# Thompson Sampling with a Mixture Prior

**Joey Hong**
UC Berkeley*

**Branislav Kveton**
Amazon*

**Manzil Zaheer**
Google DeepMind

**Mohammad Ghavamzadeh**
Google Research

**Craig Boutilier**
Google Research

## Abstract

We study Thompson sampling (TS) in online decision making, where the uncertain environment is sampled from a *mixture distribution*. This is relevant in multi-task learning, where a learning agent faces different classes of problems. We incorporate this structure in a natural way by initializing TS with a *mixture prior*, and call the resulting algorithm `MixTS`. To analyze `MixTS`, we develop a novel and general proof technique for analyzing the concentration of mixture distributions. We use it to prove Bayes regret bounds for `MixTS` in both linear bandits and finite-horizon reinforcement learning. Our bounds capture the structure of the prior, depend on the number of mixture components and their widths. We also demonstrate the empirical effectiveness of `MixTS` in synthetic and real-world experiments.

## 1 INTRODUCTION

*Thompson sampling (TS)* (Agrawal and Goyal, 2012) is arguably the most popular and practical class of exploration algorithms for *stochastic bandits* (Lattimore and Szepesvári, 2019; Agrawal and Goyal, 2012) and *reinforcement learning (RL)* (Barto and Sutton, 2018; Osband et al., 2013). However, in both settings, TS is almost exclusively applied with a *unimodal prior* over model parameters (Agrawal and Goyal, 2012, 2013; Osband et al., 2013). This is extremely limiting in a variety of settings, for instance, in a multi-task setting where a learning agent faces one of $L$ classes of bandit problems, each with a different distribution of model parameters. If this prior knowledge was expressed by a single unimodal distribution, it would generally be "wide" (hence uninformative), which can dramatically slow convergence of TS.

In this work, we incorporate *mixture models* into TS for both stochastic bandits and RL. The idea behind mixture models is using latent variables to make a model more expressive (Bishop, 2006). In supervised learning, a more expressive model can better capture a complex population of sub-populations with similar features. For instance, *Gaussian mixture models* (GMMs) (Macqueen, 1967) are commonly used to cluster features in financial markets (Wang, 2001) and to identify classes of images (Bishop, 2006). *Topic models* (Blei et al., 2003), which are mixtures of categorical distributions, are often used to analyze text data. Similarly, in online learning, algorithms can be more expressive by conditioning on a latent state (Jordan and Jacobs, 1994). In multi-task learning, where an agent faces a collection of tasks related through latent structure, we believe that TS can be substantially improved by using a more expressive prior.

We study TS with a *mixture prior*, which is a joint probability distribution over an unobserved discrete latent state and model parameters. It is unclear *a priori* if efficient algorithms exist for this problem class. From the computational perspective, the posterior distribution may not have a closed form; and thus may be hard to update efficiently. This is one reason why existing TS implementations use simple priors. Apart from computational issues, we might hope to exploit the problem structure to derive tighter regret bounds. The challenge is that the learning agent never observes the latent state. We address both challenges.

We make the following contributions. First, we propose a general algorithm, *mixture Thompson sampling* (`MixTS`), for a mixture prior with $L$ discrete latent

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. 

*The work was done while at Google Research.

states. MixTS first samples a latent state from its posterior and then samples model parameters conditioned on that state. By explicitly modeling the latent state, the posterior can be efficiently maintained for common reward distributions, such as Bernoulli and Gaussian, using conjugacy. Second, we bound the $n$-round Bayes regret of MixTS using a novel general analysis technique that accounts for jointly learning the model parameters and identifying the latent state; without ever observing it. We apply our technique in two settings: contextual linear bandit and finite-horizon RL. Finally, we evaluate MixTS empirically in synthetic bandit and RL tasks, and in a task based on image classification using the CIFAR-100 dataset (Krizhevsky, 2009).

The main theoretical contribution of this work are the first Bayes regret bounds for TS with a mixture prior that are (i) sublinear in the number of rounds and (ii) depend on how informative the prior is. Specifically, the bounds depend on the structure of the prior, the number of mixture components and their width. When the prior is unimodal, $L = 1$, our bounds match the regret bounds of classical TS (Agrawal and Goyal, 2012, 2013). On the other hand, when the mixture components have low width, the regret is determined by the cost of identifying the correct latent state (Hong et al., 2020). Hong et al. (2020) studied the same algorithm as MixTS in bandits, but proved a linear regret bound for non-zero mixture component widths. In RL, we are the first to consider and analyze a mixture prior.

## 2 SETTING

We consider an online decision-making problem where a learning agent interacts with an unknown environment sequentially over $n$ *rounds*. We start with a multi-armed bandit setting and extend it to RL in Section 5. We adopt the following notation. Random variables are capitalized. The $i$-th entry of vector $v$ is $v_i$; if a vector $v_i$ is already indexed, then we denote its $j$-th entry by $(v_i)_j$. We use $\tilde{\mathcal{O}}$ for the big O notation up to logarithmic factors.

Our setting is defined as follows. In round $t \in [n]$, the agent takes an *action* $A_t$ from an *action set* $\mathcal{A}_t$ and observes *reward* $Y_t \in \mathbb{R}$. The reward $Y_t$ is drawn i.i.d. from *reward distribution* $P(\cdot \mid A_t; \theta)$. The distribution depends on the taken action $A_t$ and *model parameters* $\theta \in \Theta$, where $\Theta$ is a set of feasible model parameters. We denote by $\mu_\theta(a) = \mathbb{E}_{Y \sim P(\cdot|a;\theta)}[Y]$ the mean reward of action $a$ under model $\theta$, and assume that all rewards are $\sigma^2$-sub-Gaussian. We subscript the action set by $t$ as $\mathcal{A}_t$. This allows us to have changing action sets, which provides additional flexibility. Specifically, in contextual bandits, the context $X_t$ in round $t$ may influence which actions are possible, a dependence captured in $\mathcal{A}_t$.

We denote by $\theta_*$ the true model parameters. In this work, we assume that $\theta_*$ is sampled from a *mixture prior* $P_0$. The mixture prior is represented using a finite set of *latent states* $\mathcal{S}$, where $|\mathcal{S}| = L$. Each latent state corresponds to a separate "hypothesis" for the parameter distribution. The model parameters $\theta_*$ are sampled as follows. First the true latent state is sampled as $S_* \sim P_0$ from the *latent state prior*, then the model parameters are sampled as $\theta_* \sim P_0(\cdot \mid S_*)$ from the *model parameter prior*. Formally, the distribution of $\theta_*$ is $\mathbb{P}(\theta_* = \theta) = \sum_{s \in \mathcal{S}} P_0(\theta \mid s) P_0(s)$.

In multi-armed bandits, a typical goal is to maximize the expected $n$-round reward, or equivalently minimize the *expected $n$-round regret*

$$\mathcal{R}(n; \theta_*) = \mathbb{E}\left[\sum_{t=1}^{n} \mu_*(A_{t,*}) - \mu_*(A_t) \mid \theta_*\right],$$

where $\mu_*(a) = \mu_{\theta_*}(a)$ is the true mean reward of action $a$, $A_{t,*} = \max_{a \in \mathcal{A}_t} \mu_*(a)$ is the optimal action in round $t$, and the expectation is taken over both the randomness in the bandit algorithm and environment. Note that $\theta_*$ is fixed in $\mathcal{R}(n, \theta_*)$. In this work, we focus on an average performance over *multiple problems*, each corresponding to different model parameters sampled from the prior. This is to capture the structure of the stochastic generative process in our analysis. By taking an expectation over $S_*$ and $\theta_*$, we obtain the *$n$-round Bayes regret* (Russo and Van Roy, 2013) $\mathcal{BR}(n) = \mathbb{E}[\mathcal{R}(n; \theta_*)]$.

## 3 ALGORITHM

Thompson sampling (Agrawal and Goyal, 2012; Russo and Van Roy, 2013) is an algorithm that takes actions proportionally to being optimal under the posterior. This is usually implemented by first sampling model parameters $\theta_t$ from the posterior, then taking action $A_t = \arg\max_{a \in \mathcal{A}_t} \mu_{\theta_t}(a)$ that maximizes the mean reward under $\theta_t$. The posterior captures agent's uncertainty over the true model parameters $\theta_*$ conditioned on history. We denote the *observation history* up to round $t$ by $H_t = (A_1, Y_1, \ldots, A_{t-1}, Y_{t-1})$, and denote the respective conditional probability and expectation by $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid H_t)$ and $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid H_t]$.

Now we describe how Thompson sampling with a mixture prior works. We first note that the posterior over model parameters at round $t$ can be obtained by marginalizing over the latent state as $\mathbb{P}_t(\theta_* = \theta) = \sum_{s \in \mathcal{S}} \mathbb{P}_t(\theta_* = \theta \mid S_* = s) \mathbb{P}_t(S_* = s)$. Because of this structure, explicit modeling of the latent state allows for tractable sampling from and updates to the posterior. We denote the posterior by $P_t$, where $P_t(s) = \mathbb{P}_t(S_* = s)$ and $P_t(\theta \mid s) = \mathbb{P}_t(\theta_* = \theta \mid S_* = s)$. Sampling $\theta_*$ from the posterior is straightforward: first a

**Algorithm 1** TS with a mixture prior (`MixTS`)

**Input:** Latent state prior $P_0$
  model parameters priors $\{P_0(\cdot \mid s)\}_{s \in \mathcal{S}}$
Initialize $P_1 \leftarrow P_0$
**for** $t \leftarrow 1, \ldots, n$ **do**
  Sample $S_t \sim P_t$ and $\theta_t \sim P_t(\cdot \mid S_t)$
  Select $A_t \leftarrow \arg\max_{a \in \mathcal{A}_t} \mu_{\theta_t}(a)$.
  Observe $Y_t$ and update
    $P_{t+1}(\theta \mid s) \propto P_t(\theta \mid s) P(Y_t \mid A_t; \theta)\,,\ \forall s \in \mathcal{S}$
    $P_{t+1}(s) \propto P_0(s) \int_\theta P_{t+1}(\theta \mid s)\, d\theta$

latent state $S_t \sim P_t$ is sampled, then $\theta_t \sim P_t(\cdot \mid S_t)$ is sampled conditioned on $S_t$. We show that each component of the posterior can be computed tractably.

The key insight is that each *model parameter posterior* has form $P_t(\theta \mid s) \propto P_0(\theta \mid s) \prod_{\ell=1}^{t-1} P(Y_t \mid A_t; \theta)$, and thus has a closed form when $P_0(\cdot \mid s)$ is conjugate to the reward distribution. This holds in many settings, such as Bernoulli rewards with a beta prior, and Gaussian rewards with a Gaussian prior. Moreover, we note that the latent state posterior can be written as $P_t(s) \propto P_0(s) \int_\theta P_t(\theta \mid s)\, d\theta$. The integral is the posterior predictive probability and can be computed efficiently when $P_0(\cdot \mid s)$ is conjugate to the reward distribution. The normalizing constant $\mathbb{P}(H_t)$ is the same for all latent states. Since $\mathcal{S}$ is finite, we normalize $P_t(s)$ by dividing it with $\sum_{s=1}^{L} P_t(s)$.

Based on the above, TS with a mixture prior can be implemented efficiently for many problems of interest. The resulting algorithm, `MixTS` (Algorithm 1), uses incremental posterior updates. In the bandit setting, our algorithm is an instance of `mmTS` (Hong et al., 2020) for latent bandits. Since `MixTS` has a mixture prior, all model parameters live in the same parameter space, a key difference from Hong et al. (2020) that allows us to analyze the concentration of the mixture posterior. In addition, we extend `MixTS` to RL in Section 5.

# 4 BAYES REGRET ANALYSIS

In this section, we prove a Bayes regret bound with a mixture prior. In Section 4.1, we provide a general analysis outline for `MixTS`. We specialize it to contextual linear bandits in Section 4.2 and extend it to RL in Section 5.

Bandit algorithms with latent variables are rare, and often lack a regret bound. The key step in our analysis is a novel construction of confidence intervals around latent variables. This is challenging because the latent variables are unobserved. Our analysis outline can be applied to any model, simply by specifying the confidence intervals. This shows the modularity and generality of our approach.

## 4.1 General Analysis Outline

Recall that $S_*$ and $\theta_*$ are the true latent state and model parameters, and let $\mu_*(a) = \mu_{\theta_*}(a)$. To simplify the sketch, we assume that $\mu_*(a) \in [0, 1]$; but Theorem 1 does not assume this.

Let $\bar{\mu}_t(a, s) = \mathbb{E}_{\theta \sim P_t(\cdot \mid s)}[\mu_\theta(a)]$ be the posterior mean reward of action $a$ under latent state $s$, and $\sigma_t(a, s)$ be a high-probability confidence width for the model parameter posteriors $P_t(\cdot \mid s)$, that is $\mathbb{P}_t\left(|\mu_*(a) - \bar{\mu}_t(a, s)| \geq \sigma_t(a, s)\right) \leq 1/n$. At a high level, our Bayes regret bounds include two terms. The first is due to concentration of the model parameter posteriors, and is bounded by the sum of confidence widths $\sum_{t=1}^{n} \sigma_t(A_t, S_t)$. The second captures the identification of the latent state, and scales with $\sqrt{Ln}$.

Let $A_{t,*} = \max_{a \in \mathcal{A}_t} \mu_*(a)$ be the optimal action in round $t$. From Russo and Van Roy (2013), we can write the Bayes regret as

$$\mathcal{BR}(n) = \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_t\left[\mu_*(A_{t,*}) - \bar{\mu}_t(A_{t,*}, S_*)\right]\right] + \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_t\left[\bar{\mu}_t(A_t, S_t) - \mu_*(A_t)\right]\right], \quad (1)$$

where we use that $\bar{\mu}_t$ is a deterministic function of history $H_t$, and that $A_t, S_t$ and $A_{t,*}, S_*$ are i.i.d. given $H_t$. To bound the Bayes regret, we can bound each term individually as follows.

