Online Restless Bandits with Unobserved States

Bowen Jiang [1] **Bo Jiang** [1] **Jian Li** [2] **Tao Lin** [3] **Xinbing Wang** [1] **Chenghu Zhou** [4]

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 pulled arm. The agent (decision maker) does not know the transition functions and reward functions, and cannot observe the states of arms even after pulling. 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 and each episode is divided into exploration and exploitation phases. During the exploration phase, samples of action-reward pairs are collected in a round-robin fashion and utilized to update the posterior distribution as a mixture of Dirichlet distributions. At the beginning of the 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. We show through simulations that TSEETC outperforms existing algorithms in regret.

1. Introduction

The restless multi-armed bandits (RMAB) is a general setup to model many sequential decision making problems ranging from wireless communication (Tekin & Liu, 2011; Sheng et al., 2014; Xiong et al., 2022b), 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 make a state transition independently. The state transitions can be action-dependent or not, and we consider the action-independent case. That is to say, the state of each arm makes one transition per time slot regardless of being pulled or not. More importantly, the transition function remains the same when pulling or not pulling. 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 assume the arms' states are known and concentrate on learning unknown parameters (Liu et al., 2010; 2011; Ortner et al., 2012; Wang et al., 2020; Xiong et al., 2022a;d;c). However, the arms' states are often unobserved 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. In the recommender system, we do not know the user's preference for the items. There are 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) or there is no discussion about the regret bound (Peng et al., 2020; Hu et al., 2020).

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; Ortner et al., 2012; Wang et al., 2020). However, existing policies may not perform close to the optimal offline policy, e.g., Liu et al. (2010) only considers the best policy that constantly pulls a fixed arm, which is not optimal for RMAB problems. Ortner et al. (2012) derives the lower bound $\tilde{\mathcal{O}}(\sqrt{T})$ for RMAB problems. Another way to estimate the unknown parameters is Thompson Sampling (TS) (Jung & Tewari, 2019; Jung et al., 2019; Jahromi et al., 2022; Hong et al., 2022). TS

[1]Shanghai Jiao Tong University, Shanghai, China. [2]SUNY-Binghamton University, Binghamton, NY, USA. [3]Communication University of China, Beijing, China. [4]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. Correspondence to: Bo Jiang <bjiang@sjtu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

algorithms do not 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 bandits 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). Mate et al. (2020), and Liu & Zhao (2010) derive the optimal index policy but they assume the parameters are known. Restless-UCB in Wang et al. (2020) achieves the regret bound of $\tilde{\mathcal{O}}(T^{2/3})$ and their algorithm is restricted to restless bandit problems with birth-death state Markov chains. A general Markov chain is considered in (Xiong et al., 2022c) with a regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$ guarantee. There are also some works that consider that the arm 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; Yemini et al., 2021) and the classical POMDP setting (Jahromi et al., 2022). Meshram et al. (2018) and Akbarzadeh & Mahajan (2021) study the RMAB problem with unobserved states but with known parameters. However, the true value of the parameters are often unavailable in practice. Some works study POMDP problem from a learning perspective, e.g., Peng et al. (2020); Hu et al. (2020), but there is no regret analysis. Under the unobserved states setting, the state-of-the-art algorithm achieves $\tilde{\mathcal{O}}(T^{2/3})$ bound on the frequentist regret (Zhou et al., 2021). Yemini et al. (2021) considers the arms are modulated by two unobserved states and with linear reward. This linear structure is quite a bit of side information that the decision maker can take advantage of and a instance-dependent regret bound of $\log(T)$ is given.

To the best of our knowledge, there is no known policy that performs close to the offline optimum with a provable regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$ for online restless bandits with unobserved states even after pulling. The unobserved states and unknown parameters bring many challenges. First, we need to control estimation error about states, which are not directly observed. Second, the error depends on the model parameters in a complex way via Bayesian updating and the parameters are still unknown. Third, since the state is not fully observable, the decision-maker cannot keep track of the number of visits to state-action pairs, a quantity that is crucial in the theoretical analysis. To deal with this challenge, we design a learning algorithm TSEETC to estimate these unknown parameters and update the posterior distribution about unknown parameters as mixture of Dirichlet distributions. We define the pseudo-count about the number of visits to state-action based on Dirichlet distribution and with this pseudo-count we obtain a bound about parameters' estimation errors. Benchmarked on a stronger oracle, we show that our algorithm achieves a bound $\tilde{\mathcal{O}}(\sqrt{T})$ in Bayesian regret. 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 of which consists of exploration and exploitation phases. During the exploration phase, we utilize a mixture of Dirichlet distributions to update posterior distributions and estimate unknown parameters. The belief state is implemented to encode previous historical information for unobserved states. In the exploitation phase, we sample parameters from the posterior distribution and derive an optimal policy based on the sampled parameter. Furthermore, we design increasing episode lengths to control the total number of episodes, 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. And we define the pseudo-count to store the state-action pairs. 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 is the first $\tilde{\mathcal{O}}(\sqrt{T})$ Bayesian regret bound in the setting with unknown parameters and unobserved states even after pulling the arm.

Experiment results. We conduct the proof-of-concept experiments, and compare our policy with existing baseline algorithms. Our results show that TSEETC outperforms existing algorithms in regret and the regret order is consistent with our theoretical result.

2. Related Work

We review the related works in two main domains: learning algorithm for unknown parameters, and methods to identify 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., 2022d;c) 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. Apart from these works, Thompson 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 for the online restless bandit with partially observed states. 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 unobserved (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 with observed states about pulled arms 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 $\tilde{\mathcal{O}}(T^{2/3})$. 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 $\tilde{\mathcal{O}}(T^{2/3})$.

