this abstract.

References

[Hut02] Marcus Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Proc. 15th Annual Conf. on Computational Learning Theory (COLT'02)*, volume 2375 of *LNAI*, pages 364–379, Sydney, 2002. Springer, Berlin.

[Hut05] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.

[LH11a] Tor Lattimore and Marcus Hutter. Asymptotically optimal agents. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011.

[LH11b] Tor Lattimore and Marcus Hutter. Time consistent discounting. In Jyrki Kivinen, Csaba Szepesvári, Esko Ukkonen, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 6925 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011.

[Ors13] Laurent Orseau. Asymptotic non-learnability of universal agents with computable horizon functions. *Theoret. Comput. Sci.*, 473:149–156, 2013. Special Issue on Learning Theory.