**Step 1.** Bound the first term of (1). For round $t$, let event

$$E_t = \{\forall a \in \mathcal{A} : |\mu_*(a) - \bar{\mu}_t(a, S_*)| \leq \sigma_t(a, S_*)\}$$

denote that the true mean is close to the posterior mean. Then

$$\mathbb{E}_t\left[\mu_*(A_{t,*}) - \bar{\mu}_t(A_{t,*}, S_*,)\right] \leq \mathbb{E}_t\left[(\mu_*(A_{t,*}) - \bar{\mu}_t(A_{t,*}, S_*))\mathbb{1}\left\{\bar{E}_t\right\}\right] + \mathbb{E}_t\left[\sigma_t(A_{t,*}, S_*)\right] \quad (2)$$

where $(\mu_*(A_{t,*}) - \bar{\mu}_t(A_{t,*}, S_*))\mathbb{1}\{E_t\} \leq \sigma_t(A_{t,*}, S_*)$ is by definition of $E_t$. The first term of (2) can be bounded using the fact that event $\bar{E}_t$ is unlikely conditioned on $H_t$. The second term can be rewritten as $\mathbb{E}_t\left[\sigma_t(A_{t,*}, S_*)\right] = \mathbb{E}_t\left[\sigma_t(A_t, S_t)\right]$, using that $A_t, S_t$ and $A_{t,*}, S_*$ are i.i.d. conditioned on $H_t$. Finally, we sum over all rounds $t \in [n]$.

**Step 2.** We want to bound the second term of (1). To do so, we first need to define *confidence sets* over latent states. Formally, for each round $t$, we construct $C_t$ such that $S_* \in C_t$ holds with a high probability. Since the latent state is unobserved, we use a frequentist construction with a proxy statistic for how well the model parameter posterior of each latent state predicts the

rewards. Let $N_t(s) = \sum_{\ell=1}^{t-1} \mathbb{1}\{S_\ell = s\}$ be the number of times $s$ was sampled from posterior up to round $t$, and

$$G_t(s) = \sum_{\ell=1}^{t-1} \mathbb{1}\{S_\ell = s\} \left(\bar{\mu}_\ell(A_\ell, s) - \eta\sigma_\ell(A_\ell, s) - Y_\ell\right)$$

be the total reward "excess" with respect to the posterior mean, where $\eta \in \mathbb{R}, \eta > 0$ is a scaling factor. Let $C_t = \{s \in \mathcal{S} : G_t(s) \leq \varepsilon\}$ be the set of latent states with at most $\varepsilon$ excess. We want to prove that $S_*$ lies in $C_t$ in round $t$ with a high probability,

$$\mathbb{P}\left(\bigcup_{t=1}^{n} \{S_* \notin C_t\}\right) \leq \sum_{t=1}^{n} \mathbb{P}\left(S_* \notin C_t\right) = \mathcal{O}(1)\,.$$

The key idea in the proof is that each $\bar{\mu}_t(A_\ell, S_*) - \eta\sigma_t(A_\ell, S_*) - Y_\ell < 0$ holds with a high probability conditioned on any history $H_t$, since we subtract the reward from its lower confidence bound. Since $\mu_*(A_\ell)$ is unknown, we substitute it with reward $Y_\ell$. We set $\varepsilon = \mathcal{O}(\sqrt{N_t(s)\log n})$ in $C_t$ to correct for reward noise. Hong et al. (2020) consider a similar construction, but used prior means and widths. We achieve better regret bounds by using the posterior.

**Step 3.** Now, we are ready to bound the second term of (1). Since regret at any round is trivially bounded by 1, we have

$$\mathbb{E}\left[\sum_{t=1}^{n} \bar{\mu}_t(A_t, S_t) - \mu_*(A_t)\right] \leq \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{S_t \notin C_t\}\right] + \mathbb{E}\left[\sum_{t=1}^{n} \left(\bar{\mu}_t(A_t, S_t) - \mu_*(A_t)\right) \mathbb{1}\{S_t \in C_t\}\right].$$

Note that the first term can be bounded as

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{S_t \notin C_t\}\right] &= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{P}_t\left(S_t \notin C_t\right)\right] \\ &= \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{P}_t\left(S_* \notin C_t\right)\right] \\ &= \sum_{t=1}^{n} \mathbb{P}\left(S_* \notin C_t\right) = \mathcal{O}(1)\,,\end{aligned}$$

where we use that $S_t$ and $S_*$ are i.i.d. conditioned on $H_t$ for the first equality, and the bound derived in Step 2 for the second. Finally, we have

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{n} \left(\bar{\mu}_t(A_t, S_t) - \mu_*(A_t)\right) \mathbb{1}\{S_t \in C_t\}\right] \qquad (3)\\ &\leq \eta\mathbb{E}\left[\sum_{t=1}^{n} \sigma_t(A_t, S_t)\right] + \\ &\mathbb{E}\left[\sum_{t=1}^{n} \left(\bar{\mu}_t(A_t, S_t) - \eta\sigma_t(A_t, S_t) - Y_t\right) \mathbb{1}\{S_t \in C_t\}\right],\end{aligned}$$

where we use that $\mathbb{E}_t\left[Y_t \mid A_t, \theta_*\right] = \mathbb{E}_t\left[\mu_*(A_t)\right]$. The first term of (3) is a sum of confidence widths, which decrease over time as the posterior concentrates. The second term of (3) can be bounded by the sum of the excesss $\sum_{s\in\mathcal{S}} G_{n+1}(s)$, which is bounded by $\mathcal{O}(\sqrt{Ln\log n} + L)$ after we trivially bound the regret in the last round where each latent state is sampled. This is because in the last round $t$ where $S_t = s$, it must be true that $s \in C_t$, and thus $G_t(s)$ is bounded.

### 4.2 Linear Bandits

The above general analysis technique can be applied in various settings. Here we specialize it to a linear bandit with $d$ dimensions. In each round $t \in [n]$, a learning agent has a potentially changing action set $\mathcal{A}_t \subseteq \mathbb{R}^d$ and takes action $A_t \in \mathcal{A}_t$. The agent observes reward $Y_t = A_t^\top \theta_* + \eta_t$, where $\theta_* \in \mathbb{R}^d$ is the unknown model parameter vector and $\eta_t \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise. We assume that $\|a\|_2 \leq \kappa$ for all rounds $t$ and $a \in \mathcal{A}_t$.

The prior is a mixture with $L$ components, indexed by latent states $s \in \mathcal{S}$. For each $s$, the model parameter prior is a Gaussian $P_0(\cdot \mid s) = \mathcal{N}(\cdot; \theta_{0,s}, \Sigma_{0,s})$, and we assume that $\theta_{0,s}$ is bounded as $\|\theta_{0,s}\|_2 \leq 1$. This is a weaker assumption than in prior works, which typically assume that $\|\theta_*\|_2$ is bounded (Abbasi-yadkori et al., 2011; Russo and Van Roy, 2013). In round $t$, `MixTS` samples $S_t \sim P_t$ and then $\theta_t \sim \mathcal{N}(\bar{\theta}_{t,S_t}, \Sigma_{t,S_t})$. Here $\bar{\theta}_{t,s}$ and $\Sigma_{t,s}$ are the posterior mean model parameter and its covariance, respectively, under latent state $s$ and are defined as

$$\begin{aligned}\Sigma_{t,s} &= (\Sigma_{0,s}^{-1} + \sigma^{-2} V_t)^{-1}\,, \\ \bar{\theta}_{t,s} &= \Sigma_{t,s}(\Sigma_{0,s}^{-1}\theta_{0,s} + \sigma^{-2} B_t)\,, \qquad (4)\end{aligned}$$

where $V_t = \sum_{\ell=1}^{t-1} A_\ell A_\ell^\top$ and $B_t = \sum_{\ell=1}^{t-1} A_\ell Y_\ell$. The posterior mean reward of action $a$ and its confidence width are given by

$$\bar{\mu}_t(a, s) = a^\top \bar{\theta}_{s,t},\ \ \sigma_t(a, s) = \sqrt{2d\log(dn)}\,\|a\|_{\Sigma_{t,s}}\,.$$

We can bound the Bayes regret of `MixTS` in this setting using the technique in Section 4.1. The proof is in Appendix A and we state the bound below.

**Theorem 1.** *Let $\lambda_{0,\max} = \max_{s\in\mathcal{S}} \lambda_{\max}(\Sigma_{0,s})$, where $\lambda_{\max}(\Sigma_{0,s})$ is the maximum eigenvalue of $\Sigma_{0,s}$ for latent state $s$. Let $\max_{a\in\mathcal{A}_t} \|a\|_2 \leq \kappa$ hold in all rounds $t \in [n]$. Then the $n$-round Bayes regret of* `MixTS` *is bounded as*

$$\mathcal{BR}(n) \leq 6\sigma d\sqrt{c_1 n\log(dn)} + 2\sigma\sqrt{Ln\log n} + c_2\,, \quad (5)$$

*where*

$$c_1 = \left(1 + \frac{\kappa^2\lambda_{0,\max}}{\sigma^2}\right)\log\left(1 + \frac{\kappa^2\lambda_{0,\max}\ n}{\sigma^2 d}\right),$$

*and $c_2$ is poly-logarithmic in $n$.*

### 4.3 Discussion

The bound has two main components: the regret for learning model parameters (Term 1) under the assumption that the latent state is known, and the regret for identifying the latent state (Term 2). Term 1 is $\tilde{\mathcal{O}}(d\sqrt{c_1 n})$ and is of the same order as in linear TS (Russo and Van Roy, 2013). The key difference is a prior-dependent constant $c_1$. Through $c_1$, Term 1 is linear in the maximum component width of the mixture prior $\sqrt{\lambda_{0,\max}}$. Term 2 is $\tilde{\mathcal{O}}(\sqrt{Ln})$ and is of the same order as identifying the true latent state among $L$ known models (Hong et al., 2020).

Our bound does not depend on the latent state prior $P_0$. This is a shortcoming of our analysis, which constructs worst-case confidence sets for latent states, and is frequentist in this respect. We defer refinements of the analysis to future work. Another shortcoming is that we do not provide a matching lower bound. Although a Bayes regret lower bound exists for $K$-armed bandits (Lai, 1987), it is unclear how to apply it to structured problems. Seminal works on Bayes regret minimization (Russo and Van Roy, 2013, 2016) also only derive upper bounds. We view deriving lower bounds as another avenue for future work.

Our analysis improves upon that of Hong et al. (2020) by analyzing concentration of the model parameter posteriors. We attain $\tilde{\mathcal{O}}(d\sqrt{c_1 n} + \sqrt{Ln})$ regret that is fully sublinear in $n$. In contrast, Hong et al. (2020) have a regret bound $\tilde{\mathcal{O}}(c'n + \sqrt{Ln})$, where $c'$ is a constant proportional to the maximum component width $\sqrt{\lambda_{0,\max}}$. This is because their analysis is agnostic to posterior improvements and treats prior uncertainty as a penalty, resulting in a linear regret bound.

Another natural comparison is to TS without the mixture prior. Since Bayes regret bounds are proved under the assumption of a correct prior, there are no other comparable Bayes regret bounds. However, we can compare to frequentist worst-case regret bounds, which hold even when the prior is misspecified. A state-of-the-art regret bound for `LinTS` is $\tilde{\mathcal{O}}(d^{3/2}\sqrt{n})$ (Abeille and Lazaric, 2017). In contrast, our bound is $\tilde{\mathcal{O}}(d\sqrt{c_1 n} + \sqrt{Ln})$, where $c_1$ scales with the maximum component width of the mixture prior $\sqrt{\lambda_{0,\max}}$ and $L$ denotes the number of latent states. With a sufficiently informative prior, $c_1 < d$; and with a small number of mixture components, $\sqrt{L} < d^{3/2}$; our bound improves over frequentist regret bounds for `LinTS`.

## 5 FINITE-HORIZON RL

Next we extend our results to *reinforcement learning (RL)* (Barto and Sutton, 2018) in *finite-horizon Markov decision processes (MDPs)* (Bellman, 1957). First, we formalize RL with a mixture prior. Then, in Section 5.1, we extend the general analysis outline from Section 4.1. Finally, in Section 5.2, we apply the outline to derive a Bayes regret bound for `MixTS` in a finite-horizon tabular MDP.

We have $n$ *episodes* indexed by $t \in [n]$. In each episode, a learning agent interacts with an MDP for $h$ *steps*. We refer to $h$ as the *horizon*. We denote a finite-horizon MDP by $M = (\mathcal{X}, \mathcal{A}, R, T, h, \rho)$, where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, $R_M(x, a) \in [0, 1]$ is the mean reward when selecting action $a$ in state $x$, $T_M(x, a, x') = \mathbb{P}\left(X_{i+1} = x' \mid X_i = x, A_i = a; M\right)$ is the probability of transitioning to state $x'$ if action $a$ is taken at state $x$, $h$ is the horizon, and $\rho$ the initial state distribution. We consider the special case of *tabular MDPs*, where both $\mathcal{X}$ and $\mathcal{A}$ are finite sets. As a shorthand, let $T_M(x, a) = (T_M(x, a, x'))_{x' \in \mathcal{X}}$ be a vector for all transitions.

A policy $\pi = (\pi^i)_{i=1}^h$ is a vector, one per step, where each $\pi^i : \mathcal{X} \to \mathcal{A}$ maps states to actions. We define the value of policy $\pi$ in MDP $M$ as $V_M(\pi) = \mathbb{E}\left[\sum_{i=1}^h R_M(X_i, A_i) \mid M, \pi\right]$, where $X_1 \sim \rho$, $A_i = \pi^i(X_i)$, and $X_{i+1} \sim \mathrm{Cat}(\cdot \mid T_M(X_i, A_i))$. The value is the expected total reward of acting under $\pi$ in $M$.

