Bayesian Reinforcement Learning with Exploration

Tor Lattimore[1] and Marcus Hutter[2]

[1] University of Alberta
tor.lattimore@gmail.com
[2] Australian National University
marcus.hutter@anu.edu.au

Abstract. We consider a general reinforcement learning problem and show that carefully combining the Bayesian optimal policy and an exploring policy leads to minimax sample-complexity bounds in a very general class of (history-based) environments. We also prove lower bounds and show that the new algorithm displays adaptive behaviour when the environment is easier than worst-case.

1 Introduction

We study the question of finding the minimax sample-complexity of reinforcement learning without making the usual Markov assumption, but where the learner has access to a finite set of reinforcement learning environments to which the truth is known to belong. This problem was tackled previously by Dyagilev et al. (2008) and Lattimore et al. (2013a). The new algorithm improves on the theoretical results in both papers and is simultaneously simpler and more elegant. Unlike the latter work, in certain circumstances the new algorithm enjoys adaptive sample-complexity bounds when the true environment is benign. We show that if $\mathcal{M} = \{\mu_1, \cdots, \mu_K\}$ is a carefully chosen finite set of history-based reinforcement learning environments, then every algorithm is necessarily ε-suboptimal for $\Omega\left(\frac{K}{\varepsilon^2(1-\gamma)^3}\log\frac{K}{\delta}\right)$ time-steps with probability at least δ where γ is the discount factor. The algorithm presented has a sample-complexity bound equal to that bound except for one factor of $\log\frac{1}{\varepsilon(1-\gamma)}$, so the minimax sample-complexity of this problem is essentially known.

Aside from the previously mentioned papers, there has been little work on this problem, although sample-complexity bounds have been proven for MDPs (Lattimore and Hutter, 2012; Szita and Szepesvári, 2010; Kearns and Singh, 2002, and references there-in), as well as partially observable and factored MDPs (Chakraborty and Stone, 2011; Even-Dar et al., 2005). There is also a significant literature on the regret criterion for MDPs (Azar et al., 2013; Auer et al., 2010, and references there-in), but meaningful results cannot be obtained without a connectedness assumption that we avoid here. Regret bounds are known if the true environment is finite-state, Markov and communicating, but where the state is not observed directly (Odalric-Ambrym et al., 2013). Less restricted settings have also

P. Auer et al. (Eds.): ALT 2014, LNAI 8776, pp. 170–184, 2014.

been studied. Sunehag and Hutter (2012) proved sample-complexity bounds for the same type of reinforcement learning problems that we do, but only for deterministic environments (for the stochastic case they gave asymptotic results). Also similar is the k-meteorologist problem studied by Diuk et al. (2009), but they consider only the 1-step problem, which is equivalent to the case where the discount factor $\gamma = 0$. In that case their algorithm is comparable to the one developed by Lattimore et al. (2013a) and suffers from the same drawbacks, most notable of which is non-adaptivity. A more detailed discussion is given in the conclusion. Recently there has been a growing interesting in algorithms based on the "near-Bayesian" Thompson sampling. See, for example, the work by Osband et al. (2013) and references there-in. Note that the aforementioned paper deals with a Bayesian regret criterion for MDPs, rather than the frequentist sample-complexity results presented here.

The new algorithm is based loosely on the universal Bayesian optimal reinforcement learning algorithm studied in depth by Hutter (2005). Unfortunately, a pure Bayesian approach may not explore sufficiently to enjoy a finite sample-complexity bound (Orseau, 2010) (some exceptions by Hutter (2002)). For this reason we add exploration periods to ensure that sufficient exploration occurs for sample-complexity bounds to become possible.

2 Notation

Due to lack of space, many of the easier proofs or omitted, along with results that are periphery to the main bound on the sample-complexity. All proofs can be found in the technical report (Lattimore and Hutter, 2014).

Strings/Sequences. A finite string of length n over non-empty alphabet $\mathcal{H}$ is a finite sequence $x_1x_2x_3\cdots x_n$ where $x_k \in \mathcal{H}$. An infinite sequence over $\mathcal{H}$ is a sequence $x_1x_2x_3\cdots$. The set of sequences over alphabet $\mathcal{H}$ of length n is denoted by $\mathcal{H}^n$. The set of finite sequences over alphabet $\mathcal{H}$ is denoted by $\mathcal{H}^* := \bigcup_{n=0}^{\infty} \mathcal{H}^n$. The set of sequences of length at most n is $\mathcal{H}^{\leq n} := \bigcup_{k=0}^{n} \mathcal{H}^k$. The uncountable set of infinite sequences is $\mathcal{H}^\infty$. For $x \in \mathcal{H}^* \cup \mathcal{H}^\infty$, the length of x is $\ell(x)$. The empty string of length zero is denoted by ϵ, which should not be confused with small constants denoted by ε. Subsequences are $x_{1:t} := x_1x_2x_3\cdots x_t$ and $x_{<t} := x_{1:t-1}$. We say x is a prefix of y and write $x \sqsubseteq y$ if $\ell(x) \leq \ell(y)$ and $x_k = y_k$ for all $k \leq \ell(x)$. The words string and sequence are used interchangeably, although the former is more likely to be finite and the latter more likely to be infinite. Strings may be concatenated in the obvious way. If $x \in \mathcal{H}^*$, then x^k is defined to be k concatenations of x. A set $A \subset \mathcal{H}^*$ is prefix free if for all $x, y \in \mathcal{H}$, $x \sqsubseteq y \implies x = y$. A prefix free set A is complete if for all infinite histories $y \in \mathcal{H}^\infty$ there exists an $x \in A$ such that $x \sqsubseteq y$.

History Sequences. Let $\mathcal{A}$, $\mathcal{O}$ and $\mathcal{R} \subset [0,1]$ be finite sets of actions, observations and rewards respectively and $\mathcal{H} := \mathcal{A} \times \mathcal{O} \times \mathcal{R}$. The set of infinite history sequences is denoted $\mathcal{H}^\infty$ while $\mathcal{H}^*$ is the set of all finite-length histories. The action/observation/reward at time-step t of history x are denoted by $a_t(x)$, $o_t(x)$, $r_t(x)$ respectively.

Environments and Policies. An environment μ is a set of conditional probability distributions $\mu(\cdot|x,a) : \mathcal{R} \times \mathcal{O} \to [0,1]$ where $x \in \mathcal{H}^*$ is a finite history and $a \in \mathcal{A}$ is an action. The value $\mu(r,o|x,a)$ is the probability of environment μ generating reward $r \in \mathcal{R}$ and observation $o \in \mathcal{O}$ given finite history $x \in \mathcal{H}^*$ has occurred and action $a \in \mathcal{A}$ has just been taken by the agent. A deterministic policy is a function $\pi : \mathcal{H}^* \to \mathcal{A}$ where $\pi(x)$ is the action taken by policy π given history x. The space of all deterministic policies is denoted by Π. A deterministic policy π is consistent with history $x \in \mathcal{H}^*$ if $\pi(x_{<t}) = a_t(x)$ for all $t \leq \ell(x)$. The set of policies consistent with history x is denoted by $\Pi(x)$.

