Non-Stationary Latent Bandits

Joey Hong  Branislav Kveton  Manzil Zaheer  Yinlam Chow
Amr Ahmed  Mohammad Ghavamzadeh  Craig Boutilier
Google Research

Abstract

Users of recommender systems often behave in a non-stationary fashion, due to their evolving preferences and tastes over time. In this work, we propose a practical approach for fast personalization to non-stationary users. The key idea is to frame this problem as a latent bandit, where the prototypical models of user behavior are learned offline and the latent state of the user is inferred online from its interactions with the models. We call this problem a non-stationary latent bandit. We propose Thompson sampling algorithms for regret minimization in non-stationary latent bandits, analyze them, and evaluate them on a real-world dataset. The main strength of our approach is that it can be combined with rich offline-learned models, which can be misspecified, and are subsequently fine-tuned online using posterior sampling. In this way, we naturally combine the strengths of offline and online learning.

1 Introduction

When users interact with recommender systems or search engines, their behavior is often guided by a latent state, a context that cannot be observed. Examples of latent states are user preferences, which persist over longer periods of time, and shorter-term user intents. As the users interact, their latent state is slowly revealed by their responses. A good recommender should cater to the user based on the latent state, which first needs to be discovered.

We formalize the problem of recommending to a user under a changing latent state as a multi-armed bandit (Lai and Robbins, 1985; Auer, 2002; Lattimore and Szepesvári, 2019). In this setting, the recommender is a learning agent and its actions are the arms of a bandit. After an arm is pulled, the agent observes a response from the user, which is also its reward. The response is a function of the observed context and an unobserved latent state. The goal of the learning agent is to maximize its cumulative reward over n interactions with the user. The challenge is that the latent state of the user is unobserved and changes. This setting is known as piecewise-stationary bandits (Hartland et al., 2007; Garivier and Moulines, 2008; Yu and Mannor, 2009).

Both non-stationary bandits (Auer et al., 2002b; Luo et al., 2018) and the special case of piecewise-stationary bandits (Hartland et al., 2007; Garivier and Moulines, 2008; Yu and Mannor, 2009) have been studied extensively in prior work. The main departures in this work are two fold. First, we assume that the latent state changes stochastically. Second, we assume that the learning agent knows, at least partially, the reward models of arms conditioned on each latent state. This assumption is realistic in most recommender domains, where a plethora of offline data allow for rich models of user behavior, conditioned on the user type, to be learned offline. Under these assumptions, the problem of learning to act can be solved efficiently by Thompson sampling (TS) (Thompson, 1933; Chapelle and Li, 2012; Russo and Van Roy, 2013) over latent states, which we propose, analyze, and extensively evaluate. To the best of our knowledge, this is the first analysis of TS in this highly practical setting.

Our approach has many benefits over prior works. Unlike adversarial techniques (Auer et al., 2002b; Luo et al., 2018), we leverage the stochastic nature of the environment, which results in practical algorithms. Unlike stochastic algorithms, which either passively (Kocsis and Szepesvari, 2006; Garivier and Moulines, 2008) or actively (Yu and Mannor, 2009; Mellor and Shapiro, 2013; Cao et al., 2019) adapt to the environment, our algorithms never forget the past or reset their model. In a sense, our approach is the most natural technique under the assumption of knowing, at least partially, the model of the environment. This assumption is natural in any domain where a plethora of offline data is available and leads to major gains over prior work.

Our paper is organized as follows. In Section 2, we introduce our setting of non-stationary latent bandits. In Section 3, we
propose two posterior sampling algorithms: one knows the exact model of the environment and the other knows a prior distribution over potential models. In Section 4, we derive gap-free bounds on the $n$-round regret of both algorithms. The algorithms are evaluated in Section 5. Finally, we discuss related work in Section 6 and conclude in Section 7.

2 Setting

We adopt the following notation. Random variables are capitalized. Greek letters denote parameters and we explicitly state beforehand when they are random. The set of arms is $\mathcal{A} = [K]$, the set of contexts is $\mathcal{X}$, and the set of latent states is $\mathcal{S}$, with $|\mathcal{S}| \ll K$.

The latent bandit (Maillard and Mannor, 2014) is an online learning problem, where the learning agent interacts with an environment over $n$ rounds as follows. In round $t \in [n]$, the agent observes context $X_t \in \mathcal{X}$, chooses action $A_t \in \mathcal{A}$, then observes reward $R_t \in \mathbb{R}$. The random variable $R_t$ depends on the context $X_t$, action $A_t$, and latent state $S_t \in \mathcal{S}$. The history up to round $t$ is $\mathcal{H}_t = (X_1, A_1, R_1, \ldots, X_{t-1}, A_{t-1}, R_{t-1})$.

The policy of the agent in round $t$ is a mapping from its history $\mathcal{H}_t$ and context $X_t$ to the choice of action $A_t$. In prior work (Maillard and Mannor, 2014; Zhou and Brunskill, 2016; Hong et al., 2020), the latent state is assumed to be constant over all rounds, which we relax in this work.

The reward is sampled from a conditional reward distribution, $P(\cdot \mid A, X, S; \theta)$, which is parameterized by reward model parameters $\theta \in \Theta$, where $\Theta$ is the space of feasible reward models of the environment. Let $\mu(a, x, s; \theta) = \mathbb{E}_{R \sim P(\cdot \mid a, x, s; \theta)}[R]$ be the mean reward of action $a$ in context $x$ and latent state $s$ under model $\theta$. We assume that the rewards are $\sigma^2$-sub-Gaussian with variance proxy $\sigma^2$,

$$\mathbb{E}_{R \sim P(\cdot \mid a, x, s; \theta)}[\exp\{\lambda(R - \mu(a, x, s; \theta))\}] \leq \exp \left[\frac{\sigma^2 \lambda^2}{2}\right],$$

for all $a, x, s$, and $\lambda > 0$. Note that we do not make strong assumptions about the form of the reward: $\mu(a, x, s; \theta)$ can be any complex function of $\theta$, and contexts can be generated by any arbitrary process.

In the non-stationary latent bandit, we additionally consider latent states that evolve over time. The initial latent state is drawn according to the prior distribution as $S_1 \sim P_1(s)$. Then, in round $t$, the underlying latent state $S_t \in \mathcal{S}$ evolves according to $S_t \sim P(\cdot \mid S_{t-1}; \phi)$, where $\phi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the transition matrix. The graphical model is shown in Figure 1. This is useful for applications in which user preferences, tasks, or intents change. For example, the latent states $\mathcal{S}$ could be different behavior modes that the user switches between over time.

Let $\theta_*, \phi_*$ be the true model parameters, so that the reward in round $t$ is sampled as $R_t \sim P(\cdot \mid A_t, X_t, S_t; \theta_*)$, and the next latent state is sampled as $S_{t+1} \sim P(\cdot \mid S_t; \phi_*)$. Note that the next round’s context and latent state are unaffected by the action chosen in the previous round. This is a specific case of POMDPs, where the actions taken by the agent do not affect the dynamics of the environment.

Performance of bandit algorithms is typically measured by regret. For a variable $X$, let $X_{i:j}$ denote its concatenation from rounds $i$ to $j$, inclusive. For a fixed latent state sequence $s_{1:n} \in \mathcal{S}^n$ and model $\theta_* \in \Theta$, let $A_{t, *} = \arg\max_{a \in \mathcal{A}} \mu(a, X_t, s_t, \theta_*)$ be the optimal arm. Then the expected $n$-round regret is defined as

$$R(n; \theta_*, s_{1:n}) = \mathbb{E} \left[ \sum_{t=1}^{n} \mu(A_{t, s}, X_t, s_t; \theta_*) - \mu(A_t, X_t, s_t; \theta_*) \right].$$

In this work, we consider the Bayes regret, which includes an expectation over latent state and/or model randomness.

