ONLINE RESTLESS BANDITS WITH UNOBSERVED STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the online restless bandit problem, where each arm evolves according to a Markov chain independently, and the reward of pulling an arm depends on both the current state of the corresponding Markov chain and the action. The agent (decision maker) does not know the transition kernels and reward functions, and cannot observe the states of arms all the time. The goal is to sequentially choose which arms to pull so as to maximize the expected cumulative rewards collected. In this paper, we propose TSEETC, a learning algorithm based on Thompson Sampling with Episodic Explore-Then-Commit. The algorithm proceeds in episodes of increasing length. In the exploration phase in each episode, action-reward samples are collected in a round robin way and then used to update the posterior as mixture of Dirichlet distributions. At the beginning of exploitation phase, TSEETC generates a sample from the posterior distribution as true parameters. It then follows the optimal policy for the sampled model for the rest of the episode. We establish the Bayesian regret bound $\tilde{\mathcal{O}}(\sqrt{T})$ for TSEETC, where T is the time horizon. This is the first bound that is close to the lower bound of restless bandits especially in unobserved states setting. We show through simulations that TSEETC outperforms existing algorithms in regret.

1 INTRODUCTION

The restless multi-armed problem (RMAB) is a general setup to model many sequential decision making problems ranging from wireless communication (Tekin & Liu, 2011; Sheng et al., 2014), sensor/machine maintenance (Ahmad et al., 2009; Akbarzadeh & Mahajan, 2021) and healthcare (Mate et al., 2020; 2021). This problem considers one agent and N arms. Each arm i is modulated by a Markov chain M^i with state transition function P^i and reward function R^i. At each time, the agent decides which arm to pull. After the pulling, all arms undergo an action-dependent Markovian state transition. The goal is to decide which arm to pull to maximize the expected reward, i.e., $\mathbb{E}[\sum_{t=1}^{T} r_t]$, where r_t is the reward at time t and T is the time horizon.

In this paper, we consider the online restless bandit problem with **unknown parameters (transition functions and reward functions)** and **unobserved states**. Many works concentrate on learning unknown parameters (Liu et al., 2010; 2011; Ortner et al., 2012; Wang et al., 2020; Xiong et al., 2022a;b) while ignoring the possibility that the states are also unknown. The unobserved states assumption is common in real-world applications, such as cache access (Paria & Sinha, 2021) and recommendation system (Peng et al., 2020). In the cache access problem, the user can only get the perceived delay but cannot know whether the requested content is stored in the cache before or after the access. Moreover, in the recommender system, we do not know the user's preference for the items. There are also some studies that consider the unobserved states. However, they often assume the parameters are known (Mate et al., 2020; Meshram et al., 2018; Akbarzadeh & Mahajan, 2021) and there is a lack of theoretical result (Peng et al., 2020; Hu et al., 2020). What is worse, the existing algorithms (Zhou et al., 2021; Jahromi et al., 2022) with theoretical guarantee do not match the lower regret bound of RMAB (Ortner et al., 2012).

One common way to handle the unknown parameters but with observed states is to use the optimism in the face of uncertainty (OFU) principle (Liu et al., 2010; 2011; Ortner et al., 2012; Wang et al., 2020; Xiong et al., 2022a). These methods sometimes do not perform close to the optimal, because of the baseline they consider, such as pulling the fixed arms (Liu et al., 2010; 2011), which is

not optimal in RMAB problem. Ortner et al. (2012) derives the lower bound $\tilde{\mathcal{O}}(\sqrt{T})$ for RMAB problem. However, it is not clear whether there is an efficient computational method to find the optimistic model in the confidence region (Lakshmanan et al., 2015). Another way to estimate the unknown parameters is Thompson Sampling (TS) method (Jung & Tewari, 2019; Jung et al., 2019; Jahromi et al., 2022; Hong et al., 2022). TS algorithm is more computationally efficient since it only needs to solve the sampled instance, without the need to solve all instances that lie within the confident sets as OFU-based algorithms (Ouyang et al., 2017). What's more, empirical studies suggest that TS algorithms outperform OFU-based algorithms in bandit and Markov decision process (MDP) problems (Scott, 2010; Chapelle & Li, 2011; Osband & Van Roy, 2017).

Some studies assume that only the states of pulled arms are observable (Mate et al., 2020; Liu & Zhao, 2010; Wang et al., 2020; Jung & Tewari, 2019). They translate the partially observable Markov decision process (POMDP) problem into a fully observable MDP by regarding the state last observed and the time elapsed as a meta-state (Mate et al., 2020; Jung & Tewari, 2019), which is much simpler due to more observations about pulled arms. Mate et al. (2020), and Liu & Zhao (2010) derive the optimal index policy but they assume the known parameters. Restless-UCB in Wang et al. (2020) achieves with no guarantee of an $\tilde{\mathcal{O}}(\sqrt{T})$ regret, and also restricted to a specific Markov model. There are also some works that consider that the arm's state is not visible even after pulling (Meshram et al., 2018; Akbarzadeh & Mahajan, 2021; Peng et al., 2020; Hu et al., 2020; Zhou et al., 2021) and the classic POMDP setting (Jahromi et al., 2022). However, there are still some challenges unresolved. Firstly, Meshram et al. (2018) and Akbarzadeh & Mahajan (2021) study the RMAB problem with unobserved states but with known parameters. Since the true value of the parameters are unavailable in practice, their contribution is limited. Secondly, the works study RMAB from a learning perspective, e.g., Peng et al. (2020); Hu et al. (2020) but there are no regret analysis. Thirdly, existing policies with regret bound $\tilde{\mathcal{O}}(T^{2/3})$ (Zhou et al., 2021; Jahromi et al., 2022) often do not have a regret guarantee that scales as $\tilde{\mathcal{O}}(\sqrt{T})$, which is the lower bound in RMAB problem (Ortner et al., 2012). To the best of our knowledge, there are no provably optimal policies that perform close to the offline optimum and match the lower bound in restless bandit, especially in unobserved states setting.

In this paper, we address the above challenges for RMAB problems with unknown parameters and unobserved states. We design a learning algorithm TSEETC to estimate these unknown parameters, and benchmarked on a stronger oracle, we show that our algorithm achieves a tighter regret bound. In summary, we make the following contributions:

Problem formulation. We consider the online restless bandit problems with unobserved states and unknown parameters. Compared with Jahromi et al. (2022), our reward functions are unknown.

Algorithmic design. We propose TSEETC, a learning algorithm based on Thompson Sampling with Episodic Explore-Then-Commit. The whole learning horizon is divided into episodes of increasing length. Each episode is split into exploration and exploitation phases. In the exploration phase, to estimate the unknown parameters, we propose the first programmable and low-complexity algorithm to update the posterior distribution as a mixture of Dirichlet distributions. For the unobserved states, we use the belief state to encode the historical information. In the exploitation phases, we sample the parameters from the posterior distribution and derive an optimal policy based on the sampled parameter. What's more, we design the determined episode length in an increasing manner to control the total episode number, which is crucial to bound the regret caused by exploration.

Regret analysis. We consider a stronger oracle which solves POMDP based on our belief state. Under a Bayesian framework, we show that the expected regret of TSEETC accumulated up to time T is bounded by $\tilde{\mathcal{O}}(\sqrt{T})$, where $\tilde{\mathcal{O}}$ hides logarithmic factors. This bound improves the exiting results (Zhou et al., 2021; Jahromi et al., 2022) and matches the theoretical lower bound for the restless bandit problem, especially in unobserved states setting.

Experiment results. We conduct the proof-of-concept experiments, and compare our policy with existing baseline algorithms. Our results show that outperforms existing algorithms and achieve a near-optimal regret bound.

2 Related Work

We review the related works in two main domains: learning algorithm for unknown parameters, and methods to unknown states.

Unknown parameters. Since the system parameters are unknown in advance, it is essential to study RMAB problems from a learning perspective. Generally speaking, these works can be divided into two categories: OFU (Ortner et al., 2012; Wang et al., 2020; Xiong et al., 2022a; Zhou et al., 2021; Xiong et al., 2022b) or TS based (Jung et al., 2019; Jung & Tewari, 2019; Jahromi et al., 2022; Hong et al., 2022). The algorithms based on OFU often construct confidence sets for the system parameters at each time, find the optimistic estimator that is associated with the maximum reward, and then select an action based on the optimistic estimator. However, these methods suffer from some issues. Firstly, it may not perform close to the offline optimum because the baseline policy they consider, such as pulling only one arm, is often a heuristic policy and not optimal. In this case, the regret bound $\mathcal{O}(\log T)$ (Liu et al., 2010) is less meaningful. Secondly, the state of the art algorithm of regret bound $\tilde{\mathcal{O}}(\sqrt{T})$ are often computationally expensive. For example, the colored-UCRL2 algorithm (Ortner et al., 2012) suffers from high computational complexity because it should find the optimistic estimator in the confidence region. Apart from these works, posterior sampling (Jung & Tewari, 2019; Jung et al., 2019) were used to solve this problem. A TS algorithm generally samples a set of MDP parameters randomly from the posterior distribution, then actions are selected based on the sampled model. Jung & Tewari (2019) and Jung et al. (2019) provide theoretical guarantee $\tilde{\mathcal{O}}(\sqrt{T})$ in the Bayesian setting. TS algorithms are confirmed to outperform optimistic algorithms in bandit and MDP problems (Scott, 2010; Chapelle & Li, 2011; Osband & Van Roy, 2017).

Unknown states. There are some works that consider the states of the pulled arm are observed (Mate et al., 2020; Liu & Zhao, 2010; Wang et al., 2020; Jung & Tewari, 2019). Mate et al. (2020) and Liu & Zhao (2010) assumes the unobserved states but with known parameters. Wang et al. (2020) constructs an offline instance and give the regret bound $\tilde{\mathcal{O}}(T^{2/3})$. Jung & Tewari (2019) considers the episodic RMAB problems and the regret bound $\tilde{\mathcal{O}}(\sqrt{T})$ is guaranteed in the Bayesian setting. Some studies assume that the states are unobserved even after pulling. Akbarzadeh & Mahajan (2021) and Meshram et al. (2018) consider the RMAB problem with unknown states but known system parameters. And there is no regret guarantee. Peng et al. (2020) and Hu et al. (2020) consider the unknown parameters but there are also no any theoretical results. The most similar to our work is Zhou et al. (2021) and Jahromi et al. (2022). Zhou et al. (2021) considers that all arms are modulated by a common unobserved Markov Chain. They proposed the estimation method based on spectral method (Anandkumar et al., 2012) and learning algorithm based on upper confidence bound (UCB) strategy (Auer et al., 2002). They also give the regret bound $\tilde{\mathcal{O}}(T^{2/3})$ and there is a gap between the lower bound $\tilde{\mathcal{O}}(\sqrt{T})$ (Ortner et al., 2012). Jahromi et al. (2022) considers the POMDP setting and propose the pseudo counts to store the state-action pairs. Their learning algorithm is based on Ouyang et al. (2017) and the regret bound is also $\tilde{\mathcal{O}}(T^{2/3})$. And their algorithm is not programmable due to the pseudo counts is conditioned on the true counts which is uncountable.

3 Problem Setting

Consider a restless bandit problem with one agent and N arms. Each arm $i \in [N] := \{1, 2, \ldots, N\}$ is associated with an independent discrete–time Markov chain $\mathcal{M}^i = (\mathcal{S}^i, P^i)$, where $\mathcal{S}^i$ is the state space and $P^i \in \mathbb{R}^{\mathcal{S}^i \times \mathcal{S}^i}$ the transition functions. Let s_t^i denote the state of arm i at time t and $s_t = (s_t^1, s_t^2, \ldots, s_t^N)$ the state of all arms. Each arm i is also associated with a reward functions $R^i \in \mathbb{R}^{\mathcal{S}^i \times \mathcal{R}}$, where $R^i(r \mid s)$ is the probability that the agent receives a reward $r \in \mathcal{R}$ when he pulls arm i in state s. We assume the state spaces $\mathcal{S}^i$ and the reward set $\mathcal{R}$ are finite and known to the agent. The parameters P^i and R^i, $i \in [N]$ are unknown, and the state s_t is also unobserved to the agent. For the sake of notational simplicity, we assume that all arms have the same state spaces with size S. Our result can be generalized in a straightforward way to allow different state spaces.

