APPENDICES

A. Proof of Lemma 3

This section centers around the proof of Lemma 3, which we reproduce below for completeness. In the main paper we present a simple sketch for the special case of $S=2$. We now extend this argument to general MDPs with $S>2$. The main strategy for this proof is to proceed via an inductive argument and consider the contribution of each component of P_k in turn. We will see that, for any choice of component, the resultant random variable is dominated by a matched Gaussian random variable just as in (12).

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*

$$w_h^P(x_{kh}) \leq 2H\sqrt{\frac{2\log(2/\delta)}{\max(n_k(x_{kh})-2,1)}} \tag{11}$$

with probability at least $1-\delta$.

Our analysis of Lemma 3 will rely heavily upon the technical analysis of Osband & Van Roy (2017). We first reproduce Lemma 2 from Osband & Van Roy (2017) in terms of stochastic optimism, rather than second order stochastic dominance.

Lemma 4 (Beta vs Dirichlet dominance)**.**
Let $X=P^\top v$ for the random variable $P\sim\text{Dirichlet}(\alpha)$ and constants $v\in\mathbb{R}^S$ and $\alpha\in\mathbb{R}^S_+$. Without loss of generality, assume $v_1\leq v_2\leq\cdots\leq v_S$. Let $\tilde{\alpha}=\sum_{i=1}^s\alpha_i(v_i-v_1)/(v_d-v_1)$ and $\tilde{\beta}=\sum_{i=1}^d\alpha_i(v_d-v_i)/(v_d-v_1)$. Then, there exists a random variable $\tilde{P}\sim\text{Beta}(\tilde{\alpha},\tilde{\beta})$ such that, for $\tilde{X}=\tilde{P}v_d+(1-\tilde{P})v_1$, $\mathbb{E}[\tilde{X}|X]=X$ and $\tilde{X}\succcurlyeq_{\text{so}}X$.

Proof. Let $\gamma_i=\text{Gamma}(\alpha,1)$ be independent and identically distributed and let $\overline{\gamma}=\sum_{i=1}^d\gamma_i$, so that $P\equiv_D\gamma/\overline{\gamma}$. Let $\alpha_i^0=\alpha_i(v_i-v_1)/(v_d-v_1)$ and $\alpha_i^1=\alpha_i(v_d-v_i)/(v_d-v_1)$ so that $\alpha=\alpha^0+\alpha^1$. Define independent random variables $\gamma^0\sim\text{Gamma}(\alpha_i^0,1)$ and $\gamma^1\sim\text{Gamma}(\alpha_i^1,1)$ so that $\gamma\equiv_D\gamma^0+\gamma^1$.

Take γ^0 and γ^1 to be independent, and couple these variables with γ so that $\gamma=\gamma^0+\gamma^1$. Note that $\tilde{\beta}=\sum_{i=1}^d\alpha_i^0$ and $\tilde{\alpha}=\sum_{i=1}^d\alpha_i^1$. Let $\overline{\gamma}^0=\sum_{i=1}^d\gamma_i^0$ and $\overline{\gamma}^1=\sum_{i=1}^d\gamma_i^1$, so that $1-\tilde{P}\equiv_D\overline{\gamma}^0/\overline{\gamma}$ and $\tilde{P}\equiv_D\overline{\gamma}^1/\overline{\gamma}$. Couple these variables so that $1-\tilde{P}=\overline{\gamma}^0/\overline{\gamma}$ and $\tilde{P}=\overline{\gamma}^1/\overline{\gamma}$. We can now say,