We use two different notions of the Bayes regret. The first one is when the true model $\theta_*, \phi_*$ is fixed, and the expectation is only over randomness in latent states. The $n$-round Bayes regret with fixed model is

$$\text{BR}(n; \theta_*, \phi_*) = \mathbb{E}_{S_{1:n} \sim \phi_*} [R(n; \theta_*, S_{1:n}) \mid \theta_*, \phi_*]$$

where $A_{t, *} = \arg\max_{a \in \mathcal{A}} \mu(a, X_t, S_t; \theta_*)$ is also a function of the random latent state. We also study the case where the true model is sampled from prior $\theta_* \sim P_1$. Then $A_{t, 1}$ depends on the random latent state and model, and the $n$-round Bayes regret is

$$\text{BR}(n) = \mathbb{E} [\text{BR}(n; \theta_*, \phi_*)].$$

It is important to note that the Bayes regret is a weaker metric than regret, which is worst case over latent sequences and models. However, we are often more concerned in practice with the average performance over a range of latent state sequences and models, that arise with multiple users or multiple sessions with the same user. This quantity is sufficiently captured by Bayes regret.

3 Model-Based Thompson Sampling

Recall that in round $t$, we have that $S_t$ is the true latent state, and model parameters $\theta_*, \phi_*$ determine the conditional rewards $P(R_t \mid A_t, X_t, S_t; \theta_*)$ and transition probabilities $P(S_t \mid S_{t-1}; \phi_*)$, respectively.
Our proposed algorithm is Thompson sampling (TS) with an offline-learned model. At a high-level, the TS algorithm operates by sampling actions stochastically according to \( P(A_t | a, H_t, X_t) = \mathbb{P}(A_t = a | H_t, X_t) \). In Section 3.1, we consider a simple case where the true model is recovered offline. In most realistic scenarios though, the true model is unknown, and we only know its uncertain estimate. In Section 3.2, we consider an agnostic case, where only priors over the reward and transition models are known, that is \( \theta_s, \phi_s \) are recovered offline. In this case, TS reduces to sampling a belief state \( B_t \in S \) from posterior distribution over latent states, and acting according to \( B_t \) and model parameters. In particular, \( A_t = \arg \max_{a \in A} \mu(a, X_t, B_t; \theta_s) \). The pseudocode of TS is detailed in Algorithm 1. In (4), we compute the posterior as detailed in Algorithm 1. In (4), we compute the posterior as
\[
\arg \max_{s_t} P_t(s_t) = \mathbb{P}(s_t = s | H_t) \]. Since the model is known exactly, this can be computed as an incremental update from \( P_{t-1} \). Then, after sampling \( B_t \) from the posterior, the algorithm simply chooses the best-performing action from the conditional reward model for \( B_t \).

Algorithm 1 mTS

1: Input:
2: Model parameters \( \theta_s, \phi_s \)
3: Prior over initial latent state \( P_t(s) \)
4: for \( t \leftarrow 1, 2, \ldots \) do
5: Sample \( B_t \sim P_t \)
6: Select \( A_t \leftarrow \arg \max_{a \in A} \mu(a, X_t, B_t; \theta_s) \)
7: Observe \( R_t \). Update posterior
\[
P_{t+1}(s_{t+1}) \propto \sum_{s_t \in S} P_t(s_t) P(s_{t+1} | s_t; \phi_s) P(R_t | A_t, X_t, s_t; \theta_s)
\]

Algorithm 2 umTS

1: Input:
2: Prior over model parameters \( P_t(\theta), P_t(\phi) \)
3: Prior over initial latent state \( P_t(s) \)
4: Initialize \( P_t(s, \theta) \propto P_t(s_t)P_t(\theta) \)
5: for \( t \leftarrow 1, 2, \ldots \) do
6: Sample \( B_t, \theta \sim P_t \)
7: Select \( A_t \leftarrow \arg \max_{a \in A} \mu(a, X_t, B_t; \theta) \)
8: Observe \( R_t \). Update joint posterior
\[
P_{t+1}(s_{t+1}, \theta) \propto \int_{\phi} P_t(\theta, \phi) \sum_{s_{t+1} \in S_t} P(s_{t+1} | \phi) P(H_{t+1} | s_{t+1: t}; \theta) d\phi
\]

Note that the joint posterior in (5) requires a summation over past latent state trajectories and is therefore intractable. We propose and analyze Algorithm 2 as a computation-inefficient algorithm, but approximate it using sequential Monte Carlo (SMC) in practice (Doucet et al., 2013).

3.3 Approximate Inference for Uncertain Models

In this section, we propose and approximate SMC algorithm to umTS. Particularly, we use particle filtering with \( N \) particles (Doucet et al., 2013; Särkkä, 2013), where each particle maintains its own latent state trajectory. At round \( t \), particle \( i \) independently samples believed state and model \( B_{t}^{(i)}(s, \theta, \phi) \sim P_{t}^{(i)}(s, \theta, \phi) \), where joint posterior
\[
P_{t}^{(i)}(s, t, \theta, \phi) = \mathbb{P}(S_t = s, \theta_s = \theta, \phi_s = \phi | H_t, B_{t-1}^{(i)})
\]
additionally depends on the particle’s past latent trajectory.

For each round \( t \), the SMC algorithm maintains a weight \( w_t \) over particles, and acts according to the weighted average of the particles’ latent state and model parameters. The weights for all particles are updated using the incremental likelihood of the resulting observations in round \( t \) as in (6) and renormalized. If the current weights \( w_t \) satisfy a resampling criterion, then the filtering algorithm resamples \( N \) particles in proportion to their weights with replacement.
The algorithm is detailed in Algorithm 3.

**Algorithm 3 umTS (Particle Filtering)**

1: **Input:**
2: 1. Prior over model parameters $P_1(\theta)$, $P_2(\phi)$
3: 2. Prior over initial latent state $P_1(s)$
4: 3. Number of particles $N$
5: 4. Set $w_1 \in \mathbb{R}^N$ s.t. $w_1,i \sim N^{-1}, i = 1, \ldots, N$
6: 5. for $t \leftarrow 1, 2, \ldots$ do
7: 6. For particle $i$, sample $B_s^{(i)} \sim \psi(\theta^{(i)})$, $\theta^{(i)} \sim \phi(s,t)$
8: 7. Set $w_{t+1,i} \leftarrow w_{t,i} P(R_t A_t, X_t | B_t^{(i)}, \phi)$
9: 8. Compute $ESS \leftarrow \left(\sum_{i=1}^N w_{t+1,i}^2\right)^{-1}$. Resample particles if $ESS$ is too small.

For a matrix (vector) $M$, we let $M_i$ denote its $i$-th row (element). Using this notation, we can write $\theta = (\theta_s)_{s \in S}$ and $\phi = (\phi_s)_{s \in S}$ as vectors of conditional parameters, one for each latent state. We can show that the sampling step for each particle can be done tractably if the reward model prior $P_1(\theta_\ast)$ and likelihood $P(\theta \mid r, a, s, \theta)$ are conjugate distributions in the exponential family. The transition prior for each latent state $P_1(\phi)$ factors as Dirichlet for each state $s$, i.e., $\phi_{s,t} \sim \text{Dir}(\alpha_{s,t})$. The key detail is that now we can obtain samples from the joint posterior using the particles and avoid the intractable sum over all possible past trajectories as in (5) in Algorithm 2. For particle $i$, we decompose the joint posterior as

$$P_t^{(i)}(s_t, \theta, \phi) \propto P(\phi \mid B_{1:t-1}^{(i)}) P(s_t \mid B_{1:t-1}^{(i)}, \phi) P(\theta \mid \mathcal{H}_t, B_{1:t-1}^{(i)})$$

Here, sampling from the joint posterior can be done by first sampling the transition parameters, then beliefs latent state, and finally reward parameters for that state.

In the case where $\phi_{s,t} \sim \text{Dir}(\alpha_{s,t})$ is Dirichlet with parameters $\alpha_{s,t} \in \mathbb{R}^S$, the posterior is also Dirichlet. For particle $i$, the posterior parameters are simply updated with the observed transitions in its latent state trajectory $B_{1:t-1}^{(i)}$. Formally, the posterior over state transitions from state $s$ would be:

$$\phi_{s}^{(i)} \mid B_{1:t-1}^{(i)} \sim \text{Dir} \left( \alpha_s + \sum_{t=1}^{t-1} \mathbb{1} \{ B_{t-1}^{(i)} = s \}, \mathbb{1} \{ B_{t}^{(i)} = s' \} \right)_{s' \in S}.$$  

The transition matrix $\phi^{(i)}$ can be tractably sampled from this Dirichlet posterior. The next latent state $B_t^{(i)}$ is easily sampled from $\phi^{(i)}$.