Probability Spaces. A policy and environment interact sequentially to stochastically generate infinite histories. In order to be rigorous, it is necessary to define a (filtered) probability space on the set of infinite histories $\mathcal{H}^\infty$. Let $x \in \mathcal{H}^*$ be a finite history, then $\Gamma_x := \{y \in \mathcal{H}^\infty : x \sqsubseteq y\}$ is the set of all infinite histories starting with x and is called the cylinder set of x. Now define σ-algebras generated by the cylinders of $\mathcal{H}^*$ and $\mathcal{H}^t$ by $\mathcal{F} := \sigma(\{\Gamma_x : x \in \mathcal{H}^*\})$ and $\mathcal{F}_{<t} := \sigma(\{\Gamma_x : x \in \mathcal{H}^{t-1}\})$. Then $(\mathcal{H}^\infty, \mathcal{F}, \{\mathcal{F}_{<t}\})$ is a filtered probability space. Throughout we use the convention that time starts at 1 with the empty history. An environment and policy interact sequentially to induce a measure $\mu^\pi : \mathcal{F} \to [0,1]$ on the filtered probability space $(\mathcal{H}^\infty, \mathcal{F}, \{\mathcal{F}_{<t}\})$. If $A \in \mathcal{F} \subseteq \mathcal{H}^\infty$, then $\mu^\pi(A)$ is the probability of the event A occurring. As is common in the literature, we abuse notation and use the short-hand $\mu^\pi(x) := \mu^\pi(\Gamma_x)$. If $x, y \in \mathcal{H}^*$, then conditional probabilities are $\mu^\pi(y|x) := \mu^\pi(xy)/\mu^\pi(x)$. Expectations with respect to μ^π are denoted by $\mathbb{E}^\pi_\mu$. If ρ is any measure on $(\mathcal{H}^\infty, \mathcal{F}, \{\mathcal{F}_{<t}\})$, then we define useful random variables:

$$\rho_{<t}(x) := \rho(x_{<t}) \qquad \rho_{1:t}(x) := \rho(x_{1:t}) \qquad \rho_{t:t+d}(x) := \frac{\rho(x_{1:t+d})}{\rho(x_{<t})}.$$

Discounting and Value Functions. Let $\gamma \in [0,1)$ be the discount factor, then the discounted value of history x is the expected discounted cumulative reward.

$$V^\pi_\mu(x;d) := \mathbb{E}^\pi_\mu \left[\sum_{t=\ell(x)+1}^{\ell(x)+d} \gamma^{t-\ell(x)-1} r_t \middle| x \right] \qquad V^\pi_\mu(x) := \lim_{d\to\infty} V^\pi_\mu(x;d),$$

where d is a horizon after which reards are not counted and we assume that $0^0 = 1$ when $\gamma = 0$. The optimal policy in environment μ is $\pi^*_\mu := \arg\max_{\pi\in\Pi} V^\pi_\mu(\epsilon)$. Since rewards are bounded in $[0,1]$ and values are discounted, the value function is also bounded: $V^\pi_\mu(x) \in [0, \frac{1}{1-\gamma}]$. The value of the optimal policy in environment μ and having observed history x is $V^*_\mu(x)$. Since the discount factor does not vary within the results we omit it from the notation for the value function, but it is important to note that all values depend on this quantity.

3 Algorithm

To begin, we consider only the prediction problem where π is fixed, but μ is unknown and the task is to predict future observations and rewards given the history. We assume that π is some fixed policy and that $\mu \in \mathcal{M} = \{\nu_1 \cdots \nu_K\}$ where

$\mathcal{M}$ is known, but not μ. The Bayesian mixture measure is $\xi^\pi := \sum_{\nu\in\mathcal{M}} w_\nu \nu^\pi$ where $w : \mathcal{M} \to [0,1]$ is a probability distribution on $\mathcal{M}$. The Bayesian optimal policy is defined by $\pi^*_\xi := \arg\max_\pi V^\pi_\xi(\epsilon) \equiv \arg\max_\pi \sum_{\nu\in\mathcal{M}} w_\nu V^\pi_\nu(\epsilon)$. It is reasonably well-known that the predictive distribution of the Bayesian mixture converges almost surely to the truth for all μ, and that it does so fast with respect to a variety of different metrics. To measure convergence we define the d-step total variation and squared Hellinger distances between predictive distributions of ξ^π and ν^π given the history at time-step t.

$$\delta^d_x(\nu^\pi,\xi^\pi) := \frac{1}{2}\sum_{y\in\mathcal{H}^d} |\nu^\pi(y|x) - \xi^\pi(y|x)| \qquad \delta^d_t(\nu^\pi,\xi^\pi)(x) := \delta^d_{x_{<t}}(\nu^\pi,\xi^\pi)$$

$$h^d_x(\nu^\pi,\xi^\pi) := \frac{1}{2}\sum_{y\in\mathcal{H}^d} \left(\sqrt{\nu^\pi(y|x)} - \sqrt{\xi^\pi(y|x)}\right)^2 \quad h^d_t(\nu^\pi,\xi^\pi)(x) := h^d_{x_{<t}}(\nu^\pi,\xi^\pi).$$

where the distances on the right hand side are defined as random variables. The following theorem by Hutter and Muchnik (2007) will be useful.

Theorem 1. *If $\mu \in \mathcal{M}$, then* $\mathbb{E}^\pi_\mu \exp\left(\frac{1}{2}\sum_{t=1}^\infty h^1_t(\mu^\pi,\xi^\pi)\right) \le \sqrt{\frac{1}{w_\mu}}$.

More usual than the Hellinger distance in the analysis of Bayesian sequence prediction is the relative entropy, but this quantity is unbounded, which somewhat surprisingly leads to weaker results (Lattimore et al., 2013b). The following theorem is a simple generalisation of Theorem 1 to the multi-step case.

Theorem 2. *Let $d \ge 1$ and $\{\tau_k\}_{k=1}^\infty$ be a sequence of $(\mathcal{H}^\infty, \mathcal{F}, \{\mathcal{F}_{<t}\})$-measurable stopping times such that $\tau_k + d \le \tau_{k+1}$ for all k. Then for all $\mu \in \mathcal{M}$*

$$\mathbb{E}^\pi_\mu \exp\left(\frac{1}{2}\sum_{k=1}^\infty h^d_{\tau_k}(\mu^\pi,\xi^\pi)\right) \le \sqrt{\frac{1}{w_\mu}}.$$

Theorem 1 is regained by choosing $\tau_k = k$ and $d = 1$. The proof of Theorem 2 can be found in the technical report. Theorem 2 shows that the predictive distribution of the Bayesian mixture converges fast to the true predictive distribution. In particular, with high probability the cumulative squared total-variation distance does not greatly exceed $\log \frac{1}{w_\mu}$.

Corollary 3. *If $\delta > 0$, then* $\mu^\pi\left(\sum_{k=1}^\infty \delta^d_{\tau_k}(\mu^\pi,\xi^\pi)^2 \ge \log\frac{1}{w_\mu} + \log\frac{1}{\delta^2}\right) \le \delta$.

