Why is Posterior Sampling Better than Optimism for Reinforcement Learning?

Ian Osband [1 2] **Benjamin Van Roy** [1]

Abstract

Computational results demonstrate that posterior sampling for reinforcement learning (PSRL) dramatically outperforms existing algorithms driven by optimism, such as UCRL2. We provide insight into the extent of this performance boost and the phenomenon that drives it. We leverage this insight to establish an $\tilde{O}(H\sqrt{SAT})$ Bayesian regret bound for PSRL in finite-horizon episodic Markov decision processes. This improves upon the best previous Bayesian regret bound of $\tilde{O}(HS\sqrt{AT})$ for any reinforcement learning algorithm. Our theoretical results are supported by extensive empirical evaluation.

1. Introduction

We consider the reinforcement learning problem in which an agent interacts with a Markov decision process with the aim of maximizing expected cumulative reward (Burnetas & Katehakis, 1997; Sutton & Barto, 1998). Key to performance is how the agent balances between exploration to acquire information of long-term benefit and exploitation to maximize expected near-term rewards. In principle, dynamic programming can be applied to compute the Bayes-optimal solution to this problem (Bellman & Kalaba, 1959). However, this is computationally intractable for anything beyond the simplest of toy problems and direct approximations can fail spectacularly poorly (Munos, 2014). As such, researchers have proposed and analyzed a number of heuristic reinforcement learning algorithms.

The literature on efficient reinforcement learning offers statistical efficiency guarantees for computationally tractable algorithms. These provably efficient algorithms (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002) predominantly address the exploration-exploitation trade-off via *optimism in the face of uncertainty* (OFU): at any state, the agent assigns to each action an optimistically biased estimate of future value and selects the action with the greatest estimate. If a selected action is not near-optimal, the estimate must be overly optimistic, in which case the agent learns from the experience. Efficiency relative to less sophisticated exploration arises as the agent avoids actions that can neither yield high value nor informative data.

An alternative approach, based on Thompson sampling (Thompson, 1933), involves sampling a statistically plausibly set of action values and selecting the maximizing action. These values can be generated, for example, by sampling from the posterior distribution over MDPs and computing the state-action value function of the sampled MDP. This approach, originally proposed in Strens (2000), is called posterior sampling for reinforcement learning (PSRL). Computational results from Osband et al. (2013) demonstrate that PSRL dramatically outperforms existing algorithms based on OFU. The primary aim of this paper is to provide insight into the extent of this performance boost and the phenomenon that drives it.

We show that, in Bayesian expectation and up to constant factors, PSRL matches the statistical efficiency of *any* standard algorithm for OFU-RL. We highlight two key shortcomings of existing state of the art algorithms for OFU (Jaksch et al., 2010) and demonstrate that PSRL does not suffer from these inefficiencies. We leverage this insight to produce an $\tilde{O}(H\sqrt{SAT})$ bound for the Bayesian regret of PSRL in finite-horizon episodic Markov decision processes where H is the horizon, S is the number of states, A is the number of actions and T is the time elapsed. This improves upon the best previous bound of $\tilde{O}(HS\sqrt{AT})$ for any RL algorithm. We discuss why we believe PSRL satisfies a tighter $\tilde{O}(\sqrt{HSAT})$, though we have not proved that. We complement our theory with computational experiments that highlight the issues we raise; empirical results match our theoretical predictions.

More importantly, we highlights a tension in OFU RL between statistical efficiency and computational tractability. We argue that any OFU algorithm that matches PSRL in statistical efficiency would likely be computationally intractable. We provide proof of this claim in a restricted setting. Our key insight, and the potential benefits of exploration guided by posterior sampling, are not restricted to the simple tabular MDPs we analyze.

[1]Stanford University, California, USA [2]Deepmind, London, UK. Correspondence to: Ian Osband <ian.osband@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

2. Problem formulation

We consider the problem of learning to optimize a random finite-horizon MDP $M^*=(\mathcal{S},\mathcal{A},R^*,P^*,H,\rho)$ over repeated episodes of interaction, where $\mathcal{S} = \{1,..,S\}$ is the state space, $\mathcal{A} = \{1,..,A\}$ is the action space, H is the horizon, and ρ is the initial state distribution. In each time period $h = 1,..,H$ within an episode, the agent observes state $s_h \in \mathcal{S}$, selects action $a_h \in \mathcal{A}$, receives a reward $r_h \sim R^*(s_h,a_h)$, and transitions to a new state $s_{h+1} \sim P^*(s_h,a_h)$. We note that this formulation, where the unknown MDP M^* is treated as itself a random variable, is often called Bayesian reinforcement learning.

A policy μ is a mapping from state $s \in \mathcal{S}$ and period $h = 1,..,H$ to action $a \in \mathcal{A}$. For each MDP M and policy μ we define the state-action value function for each period h:

$$Q^M_{\mu,h}(s,a) := \mathbb{E}_{M,\mu}\left[\sum_{j=h}^{H} \bar{r}^M(s_j,a_j) \Big| s_h = s, a_h = a\right], \quad (1)$$

where $\bar{r}^M(s,a) = \mathbb{E}[r|r \sim R^M(s,a)]$. The subscript μ indicates that actions over periods $h+1,\ldots,H$ are selected according to the policy μ. Let $V^M_{\mu,h}(s) := Q^M_{\mu,h}(s,\mu(s,h))$. We say a policy μ^M is optimal for the MDP M if $\mu^M \in \text{argmax}_\mu V^M_{\mu,h}(s)$ for all $s \in \mathcal{S}$ and $h = 1,\ldots,H$.