Recall that we assumed that the reward model prior and conditional reward distribution belong to the exponential family, which covers commonly studied reward distributions, such as Gaussian and Bernoulli. We assume that the reward likelihood is written

$$P(r \mid a, x, s; \theta) = \exp \left[ f(r, a, x)^T \kappa(\theta_s) - g(\theta_s) \right],$$

where $f(r, a, x)$ are sufficient statistics for the observed data, $\kappa(\theta_s)$ are the natural parameters, and $g(\theta_s)$ is the log-partition function. Then, the prior over $\theta_s$ is the conjugate prior of the likelihood, which has the general form of

$$P_1(\theta_s) \propto \exp \left[ \psi_s \kappa(\theta_s) - m_s g(\theta_s) \right],$$

where $\psi_s, m_s$ are parameters controlling the prior and $H(\psi_s, m_s)$ is the normalizing factor.

For particle $i$, round $t$, and state $s$, updating the posterior over $\theta_s$ simply involves updating the prior parameters with sufficient statistics from the data. Specifically, we have $m_s^{(i)} = m_s^{(i)} + \sum_{t=1}^{t-1} \mathbb{1} \{ B_{t}^{(i)} = s \}$ and

$$\psi_s^{(i)} \leftarrow \psi_s^{(i)} + \sum_{t=1}^{t-1} \mathbb{1} \{ B_{t}^{(i)} = s \} f(R_t A_t, X_t),$$

form the conditional posterior

$$P(\theta_s \mid \mathcal{H}_t, B_{1:t-1}^{(i)}) \propto \exp \left[ \psi_s^{(i)} \kappa(\theta_s) - m_s^{(i)} g(\theta_s) \right].$$

Hence, each term in the joint posterior decomposition in (7) has an analytic form, and can be tractably sampled from.

### 4 Analysis

In this section, we derive Bayes regret bounds for mTS and umTS. Recall that $A_{t,s}$ is the optimal action in round $t$. The key idea in our analysis is that the conditional distributions of $A_{t,s}$ and $A_t$, as sampled in mTS, are identical. Formally, $E[f(A_{t,s}) \mid X_t, \mathcal{H}_t] = E[f(A_t) \mid X_t, \mathcal{H}_t]$ for any function $f$ of history $\mathcal{H}_t$ and context $X_t$. Following Russo and Van Roy (2013), we design $f$ as an upper confidence bound (UCB) in a suitable UCB algorithm. In Section 4.1, we first propose that algorithm, then, in Section 4.2, we state a key regret decomposition and show how to derive Bayes regret bounds for our algorithms using the UCB algorithm. In Section 4.3, we present our regret bounds.

#### 4.1 Model-Based UCB

In this section, we propose Sw-mUCB, a model-based sliding-window UCB algorithm that uses an offline-learned model to identify non-stationary latent states. In the domain of
non-stationary bandits, Kocsis and Szepesvari (2006) and Garivier and Moulines (2008) proposed two passive adaptations to the UCB algorithm: discounting past observations or ignoring them using a sliding window. Without loss of generality, we focus on the latter due to being better suited for abrupt changes in latent state (as opposed to gradual ones). The algorithm is similar to that proposed by Maillard and Mannor (2014) and Hong et al. (2020) for stationary environments, but augmented with an additional sliding window. The novelty is that the sliding window allows for sublinear regret when the environment is non-stationary.

SW-mUCB is detailed in Algorithm 4. At a high level, it takes model parameters \( \theta_\star \) as an input. We discuss how to change SW-mUCB when \( \theta_\star \) is not known in the Appendix. SW-mUCB maintains a set of latent states \( C_t \) consistent with the rewards observed in the most recent \( \tau \) rounds, where \( \tau \) is a tunable parameter. In round \( t \), it chooses a belief state \( B_t \) from \( C_t \) and the arm \( A_t \) with the maximum expected reward in that state, \( (B_t, A_t) = \arg \max_{s \in C_t, a \in A} \mu(a, X_t, s; \hat{\theta}) \).

In SW-mUCB, the UCB for action \( a \) in round \( t \) is
\[
U_t(a) = \arg \max_{s \in C_t} \mu(a, X_t, s; \hat{\theta}).
\] (8)

The consistent latent states are determined by “gap” \( G_t(s) \), defined in (10). If \( G_t(s) \) is high, SW-mUCB marks state \( s \) as inconsistent and does not consider it in estimating UCB \( U_t \).

### 4.2 Regret Decomposition

Note that for any action \( a \in A \), the upper confidence bound \( U_t(a) \) in (8) is deterministic given \( X_t \) and \( H_t \). This observation leads to the following regret decomposition.

**Proposition 1.** The Bayes regret of mTS decomposes
\[
BR(n; \theta_\star, \phi_\star) = E \left[ \sum_{t=1}^{n} \mu(A_{t,s}, X_t, S_t; \theta_\star) - U_t(A_{t,s}) \mid \theta_\star, \phi_\star \right] + E \left[ \sum_{t=1}^{n} U_t(A_t) - \mu(A_t, X_t, S_t; \theta_\star) \mid \theta_\star, \phi_\star \right].
\] (9)

The proof is due to Russo and Van Roy (2013), and follows from rewriting the Bayes regret in terms of \( U_t \) and the observation above. Note that while we use the fixed-model formulation of the Bayes regret in (2), the proposition still holds for general (3).

Hence, though the UCBs \( U_t \) are not used by our TS algorithms, they can be used to analyze them due to the decomposition in (9). Specifically, our derivation of a Bayes regret bound for mTS proceeds according to the outline below.

**Step 1:** \( S_t \in C_t \) with high probability. We show that the true latent state is in our consistent sets with a high probability. This means that the first term in (9) is small.

**Algorithm 4 SW-mUCB**

1: Input: Model parameters \( \theta_\star \), window size \( \tau \)
2: for \( t \leftarrow 1, 2, \ldots \) do
3: Define \( N_t(s) = \sum_{\ell=\max(1, t-\tau)}^{t-1} \mathbb{I} \{ B_\ell = s \} \) and \( G_t(s) = \sum_{\ell=\max(1, t-\tau)}^{t-1} \mathbb{I} \{ B_\ell = s \} (\mu(A_t, X_t, s; \theta_\star) - R_t) \) (10)
4: Set of consistent latent states \( C_t \leftarrow \{ s \in S : G_t(s) \leq \sigma \sqrt{6N_t(s) \log n} \} \)
5: Select \( B_t, A_t \leftarrow \arg \max_{s \in C_t, a \in A} \mu(a, X_t, s; \theta_\star) \)

**Step 2:** Regret bound for SW-mUCB. This follows from bounding both terms in (9). The second term is the sum of confidence widths over time, or difference between \( U_t \) and the true mean reward. The widths decrease, under appropriate conditions, whenever an arm is pulled.

**Step 3:** Bayes regret bound for mTS. We exploit the fact that the Bayes regret decomposition for mTS in (9) can be equivalently stated for the regret of SW-mUCB. Hence, any UCB regret bound transfers to a TS Bayes regret bound.

For Step 3 to hold, our analysis in Step 2 needs to be worst-case over suboptimal latent states and actions. This is why we cannot use the fact that actions \( A_t \) maximize \( U_t \) in (8), and derive gap-free bounds.

### 4.3 Regret Bounds

In this section, we state Bayes regret bounds for mTS with known model and unTS with uncertain model. As described in Section 4.2, our bounds follow from that on SW-mUCB and Proposition 1. That bound is stated below in terms of the number of stationary segments in a horizon of \( n \) rounds, \( L = \sum_{t=2}^{n} \mathbb{I} \{ s_t \neq s_{t-1} \} + 1 \). We defer proofs of all claims to Appendix.