$$\begin{aligned}
\mathbb{E}[\tilde{X}|X] &= \mathbb{E}[(1-\tilde{P})v_1+\tilde{P}v_d|X]=\mathbb{E}\left[\frac{v_1\overline{\gamma}^0}{\overline{\gamma}}+\frac{v_d\overline{\gamma}^1}{\overline{\gamma}}\Big|X\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{v_1\overline{\gamma}^0+v_d\overline{\gamma}^1}{\overline{\gamma}}\Big|\gamma,X\right]\Big|X\right]=\mathbb{E}\left[\frac{v_1\mathbb{E}[\overline{\gamma}^0|\gamma]+v_d\mathbb{E}[\overline{\gamma}^1|\gamma]}{\overline{\gamma}}\Big|X\right]\\
&= \mathbb{E}\left[\frac{v_1\sum_{i=1}^d\mathbb{E}[\gamma_i^0|\gamma_i]+v_d\sum_{i=1}^d xp[\gamma_i^1|\gamma_i]}{\overline{\gamma}}\Big|X\right]\\
&\overset{(a)}{=} \mathbb{E}\left[\frac{v_1\sum_{i=1}^d\gamma_i\alpha_i^0/\alpha_i+v_d\sum_{i=1}^d\gamma_i\alpha_i^1/\alpha_i}{\overline{\gamma}}\Big|X\right]\\
&= \mathbb{E}\left[\frac{v_1\sum_{i=1}^d\gamma_i(v_i-v_1)+v_d\sum_{i=1}^d\gamma_i(v_d-v_i)}{\overline{\gamma}(v_d-v_1)}\Big|X\right]\\
&= \mathbb{E}\left[\frac{\sum_{i=1}^d\gamma_iv_i}{\overline{\gamma}}\Big|X\right]=\mathbb{E}\left[\sum_{i=1}^d p_iv_i\Big|X\right]=X,
\end{aligned}$$

where (a) follows from elementary properties of Gamma distribution (Osband & Van Roy, 2017). Therefore, $\tilde{X}$ is a mean-preserving spread of X and so by definition of stochastic optimism $\tilde{X}\succcurlyeq_{\text{so}}X$. □

Next, consider any fixed $P_k(x_{kh})$ and let R_k and $P_k(x \neq x_{kh})$ vary in any arbitrary way to maximize the variation from transition $w_k^P(x_{kh}) = (P_k(x_{kh}) - \hat{P}(x_{kh}))^T V_{kh+1}^k$ through their effects on the future value $V_{kh+1}^k \in [0,H]^S$. We can then upper bound the deviation from transitions by the deviation under the worst possible $v \in [0,H]^S$.

$$w_h^P(x_{kh}) \leq \max_{R_k, P_k(x \neq x_{kh})} (P_k(x_{kh}) - \hat{P}_k(x_{kh}))^T V_{kh+1}^k \leq \max_{v \in [0,H]^S} (P_k(x_{kh}) - \hat{P}_k(x_{kh}))^T v. \tag{16}$$

We can then apply Lemma 4 to (16): for any possible value of $v \in [0,H]^S$ there is a matched Beta random variable that is stochastically optimistic for $w_h^P(x_{kh})$. This means that we can then apply Lemma 2 to show that there is a matched $X \sim \left(0, \frac{H^2}{\alpha^T \mathbb{1}}\right) \succcurlyeq_{\text{so}} w_h^P(x_{kh})$. To complete the proof of Lemma 3 we apply the Gaussian tail concentration Lemma 1.

B. Conjecture of $\tilde{O}(\sqrt{HSAT})$ bounds

The key remaining loose piece of our analysis concerns the summation $\sum_{h=1}^H w_h^P(x_{kh})$. Our current proof of Theorem 2 bounds each $w_h^P(x_{kh})$ independently. Each term is $\tilde{O}(\sqrt{\frac{H}{n_k(x_{kh})}})$ and we bound the resulting sum $\tilde{O}(H\sqrt{\frac{H}{n_k(x_{kh})}})$. However, this approach is very loose and pre-supposes that *each* timestep could be maximally bad during a single episode. To repeat our geometric intuition, we have assumed a worst-case hyper-rectangle over all timesteps H when the actual geometry should be an ellipse. We therefore suffer an additional term of $\tilde{O}(\sqrt{H})$ in exactly the style of Figure 3.

In fact, it is not even possible to sequentially get the "worst-case" transitions $O(H)$ at each and every timestep during an episode, since once your sample gets one such transition then there will be no more future value to deplete. Rather than just being independent per timestep, which would be enough for us to end up with an $\tilde{O}(\sqrt{H})$ saving, they actually have some kind of anti-correlation property through the law of total variance. A very similar observation is used by recent analyses in the sample complexity setting (Lattimore & Hutter, 2012) and also finite horizon MDPs (Dann & Brunskill, 2015). This seems to suggest that it should be possible to combine the insights of Lemma 3 with, for example, Lemma 4 of (Dann & Brunskill, 2015) to remove *both* the $\sqrt{S}$ *and* the $\sqrt{H}$ from our bounds to prove Conjecture 1.

