On Thompson Sampling and Asymptotic Optimality*

Jan Leike
DeepMind
Future of Humanity Institute
University of Oxford

Tor Lattimore
Indiana University

Laurent Orseau
DeepMind

Marcus Hutter
Australian National University

Abstract

We discuss some recent results on Thompson sampling for nonparametric reinforcement learning in countable classes of general stochastic environments. These environments can be non-Markovian, non-ergodic, and partially observable. We show that Thompson sampling learns the environment class in the sense that (1) asymptotically its value converges in mean to the optimal value and (2) given a recoverability assumption regret is sublinear. We conclude with a discussion about optimality in reinforcement learning.

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 [Mnih *et al.*, 2015]. 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. The goal of this line of work is to lift these restrictions; we consider *general reinforcement learning* [Hutter, 2005; Lattimore, 2013; Leike, 2016b], 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. However, we leave computational considerations aside for now.

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 [Lattimore and Hutter, 2011] 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 for every environment from the environment class. Asymptotic optimality can be achieved through an exploration component on top of a Bayes-optimal agent [Lattimore, 2013, Ch. 5] or through optimism [Sunehag and Hutter, 2015].

Asymptotic optimality in mean is essentially a qualitative version 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 forever thereafter. 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 [Bubeck and Bianchi, 2012]. 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 [Auer *et al.*, 2010]. 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 [Nguyen *et al.*, 2013]. 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 [Hutter, 2000] and the proof that it satisfies various optimality guarantees [Hutter, 2002]. Recently it was revealed that these optimality notions are subjective [Leike and Hutter, 2015]: 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. In particular, general Bayesian agents are not asymptotically optimal [Orseau, 2013]. These negative results put the Bayesian approach to RL into question. We remedy the situation by showing that

*This is an abriged version of Leike *et al.* [2016a]

using Bayesian techniques an agent can indeed be optimal in an objective sense.

We report recent results on a strategy called *Thompson sampling*, *posterior sampling*, or the *Bayesian control rule* [Thompson, 1933]. This strategy samples an environment ρ from the posterior, follows the ρ-optimal policy for a while, and then repeats. We show that this policy is asymptotically optimal in mean. 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.

Thompson sampling was originally proposed by Thompson as a bandit algorithm [Thompson, 1933]. It is easy to implement and often achieves quite good results [Chapelle and Li, 2011]. In multi-armed bandits it attains optimal regret [Agrawal and Goyal, 2011; Kaufmann *et al.*, 2012]. Thompson sampling has also been considered for MDPs: as model-free method relying on distributions over Q-functions with convergence guarantee [Dearden *et al.*, 1998], and as a model-based algorithm without theoretical analysis [Strens, 2000]. Bayesian and frequentist regret bounds have also been established [Osband *et al.*, 2013; Osband and Van Roy, 2014; Gopalan and Mannor, 2015]. PAC guarantees have been established for an optimistic variant of Thompson sampling for MDPs [Asmuth *et al.*, 2009].

For general RL, Thompson sampling was first suggested by Ortega and Braun [2010] 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.

Thompson sampling can be viewed as inference over optimal policies [Ortega and Braun, 2012]. With each environment $\nu \in \mathcal{M}$ we associate an optimal policy π^*_ν. At time step t conditional on history $æ_{<t}$, the posterior belief over environment ν is $w(\nu \mid æ_{<t})$. A Bayesian agent averages over all environments by maximizing reward according to the Bayesian mixture $\xi(\cdot \mid æ_{<t}) = \sum_\nu w(\nu \mid æ_{<t})\nu(\cdot \mid æ_{<t})$. In contrast, Thompson sampling averages over optimal policies and we get $\pi_T = \sum_\nu w(\nu \mid æ_{<t})\pi^*_\nu$. This way no explicit reward structure is needed, only a mapping from environment ν to optimal policy π^*_ν.