**Lemma 1.** For known model parameters \( \theta_\star \) with \( \hat{\theta} = \theta_\star \), and optimal choice of \( \tau \), the \( n \)-round regret of SW-mUCB is
\[
R(n; \theta_\star, s_{1:n}) = \mathcal{O} \left( n^{2/3} \sqrt{|S| L \log n} \right).
\]

Prior derivations for sliding-window UCB without context achieved a gap-dependent bound of \( \mathcal{O}(K \sqrt{nL/\Delta^2}) \) (see Garivier and Moulines 2008, Theorem 7) after tuning \( \tau \), where \( K \) is the number of arms. A gap-free bound can be obtained by bounding \( \mathcal{O}(n^{1-\varepsilon}) \) gaps trivially. This yields a \( \mathcal{O}(n^{5/6}) \) regret bound, which is worse than Lemma 1.

In practice, the latent state sequence, and hence the number of stationary segments \( L \), is often stochastic. Given \( \phi_\star \), let \( p = 1 - \min_{s \in S} P(s \mid s; \phi_\star) \) be the maximum probability
of a change occurring. We can bound the expected value of $L$ from above by $1 + pn$. This yields the following Bayes regret bound for mTS.

**Theorem 1.** For known model parameters $\theta_*, \phi_*$, let $p = 1 - \min_{s \in S} P(s \mid s; \phi_*)$ with $L = 1 + pn$. Then, the $n$-round Bayes regret of mTS is

$$\text{BR}(n; \theta_*, \phi_*) = \mathcal{O}(n^{2/3} \sqrt{|S|L \log n})$$.

Note that recent non-stationary bandit algorithms with active change-point detection have $\mathcal{O}(\sqrt{KL})$ regret bounds (Yu and Mannor, 2009; Cao et al., 2019), where $K$ is the number of arms. However, such change-point detectors do not easily generalize to scenarios with context, and require knowledge of $n, L$ to tune their hyperparameters optimally. Our algorithm mTS handles context and does not require any parameter tuning. SW-mUCB is simply a tool to construct $U_i$ and analyze mTS; better algorithms may exist that yield tighter regret bounds for mTS. Also, while the expected number of stationary segments $\tilde{L} = pn$ appears linear in $n$, all prior works essentially assume $p = \mathcal{O}(1/n)$ by treating the number of stationary segments as a constant. Since changes are rare in many realistic applications, it is safe to assume that $\tilde{L} = \mathcal{O}(n^\delta)$, for some small $\beta > 0$.

Our next result is for umTS when only a prior over the reward and transitions is known. Our statement changes in two ways: (i) we introduce a high-probability error $\varepsilon$ in estimating the reward via a sample from the prior, and (ii) the expected number of changes $\tilde{L}$ depends on the transition prior. Recall that for any latent state $s$, we assume that the transition model $p_{\alpha,s}$ is sampled as $\phi_{s,s'} \sim \text{Dir}(\alpha_{s,s'} s' \in S)$. We define $\bar{\mu}(a, x, s) = \int_{\theta} \mu(a, x, s; \theta) P_1(\theta) d\theta$ as the mean conditional reward, marginalized with respect to $\theta$.

**Theorem 2.** Let $(\alpha_{s,s'})_{s,s' \in S \times S}$ be the prior parameters of $P_1(\phi)$, such that $\phi_{s,s'} \sim P_1(\phi)$ factors over state $s$ as $\phi_{s,s'} \sim \text{Dir}(\alpha_{s,s'} s' \in S)$. Let $p = 1 - \min_{s \in S} \alpha_{s,s}/\sum_{s' \in S} \alpha_{s,s'}$ and $L = 1 + pn$. For $\theta_*, P_1(\theta)$, choose $\varepsilon, \delta > 0$ such that

$$\{\forall a \in A, x \in X, s \in S : |\mu(a, x, s) - \bar{\mu}(a, x, s; \theta_*)| \leq \varepsilon\}$$

holds with probability at least $1 - \delta$. Then, the $n$-round Bayes regret of umTS is

$$\text{BR}(n) = \mathcal{O}\left(\delta n + \varepsilon n + n^{2/3} \sqrt{|S|L \log n}\right).$$

The bound in Theorem 2 has two linear terms in $n$, with $\delta$ and the high-probability error $\varepsilon$. Because the posterior over models is updated online, $\varepsilon$ should decrease as more rounds are observed online, meaning our bound is overly conservative. Nevertheless, some offline model-learning methods, such as tensor decomposition (Anandkumar et al., 2014), yield $\varepsilon = \mathcal{O}(1/\sqrt{n})$ for an offline dataset of size $n$. Thus our bound is not vacuous. We can formally relate $\varepsilon$ and $\delta$ using the tails of the conditional reward distributions. Let $\mu(a, x, s; \theta) - \bar{\mu}(a, x, s)$ be $v^2$-sub-Gaussian for all $a$, $x$, and $s$, where the random quantity is $\theta \sim P_1$. Then for any $\delta > 0$, we have that $\varepsilon = \mathcal{O}(\sqrt{\log(K|X||S|/\delta)})$ satisfies the conditions on $\varepsilon$ and $\delta$ needed for Theorem 2.

Among non-stationary contextual bandit algorithms, Exp4.S has near-optimal regret of $\mathcal{O}(\sqrt{S|n|L})$ for $|S|$ experts, when $L$ is known, and $\mathcal{O}(\sqrt{|S|nL})$, otherwise (Luo et al., 2018). Note that the tightness of our Bayes regret bound is limited by the sliding-window algorithm SW-mUCB. Though conceptually simple and able to yield sublinear regret, SW-mUCB likely yields a conservative Bayes regret bound. In addition, because our algorithms naturally leverage the stochasticity of the environment, we significantly outperform near-optimal algorithms, like Exp4.S, empirically. We demonstrate this in Section 5.

## 5 Experiments

In this section, we evaluate our algorithms on both synthetic and real-world datasets. We compare the following methods: (i) **CD-UCB**: UCB/LinUCB (Auer et al., 2002a; Abbasi-Yadkori et al., 2011) with a change-point detector as in Cao et al. (2019); (ii) **CD-TS**: TS/LinTS (Agrawal and Goyal, 2013; Abeille and Lazaric, 2016) with the same change-point detector; (iii) **ExpS**: Exp3.S/Exp4.S using offline reward model as experts, where each expert takes the best action as measured by its conditional reward model (Auer et al., 2002b; Luo et al., 2018); (iv) **mTS, umTS**: our proposed TS algorithms mTS, umTS.

In contrast to our method, the first two baselines do not use an offline model, but augment traditional bandit algorithms with a change-point detector that resets the algorithm when a change is detected. When there is no context, Cao et al. (2019) proposed a detector with near-optimal guarantees and state-of-the-art empirical performance. The last baseline modifies adversarial algorithms Exp3/Exp4 by enforcing a lower-bound on the expert weights; this has near-optimal regret in piecewise-stationary bandits (Auer et al., 2002b).

### 5.1 Synthetic Experiments

We artificially create a non-stationary multi-armed bandit without context, with $A = [5]$ and $S = [5]$. Mean rewards for are sampled uniformly at random $\mu(a, s) \sim \text{Uniform}(0, 1)$ for each $a \in A, s \in S$. Rewards are drawn i.i.d. from $P(c \mid a, s) = \mathcal{N}(c ; \mu(a, s), \sigma^2)$ with $\sigma = 0.5$. We use a horizon of $n = 2000$ as a primary application we are concerned with is fast personalization.