Let $\mathcal{H}_t$ denote the history of observations made *prior* to time t. To highlight this time evolution within episodes, with some abuse of notation, we let $s_{kh} = s_t$ for $t = (k-1)H + h$, so that s_{kh} is the state in period h of episode k. We define $\mathcal{H}_{kh}$ analogously. An RL algorithm is a deterministic sequence $\{\pi_k | k = 1,2,\ldots\}$ of functions, each mapping $\mathcal{H}_{k1}$ to a probability distribution $\pi_k(\mathcal{H}_{k1})$ over policies, from which the agent samples a policy μ_k for the kth episode. We define the regret incurred by an RL algorithm π up to time T to be

$$\text{Regret}(T,\pi,M^*) := \sum_{k=1}^{\lceil T/H \rceil} \Delta_k, \quad (2)$$

where Δ_k denotes regret over the kth episode, defined with respect to true MDP M^* by

$$\Delta_k := \sum_{\mathcal{S}} \rho(s)(V^{M^*}_{\mu^*,1}(s) - V^{M^*}_{\mu_k,1}(s)) \quad (3)$$

with $\mu^* = \mu^{M^*}$. We note that the regret in (2) is random, since it depends on the unknown MDP M^*, the learning algorithm π and through the history $\mathcal{H}_t$ on the sampled transitions and rewards. We define

$$\text{BayesRegret}(T,\pi,\phi) := \mathbb{E}[\text{Regret}(T,\pi,M^*) \,|\, M^* \sim \phi], \quad (4)$$

as the Bayesian expected regret for M^* distributed according to the prior ϕ. We will assess and compare algorithm performance in terms of the regret and BayesRegret.

2.1. Relating performance guarantees

For the most part, the literature on efficient RL is sharply divided between the frequentist and Bayesian perspective (Vlassis et al., 2012). By volume, most papers focus on minimax regret bounds that hold with high probability for any $M^* \in \mathcal{M}$ some class of MDPs (Jaksch et al., 2010). Bounds on the BayesRegret are generally weaker analytical statements than minimax bounds on regret. A regret bound for any $M^* \in \mathcal{M}$ implies an identical bound on the BayesReget for any ϕ with support on $\mathcal{M}$. A partial converse is available for M^* drawn with non-zero probability under ϕ, but does not hold in general (Osband et al., 2013).

Another common notion of performance guarantee is given by so-called "sample-complexity" or PAC analyses that bound the number of ϵ-sub-optimal decisions taken by an algorithm (Kakade, 2003; Dann & Brunskill, 2015). In general, optimal bounds on regret $\tilde{O}(\sqrt{T})$ imply optimal bounds on sample complexity $\tilde{O}(\epsilon^{-2})$, whereas optimal bounds on the sample complexity give only an $\tilde{O}(T^{2/3})$ bound on regret (Osband, 2016).

Our formulation focuses on the simple setting on finite horizon MDPs, but there are several other problems of interest in the literature. Common formulations include the discounted setting[1] and problems with infinite horizon under some connectedness assumption (Bartlett & Tewari, 2009). This paper may contain insights that carry over to these settings, but we leave that analysis to future work.

Our analysis focuses upon Bayesian expected regret in finite horizon MDPs. We find this criterion amenable to (relatively) simple analysis and use it obtain actionable insight to the design of practical algorithms. We absolutely do not "close the book" on the exploration/exploitation problem - there remain many important open questions. Nonetheless, our work may help to develop understanding within some of the outstanding issues of statistical and computational efficiency in RL. In particular, we shed some light on how and why posterior sampling performs so much better than existing algorithms for OFU-RL. Crucially, we believe that many of these insights extend beyond the stylized problem of finite tabular MDPs and can help to guide the design of practical algorithms for generalization and exploration via randomized value functions (Osband, 2016).

3. Posterior sampling as stochastic optimism

There is a well-known connection between posterior sampling and optimistic algorithms (Russo & Van Roy, 2014). In this section we highlight the similarity of these approaches. We argue that posterior sampling can be thought of as a *stochastically* optimistic algorithm.

[1] Discount $\gamma = 1 - 1/H$ gives an effective horizon $O(H)$.

Before each episode, a typical OFU algorithm constructs a confidence set to represent the range of MDPs that are statistically plausible given prior knowledge and observations. Then, a policy is selected by maximizing value simultaneously over policies and MDPs in this set. The agent then follows this policy over the episode. It is interesting to contrast this approach against PSRL where instead of maximizing over a confidence set, PSRL samples a single statistically plausible MDP and selects a policy to maximize value for that MDP.

Algorithm 1 OFU RL

Input: confidence set constructor Φ

for episode $k = 1,2,..$ **do**
  Construct confidence set $\mathcal{M}_k = \Phi(\mathcal{H}_{k1})$
  Compute $\mu_k \in \text{argmax}_{\mu, M \in \mathcal{M}_k} V^M_{\mu,1}$
  for timestep $h = 1,..,H$ **do**
    take action $a_{kh} = \mu_k(s_{kh}, h)$
    update $H_{kh+1} = \mathcal{H}_{kh} \cup (s_{kh}, a_{kh}, r_{kh}, s_{kh+1})$
  end for
end for

Algorithm 2 PSRL

Input: prior distribution ϕ

for episode $k = 1,2,..$ **do**
  Sample MDP $M_k \sim \phi(\cdot \,|\, \mathcal{H}_{k1})$
  Compute $\mu_k \in \text{argmax}_{\mu} V^{M_k}_{\mu,1}$
  for timestep $h = 1,..,H$ **do**
    take action $a_{kh} = \mu_k(s_{kh}, h)$
    update $H_{kh+1} = \mathcal{H}_{kh} \cup (s_{kh}, a_{kh}, r_{kh}, s_{kh+1})$
  end for
end for

3.1. The blueprint for OFU regret bounds

The general strategy for the analysis of optimistic algorithms follows a simple recipe (Strehl & Littman, 2005; Szita & Szepesvári, 2010; Munos, 2014):

1. Design confidence sets (via concentration inequality) such that $M^* \in \mathcal{M}_k$ for all k with probability $\geq 1 - \delta$.
2. Decompose the regret in each episode
$$\Delta_k = V^{M^*}_{\mu^*,1} - V^{M^*}_{\mu_k,1} = \underbrace{V^{M^*}_{\mu^*,1} - V^{M_k}_{\mu_k,1}}_{\Delta^{\text{opt}}_k} + \underbrace{V^{M_k}_{\mu_k,1} - V^{M^*}_{\mu_k,1}}_{\Delta^{\text{conc}}_k}$$
where M_k is the imagined optimistic MDP.
3. By step (1.) $\Delta^{\text{opt}}_k \leq 0$ for all k with probability $\geq 1 - \delta$.
4. Use concentration results with a pigeonhole argument over all possible trajectories $\{\mathcal{H}_{11}, \mathcal{H}_{21}, ..\}$ to bound, with probability at least $1 - \delta$,
$$\text{Regret}(T, \pi, M^*) \leq \sum_{k=1}^{\lceil T/H \rceil} \Delta^{\text{conc}}_k \,|\, M^* \in \mathcal{M}_k \leq f(S, A, H, T, \delta).$$

3.2. Anything OFU can do, PSRL can expect to do too

In this section, we highlight the connection between posterior sampling and any optimistic algorithm in the spirit of Section 3.1. Central to our analysis will be the following notion of stochastic optimism (Osband et al., 2014).