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 discrete 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 function $R^i \in \mathbb{R}^{\mathcal{S}^i \times \mathcal{R}}$, 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 $\mathcal{S}$ 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_0^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 observed, 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^i to denote the unknown P^i and R^i for arm i and denote θ as the unknown P^i and R^i for all $i \in [N]$ collectively. Since the true states are unobservable, the agent maintains a belief state $b_t^i = [b_t^i(s, \theta^i), s \in \mathcal{S}] \in \Delta_{\mathcal{S}}$ for each arm i, where

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

and $\Delta_{\mathcal{S}} := \left\{b \in \mathbb{R}_+^{\mathcal{S}} : \sum_{s \in \mathcal{S}} b(s) = 1\right\}$ is the probability simplex in $\mathbb{R}^{\mathcal{S}}$. Note that $b_t^i(s, \theta^i)$ depends on the unknown model parameter θ^i, which itself has to be learned by the agent. We aggregate all arms as a whole Markov chain $\mathcal{M}$ and denote its transition matrix and reward function as P and R, respectively. Note that the states of the arms at any given time t are independent, since the initial states are independent and they also evolve independently. As a consequence, 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). Thus the agent can base his decision at time t on b_t only. Let $\Delta_b = \bigtimes_{i=1}^N \Delta_{\mathcal{S}}$ be the state space of the overall belief state of the system. 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 of the initial distribution h as in Jahromi et al. (2022) and denote it by $J(\theta)$. We make the following assumptions.

Assumption 3.1. The smallest element ϵ_1 in the transition functions $P^i, i \in N$ is larger than zero.

Assumption 3.2. The smallest element ϵ_2 in the reward functions $R^i, i \in N$ is larger than zero.

Assumption 3.1 and Assumption 3.2 are strong in general, but they help us bound the error of belief estimation (De Castro et al., 2017). Assumption 3.1 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 (2). Since the value function $v(,\theta)$ is finite, we can bound the span function $\mathrm{sp}(\theta) := \max_b v(b,\theta) - \min_b v(b,\theta)$ as in Zhou et al. (2021). We show the details about this bound in Proposition D.1 and denote the bound by H.

We consider the Bayesian regret. 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. If an oracle knows P^i, R^i and underlying state s_t^i, the problem becomes simple as the agent would select $a_t^* = \mathrm{argmax}_{a \in \mathcal{N}} r_t R^{at}\left(r_t \mid s_t^{at}\right)$ where $R^{at}\left(r_t \mid s_t^{at}\right)$ is the reward function of the pulled arm a_t and r_t is the obtained reward. If we benchmark a learning policy against the oracle, then the regret must be linear in T, because the oracle always observes s_t while the agent cannot predict the transition based on the history. Whenever a transition occurs, there is a non-vanishing regret incurred. Since the number of transitions during the time interval $[0,T]$ is linear in T, the total regret is of the same order. Since comparing to the oracle knowing s_t is uninformative, we consider such an oracle that assumes the unknown states and known parameters. 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 it with new observation. In section 4.2, we show how to update the posterior distributions under unknown states. In section 4.3, we show the details about our learning algorithm TSEETC.

4.1. Belief Encoder for Unobserved State

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

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

where the $P^i(s' \mid s)$ is the probability of transitioning from state s at time t to state s' 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:

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

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

4.2. Mixture of Dirichlet Distribution

In this section, we estimate the unknown P^i and R^i based on 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 $f(p \mid \phi) \propto \prod_{i=1}^k p_i^{\phi_i - 1}$ (Ghavamzadeh et al., 2015).

Since the true states are unobserved, all state sequences (and their corresponding Dirichlet posteriors) should be considered, with some weight proportional to the likelihood of each state sequence (Ross et al., 2011). Denote the reward history collected from time t_1 till t_2 (not including t_2) for arm i as $r_{t_1:t_2}^i$ and similarly the states history is denoted as $s_{t_1:t_2}^i$. And the belief state history is denoted as $b_{t_1:t_2}^i$. Recall that we assume the smallest element in the transition functions and reward functions are ϵ_1 and ϵ_2, respectively. To satisfy this, we can assume the transition function P^i takes the form $P^i = \epsilon_1 \mathbf{1} + (1 - S\epsilon_1)\tilde{P}^i$, where $\tilde{P}^i$ follows the Dirichlet distribution and $\mathbf{1}$ is the vector with one in each position. Similarly, we assume the reward function R^i takes the form $R^i = \epsilon_2 \mathbf{1} + (1 - S\epsilon_2)\tilde{R}^i$, where $\tilde{R}^i$ also follows the Dirichlet distribution. The element $\mathbf{1}$ can have different lengths in correspondence with the dimension of

P^i and R^i. Then with these history information $b^i_{t_1:t_2}$ and $r^i_{t_1:t_2}$, the posterior distribution $g_t(P^i)$ and $g_t(R^i)$ at time t can be updated as in Lemma 4.1.

Lemma 4.1. *Assuming the transition function P^i has prior $g_0\left(P^i\right) = f(\frac{P^i-\epsilon_1\mathbf{1}}{1-S\epsilon_1} \mid \phi^i)$, and the reward function R^i has prior $g_0\left(R^i\right) = f(\frac{R^i-\epsilon_2\mathbf{1}}{1-S\epsilon_2} \mid \psi^i)$, given the information $r^i_{0:t}$ and $b^i_{0:t}$, the posterior distributions in the unobserved state setting are as follows:*

$$g_t\left(P^i\right) \propto \sum_{s^i_{0:t}\in\mathcal{S}^t} \left(g_0\left(P^i\right) w(s^i_{0:t})\times \prod_{s,s'}\left(\frac{P^i(s' \mid s)-\epsilon_1}{1-\epsilon_1}\right)^{N^i_{s,s'}(\bar{s}^i_t)+\phi^i_{s,s'}-1}\right), \tag{6}$$

and

$$g_t\left(R^i\right) \propto \sum_{s^i_{0:t}\in\mathcal{S}^t} \left(g_0\left(R^i\right) w(s^i_{0:t})\times \prod_{s,r}\left(\frac{R^i(r \mid s)-\epsilon_2}{1-\epsilon_2}\right)^{N^i_{s,r}(\bar{s}^i_t)+\psi^i_{s,r}-1}\right), \tag{7}$$

where $w(s^i_{0:t})$ is the likelihood of state sequence $s^i_{0:t}$ and $\mathcal{S}^t$ is the all possible states sequences from time 0 to $t-1$. ϕ^i and ψ^i are the count vectors for the transition matrix and reward function of arm i, respectively.