For CD-UCB, CD-TS, we use a change-point detector that computes the sum of the rewards for each arm in the past $\tau/2$ rounds, and the $\tau/2$ rounds before that. If the absolute value of their difference is greater than a threshold $b$, a change is detected. Following Cao et al. (2019), the window length parameter was tuned to $\tau = 100$ to minimize regret, and the threshold is chosen to be $b = \sigma \sqrt{\tau \log(2|A|n^2)/2}$. 
We also assess the performance of our algorithms on the MovieLens dataset, which contains ratings from users for movies. The dataset is divided into training and test sets, and we use the test set as our "offline" training set, and use the remaining users and movies rated by at least 200 users as our test set, giving sparse ratings matrices. We then use the remaining users and movies to create a set of diverse movies. Context matrices are used to cluster users into 5 clusters. Motivated by prior work, we create a "superuser" by randomly sampling 5 users, one from each cluster; for latent state $s$, the superuser behaves according to the user $i_s$. Note that different superusers will have a different set of behavior modes, which is often true in practice. The transition matrix that governs the dynamics of the superuser is given by the linear combination $P(s' | s; \phi_s) = 0.9J(s, s') + 0.1K(s, s')$, where $J(s, s') = 1 - p$ if $s' = s$ and $p/(|S| - 1)$ otherwise, and $K(s, s') \propto \exp(||U_{i_s} - U_{i_s}||_2^2)$. Here $J$ is used to ensure changes are infrequent, and $K$ to make transitions to similar latent states more likely. We let $p = 0.9975$ so that changes occur roughly every 400 rounds with $n = 2000$.

A run of a non-stationary contextual bandit proceeds as follows. A superuser $i_1, \ldots, i_5$ is sampled at random as described above. In each round, a latent state $S_t$ is generated according to $S_{t-1}$ and the transition matrix. Then, 20 genres, then a movie for each genre, are both uniformly sampled from the set of all genres, movies, respectively, creating a set of diverse movies. Context $X_t \in \mathbb{R}^{20 \times 20}$ is a matrix where the rows are the training feature vectors of the sampled movies, that is movie $j$ has a vector $\hat{V}_j$. The agent chooses among movies in $X_t$. The reward for recommending movie $j$ to the superuser under state $S_t = s$ is drawn from $R_t \sim \mathcal{N}(\hat{U}_j^\top V_j, 0.25)$, the product of the test user and movie vectors as its mean. Note that both $U$ and $V$ are unknown to the learning agent.

Our baselines CD-LinUCB, CD-LinTS are given movie vectors from the training set as context, and need to only learn the user vector. We could not find prior work that performed change detection in linear bandits, so we propose an adaptation of the one by Cao et al. (2019) to the linear case. Specifically, for round $t$ and window size $\tau$, the detector computes
We evaluate on We learn a model “offline” in the same way as the true model. The weighted norm is given by $\|\hat{W}_t - W^*\|_{\Sigma_t} \leq b$; for matrix $M$ and weights $\Sigma$, the weighted norm is given by $\|M\|_{\Sigma} = \sqrt{MT\Sigma M}$. Here both $\tau$ and $b$ are tuned. We maximize reward during evaluation.

We learn a model “offline” in the same way as the true model is constructed, except using the training set. Our offline model consists of $k$ clusters of users derived from $k$-means clustering on users $U$ in the training set. For each latent state, the prior given to our algorithms is a Gaussian prior with clustering on users $U$ in the training set. For each latent state, $\phi_s, \Sigma_s$ are tuned.

We evaluate on 100 superusers, and show the mean reward in Figure 4. Again, the model-based algorithms outperform finely-tuned baselines by a significant margin, especially in the short horizon. Since the offline model is misspecified due to the train-test split, umTS improves upon mTS in the long term, as it refines its model parameters online.

## 6 Related Work

### Non-stationary Bandits

This topic has been studied extensively (Kocsis and Szepesvari, 2006; Garivier and Moulines, 2008; Auer et al., 2002b). First works adapted to changes passively by weighting rewards, either by exponential discounting (Kocsis and Szepesvari, 2006) or by considering recent rewards in a sliding window (Garivier and Moulines, 2008). The latter yields a $O(K\sqrt{nL}/\Delta^2)$ gap-dependent bound when $L$ is known. In the adversarial setting (Auer et al., 2002b; Auer, 2003), adaptation can be achieved by bounding the weights of experts from below. This leads to $O(\sqrt{n|S|L})$ gap-free switching regret, where $|S|$ is the number of experts. Besbes et al. (2014) periodically reset a base bandit algorithm and attain $O(n^{2/3} V_T^{1/3})$ regret, where $V_T$ is the total variation under smooth changes. Other works monitor reward distributions and reset the bandit algorithm when a change is detected (Yu and Mannor, 2009; Liu et al., 2018). Mellor and Shapiro (2013) proposed augmenting Thompson sampling with a Bayesian change-point detector, but provide no regret guarantee. Cao et al. (2019) proposed a simple near-optimal change-point detector that yields $O(\sqrt{nKL})$ regret. In linear bandits, several recent papers studied passive adaptation of UCB algorithms (Cheung et al., 2019; Russac et al., 2019; Zhao et al., 2020). This yields $O(n^{2/3} P_T^{1/3})$ regret, where $P_T$ measures the total variation in an unknown weight vector. Luo et al. (2018) provided several contextual algorithms with similar regret to ours, with the best algorithm matching the Exp4.S bound of $O(\sqrt{n|S|L})$. All above methods forget the past, discount it, or are adversarial. This is a major drawback when the environment changes in a structured manner.

### Latent Bandits

Our work is also related to latent bandits (Maillard and Mannor, 2014; Zhou and Brunskill, 2016). Here the latent state is fixed across rounds and algorithms compete with standard bandit strategies, such as UCB (Auer et al., 2002a; Abbasi-yadkori et al., 2011) or Thompson sampling (Agrawal and Goyal, 2013; Abeille and Lazaric, 2016). Maillard and Mannor (2014) derived UCB algorithms in the multi-armed case without context under the extremes when the mean conditional rewards are either known or need to be estimated completely online. Zhou and Brunskill (2016) extended it to contextual bandits where policies are learned offline and selected online using Exp4. Bayesian policy reuse (BPR) (Rosman et al., 2016) selects offline-learned policies by maintaining a belief over the optimality of each policy, but no regret analysis exists. Recently, Hong et al. (2020) proposed and analyzed TS algorithms with complex offline-learned models. Our work is the first to extend latent bandits to non-stationary environments by considering a latent state that evolves according to a transition model, which is known or sampled from a known prior.

## 7 Conclusions

We study non-stationary latent bandits, where the conditional rewards depend on an evolving discrete latent state. Given the plethora of rich offline models, we consider a setting where an offline-learned model can be used naturally by Thompson sampling to identify the latent state online. Prior algorithms for non-stationary bandits adapt by forgetting the past, discounting it, or are adversarial. We avoid this by leveraging the stochastic latent structure of our problem and thus can outperform prior works empirically by a large margin. Our approach is contextual, aware of uncertainty, and we analyze it by a reduction to a sliding-window UCB algorithm. Though our analysis is conservative, our work can be viewed as a stepping stone for analyzing the Bayes regret of Thompson sampling in more complex graphical models than a single fixed latent state (Maillard and Mannor, 2014; Zhou and Brunskill, 2016; Hong et al., 2020).
References

Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Neural Information Processing Systems*, 2011.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Electronic Journal of Statistics*, 2016.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, 2013.

Anima Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. In *Journal of Machine Learning Research*, 2014.

Peter Auer. Finite-time analysis of the multiarmed bandit problem. In *Machine Learning*, 2002.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. In *Journal of Machine Learning Research*, 2003.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. In *SIAM journal on computing*, 2002b.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, 2014.

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *International Conference on Artificial Intelligence and Statistics*, 2019.

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

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Learning to optimize under non-stationarity. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Merlise Clyde and Edward I. George. Model uncertainty. *Statistical Science*, 2004.

Arnaud Doucet, Neil Gordon, and Nando de Freitas. *Sequential Monte Carlo Methods in Practice*. Springer New York, 2013.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. In *International Conference on Algorithmic Learning Theory*, 2008.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. In *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2015.

Cédric Hartland, Nicolas Baskiotis, Sylvain Gelly, Michèle Sebag, and Olivier Teytaud. Change point detection and meta-bandits for online learning in dynamic environments. 2007.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amir Ahmed, and Craig Boutilier. Latent bandits revisited. *CoRR*, abs/2006.08714, 2020.