Definition 1 (Stochastic optimism)**.**
Let X and Y be real-valued random variables with finite expectation. We will say that X is stochastically optimistic for Y if for any convex and increasing $u : \mathbb{R} \to \mathbb{R}$:
$$\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]. \tag{5}$$
We will write $X \succcurlyeq_{\text{so}} Y$ for this relation.

This notion of optimism is dual to second order stochastic dominance (Hadar & Russell, 1969), $X \succcurlyeq_{\text{so}} Y$ if and only if $-Y \succcurlyeq_{\text{ssd}} -X$. We say that PSRL is a stochastically optimistic algorithm since the random *imagined* value function $V^{M_k}_{\mu_k,1}$ is stochastically optimistic for the true optimal value function $V^{M^*}_{\mu^*,1}$ conditioned upon *any* possible history $\mathcal{H}_{k1}$ (Russo & Van Roy, 2014). This observation leads us to a general relationship between PSRL and the BayesRegret of *any* optimistic algorithm.

Theorem 1 (PSRL matches OFU-RL in BayesRegret)**.**
Let π^{opt} be any optimistic algorithm for reinforcement learning in the style of Algorithm 1. If π^{opt} satisfies regret bounds such that, for any M^* any $T > 0$ and any $\delta > 0$ the regret is bounded with probability at least $1 - \delta$
$$\text{Regret}(T, \pi^{\text{opt}}, M^*) \leq f(S, A, H, T, \delta). \tag{6}$$
Then, if ϕ is the distribution of the true MDP M^* and the proof of (6) follows Section 3.1, then for all $T > 0$
$$\text{BayesRegret}(T, \pi^{\text{PSRL}}, \phi) \leq 2 f(S, A, H, T, \delta = T^{-1}) + 2. \tag{7}$$

Sketch proof. This result is established in Osband et al. (2013) for the special case of $\pi^{\text{opt}} = \pi^{\text{UCRL2}}$. We include this small sketch as a refresher and a guide for high level intuition. First, note that conditioned upon any data $\mathcal{H}_{k1}$, the true MDP M^* and the sampled M_k are identically distributed. This means that $\mathbb{E}[\Delta^{\text{opt}} | \mathcal{H}_{k1}] \leq 0$ for all k. Therefore, to establish a bound upon the Bayesian regret of PSRL, we just need to bound $\sum_{k=1}^{\lceil T/H \rceil} \mathbb{E}[\Delta^{\text{conc}}_k \,|\, \mathcal{H}_k]$.

We can use that $M^* \,|\, \mathcal{H}_{k1} =^D M_k \,|\, \mathcal{H}_{k1}$ again in step (1.) from Section 3.1 to say that *both* M^*, M_k lie within $\mathcal{M}_k$ for all k with probability at least $1 - 2\delta$ via a union bound. This means we can bound the concentration error in PSRL,
$$\text{BayesRegret}(T, \pi^{\text{PSRL}}, \phi) \leq \sum_{k=1}^{\lceil T/H \rceil} \mathbb{E}[\Delta^{\text{conc}}_k \,|\, M^*, M_k \in \mathcal{M}_k] + 2\delta T$$

The final step follows from decomposing Δ^{conc}_k by adding and subtracting the imagined optimistic value $\tilde{V}_k$ generated by π^{opt}. Through an application of the triangle

inequality, $\Delta_k^{\text{conc}} \leq |V^{M_k}_{\mu_k,1} - \tilde{V}_k| + |\tilde{V}_k - V^*_{\mu_k,1}|$ we can mirror step (4.) to bound the regret from concentration, $\sum_{k=1}^{\lceil T/H \rceil} \mathbb{E}[\Delta_k^{\text{conc}} \,|\, M^*, M_k \in \mathcal{M}_k] \leq 2f(S,A,H,T,\delta)$. This result (and proof strategy) was established in multi-armed bandits by Russo & Van Roy (2014). We complete the proof of Theorem 1 with the choice $\delta = T^{-1}$ and that the regret is uniformly bounded by T. □

Theorem 1 suggest that, according to Bayesian expected regret, PSRL performs within a factor of 2 of any optimistic algorithm whose analysis follows Section 3.1. This includes the algorithms UCRL2 (Jaksch et al., 2010), UCFH (Dann & Brunskill, 2015), MORMAX (Szita & Szepesvári, 2010) and many more.

Importantly, and unlike existing OFU approaches, the algorithm performance is separated from the analysis of the confidence sets $\mathcal{M}_k$. This means that PSRL even attains the big O scaling of as-yet-undiscovered approaches to OFU, all at a computational cost no greater than solving a single known MDP - even if the matched OFU algorithm π^{opt} is computationally intractable.

4. Some shortcomings of existing OFU-RL

In this section, we discuss how and why existing OFU algorithms forgo the level of statistical efficiency enjoyed by PSRL. At a high level, this lack of statistical efficiency emerges from sub-optimal construction of the confidence sets $\mathcal{M}_k$. We present several insights that may prove crucial to the design of improved algorithms for OFU. More worryingly, we raise the question that perhaps the *optimal* statistical confidence sets $\mathcal{M}_k$ would likely be computationally intractable. We argue that PSRL offers a computationally tractable approximation to this unknown "ideal" optimistic algorithm.

Before we launch into a more mathematical argument it is useful to take intuition from a simple estimation problem, without any decision making. Consider an MDP with $A=1, H=2, S=2N+1$ as described in Figure 1. Every episode the agent transitions from $s=0$ uniformly to $s \in \{1,..,2N\}$ and receives a deterministic reward from $\{0,1\}$ depending upon this state. The simplicity of these examples means even a naive monte-carlo estimate of the value should concentrate $1/2 \pm \tilde{O}(1/\sqrt{n})$ after n episodes of interaction. Nonetheless, the confidence sets suggested by state of the art OFU-RL algorithm UCRL (Jaksch et al., 2010) become incredibly mis-calibrated as S grows.