The whole game is divided into T time steps. The initial state s_1^i for each arm $i \in [N]$ is drawn from a distribution h_i independently, which we assume to be known to the agent. At each time t, the agent chooses one arm $a_t \in [N]$ to pull and receives a reward $r_t \in \mathcal{R}$ with probability $R^{a_t}(r_t \mid s_t^{a_t})$. Note

that only the pulled arm has the reward feedback. His decision on which arm a_t to pull is based on the observed history $\mathcal{H}_t = [a_1, r_1, a_2, r_2 \cdots, a_{t-1}, r_{t-1}]$. Note that the states of the arms are never observable, even after pulling. Each arm i makes a state transition independently according to the associated P^i, whether it is pulled or not. This process continues until the end of the game. The goal of the agent is to maximize the total expected reward.

We use θ to denote the unknown P^i and R^i for $i \in [N]$ collectively. Since the true states are unobservable, the agent maintains a belief state $b_t^i = [b_t^i(s, \theta), s \in \mathcal{S}^i] \in \Delta_{\mathcal{S}^i}$ for each arm i, where

$$b_t^i(s, \theta) := \mathbb{P}\left(s_t^i = s \mid \mathcal{H}_t, \theta\right),$$

and $\Delta_{\mathcal{S}^i} := \left\{b \in \mathbb{R}_+^{\mathcal{S}^i} : \sum_{s \in \mathcal{S}^i} b(s) = 1\right\}$ is the probability simplex in $\mathbb{R}^{\mathcal{S}^i}$. Note that $b_t^i(s, \theta)$ depends on the unknown model parameter θ, which itself has to be learned by the agent. For a given θ, the overall belief state $b_t = (b_t^1, b_t^2, \cdots, b_t^N)$ is a sufficient statistic for $\mathcal{H}_{t-1}$ (Smallwood & Sondik, 1973), so the agent can base his decision at time t on b_t only. Let $\Delta_b := \Delta_{\mathcal{S}^1} \times \cdots \times \Delta_{\mathcal{S}^N}$. A deterministic stationary policy $\pi : \Delta_b \to [N]$ maps a belief state to an action. The long-term average reward of a policy π is defined as

$$J^\pi(h, \theta) := \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} r_t \;\middle|\; h, \theta\right]. \tag{1}$$

We use $J(h, \theta) = \sup_\pi J^\pi(h, \theta)$ to denote the optimal long-term average reward. We assume $J(h, \theta)$ is independent with the initial distribution h as Jahromi et al. (2022) and thus denoted by $J(\theta)$. We make the following assumption.

Assumption 1. *The smallest element ϵ in the transition functions $P_i, i \in N$ is bigger than zero.*

Assumption 1 not only helps us bound the error of belief estimation (De Castro et al., 2017) but also makes the MDP weakly communicating (Bertsekas et al., 2011). For weakly communicating MDP, it is known that there exists a bounded function $v(\cdot, \theta) : \Delta_b \to \mathbb{R}$ such that for all $b \in \Delta_b$ (Bertsekas et al., 2011),

$$J(\theta) + v(b, \theta) = \max_a \left\{r(b, a) + \sum_r P(r \mid b, a, \theta) v\left(b', \theta\right)\right\}, \tag{2}$$

where v is the relative value function, $r(b, a) = \sum_s \sum_r b^a(s, \theta) R^a(r \mid s) r$ is the expected reward, b' is the updated belief after obtaining the reward r, and $P(r \mid b, a, \theta)$ is the probability of observing r in the next step, conditioned on the current belief b and action a. The corresponding optimal policy is the maximizer of the right part in equation 2. Note that if $v(, \theta)$ satisfies equation 2, so does $v(, \theta)$ plus any constant. Therefore, without loss of generality, and since $v(, \theta)$ is bounded, we can assume that $\inf_{b \in \Delta_b} v(b, \theta) = 0$ and define the span as $\mathrm{sp}(\theta) := \sup_{b \in \Delta_b} v(b, \theta) \leq H$.

We consider the Bayesian setting for the parameters. The parameters θ^* is randomly generated from a known prior distribution Q at the beginning and then fixed but unknown to the agent. We measure the efficiency of a policy π by its regret, defined as the expected gap between the cumulative reward of an offline oracle and that of π, where the oracle is the optimal policy with the full knowledge of θ^*, but unknown states. The offline oracle is similar to Zhou et al. (2021), which is stronger than those considered in Azizzadenesheli et al. (2016) and Fiez et al. (2018). We focus on the Bayesian regret of policy π (Ouyang et al., 2017; Jung & Tewari, 2019) as follows,

$$R_T := \mathbb{E}_{\theta^* \sim Q}\left[\sum_{t=1}^{T} \left(J(\theta^*) - r_t\right)\right]. \tag{3}$$

The above expectation is with respect to the prior distribution about θ^*, the randomness in state transitions and the random reward.

4 The TSEETC Algorithm

In section 4.1, we define the belief state and show how to update is with new observation. In section 4.2, we update the posterior distributions of unknown parameters as a mixture of Dirichlet distributions. In section 4.3, we propose our learning algorithm: Thompson Sampling with Episodic Explore-Then-Commit (TSEETC) algorithm.

4.1 BELIEF ENCODER FOR UNOBSERVED STATE

Here we focus on the belief update for arm i with known parameters. At time t, the belief for arm i in state s is $b_t^i(s,\theta)$ where the θ is the true parameters. Then after the pulling of arm i, we obtain the observation r_t. The belief $b_t^i(s',\theta)$ can be update as follows:

$$b_{t+1}^i(s',\theta) = \frac{\sum_s b_t^i(s,\theta) R^i\left(r_t \mid s\right) P^i(s' \mid s)}{\sum_s b_t^i(s,\theta) R^i\left(r_t \mid s\right)}, \tag{4}$$

where the $P^i(s' \mid s)$ is the probability of transitioning from state s at time t to state s' for the next time $t+1$, and $R^i\left(r_t \mid s\right)$ is the probability of obtain reward r_t under state s.

If the arm i is not pulled, we update its belief as follows:

$$b_{t+1}^i(s',\theta) = \sum_s b_t^i(s,\theta) P^i(s' \mid s). \tag{5}$$

Then at each time, we can aggregate the belief of all arms as b_t. Based on equation 2 , we can derive the optimal action a_t for current belief b_t.

4.2 MIXTURE OF DIRICHLET DISTRIBUTION

In this section, we derive the estimation method for unknown P^i and R^i based on the mixture of Dirichlet distribution. The Dirichlet distribution is parameterized by a count vector, $\phi = (\phi_1, \ldots, \phi_k)$, where $\phi_i \geq 0$, such that the density of probability distribution $p = (p_1, \ldots, p_k)$ is defined as $f(p \mid \phi) \propto \prod_{i=1}^k p_i^{\phi_i - 1}$.

With the Dirichlet distribution, if the states are observed, the posterior distribution is just parameterized by the vector increased at each position (see Appendix C). However, we do not have access to the history $\bar{s}_t^i$ due to the unobserved states assumption. Since the states that truly visited are unknown, all state sequences (and their corresponding Dirichlet posteriors) must be considered, with some weight proportional to the likelihood of each state sequence (Ross et al., 2011). Specially, for the unknown P^i and R^i, their prior distribution are parameterized by ϕ^i and ψ^i, respectively. After observing the history $\bar{r}_t^i$, $\bar{s}_t^i$ the posterior distribution $g_t(P^i)$ and $g_t(R^i)$ at time t can be updated as in Lemma 1.

Lemma 1. *Under the unobserved state setting and the P^i, R^i follow the Dirichlet prior distributions, the posterior distribution of transition P^i and reward function R^i at time t are as follows:*

$$g_t\left(P^i\right) \propto \sum_{\bar{s}_t^i \in \mathcal{S}_i^t} g_0\left(P^i\right) w(\bar{s}_t^i) \prod_{s,s'} \left(P^i(s' \mid s)\right)^{N_{s,s'}^i\left(\bar{s}_t^i\right) + \phi_{s,s'}^i - 1}, \tag{6}$$

$$g_t\left(R^i\right) \propto \sum_{\bar{s}_t^i \in \mathcal{S}_i^t} g_0\left(R^i\right) w(\bar{s}_t^i) \prod_{s,r} \left(R^i(r \mid s)\right)^{N_{s,r}^i\left(\bar{s}_t^i\right) + \psi_{s,r}^i - 1}. \tag{7}$$

where $g_0(P^i)$ and $g_0(R^i)$ are the initial priors ; $w(\bar{s}_t^i)$ is the likelihood of state sequence $\bar{s}_t^i$.

This procedure is summarized in Algorithm 1.

Algorithm 1 Posterior Update for $R^i(s,\cdot)$ and $P^i(s,\cdot)$

1: Input: the history length τ_1, the state space $\mathcal{S}_i$, the belief history $\bar{b}_{\tau_1}^i$, $\bar{r}_{\tau_1}^i$, the initial parameters $\phi_{s,s'}^i$, $\psi_{s,r}^i$, for $s, s' \in \mathcal{S}_i$, $r \in \mathcal{R}$,
2: generate $\mathcal{S}_i^{\tau_1}$ possible state sequences
3: calculate the weight $w(j) = \prod_{t=1}^{\tau_1} b_t^i(s,\theta), j \in \mathcal{S}_i^{\tau_1}$
4: **for** j in $1, \ldots, \mathcal{S}_i^{\tau_1}$ **do**
5: count the occurence times of event (s,s') and (s,r) as $N_{s,s'}^i, N_{s,r}^i$ in sequence j
6: update $\phi_{s,s'}^i \leftarrow \phi_{s,s'}^i + N_{s,s'}^i$, $\psi_{s,r}' \leftarrow \psi_{s,r}^i + N_{s,r}^i$
7: aggregate the $\phi_{s,s'}^i$ as $\phi(j)$, $\psi_{s,r}^i$ as $\psi(j)$ for all $s, s' \in \mathcal{S}_i, r \in \mathcal{R}$
8: **end for**
9: update the mixture Dirichlet distribution
$g_{\tau_1}(P^i) \propto \sum_{j=1}^{\mathcal{S}_i^{\tau_1}} w(j) f(P^i \mid \phi(j))$, $g_{\tau_1}(R^i) \propto \sum_{j=1}^{\mathcal{S}_i^{\tau_1}} w(j) f(R^i \mid \psi(j))$

4.3 OUR ALGORITHM

Our algorithm, TSEETC, operates in episodes with different length. Each episode is split into exploration phase and exploitation phase. Denote the episode number is K_T and the first time in each episode is denoted as t_k. Let T_k be the length of the episode with the convention $T_1 = \left\lceil \frac{\sqrt{T}+1}{2} \right\rceil$. With the convention T_1 and our control manner of episode length, we can determine K_T. The length of exploration phase in each episode is fixed as τ_1 which satisfies $\tau_1 K_T = \mathcal{O}(\sqrt{T})$ and $\tau_1 \leq \frac{T_1+K_T-1}{2}$. With these notations, our whole algorithm is shown below.

