Thompson Sampling is Asymptotically Optimal in General Environments

Jan Leike
Australian National University
jan.leike@anu.edu.au

Tor Lattimore
University of Alberta
tor.lattimore@gmail.com

Laurent Orseau
Google DeepMind
lorseau@google.com

Marcus Hutter
Australian National University
marcus.hutter@anu.edu.au

Abstract

We discuss a variant of Thompson sampling for nonparametric reinforcement learning in a countable classes of general stochastic environments. These environments can be non-Markov, nonergodic, and partially observable. We show that Thompson sampling learns the environment class in the sense that (1) asymptotically its value converges to the optimal value in mean and (2) given a recoverability assumption regret is sublinear.

Keywords. General reinforcement learning, Thompson sampling, asymptotic optimality, regret, discounting, recoverability, AIXI.

1 INTRODUCTION

In reinforcement learning (RL) an agent interacts with an unknown environment with the goal of maximizing rewards. Recently reinforcement learning has received a surge of interest, triggered by its success in applications such as simple video games [MKS+15]. However, theory is lagging behind application and most theoretical analyses has been done in the bandit framework and for Markov decision processes (MDPs). These restricted environment classes fall short of the full reinforcement learning problem and theoretical results usually assume ergocity and visiting every state infinitely often. Needless to say, these assumptions are not satisfied for any but the simplest applications.

Our goal is to lift these restrictions; we consider *general reinforcement learning*, a top-down approach to RL with the aim to understand the fundamental underlying problems in their generality. Our approach to general RL is *nonparametric*: we only assume that the true environment belongs to a given countable environment class.

We are interested in agents that maximize rewards *optimally*. Since the agent does not know the true environment in advance, it is not obvious what optimality should mean. We discuss two different notions of optimality: *asymptotic optimality* and *worst-case regret*.

Asymptotic optimality requires that asymptotically the agent learns to act optimally, i.e., that the discounted value of the agent's policy π converges to the optimal discounted value, $V_\mu^* - V_\mu^\pi \to 0$ for all environments μ from the environment class. This convergence is impossible for deterministic policies since the agent has to explore infinitely often and for long stretches of time, but there are policies that converge almost surely in Cesàro average [LH11]. Bayes-optimal agents are generally not asymptotically optimal [Ors13]. However, asymptotic optimality can be achieved through an exploration component on top of a Bayes-optimal agent [Lat13, Ch. 5] or through optimism [SH15].

Asymptotic optimality in mean is essentially a weaker variant of *probably approximately correct* (PAC) that comes without a concrete convergence rate: for all $\varepsilon > 0$ and $\delta > 0$ the probability that our policy is ε-suboptimal converges to zero (at an unknown rate). Eventually this probability will be less than δ. Since our environment class can be very large and non-compact, concrete PAC/convergence rates are likely impossible.

Regret is how many expected rewards the agent forfeits by not following the best informed policy. Different problem classes have different regret rates, depending on the structure and the difficulty of the problem class. Multi-armed bandits provide a (problem-independent) worst-case regret bound of $\Omega(\sqrt{KT})$ where K is the number of arms [BB12]. In Markov decision processes (MDPs) the lower bound is $\Omega(\sqrt{DSAT})$ where S is the number of states, A the number of actions, and D the diameter of the MDP [AJO10]. For a countable class of environments given by state representation functions that map histories to MDP states, a regret of $\tilde{O}(T^{2/3})$ is achievable assuming the resulting MDP is weakly communicating [NMRO13]. A problem class is considered *learnable* if there is an algorithm that has a sublinear regret guarantee.

This paper continues a narrative that started with definition

of the Bayesian agent AIXI [Hut00] and the proof that it satisfies various optimality guarantees [Hut02]. Recently it was revealed that these optimality notions are trivial or subjective [LH15]: a Bayesian agent does not explore enough to lose the prior's bias, and a particularly bad prior can make the agent conform to any arbitrarily bad policy as long as this policy yields some rewards. These negative results put the Bayesian approach to (general) RL into question. In this paper we remedy the situation by showing that using Bayesian techniques an agent can indeed be optimal in an objective sense.

The agent we consider is known as *Thompson sampling*, *posterior sampling*, or the *Bayesian control rule* [Tho33]. It samples an environment ρ from the posterior, follows the ρ-optimal policy for one effective horizon (a lookahead long enough to encompass most of the discount function's mass), and then repeats. We show that this agent's policy is asymptotically optimal in mean (and, equivalently, in probability). Furthermore, using a recoverability assumption on the environment, and some (minor) assumptions on the discount function, we prove that the worst-case regret is sublinear. This is the first time convergence and regret bounds of Thompson sampling have been shown under such general conditions.

Thompson sampling was originally proposed by Thompson as a bandit algorithm [Tho33]. It is easy to implement and often achieves quite good results [CL11]. In multi-armed bandits it attains optimal regret [AG11, KKM12]. Thompson sampling has also been considered for MDPs: as model-free method relying on distributions over Q-functions with convergence guarantee [DFR98], and as a model-based algorithm without theoretical analysis [Str00]. Bayesian and frequentist regret bounds have also been established [ORvR13, OR14, GM15]. PAC guarantees have been established for an optimistic variant of Thompson sampling for MDPs [ALL$^+$09].

For general RL Thompson sampling was first suggested in [OB10] with resampling at every time step. The authors prove that the action probabilities of Thompson sampling converge to the action probability of the optimal policy almost surely, but require a finite environment class and two (arguably quite strong) technical assumptions on the behavior of the posterior distribution (akin to ergodicity) and the similarity of environments in the class. Our convergence results do not require these assumptions, but we rely on an (unavoidable) recoverability assumption for our regret bound.

Appendix A contains a list of notation and Appendix B contains omitted proofs.

2 PRELIMINARIES

The set $\mathcal{X}^* := \bigcup_{n=0}^{\infty} \mathcal{X}^n$ is the set of all finite strings over the alphabet $\mathcal{X}$ and the set $\mathcal{X}^\infty$ is the set of all infinite strings over the alphabet $\mathcal{X}$. The empty string is denoted by ϵ, not to be confused with the small positive real number ε. Given a string $x \in \mathcal{X}^*$, we denote its length by $|x|$. For a (finite or infinite) string x of length $\geq k$, we denote with $x_{1:k}$ the first k characters of x, and with $x_{<k}$ the first $k-1$ characters of x.

The notation $\Delta\mathcal{Y}$ denotes the set of probability distributions over $\mathcal{Y}$.

In reinforcement learning, an agent interacts with an environment in cycles: at time step t the agent chooses an *action* $a_t \in \mathcal{A}$ and receives a *percept* $e_t = (o_t, r_t) \in \mathcal{E}$ consisting of an *observation* $o_t \in \mathcal{O}$ and a real-valued *reward* r_t; the cycle then repeats for $t+1$. We assume that rewards are bounded between 0 and 1 and that the set of actions $\mathcal{A}$ and the set of percepts $\mathcal{E}$ are finite.

We fix a *discount function* $\gamma : \mathbb{N} \to \mathbb{R}$ with $\gamma_t \geq 0$ and $\sum_{t=1}^{\infty} \gamma_t < \infty$. Our goal is to maximize discounted rewards $\sum_{t=1}^{\infty} \gamma_t r_t$. The *discount normalization factor* is defined as $\Gamma_t := \sum_{k=t}^{\infty} \gamma_k$. The *effective horizon* $H_t(\varepsilon)$ is a horizon that is long enough to encompass all but an ε of the discount function's mass:

$$H_t(\varepsilon) := \min\{k \mid \Gamma_{t+k}/\Gamma_t \leq \varepsilon\} \tag{1}$$

A *history* is an element of $(\mathcal{A} \times \mathcal{E})^*$. We use $æ \in \mathcal{A} \times \mathcal{E}$ to denote one interaction cycle, and $æ_{<t}$ to denote a history of length $t-1$. We treat action, percepts, and histories both as outcomes and as random variables. A *policy* is a function $\pi : (\mathcal{A} \times \mathcal{E})^* \to \Delta\mathcal{A}$ mapping a history $æ_{<t}$ to a distribution over the actions taken after seeing this history; the probability of action a is denoted $\pi(a \mid æ_{<t})$. An *environment* is a function $\nu : (\mathcal{A} \times \mathcal{E})^* \times \mathcal{A} \to \Delta\mathcal{E}$ mapping a history $æ_{<t}$ and an action a_t to a distribution over the percepts generated after this history; the probability of percept e is denoted $\nu(e \mid æ_{<t}a_t)$.