We note that this informal argument would *not* apply Gaussian PSRL, since it generates w^P from some Gaussian posterior which does not satisfy the Bellman operators. Therefore, we should be able to find some evidence for this conjecture if we find domains where UCRL, Gaussian PSRL and PSRL all demonstrate their (unique) predicted scalings. We present some evidence of this effect in Section 5.3 and find that that our empirical results are consistent with this conjecture.

C. Estimation experiments

In this section we expand upon the simple examples given by Section 4.1 to a full decision problem with two actions. We define an MDP similar to Figures 1 and 2 but now with two actions. The first action is identical to Figure 1, but the second action modifies the transition probabilities to favor the rewarding states with probability $0.6/N$ and assigning only $0.4/N$ to the non-rewarding states.

We now investigate the *regret* of several learning algorithms which we adapt to this setting. These algorithms are based upon BEB (Kolter & Ng, 2009), BOLT (Araya et al., 2012), ϵ-greedy with $\epsilon = 0.1$, Gaussian PSRL (see Algorithm 3), Optimistic PSRL (which takes $K = 10$ samples and takes the maximum over sampled Q-values similar to BOSS (Asmuth et al., 2009)), PSRL (Strens, 2000), UCFH (Dann & Brunskill, 2015) and UCRL2 (Jaksch et al., 2010). We link to the full code for implementation in Appendix D.

We see that the loose estimates in OFU algorithms from Figures 4 and 5 lead to bad performance in a decision problem. This poor scaling with the number of successor states N occurs when *either* the rewards or the transition function is unknown. We note that in stochastic environments the PAC-Bayes algorithm BOLT, which relies upon optimistic fake prior data, can sometimes concentrate too quickly and so incur the maximum linear regret. In general, although BOLT is PAC-Bayes, it concentrates too fast to be PAC-MDP just like BEB (Kolter & Ng, 2009).

In Figure 12 we see a similar effect as we increase the episode length H. We note the second order UCFH modification improves upon UCRL2's miscalibration with H, as is reflected in their bounds (Dann & Brunskill, 2015). We note that both BEB and BOLT scale poorly with the horizon H.

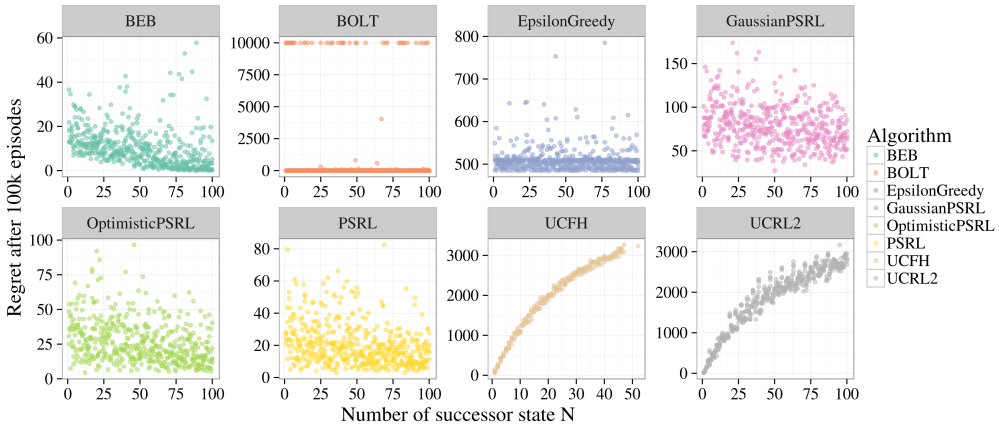

Figure 10. Known rewards R and unknown transitions P, similar to Figure 4.