Proof. We combine Markov's inequality with Theorem 2.

$$\mu^\pi\left(\sum_{k=1}^\infty \delta^d_{\tau_k}(\mu^\pi,\xi^\pi)^2 \ge \log\frac{1}{w_\mu} + \log\frac{1}{\delta^2}\right) \overset{(a)}{\le} \mu^\pi\left(\sum_{k=1}^\infty h^d_{\tau_k}(\mu^\pi,\xi^\pi) \ge \log\frac{1}{w_\mu\delta^2}\right)$$
$$\overset{(b)}{=} \mu^\pi\left(\exp\left(\frac{1}{2}\sum_{k=1}^\infty h^d_{\tau_k}(\mu^\pi,\xi^\pi)\right) \ge \frac{1}{\delta}\sqrt{\frac{1}{w_\mu}}\right) \overset{(c)}{\le} \delta,$$

where (a) follows since the Hellinger distance upper bounds the total variation distance, (b) is trivial, and (c) by Markov's inequality. □

The consequence of the above is that a Bayesian predictor quickly learns the true distribution of the rewards and observations it will receive. On first sight this might seem promising for Bayesian reinforcement learning, but there is a problem. Bayesian sequence prediction is only capable of learning to predict given a fixed policy. But in RL the agent must choose its action at each time-step, and to do this effectively it must be able to predict the consequences of *all* actions, not only the action it ultimately ends up taking. We side-step this problem in the new algorithm called BayesExp by only following the Bayesian optimal policy when it is guaranteed to be nearly optimal and exploring otherwise. The BayesExp algorithm is as follows:

Algorithm 1. BayesExp

Inputs: ε, δ and $\mathcal{M} = \{\nu_1, \nu_2, \cdots, \nu_K\}$
$\delta_1 \leftarrow \delta/2$ and $\varepsilon_1 \leftarrow \varepsilon(1-\gamma)/4$ and $\varepsilon_2 \leftarrow \varepsilon/12$ and $d \leftarrow \frac{\log \varepsilon_2(1-\gamma)}{\log \gamma}$
$x \leftarrow \epsilon$ and $t \leftarrow 1$ and $w_\nu \leftarrow 1/K$ and $D(\nu) \leftarrow 0$, $\forall \nu$
loop
  $\Pi^* \leftarrow \{\pi^*_\nu : \nu \in \mathcal{M}\} \cup \{\pi^*_\xi\}$
  $\pi \leftarrow \arg\max_{\pi \in \Pi^*} \max_{\nu \in \mathcal{M}} \delta^d_x(\nu^\pi, \xi^\pi)$
  $\Delta \leftarrow \max_{\pi \in \Pi^*, \nu \in \mathcal{M}} \delta^d_x(\nu^\pi, \xi^\pi)$
  if $\Delta > \varepsilon_1$ **then**
    $D(\nu) \leftarrow D(\nu) + \delta^d_x(\nu^\pi, \xi^\pi)^2$, $\forall \nu$
    for $j = 1 \to d$ **do**
      ACT(π)
    $\mathcal{M} \leftarrow \left\{\nu : D(\nu) \leq \log K/\delta_1^2\right\}$
  else
    $D(\nu) \leftarrow D(\nu) + \delta^1_x(\nu^{\pi^*_\xi}, \xi^{\pi^*_\xi})^2$, $\forall \nu$
    ACT(π^*_ξ)
function ACT(π)
  Take action $a = \pi(x)$ and observe $o \in \mathcal{O}$ and $r \in \mathcal{R}$ from environment
  $t \leftarrow t+1$ and $x \leftarrow xaor$

Indices. For the sake of readability the time indices have been omitted in the pseudo-code above. Throughout the analysis we write $\mathcal{M}_t$, $D_t(\nu)$ and Δ_t for the values of Δ, $D(\nu)$ and $\mathcal{M}$ as computed by BayesExp at time-step t. Similarly, $\mathcal{M}_z$, $D_z(\nu)$ and Δ_z are the values of $\mathcal{M}$, $D(\nu)$ and Δ respectively as they would be computed given the algorithm had reached history $z \in \mathcal{H}^*$.

Exploration Phases. The algorithm operates in phases of exploration and exploitation. If there exists an optimal policy π' with respect to some plausible environment such that the d-step total-variation distance between $\nu^{\pi'}$ and $\xi^{\pi'}$ is larger than ε_1, then the algorithm follows π' for exactly d time-steps. This period is called an exploration phase. The set of time-steps triggering exploring phases is denoted by $E \subseteq \mathbb{N}$. While the set of time-steps spent in exploration phases is denoted by $E_d := \bigcup_{t \in E} \{t, t+1, \cdots, t+d-1\}$.

Exploitation Time-Steps. If BayesExp is not exploring at time-step t, then t is an exploiting time-step where BayesExp is following the Bayes optimal policy. The set of all exploitation time-steps is denoted by $T := \mathbb{N} - E_d$.

Failure Phases. For the remainder of this section the policy π refers to the policy of BayesExp. A failure phase is a period of d time-steps triggered at time-step t provided t is not part of a previous exploration/failure phase and $\mu \in \mathcal{M}_t$ and $V^*_\mu(x_{<t}) - V^\pi_\mu(x_{<t}) > \varepsilon$. We denote the set of time-steps triggering failure phases by $F \subset \mathbb{N}$ and the set of time-steps spent in failure phases by $F_d := \bigcup_{t \in F} \{t, t+1, \cdots, t+d-1\}$. Failure phases depend on the unknown μ, so are not known to the algorithm and are only used in the analysis.

4 Upper Bound on Sample-Complexity

Theorem 4. *Suppose π is the policy of Algorithm 1 given input $\varepsilon > 0$, $\delta > 0$ and $\mathcal{M} = \{\nu_1, \cdots, \nu_K\}$. If $\mu \in \mathcal{M}$, then*

$$\mu^\pi \left(\sum_{t=1}^{\infty} \mathbb{1}\left\{ V^*_\mu(x_{<t}) - V^\pi_\mu(x_{<t}) > \varepsilon \right\} > \frac{416 K d}{\varepsilon^2 (1-\gamma)^2} \log \frac{4K}{\delta^2} \right) \leq \delta$$

where x is the infinite history sampled from μ^π and $d = \frac{\log(\varepsilon_2(1-\gamma))}{\log \gamma}$ is the effective horizon.

Noting that $d \in O(\frac{1}{1-\gamma} \log \frac{1}{\varepsilon(1-\gamma)})$, the sample-complexity is bounded by

$$O\left(\frac{K}{\varepsilon^2 (1-\gamma)^3} \left(\log \frac{1}{\varepsilon(1-\gamma)} \right) \left(\log \frac{K}{\delta} \right) \right).$$

Proof Overview

(a) By definition, if $V^*_\mu(x_{<t}) - V^\pi_\mu(x_{<t}) > \varepsilon$, then either $\mu \notin \mathcal{M}_t$ or t is part of an exploration/failure phase, $t \in E_d \cup F_d$.

(b) First we show that $\mu \in \mathcal{M}_t$ for all t with probability at least $1 - \delta_1$.