A policy π and an environment ν generate a probability measure ν^π over infinite histories $(\mathcal{A} \times \mathcal{E})^\infty$, defined by its values on the cylinder sets $\{h \in (\mathcal{A} \times \mathcal{E})^\infty \mid h_{<t} = æ_{<t}\}$:

$$\nu^\pi(æ_{<t}) := \prod_{k=1}^{t-1} \pi(a_k \mid æ_{<k}) \nu(e_k \mid æ_{<k}a_k)$$

When we take an expectation $\mathbb{E}_\nu^\pi$ of a random variable $X_t(æ_{<t})$ this is to be understood as the expectation of the history $æ_{<t}$ for a fixed time step t drawn from ν^π, i.e.,

$$\mathbb{E}_\nu^\pi[X_t(æ_{<t})] := \sum_{æ_{<t}} \nu^\pi(æ_{<t}) X_t(æ_{<t}).$$

We often do not explicitly add the subscript t to time-dependent random variables.

Definition 1 (Value Function)**.** The *value* of a policy π in an environment ν given history $æ_{<t}$ is defined as

$$V_\nu^\pi(æ_{<t}) := \frac{1}{\Gamma_t}\mathbb{E}_\nu^\pi\left[\sum_{k=t}^{\infty}\gamma_k r_k \,\middle|\, æ_{<t}\right],$$

if $\Gamma_t > 0$ and $V_\nu^\pi(æ_{<t}) := 0$ if $\Gamma_t = 0$. The *optimal value* is defined as $V_\nu^*(h) := \sup_\pi V_\nu^\pi(h)$.

The normalization constant $1/\Gamma_t$ ensures that values are bounded between 0 and 1. We also use the *truncated value function*

$$V_\nu^{\pi,m}(æ_{<t}) := \frac{1}{\Gamma_t}\mathbb{E}_\nu^\pi\left[\sum_{k=t}^{m}\gamma_k r_k \,\middle|\, æ_{<t}\right].$$

For each environment μ there is an *optimal policy* π_μ^* that takes an *optimal action* for each history [LH14, Thm. 10]:

$$\pi_\mu^*(a_t \mid æ_{<t}) > 0 \implies a_t \in \arg\max_a V_\mu^*(æ_{<t}a)$$

Let $\mathcal{M}$ denote a countable class of environments. We assume that $\mathcal{M}$ is large enough to contain the true environment (e.g. the class of all computable environments). Let $w \in \Delta\mathcal{M}$ be a prior probability distribution on $\mathcal{M}$ and let

$$\xi := \sum_{\nu\in\mathcal{M}} w(\nu)\nu$$

denote the corresponding Bayesian mixture over the class $\mathcal{M}$. After observing the history $æ_{<t}$ the prior w is updated to the posterior

$$w(\nu \mid æ_{<t}) := w(\nu)\frac{\nu(æ_{<t})}{\xi(æ_{<t})}.$$

We also use the notation $w(\mathcal{M}' \mid æ_{<t}) := \sum_{\nu\in\mathcal{M}'} w(\nu \mid æ_{<t})$ for a set of environments $\mathcal{M}' \subseteq \mathcal{M}$. Likewise we define $\nu(A \mid æ_{<t}) := \sum_{h\in A}\nu(h \mid æ_{<t})$ for a prefix-free set of histories $A \subseteq (\mathcal{A}\times\mathcal{E})^*$.

Let $\nu, \rho \in \mathcal{M}$ be two environments, let π_1, π_2 be two policies, and let $m \in \mathbb{N}$ be a lookahead time step. The *total variation distance* is defined as

$$D_m(\nu^{\pi_1}, \rho^{\pi_2} \mid æ_{<t}) := \sup_{A\subseteq(\mathcal{A}\times\mathcal{E})^m}\left|\nu^{\pi_1}(A \mid æ_{<t}) - \rho^{\pi_2}(A \mid æ_{<t})\right|.$$

with $D_\infty(\nu^{\pi_1}, \rho^{\pi_2} \mid æ_{<t}) := \lim_{m\to\infty} D_m(\nu^{\pi_1}, \rho^{\pi_2} \mid æ_{<t})$.

Lemma 2 (Bounds on Value Difference)**.** *For any policies π_1, π_2, any environments ρ and ν, and any horizon $t \le m \le \infty$,*

$$|V_\nu^{\pi_1,m}(æ_{<t}) - V_\rho^{\pi_2,m}(æ_{<t})| \le D_m(\nu^{\pi_1}, \rho^{\pi_2} \mid æ_{<t})$$

Proof. See Appendix B. □

3 THOMPSON SAMPLING IS ASYMPTOTICALLY OPTIMAL

Strens proposes following the optimal policy for one episode or "related to the number of state transitions the agent is likely to need to plan ahead" [Str00]. We follow Strens' suggestion and resample at the effective horizon.

Let ε_t be a monotone decreasing sequence of positive reals such that $\varepsilon_t \to 0$ as $t \to \infty$. We define our Thompson-sampling policy π_T in Algorithm 1.

Algorithm 1 Thompson sampling policy π_T

1: **while** true **do**
2: sample $\rho \sim w(\cdot \mid æ_{<t})$
3: follow π_ρ^* for $H_t(\varepsilon_t)$ steps

Note that π_T is a stochastic policy since we occasionally sample from a distribution. We assume that this sampling is independent of everything else.

Definition 3 (Asymptotic Optimality)**.** A policy π is *asymptotically optimal in an environment class $\mathcal{M}$* iff for all $\mu \in \mathcal{M}$

$$V_\mu^*(æ_{<t}) - V_\mu^\pi(æ_{<t}) \to 0 \text{ as } t \to \infty \tag{2}$$

on histories drawn from μ^π.

There are different types of asymptotic optimalities based on the type of stochastic convergence in (2). If this convergence occurs almost surely, it is called *strong asymptotic optimality* [LH11, Def. 7]; if this convergence occurs in mean, it is called *asymptotic optimality in mean*; if this convergence occurs in probability, it is called *asymptotic optimality in probability*; and if the Cesàro averages converge almost surely, it is called *weak asymptotic optimality* [LH11, Def. 7].

3.1 ASYMPTOTIC OPTIMALITY IN MEAN

This subsection is dedicated to proving the following theorem.

Theorem 4 (Thompson Sampling is Asymptotically Optimal in Mean)**.** *For all environments $\mu \in \mathcal{M}$,*

$$\mathbb{E}_\mu^{\pi_T}\left[V_\mu^*(æ_{<t}) - V_\mu^{\pi_T}(æ_{<t})\right] \to 0 \text{ as } t \to \infty.$$

This theorem immediately implies that Thompson sampling is also asymptotically optimal in probability: The convergence in mean of the random variables $X_t := V_\mu^*(æ_{<t}) - V_\mu^{\pi_T}(æ_{<t})$ stated in Theorem 4 is equivalent to convergence in probability in the sense that $\mu^{\pi_T}[X_t > \varepsilon] \to 0$ as $t \to \infty$ for all $\varepsilon > 0$ because the random variables X_t are nonnegative and bounded. However, this does not imply almost sure convergence (see Section 3.3).