Let $M_*$ be the true MDP and $\pi_*$ be the optimal policy $\pi_* = \arg\max_\pi V_{M_*}(\pi)$ (Burnetas and Katehakis, 1997). We assume that $M_*$ is sampled hierarchically from a mixture prior $P_0$: first a latent state $S_* \sim P_0$ is sampled, then the MDP $M_* \sim P_0(\cdot \mid S_*)$. This generalizes prior work on TS in RL (Osband et al., 2013; Agrawal and Jia, 2017), where the mixture prior is not considered. Recently, Ayoub et al. (2020) studied MDPs whose mean rewards and transition probabilities are linear mixtures, but assume the mean rewards and probabilities per component are known. As a shorthand, we subscript by $*$ to denote statistics related to the true MDP $M_*$, such as $V_* = V_{M_*}$, and equivalently for $R_*$ and $T_*$. The Bayes regret of an algorithm over $n$ episodes is given by $\mathcal{BR}(n) = \mathbb{E}\left[\sum_{t=1}^n V_*(\pi_*) - V_*(\pi_t)\right]$, where $\pi_t$ is the policy chosen by the algorithm in episode $t$, and the randomness is over MDP $M_*$, policies selected by the learning agent, and observations. The history is given by $H_t = ((X_{\ell,i}, A_{\ell,i}, R_{\ell,i}))_{i \in [h],\, \ell \in [t-1]}$, where $X_{\ell,i}, A_{\ell,i}, R_{\ell,i}$ are the state, action and reward for step $i$ of episode $\ell$. The reward of an episode is $Y_t = \sum_{i=1}^h R_{t,i}$.

### 5.1 General Analysis Outline

In finite-horizon RL, `MixTS` operates as Algorithm 1, but with MDP $M_t$ instead of parameters $\theta_t$ and policy $\pi_t$ instead of action $A_t$. That is, `MixTS` in episode $t$ first samples latent state $S_t \sim P_t$, then the MDP conditioned on the sampled latent state $M_t \sim P_t(\cdot \mid S_t)$.

Finally, the chosen policy in episode $t$ maximizes the value $\pi_t = \arg\max_\pi V_{M_t}(\pi)$. This algorithm is a generalization of PSRL (Osband et al., 2013), where a mixture prior is used. While bandit analyses can be often adapted to RL, we make a notable deviation. Prior works construct confidence intervals for each state of an MDP (Osband et al., 2013; Lu and Van Roy, 2019). This cannot be done with latent variables, which are shared by all states. Therefore, we construct the intervals over entire MDP trajectories.

For episode $t$, let $\overline{V}_t(\pi, s) = \mathbb{E}_{M\sim P_t(\cdot|s)}[V_M(\pi)]$ be the expected value of policy $\pi$ conditioned on $s$ and $H_t$. We have the following Bayes regret decomposition,

$$\mathcal{BR}(n) = \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_t\left[V_*(\pi_*) - \overline{V}_t(\pi_*, S_*)\right]\right] + \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_t\left[\overline{V}_t(\pi_t, S_t) - V_*(\pi_t)\right]\right], \tag{6}$$

where we use that $S_t, \pi_t$ are distributed identically to $S_*, \pi_*$ conditioned on $H_t$.

The proof sketch is similar to the one in Section 4.1, but differs in two notable aspects. We list the main differences and defer the full sketch to Appendix B. First, the expected value of a policy under a latent state $\overline{V}_t(\pi, s)$ is used in place of the mean reward $\bar{\mu}_t(a, s)$. Second, in order to construct a confidence interval around $\overline{V}_t(\pi, s)$, we use the sum of confidence widths over steps of a trajectory. Specifically, for any policy $\pi$, we have with high probability,

$$V_{M_t}(\pi) - \overline{V}_t(\pi, s) = \mathbb{E}_{M\sim P_t(\cdot|s)}[V_{M_t}(\pi) - V_M(\pi)] \le \mathbb{E}_t\left[h\sum_{i=1}^{h} c_t(X_{t,i}, A_{t,i}, s) + \phi_t(X_{t,i}, A_{t,i}, s)\right],$$

where we use the value difference lemma (Osband et al., 2013). Here, we define a high-probability confidence intervals around the mean reward and transition probabilities, $c_t(x, a, s)$ and $\phi_t(x, a, s)$, respectively, for all state-action pairs $x, a$. For $\bar{r}_t(x, a, s) = \mathbb{E}_{M\sim P_t(\cdot|s)}[R_M(x, a)]$ as the posterior mean reward, we have $\mathbb{P}_t\left(|R_M(x, a) - \bar{r}_t(x, a, s)| \ge c_t(x, a, s)\right) \le 1/n$. Similarly, for $\bar{p}_t(x, a, x', s) = \mathbb{E}_{M\sim P_t(\cdot|s)}[T_M(x, a, x')]$ as the posterior mean transition probability to state $x'$, and $\bar{p}_t(x, a, s)$ as a vector of such probabilities over all states $x' \in \mathcal{X}$, we have $\mathbb{P}_t\left(\|T_M(x, a) - \bar{p}_t(x, a, s)\|_1 \ge \phi_t(x, a, s)\right) \le 1/n$. The sum over $c_t(X_{t,i}, A_{t,i}, s)$ and $\phi_t(X_{t,i}, A_{t,i}, s)$ is used in place of $\sigma_t(A_t, S_t)$.

### 5.2 Finite-Horizon Tabular MDPs

We consider finite-horizon tabular MDPs $M$ with Bernoulli rewards. In particular, for step $i$ of episode $t$, reward $R_{t,i}$ is sampled from a Bernoulli with mean $R_M(X_{t,i}, A_{t,i})$.

Recall that MDP $M = (\mathcal{X}, \mathcal{A}, R, T, h, \rho)$ has both mean rewards and transition probabilities. Let $R_M = (R_M(x, a))_{x,a}$ and $T_M = (T_M(x, a))_{x,a}$ be their respective concatenations across all state-action pairs. For true MDP $M_*$, which is unknown to the learning agent, let $R_*, T_*$ be these quantities. We consider the following generative process in sampling $M_*$. First a latent state $S_* \sim P_0$ is sampled. Then, the mean reward for state-action $x, a$ follows a beta prior $R_*(x, a) \sim \text{Beta}(\alpha^R_{0,S_*}(x, a))$ with $\alpha^R_{0,s}(x, a) \in \mathbb{R}^2_+$ for any latent state $s$, and the transition probabilities follow a Dirichlet prior $T_*(x, a) \sim \text{Dir}(\alpha^T_{0,S_*}(x, a))$ with $\alpha^T_{0,s}(x, a) \in \mathbb{R}^{|\mathcal{X}|}_+$. Here $\mathbb{R}_+$ denotes the space of positive reals. Finally, $M_* = (\mathcal{X}, \mathcal{A}, R_*, T_*, h, \rho)$ uses these sampled quantities.

Recall that in episode $t \in [n]$, `MixTS` samples latent state $S_t \sim P_t$, then MDP $M_t \sim P_t(\cdot \mid S_t)$. Sampling $M_t$ consists of independently sampling, for each $x, a$, mean rewards $R_{M_t}(x, a) \sim \text{Beta}(\alpha^R_{t,S_t}(x, a))$ and transition probabilities $T_{M_t}(x, a) \sim \text{Dir}(\alpha^T_{t,S_t}(x, a))$. For latent state $s$, we denote by $\alpha^R_{t,s}(x, a)$, $\alpha^T_{t,s}(x, a)$ the parameters of the respective Dirichlet posteriors. Specifically,

$$\bar{r}_t(x, a, s) = \frac{(\alpha^R_{t,s}(x, a))_1}{\left\|\alpha^R_{t,s}(x, a)\right\|_1}, \qquad c_t(x, a, s) = \sqrt{\frac{2\log(2|\mathcal{X}||\mathcal{A}|n)}{\left\|\alpha^R_{t,s}(x, a)\right\|_1 + 1}}, \tag{7}$$

are the posterior mean and confidence width for the mean reward under $x, a$. Similarly, we have

$$\bar{p}_t(x, a, x', s) = \frac{(\alpha^T_{t,s}(x, a))_{x'}}{\left\|\alpha^T_{t,s}(x, a)\right\|_1}, \qquad \phi_t(x, a, s) = \sqrt{\frac{4|\mathcal{X}|\log(4|\mathcal{X}||\mathcal{A}|n)}{\left\|\alpha^T_{t,s}(x, a)\right\|_1 + 1}}, \tag{8}$$

for the transition probabilities. We simply state the Bayes regret bound and defer a full proof to Appendix B.

**Theorem 2.** *Let*

$$\Lambda_{0,s} = \min\left\{\min_{x,a}\left\|\alpha^R_{0,s}(x, a)\right\|_1, \min_{x,a}\left\|\alpha^T_{0,s}(x, a)\right\|_1\right\}$$

*represent how concentrated the reward and transition priors are for latent state $s$, where higher values correspond to lower prior widths. Let $\Lambda_{0,\min} = \min_{s\in\mathcal{S}} \Lambda_{0,s}$. Then the $n$-episode Bayes regret of* `MixTS` *is bounded*

$$\mathcal{BR}(n) \le 6|\mathcal{X}|h^{3/2}\sqrt{c_1|\mathcal{A}|n\log(4|\mathcal{X}||\mathcal{A}|n)} + \sqrt{Lhn\log n} + c_2\,.$$

*where*

$$c_1 = \log\left(1 + \frac{hn}{2|\mathcal{X}||\mathcal{A}|\Lambda_{0,\min}}\right),$$

*and $c_2$ is poly-logarithmic in $n$.*

Similarly to Theorem 1, the above regret bound decomposes into the regret due to learning the MDP under the assumption that the latent state is known (Term 1), and the regret due to identifying the correct latent state (Term 2). Term 1 is $\tilde{\mathcal{O}}(|\mathcal{X}|h^{3/2}\sqrt{c_1|\mathcal{A}|n})$ and matches classical TS bounds (Osband et al., 2013). The prior width is captured by $\Lambda_{0,\min}$ in $c_1$, which represents the minimum pseudo-counts in our beta and Dirichlet priors. Roughly speaking, the variance of beta and Dirichlet distributions is bounded by the reciprocal of these counts (Marchal and Arbel, 2017). So, when $\Lambda_{0,\min}$ is large, the beta and Dirichlet priors over mean rewards and transitions have low widths. Through $c_1$, Term 1 goes to zero in this regime. Then the regret is dominated by Term 2, which is $\tilde{\mathcal{O}}(\sqrt{Lhn})$ for identifying the correct latent state.

# 6 EXPERIMENTS

We evaluate `MixTS` in a synthetic and real-world problems. The goals of our experiments are the following: (1) assess the degree to which the Bayes regret bounds in Theorems 1 and 2 match the actual regret, (2) show that `MixTS` outperforms TS with a less-informative unimodal prior and other online model selection algorithms in a challenging real-world problem, and (3) show that `MixTS` still performs well when extended to RL settings.

## 6.1 Synthetic Linear Bandit

We begin with a synthetic $d$-dimensional Gaussian linear bandit where $d = 30$. We consider up to $L = 30$ latent states. The latent state prior is uniform, $P_0(s) = 1/L$ for each $s$. The model parameter prior is an isotropic Gaussian $P_0(\cdot \mid s) = \mathcal{N}(\cdot; \theta_{0,s}, \sigma_0^2 I_d)$. The $i$-th entry of $\theta_{0,s}$ is 0.9 when $i = s$, and 0.1 otherwise. The action set is constant over all rounds $\mathcal{A}_t = \mathcal{A} \subseteq \mathbb{R}^d$ and consists of all $d$-dimensional indicator vectors. The reward for action $A_t$ is sampled from a Gaussian $Y_t \sim \mathcal{N}(A_t^\top \theta_*, \sigma^2)$ with $\sigma = 0.1$. The horizon is $n = 1,000$ rounds. We run `MixTS` 200 times, with $S_*$ and $\theta_*$ sampled from the prior at the beginning of each run. We vary two quantities in Theorem 1, the prior width $\sigma_0 = \lambda_{0,\max}$ and number of latent states $L$, and assess their effect on regret.

For each $\sigma_0$ and $L$, we use the mean regret over multiple runs, where in each run, model parameters are drawn as $\theta_* \sim P_0$, to approximate the Bayes regret. The regret is reported in Figure 1, together with the upper bound in Theorem 1. The upper bound is multiplied by 1/30, which changes the scale but preserves the shape. We observe that our bound correctly estimates the shape of the empirical regret as a function of $\sigma_0$. In a similar experiment, where $\sigma_0 = 0.05$ is fixed and we vary the number of latent states $L$, we again observe that our bound correctly estimates the shape of the empirical regret as a function of $L$. We conclude that Theorem 1 scales correctly with the parameters of our problem class.

## 6.2 Image Classification

In our second experiment, we consider an image classification problem with a mixture of high-level tasks. We use the CIFAR-100 dataset (Krizhevsky, 2009), which consists of 60,000 images of size $32 \times 32$. There are 50,000 training and 10,000 test images. Each image belongs to one of $L = 100$ classes (image labels).

We treat each class as a *task*, so that images in class $s$ have high reward when the task is $s$. At the beginning of each run, a class is sampled as $S_* \sim P_0$, where $P_0(s) = 1/L$ for all $s$. In round $t$, the action set $\mathcal{A}_t$ consists of 10 randomly chosen images from the CIFAR-100 test set, where one image is guaranteed to be from class $S_*$. The reward of an image from class $S_*$ is Ber(0.9) and for all other classes is Ber(0.1). The horizon is $n = 500$ rounds. For such short horizons, the effect of the prior is more noticeable. We cast this problem as a linear bandit with features from a state-of-the-art EfficientNet-L2 network Xie et al. (2020); Tan and Le (2019); Foret et al. (2021). This is a convolutional neural network pretrained on both ImageNet (Russakovsky et al., 2015) and unlabeled JFT-300M (Sun et al., 2017) with input resolution 475, and fine-tuned on the CIFAR-100 training set. Each action $a \in \mathcal{A}_t$ is a 100-dimensional feature vector, the embedding after applying the network.