To see how this problem occurs, consider any algorithm for for model-based OFU-RL that builds up confidence sets for each state and action independently, such as UCRL. Even if the estimates are tight in each state and action, the resulting optimistic MDP, simultaneously optimistic across each state and action, may be far too optimistic. Geometrically

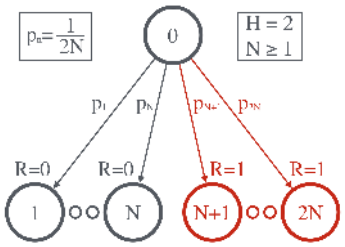

Figure 1. MDPs to illustrate the scaling with S.

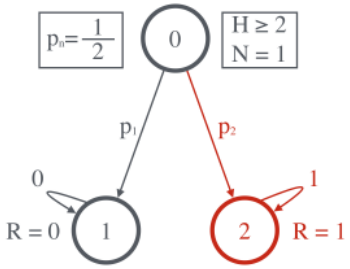

Figure 2. MDPs to illustrate the scaling with H.

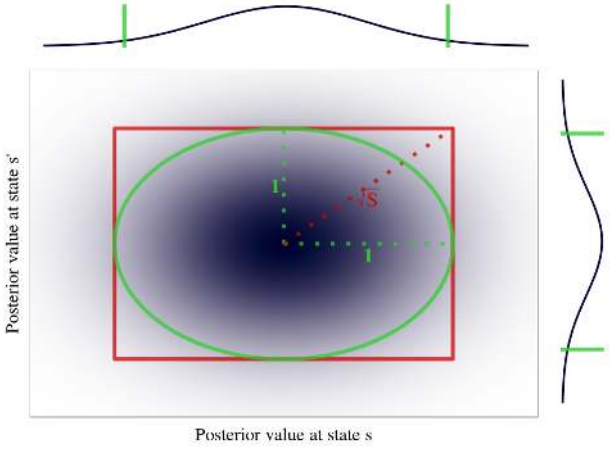

Figure 3. Union bounds give loose rectangular confidence sets.

these independent bounds form a rectangular confidence set. The corners of this rectangle will be $\sqrt{S}$ misspecified to the underlying distribution, an ellipse, when combined across S independent estimates (Figure 3).

Several algorithms for OFU-RL do exist which address this loose dependence upon S (Strehl et al., 2006; Szita & Szepesvári, 2010). However, these algorithms depend upon a partitioning of data for future value, which leads to a poor dependence upon the horizon H or equivalently the effective horizon $\frac{1}{1-\gamma}$ in discounted problems. We can use a similar toy example from Figure 2 to understand why combining independently optimistic estimates through time will contribute to a loose bound in H.

The natural question to ask is, "Why don't we simply apply these observations to design an optimistic algorithm which is simultaneously efficient in S and H?". The first impediment is that designing such an algorithm requires some new intricate concentration inequalities and analysis. Doing this rigorously may be challenging, but we believe it will be possible through a more careful application of existing tools to the insights we raise above. The bigger challenge is that, even if one were able to formally specify such an algorithm, the resulting algorithm may in general not be computationally tractable.

A similar observation to this problem of optimistic optimization has been shown in the setting of linear bandits (Dani et al., 2008; Russo & Van Roy, 2014). In these works they show that the problem of efficient optimization over ellipsoidal confidence sets can be NP-hard. This means that computationally tractable implementations of OFU have to rely upon inefficient rectangular confidence sets that give up a factor of $\sqrt{D}$ where D is the dimension of the underlying problem. By contrast, Thompson sampling approaches remain computationally tractable (since they require solving only a single problem instance) and so do not suffer from the loose confidence set construction. It remains an open question whether such an algorithm can be designed for finite MDPs. However, these previous results in the simpler bandit setting $H=1$ show that these problems with OFU-RL cannot be overcome in general.

4.1. Computational illustration

In this section we present a simple series of computational results to demonstrate this looseness in both S and H. We sample $K=1000$ episodes of data from the MDP and then examine the optimistic/sampled Q-values for UCRL2 and PSRL. We implement a version of UCRL2 optimized for finite horizon MDPs and implement PSRL with a uniform Dirichlet prior over the initial dynamics $P(0,1)=(p_1,..,p_{2N})$ and a $N(0,1)$ prior over rewards updating as if rewards had $N(0,1)$ noise. For both algorithms, if we say that R or P are *known* then we mean that we use the true R or P inside UCRL2 or PSRL. In each experiment, the estimates guided by OFU become extremely mis-calibrated, while PSRL remains stable.

The results of Figure 5 are particularly revealing. They demonstrates the potential pitfalls of OFU-RL even when the underlying transition dynamics *entirely known*. Several OFU algorithms have been proposed to remedy the loose UCRL-style L1 concentration from transitions (Filippi et al., 2010; Araya et al., 2012; Dann & Brunskill, 2015) but none of these address the inefficiency from hyper-rectangular confidence sets. As expected, these loose confidence sets lead to extremely poor performance in terms of the regret. We push full results to Appendix C along with comparison to several other OFU approaches.

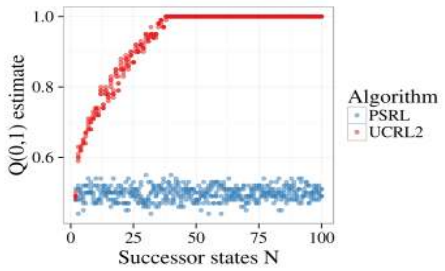

Figure 4. R known, P unknown, vary N in the MDP Figure 1.

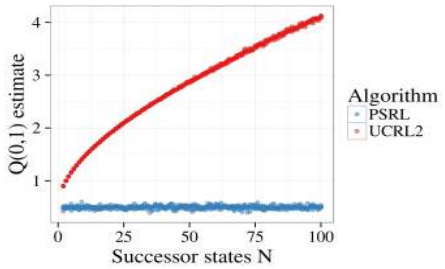

Figure 5. P known, R unknown, vary N in the MDP Figure 1.

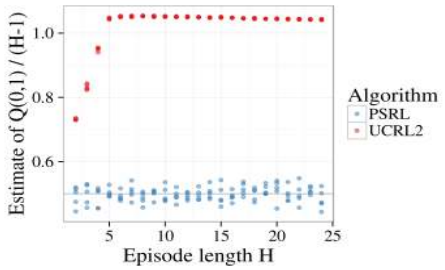

Figure 6. R,P unknown, vary H in the MDP Figure 2

5. Better optimism by sampling

Until now, all analyses of PSRL have come via comparison to some existing algorithm for OFU-RL. Previous work, in the spirit of Theorem 1, leveraged the existing analysis for UCRL2 to establish an $\tilde{O}(HS\sqrt{AT})$ bound upon the Bayesian regret (Osband et al., 2013). In this section, we present a new result that bounds the expected regret of PSRL $\tilde{O}(H\sqrt{SAT})$. We also include a conjecture that improved analysis could result in a Bayesian regret bound $\tilde{O}(\sqrt{HSAT})$ for PSRL, and that this result would be unimprovable (Osband & Van Roy, 2016).