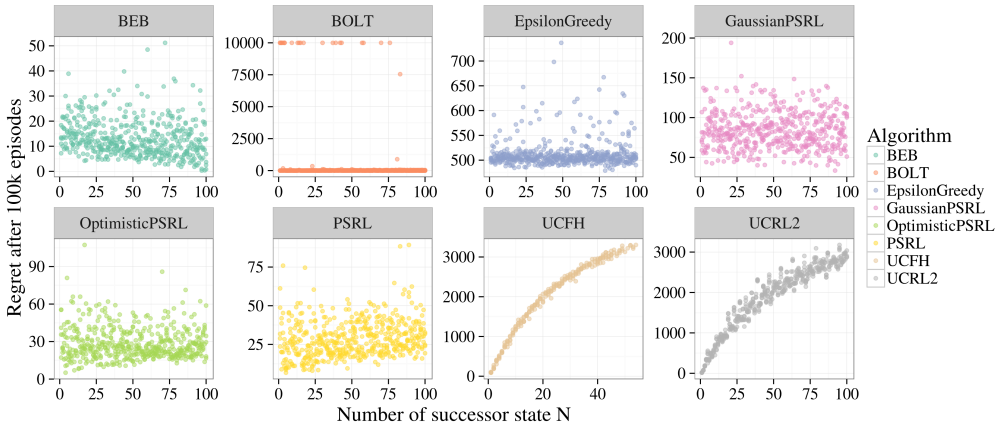

Figure 11. Unknown rewards R and known transitions P, similar to Figure 5.

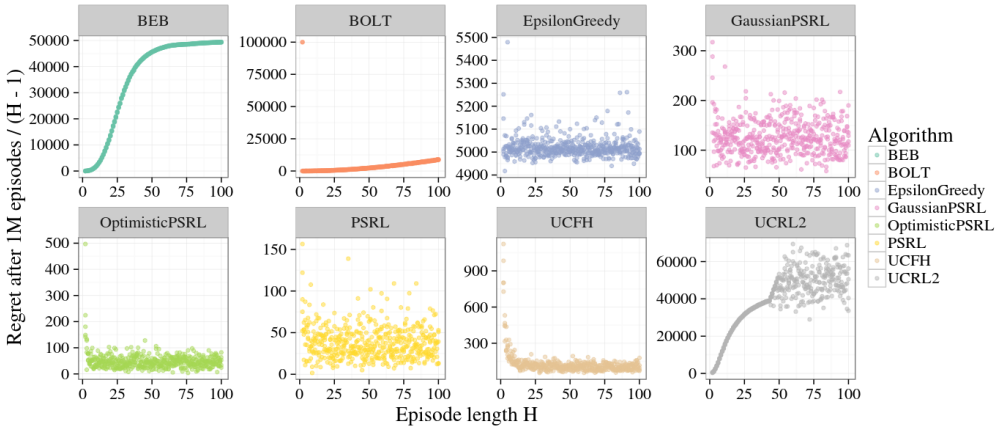

Figure 12. Unknown rewards R and transitions P, similar to Figure 6.

D. Chain experiments

All of the code and experiments used in this paper are available in full on github. As per the review request we have removed the link to this code, but instead include an anonymized excerpt of the some of the code in our submission file. We hope that researchers will find this simple codebase useful for quickly prototyping and experimenting in tabular reinforcement learning simulations.

In addition to the results already presented we also investigate the scaling of similar Bayesian learning algorithms BEB (Kolter & Ng, 2009) and BOLT (Araya et al., 2012). We see that neither algorithms scale as gracefully as PSRL, although BOLT comes close. However, as observed in Appendix C, BOLT can perform poorly in highly stochastic environments. BOLT also requires S-times more computational cost than PSRL or BEB. We include these algorithms in Figure 13.