Define the *Bayes-expected total variation distance*

$$F_m^\pi(\text{æ}_{<t}) := \sum_{\rho\in\mathcal{M}} w(\rho \mid \text{æ}_{<t}) D_m(\rho^\pi, \xi^\pi \mid \text{æ}_{<t})$$

for $m \leq \infty$.

If we replace the distance measure D_m by cross-entropy, then the quantity $F_m^\pi(\text{æ}_{<t})$ becomes the Bayes-expected *information gain* [Lat13, Eq. 3.5].

For the proof of Theorem 4 we need the following lemma.

Lemma 5 (F Vanishes On-Policy)**.** *For any policy π and any environment μ,*

$$\mathbb{E}_\mu^\pi[F_\infty^\pi(\text{æ}_{<t})] \to 0 \text{ as } t \to \infty.$$

Proof. See Appendix B. □

Proof of Theorem 4. Let $\beta, \delta > 0$ and let $\varepsilon_t > 0$ denote the sequence used to define π_T in Algorithm 1. We assume that t is large enough such that $\varepsilon_k \leq \beta$ for all $k \geq t$ and that δ is small enough such that $w(\mu \mid \text{æ}_{<t}) > 4\delta$ for all t, which holds since $w(\mu \mid \text{æ}_{<t}) \not\to 0$ μ^π-almost surely for any policy π [Hut09, Lem. 3i].

The stochastic process $w(\nu \mid \text{æ}_{<t})$ is a ξ^{π_T}-martingale since

$$\begin{aligned}
&\mathbb{E}_\xi^{\pi_T}[w(\nu \mid \text{æ}_{1:t}) \mid \text{æ}_{<t}] \\
&= \sum_{a_t e_t} \xi^{\pi_T}(\text{æ}_t \mid \text{æ}_{<t}) w(\nu) \frac{\nu^{\pi_T}(\text{æ}_{1:t})}{\xi^{\pi_T}(\text{æ}_{1:t})} \\
&= \sum_{a_t e_t} \xi^{\pi_T}(\text{æ}_t \mid \text{æ}_{<t}) w(\nu \mid \text{æ}_{<t}) \frac{\nu^{\pi_T}(\text{æ}_t \mid \text{æ}_{<t})}{\xi^{\pi_T}(\text{æ}_t \mid \text{æ}_{<t})} \\
&= w(\nu \mid \text{æ}_{<t}) \sum_{a_t e_t} \nu^{\pi_T}(\text{æ}_t \mid \text{æ}_{<t}) \\
&= w(\nu \mid \text{æ}_{<t}).
\end{aligned}$$

By the martingale convergence theorem [Dur10, Thm. 5.2.8] $w(\nu \mid \text{æ}_{<t})$ converges ξ^{π_T}-almost surely and because $\xi^{\pi_T} \geq w(\mu)\mu^{\pi_T}$ it also converges μ^{π_T}-almost surely.

We argue that we can choose t_0 to be one of π_T's resampling time steps large enough such that for all $t \geq t_0$ the following three events hold simultaneously with μ^{π_T}-probability at least $1-\delta$.

(i) There is a finite set $\mathcal{M}' \subset \mathcal{M}$ with $w(\mathcal{M}' \mid \text{æ}_{<t}) > 1-\delta$ and $w(\nu \mid \text{æ}_{<k}) \not\to 0$ as $k \to \infty$ for all $\nu \in \mathcal{M}'$.

(ii) $|w(\mathcal{M}'' \mid \text{æ}_{<t}) - w(\mathcal{M}'' \mid \text{æ}_{<t_0})| \leq \delta$ for all $\mathcal{M}'' \subseteq \mathcal{M}'$.

(iii) $F_\infty^{\pi_T}(\text{æ}_{<t}) < \delta\beta w_{\min}^2$.

where $w_{\min} := \inf\{w(\nu \mid \text{æ}_{<k}) \mid k \in \mathbb{N}, \nu \in \mathcal{M}'\}$, which is positive by (i).

(i) and (ii) are satisfied eventually because the posterior $w(\cdot \mid \text{æ}_{<t})$ converges μ^{π_T}-almost surely. Note that the set $\mathcal{M}'$ is random: the limit of $w(\nu \mid \text{æ}_{<t})$ as $t \to \infty$ depends on the history $\text{æ}_{1:\infty}$. Without loss of generality, we assume the true environment μ is contained in $\mathcal{M}'$ since $w(\mu \mid \text{æ}_{<t}) \not\to 0$ μ^{π_T}-almost surely. (iii) follows from Lemma 5 since convergence in mean implies convergence in probability.

Moreover, we define the horizon $m := t + H_t(\varepsilon_t)$ as the time step of the effective horizon at time step t. Let $\text{æ}_{<t}$ be a fixed history for which (i-iii) is satisfied. Then we have

$$\begin{aligned}
\delta\beta w_{\min}^2 &> F_\infty^{\pi_T}(\text{æ}_{<t}) \\
&= \sum_{\nu\in\mathcal{M}} w(\nu \mid \text{æ}_{<t}) D_\infty(\nu^{\pi_T}, \xi^{\pi_T} \mid \text{æ}_{<t}) \\
&= \mathbb{E}_{\nu\sim w(\cdot\mid\text{æ}_{<t})}\left[D_\infty(\nu^{\pi_T}, \xi^{\pi_T} \mid \text{æ}_{<t})\right] \\
&\geq \mathbb{E}_{\nu\sim w(\cdot\mid\text{æ}_{<t})}\left[D_m(\nu^{\pi_T}, \xi^{\pi_T} \mid \text{æ}_{<t})\right] \\
&\geq \beta w_{\min}^2 w(\mathcal{M} \setminus \mathcal{M}'' \mid \text{æ}_{<t})
\end{aligned}$$

by Markov's inequality where

$$\mathcal{M}'' := \left\{\nu \in \mathcal{M} \;\middle|\; D_m(\nu^{\pi_T}, \xi^{\pi_T} \mid \text{æ}_{<t}) < \beta w_{\min}^2\right\}.$$

For our fixed history $\text{æ}_{<t}$ we have

$$\begin{aligned}
1-\delta &< w(\mathcal{M}'' \mid \text{æ}_{<t}) \\
&\overset{(i)}{\leq} w(\mathcal{M}'' \cap \mathcal{M}' \mid \text{æ}_{<t}) + \delta \\
&\overset{(ii)}{\leq} w(\mathcal{M}'' \cap \mathcal{M}' \mid \text{æ}_{<t_0}) + 2\delta \\
&\overset{(i)}{\leq} w(\mathcal{M}'' \mid \text{æ}_{<t_0}) + 3\delta
\end{aligned}$$

and thus we get

$$1 - 4\delta < w\left[D_m(\nu^{\pi_T}, \xi^{\pi_T} \mid \text{æ}_{<t}) < \beta w_{\min}^2 \;\middle|\; \text{æ}_{<t_0}\right]. \tag{3}$$

In particular, this bound holds for $\nu = \mu$ since $w(\mu \mid \text{æ}_{<t_0}) > 4\delta$ by assumption.