The mixture prior is obtained by clustering similar tasks from the CIFAR-100 training set. This is done as follows. First, we sample 1000 random datasets of size $n = 500$ from the training set. For each dataset, we randomly choose the class $S_* \sim P_0$, and assign reward one to images from class $S_*$ and zero otherwise. Second, we fit a linear model to each dataset, where the image features are generated as above. Finally, we fit a GMM with $L$ components to the parameter vectors of the trained linear models, generating cluster means and covariances $(\theta_{0,s}, \Sigma_{0,s})_{s \in \mathcal{S}}$. The model parameter prior for $s$ is $P_0(\cdot \mid s) = \mathcal{N}(\cdot; \theta_{0,s}, \Sigma_{0,s})$.

We compare `MixTS` to four baselines: `TS`, `UniTS`, `Exp4` (Auer et al., 2002), and `CorralExp4`. `TS` is Thompson sampling with an uninformative Gaussian prior $\mathcal{N}(\mathbf{0}, I_d)$ over model parameters. `UniTS` is TS with a

unimodal Gaussian prior fit to the same data as the GMM. This baseline shows the importance of using mixtures, as opposing to just using past data. `Exp4` uses the prior means $(\theta_{0,s})_{s\in\mathcal{S}}$ as $L$ experts, where the action of expert $s$ is $\arg\max_{a\in\mathcal{A}_t} a^\top\theta_{0,s}$. The actions are a weighted vote of the experts, where better experts have higher weights. Finally, `CorralExp4` uses `Exp4` to track experts, but additionally adapts the parameters of each expert so that in round $t$, the action of expert $s$ is $\arg\max_{a\in\mathcal{A}_t} a^\top\bar{\theta}_{t,s}$, where $\bar{\theta}_{t,s}$ is defined as in (4). `CorralExp4` is an instance of a corralling bandit algorithm (Maillard and Munos, 2011; Agarwal et al., 2017; Arora et al., 2021), where a master (`Exp4`) switches between base algorithms (linear regressors). We measure the mean reward of each method, averaged over 100 independent runs. Note that all TS algorithms are misspecified in this experiment, because the models are not linear and the reward noise is not Gaussian. We use $\sigma = 0.5$ since the rewards are in $[0, 1]$. As shown in Figure 1, `MixTS` greatly outperforms `UniTS` and `TS`, especially during the cold-start regime, due to using a strong mixture prior fitted to existing data. `MixTS` also outperforms `Exp4` and `CorralExp4` by explicitly leveraging the latent state posterior to switch between models. Although `CorralExp4` uses the same model updates, it switches between the models using an adversarial algorithm.

### 6.3 Synthetic MDP

In our final experiment, we consider a synthetic finite-horizon MDP based on the *RiverSwim* environment (Osband et al., 2013). *RiverSwim* consists of $|\mathcal{X}|$ states arranged in a chain. The agent starts at the state in the middle and at every time step, can choose to swim right or left, $|\mathcal{A}| = 2$. The environment is parameterized by a latent state that denotes the direction of the current, which can be right or left, $L = 2$. At a high level, swimming with the current is always successful, but swimming against the current likely fails. If the current is to the left, the agent receives a small reward for swimming left at the leftmost state, but receives a much larger reward for swimming right at the rightmost state; if the current is to the right, the opposite holds. The optimal policy involves swimming against the current to receive the large reward. The prior mean MDP when the current is to the left is shown in Figure 2. The MDP when the current is to the right is symmetric.

In our experiments, we consider $|\mathcal{X}| = 10$ and horizon $h = 20$. The latent state prior is uniform, $P_0(s) = 1/2$ for $s$ as left or right. The MDP prior, conditioned on each latent state, consists of beta and Dirichlet priors for the mean reward and transition probabilities for each state-action pair $(x, a)$, such that the mean MDP under the prior matches the values in Figure 2. The number of episodes is $n = 1,000$ episodes, and we run `MixTS` 500 times on independent samples of the MDP from the prior. In Figure 2, we compare the mean regret over the 500 runs of `MixTS` against PSRL (Osband et al., 2013), which is a TS algorithm that uses a uniform prior over rewards and transitions. `MixTS` greatly outperforms PSRL because it identifies the correct latent state, or direction of the current, much more quickly than PSRL learns the reward and transitions from scratch.

## 7 RELATED WORK

**Thompson sampling.** Thompson sampling is known for its computational efficiency and strong empirical performance (Agrawal and Goyal, 2012; Chapelle and Li, 2012; Agrawal and Goyal, 2013). Russo and Van Roy (2013) derived first Bayes regret bounds for TS in bandits and RL (Osband et al., 2013). We build on these works by considering a mixture prior. By explicitly modeling a latent state, we can implement TS efficiently, as well as derive improved prior-dependent Bayes regret bounds. Alternatively, information theory has been used to derive Bayes regret bounds (Russo and Van Roy, 2016; Lu and Van Roy, 2019). These proofs rely on the entropy of posterior distributions, which do not have closed forms for mixtures. Recent works applied approximate TS to complex structured problems (Gopalan et al., 2014; Yu et al., 2020). Such algorithms are general, but can only be analyzed in limited settings with strong assumptions. We consider a special prior structure, and derive improved regret bounds for bandits and RL.

A related work on TS with mixture distributions is Urteaga and Wiggins (2018). The setting of this work is completely different because they study a mixture reward distribution. In comparison, we study a mixture of model parameters. To make this distinction clear, consider a linear bandit. Urteaga and Wiggins (2018) would have non-Gaussian rewards sampled from a Gaussian mixture model (GMM). We would have Gaussian rewards with model parameters sampled from a GMM. More recently, Urteaga and Wiggins (2021) proposed a non-parametric GMM over the rewards in the bandit setting. This is another instance of a mixture reward distribution.

**Online model selection.** Our work is also related to online model selection, as each latent state corresponds to a different hypothesis for the distribution of the environment. Identifying the true latent state is analogous to selecting the best-performing base model. `Exp4` (Auer et al., 2002) is one of the earliest algorithms for solving this problem in adversarial environments. Bayesian policy-reuse (BPR) (Rosman et al., 2016) could be used in stochastic environments but it

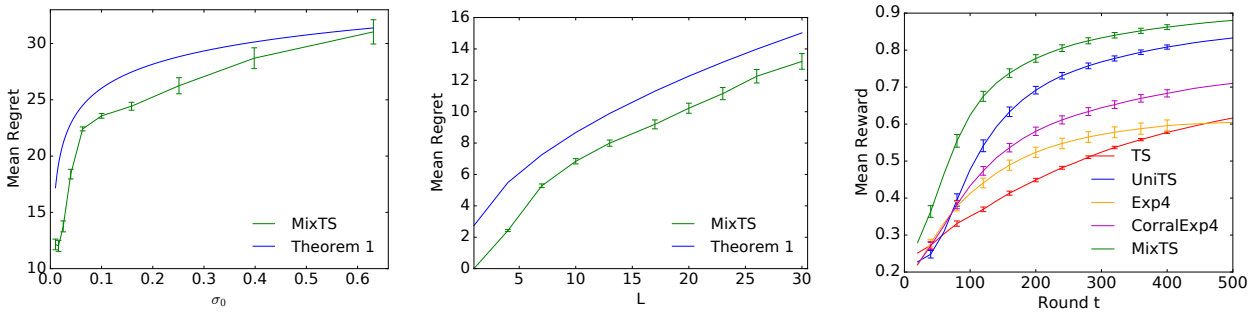


Figure 1: *Left*: Bayes regret as function of prior width $\sigma_0$. *Middle*: Bayes regret as function of the number of latent states $L$. *Right*: Mean reward on a CIFAR-100 classification bandit.

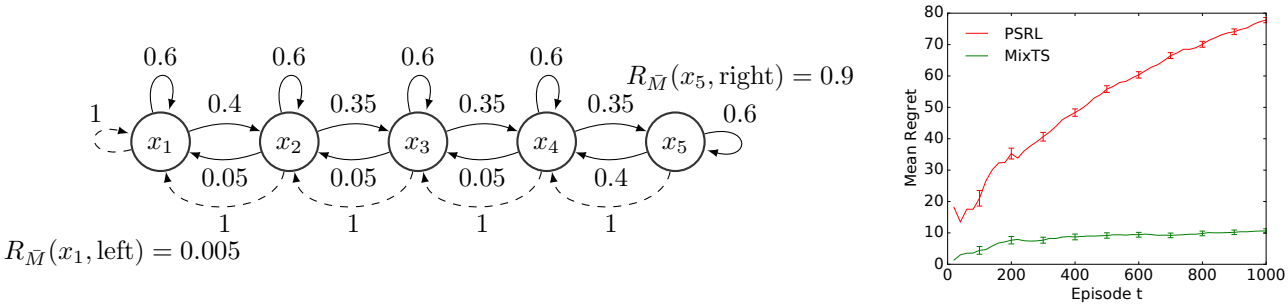


Figure 2: *Left: RiverSwim* with $|\mathcal{X}| = 5$ and current to the left. Solid and dashed arrows represent transitions under actions "left" and "right", respectively. Numbers denote the mean reward and transition probabilities of the mean MDP $\bar{M}$ under the prior. *Right:* Mean regret on finite-horizon *RiverSwim* environment.

does not have theoretical guarantees. More recently, in corralling bandits, a master algorithm learns the best-performing base bandit algorithm. Maillard and Munos (2011) proposed a modified version of `Exp4` as the master. Corralling algorithms have also been extended to the stochastic setting (Agarwal et al., 2017; Arora et al., 2021; J. Foster et al., 2019). In all above works, the base algorithm is updated only when selected by the master. In our work, because of the full Bayesian treatment, all mixture components are always updated, which increases statistical efficiency. As shown in Section 6, `MixTS` outperforms multiple online model selection baselines.

**Latent bandits.** Our setting is also an instance of latent bandits, where bandit instances are parameterized by a finite set of latent states, and each one corresponds to a different hypothesis over reward models (Maillard and Mannor, 2014; Zhou and Brunskill, 2016; Hong et al., 2020). In such structured environments, it is natural to consider a mixture prior that is learned from existing data, each component of the prior being a model distribution. However, most previous works only considered a single fixed model per latent state (Maillard and Mannor, 2014). The closest work is Hong et al. (2020), who proposed a TS algorithm `mmTS`. In a bandit, `MixTS` is an instance of `mmTS` where the conditional models are mixture components and share the same parameter space. This distinction is important, as we explicitly analyze the concentration of the mixture posterior to derive sublinear regret bounds. The regret bounds of Hong et al. (2020) are agnostic to posterior improvements and can be linear. We also apply `MixTS` to reinforcement learning, which in turn generalizes `mmTS`.

## 8 CONCLUSIONS

We propose Thompson sampling with a mixture prior (`MixTS`) for online decision making. The mixture prior is parameterized by a discrete latent state, and yields a general and tractable algorithm that can be broadly analyzed, in both bandit and RL settings. Our regret bounds reflect the structure of the prior, the number of mixture components and their widths. We evaluate `MixTS` on both synthetic and an image classification problems, and demonstrate that it performs well.

This work is a step towards analyzing TS in realistic models with latent variables. Our regret bounds depend on the number and width of the prior mixture components, but not on the latent state prior, which leaves room for improvement. We also only consider a flat discrete latent state. More expressive latent structures are an interesting direction for future work.

## References


Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *NeurIPS*, 2011.

M. Abeille and A. Lazaric. Linear thompson sampling revisited. In *AISTATS*, 2017.

A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *COLT*, 2017.

S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. 2012.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. *ICML*, 2013.

S. Agrawal and R. Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. In *NeurIPS*, 2017.

R. Arora, T. V. Marinov, and M. Mohri. Corralling stochastic bandit algorithms. In *AISTATS*, 2021.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. In *SIAM Journal of Computing*, 2002.

A. Ayoub, Z. Jia, C. Szepesvári, M. Wang, and L. F. Yang. Model-based reinforcement learning with value-targeted regression. In *ICML*, 2020.

A. Barto and R. S. Sutton. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 2018.

R. Bellman. A Markovian decision process. *Indiana University Mathematics Journal*, 6, 1957.

C. Bishop. *Pattern Recognition and Machine Learning*. Springer New York, 2006.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22, 1997.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Neural Information Processing Systems*, pages 2249–2257, 2012.

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021.

A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *ICML*, 2014.

J. Hong, B. Kveton, M. Zaheer, Y. Chow, A. Ahmed, and C. Boutilier. Latent bandits revisited. In *NeurIPS*, 2020.

D. J. Foster, A. Krishnamurthy, and H. Luo. Model selection for contextual bandits. In *NeurIPS*, 2019.

M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural Computation*, 1994.

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009.

T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, 15(3):1091 – 1114, 1987.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2019. doi: 10.1017/9781108571401.

X. Lu and B. Van Roy. Information-theoretic confidence bounds for reinforcement learning. In *NeurIPS*, 2019.

J. Macqueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

O.-A. Maillard and S. Mannor. Latent bandits. In *ICML*, 2014.

O.-A. Maillard and R. Munos. Adaptive bandits: Towards the best history-dependent strategy. In *AISTATS*, 2011.

O. Marchal and J. Arbel. On the sub-Gaussianity of the Beta and Dirichlet distributions. *CoRR*, abs/1705.00048, 2017.

I. Osband, B. Van Roy, and D. Russo. (More) efficient reinforcement learning via posterior sampling. In *NeurIPS*, 2013.

B. Rosman, M. Hawasly, and S. Ramamoorthy. Bayesian policy reuse. *Machine Learning*, 2016.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *CoRR*, abs/1301.2609, 2013.