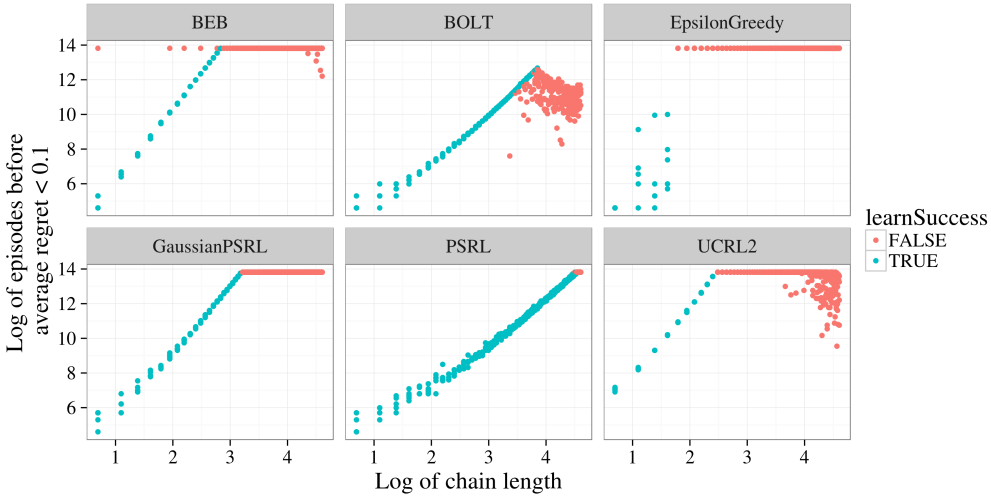

Figure 13. Scaling of more learning algorithms.

D.1. Rescaling confidence sets

It is well known that provably-efficient OFU algorithms can perform poorly in practice. In response to this observation, many practitioners suggest rescaling confidence sets to obtain better empirical performance (Szita & Szepesvári, 2010; Araya et al., 2012; Kolter & Ng, 2009). In Figure 14 we present the performance of several algorithms with confidence sets rescaled $\in \{0.01, 0.03, 0.1, 0.3, 1\}$. We can see that rescaling for tighter confidence sets can sometimes give better empirical performance. However, it does not change the fundamental scaling of the algorithm. Also, for aggressive scalings some seeds may not converge at all.

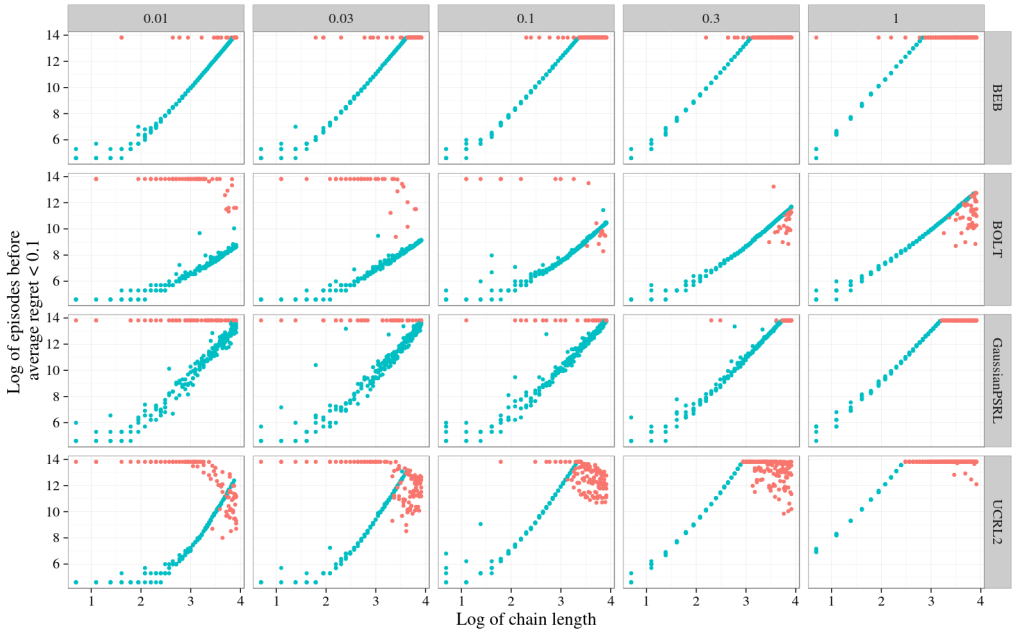

Figure 14. Rescaled proposed algorithms for more aggressive learning.

D.2. Prior sensitivities

We ran all of our Bayesian algorithms with uninformative independent priors for rewards and transitions. For rewards, we use $\overline{r}(s,a) \sim N(0,1)$ and updated as if the observed noise were Gaussian with precision $\tau = \frac{1}{\sigma^2} = 1$. For transitions, we use a uniform Dirichlet prior $P(s,a) \sim \text{Dirchlet}(\alpha)$. In Figures 15 and 16 we examine the performance of Gaussian PSRL and PSRL on a chain of length $N = 10$ as we vary τ and $\alpha = \alpha_0 \mathbb{1}$.