(c) We then use the definition of the algorithm to bound

$$|E| \leq \frac{K}{\varepsilon_1^2} \log \frac{K}{\delta_1^2} \implies |E_d| \leq \frac{K d}{\varepsilon_1^2} \log \frac{K}{\delta_1^2}.$$

(d) If $\mu \in \mathcal{M}_t$ and BayesExp is exploiting, then all plausible environments are sufficiently close under all optimal policies and so

$$V^*_\mu(x_{<t}) - V^{\pi^*_\xi}_\mu(x_{<t}) \lesssim \varepsilon. \tag{1}$$

(e) Unfortunately (1) does not imply that $V^*_\mu(x_{<t}) - V^\pi_\mu(x_{<t}) \leq \varepsilon$. A careful argument is required to ensure that the number of errors in exploitation periods is also small, which essentially means bounding the number of failure phases. This eventually follows from the fact that if BayesExp is sub-optimal while exploiting, then there must be some probability of triggering an exploration phase, which cannot happen too often.

The following lemmas are required for the proof of Theorem 4 and could be skipped until they are referred to.

Lemma 5. *Suppose t is a time-step when BayesExp is exploiting given history $x_{<t}$. Then* $V_\mu^{\pi_\xi^*}(x_{<t}) - V_\mu^\pi(x_{<t}) \leq \sum_{y \in Y} \mu^\pi(y|x_{<t}) \left(V_\mu^{\pi_\xi^*}(x_{<t}y) - V_\mu^\pi(x_{<t}y) \right) + \varepsilon_2$, *where Y is the set of finite history sequences y of length at most d such that BayesExp would explore given history $x_{<t}y$.*

$$Y = \left\{ y \in \mathcal{H}^{\leq d} : \textit{BayesExp explores given history } x_{<t}y \right\}.$$

Proof. Define $\bar{Y} = Y \cup \left\{ y \in \mathcal{H}^d : \forall z \in Y, z \not\sqsubseteq y \right\}$, which is complete and prefix free by definition. Since t is an exploitation time-step, BayesExp will follow policy π_ξ^* until such a time as it starts an exploration phase. Therefore by Lemma 13

$$\begin{aligned} V_\mu^{\pi_\xi^*}(x_{<t}) - V_\mu^\pi(x_{<t}) &\overset{(a)}{=} \sum_{y \in \bar{Y}} \mu^\pi(y|x_{<t}) \gamma^{\ell(y)} \left(V_\mu^{\pi_\xi^*}(x_{<t}y) - V_\mu^\pi(x_{<t}y) \right) \\ &\overset{(b)}{\leq} \sum_{y \in Y} \mu^\pi(y|x_{<t}) \left| V_\mu^{\pi_\xi^*}(x_{<t}y) - V_\mu^\pi(x_{<t}y) \right| + \varepsilon_2 \end{aligned}$$

where (a) follows from Lemma 13. (b) by dropping all $y \in \bar{Y} - Y$ and using the fact that for $y \in \bar{Y} - Y$ we have $\ell(y) = d$, which by the definition of the horizon $d = \frac{\log \varepsilon_2 (1-\gamma)}{\log \gamma}$ implies that the ratio $\gamma^d \leq \varepsilon_2 (1-\gamma)$. □

Lemma 6. *Let $x_{<t}$ be the history at an exploitation time-step $t \in T$ and assume $\mu \in \mathcal{M}_t$. Then* $V_\mu^*(x_{<t}; d) - V_\mu^{\pi_\xi^*}(x_{<t}; d) \leq \dfrac{2\varepsilon_1}{1-\gamma}$.

Proof. Since t is an exploitation time-step we have that $\Delta_t \leq \varepsilon_1$. Therefore

$$\begin{aligned} &V_\mu^*(x_{<t}; d) - V_\mu^{\pi_\xi^*}(x_{<t}; d) \overset{(a)}{=} V_\mu^{\pi_\mu^*}(x_{<t}; d) - V_\mu^{\pi_\xi^*}(x_{<t}; d) \\ &\overset{(b)}{\leq} V_\mu^{\pi_\mu^*}(x_{<t}; d) - V_\xi^{\pi_\mu^*}(x_{<t}; d) + V_\xi^{\pi_\mu^*}(x_{<t}; d) - V_\xi^{\pi_\xi^*}(x_{<t}; d) \\ &\qquad + V_\xi^{\pi_\xi^*}(x_{<t}; d) - V_\mu^{\pi_\xi^*}(x_{<t}; d) \\ &\overset{(c)}{\leq} \frac{1}{1-\gamma} \left(\delta^d_{x_{<t}}(\mu^{\pi_\mu^*}, \xi^{\pi_\mu^*}) + \delta^d_{x_{<t}}(\mu^{\pi_\xi^*}, \xi^{\pi_\xi^*}) \right) \overset{(d)}{\leq} \frac{2\Delta_t}{1-\gamma} \overset{(e)}{\leq} \frac{2\varepsilon_1}{1-\gamma} \end{aligned}$$

where (a) is the definition of $V_\mu^*(x_{<t}; d)$. (b) by adding and subtracting value functions. (c) by Lemma 11. (d) by the definition of Δ_t and $\mu \in \mathcal{M}_t$. (e) since t is an exploitation time-step. □

Lemma 7. *Let $x_{<t}$ be the history at an exploration time-step $t \in E$. Then*

$$V_\mu^*(x_{<t}; d) - V_\mu^\pi(x_{<t}; d) \leq \frac{\max\{4\Delta_t, \mathbb{1}\{\mu \notin \mathcal{M}_t\}\}}{1-\gamma}.$$

Proof. If $\mu \notin \mathcal{M}_t$, then we use the trivial bound of $\frac{1}{1-\gamma}$. Now assume $\mu \in \mathcal{M}_t$ and let $\pi^*_\rho = \arg\max_{\pi \in \Pi^*_t} \max_{\nu \in \mathcal{M}_t} \delta^d_t(\nu^{\pi^*_\rho}, \xi^{\pi^*_\rho})$, which means that $\rho \in \mathcal{M}_t \cup \{\xi\}$. Therefore

$$
\begin{aligned}
V^*_\mu(x_{<t}; d) - V^\pi_\mu(x_{<t}; d) &\overset{(a)}{=} V^{\pi^*_\mu}_\mu(x_{<t}; d) - V^{\pi^*_\rho}_\mu(x_{<t}; d) \\
&\overset{(b)}{\leq} V^{\pi^*_\mu}_\rho(x_{<t}; d) - V^{\pi^*_\rho}_\rho(x_{<t}; d) + \left(V^{\pi^*_\mu}_\mu(x_{<t}; d) - V^{\pi^*_\mu}_\rho(x_{<t}; d)\right) \\
&\quad + \left(V^{\pi^*_\rho}_\rho(x_{<t}; d) - V^{\pi^*_\rho}_\mu(x_{<t}; d)\right) \\
&\overset{(c)}{\leq} \frac{1}{1-\gamma}\left(\delta^d_{x_{<t}}(\rho^{\pi^*_\mu}, \mu^{\pi^*_\mu}) + \delta^d_{x_{<t}}(\rho^{\pi^*_\rho}, \mu^{\pi^*_\rho})\right) \\
&\overset{(d)}{\leq} \frac{1}{1-\gamma}\left(\delta^d_{x_{<t}}(\rho^{\pi^*_\mu}, \xi^{\pi^*_\mu}) + \delta^d_{x_{<t}}(\xi^{\pi^*_\mu}, \mu^{\pi^*_\mu}) + \delta^d_{x_{<t}}(\rho^{\pi^*_\rho}, \xi^{\pi^*_\rho}) + \delta^d_{x_{<t}}(\xi^{\pi^*_\rho}, \mu^{\pi^*_\rho})\right) \\
&\overset{(e)}{\leq} \frac{4\Delta_t}{1-\gamma}
\end{aligned}
$$