Algorithm 2 Thompson Sampling with Episodic Explore-Then-Commit

Input: prior $g_0(P)$,$g_0(R)$, initial belief b_0, exploration length τ_1, the first episode length T_1
for episode $k = 1, 2, \ldots,$ **do**
  start the first time of episode k, $t_k := t$
  generate $R(t_k) \sim g_{t_k}(R)$ and $P(t_k) \sim g_{t_k}(P)$
  for $t = t_k, t_k + 1, ..., t_k + \tau_1$ **do**
    pull the arm i for τ_1/N times in a round robin way
    receive the reward r_t
    update the belief b_t^i using $R(t_k)$, $P(t_k)$ based on equation 4
    update the belief $b_t^j, j \in N \setminus \{i\}$ using $P(t_k)$ based on equation 5
  end for
  for $i = 1, 2, \ldots, N$ **do**
    input the obtained $\bar{r}_{\tau_1}$, $\bar{b}_{\tau_1}$ to Algorithm 1 to update the posterior distribution $g_{t_k+\tau_1}(P)$, $g_{t_k+\tau_1}(R)$
  end for
  generate $R(t_k + \tau_1) \sim g_{t_k+\tau_1}(P)$, $P(t_k + \tau_1) \sim g_{t_k+\tau_1}(R)$
  for i in $0, 1, \ldots, N$ **do**
    re-update the belief b_t^i from time 0 to $t_k + \tau_1$ based on $R(t_k + \tau_1)$ and $P(t_k + \tau_1)$
  end for
  compute $\pi_k^*(\cdot) = \text{Oracle}\left(\cdot, R(t_k + \tau_1), P(t_k + \tau_1)\right)$
  for $t = t_k + \tau_1 + 1, \cdots, t_{k+1} - 1$ **do**
    apply action $a_t = \pi_k^*\left(b_t\right)$
    observe new reward r_{t+1}
    update the belief b_t of all arms based equation 4, equation 5
  end for
end for

In episode k, for the exploration phase, we first sampled the θ_{t_k} from the distribution $g_{t_k}(P)$ and $g_{t_k}(R)$. We pull each arm for τ_1/N times in a round robin way. For the pulled arm, we update its belief based on equation 4 using θ_{t_k}. For the arms that are not pulled, we update its belief based on equation 5 using θ_{t_k}. The reward and belief history of each arm are input into Algorithm 1 to update the posterior distribution after the exploration phase. Then we sample the new $\theta_{t_k+\tau_1}$ from the posterior distribution, and re-calibrate the belief b_t based on the most recent estimated $\theta_{t_k+\tau_1}$. Next we enter into the exploitation phase . Firstly we derive the optimal policy π_k for the sampled parameter $\theta_{t_k+\tau_1}$. Then we use policy π_k for the rest of the episode k.

We control the increasing of episode length in a deterministic manner. Specially, the length for episode k is just one more than the last episode k. That's also to say, the first time t_{k+1} is determined by $t_{k+1} = kT_1 + k$. The intuitions behind such different episode lengths are in two folds. Firstly, as the Dirichlet counts grow larger and larger, the transition functions and reward functions defined by these counts do not change much. Then the sampled parameter θ_k is more concentrated on true values. Therefore we should increase the length of exploitation phase to minimize the total regret. Secondly, in such a deterministic increasing manner, the episode number K_T is bounded by $\mathcal{O}(\sqrt{T})$ as in Lemma 2. Then the regret caused by the exploration phases can be bound by $\mathcal{O}(\sqrt{T})$, which is an crucial part in Theorem 1.

Remark 1. *We use an Oracle to derive the optimal policy for the sampled parameters in Algorithm 2. The Oracle can be the Bellman equation for POMDP as we introduced in equation 2, or the*

approximation methods (Pineau et al., 2003; Silver & Veness, 2010) , etc. The approximation error is discussed in section 5.

5 Performance Analysis

In section 5.1, we show our theoretical results and some discussions. In section 5.2, we provide a proof sketch and the detailed proof is in Appendix B.

5.1 Regret bound and discussions

Theorem 1. *Suppose Assumptions 1 to 2 hold and the Oracle returns the optimal plocy. The Bayesian regret of TSEETC satisfies*

$$R_T \leq 48C_1C_2S\sqrt{AT\log(AT)} + (\tau_1\Delta R + H + 4C_1C_2SA)\sqrt{T} + C_1C_2,$$

where $C_1 = L_1 + L_2SN + N + S^2, C_2 = r_{max} + H$ are constants independent with time horizon T, $L_1 = 4M\left(\frac{1-\epsilon}{\epsilon}\right)^2 / \min\{R^*_{\min}, 1 - R^*_{\max}\}$, $L_2 = 4M(1-\epsilon)^2/\epsilon^3 + \sqrt{M}$, M is the Cartesian product of each arms' state space, $R^*_{\max}$ and $R^*_{\min}$ are the maximum and minimum element of the functions R^* respectively, τ_1 is the fixed exploration length in each episode, ΔR is the biggest gap of the reward obtained at each two different time, H is the bounded span, r_{max} is the maximum reward obtain each time.

Remark 2. *The Theorem 1 shows that the regret of TSEETC is upper bound by $\tilde{\mathcal{O}}(\sqrt{T})$. This is the first bound that matches the lower bound in restless bandit problem (Ortner et al., 2012) in such unobserved state setting. Although TSEETC looks similar to explore-then-commit (Lattimore & Szepesvári, 2020), a key novelty of TSEETC lies in using the approach of posterior sampling (Jung & Tewari, 2019; Russo & Van Roy, 2014) to balance exploration and exploitation in a deterministic-episode manner. Our algorithm ensures the episode length grows at a linear rate, which guarantees that the total episode number is bounded by $\mathcal{O}(\sqrt{T})$. Therefore the total regret caused by exploration is well controlled by $\mathcal{O}(\sqrt{T})$ and this is better than the bound $\mathcal{O}(T^{2/3})$ in Zhou et al. (2021). What's more, in the exploitation phase, our regret bound $\tilde{\mathcal{O}}(\sqrt{T})$ is also better than $\tilde{\mathcal{O}}(T^{2/3})$ (Zhou et al., 2021). This shows our posterior sampling based method is superior to UCB based solution (Osband & Van Roy, 2017). In Jahromi et al. (2022), their pseudo count of state-action pair is always smaller than the true counts with some probability at any time. However, in our algorithm, the sampled parameter is becoming more concentrated on true values. Therefore, our pseudo count(defined in equation 13) based on belief approximates the true counts more closely, which helps us obtain a tighter bound.*

Remark 3. *(Approximation error.) If the oracle returns an ϵ_k-approximate policy $\tilde{\pi}_k$ in each episode instead of the optimal policy. That is to say, $r(b, \tilde{\pi}_k(b)) + \sum_r P(r \mid b, \tilde{\pi}_k(b), \theta)v(b', \theta) \leq \max_a \{r(b,a) + \sum_r P(r \mid b, a, \theta)v(b', \theta)\} - \epsilon_k$. Then we should consider the extra regret $\mathbb{E}\left[\sum_{k:t_k \leq T}(T_k - \tau_1)\epsilon_k\right]$ in exploitation phase. If we control the error as $\epsilon_k \leq \frac{1}{T_k - \tau_1}$, then we can bound the extra regret as $\mathbb{E}\left[\sum_{k:t_k \leq T}(T_k - \tau_1)\epsilon_k\right] \leq k_T = \mathcal{O}(\sqrt{T})$(Lemma 2). Thus the approximation error in the computation of optimal policy is only additive to the regret of our algorithm.*

5.2 Proof Sketch

In our algorithm, the total regret can be decomposed as follows:

$$R_T = \underbrace{\mathbb{E}_{\theta_*}\left[\sum_{k=1}^{k_T}\sum_{t_k}^{t_k+\tau_1} J(\theta^*) - r_t\right]}_{\text{Regret (A)}} + \underbrace{\mathbb{E}_{\theta_*}\left[\sum_{k=1}^{k_T}\sum_{t_k+\tau_1+1}^{t_{k+1}-1} J(\theta^*) - r_t\right]}_{\text{Regret (B)}}. \tag{8}$$

Bounding Regret (A). The Regret (A) is the regret caused in the exploration phase of each episode. This term can be simply bounded as follows:

$$\text{Regret (A)} \leq \mathbb{E}_{\theta_*}\left[\sum_{k=1}^{k_T}\tau_1\Delta R\right] \leq \tau_1\Delta R k_T \tag{9}$$

where $\Delta R = r_{max} - r_{min}$ is the biggest gap of the reward received at each two different times. The regret in equation 9 is related with the episode number k_T, which can be bound in Lemma 2.

Lemma 2. *(Bound the episode number) With the convention* $T_1 = \left\lceil \frac{\sqrt{T}+1}{2} \right\rceil$ *and* $T_k = T_{k-1} + 1$, *the episode number is bounded by* $K_T = \mathcal{O}(\sqrt{T})$.

Bounding Regret (B). Next we bound Regret(B) in the exploitation phase. Define $\hat{b}_t$ is the belief updated with parameter θ_k, and b_t^* represents the belief with θ^*. During episode k, based on equation 2 for the sampled parameter θ_k and that $a_t = \pi^*(\hat{b}_t)$, we can write:

$$J(\theta_k) + v(\hat{b}_t, \theta_k) = r(\hat{b}_t, a_t) + \sum_r P(r \mid \hat{b}_t, a_t, \theta_k) v(\hat{b}_{t+1}, \theta_k). \tag{10}$$

With this equation, we proceed by decomposing the regret as:

$$\text{Regret(B)} = R_1 + R_2 + R_3 + R_4 \tag{11}$$

where each term is defined as follows:

$$\begin{aligned}
R_1 &= \mathbb{E}_{\theta_*} \sum_{k=1}^{k_T} \left[(T_k - \tau_1 - 1) \left(J(\theta^*) - J(\theta_k) \right) \right], \\
R_2 &= \mathbb{E}_{\theta_*} \sum_{k=1}^{k_T} \left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1} \left(v(\hat{b}_{t+1}, \theta_k) - v(\hat{b}_t, \theta_k) \right) \right], \\
R_3 &= \mathbb{E}_{\theta_*} \sum_{k=1}^{k_T} \left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1} \left(\sum_r P\left[r \mid \hat{b}_t, a_t, \theta_k \right] v(\hat{b}_{t+1}, \theta_k) - v(\hat{b}_{t+1}, \theta_k) \right) \right], \\
R_4 &= \mathbb{E}_{\theta_*} \sum_{k=1}^{k_T} \left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1} \left(r(\hat{b}_t, a_t) - r(b_t^*, a_t) \right) \right].
\end{aligned}$$

Bounding R_1. One key property of Posterior Sampling algorithms is that for given the history $\mathcal{H}_{t_k}$, the true parameter θ^* and sampled θ_k are identically distributed at the time t_k. Due to the length T_k determined and independent with θ_k, then R_1 is zero thanks to Lemma 3.

Lemma 3. *(Posterior Sampling (Ouyang et al., 2017)). In TSEETC,* t_k *is an almost surely finite* $\sigma(\mathcal{H}_{t_k})$*-stopping time. If the prior distribution* $g_0(P)$,$g_0(R)$ *is the distribution of* θ^*, *then for any measurable function* g,

$$\mathbb{E}\left[g(\theta^*) \mid \mathcal{H}_{t_k} \right] = \mathbb{E}\left[g(\theta_k) \mid \mathcal{H}_{t_k} \right].$$

Bounding R_2. The regret R_2 is the telescopic sum of value function and can be bounded as $R_2 \leq HK_T$. It solely depends on the episode number and the upper bound H of span function. As a result, R_2 reduce to a finite bound over the number of episodes k_T, which can be bounded in Lemma 2.