This procedure is summarized in Algorithm 1. In line 2-3, we consider all the possible state transition sequences and calculate their corresponding weights. Then we derive the Dirichlet distribution related to the specific sequence (in line 4-8). In line 9, we update the posterior distribution as the mixture Dirichlet distribution.

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

1: Input: the history length τ_1, the state space $\mathcal{S}$, the belief history $b^i_{0:\tau_1}$, the reward history $r^i_{0:\tau_1}$, the initial parameters $\phi^i_{s,s'}$, $\psi^i_{s,r}$, for $s, s' \in \mathcal{S}$, $r \in \mathcal{R}$,
2: generate $\mathcal{S}^{\tau_1}$ possible state sequences
3: calculate the weight $w(j) = \prod_{t=0}^{\tau_1-1} b^i_t(s,\theta), j \in \mathcal{S}^{\tau_1}$
4: **for** j in $1,\dots,\mathcal{S}^{\tau_1}$ **do**
5: count the occurence times of event (s,s') and (s,r) as $N^i_{s,s'}, N^i_{s,r}$ in sequence j
6: update $\phi^i_{s,s'} \leftarrow \phi^i_{s,s'} + N^i_{s,s'}$, $\psi^i_{s,r} \leftarrow \psi^i_{s,r} + N^i_{s,r}$
7: aggregate the $\phi^i_{s,s'}$ as $\phi(j)$, $\psi^i_{s,r}$ as $\psi(j)$ for all $s, s' \in \mathcal{S}, r \in \mathcal{R}$
8: **end for**
9: update the mixture Dirichlet distribution $g_{\tau_1}(P^i) \propto \sum_{j=1}^{\mathcal{S}^{\tau_1}} w(j) f(\frac{P^i-\epsilon_1\mathbf{1}}{1-S\epsilon_1} \mid \phi(j))$, $g_{\tau_1}(R^i) \propto \sum_{j=1}^{\mathcal{S}^{\tau_1}} w(j) f(\frac{R^i-\epsilon_2\mathbf{1}}{1-S\epsilon_2} \mid \psi(j))$

Algorithm 2 Thompson Sampling with Episodic Explore-Then-Commit

1: Input: prior $g_0(P)$,$g_0(R)$, initial belief b_0, exploration length τ_1, the first episode length T_1
2: **for** episode $k = 1, 2, \dots,$ **do**
3: start the first time of episode k, $t_k := t$
4: sample $R_{t_k} \sim g_{t_{k-1}+\tau_1}(R)$ and $P_{t_k} \sim g_{t_{k-1}+\tau_1}(P)$
5: **for** $t = t_k, t_k+1, ..., t_k+\tau_1$ **do**
6: pull the arm i for τ_1/N times in a round robin way
7: receive the reward r_t
8: update the belief b^i_t using R_{t_k}, P_{t_k} according to (4)
9: update the belief $b^j_t, j \in N \setminus \{i\}$ using P_{t_k} according to (5)
10: **end for**
11: **for** $i = 1, 2, \dots, N$ **do**
12: input the obtained $r_{t_1:t_1+\tau_1}, ..., r_{t_k:t_k+\tau_1}$, $b_{t_1:t_1+\tau_1}, ..., b_{t_k:t_k+\tau_1}$ to Algorithm 1 to update the posterior distribution $g_{t_k+\tau_1}(P)$, $g_{t_k+\tau_1}(R)$
13: **end for**
14: sample $R_{t_k+\tau_1} \sim g_{t_k+\tau_1}(P)$, $P_{t_k+\tau_1} \sim g_{t_k+\tau_1}(R)$
15: **for** i in $0, 1, \dots, N$ **do**
16: re-update the belief b^i_t from time 0 to $t_k+\tau_1$ according to $R_{t_k+\tau_1}$ and $P_{t_k+\tau_1}$
17: **end for**
18: compute $\pi^*_k(\cdot) = \text{Oracle}\left(\cdot, R_{t_k+\tau_1}, P_{t_k+\tau_1}\right)$
19: **for** $t = t_k+\tau_1+1, \cdots, t_{k+1}-1$ **do**
20: apply action $a_t = \pi^*_k(b_t)$
21: observe new reward r_{t+1}
22: update the belief b_t of all arms using (4), (5)
23: **end for**
24: **end for**

4.3. Our Algorithm

In this section, we present the details about our TSEETC algorithm. TSEETC operates in episodes with different lengths. The total number of episodes is denoted by k_T. The length of episode k is denoted as T_k and is determined by $T_k = T_1 + k - 1$, where $T_1 = \left\lceil \frac{\sqrt{T}+1}{2} \right\rceil$. Denote the first time of the episode k by t_k. Each episode is split into an exploration phase and an exploitation phase. The length of exploration phase in each episode is fixed as τ_1 such that $\tau_1 K_T = \mathcal{O}(\sqrt{T})$ and $\tau_1 \le \frac{T_1+K_T-1}{2}$. Define the sampled parameters at time t as R_t and P_t. With these notations, we show the pseudo-code about TSEETC in Algorithm 2.

In episode k, for the exploration phase (line 3-17), we first sample the R_{t_k}, P_{t_k} from the distribution $g_{t_{k-1}+\tau_1}(P)$ and $g_{t_{k-1}+\tau_1}(R)$ to update the belief states. We pull each arm for τ_1/N times in a round-robin way. For the pulled arm, we update its belief according to (4) using R_{t_k} and P_{t_k}. For the arms that are not pulled, we update its belief according to (5) using P_{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 $R_{t_k+\tau_1}$, $P_{t_k+\tau_1}$ from the posterior distribution, and re-calibrate the belief b_t based on the most recent sampled $R_{t_k+\tau_1}$, $P_{t_k+\tau_1}$. Next, we enter into the exploitation phase (line 18-23). First, we use an Oracle to derive the optimal policy π_k for the sampled parameters $R_{t_k+\tau_1}$, $P_{t_k+\tau_1}$. 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 Remark 4.2. Then we use policy π_k for the rest of the episode k.