Levente Kocsis and Csaba Szepesvari. Discounted ucb. In *2nd PASCAL Challenges Workshop*, 2006.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. In *Advances in applied mathematics*, 1985.

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

Fang Liu, Joohyun Lee, and Ness B. Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *AAAI Conference on Artificial Intelligence*, 2018.

Haipeng Luo, Chen-Yu Wei, Alekh Agarwal, and John Langford. Efficient contextual bandits in non-stationary worlds. In *Conference on Learning Theory*, 2018.

Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. In *International Conference on Machine Learning*, 2014.

Joseph Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change detection. In *International Conference on Artificial Intelligence and Statistics*, 2013.

Benjamin Rosman, Majd Hawasly, and Subramanian Ramamoorthy. Bayesian policy reuse. In *Machine Learning*, 2016.

Yoan Russac, Claire Vernade, and Olivier Cappé. Weighted linear bandits for non-stationary environments. In *Neural Information Processing Systems*, 2019.

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

Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. *Neural Information Processing Systems*, 2008.

Simo Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

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

Qingyun Wu, Naveen Iyer, and Hongning Wang. Learning contextual bandits in a non-stationary environment. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018.

Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *International Conference on Machine Learning*, 2009.
Peng Zhao, Lijun Zhang, Yuan Jiang, and Zhi-Hua Zhou. 
A simple approach for non-stationary linear bandits. In 
*International Conference on Artificial Intelligence and 
Statistics*, 2020.

Li Zhou and Emma Brunskill. Latent contextual bandits 
and their application to personalized recommendations 
for new users. In *International Joint Conferences on 
Artificial Intelligence*, 2016.
A Proofs

Our proofs rely on the following concentration inequality, which is a straightforward extension of the Azuma-Hoeffding inequality to sub-Gaussian random variables. This was used and proved by Hong et al. (2020).

**Proposition 2.** Let \((Y_t)_{t \in [n]}\) be a martingale difference sequence with respect to filtration \((\mathcal{F}_t)_{t \in [n]}\), that is \(\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = 0\) for any \(t \in [n]\). Let \(Y_t \mid \mathcal{F}_{t-1}\) be \(\sigma^2\)-sub-Gaussian for any \(t \in [n]\). Then for any \(\varepsilon > 0\),

\[
\mathbb{P} \left( \sum_{t=1}^{n} Y_t \geq \varepsilon \right) \leq 2 \exp \left( -\frac{\varepsilon^2}{2n\sigma^2} \right).
\]

A.1 Proof of Lemma 1

Recall that we have the following fixed quantities: true reward parameters \(\theta_s\), latent state sequence \(s_{1:n}\), and the number of stationary segments \(L\). Note that we can decompose the regret as

\[
\mathcal{R}(n; \theta_s, s_{1:n}) = \mathbb{E} \left[ \sum_{t=1}^{n} (\mu(A_t, X_t, s_t; \theta_s) - \mu(A_t, X_t, s_t; \theta_s)) \right]
= \mathbb{E} \left[ \sum_{t=1}^{n} (\mu(A_t, X_t, s_t; \theta_s) - U_t(A_t)) \right] + \mathbb{E} \left[ \sum_{t=1}^{n} (U_t(A_t) - \mu(A_t, X_t, s_t; \theta_s)) \right]
\leq \mathbb{E} \left[ \sum_{t=1}^{n} (\mu(A_t, X_t, s_t; \theta_s) - U_t(A_t)) \right] + \mathbb{E} \left[ \sum_{t=1}^{n} (U_t(A_t) - \mu(A_t, X_t, s_t; \theta_s)) \right].
\]

This is because for all rounds \(t \in [n]\), we choose \(A_t = \arg \max_{a \in A} U_t(a)\), which means \(U_t(A_t) \geq U_t(A_{t,s})\).

Let \(T\) be a set of all rounds \(t\) that are not close to any change-point, that is \(s_t = s_{t-1}\) for all \(t \in \{t - \tau + 1, \ldots, t\}\). Note that this includes all rounds where the last \(\tau\) rounds have the same latent state as that round. Let

\[
E_t = \left\{ \forall s \in \mathcal{S} : \sum_{t=1}^{n} 1 \{B_t = s\} (\mu(A_t, X_t, s_t; \theta_s) - R_t) \leq \sigma \sqrt{6N_t(s) \log n} \right\}
\]

be the event that the total realized reward under each played latent state is close to its expectation. Let \(E = \cap_{t \in T} E_t\) be the event that this holds for all rounds not close to a change-point, and \(\bar{E}\) be its complement. Then we can bound the expected \(n\)-round regret as

\[
\mathcal{R}(n; \theta_s, s_{1:n}) \leq L \tau + \mathbb{E} \left[ \sum_{t \in T} (\mu(A_t, X_t, s_t; \theta_s) - U_t(A_t)) \right]
= L \tau + \mathbb{E} \left[ \mathbb{1}\{\bar{E}\} \sum_{t \in T} (\mu(A_t, X_t, s_t; \theta_s) - \mu(A_t, X_t, s_t; \theta_s)) \right] + \mathbb{E} \left[ \mathbb{1}\{E\} \sum_{t \in T} (\mu(A_t, X_t, s_t; \theta_s) - \mu(A_t, X_t, s_t; \theta_s)) \right]
\leq L \tau + \mathbb{E} \left[ \mathbb{1}\{\bar{E}\} \sum_{t \in T} (\mu(A_t, X_t, s_t; \theta_s) - \mu(A_t, X_t, s_t; \theta_s)) \right]
+ \mathbb{E} \left[ \mathbb{1}\{E\} \sum_{t \in T} (U_t(A_t) - \mu(A_t, X_t, s_t; \theta_s)) \right],
\]

where for the first inequality we upper bound the regret in rounds close to change-points by \(L \tau\), and in the second we use the regret decomposition in (11). We ignore the rounds within \(\tau\) rounds of change-points because the empirical mean reward estimates over the those rounds are biased.

We first show that the probability of \(\bar{E}\) occurring is low. Without context, this would follow immediately from Hoeffding’s inequality. Since we have context generated by some random process, we instead turn to martingales.
Proposition 3. Let $E_t$ be defined as in (12) for all rounds $t$, $E = \cap_{t \in T} E_t$, and $\bar{E}$ be its complement. Then $\Pr (\bar{E}) \leq 2 |S| n^{-1}$.

Proof. Because the UCBs depend on which latent states are eliminated, the UCBs depend on the history, and the conditional action given observed context also depends on the history. For each latent state $s$ and round $t$, let $T_{t,s}$ be the rounds where state $s$ was chosen among the past $\tau$ rounds. For round $\ell \in T_{t,s}$, let $Y_{\ell}(s) = \mu(A_{\ell}, X_{\ell}, s; \theta_s) - R_{\ell}$. Observe that $Y_{\ell}(s) \mid X_{\ell}, H_\ell$ is $\sigma^2$-sub-Gaussian. This implies that $(Y_{\ell}(s))_{\ell \in T_{t,s}}$ is a martingale difference sequence with respect to context and history $(X_{\ell}, H_\ell)_{\ell \in T_{t,s}}$, or $E [Y_{\ell}(s) \mid X_{\ell}, H_\ell] = 0$ for all rounds $\ell \in T_{t,s}$.