where (a) follows since BayesExp follows policy π^*_ρ while exploring. (b) by expanding the values. (c) by Lemma 11. (d) by the triangle inequality. (e) by the definition of Δ_t and because $\rho, \mu \in \mathcal{M} \cup \{\xi\}$. □

The proof of Theorem 4 uses a number of constants that are functions of each other. For convenience they are described in the table below.

Table 1. Constants for Theorem 4

constant	ε_1	ε_2	ε_3	ε_4	δ_1	d
constraint			$= \varepsilon_2 + \frac{2\varepsilon_1}{1-\gamma}$	$= (\varepsilon - \varepsilon_3 - 2\varepsilon_2)(1-\gamma)$		
value	$\varepsilon(1-\gamma)/4$	$\varepsilon/12$	$7\varepsilon/12$	$\varepsilon(1-\gamma)/4$	$\delta/2$	$\frac{\log \varepsilon_2}{\log \gamma}$

Proof (of Theorem 4). Following the plan, we start by bounding the probability that μ is removed from $\mathcal{M}_t$.

Step 1: Bounding Inconsistency Probability. Let A_1 be the event that $\mu \in \mathcal{M}_t$ for all time-steps t. Environment μ is removed from the model class $\mathcal{M}_t$ only once the counter $D(\mu)$ exceeds $\log K/\delta_1^2$. But $D(\mu)$ is the cumulative squared total variation distance between μ^π and ξ^π, which by Corollary 3 is bounded by $\log K/\delta_1^2$ with μ^π-probability at least $1-\delta_1$ and so $\mu^\pi(A_1) \geq 1-\delta_1$.

Step 2: Bounding Exploration Phases. Let t be the start of an exploration phase. Then by definition there exists a $\nu \in \mathcal{M}_t$ such that $\delta^d_t(\nu^\pi, \xi^\pi) > \varepsilon_1$ and so $D(\nu)$ is incremented by at least ε_1^2. Since an environment is removed from $\mathcal{M}$ once $D(\nu)$ exceeds $\log K/\delta_1^2$, the number of exploration phases is bounded by $E_{\max} := \frac{K}{\varepsilon_1^2} \log \frac{K}{\delta_1^2}$. Since each exploration phase is exactly d time-steps long, the number of time-steps spent in exploration phases satisfies

$$
|E_d| \leq \frac{Kd}{\varepsilon_1^2} \log \frac{K}{\delta_1^2}. \tag{2}
$$

By identical reasoning it holds that $\sum_{t \in E} \Delta_t^2 \leq K \log \frac{K}{\delta_1^2}$. (3)

Note that both (2) and (3) hold surely over all history trajectories.

Step 3: Exploitation Success. Assume that event A_1 is true, which means that $\mu \in \mathcal{M}_t$ for all time-steps. Let $t \in T$ be a time-step when BayesExp is exploiting. Therefore

$$V_\mu^*(x_{<t}) - V_\mu^{\pi_\xi^*}(x_{<t}) \overset{(a)}{\leq} \varepsilon_2 + V_\mu^*(x_{<t}; d) - V_\mu^{\pi_\xi^*}(x_{<t}; d) \overset{(b)}{\leq} \frac{2\varepsilon_1}{1-\gamma} + \varepsilon_2 =: \varepsilon_3 < \varepsilon$$

where (a) follows by truncating the horizon (Lemma 12) and (b) by Lemma 6.

Step 4: Connecting the Policies. We now bound the number of failure phases. The intuition is that if BayesExp is exploiting at time-step t, then the Bayes-optimal policy π_ξ^* is near-optimal. Since BayesExp follows this policy until an exploration phase, $V_\mu^*(x) - V_\mu^\pi(x)$ can only be large if there is a reasonable probability of encountering an exploration phase within the next d time-steps. By some form of concentration inequality this cannot happen too often before an exploration phase actually occurs, which will lead to the correct bound on the number of time-steps when $V_\mu^*(x_{<t}) - V_\mu^\pi(x_{<t}) > \varepsilon$. Let $F = \{t_1, t_2, \cdots\}$ be the set of time-steps triggering failure phases with corresponding histories $x_{<t_k}$. For $k > |F|$ define $t_k = \infty$. At time-step t_k having observed history $x_{<t_k}$ define Y as in the statement of Lemma 5 to be the set of finite histories of length at most d such that BayesExp would explore upon reaching history $x_{<t_k} y$.

$$Y := \left\{ y \in \mathcal{H}^{\leq d} : \text{BayesExp explores given history } x_{<t_k} y \right\}.$$

For $t_k < \infty$ we have that

$$\begin{aligned}
\varepsilon &\overset{(a)}{<} V_\mu^*(x_{<t_k}) - V_\mu^\pi(x_{<t_k}) \overset{(b)}{=} V_\mu^*(x_{<t_k}) - V_\mu^{\pi_\xi^*}(x_{<t_k}) + V_\mu^{\pi_\xi^*}(x_{<t_k}) - V_\mu^\pi(x_{<t_k}) \\
&\overset{(c)}{\leq} \varepsilon_3 + V_\mu^{\pi_\xi^*}(x_{<t_k}) - V_\mu^\pi(x_{<t_k}) \\
&\overset{(d)}{\leq} \varepsilon_3 + \varepsilon_2 + \sum_{y \in Y} \mu^\pi(y|x_{<t_k}) \left(V_\mu^{\pi_\xi^*}(x_{<t_k} y) - V_\mu^\pi(x_{<t_k} y) \right) \\
&\overset{(e)}{\leq} \varepsilon_3 + 2\varepsilon_2 + \sum_{y \in Y} \mu^\pi(y|x_{<t_k}) \left(V_\mu^*(x_{<t_k} y; d) - V_\mu^\pi(x_{<t_k} y; d) \right) \\
&\overset{(f)}{\leq} \varepsilon_3 + 2\varepsilon_2 + \sum_{y \in Y} \mu^\pi(y|x_{<t_k}) \left(\frac{\max\left\{ 4\Delta_{x_{<t_k} y}, \mathbb{1}\left\{ \mu \notin \mathcal{M}_{x_{<t_k} y} \right\} \right\}}{1-\gamma} \right)
\end{aligned} \tag{4}$$

where (a) follows from the definition of t_k as a time-step when π is ε-suboptimal. (b) by splitting the difference sum. (c) by the fact that π_ξ^* is at worst ε_3-suboptimal when BayesExp is exploiting and $\mu \in \mathcal{M}_t$ (Step 3). Note that $\mu \in \mathcal{M}_{t_k}$ is assumed in the definition of a failure phase. (d) by Lemma 5.