It remains to show that with high probability the value $V_\mu^{\pi_\rho^*}$ of the sample ρ's optimal policy π_ρ^* is sufficiently close to the μ-optimal value V_μ^*. The worst case is that we draw the worst sample from $\mathcal{M}' \cap \mathcal{M}''$ twice in a row. From now on, let ρ denote the sample environment we draw at time step t_0, and let t denote some time step between t_0 and $t_1 := t_0 + H_{t_0}(\varepsilon_{t_0})$ (before the next resampling). With probability $w(\nu' \mid \text{æ}_{<t_0}) w(\nu' \mid \text{æ}_{<t_1})$ we sample ν' both at t_0 and t_1 when following π_T. Therefore we have for all $\text{æ}_{t:m}$ and all $\nu \in \mathcal{M}$

$$\begin{aligned}
&\nu^{\pi_T}(\text{æ}_{1:m} \mid \text{æ}_{<t}) \\
&\geq w(\nu' \mid \text{æ}_{<t_0}) w(\nu' \mid \text{æ}_{<t_1}) \nu^{\pi_{\nu'}^*}(\text{æ}_{1:m} \mid \text{æ}_{<t}).
\end{aligned}$$

Thus we get for all $\nu \in \mathcal{M}'$ (in particular ρ and μ)

$$\begin{aligned}
&D_m(\mu^{\pi_T}, \rho^{\pi_T} \mid æ_{<t}) \\
&\geq \sup_{\nu' \in \mathcal{M}} \sup_{A \subseteq (\mathcal{A}\times\mathcal{E})^m} \Big| w(\nu' \mid æ_{<t_0}) w(\nu' \mid æ_{<t_1}) \\
&\qquad (\mu^{\pi^*_{\nu'}}(A \mid æ_{<t}) - \rho^{\pi^*_{\nu'}}(A \mid æ_{<t})) \Big| \\
&\geq w(\nu \mid æ_{<t_0}) w(\nu \mid æ_{<t_1}) \\
&\qquad \sup_{A \subseteq (\mathcal{A}\times\mathcal{E})^m} \Big| \mu^{\pi^*_\nu}(A \mid æ_{<t}) - \rho^{\pi^*_\nu}(A \mid æ_{<t}) \Big| \\
&\geq w_{\min}^2 D_m(\mu^{\pi^*_\nu}, \rho^{\pi^*_\nu} \mid æ_{<t}).
\end{aligned}$$

For $\rho \in \mathcal{M}''$ we get

$$\begin{aligned}
&D_m(\mu^{\pi_T}, \rho^{\pi_T} \mid æ_{<t}) \\
&\leq D_m(\mu^{\pi_T}, \xi^{\pi_T} \mid æ_{<t}) + D_m(\rho^{\pi_T}, \xi^{\pi_T} \mid æ_{<t}) \\
&\overset{(3)}{<} \beta w_{\min}^2 + \beta w_{\min}^2 = 2\beta w_{\min}^2,
\end{aligned}$$

which implies together with Lemma 2 and the fact that rewards in $[0,1]$

$$\begin{aligned}
&\Big| V_\mu^{\pi^*_\nu}(æ_{<t}) - V_\rho^{\pi^*_\nu}(æ_{<t}) \Big| \\
&\leq \frac{\Gamma_{t+H_t(\varepsilon_t)}}{\Gamma_t} + \Big| V_\mu^{\pi^*_\nu, m}(æ_{<t}) - V_\rho^{\pi^*_\nu, m}(æ_{<t}) \Big| \\
&\leq \varepsilon_t + D_m(\mu^{\pi^*_\nu}, \rho^{\pi^*_\nu} \mid æ_{<t}) \\
&\leq \varepsilon_t + \tfrac{1}{w_{\min}^2} D_m(\mu^{\pi_T}, \rho^{\pi_T} \mid æ_{<t}) \\
&< \beta + 2\beta = 3\beta.
\end{aligned}$$

Hence we get (omitting history arguments $æ_{<t}$ for simplicity)

$$\begin{aligned}
V_\mu^* = V_\mu^{\pi^*_\mu} &< V_\rho^{\pi^*_\mu} + 3\beta \leq V_\rho^* + 3\beta \\
&= V_\rho^{\pi^*_\rho} + 3\beta < V_\mu^{\pi^*_\rho} + 3\beta + 3\beta = V_\mu^{\pi^*_\rho} + 6\beta.
\end{aligned} \tag{4}$$

With μ^{π_T}-probability at least $1-\delta$ (i), (ii), and (iii) are true, with μ^{π_T}-probability at least $1-\delta$ our sample ρ happens to be in $\mathcal{M}'$ by (i), and with $w(\,\cdot \mid æ_{<t_0})$-probability at least $1-4\delta$ the sample is in $\mathcal{M}''$ by (3). All of these events are true simultaneously with probability at least $1-(\delta+\delta+4\delta) = 1-6\delta$. Hence the bound (4) transfers for π_T such that with μ^{π_T}-probability $\geq 1-6\delta$ we have

$$V_\mu^*(æ_{<t}) - V_\mu^{\pi_T}(æ_{<t}) < 6\beta.$$

Therefore $\mu^{\pi_T}[V_\mu^*(æ_{<t}) - V_\mu^{\pi_T}(æ_{<t}) \geq 6\beta] < 6\delta$ and with $\delta \to 0$ we get that $V_\mu^*(æ_{<t}) - V_\mu^{\pi_T}(æ_{<t}) \to 0$ as $t \to \infty$ in probability. The value function is bounded, thus it also converges in mean by the dominated convergence theorem. □

3.2 WEAK ASYMPTOTIC OPTIMALITY

It might appear that convergence in mean is more natural than the convergence of Cesàro averages of weak asymptotic optimality. However, both notions are not so fundamentally different because they both allow an infinite number of bad mistakes (actions that lead to $V_\mu^* - V_\mu^\pi$ being large). Asymptotic optimality in mean allows bad mistakes as long as their probability converges to zero; weak asymptotic optimality allows bad mistakes as long as the total time spent on bad mistakes grows sublinearly.

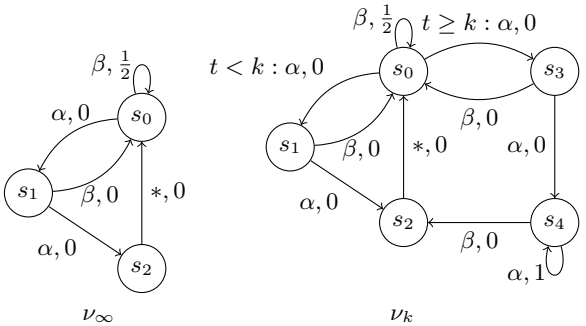
Lattimore and Hutter show that weak asymptotic optimality is possible in a countable class of deterministic environments using an MDL-agent that explores through bursts of random walks [LH11, Def. 10]. For classes of stochastic environments, BayesExp is weakly asymptotically optimal [Lat13, Ch. 5]. However, this requires the additional condition that the effective horizon grows sublinearly, $H_t(\varepsilon_t) \in o(t)$, while Theorem 4 does not require this condition.

Generally, weak asymptotic optimality and asymptotic optimality in mean are incomparable because the notions of convergence are incomparable for (bounded) random variables. First, for deterministic sequences (i.e. deterministic policies in deterministic environments), convergence in mean is equivalent to (regular) convergence, which implies convergence in Cesàro average, but not vice versa. Second, convergence in probability (and hence convergence in mean for bounded random variables) does not imply almost sure convergence of Cesàro averages [Sto13, Sec. 14.18]. We leave open the question whether the policy π_T is weakly asymptotically optimal.

3.3 STRONG ASYMPTOTIC OPTIMALITY

Strong asymptotic optimality is known to be impossible for deterministic policies [LH11, Thm. 8.1], but whether it is possible for stochastic policies is an open question. However, we show that Thompson sampling is not strongly asymptotically optimal.

Example 6 (Thompson Sampling is not Strongly Asymptotically Optimal)**.** Define $\mathcal{A} := \{\alpha, \beta\}$, $\mathcal{E} := \{0, 1/2, 1\}$, and assume geometric discounting, $\gamma_t := \gamma^t$ for $\gamma \in (0,1)$. Consider the following class of environments $\mathcal{M} := \{\nu_\infty, \nu_1, \nu_2, \ldots\}$ (transitions are labeled with action, reward):



Environment ν_k works just like environment ν_∞ except

that after time step k, the path to state s_3 gets unlocked and the optimal policy is to take action α twice from state s_0. The class $\mathcal{M}$ is a class of deterministic weakly communicating MDPs (but as an MDP ν_k has more than 5 states). The optimal policy in environment ν_∞ is to always take action β, the optimal policy for environment ν_k is to take action β for $t < k$ and then take action β in state s_1 and action α otherwise.