Bounding R_3 and R_4. The regret terms R_3 and R_4 is related with estimation error about θ. Thus we should bound the parameters' error especially in our unobserved state setting. Note that when the states are observed, it is easy to count the state-action pairs and then the confidence interval can be also constructed such as in Wang et al. (2020); Xiong et al. (2022a). Recall the definition of ϕ, ψ, we can define the empirical estimation of $\hat{P}(s' \mid s)$ and $\hat{R}(r \mid s)$ as follows:

$$\hat{P}(s' \mid s) = \frac{\phi_{s,s'}}{\|\phi_{s,\cdot}\|_1}, \hat{R}(r \mid s) = \frac{\psi_{s,r}}{\|\psi_{s,\cdot}\|_1} \tag{12}$$

We also define the pesudo count of the state-action pair (s, a) before the episode k as

$$N_{t_k}(s, a) = \|\psi_{s,\cdot}(t_k)\|_1 - \|\psi_{s,\cdot}(0)\|_1 \tag{13}$$

where $\psi_{s,\cdot}(t_k)$ represents the count of state-action $z = (s, a)$ pair before the episode k. Then we define the confidence set for episode k, for all state-action pairs, the sampled P and R satisfy,

$$\mathcal{M}_k := \left\{ P : \sum_{s' \in \mathcal{S}} \left| P(s' \mid z) - \hat{P}_k(s' \mid z) \right| \leq \beta_k(z), R : \sum_{r \in \mathcal{R}} \left| R(r \mid z) - \hat{R}_k(r \mid z) \right| \leq \beta_k(z) \right\}$$

where $\beta_k(s,a) := \sqrt{\frac{14S\log(2SAt_k)}{\max\{1, N_{t_k}(s,a)\}}}$ is chosen conservatively (Auer et al., 2008) so that $\mathcal{M}_k$ contains both M^* and M_k with high probability. Specially, for the unobserved state setting, the belief error under different parameters is upper bounded by the gap between the estimators as in Proposition 1. Then the core of the proofs lies in deriving a high-probability confidence set with our pesudo counts and show that the estimated error accumulated to T is bounded by $\sqrt{T}$. With $C_1 = L_1 + L_2SN + N + S^2$, We show the final results here and the detailed proof in Appendix B.3,B.4.

Lemma 4. *R_3 satisfies the following bound*

$$R_3 \leq 48C_1SH\sqrt{AT\log AT} + +4C_1SAH\sqrt{T} + C_1H.$$

Lemma 5. *R_4 satisfies the following bound*

$$R_4 \leq 48C_1Sr_{max}\sqrt{AT\log(AT)} + 4C_1SAr_{max}\sqrt{T} + C_1r_{max}.$$

6 Numerical Experiments

In this section, we present proof-of-concept experiments. We consider two arms and there are two hidden states for each arm. We pull just one arm each time. The learning horizon $T = 50000$, and each algorithm runs 100 iterations. The transition functions and reward functions for all arms are the same. We initialize the algorithm with uninformed Dirichlet prior on the unknown parameters. We compare our algorithm with simple heuristics ϵ-greedy (Lattimore & Szepesvári, 2020) ($\epsilon = 0.01$), and Sliding-Window UCB (Garivier & Moulines, 2011) with specified window size, Q-learning (Hu et al., 2020) and SEEU (Zhou et al., 2021). The results are shown in Figure 1. We can find that TSEETC has the minimum regret among the five algorithms.

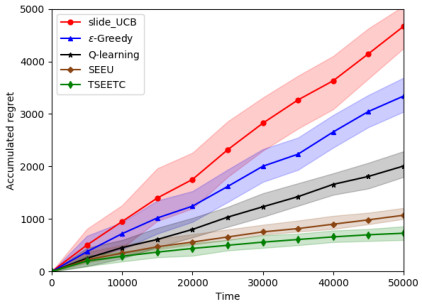

Figure 1: The cumulative regret

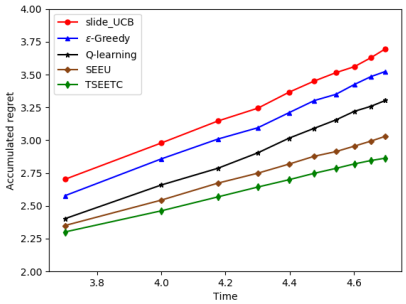

Figure 2: The log-log regret

In Figure 2, we plot the cumulative regret versus T of the five algorithms in log-log scale. We observe that the slopes of all algorithms except for our TSEETC and SEEU are close to one, suggesting that they incur linear regrets. What is more, the slope of TSEETC is close to 0.5, which is better than SEEU. This is consistent with our theoretical result.

7 Conclusion

In this paper, we consider the restless bandit with unknown states and unknown dynamics. We propose the TSEETC algorithm to estimate these unknown parameters and derive the optimal policy. We also establish the Bayesian regret of our algorithm as $\tilde{\mathcal{O}}(\sqrt{T})$ which is the first bound that matches the lower bound especially in restless bandit problems with unobserved states . Numerical results validate that the TSEETC algorithm outperforms other learning algorithms in regret. A related open question is whether our method can be applied to the setting where the transition functions are action dependent. We leave it for future research.

REFERENCES

Sahand Haji Ali Ahmad, Mingyan Liu, Tara Javidi, Qing Zhao, and Bhaskar Krishnamachari. Optimality of myopic sensing in multichannel opportunistic access. *IEEE Transactions on Information Theory*, 55(9):4040–4050, 2009.

Nima Akbarzadeh and Aditya Mahajan. Maintenance of a collection of machines under partial observability: Indexability and computation of whittle index. *arXiv preprint arXiv:2104.05151*, 2021.

Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pp. 33–1. JMLR Workshop and Conference Proceedings, 2012.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of pomdps using spectral methods. In *Conference on Learning Theory*, pp. 193–256. PMLR, 2016.

Dimitri P Bertsekas et al. Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*, 2011.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Yohann De Castro, Elisabeth Gassiat, and Sylvain Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden markov models. *IEEE Transactions on Information Theory*, 63(8):4758–4777, 2017.

Tanner Fiez, Shreyas Sekar, and Lillian J Ratliff. Multi-armed bandits for correlated markovian environments with smoothed reward feedback. *arXiv preprint arXiv:1803.04008*, 2018.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pp. 174–188. Springer, 2011.

Joey Hong, Branislav Kveton, Manzil Zaheer, Mohammad Ghavamzadeh, and Craig Boutilier. Thompson sampling with a mixture prior. In *International Conference on Artificial Intelligence and Statistics*, pp. 7565–7586. PMLR, 2022.

Zhisheng Hu, Minghui Zhu, and Peng Liu. Adaptive cyber defense against multi-stage attacks using learning-based pomdp. *ACM Transactions on Privacy and Security (TOPS)*, 24(1):1–25, 2020.

Mehdi Jafarnia Jahromi, Rahul Jain, and Ashutosh Nayyar. Online learning for unknown partially observable mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 1712–1732. PMLR, 2022.

Young Hun Jung and Ambuj Tewari. Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems*, 32, 2019.

Young Hun Jung, Marc Abeille, and Ambuj Tewari. Thompson sampling in non-episodic restless bandits. *arXiv preprint arXiv:1910.05654*, 2019.

Kailasam Lakshmanan, Ronald Ortner, and Daniil Ryabko. Improved regret bounds for undiscounted continuous reinforcement learning. In *International Conference on Machine Learning*, pp. 524–532. PMLR, 2015.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Haoyang Liu, Keqin Liu, and Qing Zhao. Learning in a changing world: Non-bayesian restless multi-armed bandit. Technical report, CALIFORNIA UNIV DAVIS DEPT OF ELECTRICAL AND COMPUTER ENGINEERING, 2010.

Haoyang Liu, Keqin Liu, and Qing Zhao. Logarithmic weak regret of non-bayesian restless multi-armed bandit. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1968–1971. IEEE, 2011.

Keqin Liu and Qing Zhao. Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access. *IEEE Transactions on Information Theory*, 56(11):5547–5567, 2010.

Aditya Mate, Jackson Killian, Haifeng Xu, Andrew Perrault, and Milind Tambe. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Aditya Mate, Andrew Perrault, and Milind Tambe. Risk-aware interventions in public health: Planning with restless multi-armed bandits. In *20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS). London, UK*, volume 10, 2021.

Rahul Meshram, D Manjunath, and Aditya Gopalan. On the whittle index for restless multiarmed hidden markov bandits. *IEEE Transactions on Automatic Control*, 63(9):3046–3053, 2018.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory*, pp. 214–228. Springer, 2012.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pp. 2701–2710. PMLR, 2017.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. *Advances in neural information processing systems*, 30, 2017.

Debjit Paria and Abhishek Sinha. Leadcache: Regret-optimal caching in networks. *Advances in Neural Information Processing Systems*, 34:4435–4447, 2021.

Zhaoqing Peng, Junqi Jin, Lan Luo, Yaodong Yang, Rui Luo, Jun Wang, Weinan Zhang, Haiyang Xu, Miao Xu, Chuan Yu, et al. Learning to infer user hidden states for online sequential advertising. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2677–2684, 2020.

Joelle Pineau, Geoff Gordon, Sebastian Thrun, et al. Point-based value iteration: An anytime algorithm for pomdps. In *IJCAI*, volume 3, pp. 1025–1032. Citeseer, 2003.

Stéphane Ross, Joelle Pineau, Brahim Chaib-draa, and Pierre Kreitmann. A bayesian approach for learning and planning in partially observable markov decision processes. *Journal of Machine Learning Research*, 12(5), 2011.

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

Steven L Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

Shang-Pin Sheng, Mingyan Liu, and Romesh Saigal. Data-driven channel modeling using spectrum measurement. *IEEE Transactions on Mobile Computing*, 14(9):1794–1805, 2014.

David Silver and Joel Veness. Monte-carlo planning in large pomdps. *Advances in neural information processing systems*, 23, 2010.

Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

Cem Tekin and Mingyan Liu. Online learning in opportunistic spectrum access: A restless bandit approach. In *2011 Proceedings IEEE INFOCOM*, pp. 2462–2470. IEEE, 2011.

Siwei Wang, Longbo Huang, and John Lui. Restless-ucb, an efficient and low-complexity algorithm for online restless bandits. *Advances in Neural Information Processing Systems*, 33:11878–11889, 2020.

Guojun Xiong, Shufan Wang, Gang Yan, and Jian Li. Reinforcement learning for dynamic dimensioning of cloud caches: A restless bandit approach. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 2108–2117. IEEE, 2022a.

Guojun Xiong, Shufan Wang, Gang Yan, and Jian Li. Reinforcement Learning for Dynamic Dimensioning of Cloud Caches: A Restless Bandit Approach. In *Proc. of IEEE INFOCOM*, 2022b.

Xiang Zhou, Yi Xiong, Ningyuan Chen, and Xuefeng Gao. Regime switching bandits. *Advances in Neural Information Processing Systems*, 34, 2021.

A TABLE OF NOTATIONS

Notation	Description
T	The length of horizon
K_T	The episode number of time T
T_k	The episode length of episode k
τ_1	The fixed exploration length in each episode
P^i	The transition functions for arm i
R^i	The reward function for arm i
P_k	The sampled transition functions
R_k	The sampled reward function
r_t	The reward obtained at time t
$b_t^i(s,\theta)$	The belief state for being in state s at time t for arm i with parameter θ
$\hat{b}_t$	The belief of all arms at time t with parameter θ_k
b_t^*	The belief of all arms at time t with parameter θ^*
a_t	The action at time t
$J(\theta_k)$	The optimal long term average reward with parameter θ_k
r_{max}	The maximum reward obtained each time
r_{min}	The minimum reward obtained each time
ΔR	The biggest gap of the obtained reward

B PROOF OF THEOREM 1

Recall that our goal is to minimize the regret :

$$R_T := \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left[J(\theta^*) - r_t\right]\right] \tag{14}$$

r_t depends on the state s_t and a_t. Due to $\mathbb{E}_{\theta_*}\left[r\left(s_t, a_t\right) \mid \mathcal{H}_{t-1}\right] = r(b_t^*, a_t)$ for any t, we have,

$$R_T := \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left[J(\theta^*) - r(b_t^*, a_t)\right]\right] \tag{15}$$

In our algorithm, each episode is split into the exploration and exploitation phase then we can rewrite the regret as:

$$R_T = \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t_k}^{t_k+\tau_1} J(\theta^*) - r\left(b_t^*, a_t\right) + \sum_{k=1}^{k_T}\sum_{t_k+\tau_1+1}^{t_{k+1}-1} J(\theta^*) - r\left(b_t^*, a_t\right)\right] \tag{16}$$

where τ_1 is the exploration length for each episode. τ_1 is a constant. t_k is the start time of episode k. Define the first part as Regret (A) which is caused by the exploration operations. The another part is defined as Regret (B) .

$$\text{Regret (A)} = \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t_k}^{t_k+\tau_1} J(\theta^*) - r\left(b_t^*, a_t\right)\right]$$

$$\text{Regret (B)} = \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t_k+\tau_1+1}^{t_{k+1}-1} J(\theta^*) - r\left(b_t^*, a_t\right)\right]$$

Recall that the reward set is $\mathcal{R}$ and we define the maximum reward gap in $\mathcal{R}$ as $\Delta R = r_{max} - r_{min}$. Then we get:

$$J(\theta^*) - r\left(b_t^*, a_t\right) \leq \Delta R$$

Then Regret (A) can be simply upper bounded as follows:

$$\text{Regret (A)} \leq \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\tau_1 \Delta R\right] \leq \tau_1 \Delta R k_T$$

Regret (A) is related with the episode number k_T obviously, which is bounded in Lemma 11. Next we should bound the term Regret (B).