Our deterministic linear increment of episode length guarantees the episode number k_T is order $\mathcal{O}(\sqrt{T})$ as in Lemma B.6. Then the regret of the exploration phases can be bound by $\mathcal{O}(\sqrt{T})$, which is an crucial part in Theorem 5.1.

Remark 4.2. If the oracle returns an ϵ_k-approximate policy $\tilde{\pi}_k$ in each episode instead of the optimal policy, i.e., $r(b,\tilde{\pi}_k(b)) + \sum_r P(r \mid b,\tilde{\pi}_k(b),\theta)v(b',\theta) \leq \max_a \{r(b,a) + \sum_r P(r \mid b,a,\theta)v(b',\theta)\} - \epsilon_k$, then there will be an extra regret term$\mathbb{E}\left[\sum_{k:t_k\leq T}(T_k-\tau_1)\epsilon_k\right]$ in the exploitation phase. If we control the error as $\epsilon_k \leq \frac{1}{T_k-\tau_1}$, then this extra regret can be bounded as $\mathbb{E}\left[\sum_{k:t_k\leq T}(T_k-\tau_1)\epsilon_k\right] \leq k_T = \mathcal{O}(\sqrt{T})$ by Lemma B.6. Thus the approximation error in the computation of optimal policy does not affect the order of our regret bound.

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 5.1. *Suppose Assumptions 3.1,3.2 hold and the Oracle returns the optimal policy in each episode. The Bayesian regret of our algorithm satisfies*

$$R_T \leq 48C_1C_2S\sqrt{NT\log(NT)} + C_1C_2 + (\tau_1\Delta R + H + 4C_1C_2SN)\sqrt{T},$$

where $C_1 = L_1 + L_2N + N^2 + S^2, C_2 = r_{max} + H$ *are constants independent with time horizon* T, $L_1 = \frac{4(1-\epsilon_1)^2}{N\epsilon_1^2\epsilon_2}, L_2 = \frac{4(1-\epsilon_1)^2}{\epsilon_1^3}$, ϵ_1 *and* ϵ_2 *are the lower bounds of the functions* P^* *and* R^**, respectively.* τ_1 *is the fixed exploration length in each episode,* ΔR *is is the gap between the maximum and the minimum rewards,* H *is the bounded span,* r_{max} *is the maximum reward obtain each time,* N *is the number of arms and* S *is the state size for each arm.*

The best existing bounds on both the frequentist regret and the Bayesian regret are $\tilde{\mathcal{O}}(T^{2/3})$ in the setting with both unobserved state and unknown parameters. Our algorithm is the first to achieves the $\tilde{\mathcal{O}}(\sqrt{T})$ Bayesian regret bound on average. Whether one can achieve the $\tilde{\mathcal{O}}(\sqrt{T})$ frequentist regret bound is still open.

The key ingredients that allow us to obtain the $\tilde{\mathcal{O}}(\sqrt{T})$ bound are as follows. First, we estimate the unknown parameters based on Thompson sampling to update the posterior distribution of unknown parameters as the mixture of each combined distribution. Second, to control the regret caused by the exploration phases, we use an episodic algorithm and increase the episode length in a deterministic manner that guarantees the total episode number is order $\mathcal{O}(\sqrt{T})$, so the regret of the exploration phases is bounded by $\mathcal{O}(\sqrt{T})$. Third, we propose a novel pseudo count of the state-action pairs based on Dirichlet distribution, which allows us to bound the total estimation errors about unknown parameters and unobserved states in the exploitation phase by $\tilde{\mathcal{O}}(\sqrt{T})$.

The algorithm in Zhou et al. (2021) is also episodic and each episode is divided into an exploration phase and an exploitation phase as ours. Their cumulative regret bounds in each phase are both $\tilde{\mathcal{O}}(T^{2/3})$ in the frequentist sense. The bottleneck of their method is the spectral estimator used for parameter estimation, which has an error bound of order $1/\sqrt{k}$, where k is the episode index. To control this error, they had to use a longer exploration phase than we do, which results in a larger regret. In contrast, the regrets of our algorithm in both phases are well controlled by $\tilde{\mathcal{O}}(\sqrt{T})$, in the Bayesian sense though. Jahromi et al. (2022) considers a similar problem in the POMDP setting and obtain a Bayesian regret bound of $\tilde{\mathcal{O}}(T^{2/3})$. They define the pseudo counts of state-action pairs, but their pseudo counts are always smaller than the true counts with a nonzero probability at any time. On the other hand, in our algorithm, the sampled parameter is more concentrated around the true values with the posterior update. Therefore, our belief-based pseudo counts defined in (13) approximate the true counts more closely, which helps us obtain the final $\tilde{\mathcal{O}}(\sqrt{T})$ regret bound.

The existing work in restless bandits provide regret bounds depending on the mixing time T_{mix}. To guarantee an accuracy of $\frac{1}{T}$, the mixing time T_{mix} can be bound by $O(\log T)$ (Jung et al., 2019). Therefore, bounds depending on T_{mix} can be bound as $\mathcal{O}(\log T\sqrt{T})$ (Jung et al., 2019) and $\mathcal{O}\left(\log^{7/2} T\sqrt{T}\right)$ (Ortner et al., 2012). Our regret bound depends on the lower bounds ϵ_1 and ϵ_2 in Assumptions 3.1,3.2, which is independent with the time horizon T. Then our regret bound is $\mathcal{O}(\sqrt{T\log T})$. Therefore our regret bound improves those two bounds by a logarithmic factor.

Remark 5.2. (Continuous reward functions). We assume the reward set is finite as Jung & Tewari (2019); Zhou et al. (2021); Singh et al. (2022). However, our TSEETC algo-

rithm can be extended to handle continuous rewards. First, for unknown states, we can update the belief states incorporated with such continuous reward function. After pulling the arm and observing r, the belief state b is revised according to Bayes' theorem: $b(s') \propto \sum_s b(s)R(r \mid s)P(s' \mid s)$ (Hoey & Poupart, 2005), where $R(r \mid s)$ is a probability density function. Second, the posterior distribution $g(R)$ should be updated with continuous reward. For the case where each arm has two states, $g(R)$ is the Beta distribution. We can accordingly modify TSEETC so that after observing the reward $r_t \in [0,1]$ at time t, it performs a Bernoulli trial with success probability r_t. Let the random variable $\tilde{r}_t$ denote the outcome of this Bernoulli trial, and let $S_i(t), F_i(t)$ be the number of successes and failures in the Bernoulli trials until time t. If $\tilde{r}_t = 1$, we set $S_i(t) = S_i(t) + 1$. Otherwise, we let $F_i(t) = F_i(t) + 1$. Then we can update the parameters in $g(R)$ accordingly. We leave the theoretical analysis for continuous reward as future work.