(e) by the fact that $V_\mu^* \geq V_\mu^\pi$ for all π and by Lemma 12. (f) by Lemma 7. Define random variable X_k by

$$X_k := \sum_{t=t_k}^{t_k+d} \mathbb{1}\{t \in E\} \left(\max\left\{4\Delta_t, \mathbb{1}\{\mu \notin \mathcal{M}_t\}\right\}\right) \in [0,4].$$

By the definition of X_k and (4), if $t_k < \infty$, then

$$\begin{aligned}\mathbb{E}_\mu^\pi[X_k|x_{<t_k}] &= \sum_{y\in Y} \mu^\pi(y|x_{<t_k}) \left(4\Delta_{x_{<t_k}y} + \mathbb{1}\{\mu \notin \mathcal{M}_t\}\right) \\ &\geq (\varepsilon - \varepsilon_3 - 2\varepsilon_2)(1-\gamma) =: \varepsilon_4 \equiv \varepsilon_1. \end{aligned} \tag{5}$$

Using the bounds on the number of exploration phases given in Step 2 we have

$$\begin{aligned}\sum_{k=1}^{\infty} X_k &\overset{(a)}{\leq} 1 + \sum_{t\in E} 4\Delta_t \overset{(b)}{\leq} 1 + 4\sqrt{|E| \sum_{t\in E} \Delta_t^2} \\ &\overset{(c)}{\leq} 1 + 4\sqrt{\frac{K}{\varepsilon_1^2} \log\frac{K}{\delta_1^2} \cdot K \log\frac{K}{\delta_1^2}} \leq \frac{5K}{\varepsilon_1} \log\frac{K}{\delta_1^2} \end{aligned} \tag{6}$$

where (a) follows from the definition of X_k and the fact that $\mu \in \mathcal{M}_{t_k}$ for all $t_k < \infty$. (b) by Jensen's inequality. (c) by Equations (2) and (3). Finally we can apply concentration inequalities by noting that $\sum_{k=1}^n \mathbb{E}_\mu^\pi[X_k|x_{<t_k}] - X_k$ is a martingale with zero expectation and differences bounded by 4. Let $F_{\max} \in \mathbb{N}$ be a constant to be defined shortly and let A_2 be the event that:

$$\sum_{k=1}^{F_{\max}} \mathbb{E}_\mu^\pi[X_k|x_{<t_k}] \leq \sum_{k=1}^{F_{\max}} X_k + \sqrt{2\cdot 4^2 \cdot F_{\max} \log\frac{1}{\delta_1}}.$$

By Azuma's inequality $\mu^\pi(A_2) \geq 1 - \delta_1$. If A_2 occurs, then

$$\frac{1}{F_{\max}} \sum_{k=1}^{F_{\max}} \mathbb{E}[X_k|X_{k-1}] \overset{(a)}{\leq} \frac{5K}{\varepsilon_1 F_{\max}} \log\frac{K}{\delta_1^2} + \sqrt{\frac{2\cdot 4^2}{F_{\max}} \log\frac{1}{\delta_1}} \overset{(b)}{<} \varepsilon_4 \tag{7}$$

where (a) follows by substituting (6) and (b) by choosing

$$F_{\max} := \frac{25K}{\varepsilon_1\varepsilon_4} \log\frac{K}{\delta_1^2} \equiv \frac{400K}{\varepsilon^2(1-\gamma)^2} \log\frac{4K}{\delta^2}.$$

But (7) implies that there exists a $k < F_{\max}$ such that $\mathbb{E}[X_k|x_{<t_k}] \leq \varepsilon_4$, which by (5) implies that $t_k = \infty$ and so $|F| \leq F_{\max}$ and $|F_d| \leq F_{\max} d$.

Step 5: Finishing Up. Assuming events A_1 and A_2 both occur, then it holds that both $|E_d| \leq dE_{\max}$ and $|F_d| \leq dF_{\max}$. Since $V_\mu^*(x_{<t}) - V_\mu^\pi(x_{<t}) > \varepsilon$ implies that $t \in F_d \cup E_d$ or $\mu \notin \mathcal{M}_t$ it follows that

$$\begin{aligned}&\mu^\pi\left(\sum_{t=1}^{\infty} \mathbb{1}\left\{V_\mu^*(x_{<t}) - V_\mu^\pi(x_{<t}) > \varepsilon\right\} \leq d(E_{\max} + F_{\max})\right) \\ &\geq \mu^\pi(A_1 \cap A_2) \geq 1 - 2\delta_1 = 1 - \delta. \end{aligned}$$

Substituting $E_{\max} = \frac{16K}{\varepsilon^2(1-\gamma)^2}\log\frac{4K}{\delta^2}$ and $F_{\max} = \frac{400K}{\varepsilon^2(1-\gamma)^2}\log\frac{4K}{\delta^2}$ completes the proof that

$$\mu^\pi\left(\sum_{t=1}^{\infty}\mathbb{1}\{V_\mu^*(x_{<t}) - V_\mu^\pi(x_{<t}) > \varepsilon\} > \frac{416Kd}{\varepsilon^2(1-\gamma)^2}\log\frac{4K}{\delta^2}\right) \leq \delta$$

as required. □

Remark 8. The constant can be reduced to ~ 200 by making ε_2 significantly smaller (and paying only a log cost) and increasing $\varepsilon_1 = \varepsilon_4 \approx \varepsilon(1-\gamma)/3$.

5 Lower Bound on Sample-Complexity

In the last section we showed for any finite environment class $\mathcal{M}$ of size K that the algorithm BayesExp is ε-optimal except for at most

$$O\left(\frac{K}{\varepsilon^2(1-\gamma)^3}\left(\log\frac{1}{\varepsilon(1-\gamma)}\right)\left(\log\frac{K}{\delta}\right)\right) \qquad (\star)$$

time-steps with probability at least $1-\delta$. We now describe the counter-example leading to a nearly-matching lower-bound in the sense that there exist environment classes where no algorithm has sample-complexity much better than $(\star)$. We do not claim that BayesExp achieves the optimal sample-complexity bound in all classes (it does not), only that there exists a class where it (very nearly) does. The gap between the lower and upper bounds is only a $\log\frac{1}{\varepsilon(1-\gamma)}$ factor. The most natural approach to proving a lower bound on the sample-complexity would be to use the famous result by Mannor and Tsitsiklis (2004) on the sample-complexity of exploration for multi-armed bandits. But environment classes based on stationary bandit-like environments lead only to an $\Omega(K)$ bounds on the sample-complexity rather than the desired $\Omega(K\log K)$. The reason is that for such environments the median elimination algorithm for minimising bandit sample-complexity can be used, which achieves the $O(K)$ bound (Even-Dar et al., 2002). This highlights a distinction between the two settings. Even if $\gamma = 0$ (1 step lookahead), the non-stationary version of the problem considered here is harder than the (stationary) bandit case.