5.1. From S to $\sqrt{S}$

In this section we present a new analysis that improves the bound on the Bayesian regret from S to $\sqrt{S}$. The proof of this result is somewhat technical, but the essential argument comes from the simple observation of the loose rectangular confidence sets from Section 4. The key to this analysis is a technical lemma on Gaussian-Dirichlet concentration (Osband & Van Roy, 2017).

Theorem 2. *Let M^* be the true MDP distributed according to prior ϕ with any independent Dirichlet prior over transitions. Then the regret for PSRL is bounded*

$$\text{BayesRegret}(T,\pi^{\text{PSRL}},\phi)=\tilde{O}\left(H\sqrt{SAT}\right). \quad (8)$$

Our proof of Theorem 2 mirrors the standard OFU-RL analysis from Section 3.1. To condense our notation we write $x_{kh}:=(s_{kh},a_{kh})$ and $V^k_{k,h}:=V^{M_k}_{\mu_k,h}$. Let the posterior mean of rewards $\hat{r}_k(x):=\mathbb{E}[\overline{r}^*(x)|\mathcal{H}_{k1}]$, transitions $\hat{P}_k(x):=\mathbb{E}[P^*(x)|\mathcal{H}_{k1}]$ with respective deviations from sampling noise $w^R(x):=\overline{r}_k(x)-\hat{r}_k(x)$ and $w^P_h(x):=(P_k(x)-\hat{P}_k(x))^TV^k_{kh+1}$.

We note that, conditional upon the data $\mathcal{H}_{k1}$ the true reward and transitions are independent of the rewards and transitions sampled by PSRL, so that $\mathbb{E}[\overline{r}^*(x)|\mathcal{H}_{k1}]=\hat{r}_k(x),\mathbb{E}[P^*(x)|\mathcal{H}_{k1}]=\hat{P}_k(x)$ for any x. However, $\mathbb{E}[w^R(x)|\mathcal{H}_{k1}]$ and $\mathbb{E}[w^P_h(x)|\mathcal{H}_{k1}]$ are generally non-zero, since the agent chooses its policy to optimize its reward under M_k. We can rewrite the regret from concentration via the Bellman operator (section 5.2 of Osband et al. (2013)),

$$\begin{aligned}
&\mathbb{E}\left[V^k_{k1}-V^*_{k1}|\mathcal{H}_{k1}\right]\\
=\;&\mathbb{E}\left[(\overline{r}_k-\overline{r}^*)(x_{k1})+P_k(x_{k1})^TV^k_{k2}-P^*(x_{k1})^TV^*_{k2}\mid\mathcal{H}_{k1}\right]\\
=\;&\mathbb{E}\Bigg[(\overline{r}_k-\overline{r}^*)(x_{k1})+\left(P_k(x_{k1})-\hat{P}_k(x_{k1})\right)^TV^k_{k2}\\
&+\mathbb{E}\left[\left(V^k_{k2}-V^*_{k2}\right)(s')|s'\sim P^*(x_{k1})\right]\mid\mathcal{H}_{k1}\Bigg]\\
=\;&\dots\\
=\;&\mathbb{E}\Bigg[\textstyle\sum_{h=1}^H\{\overline{r}_k(x_{k1})-\hat{r}^*(x_{k1})\}\\
&+\textstyle\sum_{h=1}^H\left\{\left(P_k(x_{kh})-\hat{P}_k(x_{kh})\right)^TV^k_{kh}\right\}\mid\mathcal{H}_{k1}\Bigg]\\
\le\;&\mathbb{E}\left[\textstyle\sum_{h=1}^H|w^R(x_{kh})|+\sum_{h=1}^H|w^P_h(x_{kh})|\mid\mathcal{H}_{k1}\right]. \quad (9)
\end{aligned}$$

We can bound the contribution from unknown rewards $w^R_k(x_{kh})$ with a standard argument from earlier work (Buldygin & Kozachenko, 1980; Jaksch et al., 2010).

Lemma 1 (Sub-Gaussian tail bounds)**.** *Let $x_1,..,x_n$ be independent samples from sub-Gaussian random variables. Then, for any $\delta>0$*

$$\mathbb{P}\left(\frac{1}{n}|\sum_{i=1}^n x_i|\ge\sqrt{\frac{2\log(2/\delta)}{n}}\right)\le\delta. \quad (10)$$

The key piece of our new analysis will be to show that the contribution from the transition estimate $\sum_{h=1}^H|w^P(x_{kh})|$ concentrates at a rate independent of S. At the root of our argument is the notion of stochastic optimism (Osband, 2016), which introduces a partial ordering over random variables. We make particular use of Lemma 2, that relates the concentration of a Dirichlet posterior with that of a matched Gaussian distribution (Osband & Van Roy, 2017).

Lemma 2 (Gaussian-Dirichlet dominance)**.** *For all fixed $V\in[0,1]^N$, $\alpha\in[0,\infty)^N$ with $\alpha^T\mathbb{1}\ge2$, if $X\sim N(\alpha^\top V/\alpha^\top\mathbf{1},1/\alpha^\top\mathbf{1})$ and $Y=P^TV$ for $P\sim\text{Dirichlet}(\alpha)$ then $X\succcurlyeq_{\text{so}}Y$.*

We can use Lemma 2 to establish a similar concentration bound on the error from sampling $w^P_h(x)$.

Lemma 3 (Transition concentration)**.** *For any independent prior over rewards with $\overline{r}\in[0,1]$, additive sub-Gaussian noise and an independent Dirichlet prior over transitions at state-action pair x_{kh}, then*

$$w^P_h(x_{kh})\le2H\sqrt{\frac{2\log(2/\delta)}{\max(n_k(x_{kh})-2,1)}} \quad (11)$$

with probability at least $1-\delta$.