For any round $t$, and state $s \in S$, we have that $T_{t,s}$ is a random quantity. First, we fix $|T_{t,s}| = N_t(s) = u$ where $u \leq \tau$ and yield the following due to Proposition 2,

$$
P \left( \left| \sum_{\ell \in T_{t,s}} Y_\ell(s) \right| \geq \sigma \sqrt{6u \log n} \right) \leq 2 \exp \left[ -3 \log n \right] = 2n^{-3}.
$$

So, by the union bound, we have

$$
P (\bar{E}) \leq \sum_{t \in T} \sum_{s \in S} \sum_{u=1}^{\tau} P \left( \left| \sum_{\ell \in T_{t,s}} Y_\ell(s) \right| \geq \sigma \sqrt{6u \log n} \right) \leq 2 |S| n^{-1}.
$$

This concludes the proof. \[\square\]

We can show that the second term in (13) is small because the probability of $\bar{E}$ is small. Specifically, from Proposition 3, and that total regret is bounded by $n$, we have that the second term in (13) is bounded by $n \Pr (\bar{E}) \leq 2 |S|$. Next, we bound the third term in (13). For round $t \in T$, the event $\mu(A_{t,s}, X_t, s; \theta_s) > U_t(A_{t,s})$ occurs only if $s_t \notin C_t$ also occurs. By the design of $C_t$ in S\small$\scriptsize \text{W-}\$UCB, this happens only if $G_t(s_t) > \sigma \sqrt{6N_t(s_t)} \log n$. Event $E_t$ says that the opposite is true for all states, including true state $s_t$. So the third term in (13) is at most 0.

Now we consider the last term in (13). We know that $T$ is composed of $L$ stationary segments. We bound the last term for each segment individually as follows.

Proposition 4. Let $I \subseteq T$ be a stationary segment containing $m$ rounds. Then

$$
E \left[ \mathbb{1} \{ E \} \sum_{t \in I} (U_t(A_t) - \mu(A_t, X_t, s_t; \theta_s)) \right] \leq |S| |m/\tau| + 2\sigma \sqrt{6|S| |m/\tau| m \log n}.
$$

Proof. To ease exposition, let the $m$ rounds in $I$ be denoted $1, \ldots, m$. We can further divide $I$ into intervals of length $\tau$ and the last with length of at most $\tau$. Let $1 = t_0 \leq t_1 \leq \ldots \leq t_{\lfloor m/\tau \rfloor} = m$ partition $I$ into such intervals. We can write,

$$
E \left[ \mathbb{1} \{ E \} \sum_{t \in I} (U_t(A_t) - \mu(A_t, X_t, s_t; \theta_s)) \right] = E \left[ \mathbb{1} \{ E \} \sum_{i=1}^{\lfloor m/\tau \rfloor} \sum_{\ell = t_i-1}^{t_i} (\mu(A_{\ell}, X_{\ell}, B_\ell; \theta_s) - R_{\ell}) \right] + E \left[ \mathbb{1} \{ E \} \sum_{i=1}^{\lfloor m/\tau \rfloor} \sum_{\ell = t_i-1}^{t_i} (R_{\ell} - \mu(A_{\ell}, X_{\ell}, s_{\ell}; \theta_s)) \right]
$$

$$
\leq E \left[ \sum_{i=1}^{\lfloor m/\tau \rfloor} \sum_{s \in S} (G_{t_i}(s) + 1) \right] + \sum_{i=1}^{\lfloor m/\tau \rfloor} \sum_{s \in S} \sigma \sqrt{6N_{t_i}(s)} \log n
$$

$$
\leq |S| \lfloor m/\tau \rfloor + \sum_{s \in S} \sum_{i=1}^{\lfloor m/\tau \rfloor} 2\sigma \sqrt{6N_{t_i}(s)} \log n.
$$

For each window $i$ of length $\tau$ and latent state $s$, we use that until the last round before $t_i$ where $s$ is selected, we have an upper bound on the total prediction error, given by the upper bound on the gap $G_{t_i}(s) \leq \sigma \sqrt{6N_{t_i}(s)} \log n$, where $G_{t_i}(s)$ is defined as in (10). Recall that $E_{t,s}$, as defined in (12), occurring implies that the deviation of the realized reward from the true
means bounded by $\sigma \sqrt{6N_t(s) \log n}$. Accounting for the last round where $s$ was chosen in window $i$ yields the right-hand side of the inequality. Applying the Cauchy-Schwarz inequality yields,

$$\mathbb{E} \left[ \mathbb{1} \{ E \} \sum_{t \in T} U_t(A_t) - \mu(A_t, X_t, s_t; \theta_s) \right] \leq \left| S \right| \left[ m/\tau \right] + 2\sigma \sqrt{6|S|m/\tau|m \log n},$$

which is the desired upper bound.

Now we can bound the last term in (13) by combining Proposition 4 across all $L$ stationary segments. Let $(i_t)_{t \in [L]}$ denote the stationary segments, and segment $i_t$ have length $m_i$. We have,

$$\mathbb{E} \left[ \mathbb{1} \{ E \} \sum_{t \in \tau} (U_t(A_t) - \mu(A_t, X_t, s_t; \theta_s)) \right] = \mathbb{E} \left[ \mathbb{1} \{ E \} \sum_{i=1}^{L} \sum_{t \in i_t} (U_t(A_t) - \mu(A_t, X_t, s_t; \theta_s)) \right] \leq \sum_{i=1}^{L} \left| S \right| \left[ m_i/\tau \right] + 2\sigma \sqrt{6|S|m_i/\tau|m_i \log n} \leq |S|(n/\tau) + 2\sigma \sqrt{6|S|(n/\tau)n \log n}.$$  

Here we use that for any segment $i$, we have $[m_i/\tau] \leq (m_i + \tau)/\tau$ for any number of rounds $m_i$, and that $\sum m_i = n - L\tau$ because we omitted rounds to close to a change-point. Combining the bounds for all terms in (13) yields,

$$\mathcal{R}(n; \theta_s, s_{1:n}) \leq L\tau + 2|S| + |S|(n/\tau) + 2\sigma \sqrt{6|S|(n/\tau)n \log n},$$

When $L$ is known, we can solve for the optimal window length $\tau = \mathcal{O}(n^{2/3}\sqrt{|S| \log n/L})$, which when substituted into the regret bound yields $\mathcal{R}(n; \theta_s, s_{1:n}) = \mathcal{O}(n^{2/3}\sqrt{|S|L \log n})$, as desired.

### A.2 Proof of Theorem 1

From the Bayes regret formulation in (2), the true latent state sequence $S_{1:n} \in S^n$ is random for a fixed transition model $\phi_s$. Here we still assume a fixed reward model $\theta_s$. We have that the optimal action $A_{t,*} = \arg \max_{a,X_t,S_t; \theta_s} \mu(a, X_t, S_t; \theta_s)$ is random not only due to context, but also latent state $S_t$. We also have that $L = \sum_{i=1}^{n} \mathbb{1} \{ S_t \neq S_{t-1} \}$ is random due to latent state sequence $S_{1:n}$.

Similar to Russo and Van Roy (2013), we reduce our analysis of mTS to analysis of SW-mUCB as done in Lemma 1. We define $U_t(a) = \max_{a \in C_{t}} \mu(a, X_t, s; \theta) \in C_{t}$ where the $C_t$ is as in SW-mUCB. Recall that the Bayes regret is given by (2), and can be decomposed as (9). In Appendix A.1, we bounded an equivalent regret decomposition for any $\theta_s, S_{1:n}$ and therefore also in expectation over $S_{1:n} \sim \phi_s$. We have the Bayes regret bound,

$$\mathbb{E}_{S_{1:n} \sim \phi_s} \left[ L\tau + 2|S| + |S|(n/\tau) + 2\sigma \sqrt{6|S|(n/\tau)n \log n} \right],$$

where we directly substitute the upper bound in Lemma 1 inside the expectation.

Since $\phi_s$ is known, we can define $p = 1 - \min_{s \in S} P(s | s; \phi_s)$ as the maximum probability of a change occurring. Then number of change-points $L - 1$ is a binomial random variable, so that $\mathbb{E}_{S_{1:n} \sim \phi_s} [L] = 1 + pn = \hat{L}$. For optimal choice of $\tau = \mathcal{O}(n^{2/3}\sqrt{|S| \log n/L})$, we can simplify the expectation over random $L$ to yield,

$$\mathbb{E}_{S_{1:n} \sim \phi_s} \left[ L\tau + 2|S| + |S|(n/\tau) + 2\sigma \sqrt{6|S|(n/\tau)n \log n} \right],$$

where we use Jensen’s inequality and that the expression inside the expectation is concave in $L$. 