Remark 5.3. (Thompson sampling approximation error). The establishment of the final regret bound and the total estimation error for belief states requires satisfying not only Assumption 3.1 and Assumption 3.2, but also relies crucially on the Oracle returning the optimal policy in each episode and exact posterior updates. However, in practice, marginalizing the full state sequence in Algorithm 1 is an exponentially costly task. Therefore, we approximate the posterior distribution (Urteaga & Wiggins, 2018; Lu & Van Roy, 2017) by sampling M state transition sequences, where M is a hyperparameter. As such, it is necessary to consider the impact of approximation errors on the posterior distribution (Phan et al., 2019; Mazumdar et al., 2020) in relation to the regret bound. A desired final regret bound comprises two terms: the first term is the regret bound achieved by Thompson sampling with an exact posterior, while the second term is an incremental term and accounts for posterior mismatch. Importantly, the second term converges to zero as the size of sequences considered M approaches infinity. To achieve this, we need to study the divergence between two distributions $g(P)$ and $g'(P)$ defined by different Dirichlet counts and bound the gap between the sampled transition matrix and the true parameters based on such a divergence. This is nontrivial and it deserves further study.

5.2. Proof Sketch

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:

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

where $\Delta R = r_{max} - r_{min}$ is the gap between the maximum and the minimum rewards. The regret in (9) is related to the episode number k_T. Since the first episode has length of order $\mathcal{O}(\sqrt{T})$ and the episode length is increasing linearly with the episode index, we can easily bound the total episode number by $\mathcal{O}(\sqrt{T})$ as in Lemma B.6.

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

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

With (10), 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:

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

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

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

$$R_4 = \mathbb{E}_{\theta^*}\sum_{k=1}^{k_T}\left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1}\left(r(\hat{b}_t, a_t) - r(b_t^*, a_t)\right)\right].$$

Bounding R_1. A key property of TS algorithms is that when the prior distribution g_0 coincides with that of the true parameter θ^*, given the history $\mathcal{H}_{t_k}$, the sampled θ_k has the same distribution as θ^* at time t_k as stated in Lemma 5.4. Since the length T_k is deterministic and independent of θ_k, R_1 is zero thanks to this property as stated in Lemma 5.5 .

Lemma 5.4. *(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,*

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

Lemma 5.5. *R_1 satisfies that $R_1 = 0$.*

Bounding R_2. The inner sum in R_2 is a telescopic sum that reduces to the difference of two value functions, which is upper bounded by the span and hence by H. Since the number of episodes k_T is deterministic and $O(\sqrt{T})$, we have $R_2 \leq Hk_T = O(H\sqrt{T})$ and we have Lemma 5.6.

Lemma 5.6. *R_2 is bounded by $R_2 \leq O(H\sqrt{T})$.*

Bounding R_3 and R_4. The regret terms R_3 and R_4 are related to the estimation error of θ (including the transition function P and reward function R) and estimation error of belief state. The belief estimation error can be bounded in terms of the estimation errors of θ by a result of Xiong et al. (2022e), reproduced in Proposition D.2. Thus the key is to bound the estimation error of θ. Recalling the definition of ϕ, ψ in Lemma 4.1, we define the posterior mean of $\hat{P}^i(s' \mid s)$ and $\hat{R}^i(r \mid s)$ for arm i at time t as follows:

$$\begin{aligned}\hat{P}^i(s' \mid s) =& \frac{\epsilon_1 + (1-\epsilon_1)\phi^i_{s,s'}(t)}{S\epsilon_1 + (1-\epsilon_1)\left\|\phi^i_{s,\cdot}(t)\right\|_1} \\ \hat{R}^i(r \mid s) =& \frac{\epsilon_2 + (1-\epsilon_2)\psi^i_{s,r}(t)}{S\epsilon_2 + (1-\epsilon_2)\left\|\psi^i_{s,\cdot}(t)\right\|_1}.\end{aligned} \tag{12}$$

For a fixed arm i, it can be pulled or not each time. The action a is 1 or 0 depending on whether the arm is pulled or not. Then we define the pesudo count of the state-action pair (s, a) before the episode k as

$$N^i_{t_k}(s,1) = \left\|\psi^i_{s,\cdot}(t_k)\right\|_1 - \left\|\psi^i_{s,\cdot}(0)\right\|_1, \tag{13}$$

$$N^i_{t_k}(s,0) = (\sum_{j=1}^{k-1} T_j) - N^i_{t_k}(s,1), \tag{14}$$

where $\psi^i_{s,\cdot}(t_k)$ is the parameter in the Dirichlet distribution at time t_k about reward function of arm i. Let $\mathcal{M}^i_k$ be the set of plausible MDPs in episode k with reward function $R^i(r \mid z)$ and transition function $P^i(s' \mid z)$ satisfying,

$$\begin{aligned}\sum_{s' \in \mathcal{S}} \left| P^i(s' \mid z) - \hat{P}^i_k(s' \mid z) \right| \leq \beta_k(z) \\ \sum_{r \in \mathcal{R}} \left| R^i(r \mid z) - \hat{R}^i_k(r \mid z) \right| \leq \beta_k(z),\end{aligned} \tag{15}$$

where $\beta^i_k(s,a) := \sqrt{\frac{14S\log(2Nt_kT)}{\max\{1, N^i_{t_k}(s,a)\}}}$ is chosen conservatively (Auer et al., 2008) so that $\mathcal{M}^i_k$ contains both P^i_* and P^i_k, R^i_* and R^i_k with high probability. P^i_* and R^i_* are the true parameters as we defined in Section 4.1.