Suppose the policy π_T is acting in environment ν_∞. Since it is asymptotically optimal in the class $\mathcal{M}$, it has to take actions $\alpha\alpha$ from s_0 infinitely often: for $t < k$ environment ν_k is indistinguishable from ν_∞, so the posterior for ν_k is larger or equal to the prior. Hence there is always a constant chance of sampling ν_k until taking actions $\alpha\alpha$, at which point all environments ν_k for $k \leq t$ become falsified.

If the policy π_T decides to explore and take the first action α, it will be in state s_1. Let $æ_{<t}$ denote the current history. Then the ν_∞-optimal action is β and

$$V^*_{\nu_\infty}(æ_{<t}) = (1-\gamma)\left(0 + \gamma\frac{1}{2} + \gamma^2\frac{1}{2} + \dots\right) = \frac{\gamma}{2}.$$

The next action taken by π_T is α since any optimal policy for any sampled environment that takes action α once, takes that action again (and we are following that policy for an ε_t-effective horizon). Hence

$$V^{\pi_T}_{\nu_\infty}(æ_{<t}) \leq (1-\gamma)\left(0 + 0 + \gamma^2\frac{1}{2} + \gamma^3\frac{1}{2} + \dots\right) = \frac{\gamma^2}{2}.$$

Therefore $V^*_{\nu_\infty} - V^{\pi_T}_{\nu_\infty} \geq (\gamma - \gamma^2)/2 > 0$. This happens infinitely often with probability one and thus we cannot get almost sure convergence. ♢

We expect that strong asymptotic optimality can be achieved with Thompson sampling by resampling at every time step (with strong assumptions on the discount function).

4 REGRET

4.1 SETUP

In general environments classes worst-case regret is linear because the agent can get caught in a trap and be unable to recover [Hut05, Sec. 5.3.2]. To achieve sublinear regret we need to ensure that the agent can recover from mistakes. Formally, we make the following assumption.

Definition 7 (Recoverability)**.** An environment ν satisfies the *recoverability assumption* iff

$$\sup_\pi \left|\mathbb{E}^{\pi^*_\nu}_\nu[V^*_\nu(æ_{<t})] - \mathbb{E}^{\pi}_\nu[V^*_\nu(æ_{<t})]\right| \to 0 \text{ as } t \to \infty.$$

Recoverability compares following the worst policy π for $t-1$ time steps and then switching to the optimal policy π^*_ν to having followed π^*_ν from the beginning. The recoverability assumption states that switching to the optimal policy at any time step enables the recovery of most of the value.

Note that Definition 7 demands that it becomes less costly to recover from mistakes as time progresses. This should be regarded as an effect of the discount function: if the (effective) horizon grows, recovery becomes easier because the optimal policy has more time to perform a recovery. Moreover, recoverability is on the optimal policy, in contrast to the notion of ergodicity in MDPs which demands returning to a starting state regardless of the policy.

Remark 8 (Weakly Communicating POMDPs are Recoverable)**.** If the effective horizon is growing, $H_t(\varepsilon) \to \infty$ as $t \to \infty$, then any weakly communicating finite state partially observable MDP satisfies the recoverability assumption.

Definition 9 (Regret)**.** The *regret* of a policy π in environment μ is

$$R_m(\pi, \mu) := \sup_{\pi'} \mathbb{E}^{\pi'}_\mu\left[\sum_{t=1}^m r_t\right] - \mathbb{E}^{\pi}_\mu\left[\sum_{t=1}^m r_t\right].$$

Note that regret is undiscounted and always nonnegative. Moreover, the supremum is always attained by some policy (not necessarily the (V_μ-)optimal policy π^*_μ because that policy uses discounting), since the space of possible different policies for the first m actions is finite since we assumed the set of actions $\mathcal{A}$ and the set of percepts $\mathcal{E}$ to be finite.

Assumption 10 (Discount Function)**.** *Let the discount function γ be such that*

(a) $\gamma_t > 0$ *for all t,*

(b) γ_t *is monotone decreasing in t, and*

(c) $H_t(\varepsilon) \in o(t)$ *for all* $\varepsilon > 0$.

This assumption demands that the discount function is somewhat well-behaved: the function has no oscillations, does not become 0, and the horizon is not growing too fast.

Assumption 10 is satisfied by geometric discounting: $\gamma_t := \gamma^t > 0$ (a) for some fixed constant $\gamma \in (0,1)$ is monotone decreasing (b), $\Gamma_t = \gamma^t/(1-\gamma)$, and $H_t(\varepsilon) = \lceil \log_\gamma \varepsilon \rceil \in o(t)$ (c).

The problem with geometric discounting is that it makes the recoverability assumption very strong: since the horizon is not growing, the environment has to enable *faster recovery* as time progresses; in this case weakly communicating partially observable MDPs are *not* recoverable.

A choice with $H_t(\varepsilon) \to \infty$ that satisfies Assumption 10 is $\gamma_t := e^{-\sqrt{t}}/\sqrt{t}$ [Lat13, Sec. 2.3.1]. For this discount function $\Gamma_t \approx 2e^{-\sqrt{t}}$, $H_t(\varepsilon) \approx -\sqrt{t}\log\varepsilon + (\log\varepsilon)^2 \in o(t)$, and thus $H_t(\varepsilon) \to \infty$ as $t \to \infty$.

4.2 SUBLINEAR REGRET

This subsection is dedicated to the following theorem.

Theorem 11 (Sublinear Regret)**.** *If the discount function γ satisfies Assumption 10, the environment $\mu \in \mathcal{M}$ satisfies the recoverability assumption, and π is asymptotically optimal in mean, i.e.,*

$$\mathbb{E}_\mu^\pi\big[V_\mu^*(\text{æ}_{<t}) - V_\mu^\pi(\text{æ}_{<t})\big] \to 0 \text{ as } t \to \infty,$$

then $R_m(\pi, \mu) \in o(m)$.

If the items in Assumption 10 are violated, Theorem 11 can fail:

- If $\gamma_t = 0$ for some time steps t, our policy does not care about those time steps and might take actions that have large regret.
- Similarly if γ oscillates between high values and very low values: our policy might take high-regret actions in time steps with comparatively lower γ-weight.
- If the horizon grows linearly, infinitely often our policy might spend some constant fraction of the current effective horizon exploring, which incurs a cost that is a constant fraction of the total regret so far.

To prove Theorem 11, we apply the following technical lemma.

Lemma 12 (Value and Regret)**.** *Let $\varepsilon > 0$ and assume the discount function γ satisfies Assumption 10. Let $(d_t)_{t\in\mathbb{N}}$ be a sequence of numbers with $|d_t| \leq 1$ for all t. If there is a time step t_0 with*

$$\frac{1}{\Gamma_t}\sum_{k=t}^{\infty}\gamma_k d_k < \varepsilon \quad \forall t \geq t_0 \tag{5}$$

then

$$\sum_{t=1}^{m} d_t \leq t_0 + \varepsilon(m - t_0 + 1) + \frac{1+\varepsilon}{1-\varepsilon}H_m(\varepsilon)$$

Proof. This proof essentially follows the proof of [Hut06, Thm. 17]; see Appendix B. □

Proof of Theorem 11. Let $(\pi_m)_{m\in\mathbb{N}}$ denote any sequence of policies, such as a sequence of policies that attain the supremum in the definition of regret. We want to show that

$$\mathbb{E}_\mu^{\pi_m}\left[\sum_{t=1}^{m} r_t\right] - \mathbb{E}_\mu^{\pi}\left[\sum_{t=1}^{m} r_t\right] \in o(m).$$