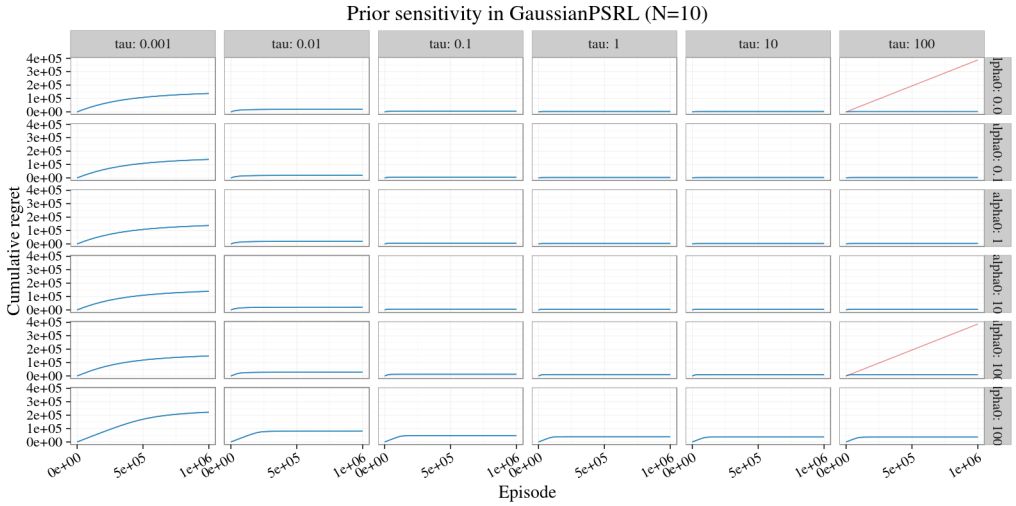

Figure 15. Prior sensitivity in Gaussian PSRL.

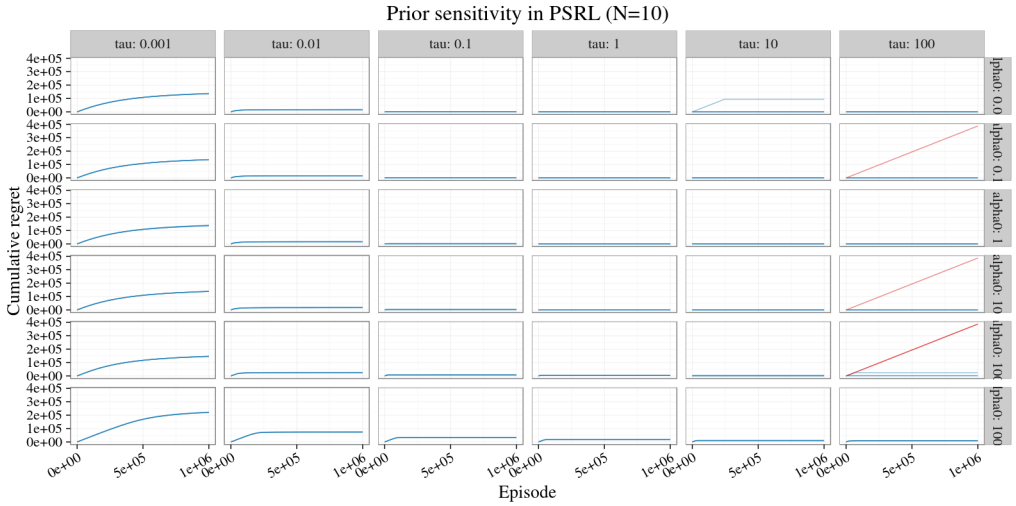

Figure 16. Prior sensitivity in PSRL.

We find that both of the algorithms are extremely robust over several orders of magnitude. Only large values of τ (which means that the agent updates it reward prior too quickly) caused problems for some seeds in this environment. Developing a more clear frequentist analysis of these Bayesian algorithms is a direction for important future research.