During the episode k, based on equation 2, we get:

$$J(\theta_k) + v(\hat{b}_t, \theta_k) = r(\hat{b}_t, a_t) + \sum_r P(r \mid \hat{b}_t, a_t, \theta_k) v(b', \theta_k) \tag{17}$$

where $J(\theta_k)$ is the optimal long-term average reward when the system parameter is θ_k and $r(\hat{b}_t, a_t)$ is the expected reward we can get when the action a_t is taken for the current belief $\hat{b}_t$. b' is the updated belief based on equation 4 with parameter θ_k when the reward r is returned.

Using this equation, we proceed by decomposing the regret as:

$$\text{Regret(B)} = R_1 + R_2 + R_3 + R_4 \tag{18}$$

where

$$\begin{aligned}
R_1 &= \mathbb{E}_{\theta_*} \sum_{k=1}^{k_T} \left[(T_k - \tau_1 - 1) \left(J(\theta^*) - J(\theta_k) \right) \right] \\
R_2 &= \mathbb{E}_{\theta_*} \sum_{k=1}^{k_T} \left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1} \left(v(\hat{b}_{t+1}, \theta_k) - v(\hat{b}_t, \theta_k) \right) \right] \\
R_3 &= \mathbb{E}_{\theta_*} \sum_{k=1}^{k_T} \left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1} \left(\sum_r P\left[r \mid \hat{b}_t, a_t, \theta_k \right] v(\hat{b}_{t+1}, \theta_k) - v(\hat{b}_{t+1}, \theta_k) \right) \right] \\
R_4 &= \mathbb{E}_{\theta_*} \sum_{k=1}^{k_T} \left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1} \left(r(\hat{b}_t, a_t) - r(b_t^*, a_t) \right) \right]
\end{aligned}$$

Next we bound the four parts one by one.

B.1 Bound R_1

Lemma 6. *R_1 satisfies that $R_1 = 0$.*

Proof. Recall that:

$$R_1 = \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[(T_k - \tau_1 - 1) \left(J(\theta^*) - J(\theta_k) \right] \right.$$

For each episode, T_k is determined and is independent with θ_k. Based on Lemma 3, we know that,

$$\mathbb{E}_{\theta^*}[J(\theta^*)] = \mathbb{E}_{\theta^*}[J(\theta_k)]$$

therefore, the part R_1 is 0.

□

B.2 Bound R_2

Lemma 7. *R_2 satisfies the following bound*

$$R_2 \leq H K_T$$

where K_T is the total number of episodes until time T.

Proof. Recall that R_2 is the telescoping sum of value function at time $t+1$ and t.

$$R_2 = \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} \left[v(\hat{b}_{t+1}, \theta_k) - v(\hat{b}_t, \theta_k) \right] \right] \tag{19}$$

We consider the whole sum in episode k, then the R_2 can be rewrite as:

$$R_2 = \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[v(\hat{b}_{t_{k+1}}, \theta_k) - v(\hat{b}_{t_k+\tau_1+1}, \theta_k) \right]$$

Due to the span of $v(b, \theta)$ is bounded by H, then we can obtain the final bound,

$$R_2 \leq HK_T$$

□

B.3 Bound R_3

In this section, we first rewrite the R_3 in section B.3.1. In section B.3.2, we show the details about how to bound R_3.

B.3.1 Rewrite R_3

Lemma 8. *(Rewrite R_3) The regret R_3 can be bounded as follows:*

$$R_3 \leq H\mathbb{E}_{\theta^*} \left[\sum_{k=1}^{K_T} \sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} ||P^* - P_k||_1 \right] + H\mathbb{E}_{\theta^*} \left[\sum_{k=1}^{K_T} \sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} ||b_t^* - \hat{b}_t||_1 \right]$$
$$+ S^2 H\mathbb{E}_{\theta^*} \left[\sum_{k=1}^{k_T} \sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} \|R^* - R_k\|_1 \right]$$

where P_k is the sampled transition functions in episode k, R_k is the sampled reward functions in episode k, b_t^* is the belief at time t updated with true P^* and R^*, $\hat{b}_t$ is the belief at time t updated with sampled P_k, R_k.

Proof. The most part is similar to Jahromi et al. (2022), except that we should handle the unknown reward functions.

Recall that $R_3 = \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} \left(\sum_r P(r \mid \hat{b}_t, a_t, \theta_k) v\left(b', \theta_k\right) - v(\hat{b}_{t+1}, \theta_k) \right) \right]$

Recall that $\mathcal{H}_t$ is the history of actions and observations prior to action a_t. Conditioned on $\mathcal{H}_t$, θ^* and θ_k, the only random variable in $\hat{b}_{t+1}$ is r_{t+1}, then we can get,

$$\mathbb{E}_{\theta^*} \left[v(\hat{b}_{t+1}, \theta_k) \mid \mathcal{H}_t, \theta_k \right] = \sum_{r \in R} v\left(b', \theta_k\right) \cdot P_{\theta^*}\left(r_{t+1} = r \mid \mathcal{H}_t, \theta_k\right) \tag{20}$$

Note that $P_{\theta^*}\left(r_{t+1} = r \mid \mathcal{H}_t, \theta_k\right) = P(r \mid b_t^*, a_t, \theta^*)$, which is the probability of getting reward r given $\mathcal{H}_t$ and θ_k and can be written as follows,

$$\begin{aligned}
P_{\theta^*}\left(r_{t+1} \mid \mathcal{H}_t, \theta_k\right) &= \sum_{s'} R^*\left(r \mid s'\right) P_{\theta_*}\left(s_{t+1} = s' \mid \mathcal{H}_t, \theta_k\right) \\
&= \sum_{s'} R^*\left(r \mid s'\right) \sum_{s} P_{\theta_*}\left(s_{t+1} = s' \mid s_t = s, \mathcal{H}_t, a_t, \theta_k\right) P_{\theta_*}\left(s_t = s \mid \mathcal{H}_t, \theta_k\right) \\
&= \sum_{s} \sum_{s'} b_t^*(s) P^*\left(s' \mid s\right) R^*\left(r \mid s'\right) \\
&= P(r \mid b_t^*, a_t, \theta^*)
\end{aligned} \tag{21}$$

Therefore, we can rewrite the R_3 as follows,

$$R_3 = \mathbb{E}_{\theta^*} \left[\sum_{k=1}^{k_T} \sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} \left(\sum_{r \in R} (P(r \mid \hat{b}_t, a_t, \theta_k) - P(r \mid b_t^*, a_t, \theta^*) v\left(b', \theta_k\right) \right) \right]$$

Based on equation 21, we get

$$
\begin{aligned}
R_3 &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s \hat{b}_t(s)P_k(s' \mid s)\right)\right] \\
&\quad - \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R^*\left(r \mid s'\right)\sum_s b_t^*(s)P^*(s' \mid s)\right)\right] \\
&= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s \hat{b}_t(s)P_k(s' \mid s)\right)\right] \\
&\quad - \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s b_t^*(s)P^*(s' \mid s)\right)\right] \\
&\quad + \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s b_t^*(s)P^*(s' \mid s)\right)\right] \\
&\quad - \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R^*\left(r \mid s'\right)\sum_s b_t^*(s)P^*(s' \mid s)\right)\right]
\end{aligned}
\tag{22}
$$

Define

$$
R_3' = \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\left[\sum_s \hat{b}_t(s)P_k(s' \mid s) - \sum_s b_t^*(s)P^*(s' \mid s)\right]\right)\right]
$$

$$
R_3'' = \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right)\left[R_k\left(r \mid s'\right) - R^*\left(r \mid s'\right)\right]\sum_s b_t^*(s)P^*(s' \mid s)\right)\right]
$$

Bounding R_3'. The part R_3' can be bounded as Jahromi et al. (2022).

$$
\begin{aligned}
R_3' &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\left[\sum_s \hat{b}_t(s)P_k(s' \mid s) - \sum_s b_t^*(s)P^*(s' \mid s)\right]\right)\right] \\
&= R_3'(0) + R_3'(1)
\end{aligned}
$$

where

$$
\begin{aligned}
R_3'(0) &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s \hat{b}_t(s)P_k(s' \mid s)\right)\right] \\
&\quad - \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s b_t^*(s)P_k(s' \mid s)\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
R_3'(1) &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s b_t^*(s)P_k(s' \mid s)\right)\right] \\
&\quad - \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s b_t^*(s)P^*(s' \mid s)\right)\right]
\end{aligned}
$$

For $R_3'(0)$, because $\sum_r R_k(r \mid s') = 1$,$\sum_{s'} P_k(s' \mid s) = 1$,$v(b', \theta_k) \leq H$, we have

$$
\begin{aligned}
R_3'(0) &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v(b', \theta_k) R_k(r \mid s') \sum_s \hat{b}_t(s) P_k(s' \mid s)\right)\right] \\
&\quad - \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v(b', \theta_k) R_k(r \mid s') \sum_s b_t^*(s) P_k(s' \mid s)\right)\right] \\
&= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v(b', \theta_k) R_k(r \mid s') \sum_s (\hat{b}_t(s) - b_t^*(s) P_k(s' \mid s)\right)\right] \\
&\leq \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v(b', \theta_k) R_k(r \mid s') \sum_s |\hat{b}_t(s) - b_t^*(s)| P_k(s' \mid s)\right)\right] \\
&\leq H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_s |\hat{b}_t(s) - b_t^*(s)|\right)\right] \\
&= H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\left\|\hat{b}_t(s) - b_t^*(s)\right\|_1\right)\right]
\end{aligned}
$$

where the first inequality is due to $\hat{b}_t(s) - b_t^*(s) \leq |\hat{b}_t(s) - b_t^*(s)|$.

For the first term in $R_3'(1)$, note that conditioned on $\mathcal{H}_t, \theta_*$, the distribution of s_t is b_t^*. Furthermore, a_t is measurable with respect to the sigma algebra generated by $\mathcal{H}_t, \theta_k$ since $a_t = \pi^*(\hat{b}_t, \theta_k)$. Thus, we have

$$
\mathbb{E}_{\theta^*}\left[v(b', \theta_k) \sum_s P^*(s' \mid s) b^*(s) \mid H_t, \theta_k\right] = v(b', \theta_k) \mathbb{E}_{\theta^*}\left[P^*(s' \mid s) \mid H_t, \theta_k\right] \tag{23}
$$

$$
\mathbb{E}_{\theta^*}\left[v(b', \theta_k) \sum_s P_k(s' \mid s) b^*(s) \mid H_t, \theta_k\right] = v(b', \theta_k) \mathbb{E}_{\theta^*}\left[P_k(s' \mid s) \mid H_t, \theta_k\right] \tag{24}
$$

Substitute equation 23, equation 24 into $R_3'(1)$, we have

$$
\begin{aligned}
R_3'(1) &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v(b', \theta_k) R_k(r \mid s') (P_k(s' \mid s) - P^*(s' \mid s))\right)\right] \\
&\leq \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v(b', \theta_k) R_k(r \mid s') |P_k(s' \mid s) - P^*(s' \mid s)|\right)\right] \\
&\leq H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_{s'} |P_k(s' \mid s) - P^*(s' \mid s)|\right)\right] \\
&\leq H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\|P_k - P^*\|_1\right)\right]
\end{aligned}
$$

where the second inequality is due to $v(b', \theta_k) \leq H$ and $\sum_r R_k(r \mid s') = 1$.