The core of the proof lies in deriving a high-probability confidence set with our pseudo counts and showing that the estimated error accumulated to T for each arm is bounded by $\sqrt{T}$. Thus, we can derive the final error bound about the MDP aggregated by all arms as stated in Lemma 5.7 . The proof of Lemma 5.7 is in the Appendix B.3.2.

Lemma 5.7. *(Estimation errors of unknown parameters). Suppose Assumptions 3.1,3.2 hold and the posterior distributions are exactly updated, then the total estimation error about unknown parameters accumulated by all exploitation phases satisfies the following bound*

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} \|P^* - P_k\|_1\right] \leq& 48SN\sqrt{NT\log(NT)} \\ &+ 4SN^2\sqrt{T} + N,\end{aligned}$$

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} \|R^* - R_k\|_1\right] \leq& 48S\sqrt{NT\log(NT)} \\ &+ 4SN\sqrt{T} + 1,\end{aligned}$$

where P_k,R_k are the sampled parameters in episode k.

With Lemma 5.7, we can bound the estimation errors about belief states as stated in Lemma 5.8. The proof of Lemma 5.8 is in the Appendix B.3.2.

Lemma 5.8. *(Control belief error). Suppose Assumptions 3.1,3.2 hold and the posterior distributions are exactly updated, then the total estimation error about belief states accumulated by all exploitation phases satisfies the following bound*

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{K_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} ||b^*_t - \hat{b}_t||_1\right] \leq& 4C_1SN\sqrt{T} \\ &+ 48C_1S\sqrt{NT\log(NT)} + C_1\end{aligned}$$

The Lemma 5.8 shows that the accumulated belief errors about unobserved states is also bounded by $\tilde{\mathcal{O}}(\sqrt{T})$. Then, we can obtain the final bound about R_3, R_4 and the detailed proof in Appendix B.3,B.4.

Lemma 5.9. *R_3 satisfies the following bound*

$$R_3 \leq 48C_1SH\sqrt{NT\log NT} + 4C_1SNH\sqrt{T} + C_1H.$$

Lemma 5.10. *R_4 satisfies the following bound*

$$\begin{aligned}R_4 \leq& 48C_1Sr_{max}\sqrt{NT\log(NT)} \\ &+4C_1SNr_{max}\sqrt{T} + C_1r_{max}.\end{aligned}$$

Then the claim of the Theorem 5.1 directly follows from Lemma 5.5, Lemma 5.6, Lemma 5.9, 5.10.

6. Numerical Experiments

In this section, we present proof-of-concept experiments. To implement TSEETC efficiently, we just consider the most possible states transition sequences in the posterior update about unknown parameters. This approximation reduce the

Table 1. The average accumulated regrets of different algorithms with different arms and states

(ARMS, STATES)	TSEETC	SEEU	RUCB	Q-LEARNING	ϵ-GREEDY	SLIDE-UCB
(2, 2)	580	871	1259	1710	2653	4039
(4, 2)	9968	10253	13520	14932	16684	17690
(6, 2)	14640	25940	26932	29875	30260	33894
(8, 2)	27252	34614	35650	42261	44962	46541
(10, 2)	39635	42600	44506	49580	51540	54652
(2, 3)	4654	6065	7420	7976	8598	9590
(2, 4)	10080	11652	14064	15648	17895	18953

computational complexity and the final simulation results show that this approximated algorithm can still achieve better performance than the existing algorithms. We consider two arms and there are two hidden states (0 and 1) for each arm. We pull just one arm each time. The learning horizon $T = 50000$, and each algorithm runs 100 iterations. At state 1, the reward set is $\{10, 20\}$ and the reward set is $\{-10, 10\}$ at state 0. The transition functions and reward functions for all arms are the same. We initialize the algorithm with uninformed Dirichlet prior on the unknown parameters. The baselines include ϵ-greedy (Lattimore & Szepesvári, 2020) with $\epsilon = 0.01$, Sliding-Window UCB (Garivier & Moulines, 2011) with specified window size(equal to 50), RUCB (Liu et al., 2010), Q-learning (Hu et al., 2020), and SEEU (Zhou et al., 2021). The pseudo-counts in Jahromi et al. (2022) are related with the expectation of true counts, which can not be obtained due to the unknown states. Thus we excluded it in our experiments. The results are shown in Figure 1. We observe that approximate TSEETC has the minimum regret among these algorithms.

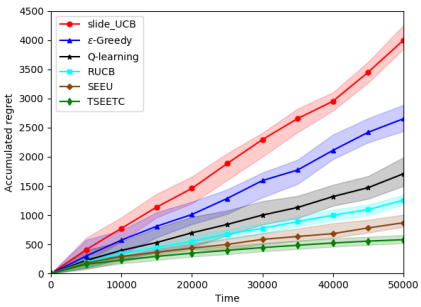

Figure 1. The cumulative regret

In Figure 2, we plot the cumulative regret versus T of the six 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.

Next, we show the robustness of TSEETC to other action and state dimensionalities. We first consider the setting with

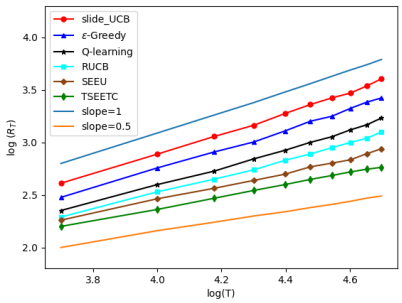

Figure 2. The log-log regret

different arms and each with the same state space. Secondly, we consider the case where the number of arms is equal, but the state spaces of each arm are different. The results are shown in Table 1. It shows that our TSEETC achieves minimal cumulative regret among all compared algorithms under different settings.

7. Conclusion

In this paper, we consider restless bandits 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 $\tilde{\mathcal{O}}(\sqrt{T})$. 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 work.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (No. 62072302, 42050105, 62262018), and the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China, China (No.SKLMCC2021KF011). We thank all reviewers for their constructive feedback.

References

Ahmad, S. H. A., Liu, M., Javidi, T., Zhao, Q., and Krishnamachari, B. Optimality of myopic sensing in multichannel opportunistic access. *IEEE Transactions on Information Theory*, 55(9):4040–4050, 2009.