D.3. Optimistic posterior sampling

We compare our implementation of PSRL with a similar optimistic variant which samples $K \geq 1$ samples from the posterior and forms the optimistic Q-value over the envelope of sampled Q-values. This algorithm is sometimes called "optimistic posterior sampling" (Fonteneau et al., 2013). We experiment with this algorithm over several values of K but find that the resultant algorithm performs very similarly to PSRL, but at an increased computational cost. We display this effect over several magnitudes of K in Figures 17 and 18.

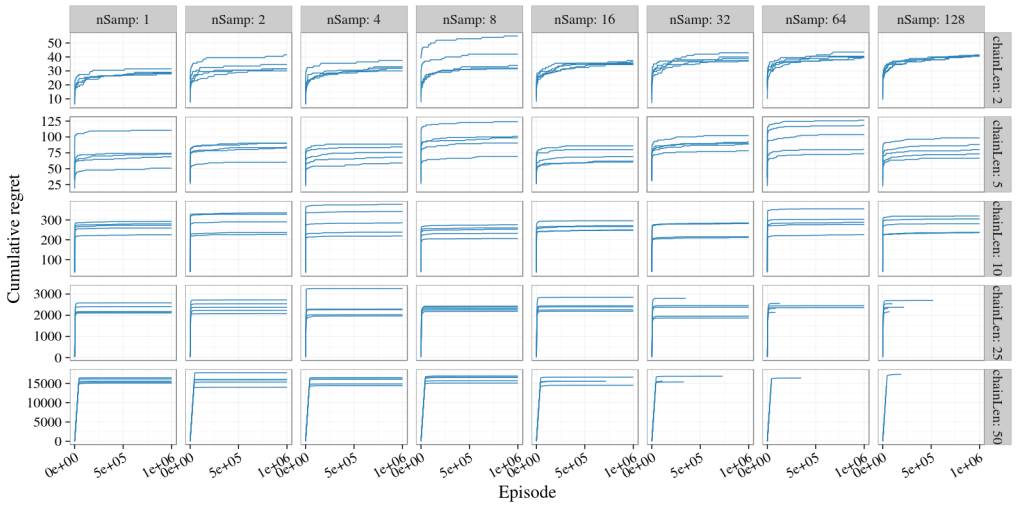

Figure 17. PSRL with multiple samples is almost indistinguishable.

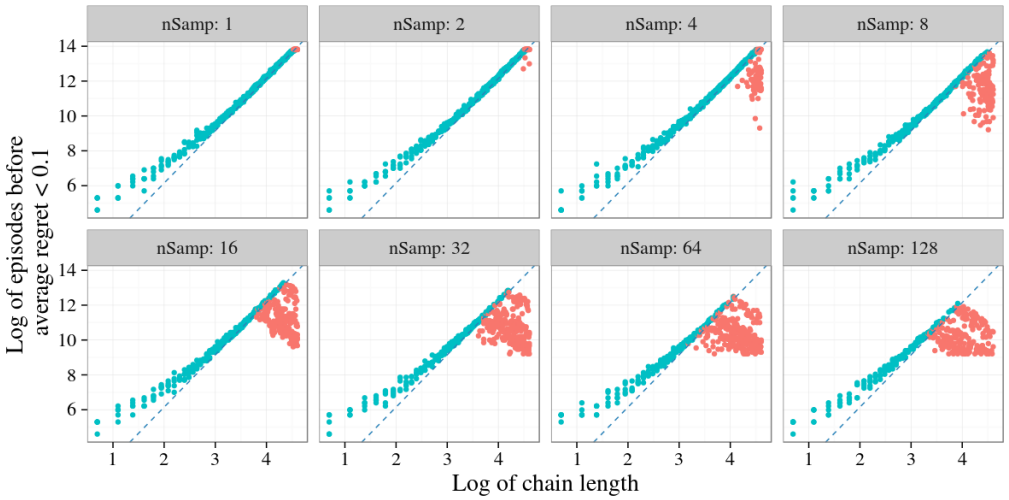

Figure 18. PSRL with multiple samples is almost indistinguishable.

This algorithm "Optimistic PSRL" is spiritually very similar to BOSS (Asmuth et al., 2009) and previous work had suggested that $K > 1$ could lead to improved performance. We believe that an important difference is that PSRL, unlike Thompson sampling, should not resample every timestep but previous implementations had compared to this faulty benchmark (Fonteneau et al., 2013).