Therefore we obtain the final results,

$$
R_3' \leq H\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} ||P^* - P_k||_1\right] + H\mathbb{E}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} ||b_t^* - \hat{b}_t||_1\right]
$$

Bounding R_3''. For part R_3'', note that for any fixed s', $\sum_s b_t^*(s)P^*(s' \mid s) \leq S$, therefore we can bound R_3'' as follows,

$$\begin{aligned}
R_3'' &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right)\left[R_k\left(r \mid s'\right)-R^*\left(r \mid s'\right)\right]\sum_s b_t^*(s)P^*(s' \mid s)\right)\right] \\
&\leq SH\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_{s'}\sum_r\left[R_k\left(r \mid s'\right)-R^*\left(r \mid s'\right)\right]\right)\right] \\
&\leq SH\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} S\left\|R_k-R^*\right\|_1\right] \\
&\leq S^2H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left\|R_k-R^*\right\|_1\right]
\end{aligned} \tag{25}$$

where the first inequality is due to $v\left(b',\theta_k\right) \leq H$ and $\sum_s b_t^*(s)P^*(s' \mid s) \leq S$, the second inequality is due to for any fixed s', $\sum_r \left[R_k\left(r \mid s'\right)-R^*\left(r \mid s'\right)\right] \leq \left\|R_k-R^*\right\|_1$.

B.3.2 Bound R_3

Lemma 9. *R_3 satisfies the following bound*

$$\begin{aligned}
R_3 \leq\ & 48(L_1+L_2SN+N+S^2)SH\sqrt{AT\log AT}+(L_1+L_2SN+N+S^2)H \\
&+4(L_1+L_2SN+N+S^2)SAH(T_1+K_T-\tau_1-1)
\end{aligned}$$

Proof. Recall that the R_3 is as follows:

$$R_3=\mathbb{E}_{\theta^*}\sum_{k=1}^{k_T}\left[\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r P[r \mid \hat{b}_t,a_t,\theta_k]v\left(b',\theta_k\right)-v(\hat{b}_{t+1},\theta_k)\right)\right]$$

This regret terms are dealing with the model estimation errors. That is to say, they depend on the on-policy error between the empirical estimated and the true transition functions, the estimated and the true reward functions. Thus we should bound the parameters' error especially in our unobserved state setting. Based on the parameters in our Dirichlet distribution, we can define the empirical estimation of reward function and transition functions for arm i as follows:

$$\hat{P}^i(s' \mid s)(t)=\frac{\phi_{s,s'}^i(t)}{\left\|\phi_{s,\cdot}^i(t)\right\|_1}, \qquad \hat{R}^i(r \mid s)(t)=\frac{\psi_{s,r}^i(t)}{\left\|\psi_{s,\cdot}^i(t)\right\|_1} \tag{26}$$

where $\phi_{s,s'}^i(t)$ is the parameters in the posterior distribution of P^i at time t, $\psi_{s,r}^i(t)$ is the parameters in the posterior distribution of R^i at time t. We also define the pseudo count $N_{t_k}^i(s,a)$ of the state-action pair (s,a) before the episode k for arm i as

$$N_{t_k}^i(s,a)=\left\|\psi_{s,\cdot}^i(t_k)\right\|_1-\left\|\psi_{s,\cdot}^i(0)\right\|_1$$

For notational simplicity, we use $z=(s,a)\in\mathcal{S}\times\mathcal{A}$ and $z_t=(s_t,a_t)$ to denote the corresponding state-action pair. Then based on Lemma 8 we can decompose the R_3 as follows,

$$\begin{aligned}
R_3 &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r P[r \mid \hat{b}_t,a_t,\theta_k]v\left(b',\theta_k\right)-v(\hat{b}_{t+1},\theta_k)\right)\right] \\
&= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left[\sum_r\left(P(r \mid \hat{b}_t,a_t,\theta_k)-P(r \mid b_t^*,a_t,\theta^*)\right)v(b',\theta_k)\right]\right] \\
&\leq R_3^0+R_3^1+R_3^2
\end{aligned}$$

where

$$R_3^0=H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left\|P^*-P_k\right\|_1\right]$$

$$R_3^1 = H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}||b_t^* - \hat{b}_t||_1\right]$$

$$R_3^2 = S^2H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\|R^* - R_k\|_1\right]$$

Note that the following results are all focused on one arm. We drop the arm index i, and P^* is the true transition functions, P_k is the sampled transition functions for this just one arm. We can extend the results on a arm to the aggregated large MDP based on the Lemma 12.

Bounding R_3^0. Since $0 \leq v(b', \theta_k) \leq H$ from our assumption , each term in the inner summation is bounded by

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} |\left(P^*\left(s' \mid z_t\right) - P_k\left(s' \mid z_t\right)\right)| v\left(s', \theta_k\right) \\
\leq & H\sum_{s'\in\mathcal{S}} \left|P^*\left(s' \mid z_t\right) - P_k\left(s' \mid z_t\right)\right| \\
\leq & H\sum_{s'\in\mathcal{S}} \left|P^*\left(s' \mid z_t\right) - \hat{P}_k\left(s' \mid z_t\right)\right| + H\sum_{s'\in\mathcal{S}} \left|P_k\left(s' \mid z_t\right) - \hat{P}_k\left(s' \mid z_t\right)\right|.
\end{aligned}$$

where $P^*\left(s' \mid z_t\right)$ is the true functions, $P_k\left(s' \mid z_t\right)$ is the sampled functions and $\hat{P}_k\left(s' \mid z_t\right)$ is the empirical estimation. The second inequality above in due to triangle inequality. Note that for any M_k in confidence set $\mathcal{M}_k$, for all state-action pairs, the sampled P and R satisfy,

$$\mathcal{M}_k = \left\{P : \sum_{s'\in\mathcal{S}} \left|P\left(s' \mid z\right) - \hat{P}_k\left(s' \mid z\right)\right| \leq \beta_k(z), R : \sum_{r\in\mathcal{R}} \left|R\left(r \mid z\right) - \hat{R}_k\left(r \mid z\right)\right| \leq \beta_k(z)\right\}$$

where $\beta_k(z) = \sqrt{\frac{14S\log(2At_kT)}{\max\left(1, N_{t_k}(z)\right)}}$ (Auer et al., 2008) and $N_{t_k}(s,a) = \|\psi_{s,\cdot}(t_k)\|_1 - \|\psi_{s,\cdot}(0)\|_1$. Note that $\beta_k(z)$ is the confidence set with $\delta = 1/t_k$.

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}} \left|P^*\left(s' \mid z_t\right) - \hat{P}_k\left(s' \mid z_t\right)\right| + \sum_{s'\in\mathcal{S}} \left|P_k\left(s' \mid z_t\right) - \hat{P}_k\left(s' \mid z_t\right)\right| \\
&\leq 2\beta_k\left(z_t\right) + 2\left(\mathbb{I}_{\{P^*\notin B_k\}} + \mathbb{I}_{\{P_k\notin B_k\}}\right).
\end{aligned} \tag{27}$$

We assume the length of the last episode is the biggest. Note that even the assumption does not hold, we can enlarge the sum items as $T_{K_T-1} - \tau_1$. This does not affect the order of our regret bound. With our assumption, because the all episode length is not bigger than the last episode, that is $t_{k+1} - 1 - (t_k + \tau_1) \leq T_{K_T} - \tau_1$, then we can obtain,

$$\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1}^{t_{k+1}-1} \beta_k\left(z_t\right) \leq \sum_{k=1}^{K_T}\sum_{t=1}^{T_{k_T}-\tau_1} \beta_k\left(z_t\right) \tag{28}$$

Due to $\sum_{s'\in\mathcal{S}} \left|P^*\left(s' \mid z_t\right) - \hat{P}_k\left(s' \mid z_t\right)\right| \leq 2$ is always true, then we can obtain,

$$\begin{aligned}
\sum_{k=1}^{K_T}\sum_{t=1}^{T_{k_T}-\tau_1} \min\{2, \beta_k\left(z_t\right)\} \leq & \sum_{k=1}^{K_T}\sum_{t=1}^{T_{k_T}-\tau_1} 2\mathbb{I}(N_{t_k} < T_{k_T} - \tau_1) \\
& + \sum_{k=1}^{K_T}\sum_{t=1}^{T_{k_T}-\tau_1} \mathbb{I}(N_{t_k} \geq T_{k_T} - \tau_1)\sqrt{\frac{14S\log\left(2At_kT\right)}{\max\left(1, N_{t_k}\left(z_t\right)\right)}}
\end{aligned} \tag{29}$$

Consider the first part in equation 29. Obviously, the maximum of N_{t_k} is $T_{k_T} - \tau_1$. Because there are totally SA state-action pairs, therefore, the first part in equation equation 29 can be bounded as,

$\sum_{k=1}^{K_T}\sum_{t=1}^{T_{k_T}-\tau_1} 2\mathbb{I}(N_{t_k} < T_{k_T} - \tau_1) \leq 2(T_{k_T} - \tau_1)SA$. Due to $T_{k_T} = T_1 + K_T - 1$ and Lemma 11, we get ,

$$2(T_{k_T} - \tau_1)SA = 2(T_1 + K_T - \tau_1 - 1)SA = \mathcal{O}(\sqrt{T})$$

Consider the second part in 29. Denote the $N_t(s,a)$ is the count of (s,a) before time t(not including t). Due to we just consider the exploration phase in each episode, then $N_t(s,a)$ can be calculated as follows,

$$N_t(s,a) = |\{\tau < t, \tau \in [t_k, t_k + \tau_1], k \leq k(t) : (s_\tau, a_\tau) = (s,a)\}|$$

where $k(t)$ is the episode number where the time t is in.

In the second part in equation 29, when $N_{t_k} \geq T_{k_T} - \tau_1$, based on our assumption $\tau_1 \leq \frac{T_1+K_T-1}{2}$, we can get,

$$\tau_1 \leq \frac{T_1 + K_T - 1}{2}$$
$$2\tau_1 \leq T_1 + K_T - 1 = T_{k_T}$$

therefore, $T_{k_T} - \tau_1 \geq \tau_1$. Because $N_{t_k} \geq T_{k_T} - \tau_1$, then $N_{t_k}(s,a) \geq \tau_1$. For any $t \in [t_k, t_k + \tau_1]$,we have

$$N_t(s,a) \leq N_{t_k}(s,a) + \tau_1 \leq 2N_{t_k}(s,a)$$

Therefore $N_t(s,a) \leq 2N_{t_k}(s,a)$. Next we can bound the confidence set when $N_t(s,a) \leq 2N_{t_k}(s,a)$ as follows,

$$\begin{aligned}
\sum_{k=1}^{K_T}\sum_{t=1}^{T_{k_T}-\tau_1} \beta_k(z_t) &\leq \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \sqrt{\frac{14S\log(2At_kT)}{\max(1, N_{t_k}(z_t))}} \\
&\leq \sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \sqrt{\frac{14S\log(2AT^2)}{\max(1, N_{t_k}(z_t))}} \\
&= \sum_{t=1}^{T} \sqrt{\frac{28S\log(2AT^2)}{\max(1, N_t(z_t))}} \\
&\leq \sqrt{56S\log(2AT)} \sum_{t=1}^{T} \frac{1}{\sqrt{\max(1, N_t(z_t))}}
\end{aligned} \tag{30}$$

where the second inequality in equation 30 is due to $t_k \leq T$ for all episodes and the first equality is due to $N_t(s,a) \leq 2N_{t_k}(s,a)$.