Theorem 9. *For each $K > 1$ and $\gamma > 0$ there exists an environment class $\mathcal{M}$ such that for all policies π there exists a $\mu \in \mathcal{M}$ where*

$$\mu^\pi\left(\sum_{t=1}^{\infty}\mathbb{1}\{V_\mu^*(x_{<t}) - V_\mu^\pi(x_{<t}) > \varepsilon\} > c\cdot\left(\frac{\log\frac{1}{2}}{\log\gamma}\right)\frac{K}{\varepsilon^2(1-\gamma)^2}\log\frac{K}{\delta}\right) > \delta.$$

for some $c > 0$ independent of K, π and γ.

The complete proof is left for the technical report, but we describe the counter-example and justify the bound.

Counter-Example. Let $\mathcal{A} = \{\rightarrow, \rightsquigarrow\}$ consist of two actions, $\mathcal{O}$ be a singleton and $\mathcal{R} = \{0, \frac{1}{2}, 1\}$. Let $K \geq 2$ and $\varepsilon, \delta > 0$ be sufficiently small. Define environment class $\mathcal{M} = \{\mu_1, \mu_2, \cdots, \mu_K\}$ as in Figure 1. The parameter $\varepsilon_{t,k}$ determines the optimal action at each time-step. Let L be some large constant, then define $\varepsilon_{t,k}$ in environment μ_k by $\varepsilon_{t,k} = \frac{\varepsilon}{2} \operatorname{sign}\{kL - t\}$. So if the learner chooses action $\rightsquigarrow$, then with probability $\frac{1}{2} + \varepsilon_{t,k}$ it receives reward 1 and otherwise no reward. For $\rightarrow$ it deterministically receives reward $\frac{1}{2}$ regardless of the time-step or environment.

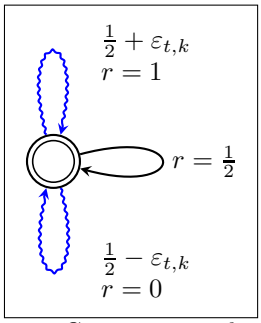

Fig. 1. Counter-example for lower bound. Environment μ_k

Explanation of the Bound. The optimal action in environment μ_k is to take action $\rightarrow$ until time-step kL and there-after take action $\rightsquigarrow$. We call each period of L time-steps a phase and consider the number of ε-errors made in the first $K-1$ phases. The difficulty arises because at the start of the ℓth phase an agent cannot distinguish between environment μ_ℓ and $\mu_{\ell+1}$. But in environment μ_ℓ the agent should take action $\rightsquigarrow$ while in environment $\mu_{\ell+1}$ the agent should take action $\rightarrow$. Now $\rightarrow$ is uninformative, so the only question is how many times a policy must sample action $\rightsquigarrow$ before switching to action $\rightarrow$. In order to guarantee that it is correct in phase ℓ with probability δ/K it should take action $\rightarrow$ approximately $\frac{1}{\varepsilon^2} \log \frac{K}{\delta}$ times. Since it must be correct in all phases, which are essentially independent, the number of times $\rightsquigarrow$ must be taken in environment μ_K in phases $\ell < K$ is $O(\frac{K}{\varepsilon^2} \log \frac{K}{\delta})$. In order to add the dependence on γ we must make two modifications.

1. Add a near-absorbing state corresponding to the times when the agent receives rewards 1, 0 and 1/2 respectively. If the agent stays in these states for $O(\frac{1}{1-\gamma})$, then the cost of a mistake becomes $\varepsilon/(1-\gamma)$ and the mistake bound will depend on $\varepsilon^{-2}(1-\gamma)^{-2}$.
2. To obtain an additional factor of the horizon we proceed in the same fashion as the lower bound given by Lattimore and Hutter (2012). Adapt the environment again so that the agent stays in the decision node for exactly $O(\frac{1}{1-\gamma})$ time-steps, regardless of its action. Only the action at the end of this period decides whether or not the agent gets reward 1/2 or 0 or 1. But if the agent is following a policy that makes an error, then this is counted for $O(\frac{1}{1-\gamma})$ time-steps before the error actually occurs, which multiplies the total number of errors by this quantity.

6 Adaptivity of BayesExp

We now show that the algorithm may learn faster when environments are easy to distinguish. Assume $\gamma = 0$, which implies that the effective horizon $d = 1$. A K-armed Bernoulli bandit is characterised by a vector $p \in [0,1]^K$. At each time-step the learner chooses arm $I_t \in \{1, \cdots, K\}$ and receives reward 1 with

probability p_{I_t} and reward 0 otherwise. The value p_k is called the bias of the kth arm. There is now a huge literature on bandits, which we will not discuss, but see Bubeck and Cesa-Bianchi (2012) and references there-in for a good introduction. Choose $\mathcal{M} = \{\nu_1, \cdots, \nu_K\}$ to be the set of K-armed Bernoulli bandits where in environment ν_k the bias of the k arm is $\frac{1}{2}$ while for all other arms it is equal to $\frac{1}{2} - \Delta_k$ where $\Delta_k \geq \varepsilon$. Thus the optimal action in environment ν_k is to always choose arm k. Note that in this setting there are no observations ($\mathcal{O} \equiv$ singleton) and $\mathcal{R} = \{0, 1\}$. We show that the performance of BayesExp is substantially improved for large Δ_k where the environments are more easily distinguished.

Theorem 10. *If BayesExp is run on the environment class described above, then*

$$\mu^\pi \left(\sum_{t=1}^{\infty} \mathbb{1}\left\{ V_\mu^*(x_{<t}) - V_\mu^\pi(x_{<t}) > \varepsilon \right\} > \sum_{k:\nu_k \neq \mu} \frac{4}{\Delta_k^2} \log \frac{K}{\delta^2} \right) \leq \delta.$$

The proof may be found in the associated technical report.

7 Conclusion

We adapted the Bayesian optimal agent studied by Hutter (2005) and others by adding an exploration component. The new algorithm achieves minimax finite sample-complexity bounds for finite environment classes. The theoretical results improve substantially on those given for the MERL algorithm by Lattimore et al. (2013a). In that work only two environments are compared in each exploration phase and models were discarded based on rewards alone, with observations completely ignored. Like the k-meteorologist algorithm (Diuk et al., 2009) models were only removed in discrete blocks. In contrast, the approach used here eliminates environments smoothly, which in benign environments may occur significantly faster than the worst-case bounds suggest. An example of this adaptivity is given for bandit environments in Section 6. There is another benefit of BayesExp illustrated by the example in Section 6. While the analysis in the proof of Theorem 4 leads to a largish constant, it is not used by the algorithm, which means that in simple cases the analysis can be improved substantially.

Future work could focus on proving more general problem-dependent bounds on the sample-complexity of algorithms like BayesExp, and characterising the difficulty of reinforcement learning environments and classes. This problem is now reasonably understood for bandit environments, but even for MDPs there is only limited work on problem-dependent bounds, and nothing for general RL as far as we are aware. Larger environment classes are also worth considering, including countable or separable spaces where uniform sample-complexity bounds are not possible, but problem-dependent asymptotic bounds are. We are optimistic that BayesExp can be extended to these cases.