For

$$d_k^{(m)} := \mathbb{E}_\mu^{\pi_m}[r_k] - \mathbb{E}_\mu^{\pi}[r_k] \tag{6}$$

we have $-1 \leq d_k^{(m)} \leq 1$ since we assumed rewards to be bounded between 0 and 1. Because the environment μ satisfies the recoverability assumption we have

$$\left|\mathbb{E}_\mu^{\pi_\mu^*}[V_\mu^*(\text{æ}_{<t})] - \mathbb{E}_\mu^{\pi}[V_\mu^*(\text{æ}_{<t})]\right| \to 0 \text{ as } t \to \infty, \text{ and}$$
$$\sup_m \left|\mathbb{E}_\mu^{\pi_\mu^*}[V_\mu^*(\text{æ}_{<t})] - \mathbb{E}_\mu^{\pi_m}[V_\mu^*(\text{æ}_{<t})]\right| \to 0 \text{ as } t \to \infty,$$

so we conclude that

$$\sup_m \left|\mathbb{E}_\mu^{\pi}[V_\mu^*(\text{æ}_{<t})] - \mathbb{E}_\mu^{\pi_m}[V_\mu^*(\text{æ}_{<t})]\right| \to 0$$

by the triangle inequality and thus

$$\sup_m \mathbb{E}_\mu^{\pi_m}[V_\mu^*(\text{æ}_{<t})] - \mathbb{E}_\mu^{\pi}[V_\mu^*(\text{æ}_{<t})] \to 0 \text{ as } t \to \infty. \tag{7}$$

By assumption the policy π is asymptotically optimal in mean, so we have

$$\mathbb{E}_\mu^{\pi}[V_\mu^*(\text{æ}_{<t})] - \mathbb{E}_\mu^{\pi}[V_\mu^{\pi}(\text{æ}_{<t})] \to 0 \text{ as } t \to \infty,$$

and with (7) this combines to

$$\sup_m \mathbb{E}_\mu^{\pi_m}[V_\mu^*(\text{æ}_{<t})] - \mathbb{E}_\mu^{\pi}[V_\mu^{\pi}(\text{æ}_{<t})] \to 0 \text{ as } t \to \infty.$$

From $V_\mu^*(\text{æ}_{<t}) \geq V_\mu^{\pi_m}(\text{æ}_{<t})$ we get

$$\limsup_{t\to\infty}\left(\sup_m \mathbb{E}_\mu^{\pi_m}[V_\mu^{\pi_m}(\text{æ}_{<t})] - \mathbb{E}_\mu^{\pi}[V_\mu^{\pi}(\text{æ}_{<t})]\right) \leq 0. \tag{8}$$

For $\pi' \in \{\pi, \pi_1, \pi_2, \ldots\}$ we have

$$\begin{aligned}
\mathbb{E}_\mu^{\pi'}[V_\mu^{\pi'}(\text{æ}_{<t})] &= \mathbb{E}_\mu^{\pi'}\left[\frac{1}{\Gamma_t}\mathbb{E}_\mu^{\pi'}\left[\sum_{k=t}^{\infty}\gamma_k r_k \,\middle|\, \text{æ}_{<t}\right]\right] \\
&= \mathbb{E}_\mu^{\pi'}\left[\frac{1}{\Gamma_t}\sum_{k=t}^{\infty}\gamma_k r_k\right] \\
&= \frac{1}{\Gamma_t}\sum_{k=t}^{\infty}\gamma_k \mathbb{E}_\mu^{\pi'}[r_k],
\end{aligned}$$

so from (6) and (8) we get

$$\limsup_{t\to\infty}\sup_m \frac{1}{\Gamma_t}\sum_{k=t}^{\infty}\gamma_k d_k^{(m)} \leq 0.$$

Let $\varepsilon > 0$ and choose t_0 independent of m and large enough such that $\sup_m \sum_{k=t}^{\infty}\gamma_k d_k^{(m)}/\Gamma_t < \varepsilon$ for all $t \geq t_0$. Now we let $m \in \mathbb{N}$ be given and apply Lemma 12 to get

$$\begin{aligned}
\frac{R_m(\pi,\mu)}{m} &= \frac{\sum_{k=1}^{m} d_k^{(m)}}{m} \\
&\leq \frac{t_0 + \varepsilon(m - t_0 + 1) + \frac{1+\varepsilon}{1-\varepsilon}H_m(\varepsilon)}{m}.
\end{aligned}$$

Since $H_t(\varepsilon) \in o(t)$ according to Assumption 10c we get $\limsup_{m\to\infty} R_m(\pi,\mu)/m \leq 0$. □

Example 13 (Converse of Theorem 11 is False)**.** Let μ be a two-armed Bernoulli bandit with means 0 and 1 and suppose we are using geometric discounting with discount factor $\gamma \in [0, 1)$. This environment is recoverable. If our policy π pulls the suboptimal arm exactly on time steps $1, 2, 4, 8, 16, \ldots$, regret will be logarithmic. However, on time steps $t = 2^n$ for $n \in \mathbb{N}$ the value difference $V^*_\mu - V^\pi_\mu$ is deterministically at least $1 - \gamma > 0$. $\lozenge$

4.3 IMPLICATIONS

We get the following immediate consequence.

Corollary 14 (Sublinear Regret for the Optimal Discounted Policy)**.** *If the discount function γ satisfies Assumption 10 and the environment μ satisfies the recoverability assumption, then $R_m(\pi^*_\mu, \mu) \in o(m)$.*

Proof. From Theorem 11 since the policy π^*_μ is (trivially) asymptotically optimal in μ. $\square$

If the environment does not satisfy the recoverability assumption, regret may be linear *even on the optimal policy*: the optimal policy maximizes discounted rewards and this short-sightedness might incur a tradeoff that leads to linear regret later on if the environment does not allow recovery.

Corollary 15 (Sublinear Regret for Thompson Sampling)**.** *If the discount function γ satisfies Assumption 10 and the environment $\mu \in \mathcal{M}$ satisfies the recoverability assumption, then $R_m(\pi_T, \mu) \in o(m)$ for the Thompson sampling policy π_T.*

Proof. From Theorem 4 and Theorem 11. $\square$

5 DISCUSSION

In this paper we introduced a reinforcement learning policy π_T based on Thompson sampling for general countable environment classes (Algorithm 1). We proved two asymptotic statements about this policy. Theorem 4 states that π_T is asymptotically optimal in mean: the value of π_T in the true environment converges to the optimal value. Corollary 15 states that the regret of π_T is sublinear: the difference of the expected average rewards between π_T and the best informed policy converges to 0. Both statements come without a concrete convergence rate because of the weak assumptions we made on the environment class.

Asymptotic optimality has to be taken with a grain of salt. It provides no incentive to the agent to avoid traps in the environment. Once the agent gets caught in a trap, all actions are equally bad and thus optimal: asymptotic optimality has been achieved. Even worse, an asymptotically optimal agent has to explore all the traps because they might contain hidden treasure. Overall, there is a dichotomy between the asymptotic nature of asymptotic optimality and the use of discounting to prioritize the present over the future. Ideally, we would want to give finite guarantees instead, but without additional assumptions this is likely impossible in this general setting. Our regret bound could be a step in the right direction, even though itself asymptotic in nature.

For Bayesians asymptotic optimality means that the posterior distribution $w(\cdot \mid æ_{<t})$ concentrates on environments that are indistinguishable from the true environment (but generally not on the true environment). This is why Thompson sampling works: any optimal policy of the environment we draw from the posterior will, with higher and higher probability, also be (almost) optimal in the true environment.

If the Bayesian mixture ξ is inside the class $\mathcal{M}$ (as it is the case for the class of lower semicomputable chronological semimeasures [Hut05]), then we can assign ξ a prior probability that is arbitrarily close to 1. Since the posterior of ξ is the same as the prior, Thompson sampling will act according to the Bayes-optimal policy most of the time. This means the Bayes-value of Thompson sampling can be very good; formally, $V^*_\xi(\epsilon) - V^{\pi_T}_\xi(\epsilon)$ can be made arbitrarily small, and thus Thompson sampling can have near-optimal Legg-Hutter intelligence [LH07].