Then similar to Ouyang et al. (2017), since $N_t(z_t)$ is the count of visits to z_t, we have

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{\sqrt{\max(1, N_t(z_t))}} &= \sum_z \sum_{t=1}^{T} \frac{\mathbb{I}_{\{z_t = z\}}}{\sqrt{\max(1, N_t(z))}} \\
&= \sum_z \left(\mathbb{I}_{\{N_{T+1}(z)>0\}} + \sum_{j=1}^{N_{T+1}(z)-1} \frac{1}{\sqrt{j}} \right) \\
&\leq \sum_z \left(\mathbb{I}_{\{N_{T+1}(z)>0\}} + 2\sqrt{N_{T+1}(z)} \right) \leq 3\sum_z \sqrt{N_{T+1}(z)}
\end{aligned}$$

Since $\sum_z N_{T+1}(z) \leq T$, we have

$$3\sum_z \sqrt{N_{T+1}(z)} \leq 3\sqrt{SA\sum_z N_{T+1}(z)} = 3\sqrt{SAT} \tag{31}$$

With equation 30 and equation 31 we get

$$2H\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1} \beta_k(z_t) \leq 6\sqrt{56}HS\sqrt{AT\log(AT)} \leq 48HS\sqrt{AT\log(AT)}$$

Then we can bound the equation 28 as follows,

$$\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \beta_k(z_t) \leq 24S\sqrt{AT\log(AT)} + 2SA(T_1 + K_T - \tau_1 - 1) \tag{32}$$

Choose the $\delta = 1/T$ in Lemma 13, and based by Lemma 3, we obtain that

$$\mathbb{P}(P_k \notin B_k) = \mathbb{P}(P^* \notin B_k) \leq \frac{1}{15Tt_k^6}.$$

Then we can obtain,

$$2\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T} T_k \left(\mathbb{I}_{\{\theta^* \notin B_k\}} + \mathbb{I}_{\{\theta_k \notin B_k\}}\right)\right] \leq \frac{4}{15}\sum_{k=1}^{\infty} t_k^{-6} \leq \frac{4}{15}\sum_{k=1}^{\infty} k^{-6} \leq 1. \tag{33}$$

Therefore we obtain

$$2H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T} T_k \left(\mathbb{I}_{\{\theta^* \notin B_k\}} + \mathbb{I}_{\{\theta_k \notin B_k\}}\right)\right] \leq H. \tag{34}$$

Therefore, we can obtain the bound for one arm as follows,

$$\begin{aligned} &\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T} \sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} \left(\sum_{s' \in \mathcal{S}} \left(P^*(s' \mid z_t) - P_k(s' \mid z_t)\right) v(s', \theta_k)\right)\right] \\ &\leq H + 4SAH(T_1 + K_T - \tau_1 - 1) + 48HS\sqrt{AT\log(AT)} \end{aligned} \tag{35}$$

Next we consider the state transition of all arms. Recall that the states of all arms at time t is s_t. Because every arm evolves independently, then the transition probability from state s_t to state s_{t+1} is as follows,

$$P(s_{t+1} \mid s_t, \theta^*) = \prod_{i=1}^{N} P_i^*\left(s_{t+1}^i \mid s_t^i\right)$$

where P_i^* is the true transition functions of arm i. Based by the Lemma 12 and our assumption that all arms have the same state space S, we can obtain

$$\begin{aligned} \sum_{s_{t+1}} \left|P(s_{t+1} \mid s_t, \theta^*) - P(s_{t+1} \mid s_t, \theta_k)\right| &\leq \sum_{i}^{N} \left\|P_i^*\left(s_{t+1}^i \mid s_t^i\right) - P_k^i\left(s_{t+1}^i \mid s_t^i\right)\right\|_1 \\ &\leq N\left\|P_i^*\left(s_{t+1}^i \mid s_t^i\right) - P_k^i\left(s_{t+1}^i \mid s_t^i\right)\right\|_1 \end{aligned} \tag{36}$$

where P_i^* is the true transition functions of arm i. Therefore, we can bound the R_3^0 as follows:

$$R_3^0 \leq NH + 4SANH(T_1 + K_T - \tau_1 - 1) + 48SNH\sqrt{AT\log(AT)} \tag{37}$$

Bounding R_3^1. Based on the Proposition 1, we know that

$$\left\|b_t^* - \hat{b}_t\right\|_1 \leq L_1\|R^* - R_k\|_1 + L_2\|P^* - P_k\|_F$$

Based on the facts about the functions norm, we known that,

$$\|P^* - P_k\|_F \leq S\|P^* - P_k\|_1$$

Therefore , we can bound the belief error at any time as follows:

$$\left\|b_t^* - \hat{b}_t\right\|_1 \leq L_1\|R^* - R_k\|_1 + L_2 S\|P^* - P_k\|_1 \tag{38}$$

Recall in the confidence for M_k, the error bound is the same for $\|R^* - R_k\|_1$ and $\|P^* - P_k\|_1$, and based by the bound in equation 32 and equation 33, we can bound the R_3^1 as follows:

$$\begin{aligned} R_3^1 &\leq H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(L_1\|R^* - R_k\|_1 + L_2 S\|P^* - P_k\|_1\right)\right] \\ &\leq (L_1 + L_2 SN)H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\left(2\beta_k\left(z_t\right) + 2\left(\mathbb{I}_{\{P^*\notin B_k\}} + \mathbb{I}_{\{P_k\notin B_k\}}\right)\right)\right] \\ &\leq 48(L_1 + L_2 SN)SH\sqrt{AT\log AT} + (L_1 + L_2 SN)H \\ &\quad 4(L_1 + L_2 SN)SAH(T_1 + K_T - \tau_1 - 1) \end{aligned} \tag{39}$$

Bounding R_3^2. Based on equation 32 and equation 33, we can bound R_3^2 as follows,

$$\begin{aligned} R_3^2 &= S^2 H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\|R^*(\cdot \mid s) - R_k(\cdot \mid s)\|_1\right] \\ &\leq S^2 H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(2\beta_k\left(z_t\right) + 2\left(\mathbb{I}_{\{R^*\notin B_k\}} + \mathbb{I}_{\{R_k\notin B_k\}}\right)\right)\right] \\ &\leq HS^2 + 4S^3 AH(T_1 + K_T - \tau_1 - 1) + 48HS^3\sqrt{AT\log(AT)} \end{aligned} \tag{40}$$

where the second inequality is because the error bound is the same for $\|R^* - R_k\|_1$ and $\|P^* - P_k\|_1$.

Combine the bound in equation 37, equation 39 and equation 40, we bound the term R_3 as follows:

$$\begin{aligned} R_3 &\leq 48(L_1 + L_2 SN)SH\sqrt{AT\log AT} + 4(L_1 + L_2 SN)SAH(T_1 + K_T - \tau_1 - 1) \\ &\quad + (L_1 + L_2 SN)H + NH + 4SANH(T_1 + K_T - \tau_1 - 1) + 48SNH\sqrt{AT\log(AT)} \\ &\quad + HS^2 + 4S^3 AH(T_1 + K_T - \tau_1 - 1) + 48HS^3\sqrt{AT\log(AT)} \\ &= 48(L_1 + L_2 SN + N + S^2)SH\sqrt{AT\log AT} + (L_1 + L_2 SN + N + S^2)H \\ &\quad + 4(L_1 + L_2 SN + N + S^2)SAH(T_1 + K_T - \tau_1 - 1) \end{aligned} \tag{41}$$

□

B.4 Bound R_4

Lemma 10. *R_4 satisfies the following bound*

$$\begin{aligned} R_4 &\leq 48(L_1 + L_2 SN + N + S^2)Sr_{max}\sqrt{AT\log(AT)} + (L_1 + L_2 SN + N + S^2)r_{max} \\ &\quad + 4(L_1 + L_2 SN + N + S^2)SAr_{max}(T_1 + K_T - \tau_1 - 1) \end{aligned}$$

Proof. We can rewrite the R_4 as follows:

$$\begin{aligned} R_4 &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_s r_k\left(s, a_t\right)\hat{b}_t(s) - \sum_s r^*\left(s, a_t\right)b_t^*(s)\right)\right] \\ &\leq \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s, a_t\right)\hat{b}_t(s) - \sum_s r_k\left(s, a_t\right)b_t^*(s) + \sum_s r_k\left(s, a_t\right)b_t^*(s) - \sum_s r^*\left(s, a_t\right)b_t^*(s)\right)\right] \end{aligned} \tag{42}$$

where $r_k\left(s, a_t\right) = \sum_r rR_k(r \mid s)$ is the expect reward conditioned on the state s and a_t, when the reward function is R_k. And $r^*\left(s, a_t\right) = \sum_r rR^*(r \mid s)$ is the expect reward conditioned on the state s and a_t,with the true reward function. The equation 42 is due to the add the term $\sum_s r_k\left(s, a_t\right)b_t^*(s)$ and subtract it.

Denote

$$R_4^0 = \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s, a_t\right)\hat{b}_t(s) - \sum_s r_k\left(s, a_t\right)b_t^*(s)\right)\right]$$

$$R_4^1 = \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s,a_t\right) b_t^*(s) - \sum_s r^*\left(s,a_t\right) b_t^*(s)\right)\right]$$

For R_4^0,

$$\begin{aligned} R_4^0 &= \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s,a_t\right) \hat{b}_t(s) - \sum_s r_k\left(s,a_t\right) b_t^*(s)\right)\right] \\ &= \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s,a_t\right) \left(\hat{b}_t(s) - b_t^*(s)\right)\right)\right] \\ &\leq r_{max}\mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s \left|\hat{b}_t(s) - b_t^*(s)\right|\right)\right] \end{aligned} \tag{43}$$

where the last inequality is due to the fact $r_k\left(s,a_t\right) \leq r_{max}$.

For R_4^1,

$$\begin{aligned} R_4^1 &= \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s,a_t\right) b_t^*(s) - \sum_s r^*\left(s,a_t\right) b_t^*(s)\right)\right] \\ &= \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s \left[r_k\left(s,a_t\right) - r^*\left(s,a_t\right)\right] b_t^*(s)\right)\right] \\ &\leq \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s \left|r_k\left(s,a_t\right) - r^*\left(s,a_t\right)\right|\right)\right] \\ &\leq \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s \sum_r r\left|R_k\left(r \mid s\right) - R^*\left(r \mid s\right)\right|\right)\right] \\ &\leq S r_{max}\mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\left\|R_k - R^*\right\|_1\right)\right] \end{aligned} \tag{44}$$

where the first inequality in 44 is due to $b_t^*(s) \leq 1$, $r_k\left(s,a_t\right) - r^*\left(s,a_t\right) \leq \left|r_k\left(s,a_t\right) - r^*\left(s,a_t\right)\right|$ and the second inequality is due to $\sum_r \left[R_k\left(r \mid s\right) - R^*\left(r \mid s\right)\right] \leq \left\|R_k - R^*\right\|_1$.

Based on the equation 39, we can bound the R_4^0,

$$\begin{aligned} R_4^0 \leq\ & 48(L_1 + L_2SN)Sr_{max}\sqrt{AT\log AT} + (L_1 + L_2SN)r_{max} \\ & + 4(L_1 + L_2SN)SAr_{max}(T_1 + K_T - \tau_1 - 1) \end{aligned}$$

Note that for any M_k in confidence set $\mathcal{M}_k$, the reward function satisfies, :

$$\mathcal{M}_k := \left\{R : \sum_{r\in\mathcal{R}}\left|R\left(r \mid z\right) - \hat{R}_k\left(r \mid z\right)\right| \leq \beta_k(z), \forall(s,a)\right\}$$

Then based on equation 40, we get

$$R_4^1 \leq 48S^2 r_{max}\sqrt{AT\log AT} + 2S^2Ar_{max}(T_1 + K_T - \tau_1 - 1) + Sr_{max}$$

Then we can obtain the final bound:

$$\begin{aligned} R_4 \leq\ & 48(L_1 + L_2SN + S)Sr_{max}\sqrt{AT\log(AT)} + 4(L_1 + L_2SN + S)SAr_{max}(T_1 + K_T - \tau_1 - 1) \\ & + (L_1 + L_2SN + S)r_{max} \\ \leq\ & 48(L_1 + L_2SN + N + S^2)Sr_{max}\sqrt{AT\log(AT)} + (L_1 + L_2SN + N + S^2)r_{max} \\ & + 4(L_1 + L_2SN + N + S^2)SAr_{max}(T_1 + K_T - \tau_1 - 1) \end{aligned}$$

where the last inequality is due to $S \leq N + S^2$. □

B.5 THE TOTAL REGRET

Next we bound the episode number.