D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. In *Journal of Machine Learning Research*, 2016.

C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. doi: 10.1109/ICCV.2017.97.

M. Tan and Q. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *ICML*, pages 6105–6114, 2019.

I. Urteaga and C. H. Wiggins. Variational inference for the multi-armed contextual bandit. In *AISTATS*, 2018.

I. Urteaga and C. H. Wiggins. Nonparametric Gaussian mixture models for the multi-armed contextual bandit. *CoRR*, abs/1808.02932, 2021.

J. Wang. Generating daily changes in market variables using a multivariate mixture of normal distributions. In *Proceedings of the 33rd Winter Conference on Simulation*, 2001.

Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

T. Yu, B. Kveton, Z. Wen, R. Zhang, and O. J. Mengshoel. Graphical models meet bandits: A variational Thompson sampling approach. In *ICML*, 2020.

L. Zhou and E. Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. In *IJCAI*, 2016.

# A Linear Bandit Proofs

## A.1 Useful Lemmas

**Lemma 1.** *Let $X \in \mathbb{R}^d$ be a random vector sampled from the multivariate Gaussian $X \sim \mathcal{N}(0, \Sigma)$. For any $\varepsilon \geq 0$, define event $E = \{\|X\|_{\Sigma^{-1}} \geq \varepsilon\}$. Then,*

$$\mathbb{E}\left[\|X\|_{\Sigma^{-1}} \mathbb{1}\{E\}\right] \leq \frac{1}{\sqrt{2\pi}} d^{3/2} \exp\left(-\frac{\varepsilon^2}{2d}\right)$$

*Proof.* Using that $X \sim \mathcal{N}(0, \Sigma)$, we can conclude that $\Sigma^{-1/2} X \sim \mathcal{N}(0, I_d)$ has independent Gaussian entries. We have

$$\begin{aligned}
\mathbb{E}\left[\|X\|_{\Sigma^{-1}} \mathbb{1}\{E\}\right] = \mathbb{E}\left[\left\|\Sigma^{-1/2} X\right\|_2 \mathbb{1}\{E\}\right] &\leq \sqrt{d}\, \mathbb{E}\left[\left\|\Sigma^{-1/2} X\right\|_\infty \mathbb{1}\{E\}\right] \\
&\leq \sqrt{d} \sum_{i=1}^{d} \frac{1}{\sqrt{2\pi}} \int_{u=\varepsilon/\sqrt{d}}^{\infty} u \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \\
&= \sqrt{d} \sum_{i=1}^{d} -\frac{1}{\sqrt{2\pi}} \int_{u=\varepsilon/\sqrt{d}}^{\infty} \left(\exp\left(-\frac{u^2}{2}\right)\right)' \mathrm{d}u \\
&= \frac{1}{\sqrt{2\pi}} d^{3/2} \exp\left(-\frac{\varepsilon^2}{2d}\right),
\end{aligned}$$

where we use that $\left\|\Sigma^{-1/2} X\right\|_2 \geq \varepsilon$ implies $\left\|\Sigma^{-1/2} X\right\|_\infty \geq \varepsilon/\sqrt{d}$, and consider each entry of $\Sigma^{-1/2} X$ separately. □

**Lemma 2.** *For round $t$ and latent state $s$, let $\theta_{t,s}, \Sigma_{t,s}$ be defined as in (4). If $\|A_t\|_2 \leq \kappa$ for all $t$, then for any $C > 0$ such that $\lambda_{\max}(\Sigma_{0,s}) \leq \sigma^2 C/\kappa^2$, then*

$$\sum_{t=1}^{n} \|A_t\|^2_{\Sigma_{t,s}} \leq \sigma^2 (1 + C) \log \frac{\det\left(\Sigma^{-1}_{n+1,s}\right)}{\det\left(\Sigma^{-1}_{0,s}\right)}.$$

*Proof.* The proof is similar to that done for Lemma 11 of Abbasi-yadkori et al. (2011). Instead, we consider the norm with respect to posterior covariance $\Sigma_{t,s}$ rather than empirical covariance $V_t^{-1}$.

We have,

$$\begin{aligned}
\det\left(\Sigma^{-1}_{n+1,s}\right) = \det\left(\Sigma^{-1}_{n,s} + \sigma^{-2} A_n A_n^\top\right) &= \det\left(\Sigma^{-1}_{n,s}\right)\left(1 + \left\|\sigma^{-2} A_n\right\|^2_{\Sigma_{n,s}}\right) \\
&= \det\left(\Sigma^{-1}_{0,s}\right) \prod_{t=1}^{n} \left(1 + \sigma^{-2} \|A_t\|^2_{\Sigma_{t,s}}\right),
\end{aligned}$$

where we use the matrix determinant lemma, which says $\det\left(A + uu^\top\right) = \det(A)\left(1 + \|u\|^2_{A^{-1}}\right)$ for matrix $A$ and vector $u$. Note that

$$\|A_t\|^2_{\Sigma_{t,s}} \leq \lambda_{\max}(\Sigma_{t,s}) \|A_t\|_2^2 \leq \kappa^2 \lambda_{\max}(\Sigma_{0,s}),$$

so if $\lambda_{\max}(\Sigma_{0,s}) \leq \sigma^2 C/\kappa^2$, then $\sigma^{-2} \|A_t\|^2_{\Sigma_{t,s}} \leq C$. Using that $x \leq (1 + C)\log(1 + x)$ for $x \in [0, C]$, we get,

$$\begin{aligned}
\sum_{t=1}^{n} \sigma^{-2} \|A_{t,s}\|^2_{\Sigma_{t,s}} &\leq (1 + C) \sum_{t=1}^{n} \log\left(1 + \sigma^{-2} \|A_{t,s}\|^2_{\Sigma_{t,s}}\right) \\
&\leq (1 + C) \log \frac{\det\left(\Sigma^{-1}_{n+1,s}\right)}{\det\left(\Sigma^{-1}_{0,s}\right)}.
\end{aligned}$$

This yields

$$\sum_{t=1}^{n} \|A_t\|^2_{\Sigma_{t,s}} \leq \sigma^2(1+C)\log\frac{\det\left(\Sigma^{-1}_{n+1,s}\right)}{\det\left(\Sigma^{-1}_{0,s}\right)},$$

as desired. □

### A.2 Proof of Theorem 1

In the outline, we were able to trivially bound the regret of each round by 1; this is no longer the case since $\theta_*$ is a sample from a Gaussian. To handle this, we introduce event

$$E_0 = \left\{\|\theta_* - \theta_{0,S_*}\|_{\Sigma^{-1}_{0,S_*}} \leq \sqrt{2d\log(dn)}\right\},$$

which occurs when $\theta_*$ is not far from its prior mean. We can bound the regret by,

$$\mathbb{E}\left[\sum_{t=1}^{n} A_{t,*}^\top\theta_* - A_t^\top\theta_*\right] \leq \mathbb{E}\left[\sum_{t=1}^{n}(A_{t,*}^\top\theta_* - A_t^\top\theta_*)\mathbb{1}\{E_0\}\right] + \mathbb{E}\left[\sum_{t=1}^{n}(A_{t,*}^\top\theta_* - A_t^\top\theta_*)\mathbb{1}\{\bar{E}_0\}\right].$$

When $\bar{E}_0$ occurs, the regret for a round can be arbitrarily large; to handle this, we factor in that $\bar{E}_0$ is unlikely. Fix round $t$. We bound the regret in round $t$ as

$$\begin{aligned}
\mathbb{E}_t\left[(A_{t,*}^\top\theta_* - A_t^\top\theta_*)\mathbb{1}\{\bar{E}_0\}\right] &\leq \mathbb{E}_t\left[A_{t,*}^\top(\theta_* - \theta_{0,S_*})\mathbb{1}\{\bar{E}_0\}\right] + \mathbb{E}_t\left[A_{t,*}^\top\theta_{0,S_*}\mathbb{1}\{\bar{E}_0\}\right] \\
&\leq \mathbb{E}_t\left[\|A_{t,*}\|_{\Sigma_{0,S_*}}\|\theta_* - \theta_{0,S_*}\|_{\Sigma^{-1}_{0,S_*}}\mathbb{1}\{\bar{E}_0\}\right] + \mathbb{E}_t\left[\|A_{t,*}\|_2\|\theta_{0,S_*}\|_2\mathbb{1}\{\bar{E}_0\}\right] \\
&\leq \sqrt{\kappa^2\lambda_{0,\max}}\,\mathbb{E}\left[\|\theta_* - \theta_{0,S_*}\|_{\Sigma^{-1}_{0,S_*}}\mathbb{1}\{\bar{E}_0\}\right] + \kappa\mathbb{P}\left(\bar{E}_0\right),
\end{aligned}$$

where we use the Cauchy-Schwartz inequality, and $\|a\|_{\Sigma_{0,s}} \leq \sqrt{\lambda_{\max}(\Sigma_{0,s})}\,\|a\|_2 \leq \sqrt{\kappa^2\lambda_{0,\max}}$ for any action $a$ and latent state $s$. Since $\theta_* - \theta_{0,S_*} \sim \mathcal{N}(0, \Sigma_{0,S_*})$, we have $\mathbb{P}\left(\bar{E}_0\right) \leq n^{-1}$ and

$$\mathbb{E}\left[\|\theta_* - \theta_{0,S_*}\|_{\Sigma^{-1}_{0,S_*}}\mathbb{1}\{\bar{E}_0\}\right] \leq \sqrt{\frac{d}{2\pi}}n^{-1},$$

where we apply Lemma 1 with $\varepsilon = \sqrt{2d\log(dn)}$. Hence, we can bound the Bayes regret as

$$\mathbb{E}\left[\sum_{t=1}^{n} A_{t,*}^\top\theta_* - A_t^\top\theta_*\right] \leq \mathbb{E}\left[\sum_{t=1}^{n}(A_{t,*}^\top\theta_* - A_t^\top\theta_*)\mathbb{1}\{E_0\}\right] + \sqrt{\frac{\kappa^2\lambda_{0,\max}d}{2\pi}} + \kappa.$$

When $E_0$ occurs, we have $M = \sqrt{2\kappa^2\lambda_{0,\max}d\log(dn)} + \kappa$ is an upper-bound on regret for a round. We use $\langle\cdot\rangle_M = \min\{\cdot, M\}$.

From here, we can follow the analysis outline in Section 4.1 using $\bar{\mu}_t, \sigma_t$ defined as

$$\bar{\mu}_t(a,s) = a^\top\bar{\theta}_{s,t},\ \ \sigma_t(a,s) = \|a\|_{\Sigma_{t,s}}\sqrt{2d\log(dn)}.$$

Using Equation (1), we can decompose the Bayes regret in a linear bandit as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n}(A_{t,*}^\top\theta_* - A_t^\top\theta_*)\mathbb{1}\{E_0\}\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_t\left[(A_{t,*}^\top\theta_* - A_{t,*}^\top\bar{\theta}_{t,S_*})\mathbb{1}\{E_0\}\right]\right] + \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_t\left[(A_t^\top\bar{\theta}_{t,S_t} - A_t^\top\theta_*)\mathbb{1}\{E_0\}\right]\right]. \qquad (9)
\end{aligned}$$

We bound each term individually.

**Step 1.** Let us first consider the first term of (9). Fix round $t$. Let us define event

$$E_t = \left\{ \left\| \theta_* - \bar{\theta}_{t,S_*} \right\|_{\Sigma^{-1}_{t,S_*}} \leq \sqrt{2d \log(dn)} \right\},$$

which occurs when $\theta_*$ is not far from the mean of conditional posterior $P_t(\cdot \mid S_*) = \mathcal{N}(\bar{\theta}_{t,S_*}, \Sigma_{t,S*})$. Note that $E_1 = E_0$ from earlier. We can bound

$$\begin{aligned}
\mathbb{E}_t \left[ A_{t,*}^\top \theta_* - A_{t,*}^\top \bar{\theta}_{t,S_*} \right] &= \mathbb{E}_t \left[ \left( A_{t,*}^\top \theta_* - A_{t,*}^\top \bar{\theta}_{t,S_*} \right) \mathbb{1}\{E_t\} \right] + \mathbb{E}_t \left[ \left( A_{t,*}^\top \theta_* - A_{t,*}^\top \bar{\theta}_{t,S_*} \right) \right) \mathbb{1}\left\{\bar{E}_t\right\} \right] \\
&\leq \mathbb{E}_t \left[ \|A_{t,*}\|_{\Sigma_{t,S_*}} \sqrt{2d \log(dn)} \right] + \mathbb{E}_t \left[ \left( A_{t,*}^\top \theta_* - A_{t,*}^\top \bar{\theta}_{t,S_*} \right) \mathbb{1}\left\{\bar{E}_t\right\} \right] ,
\end{aligned}$$

where we use that when $E_t$ occurs, we have

$$(A_{t,*}^\top \theta_* - A_{t,*}^\top \bar{\theta}_{t,S_*}) \mathbb{1}\{E_t\} \leq \|A_{t,*}\|_{\Sigma_{t,S_*}} \left\| \theta_* - \bar{\theta}_{t,s} \right\|_{\Sigma^{-1}_{t,S_*}} \mathbb{1}\{E_t\} \leq \|A_{t,*}\|_{\Sigma_{t,S_*}} \sqrt{2d \log(dn)} .$$

Now, when $\bar{E}_t$ occurs, we have

$$\mathbb{E}_t \left[ \left( A_{t,*}^\top \theta_* - A_{t,*}^\top \bar{\theta}_{t,S_*} \right) \mathbb{1}\left\{\bar{E}_t\right\} \right] \leq \kappa \sqrt{\lambda_{0,\max}}\, \mathbb{E}_t \left[ \left\| \theta_* - \bar{\theta}_{t,S_*} \right\|_{\Sigma^{-1}_{t,S_*}} \mathbb{1}\left\{\bar{E}_t\right\} \right] ,$$

where we again use Cauchy-Schwartz and $\|a\|_{\Sigma_{t,s}} \leq \sqrt{\kappa^2 \lambda_{0,\max}}$ for any action $a$ and latent state $s$.