In contrast, the Bayes-value of Thompson sampling can also be very bad: Suppose you have a class of $(n+1)$-armed bandits indexed $1, \ldots, n$ where bandit i gives reward $1 - \varepsilon$ on arm 1, reward 1 on arm $i + 1$, and reward 0 on all other arms. For geometric discounting and $\varepsilon < (1-\gamma)/(2-\gamma)$, it is Bayes-optimal to pull arm 1 while Thompson sampling will explore on average $n/2$ arms until it finds the optimal arm. The Bayes-value of Thompson sampling is $1/(n - \gamma_{n-1})$ in contract to $(1-\varepsilon)$ achieved by Bayes. For a horizon of n, the Bayes-optimal policy suffers a regret of εn and Thompson sampling a regret of $n/2$, which is much larger for small ε.

The exploration performed by Thompson sampling is qualitatively different from the exploration by BayesExp [Lat13, Ch. 5]. BayesExp performs phases of exploration in which it maximizes the expected information gain. This explores the environment class completely, even achieving off-policy prediction [OLH13, Thm. 7]. In contrast, Thompson sampling only explores on the optimal policies, and in some environment classes this will not yield off-policy prediction. So in this sense the exploration mechanism of Thompson sampling is more reward-oriented than maximizing information gain.

Possible avenues of future research are providing concrete convergence rates for specific environment classes and results for uncountable (parameterized) environment classes. For the latter, we have to use different analysis techniques because the true environment μ is typically assigned a prior probability of 0 (only a positive density) but the proofs of Lemma 5 and Theorem 4 rely on dividing by or tak-

ing a minimum over prior probabilities. We also left open whether Thompson sampling is weakly asymptotically optimal.

Acknowledgements

Example 6 was developed jointly with Stuart Armstrong. We thank Tom Everitt and Djallel Bouneffouf for proofreading.

REFERENCES

[AG11] Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, 2011.

[AJO10] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[ALL+09] John Asmuth, Lihong Li, Michael L Littman, Ali Nouri, and David Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 19–26, 2009.

[BB12] Sébastien Bubeck and Cesa-Nicolò Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[BD62] David Blackwell and Lester Dubins. Merging of opinions with increasing information. *The Annals of Mathematical Statistics*, pages 882–886, 1962.

[CL11] Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson sampling. In *Neural Information Processing Systems*, pages 2249–2257, 2011.

[DFR98] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian Q-learning. In *AAAI*, pages 761–768, 1998.

[Dur10] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4th edition, 2010.

[GM15] Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized Markov decision processes. In *Conference on Learning Theory*, pages 861–898, 2015.

[Hut00] Marcus Hutter. A theory of universal artificial intelligence based on algorithmic complexity. Technical report, 2000. `http://arxiv.org/abs/cs.AI/0004001`.

[Hut02] Marcus Hutter. Self-optimizing and Pareto-optimal policies in general environments based on Bayes-mixtures. In *Computational Learning Theory*, pages 364–379. Springer, 2002.

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

[Hut06] Marcus Hutter. General discounting versus average reward. In *Algorithmic Learning Theory*, pages 244–258. Springer, 2006.

[Hut09] Marcus Hutter. Discrete MDL predicts in total variation. In *Neural Information Processing Systems*, pages 817–825, 2009.

[KKM12] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

[Lat13] Tor Lattimore. *Theory of General Reinforcement Learning*. PhD thesis, Australian National University, 2013.

[LH07] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds & Machines*, 17(4):391–444, 2007.

[LH11] Tor Lattimore and Marcus Hutter. Asymptotically optimal agents. In *Algorithmic Learning Theory*, pages 368–382. Springer, 2011.

[LH14] Tor Lattimore and Marcus Hutter. General time consistent discounting. *Theoretical Computer Science*, 519:140–154, 2014.

[LH15] Jan Leike and Marcus Hutter. Bad universal priors and notions of optimality. In *Conference on Learning Theory*, pages 1244–1259, 2015.

[MKS+15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[NMRO13] Phuong Nguyen, Odalric-Ambrym Maillard, Daniil Ryabko, and Ronald Ortner. Competing with an infinite set of models in reinforcement learning. In *Artificial Intelligence and Statistics*, pages 463–471, 2013.

[OB10] Pedro A Ortega and Daniel A Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, pages 475–511, 2010.

[OLH13] Laurent Orseau, Tor Lattimore, and Marcus Hutter. Universal knowledge-seeking agents for stochastic environments. In *Algorithmic Learning Theory*, pages 158–172. Springer, 2013.

[OR14] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Neural Information Processing Systems*, pages 1466–1474, 2014.

[Ors13] Laurent Orseau. Asymptotic non-learnability of universal agents with computable horizon functions. *Theoretical Computer Science*, 473:149–156, 2013.

[ORvR13] Ian Osband, Dan Russo, and Benjamin van Roy. (More) efficient reinforcement learning via posterior sampling. In *Neural Information Processing Systems*, pages 3003–3011, 2013.

[SH15] Peter Sunehag and Marcus Hutter. Rationality, optimism and guarantees in general reinforcement learning. *Journal of Machine Learning Research*, 16:1345–1390, 2015.

[Sto13] Jordan M Stoyanov. *Counterexamples in Probability*. Courier Corporation, 3rd edition, 2013.

[Str00] Malcolm Strens. A Bayesian framework for reinforcement learning. In *International Conference on Machine Learning*, pages 943–950, 2000.

[Tho33] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

A LIST OF NOTATION

$:=$	defined to be equal
$\mathbb{N}$	the natural numbers, starting with 0
$\Delta\mathcal{Y}$	the set of all probability distributions on $\mathcal{Y}$
$\mathcal{X}^*$	the set of all finite strings over the alphabet $\mathcal{X}$
$\mathcal{X}^\infty$	the set of all infinite strings over the alphabet $\mathcal{X}$
$\mathcal{A}$	the (finite) set of possible actions
$\mathcal{E}$	the (finite) set of possible percepts
α, β	two different actions, $\alpha, \beta \in \mathcal{A}$
a_t	the action in time step t
e_t	the percept in time step t
r_t	the reward in time step t, bounded between 0 and 1
$\text{æ}_{<t}$	the history up to time $t-1$, i.e., the first $t-1$ interactions, $a_1e_1a_2e_2\ldots a_{t-1}e_{t-1}$
ϵ	the history of length 0
ε	a small positive real number
γ	the discount function $\gamma : \mathbb{N} \to \mathbb{R}_{\geq 0}$
Γ_t	a discount normalization factor, $\Gamma_t := \sum_{i=t}^\infty \gamma_i$
$H_t(\varepsilon)$	the ε-effective horizon, defined in (1)
π	a (stochastic) policy, i.e., a function $\pi : (\mathcal{A}\times\mathcal{E})^* \to \Delta\mathcal{A}$
π^*_ν	an optimal policy for environment ν
V^π_ν	value of the policy π in environment ν
n, k, i	natural numbers
t	(current) time step
m	time step at the end of an effective horizon
$\mathcal{M}$	a countable class of environments
ν, μ, ρ	environments from $\mathcal{M}$, i.e., functions $\nu : (\mathcal{A}\times\mathcal{E})^*\times\mathcal{A} \to \Delta\mathcal{E}$; μ is the true environment
ξ	Bayesian mixture over all environments in $\mathcal{M}$

B OMITTED PROOFS

Let P and Q be two probability distributions. We say P is *absolutely continuous with respect to* Q ($P \ll Q$) iff $Q(E) = 0$ implies $P(E) = 0$ for all measurable sets E. If $P \ll Q$ then there is a function dP/dQ called *Radon-Nikodym derivative* such that

$$\int f dP = \int f \frac{dP}{dQ} dQ$$

for all measurable functions f. This function dP/dQ can be seen as a density function of P with respect to the background measure Q.