Sketch proof. Our proof relies heavily upon some technical results from the note from Osband & Van Roy (2017). We cannot apply Lemma 2 directly to w^P, since the future value V^k_{kh+1} is itself be a random variable whose value depends on the sampled transition $P_k(x_{kh})$. However, although V^k_{kh+1} can vary with P_k, the structure of the MDP means that resultant $w^P(x_{kh})$ is still no more optimistic than the most optimistic possible *fixed* $V\in[0,H]^S$.

We begin this proof only for the simply family of MDPs with $S=2$, which we call $\mathcal{M}_2$. We write $p:=P_k(x_{kh})(1)$ for the first component of the unknown transition at x_{kh} and similarly $\hat{p}:=\hat{P}_k(x_{kh})(1)$. We can then bound the transition concentration,

$$\begin{aligned}
|w^P_h(x_{kh})| &= |(P_k(x_{kh})-\hat{P}_k(x_{kh}))^TV^k_{kh+1}|\\
&\le |(p-\hat{p})||(V^k_{kh+1}(1)-V^k_{kh+1}(2))|\\
&\le |p-\hat{p}|\sup_{R_k,P_k}|(V^k_{kh+1}(1)-V^k_{kh+1}(2))|\\
&\le |(p-\hat{p})|H \quad (12)
\end{aligned}$$

Lemma 2 now implies that for any $\alpha\in\mathbb{R}_+$ with $\alpha^T\mathbb{1}\ge2$, the random variables $p\sim\text{Dirichlet}(\alpha)$ and $X\sim N(0,\sigma^2=1/\alpha^T\mathbb{1})$ are ordered,

$$X\succcurlyeq_{\text{so}}p-\hat{p}\implies|X|H\succcurlyeq_{\text{so}}|p-\hat{p}|H\succcurlyeq_{\text{so}}|w^P_h(x_{kh})|. \quad (13)$$

We conclude the proof for $M\in\mathcal{M}_2$ through an application of Lemma 1. To extend this argument to multiple states $S>2$ we consider the marginal distribution of P_k over any subset of states, which is Beta distributed similar to (12). We push the details to Appendix A. □

To complete the proof of Theorem 2 we combine Lemma 1 with Lemma 3. We rescale $\delta\leftarrow\delta/2SAT$ so that these confidence sets hold at each $R(s,a),P(s,a)$ via union bound with probability at least $1-\frac{1}{T}$,

$$\begin{aligned}
&\mathbb{E}\left[\textstyle\sum_{h=1}^H\left\{|w^R(x_{kh})|+|w^P_h(x_{kh})|\right\}\mid\mathcal{H}_{k1}\right]\\
\le\;&\textstyle\sum_{h=1}^H2(H+1)\sqrt{\frac{2\log(4SAT)}{\max(n_k(x_{kh})-2,1)}}. \quad (14)
\end{aligned}$$

We can now use (14) together with a pigeonhole principle over the number of visits to each state and action:

$$\begin{aligned} & \text{BayesRegret}(T,\pi^{\text{PSRL}},\phi) \\ \leq\ & \sum_{k=1}^{\lceil T/H\rceil}\sum_{h=1}^{H} 2(H+1)\sqrt{\tfrac{2\log(4SAT)}{n_k(x_{kh})}} + 2SA + 1 \\ \leq\ & 10H\sqrt{SAT\log(4SAT)}. \end{aligned}$$

This completes the proof of Theorem 2. □

Prior work has designed similar OFU approaches that improve the learning scaling with S. MORMAX (Szita & Szepesvári, 2010) and delayed Q-learning (Strehl et al., 2006), in particular, come with sample complexity bounds that are linear in S, and match lower bounds. But even in terms of sample complexity, these algorithms are not necessarily an improvement over UCRL2 or its variants (Dann & Brunskill, 2015). For clarity, we compare these algorithms in terms of $T^{\pi}(\epsilon) := \min\left\{T \mid \frac{1}{T}\text{BayesRegret}(T,\pi,\phi) \leq \epsilon\right\}$.

DelayQ	MORMAX	UCRL2	PSRL Theorem 2
$\tilde{O}\left(\frac{H^9SA}{\epsilon^4}\right)$	$\tilde{O}\left(\frac{H^7SA}{\epsilon^2}\right)$	$\tilde{O}\left(\frac{H^2S^2A}{\epsilon^2}\right)$	$\tilde{O}\left(\frac{H^2SA}{\epsilon^2}\right)$

Table 1. Learning times compared in terms of $T^{\pi}(\epsilon)$.

Theorem 1 implies $T^{\text{PSRL}}(\epsilon) = \tilde{O}(\frac{H^2SA}{\epsilon^2})$. MORMAX and delayed Q-learning reduces the S-dependence of UCRL2, but this comes at the expense of worse dependence on H, and the resulting algorithms are not practical.

5.2. From H to $\sqrt{H}$

Recent analyses (Lattimore & Hutter, 2012; Dann & Brunskill, 2015) suggest that simultaneously reducing the dependence of H to $\sqrt{H}$ may be possible. They note that "local value variance" satisfies a Bellman equation. Intuitively this captures that if we transition to a bad state $V \simeq 0$, then we cannot transition anywhere much worse during this episode. This relation means that $\sum_{h=1}^{H} w_h^P(x_{kh})$ should behave more as if they were independent and grow $O(\sqrt{H})$, unlike our analysis which crudely upper bounds them each in turn $O(H)$. We present a sketch towards an analysis of Conjecture 1 in Appendix B.

Conjecture 1. *For any prior over rewards with $\overline{r} \in [0,1]$, additive sub-Gaussian noise and any independent Dirichlet prior over transitions, we conjecture that*

$$\mathbb{E}\left[\text{Regret}(T,\pi^{\text{PSRL}},M^*)\right] = \tilde{O}\left(\sqrt{HSAT}\right), \qquad (15)$$

and that this matches the lower bounds for any algorithm up to logarithmic factors.

The results of (Bartlett & Tewari, 2009) adapted to finite horizon MDPs would suggest a lower bound $\Omega(H\sqrt{SAT})$ on the minimax regret for any algorithm. However, the associated proof is incorrect (Osband & Van Roy, 2016). The strongest lower bound with a correct proof is $\Omega(\sqrt{HSAT})$ (Jaksch et al., 2010). It remains an open question whether such a lower bound applies to Bayesian regret over the class of priors we analyze in Theorem 2.