Using that that $\theta_* - \bar{\theta}_{t,s} \mid H_t \sim \mathcal{N}(0, \Sigma_{t,s})$, we can use Lemma 1 with $\varepsilon = \sqrt{2d \log(dn)}$ to bound

$$\mathbb{E}_t \left[ \left\| \theta_* - \bar{\theta}_{t,S_*} \right\|_{\Sigma^{-1}_{t,S_*}} \mathbb{1}\left\{\bar{E}_t\right\} \right] \leq \sqrt{\frac{d}{2\pi}} n^{-1} .$$

Hence, we can bound the first term of (9) by

$$\mathbb{E}\left[ \sum_{t=1}^n \mathbb{E}_t \left[ A_{t,*}^\top \theta_* - A_{t,*}^\top \bar{\theta}_{t,S_*} \right] \right] \leq \sqrt{2d \log(dn)}\, \mathbb{E}\left[ \sum_{t=1}^n \|A_{t,*}\|_{\Sigma_{t,S_*}} \right] + \sqrt{\frac{\kappa^2 \lambda_{0,\max} d}{2\pi}} .$$

**Step 2.** We define $C_t$ as a high-probability set around latent states using the following construction:

$$C_t = \left\{ s \in \mathcal{S} : G_t(s) \leq 2\sigma \sqrt{N_t(s) \log n} \right\} ,$$

where $N_t(s) = \sum_{\ell=1}^{t-1} \mathbb{1}\{S_t = s\}$ and

$$G_t(s) = \sum_{\ell=1}^{t-1} \mathbb{1}\{S_t = s\} \left( A_t^\top \bar{\theta}_{t,s} - \|A_t\|_{\Sigma_{t,s}} \sqrt{2d \log n} - Y_t \right)$$

is the "over-estimation" of the predicted rewards under a latent state and the realized reward. We show that $S_* \in C_t$ holds with high probability for any round via the following lemma.

**Lemma 3.** *For any round $t$, $\mathbb{P}(S_* \notin C_t) \leq 2Ln^{-1}$.*

*Proof.* We know that $S_* \in C_t$ occurs if $G_t(S_*)$ is not too large. On a high-level, our goal is to upper-bound $G_t(S_*)$ by a martingale with respect to history, then bound the probability that $G_t(S_*)$ is too large using Azuma's inequality for concentration of martingales.

For $\ell < t$, we know that $\theta_* - \bar{\theta}_{\ell,s} \mid H_\ell \sim \mathcal{N}(0, \Sigma_{\ell,s})$. Let us define

$$\mathcal{E}_\ell = \left\{ \left\| \theta_* - \bar{\theta}_{\ell,S_*} \right\|_{\Sigma^{-1}_{\ell,S_*}} \leq 2\sqrt{d \log n}, \right\}$$

as the event that $\theta_*$ is not too far from its posterior mean. Let $\mathcal{E} = \cap_{\ell=1}^{t-1} \{\mathcal{E}_\ell\}$ be the event that this holds for all rounds up to round $t$ and $\bar{\mathcal{E}}$ be the complement. We know that

$$\mathbb{1}\{S_* \notin C_t\} = \mathbb{1}\left\{ G_t(S_*) \geq 2\sigma \sqrt{N_t(S_*) \log n} \right\} \leq \mathbb{1}\left\{\bar{\mathcal{E}}\right\} + \mathbb{1}\{\mathcal{E}\}\, \mathbb{1}\left\{ G_t(S_*) \geq 2\sigma \sqrt{N_t(S_*) \log n} \right\} ,$$

which implies that

$$\mathbb{P}\left(S_* \notin C_t\right) \leq \mathbb{P}\left(\bar{\mathcal{E}}\right) + \mathbb{P}\left(G_t(S_*)\mathbb{1}\{\mathcal{E}\} \geq 2\sigma\sqrt{N_t(S_*)\log n}\right). \tag{10}$$

We will bound each probability individually. For the first probability of (10), we simply have

$$\mathbb{P}\left(\bar{\mathcal{E}}\right) \leq \sum_{s\in\mathcal{S}}\sum_{\ell=1}^{t-1}\mathbb{E}\left[\mathbb{P}_\ell\left(\left\|\theta_* - \bar{\theta}_{\ell,s}\right\|_{\Sigma_{\ell,s}^{-1}} \geq 2\sqrt{d\log n}\right)\right] \leq Ln^{-1},$$

where we use that for $S_* = s$ and round $\ell$, we have $\left\|\theta_* - \bar{\theta}_{\ell,s}\right\|_{\Sigma_{\ell,s}^{-1}} \mid H_\ell$ is the sum of independent Gaussians. Then, we take an expectation over histories, and use a union bound over latent states and rounds.

Now, consider the second probability in (10). Fix $S_* = s$, and let $\mathcal{T}_{t,s} = \{\ell < t : S_\ell = s\}$ be the rounds where $s$ is sampled up to round $t$. Also, let $Z_\ell = A_\ell^\top\theta_* - Y_\ell$. Observe that $Z_\ell \sim \mathcal{N}(0,\sigma^2)$, so that $(Z_\ell)_{t\in\mathcal{T}_{t,s}}$ is a martingale difference sequence with respect to histories $(H_\ell)_{t\in\mathcal{T}_{t,s}}$. We have

$$\begin{aligned}
&(A_\ell^\top\bar{\theta}_{t,\ell} - \|A_\ell\|_{\Sigma_{t,\ell}}\sqrt{2d\log n} - Y_\ell)\mathbb{1}\{\mathcal{E}_\ell\} \\
&= (A_\ell^\top\theta_* + A_\ell^\top(\bar{\theta}_{t,\ell} - \theta_*) - \|A_\ell\|_{\Sigma_{t,\ell}}\sqrt{2d\log n} - Y_\ell)\mathbb{1}\{\mathcal{E}_\ell\} \\
&\leq (A_\ell^\top\theta_* + \|A_\ell\|_{\Sigma_{t,\ell}}\left\|\bar{\theta}_{t,\ell} - \theta_*\right\|_{\Sigma_{t,\ell}^{-1}} - 2\|A_\ell\|_{\Sigma_{t,\ell}}\sqrt{d\log n} - Y_\ell)\mathbb{1}\{\mathcal{E}_\ell\} \leq Z_\ell,
\end{aligned}$$

where we use Cauchy-Schwartz in the inequality. This implies that

$$G_t(s)\mathbb{1}\{\mathcal{E}\} = \sum_{\ell\in\mathcal{T}_{t,s}}(A_\ell^\top\bar{\theta}_{t,\ell} - 2\|A_\ell\|_{\Sigma_{t,\ell}}\sqrt{d\log n} - Y_\ell)\mathbb{1}\{\mathcal{E}_\ell\} \leq \sum_{\ell\in\mathcal{T}_{t,s}} Z_\ell.$$

For any round $t$, and latent state $s$, we have that $\mathcal{T}_{t,s}$ is a random quantity. First, we fix $|\mathcal{T}_{t,s}| = N_t(s) = u$ where $u < t$ and yield the following due to Azuma's inequality,

$$\mathbb{P}\left(G_t(s)\mathbb{1}\{\mathcal{E}\} \geq 2\sigma\sqrt{u\log n}\right) \leq \mathbb{P}\left(\sum_{\ell\in\mathcal{T}_{t,s}} Z_\ell(s) \geq 2\sigma\sqrt{u\log n}\right) \leq \exp\left[-2\log n\right] = n^{-2}.$$

Finally, by the union bound, we have

$$\mathbb{P}\left(G_t(S_*)\mathbb{1}\{\mathcal{E}\} \geq 2\sigma\sqrt{N_t(S_*)\log n}\right) \leq \sum_{s\in\mathcal{S}}\sum_{u=1}^{t-1}\mathbb{P}\left(G_t(s)\mathbb{1}\{\mathcal{E}\} \geq 2\sigma\sqrt{u\log n}\right) \leq Ln^{-1}.$$

Combining the two bounds completes the proof. □

**Step 3.** Now, we consider the second term of (9). We have,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^n\langle A_t^\top\bar{\theta}_{t,S_t} - A_t^\top\theta_*\rangle_M\right] &\leq M\sum_{t=1}^n\mathbb{P}\left(S_t \notin C_t\right) + \mathbb{E}\left[\sum_{t=1}^n\langle A_t^\top\bar{\theta}_{t,S_t} - A_t^\top\theta_*\rangle_M\mathbb{1}\{S_t\in C_t\}\right] \\
&\leq M\sum_{t=1}^n\mathbb{P}\left(S_* \notin C_t\right) + \mathbb{E}\left[\sum_{t=1}^n\langle A_t^\top\bar{\theta}_{t,S_t} - A_t^\top\theta_*\rangle_M\mathbb{1}\{S_t\in C_t\}\right]
\end{aligned}$$

where we use that conditioned on $H_t$, $S_t, S_*$ are i.i.d. to get $\mathbb{P}\left(S_t \in C_t\right) = \mathbb{E}\left[\mathbb{P}_t\left(S_t \in C_t\right)\right] = \mathbb{E}\left[\mathbb{P}_t\left(S_* \in C_t\right)\right] = \mathbb{P}\left(S_* \in C_t\right)$ From Lemma 3, the first term is $2LM$. From the outline in Section 4.1, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^n\langle A_t^\top\bar{\theta}_{t,S_t} - A_t^\top\theta_*\rangle_M\mathbb{1}\{S_t\in C_t\}\right] \\
&\leq 2\sqrt{d\log(dn)}\,\mathbb{E}\left[\sum_{t=1}^n\|A_t\|_{\Sigma_{t,S_t}}\right] + \mathbb{E}\left[\sum_{t=1}^n\langle A_t^\top\bar{\theta}_{t,S_t} - 2\|A_t\|_{\Sigma_{t,S_t}}\sqrt{d\log(dn)} - Y_t\rangle_M\mathbb{1}\{S_t\in C_t\}\right].
\end{aligned}$$

The last term can be bounded as

$$\sum_{t=1}^{n}\langle A_t^\top \bar{\theta}_{t,S_t} - 2\left\|A_t\right\|_{\Sigma_{t,S_t}}\sqrt{d\log(dn)} - Y_t\rangle_M \mathbb{1}\{S_t \in C_t\} \le \sum_{s\in\mathcal{S}} G_n(s) + LM \le 2\sigma\sqrt{Ln\log n} + LM\,,$$

where for latent state $s$, and $t' = \max_{t\in[n]}\{S_t = s\}$ as the last round that a latent state $s$ is acted upon, we use that there is an upper-bound on $G_{t'}(s) \le G_n(s)$ by definition of $s \in C_t$. We trivially bound the regret by $M$ for the last round $s$ is acted upon.

Hence, we can bound the second term of (9) by

$$\mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_t\left[(A_t^\top\bar{\theta}_{t,S_t} - A_t^\top\theta_*)\mathbb{1}\{E_0\}\right]\right] \le 2\sqrt{d\log n}\,\mathbb{E}\left[\sum_{t=1}^{n}\left\|A_t\right\|_{\Sigma_{t,S_t}}\right] + 2\sigma\sqrt{Ln\log n} + 3LM\,.$$

What remains is bounding the sum of confidence widths. We have

$$\begin{aligned}
\sum_{t=1}^{n}\left\|A_t\right\|_{\Sigma_{t,S_t}} \le \sum_{t=1}^{n}\max_{s\in\mathcal{S}}\left\|A_t\right\|_{\Sigma_{t,s}} &\le \max_{s\in\mathcal{S}}\sqrt{n\sum_{t=1}^{n}\left\|A_t\right\|^2_{\Sigma_{t,s}}} \\
&\le \max_{s\in\mathcal{S}}\sqrt{\sigma^2\left(1+\frac{\kappa^2\lambda_{0,\max}}{\sigma^2}\right)n\log\left(\frac{\det\left(\Sigma_{n+1,s}^{-1}\right)}{\det\left(\Sigma_{0,s}^{-1}\right)}\right)} \\
&\le \sqrt{\sigma^2\left(1+\frac{\kappa^2\lambda_{0,\max}}{\sigma^2}\right)nd\log\left(1+n\frac{\kappa^2\lambda_{0,\max}}{\sigma^2 d}\right)}\,,
\end{aligned}$$

where we first use that $\|A_t\|_{\Sigma_{t,s}}$ for latent states $s$ differ among one another only through their prior, and then use Lemma 2 to bound the sum of norms. We use the determinant-trace inequality to bound,

$$\log\left(\frac{\det\left(\Sigma_{n+1,s}^{-1}\right)}{\det\left(\Sigma_{0,s}^{-1}\right)}\right) \le d\log\left(\frac{\mathsf{trace}\left(\Sigma_{0,s}^{-1}\right)+n\sigma^{-2}\kappa^2}{\mathsf{trace}\left(\Sigma_{0,s}^{-1}\right)}\right) \le d\log\left(1+n\frac{\kappa^2\lambda_{0,\max}}{\sigma^2 d}\right)\,,$$

where we use that

$$\mathsf{trace}\left(\Sigma_{0,s}^{-1}\right) \ge \lambda_{\min}(\Sigma_{0,s}^{-1})d = \lambda_{\max}^{-1}(\Sigma_{0,s})d \ge \lambda_{0,\max}^{-1}d\,.$$