Proof of Lemma 2. Let P, R, and Q be probability measures with $P \ll Q$ and $R \ll Q$ (we can take $Q := P/2 + R/2$), let dP/dQ and dR/dQ denote their Radon-Nikodym derivative with respect to Q, and let X denote a random variable with values in $[0,1]$. Then

$$\begin{aligned}\int X dP - \int X dR &= \int \left(X\frac{dP}{dQ} - X\frac{dR}{dQ}\right) dQ \\ &\leq \int_A X\left(\frac{dP}{dQ} - \frac{dR}{dQ}\right) dQ\end{aligned}$$

with $A := \left\{x \;\middle|\; \frac{dP}{dQ}(x) - \frac{dR}{dQ}(x) \geq 0\right\}$

$$\begin{aligned}&\leq \int_A \left(\frac{dP}{dQ} - \frac{dR}{dQ}\right) dQ \\ &= P(A) - R(A) \\ &\leq \sup_A |P(A) - R(A)| = D(P,R)\end{aligned}$$

From this also follows $\int X dR - \int X dP \leq D(R,P)$, and since D is symmetric we get

$$\left|\int X dP - \int X dR\right| \leq D(P,R). \tag{9}$$

According to Definition 1, the value function is the expectation of the random variable $\sum_{k=t}^m \gamma_k r_k/\Gamma_t$ that is bounded between 0 and 1. Therefore we can use (9) with $P := \nu^{\pi_1}(\cdot \mid \text{æ}_{<t})$ and $R := \rho^{\pi_2}(\cdot \mid \text{æ}_{<t})$ on the space $(\mathcal{A}\times\mathcal{E})^m$ of the histories of length $\leq m$ to conclude that $|V^{\pi_1,m}_\nu(\text{æ}_{<t}) - V^{\pi_2,m}_\rho(\text{æ}_{<t})|$ is bounded by $D_m(\nu^{\pi_1}, \rho^{\pi_2} \mid \text{æ}_{<t})$. □

Proof of Lemma 5. From Blackwell-Dubins' theorem [BD62] we get $D_\infty(\mu^\pi, \xi^\pi \mid \text{æ}_{<t}) \to 0$ μ^π-almost surely, and since D is bounded, this convergence also occurs in mean. Thus for every environment $\nu \in \mathcal{M}$,

$$\mathbb{E}^\pi_\nu\big[D_\infty(\nu^\pi, \xi^\pi \mid \text{æ}_{<t})\big] \to 0 \text{ as } t \to \infty. \tag{10}$$

Now

$$\begin{aligned}&\mathbb{E}^\pi_\mu[F^\pi_\infty(\text{æ}_{<t})] \\ &\leq \frac{1}{w(\mu)}\mathbb{E}^\pi_\xi[F^\pi_\infty(\text{æ}_{<t})] \\ &= \frac{1}{w(\mu)}\mathbb{E}^\pi_\xi\left[\sum_{\nu\in\mathcal{M}} w(\nu \mid \text{æ}_{<t}) D_\infty(\nu^\pi, \xi^\pi \mid \text{æ}_{<t})\right] \\ &= \frac{1}{w(\mu)}\mathbb{E}^\pi_\xi\left[\sum_{\nu\in\mathcal{M}} w(\nu)\frac{\nu^\pi(\text{æ}_{<t})}{\xi^\pi(\text{æ}_{<t})} D_\infty(\nu^\pi, \xi^\pi \mid \text{æ}_{<t})\right] \\ &= \frac{1}{w(\mu)}\sum_{\nu\in\mathcal{M}} w(\nu)\mathbb{E}^\pi_\nu\big[D_\infty(\nu^\pi, \xi^\pi \mid \text{æ}_{<t})\big] \to 0\end{aligned}$$

by [Hut05, Lem. 5.28ii] since total variation distance is bounded. □

Proof of Lemma 12. By Assumption 10a we have $\gamma_t > 0$ for all t and hence $\Gamma_t > 0$ for all t. By Assumption 10b have that γ is monotone decreasing, so we get for all $n \in \mathbb{N}$

$$\Gamma_t = \sum_{k=t}^{\infty} \gamma_k \leq \sum_{k=t}^{t+n-1} \gamma_t + \sum_{k=t+n}^{\infty} \gamma_k = n\gamma_t + \Gamma_{t+n}.$$

And with $n := H_t(\varepsilon)$ this yields

$$\frac{\gamma_t H_t(\varepsilon)}{\Gamma_t} \geq 1 - \frac{\Gamma_{t+H_t(\varepsilon)}}{\Gamma_t} \geq 1 - \varepsilon > 0. \tag{11}$$

In particular, this bound holds for all t and $\varepsilon > 0$.

Next, we define a series of nonnegative weights $(b_t)_{t \geq 1}$ such that

$$\sum_{t=t_0}^{m} d_k = \sum_{t=t_0}^{m} \frac{b_t}{\Gamma_t} \sum_{k=t}^{m} \gamma_k d_k.$$

This yields the constraints

$$\sum_{k=t_0}^{t} \frac{b_k}{\Gamma_k} \gamma_t = 1 \quad \forall t \geq t_0.$$

The solution to these constraints is

$$b_{t_0} = \frac{\Gamma_{t_0}}{\gamma_{t_0}}, \text{ and } b_t = \frac{\Gamma_t}{\gamma_t} - \frac{\Gamma_t}{\gamma_{t-1}} \text{ for } t > t_0. \tag{12}$$

Thus we get

$$\begin{aligned}
\sum_{t=t_0}^{m} b_t &= \frac{\Gamma_{t_0}}{\gamma_{t_0}} + \sum_{t=t_0+1}^{m} \left(\frac{\Gamma_t}{\gamma_t} - \frac{\Gamma_t}{\gamma_{t-1}} \right) \\
&= \frac{\Gamma_{m+1}}{\gamma_m} + \sum_{t=t_0}^{m} \left(\frac{\Gamma_t}{\gamma_t} - \frac{\Gamma_{t+1}}{\gamma_t} \right) \\
&= \frac{\Gamma_{m+1}}{\gamma_m} + m - t_0 + 1 \\
&\leq \frac{H_m(\varepsilon)}{1-\varepsilon} + m - t_0 + 1
\end{aligned}$$

for all $\varepsilon > 0$ according to (11).

Finally,

$$\begin{aligned}
\sum_{t=1}^{m} d_t &\leq \sum_{t=1}^{t_0} d_t + \sum_{t=t_0}^{m} \frac{b_t}{\Gamma_t} \sum_{k=t}^{m} \gamma_k d_k \\
&\leq t_0 + \sum_{t=t_0}^{m} \frac{b_t}{\Gamma_t} \sum_{k=t}^{\infty} \gamma_k d_k - \sum_{t=t_0}^{m} \frac{b_t}{\Gamma_t} \sum_{k=m+1}^{\infty} \gamma_k d_k
\end{aligned}$$

and using the assumption (5) and $d_t \geq -1$,

$$\begin{aligned}
&< t_0 + \sum_{t=t_0}^{m} b_t \varepsilon + \sum_{t=t_0}^{m} \frac{b_t \Gamma_{m+1}}{\Gamma_t} \\
&\leq t_0 + \frac{\varepsilon H_m(\varepsilon)}{1-\varepsilon} + \varepsilon(m - t_0 + 1) + \sum_{t=t_0}^{m} \frac{b_t \Gamma_{m+1}}{\Gamma_t}
\end{aligned}$$

For the latter term we substitute (12) to get

$$\begin{aligned}
\sum_{t=t_0}^{m} \frac{b_t \Gamma_{m+1}}{\Gamma_t} &= \frac{\Gamma_{m+1}}{\gamma_{t_0}} + \sum_{t=t_0+1}^{m} \left(\frac{\Gamma_{m+1}}{\gamma_t} - \frac{\Gamma_{m+1}}{\gamma_{t-1}} \right) \\
&= \frac{\Gamma_{m+1}}{\gamma_m} \leq \frac{H_m(\varepsilon)}{1-\varepsilon}
\end{aligned}$$

with (11). □