Osband and van Roy [2016] show that Thompson sampling is better than optimism because of the shape of the confidence sets in tabular MDPs. However, it can be argued that this is not an inherent flaw of the strategy of optimism, but rather of the way that confidence bounds are typically calculated.

More generally, Lattimore and Szepesvári [2016] point out that there seems to be something fundamentally flawed about both Thompson sampling and optimism. This is exhibited in a linear bandit where the most efficient exploration strategy involves taking actions that can be confidently judged as suboptimal. Optimistic strategies refrain from taking actions they know to be suboptimal even if they are informative. Thompson sampling is similar in this respect: the posterior concentrates around the likely optimal actions, so sampling a policy that takes the suboptimal action is very unlikely. This has been a known effect in the context of partial monitoring problems [Bartók *et al.*, 2014], that commonly involve information that can only be gained by taking suboptimal actions. However, in the most common theoretical frameworks for RL, multi-armed bandits and tabular MDPs, this problem does not exist and thus has gone unnoticed so far by the theoretical literature.

2 Preliminaries and Notation

In reinforcement learning, an agent interacts with an environment in cycles: at time step t the agent chooses an *action* a_t and receives a *percept* $e_t = (o_t, r_t)$ consisting of an *observation* o_t and a real-valued *reward* r_t; the cycle then repeats for time step $t+1$. A *history* is a sequence of actions and percepts: we use $æ_{<t}$ to denote a history of length $t-1$. In the following we assume that rewards are bounded between 0 and 1.

In contrast to most of the literature on reinforcement learning, we are agnostic towards the discounting strategy. Our goal is to maximize discounted rewards $\sum_{t=1}^\infty \gamma_t r_t$ for a fixed *discount function* $\gamma : \mathbb{N} \to \mathbb{R}$ with $\gamma_t \geq 0$ and $\sum_{t=1}^\infty \gamma_t < \infty$. Geometric discounting ($\gamma_t = \gamma^t$ for some constant $\gamma \in (0,1)$) is the most common form of discounting, although other forms can be used [Lattimore and Hutter, 2014]. The *discount normalization factor* is defined as $\Gamma_t := \sum_{k=t}^\infty \gamma_k$.

An *ε-effective horizon* $H_t(\varepsilon)$ is a horizon that is long enough to encompass all but an ε of the discount function's mass:

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

An ε-effective horizon is a central quantity in online reinforcement learning, and has a similar function to an episode in the episodic setting. It is the amount of time that an agent needs to plan ahead while losing only a fraction ε of the possible value. It can be used to constrain the planning horizon to a finite number of steps regardless of the discount function used. For geometric discounting, the horizon is $\lceil \log_\gamma \varepsilon \rceil$ (see Leike, 2016b, Tab. 4.1).

A *policy* is a function $\pi(a \mid æ_{<t})$ specifying the probability of taking action a after seeing the history $æ_{<t}$. Likewise, an *environment* is a function $\nu(e \mid æ_{<t}a_t)$ specifying the probability of emitting percept e after seeing the history $æ_{<t}$ and the action a_t. Together, a policy π and an environment ν generate a probability measure over histories denoted ν^π. We use $\mathbb{E}^\pi_\nu$ to denote the expectation over the history $æ_{<t}$ drawn from ν^π.

The *value* of a policy π in an environment ν given history $æ_{<t}$ is defined as

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

The *optimal value* is defined as $V^*_\nu(h) := \sup_\pi V^\pi_\nu(h)$, and the *optimal policy* is $\pi^*_\nu \in \arg\max_\pi V^\pi_\nu$.

Let $\mathcal{M}$ denote a countable class of environments. We assume that $\mathcal{M}$ is large enough to contain the true environment

Algorithm 1 Thompson sampling policy π_T

Require: ε_t, a monotone decreasing sequence of positive reals with $\varepsilon_t \to 0$ as $t \to \infty$.

while true **do**
  sample $\rho \sim w(\cdot \mid æ_{<t})$
  follow π^*_ρ for $H_t(\varepsilon_t)$ steps

(e.g. the class of all computable environments; Hutter, 2005). Let w 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

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

Finally, the *regret* of a policy π in environment μ is how much reward the agent has lost at time step m by not having followed the optimal policy from the beginning:

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

Note that regret is undiscounted and always nonnegative. Moreover, the supremum is always attained by some policy, which is usually not by the (V_μ-)optimal policy π^*_μ because that policy uses discounting.