One particularly interesting aspect of Conjecture 1 is that we can construct another algorithm that satisfies the proof of Theorem 2 but would not satisfy the argument for Conjecture 1 of Appendix B. We call this algorithm Gaussian PSRL, since it operates in a manner similar to PSRL but actually uses the Gaussian sampling we use for the *analysis* of PSRL in its algorithm.

Algorithm 3 Gaussian PSRL

Input: Posterior MAP estimates $\overline{r}_k$, $\hat{P}_k$, visit counts n_k
Output: Random $Q_{k,h}(s,a) \succcurlyeq_{\text{so}} Q_h^*(s,a)$ for all (s,a,h)

```
Initialize Q_{k,H+1}(s,a) ← 0 for all (s,a)
for timestep h = H, H-1, .., 1 do
    V_{k,h+1}(s) ← max_α Q_{k,h+1}(s,α)
    Sample w_k(s,a,h) ~ N(0, (H+1)^2 / max(n_k(s,a)-2, 1))
    Q_{k,h}(s,a) ← r̄_k(s,a) + P̂_k(s,a)^T V + w_k(s,a,h)  ∀(s,a)
end for
```

Algorithm 3 presents the method for sampling random Q-values according to Gaussian PSRL, the algorithm then follows these samples greedily for the duration of the episode, similar to PSRL. Interestingly, we find that our experimental evaluation is consistent with $\tilde{O}(HS\sqrt{AT})$, $\tilde{O}(H\sqrt{SAT})$ and $\tilde{O}(\sqrt{HSAT})$ for UCRL2, Gaussian PSRL and PSRL respectively.

5.3. An empirical investigation

We now discuss a computational study designed to illustrate how learning times scale with S and H, and to empirically investigate Conjecture 1. The class of MDPs we consider involves a long chain of states with $S = H = N$ and with two actions: left and right. Each episode the agent begins in state 1. The optimal policy is to head right at every timestep, all other policies have zero expected reward. Inefficient exploration strategies will take $\Omega(2^N)$ episodes to learn the optimal policy (Osband et al., 2014).

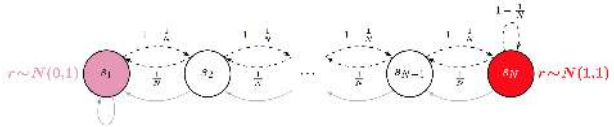

Figure 7. MDPs that highlight the need for efficient exploration.

We evaluate several learning algorithms from ten random seeds and $N = 2,..,100$ for up to ten million episodes each. Our goal is to investigate their empirical performance and scaling. We believe this is the first ever large scale empirical investigation into the scaling properties of algorithms for efficient exploration.

We highlight results for three algorithms with $\tilde{O}(\sqrt{T})$ Bayesian regret bounds: UCRL2, Gaussian PSRL and PSRL. We implement UCRL2 with confidence sets optimized for finite horizon MDPs. For the Bayesian algorithms we use a uniform Dirichlet prior for transitions and $N(0,1)$ prior for rewards. We view these priors as simple ways to encode very little prior knowledge. Full details and a link to source code are available in Appendix D.

Figure 8 display the regret curves for these algorithms for $N \in \{5,10,30,50\}$. As suggested by our analysis, PSRL outperforms Gaussian PSRL which outperforms UCRL2. These differences seems to scale with the length of the chain N and that even for relatively small MDPs, PSRL is many orders of magnitude more efficient than UCRL2.

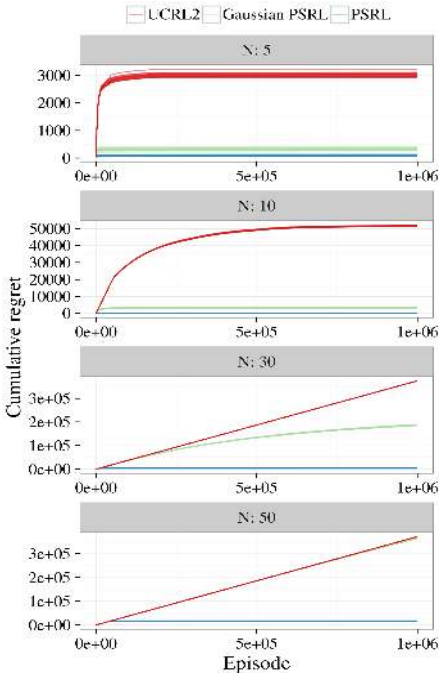

Figure 8. PSRL outperforms other methods by large margins.

We investigate the empirical scaling of these algorithms with respect to N. The results of Theorem 2 and Conjecture 1 only bound the Bayesian regret according to the prior ϕ. The family of environments we consider in this example are decidedly *not* from this uniform distribution; in fact they are chosen to be as difficult as possible. Nevertheless, the results of Theorem 2 and Conjecture 1 provide remarkably good description for the behavior we observe.

Define $\text{learning time}(\pi,N) := \min\left\{K \mid \frac{1}{K}\sum_{k=1}^{K}\Delta_k \leq 0.1\right\}$ for the algorithm π on the MDP from Figure 7 with size N. For any $B_\pi > 0$, the regret bound $\tilde{O}(\sqrt{B_\pi T})$ would imply $\log(\text{learning time})(\pi,N) = B_\pi H \times \log(N) + o(\log(N))$. In the cases of Figure 7 with $H = S = N$ then the bounds $\tilde{O}(HS\sqrt{AT})$, $\tilde{O}(H\sqrt{SAT})$ and $\tilde{O}(\sqrt{HSAT})$ would suggest a slope B_π of $5, 4$ and 3 respectively.

Remarkably, these high level predictions match our empirical results almost exactly, as we show in Figure 9. These results provide some support to Conjecture 1 and even, since the spirit of these environments is similar example used in existing proofs, the ongoing questions of fundamental lower bounds (Osband & Van Roy, 2016). Further, we note that every single seed of PSRL and Gaussian PSRL learned the optimal policy for every single N. We believe that this suggests it may be possible to extend our Bayesian analysis to provide minimax regret bounds of the style in UCRL2 for suitable choice of diffuse uninformative prior.