Lemma 11. *(Bound the episode number) With the convention* $T_1 = \left\lceil \frac{\sqrt{T}+1}{2} \right\rceil$ *and* $T_k = T_{k-1} + 1$, *the episode number is bounded by* $K_T = \mathcal{O}(\sqrt{T})$.

Proof. Note that the total horizon is T. The length of episode k is $T_k = T_1 + k - 1$. Then we can get,

$$\begin{aligned} T &= T_1 + T_2 + ... + T_{k_T} \\ &= T_1 + (T_1 + 1) + ... + (T_1 + K_T - 1) \\ &= K_T T_1 + (1 + 2 + ... + K_T - 1) \\ &= K_T T_1 + \frac{K_T(K_T - 1)}{2} \end{aligned} \tag{45}$$

Therefore,

$$K_T^2 + (2T_1 - 1)K_T - 2T = 0 \tag{46}$$

With the convention $T_1 = \left\lceil \frac{\sqrt{T}+1}{2} \right\rceil$, then we can get $K_T = \mathcal{O}(\sqrt{T})$ □

Denote $C_1 = L_1 + L_2SN + N + S^2$,$C_2 = H + r_{max}$ and $C_3 = T_1 + K_T - \tau_1 - 1$, then we can get the final regret:

$$\begin{aligned} R_T &= \text{Regret(A)} + R_1 + R_2 + R_3 + R_4 \\ &\leq \tau_1 \Delta R K_T + H K_T + 48 C_1 S H \sqrt{AT \log(AT)} + 4 C_1 C_3 S A H + C_1 H \\ &\quad + 48 C_1 S r_{max} \sqrt{AT \log(AT)} + 4 C_1 C_3 S A r_{max} + C_1 r_{max} \\ &\leq (\tau_1 \Delta R + H)\sqrt{T} + 48 C_1 S (H + r_{max}) \sqrt{AT \log(AT)} \\ &\quad + 4 C_1 S A (r_{max} + H)\sqrt{T} + C_1 (H + r_{max}) \\ &= 48 C_1 C_2 S \sqrt{AT \log(AT)} + (\tau_1 \Delta R + H + 4 C_1 C_2 S A)\sqrt{T} + C_1 C_2 \end{aligned}$$

Thus, we get the final Theorem.

Theorem 2. *Suppose Assumptions 1 and 2 hold and the Oracle returns the optimal policy in each episode. The Bayesian regret of our algorithm satisfies*

$$R_T \leq 48 C_1 C_2 S \sqrt{AT \log(AT)} + (\tau_1 \Delta R + H + 4 C_1 C_2 S A)\sqrt{T} + C_1 C_2$$

where $C_1 = L_1 + L_2SN + N + S^2, C_2 = r_{max} + H$ *are constants independent with time horizon* T, L_1 *and* L_2 *are given in Proposition 1,* τ_1 *is the exploration length in each episode,* ΔR *is the biggest gap of the reward obtained at each two different time,* H *is the bounded span,* r_{max} *is the maximum reward obtain each time.*

□

C POSTERIOR DISTRIBUTION

Note that we assume the state transition is independent of the action for each arm. Denote the states visited history from time 0 till t of arm i as $\bar{s}_t^i$ and the reward collected history is $\bar{r}_t^i$. And $N_{s,s'}^i\left(\bar{s}_t^i\right)$ is the occurence time of state evolves from s to s' for arm i in the state history $\bar{s}_t^i$. Hence, if the prior $g\left(P_i(s,\cdot)\right)$ is Dirichlet $\left(\phi_{s,s_1}^i, \ldots, \phi_{s,\mathcal{S}_i}^i\right)$, then after the observation of history $\bar{s}_t^i$, the posterior $g\left(P_i(s,\cdot) \mid \bar{s}_t^i\right)$ is Dirichlet $\left(\phi_{s,s_1}^i + N_{s,s_1}^i\left(\bar{s}_t^i\right), \ldots, \phi_{s,\mathcal{S}_i}^i + N_{s,\mathcal{S}_i}^i\left(\bar{s}_t^i\right)\right)$ (Ross et al., 2011). The posterior update for $R_i(s,\cdot)$ is similar to $P_i(s,\cdot)$. where $N_{ss'}\left(\bar{s}_t, \bar{a}_{t-1}\right) = \sum_{i=0}^{t-1} I_{\{(s,s')\}}\left(s_i, s_{i+1}\right)$ is the number of times the transition (s, s') appears in the history.

Similarly, if the prior $g\left(R_i(s,\cdot)\right)$ is Dirichlet $\left(\psi^i_{s,r_1},\ldots,\psi^i_{s,r_k}\right)$, then after the observation of reward history $\bar{r}^i_t$ and $\bar{s}^i_t$, the posterior $g\left(R_i(s,\cdot)\mid \bar{r}^i_t,\bar{s}^i_t\right)$ is Dirichlet $\left(\psi^i_{s,r_1}+N^i_{s,r_1}\left(\bar{s}^i_t,\bar{r}^i_t\right),\ldots,\psi^i_{s,r_k}+N^i_{s,r_k}\left(\bar{s}^i_t,\bar{r}^i_t\right)\right)$, and $N^i_{s,r}$ is the number of times the observation (s,r) appears in the history $\left(\bar{s}^i_t,\bar{r}^i_t\right)$.

Here we drop the arm index and consider a fixed arm. For the unknown transition functions, assume it follows the Dirichlet proir distribution . Let $\bar{r}_t=(r_1,r_2,\ldots,r_t)$ be the history of observations of the agent up to time t. Recall also that we denote $\bar{s}_t=(s_1,s_1,\ldots,s_t)$ and $\bar{a}_{t-1}=(a_1,a_2,\ldots,a_{t-1})$ the history of visited states and actions respectively. Next we show the details that how to update the posterior distribution for unknown P.

$$
\begin{aligned}
g\left(P\mid \bar{a}_{t-1},\bar{r}_{t-1}\right) =& \frac{P\left(\bar{r}_{t-1},s_t\mid P,\bar{a}_{t-1}\right)g\left(P,\bar{a}_{t-1}\right)}{\int P\left(\bar{r}_{t-1},s_t\mid P,\bar{a}_{t-1}\right)g\left(P,\bar{a}_{t-1}\right)dP}\\
=& \frac{\sum_{\bar{s}_{t-1}\in S^t}P\left(\bar{r}_{t-1},\bar{s}_t\mid P,\bar{a}_{t-1}\right)g(P)}{\int P\left(\bar{r}_{t-1},s_t\mid P,\bar{a}_{t-1}\right)g\left(P,\bar{a}_{t-1}\right)dP}\\
=& \frac{\sum_{\bar{s}_{t-1}\in S^t}g\left(P\right)\prod_{i=1}^{t}P(s_i\mid s_{i-1})}{\int P\left(\bar{r}_{t-1},s_t\mid P,\bar{a}_{t-1}\right)g\left(P,\bar{a}_{t-1}\right)dP}\\
=& \frac{\sum_{\bar{s}_{t-1}\in S^t}g\left(P\right)\left[\prod_{s,s'}P(s'\mid s)^{N_{ss'}(\bar{s}_t)}\right]}{\int P\left(\bar{r}_{t-1},s_t\mid P,\bar{a}_{t-1}\right)g\left(P,\bar{a}_{t-1}\right)dP}
\end{aligned}
$$

Next we show the Bayesian approach to learning unknown P and R with the history $(\bar{a}_{t-1},\bar{r}_t)$. Since the current state s_t of the agent at time t is unknown in the POMDP, we consider a joint posterior $g\left(s_t,P,R\mid \bar{a}_{t-1},\bar{r}_t\right)$ over s_t, P, and R. By the laws of probability and Markovian assumption of the POMDP, we have:

$$
\begin{aligned}
g\left(s_t,P,R\mid \bar{a}_{t-1},\bar{r}_{t-1}\right) \propto& P\left(\bar{r}_t,s_t\mid P,R,\bar{a}_{t-1}\right)g\left(P,R,\bar{a}_{t-1}\right)\\
\propto& \sum_{\bar{s}_{t-1}\in S^t}P\left(\bar{r}_t,\bar{s}_t\mid P,R,\bar{a}_{t-1}\right)g(P,R)\\
\propto& \sum_{\bar{s}_{t-1}\in S^t}g\left(s_0,P,R\right)\prod_{i=1}^{t}P(s_i\mid s_{i-1})R(r_i\mid s_i)\\
\propto& \sum_{\bar{s}_{t-1}\in S^t}g\left(s_0,P,R\right)\left[\prod_{s,s'}\left(P(s'\mid s)\right)^{N_{ss'}(\bar{s}_t)}\right]\times\\
& \left[\prod_{s,r}\left(R(r\mid s)\right)^{N_{sr}(\bar{s}_t,\bar{r}_{t-1})}\right]
\end{aligned}
$$

where $g\left(s_0,P,R\right)$ is the joint prior over the initial state s_0, transition function P, and reward function R; $N_{ss'}\left(\bar{s}_t\right)=\sum_{i=0}^{t-1}I_{\{(s,s')\}}\left(s_i,s_{i+1}\right)$ is the number of times the transition (s,s') appears in the history of state-action $(\bar{s}_t)$; and $N_{sr}\left(\bar{s}_t,\bar{r}_{t-1}\right)=\sum_{i=1}^{t}I_{\{(s,r)\}}\left(s_i,r_i\right)$ is the number of times the observation (s,r) appears in the history of state-rewards $(\bar{s}_t,\bar{r}_{t-1})$.

D Technical Results

Proposition 1. *(Controlling the belief error (Zhou et al., 2021)). Suppose Assumption 1 holds. Given (R_k,P_k), an estimator of the true model parameters (R^*,P^*). For an arbitrary reward-action sequence $\bar{r}_t,\bar{a}_t$, let $\hat{b}_t(\cdot,R_k,P_k)$ and $b_t(\cdot,R^*,P^*)$ be the corresponding beliefs in period t under (R_k,P_k) and (R^*,P^*) respectively. Then there exists constants L_1,L_2 such that*

$$
\left\|b_t(\cdot,R^*,P^*)-\hat{b}_t(\cdot,R_k,P_k)\right\|_1\leq L_1\|R_k-R^*\|_1+L_2\|P_k-P^*\|_F
$$

where $L_1=4M\left(\frac{1-\epsilon}{\epsilon}\right)^2/\min\left\{R^*_{\min},1-R^*_{\max}\right\}$, $L_2=4M(1-\epsilon)^2/\epsilon^3+\sqrt{M}$, M is the Cartesian product of each arms' state space, $R^*_{\max}$ and $R^*_{\min}$ are the maximum and minimum element of the functions R^* respectively.

Lemma 12. *(Lemma 13 in Jung et al. (2019)) Suppose a_k and b_k are probability distributions over a set $[n_k]$ for $k \in [K]$. Then we have*

$$\sum_{x \in \otimes_{k=1}^{K}[n_k]} \left| \prod_{k=1}^{K} a_{k,x_k} - \prod_{k=1}^{K} b_{k,x_k} \right| \leq \sum_{k=1}^{K} \|a_k - b_k\|_1$$

Lemma 13. *(Lemma 17 in Auer et al. (2008)) For any $t \geq 1$, the probability that the true MDP M is not contained in the set of plausible MDPs $\mathcal{M}(t)$ at time t is at most $\frac{\delta}{15t^6}$, that is*

$$\mathbb{P}\{M \notin \mathcal{M}(t)\} < \frac{\delta}{15t^6}$$