3 Results

When introducing Thompson sampling for MDPs, 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" [Strens, 2000]. We follow Strens' suggestion and resample our policy at the *effective horizon*. The Thompson sampling policy π_T is defined in Algorithm 1. It is a stochastic policy since it occasionally involves sampling from a distribution. We prove the following main result.

Theorem 1 (Thompson Sampling is Asymptotically Optimal in Mean; Leike *et al.*, 2016a, Thm. 4)**.** *The policy π_T is asymptotically optimal in mean, i.e., for all environments μ from the countable class $\mathcal{M}$,*

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

Note that in contrast to previous results [Lattimore and Hutter, 2011], no assumptions on the discount function are required.

In general environments classes worst-case regret is linear because the agent can get caught in a trap and be unable to recover [Hutter, 2005, Sec. 5.3.2]. To achieve sublinear regret we need to ensure that the agent can recover from mistakes. We introduce the following technical assumption.

Definition 2 (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.

Our notion of recoverability 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. For growing effective horizons any weakly communicating finite state partially observable MDP is recoverable. Moreover, recovery is performed on the optimal policy, in contrast to the stronger notion of ergodicity in MDPs which demands returning to a starting state regardless of the policy.

Theorem 3 (Sublinear Regret; Leike *et al.*, 2016a, Thm. 11)**.** *Under suitable assumptions on the discount function, if the environment $\mu \in \mathcal{M}$ satisfies the recoverability assumption, and π is asymptotically optimal in mean, then regret is sublinear: $R_m(\pi, \mu) \in o(m)$.*

Together with Theorem 1 we get the following corollary.

Corollary 4 (Sublinear Regret for Thompson Sampling)**.** *Under suitable assumptions on the discount function, if the environment μ satisfies the recoverability assumption, then $R_m(\pi_T, \mu) \in o(m)$ for the Thompson sampling policy π_T.*

The assumptions on the discount function in Theorem 3 and Corollary 4 are satisfied by geometric discounting. However, since geometric discounting has a constant horizon, it makes the recoverability assumption very strong: the environment has to enable *faster recovery* as time progresses; in this case weakly communicating partially observable MDPs are *not* recoverable. An alternative discount function choice that satisfies the aforementioned assumptions but has a growing horizon is $\gamma_t := e^{-\sqrt{t}}/\sqrt{t}$ [Lattimore, 2013, Sec. 2.3.1].

Our recoverability assumption is necessary: if it is not satisfied, regret may be linear *even on the optimal policy*: the optimal policy maximizes discounted rewards and this shortsightedness might incur a tradeoff that leads to linear regret later on if the environment does not allow for recovery.

4 Discussion

A policy is asymptotically optimal if the agent learns to act optimally in any environment from the class $\mathcal{M}$. We showed that Thompson sampling is asymptotically optimal in mean (Theorem 1). Similar to BayesExp which is weakly asymptotically optimal if the horizon grows sublinearly [Lattimore, 2013, Ch. 5], both policies commit to exploration for several steps. This is necessary for optimality [Leike, 2016b, Ex. 5.19]:

> To achieve asymptotic optimality, the agent needs to explore infinitely often for an entire effective horizon.

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. Concisely, we can state the following impossibility result for non-recoverable environment classes [Leike, 2016b, Sec. 5.6.3]:

> Either the agent gets caught in a trap or it is not asymptotically optimal.

Note that for our optimality notions any finite number of time steps are irrelevant (optimality is a *tail event*): asymptotic optimality requires only convergence in the limit and sublinear regret is about the asymptotic behaviour of regret as a function of the horizon m. Hence an optimal agent can be arbitrarily lazy. Overall, there is a dichotomy between the asymptotic nature of optimality and the use of discounting to prioritize the present over the future. Ideally, we would want to give finite regret guarantees instead, but without additional assumptions this is likely impossible in this general setting.