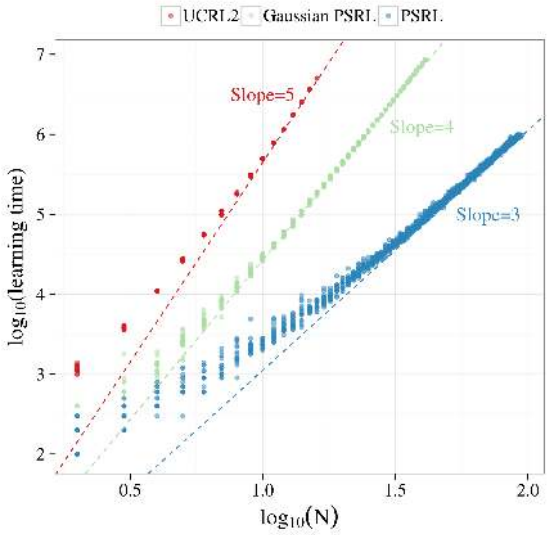

Figure 9. Empirical scaling matches our conjectured analysis.

6. Conclusion

PSRL is orders of magnitude more statistically efficient than UCRL *and* the same computational cost as solving a known MDP. We believe that analysts will be able to formally specify an OFU approach to RL whose statistical efficiency matches PSRL. However, we argue that the resulting confidence sets which address both the coupling over H and S may result in a computationally intractable optimization problem. Posterior sampling offers a computationally tractable approach to statistically efficient exploration.

We should stress that the finite tabular setting we analyze is not a reasonable model for most problems of interest. Due to the curse of dimensionality, RL in practical settings will require generalization between states and actions. The goal of this paper is not just to improve a mathematical bound in a toy example (although we do also do that). Instead, we hope this simple setting can highlight some shortcomings of existing approaches to "efficient RL" and provide insight into why algorithms based on sampling may offer important advantages. We believe that these insights may prove valuable as we move towards algorithms that solve the problem we really care about: synthesizing efficient exploration with powerful generalization.

Acknowledgements

This work was generously supported by DeepMind, a research grant from Boeing, a Marketing Research Award from Adobe, and a Stanford Graduate Fellowship, courtesy of PACCAR. The authors would like to thank Daniel Russo for many hours of discussion and insight leading to this research, Shipra Agrawal and Tor Lattimore for pointing out several flaws in some early proof steps, anonymous reviewers for their helpful comments and many more colleagues at DeepMind including Remi Munos, Mohammad Azar and more for inspirational conversations.

References

Araya, Mauricio, Buffet, Olivier, and Thomas, Vincent. Near-optimal brl using optimistic local transitions. *arXiv preprint arXiv:1206.4613*, 2012.

Asmuth, John, Li, Lihong, Littman, Michael L, Nouri, Ali, and Wingate, David. A Bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 19–26. AUAI Press, 2009.

Bartlett, Peter L. and Tewari, Ambuj. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI2009)*, pp. 35–42, June 2009.

Bellman, Richard and Kalaba, Robert. On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2): 1–9, 1959.

Brafman, Ronen I. and Tennenholtz, Moshe. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Buldygin, Valerii V and Kozachenko, Yu V. Sub-gaussian random variables. *Ukrainian Mathematical Journal*, 32 (6):483–489, 1980.

Burnetas, Apostolos N and Katehakis, Michael N. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Dani, Varsha, Hayes, Thomas P., and Kakade, Sham M. Stochastic linear optimization under bandit feedback. In *COLT*, pp. 355–366, 2008.

Dann, Christoph and Brunskill, Emma. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. TBA, 2015.

Filippi, Sarah, Cappé, Olivier, and Garivier, Aurélien. Optimism in reinforcement learning and kullback-leibler divergence. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pp. 115–122. IEEE, 2010.

Fonteneau, Raphaël, Korda, Nathan, and Munos, Rémi. An optimistic posterior sampling strategy for Bayesian reinforcement learning. In *NIPS 2013 Workshop on Bayesian Optimization (BayesOpt2013)*, 2013.

Gopalan, Aditya and Mannor, Shie. Thompson sampling for learning parameterized Markov decision processes. *arXiv preprint arXiv:1406.7498*, 2014.

Hadar, Josef and Russell, William R. Rules for ordering uncertain prospects. *The American Economic Review*, pp. 25–34, 1969.

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

Kakade, Sham. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, University College London, 2003.

Kearns, Michael J. and Singh, Satinder P. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

Kolter, J Zico and Ng, Andrew Y. Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 513–520. ACM, 2009.

Lattimore, Tor and Hutter, Marcus. PAC bounds for discounted MDPs. In *Algorithmic learning theory*, pp. 320–334. Springer, 2012.

Munos, Rémi. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. 2014.

Osband, Ian. *Deep Exploration via Randomized Value Functions*. PhD thesis, Stanford, 2016.

Osband, Ian and Van Roy, Benjamin. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pp. 1466–1474, 2014a.

Osband, Ian and Van Roy, Benjamin. Near-optimal reinforcement learning in factored MDPs. In *Advances in Neural Information Processing Systems*, pp. 604–612, 2014b.

Osband, Ian and Van Roy, Benjamin. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Osband, Ian and Van Roy, Benjamin. Gaussian-dirichlet posterior dominance in sequential learning. *arXiv preprint arXiv:1702.04126*, 2017.

Osband, Ian, Russo, Daniel, and Van Roy, Benjamin. (More) efficient reinforcement learning via posterior sampling. In *NIPS*, pp. 3003–3011. Curran Associates, Inc., 2013.

Osband, Ian, Van Roy, Benjamin, and Wen, Zheng. Generalization and exploration via randomized value functions. *arXiv preprint arXiv:1402.0635*, 2014.

Russo, Daniel and Van Roy, Benjamin. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Strehl, Alexander L and Littman, Michael L. A theoretical analysis of model-based interval estimation. In *Proceedings of the 22nd international conference on Machine learning*, pp. 856–863. ACM, 2005.

Strehl, Alexander L., Li, Lihong, Wiewiora, Eric, Langford, John, and Littman, Michael L. PAC model-free reinforcement learning. In *ICML*, pp. 881–888, 2006.

Strens, Malcolm J. A. A Bayesian framework for reinforcement learning. In *ICML*, pp. 943–950, 2000.

Sutton, Richard and Barto, Andrew. *Reinforcement Learning: An Introduction*. MIT Press, March 1998.

Szita, István and Szepesvári, Csaba. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1031–1038, 2010.

Thompson, W.R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Vlassis, Nikos, Ghavamzadeh, Mohammad, Mannor, Shie, and Poupart, Pascal. Bayesian reinforcement learning. In *Reinforcement Learning*, pp. 359–386. Springer, 2012.