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

Anandkumar, A., Hsu, D., and Kakade, S. M. 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.

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

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

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

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

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

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

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

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

Ghavamzadeh, M., Mannor, S., Pineau, J., Tamar, A., et al. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015.

Hoey, J. and Poupart, P. Solving pomdps with continuous or large discrete observation spaces. In *IJCAI*, pp. 1332–1338, 2005.

Hong, J., Kveton, B., Zaheer, M., Ghavamzadeh, M., and Boutilier, C. Thompson sampling with a mixture prior. In *International Conference on Artificial Intelligence and Statistics*, pp. 7565–7586. PMLR, 2022.

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

Jahromi, M. J., Jain, R., and Nayyar, A. Online learning for unknown partially observable mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 1712–1732. PMLR, 2022.

Jung, Y. H. and Tewari, A. Regret bounds for thompson sampling in episodic restless bandit problems. *Advances in Neural Information Processing Systems*, 32, 2019.

Jung, Y. H., Abeille, M., and Tewari, A. Thompson sampling in non-episodic restless bandits. *arXiv preprint arXiv:1910.05654*, 2019.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Liu, H., Liu, K., and Zhao, Q. Learning in a changing world: Non-bayesian restless multi-armed bandit. Technical report, California Univ Davis Dept of Electrical and Computer Engineering, 2010.

Liu, H., Liu, K., and Zhao, Q. 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.

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

Lu, X. and Van Roy, B. Ensemble sampling. *Advances in neural information processing systems*, 30, 2017.

Mate, A., Killian, J., Xu, H., Perrault, A., and Tambe, M. Collapsing bandits and their application to public health intervention. *Advances in Neural Information Processing Systems*, 33:15639–15650, 2020.

Mate, A., Perrault, A., and Tambe, M. 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.

Mazumdar, E., Pacchiano, A., Ma, Y., Jordan, M., and Bartlett, P. On approximate thompson sampling with langevin algorithms. In *International Conference on Machine Learning*, pp. 6797–6807. PMLR, 2020.

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

Ortner, R., Ryabko, D., Auer, P., and Munos, R. Regret bounds for restless markov bandits. In *International conference on algorithmic learning theory*, pp. 214–228. Springer, 2012.

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

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown markov decision processes: A thompson sampling approach. *Advances in neural information processing systems*, 30, 2017.

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

Peng, Z., Jin, J., Luo, L., Yang, Y., Luo, R., Wang, J., Zhang, W., Xu, H., Xu, M., Yu, C., 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.

Phan, M., Abbasi Yadkori, Y., and Domke, J. Thompson sampling and approximate inference. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

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

Sheng, S.-P., Liu, M., and Saigal, R. Data-driven channel modeling using spectrum measurement. *IEEE Transactions on Mobile Computing*, 14(9):1794–1805, 2014.

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

Singh, S. K., Borkar, V. S., and Kasbekar, G. S. User association in dense mmwave networks as restless bandits. *IEEE Transactions on Vehicular Technology*, 71(7):7919–7929, 2022.

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

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

Urteaga, I. and Wiggins, C. Variational inference for the multi-armed contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, pp. 698–706. PMLR, 2018.

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

Xiong, G., Li, J., and Singh, R. Reinforcement learning augmented asymptotically optimal index policy for finite-horizon restless bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8726–8734, 2022a.

Xiong, G., Qin, X., Li, B., Singh, R., and Li, J. Index-aware reinforcement learning for adaptive video streaming at the wireless edge. In *Proceedings of the Twenty-Third International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 81–90, 2022b.

Xiong, G., Wang, S., and Li, J. Learning infinite-horizon average-reward restless multi-action bandits via index awareness. *Advances in Neural Information Processing Systems*, 35:17911–17925, 2022c.

Xiong, G., Wang, S., Yan, G., and Li, J. Reinforcement Learning for Dynamic Dimensioning of Cloud Caches: A Restless Bandit Approach. In *Proc. of IEEE INFOCOM*, 2022d.

Xiong, Y., Chen, N., Gao, X., and Zhou, X. Sublinear regret for learning pomdps. *Production and Operations Management*, 31(9):3491–3504, 2022e.

Yemini, M., Leshem, A., and Somekh-Baruch, A. The restless hidden markov bandit with linear rewards and side information. *IEEE Transactions on Signal Processing*, 69:1108–1123, 2021.

Zhou, X., Xiong, Y., Chen, N., and Gao, X. 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 function for aggregated MDP
R_k	The sampled reward function for aggregated MDP
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
$r(b_t,a_t)$	The expected reward obtained when the belief state is b_t and the action is a_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 5.1

Recall that our goal is to minimize the regret :

$$R_T := \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(J(\theta^*) - r_t\right)\right]. \tag{16}$$

r_t depends on the state s_t and a_t. Thus r_t can be written as $r(s_t,a_t)$. 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,

$$R_T := \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(J(\theta^*) - r(b_t^*,a_t)\right)\right]. \tag{17}$$

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

$$R_T = \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t_k}^{t_k+\tau_1}\left(J(\theta^*) - r\left(b_t^*,a_t\right)\right) + \sum_{k=1}^{k_T}\sum_{t_k+\tau_1+1}^{t_{k+1}-1}\left(J(\theta^*) - r\left(b_t^*,a_t\right)\right)\right], \tag{18}$$

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 Regret (B) is as follows.

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

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

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:

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

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 B.6. Next we should bound the term Regret (B).

During the episode k, based on (2), we get:

$$J\left(\theta_k\right) + v(\hat{b}_t, \theta_k) = r(\hat{b}_t, a_t) + \sum_r P(r \mid \hat{b}_t, a_t, \theta_k) v\left(b', \theta_k\right), \tag{19}$$

where $J\left(\theta_k\right)$ is the optimal long-term average reward when the system parameter is θ_k, $\hat{b}_t$ is the belief at time t updated with parameter θ_k, $r(\hat{b}_t, a_t)$ 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 (4) with parameter θ_k when the reward r is received.