A.3 Proof of Theorem 2

From the Bayes regret formulation in (3), both the reward and transition model parameters $\theta_*$, $\phi_*$ are now random according to priors $P_\theta(P_{\phi})$, respectively. We have that the optimal action $A_{t,*} = \arg\max_{a \in A} \mu(a, X_t, S_t; \theta_*)$ is random due to context $X_t$, latent state $S_t$, and model $\theta_*$. Recall that given prior $P_\theta$, we have that $\bar{\mu}(a, x, s) = \int_\theta \mu(a, x, s; \theta) P_\theta d\theta$ is the mean conditional reward marginalized with respect to the prior. We make one small change to (10) in Sw-umUCB, which accounts for uncertainty: for round $t$ and state $s$, instead of acting according to true means $\mu(A_t, X_t, s; \theta_*)$, we act conservatively according to the mean marginalized over the prior $\bar{\mu}(A_t, X_t, s)$. Formally, the “gap” in Sw-umUCB is redefined as

$$G_t(s) = \sum_{\ell = \max\{1, t-\tau\}}^{t-1} \mathbb{1}\{B_\ell = s\} (\bar{\mu}(A_\ell, X_\ell, s) - \varepsilon - R_\ell).$$

(14)

The additional $\varepsilon$ ensures that we do not mistakenly eliminate the true latent state from $C_t$ due to a prediction error.

Our proof uses the following regret bound for Sw-umUCB, which is for a fixed $\theta_*$ sampled from the prior.

**Lemma 2.** For fixed model parameters $\theta_*$, assume that there exist $\varepsilon > 0$ such that $\theta_*$ satisfies the following: $\{\forall a \in A, x \in X, s \in S : |\bar{\mu}(a, x, s) - \mu(a, x, s; \theta_*)| \leq \varepsilon\}$. Then for optimal $\tau$, the $n$-round regret of Sw-umUCB is

$$\mathcal{R}(n; \theta_*, s_{1:n}) = \mathcal{O}\left(\varepsilon n + n^{2/3} \sqrt{|S|L \log n}\right).$$

**Proof.** We have the same regret decomposition for $n$-round regret, stated in (13). The analysis proceeds similarly to Appendix A.1, only we need to additionally account for prediction error in the conditional mean rewards. We only highlight the differences, and defer other details of the proof to Appendix A.1.

Using Proposition 3, and that the total regret is bounded by $n$, we again have the second term in (13) can be bounded by, $n \mathbb{E} (\mathbb{E}) \leq 2|S|$. Bounding the third term in (13) requires a slight change. For round $t \in T$, we have that the event $\mu(A_t, X_t, s_1; \theta_*) - U_t(A_t, s_1) > \varepsilon$ occurs only if $s_t \notin C_t$. By the design of $C_t$ in Sw-umUCB, this happens only if $G_t(s_t) > \sigma \sqrt{6N_t(s) \log n}$, since

$$G_t(s_t) = \sum_{\ell = \max\{1, t-\tau\}}^{t-1} \mathbb{1}\{B_\ell = s_t\} (\bar{\mu}(A_\ell, X_\ell, s_t) - \varepsilon - R_\ell) \leq \sum_{\ell = \max\{1, t-\tau\}}^{t-1} \mathbb{1}\{B_\ell = s_t\} (\mu(A_\ell, X_\ell, s_t; \theta_*) - R_\ell).$$

Event $E_t$ says that the opposite is true for all states, including true state $s_t$. So the third term in (13) is at most $\varepsilon n$.

For the last term in (13), we need to account for the fact that $\varepsilon$ is included in the gap $G_t(s)$ for every round $t$ and state $s$. To do so, we introduce a $\varepsilon n$ in the expression as,

$$\mathbb{E}\left[\mathbb{1}\{E\} \sum_{t \in T} (U_t(A_t) - \mu(A_t, X_t, s_t; \theta_*)\right] \leq \varepsilon n + \mathbb{E}\left[\mathbb{1}\{E\} \sum_{t \in T} (U_t(A_t) - \varepsilon + \mu(A_t, X_t, s_t; \theta_*)\right]

$$

The second term on the right-hand side can be bounded the same way as in Appendix A.1 by introducing the realized reward, and bounding the sum of confidence widths using the gap given in (14).

This yields the bound on total regret,

$$\mathcal{R}(n; \theta_*, s_{1:n}) \leq L\tau + 2\varepsilon n + |S|(n/\tau + 2) + 2\sigma \sqrt{6|S|(n/\tau)n \log n}.$$ Solving for optimal window length in terms of $L$ yields $\tau = \mathcal{O}(n^{2/3} \sqrt{|S| \log n / L})$, which when substituted into the regret gives $\mathcal{R}(n; \theta_*, s_{1:n}) = \mathcal{O}(\varepsilon n + n^{2/3} \sqrt{|S|L \log n})$, as desired.

In order to prove Theorem 2, we again reduce to the proof of Lemma 2 for Sw-umUCB. We define $U_t(a) = \arg\max_{s \in C_t} \mu(a, X_t, s; \theta_*)$ as in Sw-umUCB. We also define event

$$\mathcal{E} = \{\forall a \in A, x \in X, s \in S : |\bar{\mu}(a, x, s) - \mu(a, x, s; \theta_*)| \leq \varepsilon\}.$$
for when the sampled true model $\theta^*$ behaves close to expected, and $\bar{E}$ as its complement. If $E$ does not hold, then the best possible upper bound on regret is $n$; fortunately, we assume in the statement of the theorem that the probability of that occurring is bounded by $\delta$. So we can bound the $n$-round Bayes regret as

$$BR(n) = \mathbb{E} \left[ \mathbb{1} \{ \bar{E} \} R(n; \theta^*, s_{1:n}) \right] + \mathbb{E} \left[ \mathbb{1} \{ E \} R(n; \theta^*, s_{1:n}) \right] \leq \delta n + \mathbb{E} \left[ \mathbb{1} \{ E \} R(n; \theta^*, s_{1:n}) \right].$$

The second term can be decomposed as in (9) and bounded by Lemma 2 as the bound in the lemma is worst-case over any model parameters $\theta^*$ and sequence $S_{1:n}$. We have the Bayes regret bound,

$$BR(n) \leq \delta n + 2\varepsilon n + \mathbb{E}_{\phi^* \sim P_1} \left[ L\tau + 2|S| + |S|(n/\tau) + 2\sigma \sqrt{6|S|(n/\tau)n \log n} \right].$$

Here $\phi^*$ is random, and hence number of stationary segments $L$ is also random. Let $p$ denote the maximum probability of change, i.e., for fixed $\phi^*$, we have $p = 1 - \min_{s \in S} P(s \mid s; \phi^*)$ as in Theorem 1. Unlike in Theorem 1, we have that $p$ is random as well due to randomness in $\phi^*$. Recall from the statement of Theorem 2 that $(\alpha_{s,s'})_{s,s' \in S \times S}$ are the prior parameters of $P_1(\phi)$. We can write $\mathbb{E}_{\phi^* \sim P_1}[p] \leq 1 - \min_{s \in S} \alpha_{s,s}/\sum_{s' \in S} \alpha_{s,s'} = \hat{p}$. This means we can bound the expected value of $L$ as,

$$\mathbb{E}_{\phi^* \sim P_1} \left[ L \right] = \mathbb{E}_{\phi^* \sim P_1} \left[ \mathbb{E}_{S_{1:n} \sim \phi^*} \left[ L \mid \phi^* \right] \right] \leq \mathbb{E}_{\phi^* \sim P_1} \left[ 1 + np \right] \leq 1 + \hat{p}n = \hat{L}.$$

Since the Bayes regret is still concave in $L$, we can apply the same trick as in Appendix A.2 using Jensen’s inequality, and yield, for optimal choice of $\tau = O(n^{2/3} \sqrt{|S| \log n/\hat{L}})$, the desired Bayes regret bound

$$BR(n) = O \left( \delta n + \varepsilon n + n^{2/3} \sqrt{|S| \mathbb{E}_{\phi^* \sim P_1}[L] \log n} \right) = O \left( \delta n + \varepsilon n + n^{2/3} \sqrt{|S| \hat{L} \log n} \right).$$