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.

Interestingly, the Bayes-value of Thompson sampling can be very bad: Consider a class of $(n+1)$-armed bandit environments indexed $1, \dots, n$ 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 contrast 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 [Lattimore, 2013, 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 [Orseau *et al.*, 2013, Thm. 7]. However, off-policy prediction is too strong for reinforcement learning because it does not take the reward structure into account: it requires the agent to understand all parts of the environment, even the ones that are known to have low reward. A reward-oriented exploration strategy is information-directed sampling [Russo and Van Roy, 2014]. This bandit strategy trades off rewards with information gain about the optimal policy. However, those two quantities which are measured in two different units, which makes the tradeoff artificial. The exploration mechanism of Thompson sampling is reward-oriented and does not lead to off-policy prediction. As such, it leads to an exploration strategy that is measures on the value scale. In fact, the *exploration potential* proposed by Leike [2016a] is derived from Thompson sampling. It is a quantity that measures exploration on a value scale, making it commensurable with reward.

The result given in Theorem 1 is used by Leike *et al.* [2016b] to prove that in an arbitrary multi-agent environment, if all players act according to Algorithm 1 and their hypothesis class satisfies the *grain of truth assumption*, then they converge to a Nash equilibrium. The grain of truth assumption requires that the environment each player interacts with (the game combined with the other players), is in its hypothesis class $\mathcal{M}$. However, the result is not tied to Thompson sampling; the contribution of Leike *et al.* is to construct a general class that satisfies the grain of truth assumption. This class can then be combined with any asymptotically optimal policy, such as Lattimore and Hutter [2011]. However, the lack of additional assumptions on the discount function in Theorem 1 leads to a clean convergence theorem.

This paper only provides a brief introduction into general reinforcement learning. All proofs and further details can be found in Leike *et al.* [2016a]. The original work in general reinforcement learning is by Hutter [2005]. For a more recent introduction, in particular a longer discussion on optimality, we refer the reader to Leike [2016b]. For an empirical illustration of these and related results, see Aslanides *et al.* [Forthcoming].

Acknowledgements

We are grateful to Pedro Ortega for sharing his deep understanding of Thompson sampling and to Toby Ord for insightful discussions about asymptotic optimality.

References

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

John Aslanides, Jan Leike, and Marcus Hutter. General reinforcement learning algorithms: Survey and experiments. Forthcoming.

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.

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

Gábor Bartók, Dean Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

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.

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

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

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

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

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

Marcus Hutter. *Universal Artificial Intelligence*. Springer, 2005.

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

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

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

Tor Lattimore and Csaba Szepesvári. The end of optimism? an asymptotic analysis of finite-armed linear bandits. *arXiv preprint arXiv:1610.04491*, 2016.

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

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

Jan Leike, Tor Lattimore, Laurent Orseau, and Marcus Hutter. Thompson sampling is asymptotically optimal in general environments. In *Uncertainty in Artificial Intelligence*, pages 417–426, 2016.

Jan Leike, Jessica Taylor, and Benya Fallenstein. A formal solution to the grain of truth problem. In *Uncertainty in Artificial Intelligence*, pages 427–436, 2016.

Jan Leike. Exploration potential. In *European Workshop on Reinforcement Learning*, 2016.

Jan Leike. *Nonparametric General Reinforcement Learning*. PhD thesis, Australian National University, 2016.

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.

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.

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

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

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.

Pedro Ortega and Daniel Braun. Adaptive coding of actions and observations. In *NIPS Workshop on Information in Perception and Action*, 2012.

Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the Eluder Dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

Ian Osband and Benjamin van Roy. Why is posterior sampling better than optimism for reinforcement learning. In *European Workshop on Reinforcement Learning*, 2016.

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

Dan Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2014.

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

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

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.