Combining the bounds across all steps yields

$$\begin{aligned}
\mathcal{BR}(n) \le 4d\sqrt{\sigma^2\left(1+\frac{\kappa^2\lambda_{0,\max}}{\sigma^2}\right)n\log(dn)\log\left(1+n\frac{\kappa^2\lambda_{0,\max}}{\sigma^2 d}\right)} + 2\sqrt{\sigma^2 Ln\log n} \\
+ 3L\sqrt{2\kappa^2\lambda_{0,\max}d\log(dn)} + 2\sqrt{\frac{\kappa^2\lambda_{0,\max}d}{2\pi}} + 4L\kappa\,.
\end{aligned}$$

□

# B Tabular MDP Proofs

## B.1 Useful Lemmas

**Lemma 4** (Theorem 1 and 3 of Marchal and Arbel (2017)). *Let $X \sim \text{Beta}(\alpha, \beta)$ for $\alpha, \beta > 0$. Then $X - \mathbb{E}[X]$ is $\sigma^2$-sub-Gaussian with $\sigma^2 = 1/(4(\alpha + \beta + 1))$. Similarly, let $X \sim \text{Dir}(\alpha)$ for $\alpha \in \mathbb{R}_+^d$. Then $X - \mathbb{E}[X]$ is $\sigma^2$-sub-Gaussian with $\sigma^2 = 1/(4(\|\alpha\|_1 + 1))$.*

**Lemma 5** (Value difference lemma). *For any MDPs $M'$, $M$, and policy $\pi$,*

$$V_{M'}(\pi) - V_M(\pi) \leq \mathbb{E}\left[\sum_{i=1}^{h} R_{M'}(X_i, A_i) - R_M(X_i, A_i) h \left\|T_{M'}(X_i, A_i) - T_M(X_i, A_i)\right\|_1\right].$$

**Lemma 6.** *For episode $t$ and state $s$, let $\beta_t(s, x, a) = c_t(s, x, a) + \phi_t(s, x, a)$ for any $(x, a)$ as in (7), (8), respectively. Let $\Lambda_{0,s} = \min\{\min_{x,a} \left\|\alpha_{0,s}^R(x, a)\right\|_1, \min_{x,a} \left\|\alpha_{0,s}^T(x, a)\right\|_1\}$ represent at least how concentrated the reward and transition priors are for latent state $s$, where higher values correspond to lower prior widths. Then we have that*

$$h \sum_{t=1}^{n} \sum_{i=1}^{h} \beta_t(X_{t,i}, A_{t,i}, s) \leq 4|\mathcal{X}|h\sqrt{|\mathcal{A}|nh \log(4|\mathcal{X}||\mathcal{A}|n) \log\left(1 + \frac{nh}{2|\mathcal{X}||\mathcal{A}|\Lambda_{0,s}}\right)} + |\mathcal{X}||\mathcal{A}|h^2.$$

*Proof.* The proof is similar to that done in Osband et al. (2013). However, we use prior-dependent definitions for the confidence width $\beta_t$. First, we define $N_t(x, a) = \sum_{\ell=1}^{t-1} \sum_{i=1}^{h} \mathbb{1}\{X_{\ell,i} = x, A_{\ell,i} = a\}$ as the number of times $x, a$ were sampled up to episode $t$. We can decompose the sum as

$$\sum_{t=1}^{n} \sum_{i=1}^{h} \beta_t(X_{t,i}, A_{t,i}, s) \leq \sum_{t=1}^{n} \sum_{i=1}^{h} \mathbb{1}\{N_t(X_{t,i}, A_{t,i}) \leq h\} + \sum_{t=1}^{n} \sum_{i=1}^{h} \mathbb{1}\{N_t(X_{t,i}, A_{t,i}) > h\} \beta_t(X_{t,i}, A_{t,i}, s),$$

where we trivially bound the regret in a step of an episode by 1. Therefore, the first term is bounded as $|\mathcal{X}||\mathcal{A}|h$.

For the second term, let us additionally define $N_{t,i}(x, a) = N_t(x, a) + \sum_{k=1}^{i-1} \mathbb{1}\{X_{t,k} = x, A_{t,k} = a\}$ as the number of times $x, a$ were sampled up to step $i$ of episode $t$. Now, if $N_t(x, a) > h$, then we know that $N_{t,i}(x, a) \leq N_t(x, a) + h \leq 2N_t(x, a)$. We consider $c_t, \phi_t$ of $\beta_t$ individually. We have

$$\begin{aligned}
&\sum_{t=1}^{n} \sum_{i=1}^{h} \mathbb{1}\{N_t(X_{t,i}, A_{t,i}) > h\} c_t(X_{t,i}, A_{t,i}, s) \\
&= \sum_{t=1}^{n} \sum_{i=1}^{h} \mathbb{1}\{N_t(X_{t,i}, A_{t,i}) > h\} \sqrt{\frac{2\log(2|\mathcal{X}||\mathcal{A}|n)}{\left\|\alpha_{t,s}^R(x, a)\right\|_1 + 1}} \\
&= \sum_{x,a} \sum_{t=1}^{n} \sum_{i=1}^{h} \mathbb{1}\{N_t(x, a) > h\} \sqrt{\frac{2\log(2|\mathcal{X}||\mathcal{A}|n)}{\left\|\alpha_{0,s}^R(x, a)\right\|_1 + N_t(x, a) + 1}} \\
&\leq \sum_{x,a} \sum_{t=1}^{n} \sum_{i=1}^{h} \sqrt{\frac{4\log(2|\mathcal{X}||\mathcal{A}|n)}{2\left\|\alpha_{0,s}^R(x, a)\right\|_1 + N_{t,i}(x, a)}} \\
&\leq 2\sqrt{\log(2|\mathcal{X}||\mathcal{A}|n)} \sum_{x,a} \sqrt{N_{n+1}(x, a) \sum_{u=1}^{N_{n+1}(x,a)} \frac{1}{2\left\|\alpha_{0,s}^R(x, a)\right\|_1 + u}} \\
&\leq 2\sqrt{|\mathcal{X}||\mathcal{A}|nh \log(2|\mathcal{X}||\mathcal{A}|n)} \sqrt{\sum_{u=1}^{nh/|\mathcal{X}||\mathcal{A}|} \frac{1}{2\Lambda_{0,s} + u}} \\
&\leq 2\sqrt{|\mathcal{X}||\mathcal{A}|nh \log(2|\mathcal{X}||\mathcal{A}|n) \log\left(1 + \frac{nh}{2|\mathcal{X}||\mathcal{A}|\Lambda_{0,s}}\right)},
\end{aligned}$$

where for the last inequality, we use that for any $x > 0$,

$$\sum_{u=1}^{nh/|\mathcal{X}||\mathcal{A}|} \frac{1}{x+u} \leq \int_{u=x}^{x+nh/|\mathcal{X}||\mathcal{A}|} u^{-1} du \leq \log\left(1 + \frac{nh}{|\mathcal{X}||\mathcal{A}|x}\right).$$

Similarly, we have

$$\sum_{t=1}^{n}\sum_{i=1}^{h} \mathbb{1}\{N_t(X_{t,i}, A_{t,i}) > h\}\, \phi_t(s, X_{t,i}, A_{t,i})$$
$$\leq 2|\mathcal{X}|h\sqrt{2|\mathcal{A}|nh\log(4|\mathcal{X}||\mathcal{A}|n)\log\left(1 + \frac{nh}{2|\mathcal{X}||\mathcal{A}|\Lambda_{0,s}}\right)}.$$

Combining the two bounds yields

$$h\sum_{t=1}^{n}\sum_{i=1}^{h} \beta_t(s, X_{t,i}, A_{t,i}) \leq 4|\mathcal{X}|h\sqrt{2|\mathcal{A}|nh\log(4|\mathcal{X}||\mathcal{A}|n)\log\left(1 + \frac{nh}{2|\mathcal{X}||\mathcal{A}|\Lambda_{0,s}}\right)} + |\mathcal{X}||\mathcal{A}|h^2.$$

□

### B.2 General Analysis Outline

**Step 1.** Bound the Bayes regret due to the first term of (6). For episode $t$, we introduce event

$$E_t = \{\forall(x,a) : |R_{M_t}(x,a) - \bar{r}_t(x,a,S_t)| \leq c_t(x,a,S_t)\,,\ \|T_{M_t}(x,a) - \bar{p}_t(x,a,S_t)\|_1 \leq \phi_t(x,a,S_t)\}$$

to denote when the sampled mean rewards and transitions are not far from their posterior means for all state-action pairs. Using Lemma 5, we know that

$$\begin{aligned}
&\mathbb{E}_t\left[V_*(\pi_*) - \overline{V}_t(\pi_*, S_*)\right] \\
&= \mathbb{E}_t\left[\mathbb{E}_{M\sim P_t(\cdot|S_*)}\left[V_*(\pi_*) - V_M(\pi_*)\right]\right] \\
&\leq \mathbb{E}_t\left[\sum_{i=1}^{h}(R_*(X_{t,i},A_{t,i}) - r_t(X_{t,i},A_{t,i},S_*)) + h\left\|T_*(X_{t,i},A_{t,i}) - \bar{p}_t(X_{t,i},A_{t,i},S_*)\right\|_1\right] \\
&\leq \mathbb{E}_t\left[h\sum_{i=1}^{h}(R_*(X_{t,i},A_{t,i}) - r_t(X_{t,i},A_{t,i},S_*) + \|T_*(X_{t,i},A_{t,i}) - \bar{p}_t(X_{t,i},A_{t,i},S_*)\|_1)\mathbb{1}\{\bar{E}_t\}\right] \\
&\quad + \mathbb{E}_t\left[h\sum_{i=1}^{h}\beta_t(X_{t,i},A_{t,i},S_t)\right],
\end{aligned}$$

where $\beta_t(x,a,s) = c_t(x,a,s) + \phi_t(x,a,s)$. Here, we take an expectation over MDPs to apply Lemma 5, then condition on $E_t$ occurring. The second term can be bounded as a sum of confidence widths, and the remaining term can be bounded by using that conditioned on $H_t$, $\bar{E}_t$ is unlikely.

**Step 2.** For each episode $t$, construct $C_t$ such that $S_* \in C_t$ with high probability. To do so, we define $N_t(s) = \sum_{\ell=1}^{k-1}\mathbb{1}\{S_\ell = s\}$ as the number of times $s$ was acted upon and

$$G_t(s) = \sum_{\ell=1}^{t-1}\mathbb{1}\{S_\ell = s\}\left(\overline{V}_t(\pi_t, s) - \eta h\sum_{i=1}^{h}\beta_\ell(X_{\ell,i},A_{\ell,i},s) - \sum_{i=1}^{h}R_{\ell,i}\right)$$

as the total over-estimation of observed returns by assuming that $s$ is the true latent state, where $\eta \in \mathbb{R}$ is a scaling factor. Here, we use the shorthand $\beta_t(x,a,s) = c_t(x,a,s) + \phi_t(x,a,s)$. Then we define $C_t$ as containing all latent states $s$ where $G_t(s) = \mathcal{O}(\sqrt{N_t(s)h\log n})$. Note that we scale by an additional $\mathcal{O}(\sqrt{h})$ over the outline for bandits in Section 4.1 to account for taking the summation over a trajectory. We show that for any episode $t$, $\mathbb{P}\left(S_* \notin C_t\right) = \mathcal{O}(1/n)$. This means that with high probability, the true latent state lies in $C_t$.

**Step 4.** We can decompose the second term of (6) as

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{n}\overline{V}_t(\pi_t,S_t)-V_*(\pi_t)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_t\left[(\overline{V}_t(\pi_t,S_t)-V_*(\pi_t))\mathbb{1}\{S_t\in C_t\}\right]\right]+h\sum_{t=1}^{n}\mathbb{P}\left(S_t\notin C_t\right)\\ &\leq \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{E}_t\left[(\overline{V}_t(\pi_t,S_t)-V_*(\pi_t))\mathbb{1}\{S_t\in C_t\}\right]\right]+h\sum_{t=1}^{n}\mathbb{P}\left(S_*\notin C_t\right),\end{aligned}$$

where we use that conditioned on $H_t$, latent states $S_*, S_t$ are identically distributed. From Step 1 and 2, we know that the second term is bounded as $2Lh$. Finally, the remaining term can be bounded as

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{n}(\overline{V}_t(\pi_t,S_t)-V_*(\pi_t))\mathbb{1}\{S_t\in C_t\}\right]\\ &=\eta h\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{h}\beta_t(X_{t,i},A_{t,i},S_t)\right]+\mathbb{E}\left[\sum_{t=1}^{n}\left(\overline{V}_t(\pi_t,S_t)-\eta h\sum_{i=1}^{h}\beta_t(X_{t,i},A_{t,i},S_t)-\sum_{i=1}^{h}R_{t,i}\right)\mathbb{1}\{S_t\in C_t\}\right],\end{aligned}$$

where we use that $\mathbb{E}_t\left[\sum_{i=1}^{h}R_{t,i}\mid\pi_t,M_*\right]=\mathbb{E}_t\left[V_*(\pi_t)\right]$. The second can be bounded as the sum of confidence widths, which concentrate over time. The remaining term can be bounded by the sum of gaps $\sum_{s\in\mathcal{S}}G_{n+1}(s)$, which we know is bounded by $\mathcal{O}(\sqrt{Lnh\log n}+Lh)$ after trivially bounding the regret the last time each latent state is acted upon by $h$.

### B.3 Proof of Theorem 2

Recall from the proof sketch in Section 5.1 that $\overline{V}_t(s,\pi)=\mathbb{E}_{M\sim\bar{p}_t(\cdot|s)}\left[V_M(\pi)\right]$ is the expected value of a policy under state $s$, marginalized over MDPs sampled from its conditional posterior for episode $t$. We want to bound each term of the regret decomposition in (6) separately.