Acknowledgements. This work was supported by the Alberta Innovates Technology Futures and NSERC.

References

Auer, P., Jaksch, T., Ortner, R.: Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research 99, 1532–4435 (2010) ISSN 1532-4435

Azar, M.G., Lazaric, A., Brunskill, E.: Regret bounds for reinforcement learning with policy advice. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013, Part I. LNCS, vol. 8188, pp. 97–112. Springer, Heidelberg (2013)

Bubeck, S., Cesa-Bianchi, N.: Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Foundations and Trends in Machine Learning. Now Publishers Incorporated (2012) ISBN 9781601986269

Chakraborty, D., Stone, P.: Structure learning in ergodic factored mdps without knowledge of the transition function's in-degree. In: Proceedings of the Twenty Eighth International Conference on Machine Learning (2011)

Diuk, C., Li, L., Leffler, B.: The adaptive k-meteorologists problem and its application to structure learning and feature selection in reinforcement learning. In: Danyluk, A.P., Bottou, L., Littman, M.L. (eds.) Proceedings of the 26th Annual International Conference on Machine Learning, pp. 249–256. ACM (2009)

Dyagilev, K., Mannor, S., Shimkin, N.: Efficient reinforcement learning in parameterized Models: Discrete parameter case. In: Girgin, S., Loth, M., Munos, R., Preux, P., Ryabko, D. (eds.) EWRL 2008. LNCS (LNAI), vol. 5323, pp. 41–54. Springer, Heidelberg (2008)

Even-Dar, E., Mannor, S., Mansour, Y.: PAC Bounds for Multi-armed Bandit and Markov Decision Processes. In: Kivinen, J., Sloan, R.H. (eds.) COLT 2002. LNCS (LNAI), vol. 2375, pp. 255–270. Springer, Heidelberg (2002)

Even-Dar, E., Kakade, S., Mansour, Y.: Reinforcement learning in POMDPs without resets. In: International Joint Conference on Artificial Intelligence, pp. 690–695 (2005)

Hutter, M.: Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In: Kivinen, J., Sloan, R.H. (eds.) COLT 2002. LNCS (LNAI), vol. 2375, pp. 364–379. Springer, Heidelberg (2002)

Hutter, M.: Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability. Springer, Berlin (2005)

Hutter, M., Muchnik, A.: On semimeasures predicting Martin-Löf random sequences. Theoretical Computer Science 382(3), 247–261 (2007)

Kearns, M., Singh, S.: Near-optimal reinforcement learning in polynomial time. Machine Learning 49(2-3), 209–232 (2002)

Lattimore, T., Hutter, M.: PAC bounds for discounted MDPs. In: Bshouty, N.H., Stoltz, G., Vayatis, N., Zeugmann, T. (eds.) ALT 2012. LNCS, vol. 7568, pp. 320–334. Springer, Heidelberg (2012)

Lattimore, T., Hutter, M.: Bayesian reinforcement learning with exploration. arxiv (2014)

Lattimore, T., Hutter, M., Sunehag, P.: The sample-complexity of general reinforcement learning. In: Proceedings of the 30th International Conference on Machine Learning (2013a)

Lattimore, T., Hutter, M., Sunehag, P.: Concentration and confidence for discrete bayesian sequence predictors. In: Jain, S., Munos, R., Stephan, F., Zeugmann, T. (eds.) ALT 2013. LNCS, vol. 8139, pp. 324–338. Springer, Heidelberg (2013)

Mannor, S., Tsitsiklis, J.: The sample complexity of exploration in the multi-armed bandit problem. Journal of Machine Learning Research 5, 623–648 (2004)

Odalric-Ambrym, M., Nguyen, P., Ortner, R., Ryabko, D.: Optimal regret bounds for selecting the state representation in reinforcement learning. In: Proceedings of the Thirtieth International Conference on Machine Learning (2013)

Orseau, L.: Optimality issues of universal greedy agents with static priors. In: Hutter, M., Stephan, F., Vovk, V., Zeugmann, T. (eds.) ALT 2010. LNCS (LNAI), vol. 6331, pp. 345–359. Springer, Heidelberg (2010)

Osband, I., Russo, D., Van Roy, B.: (More) efficient reinforcement learning via posterior sampling. In: Advances in Neural Information Processing Systems, pp. 3003–3011 (2013)

Sunehag, P., Hutter, M.: Optimistic agents are asymptotically optimal. In: Thielscher, M., Zhang, D. (eds.) AI 2012. LNCS, vol. 7691, pp. 15–26. Springer, Heidelberg (2012)

Szita, I., Szepesvári, C.: Model-based reinforcement learning with nearly tight exploration complexity bounds. In: Proceedings of the 27th International Conference on Machine Learning, pp. 1031–1038. ACM, New York (2010)

A Properties of Value Functions

Lemma 11. *Let $\pi \in \Pi$ and μ and ν be two environments. Then*

$$V_\mu^\pi(x;d) - V_\nu^\pi(x;d) \leq \frac{\delta_x^d(\mu^\pi, \nu^\pi)}{1-\gamma}.$$

Proof. The difference in value functions is a difference in expected returns with respect to μ^π and ν^π. This is bounded by the total variation distance multiplied by the maximum return, which is $1/(1-\gamma)$. □

Lemma 12. *If x is a history at time-step t and $\varepsilon > 0$ and $d \geq \left\lceil \frac{\log \varepsilon(1-\gamma)}{\log \gamma} \right\rceil$, then $V_\mu^\pi(x) \geq V_\mu^\pi(x;d)$ and $V_\mu^\pi(x) - V_\mu^\pi(x;d) \leq \varepsilon$.*

Proof. That $V_\mu^\pi(x) \geq V_\mu^\pi(x;d)$ is trivial. For the second claim:

$$V_\mu^\pi(x) - V_\mu^\pi(x;d) \overset{(a)}{=} \mathbb{E}_\mu^\pi \left[\sum_{k=t+d}^{\infty} \gamma^{k-t} r_k \middle| x \right] \overset{(b)}{\leq} \sum_{k=t+d}^{\infty} \gamma^{k-t} \overset{(c)}{=} \frac{\gamma^d}{1-\gamma} \overset{(d)}{\leq} \varepsilon$$

where (a) follows by adding and subtracting the tail sum. (b) because $r_k \in [0,1]$. (c) is trivial while (d) follows from the definition of d. □

Lemma 13. *Let μ be an environment, $x \in \mathcal{H}^*$ a history and $Y \subset \mathcal{H}^*$ be complete and prefix free. If π_1 and π_2 are policies such that $\pi_1(xz) = \pi_2(xz)$ for all y, z for which $z \sqsubset y$. Then*

$$V_\mu^{\pi_1}(x) - V_\mu^{\pi_2}(x) = \sum_{y \in Y} \mu^{\pi_1}(y|x) \gamma^{\ell(y)} \left(V_\mu^{\pi_1}(xy) - V_\mu^{\pi_2}(xy) \right).$$