**Step 1.** For episode $t$, let

$$E_t=\{\forall(x,a):|R_{M_t}(x,a)-\bar{r}_t(x,a,S_t)|\leq c_t(x,a,S_t)\,,\ \|T_{M_t}(x,a)-\bar{p}_t(x,a,S_t)\|_1\leq\phi_t(x,a,S_t)\}$$

denote the event that the sampled mean rewards and transitions are not far from their posterior means for all state-action pairs. From the sketch in Appendix B.2, we rewrite the first term of (6) as

$$\begin{aligned}&\mathbb{E}_t\left[V_{M_t}(\pi_t)-\overline{V}_t(\pi_t,S_t)\right]\\ &\leq h\sum_{i=1}^{h}\mathbb{E}_t\left[(R_{M_t}(X_{t,i},A_{t,i})-\bar{r}_t(X_{t,i},A_{t,i},S_t)+\|T_{M_t}(X_{t,i},A_{t,i})-\bar{p}_t(X_{t,i},A_{t,i},S_t)\|_1)\mathbb{1}\left\{\bar{E}_t\right\}\right]\\ &\quad+h\sum_{i=1}^{h}\mathbb{E}_t\left[\beta_t(X_{t,i},A_{t,i},S_t)\right].\end{aligned}$$

For each episode $t$, we can use that $R_{M_t}(x,a)\mid H_t\sim\text{Beta}(\alpha^R_{t,S_t}(x,a))$ to yield

$$\begin{aligned}\mathbb{E}_t\left[(R_{M_t}(X_{t,i},A_{t,i})-\bar{r}_t(X_{t,i},A_{t,i},S_t))\mathbb{1}\left\{\bar{E}_t\right\}\right] &\leq\sum_{x,a}\int_{r=c_t(S_t,x,a)}^{\infty}r\mathbb{P}_t\left(R_{M_t}(x,a)-\bar{r}_t(x,a,S_t)=r\right)dr\\ &\leq\sum_{x,a}\mathbb{P}_t\left(R_{M_t}(x,a)-\bar{r}_t(x,a,S_t)\geq c_t(x,a,S_t)\right)\\ &\leq\sum_{x,a}\exp\left[-\frac{c_t(x,a,S_t)^2}{2/\left(4\left(\left\|\alpha^R_{t,S_t}(x,a)\right\|_1+1\right)\right)}\right]\\ &\leq 1/(2n)\,,\end{aligned}$$

where the second inequality uses that $R_{M_t}(x,a) \le 1$ and the third uses the sub-Gaussian parameter given in Lemma 4. Similarly, since $T_{M_t}(x,a) \mid H_t \sim \mathrm{Dir}(\alpha^T_{k,S_t}(x,a))$, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\|T_{M_t}(X_{t,i},A_{t,i}) - \bar{p}_t(X_{t,i},A_{t,i},S_t)\|_1 \mathbb{1}\{\bar{E}_t\}\right] \\
&\le |\mathcal{X}|\mathbb{E}_t\left[\max_x |T_{M_t}(X_{t,i},A_{t,i},x) - \bar{p}_t(X_{t,i},A_{t,i},x,S_t)|\,\mathbb{1}\{\bar{E}_t\}\right].
\end{aligned}$$

Now, using Lemma 4 for Dirichlet distributions, we have

$$\begin{aligned}
&\mathbb{E}_t\left[\max_x |T_{M_t}(X_{t,i},A_{t,i},x) - \bar{p}_t(X_{t,i},A_{t,i},x,S_t)|\,\mathbb{1}\{\bar{E}_t\}\right] \\
&\le \sum_{(x,a,x')} \int_{p=\phi_t(x,a,S_t)/\sqrt{|\mathcal{X}|}}^{\infty} p\mathbb{P}_t\left(|T_{M_t}(x,a,x') - \bar{p}_t(x,a,x',S_t)| = p\right) dp \\
&\le \sum_{(x,a,x')} 2\mathbb{P}_t\left(|T_{M_t}(x,a,x') - \bar{p}_t(x,a,x',S_t)| \ge \phi_t(x,a,S_t)/\sqrt{|\mathcal{X}|}\right) \\
&\le \sum_{(x,a,x')} 2\exp\left[-\frac{\phi_t(x,a,S_t)^2}{2|\mathcal{X}|/\left(4\left(\left\|\alpha^T_{k,S_t}(x,a)\right\|_1 + 1\right)\right)}\right] \\
&\le 1/(2n),
\end{aligned}$$

So, we can bound the first term of (6) by

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_t\left[V_{M_t}(\pi_t) - \overline{V}_t(S_t,\pi_t)\right]\right] \le |\mathcal{X}|h^2 + h\sum_{t=1}^{n}\sum_{i=1}^{h}\mathbb{E}_t\left[\beta_t(X_{t,i},A_{t,i},S_t)\right].$$

**Step 2.** For each episode $t$, we define $C_t$ as follows:

$$C_t = \left\{s \in \mathcal{S} : G_t(s) \le \sqrt{hN_t(s)\log n}\right\},$$

where $N_t(s) = \sum_{\ell=1}^{t-1}\mathbb{1}\{S_\ell = s\}$ is the number of times $s$ was sampled from the posterior and $G_t(s)$ is defined as

$$G_t(s) = \sum_{\ell=1}^{k-1}\mathbb{1}\{S_\ell = s\}\left(\overline{V}_t(\pi_t,s) - h\sqrt{2}\sum_{t=1}^{h}\beta_t(X_{\ell,t},A_{\ell,t},s) - \sum_{t=0}^{h-1}R_{\ell,t}\right).$$

We show that $S_* \in C_t$ holds with high probability for any episode via the following lemma.

**Lemma 7.** *For any episode $t$, $\mathbb{P}(S_* \notin C_t) \le 2Lhn^{-1}$.*

*Proof.* Fix $S_* = s$. We know that $s \in C_t$ occurs as long as $G_t(s)$ is not too large. Let us define $\mathcal{T}_{t,s} = \{\ell < t : S_\ell = s\}$ as the episodes where $s$ is sampled until episode $t$. We want to upper-bound $G_t(s)$ by a martingale with respect to history, then bound the probability that $G_t(s)$ is too large using Azuma's inequality for concentration of martingales.

Let us define

$$\begin{aligned}
\mathcal{E}_{t,i} = \big\{ &|\bar{r}_t(X_{t,i},A_{t,i},S_t) - R_{M_t}(X_{t,i},A_{t,i})| \le \sqrt{2}c_t(X_{t,i},A_{t,i},S_t), \\
&\|\bar{p}_t(X_{t,i},A_{t,i},S_t) - T_{M_t}(X_{t,i},A_{t,i})\|_1 \le \sqrt{2}\phi_t(X_{t,i},A_{t,i},S_t)\big\}
\end{aligned}$$

as the event that the mean reward and transition probabilities for episode $t$ of episode $k$ are not far from their posterior means. Let $\mathcal{E} = \cap_{t=1}^{n}\cap_{i=1}^{h}\mathcal{E}_{t,i}$ be the event that this holds for all episodes and steps and $\bar{\mathcal{E}}$ be the complement. We know that

$$\mathbb{P}\left(\bar{\mathcal{E}}\right) \le \sum_{t=1}^{n}\sum_{i=1}^{h}\sum_{s\in\mathcal{S}}\sum_{x,a}\mathbb{E}\left[\mathbb{P}_t\left(\mathcal{E}_{t,i}\right)\right] \le \sum_{t=1}^{n}\sum_{i=1}^{h}\sum_{s\in\mathcal{S}}\sum_{x,a}\left(|\mathcal{X}||\mathcal{A}|n\right)^{-2} \le Lhn^{-1},$$

where we use that we have $R_{M_t}(x,a)$ and $T_{M_t}(x,a)$ follow a Beta and Dirichlet distribution, respectively, which are sub-Gaussian from Lemma 4.

For episode $\ell \in \mathcal{T}_{t,s}$, let $Z_\ell = V_*(\pi_\ell) - \sum_{i=1}^h R_{\ell,t}$. Observe that $\mathbb{E}_\ell[Z_\ell] = 0$, so that $(Z_\ell)_{\ell \in \mathcal{T}_{t,s}}$ is a martingale difference sequence with respect to histories $(H_\ell)_{\ell \in \mathcal{T}_{t,s}}$. Also note that since $Z_\ell$ the sum of $h$ Bernoulli random variables and is therefore $\sigma^2$-sub-Gaussian with $\sigma^2 = h/4$. We know that conditioned on $H_\ell$,

$$\overline{V}_\ell(\pi_\ell, s) - h\sqrt{2}\sum_{t=1}^{h} \beta_\ell(X_{\ell,t}, A_{\ell,t}, s)\mathbb{1}\{\mathcal{E}_{\ell,t}\} - \sum_{t=0}^{h-1} R_{\ell,t} \leq V_*(\pi_\ell) - \sum_{t=0}^{h-1} R_{\ell,t} = Z_\ell$$

where we use Lemma 5 to bound $V_*(\pi_\ell) - \overline{V}_\ell(\pi_\ell, s)$. This implies that conditioned on $(H_\ell)_{\ell \in \mathcal{T}_{t,s}}$, we have

$$G_t(s)\mathbb{1}\{\mathcal{E}\} = \sum_{\ell \in \mathcal{T}_{t,s}} \left( \overline{V}_\ell(s, \pi_\ell) - h\sqrt{2}\sum_{t=1}^{h} \beta_\ell(s, X_{\ell,t}, A_{\ell,t})\mathbb{1}\{\mathcal{E}_{\ell,t}\} - \sum_{t=0}^{h-1} R_{\ell,t} \right) \leq \sum_{\ell \in \mathcal{T}_{t,s}} Z_\ell \,.$$

For any episode $t$, we have that $\mathcal{T}_{t,s}$ is a random quantity. First, we fix $|\mathcal{T}_{t,s}| = N_t(s) = u$ where $u < t$ and yield the following due to Azuma's inequality,

$$\mathbb{P}_t\left(G_t(s)\mathbb{1}\{\mathcal{E}\} \geq \sqrt{4(h/4)u\log n}\right) \leq \mathbb{P}\left(\sum_{\ell \in \mathcal{T}_{t,s}} Z_\ell \geq \sqrt{4(h/4)u\log n}\right) \leq \exp\left[-2\log n\right] = n^{-2}\,.$$

Finally, by the union bound, we have

$$\begin{aligned}
\mathbb{P}\left(S_* \notin C_t\right) &\leq \sum_{s \in \mathcal{S}} \sum_{u=1}^{t-1} \mathbb{P}\left(G_t(s) \geq \sqrt{hu\log n}\right) \\
&\leq \mathbb{P}\left(\bar{\mathcal{E}}\right) + \sum_{s \in \mathcal{S}} \sum_{u=1}^{t-1} \mathbb{P}\left(G_t(s)\mathbb{1}\{\mathcal{E}\} \geq \sqrt{hu\log n}\right) \leq 2Lhn^{-1}\,.
\end{aligned}$$

This completes the proof. □

**Step 4.** Following the sketch of Appendix B.2, we can rewrite the second term of (6) as

$$\mathbb{E}\left[\sum_{t=1}^{n} \mathbb{E}_t\left[\overline{V}_t(S_t, \pi_t) - V_{M_*}(\pi_t)\right]\right] \leq \mathbb{E}\left[\sum_{t=1}^{n} (\overline{V}_t(S_t, \pi_t) - V_{M_*}(\pi_t))\mathbb{1}\{S_t \in C_t\}\right] + h\sum_{t=1}^{n} \mathbb{P}\left(S_* \notin C_t\right)\,.$$

From Step 1 and 2, the second term is bounded as $2Lh$. Finally, the remaining term can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{n} (\overline{V}_t(S_t, \pi_t) - V_{M_*}(\pi_t)\mathbb{1}\{S_t \in C_t\}\right] \\
&= h\sqrt{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{t=1}^{h} \beta_t(X_{t,i}, A_{t,i}, S_t)\right] + \mathbb{E}\left[\sum_{t=1}^{n}\left(\overline{V}_t(S_t, \pi_t) - h\sqrt{2}\sum_{i=1}^{h} \beta_t(X_{t,i}, A_{t,i}, S_t) - \sum_{t=1}^{h} R_{t,i}\right)\mathbb{1}\{S_t \in C_t\}\right] \\
&\leq h\sqrt{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{t=1}^{h} \beta_t(X_{t,i}, A_{t,i}, S_t)\right] + \mathbb{E}\left[\sum_{s \in \mathcal{S}} G_{n+1}(s) + Lh\right] \\
&\leq h\sqrt{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{t=1}^{h} \beta_t(X_{t,i}, A_{t,i}, S_t)\right] + \sqrt{Lnh\log n} + Lh\,,
\end{aligned}$$

where we use that up until the last episode $t' = \max_{t \in [m]}\{S_t = s\}$ a latent state $s$ is sampled from the posterior, there is an upper-bound on its overestimation $G_{t'}(s)$.

Let $\Lambda_{0,s} = \min\{\min_{x,a} \left\|\alpha^R_{0,s}(x,a)\right\|_1, \min_{x,a} \left\|\alpha^T_{0,s}(x,a)\right\|_1\}$ represent at least how concentrated the reward and transition priors are for latent state $s$. Let $\Lambda_{0,\min} = \min_{s \in \mathcal{S}} \Lambda_{0,s}$ be the minimum over latent states. What remains

the bounding the sum of confidence widths, which is done in Lemma 6. Combining the regret due to both terms gives,

$$\begin{aligned}\mathcal{BR}(m) \leq 4|\mathcal{X}|h\sqrt{2|\mathcal{A}|nh\log(4|\mathcal{X}||\mathcal{A}|n)\log\left(1+\frac{nh}{2|\mathcal{X}||\mathcal{A}|\Lambda_{0,\min}}\right)} + 2|\mathcal{X}||\mathcal{A}|h^2 \\ + \sqrt{Lnh\log n} + 3Lh\,.\end{aligned}$$

□