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

$$\text{Regret(B)} = R_1 + R_2 + R_3 + R_4, \tag{20}$$

where

$$\begin{aligned}
R_1 &= \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[\left(T_k - \tau_1 - 1\right)\left(J(\theta^*) - J(\theta_k)\right)\right], \\
R_2 &= \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1} \left(v(\hat{b}_{t+1}, \theta_k) - v(\hat{b}_t, \theta_k)\right)\right], \\
R_3 &= \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1} \left(\sum_r P(r \mid \hat{b}_t, a_t, \theta_k) v(b', \theta_k) - v(\hat{b}_{t+1}, \theta_k)\right)\right], \\
R_4 &= \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[\sum_{t_k+\tau_1+1}^{t_{k+1}-1} \left(r(\hat{b}_t, a_t) - r(b_t^*, a_t)\right)\right].
\end{aligned}$$

Next we bound the four parts one by one.

B.1. Bound R_1

Lemma B.1. *R_1 satisfies that $R_1 = 0$.*

Proof. Recall that:

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

For each episode, T_k is determined and is independent with θ_k. Based on Lemma 5.4, 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 B.2. *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.

$$R_2 = \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} \left[v(\hat{b}_{t+1}, \theta_k) - v(\hat{b}_t, \theta_k) \right] \right]. \tag{21}$$

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 as in proposition D.1 , then we can obtain the final bound,

$$R_2 \leq H k_T.$$

□

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 B.3. *(Rewrite R_3) The regret R_3 can be bounded as follows:*

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

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,

$$\mathbb{E}_{\theta^*} \left[v(\hat{b}_{t+1}, \theta_k) \mid \mathcal{H}_t, \theta_k \right] = \sum_{r \in R} v\left(b', \theta_k\right) P(r \mid b_t^*, a_t, \theta^*), \tag{22}$$

where $P(r \mid b_t^*, a_t, \theta^*)$ is the probability of getting reward r given b_t^*, a_t, θ^*. By the law of probability, $P(r \mid b_t^*, a_t, \theta^*)$ can be written as follows,

$$\begin{aligned} P(r \mid b_t^*, a_t, \theta^*) &= \sum_{s'} R^*\left(r \mid s'\right) P\left(s_{t+1} = s' \mid \mathcal{H}_t, \theta^*\right) \\ &= \sum_{s'} R^*\left(r \mid s'\right) \sum_{s} P^*\left(s_{t+1} = s' \mid s_t = s, \mathcal{H}_t, a_t, \theta^*\right) P\left(s_t = s \mid \mathcal{H}_t, \theta^*\right) \\ &= \sum_{s} \sum_{s'} b_t^*(s) P^*\left(s' \mid s\right) R^*\left(r \mid s'\right), \end{aligned} \tag{23}$$

where P^* is the transition functions for the MDP aggregated by all arms, R^* is the reward function for the aggregated MDP. 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 (23), we get

$$
\begin{aligned}
R_3 = &\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s \hat{b}_t(s) P_k(s' \mid s)\right)\right] \\
&- \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R^*\left(r \mid s'\right)\sum_s b_t^*(s) P^*(s' \mid s)\right)\right] \\
= &\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s \hat{b}_t(s) P_k(s' \mid s)\right)\right] \\
&- \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s b_t^*(s) P^*(s' \mid s)\right)\right] \\
&+ \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R_k\left(r \mid s'\right)\sum_s b_t^*(s) P^*(s' \mid s)\right)\right] \\
&- \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right) R^*\left(r \mid s'\right)\sum_s b_t^*(s) P^*(s' \mid s)\right)\right].
\end{aligned}
\tag{24}
$$

where R_k is the sampled reward function for aggregated MDP, P_k is the sampled transition function for aggregated MDP.

Define

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

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

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

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

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

where the first inequality is due to $\hat{b}_t(s) - b_t^*(s) \leq |\hat{b}_t(s) - b_t^*(s)|$ and the second inequality is because $\sum_r R_k\left(r \mid s'\right)=1$, $\sum_{s'} P_k(s' \mid s)=1$, $v\left(b', \theta_k\right) \leq H$.

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

$$
\mathbb{E}_{\theta^*}\left[v\left(b', \theta_k\right) \sum_s P^*\left(s' \mid s\right) b^*(s) \mid H_t, \theta_k\right] = v\left(b', \theta_k\right) \mathbb{E}_{\theta^*}\left[P^*\left(s' \mid s\right) \mid H_t, \theta_k\right]. \tag{25}
$$

$$
\mathbb{E}_{\theta^*}\left[v\left(b', \theta_k\right) \sum_s P_k\left(s' \mid s\right) b^*(s) \mid H_t, \theta_k\right] = v\left(b', \theta_k\right) \mathbb{E}_{\theta^*}\left[P_k\left(s' \mid s\right) \mid H_t, \theta_k\right]. \tag{26}
$$

Substitute (25), (26) into $R_3'(1)$, we have

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

where the first inequality is because $P_k(s' \mid s) - P^*(s' \mid s) \leq |P_k(s' \mid s) - P^*(s' \mid s)|$, the second inequality is due to $v\left(b', \theta_k\right) \leq H$ and $\sum_r R_k\left(r \mid s'\right)=1$.

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,

$$\begin{aligned}
R_3'' &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_r\sum_{s'} v\left(b',\theta_k\right)\left[R_k\left(r \mid s'\right)-R^*\left(r \mid s'\right)\right]\sum_s b_t^*(s)P^*(s' \mid s)\right)\right] \\
&\leq SH\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_{s'}\sum_r\left[R_k\left(r \mid s'\right)-R^*\left(r \mid s'\right)\right]\right)\right] \\
&\leq SH\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1} S\left\|R_k-R^*\right\|_1\right] \\
&\leq S^2H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left\|R_k-R^*\right\|_1\right],
\end{aligned} \tag{27}$$

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 B.4. *R_3 satisfies the following bound*

$$\begin{aligned}
R_3 \leq\ & 48(L_1+L_2N+N+S^2)SH\sqrt{NT\log(NT)}+(L_1+L_2N+N+S^2)H \\
&+4(L_1+L_2N+N^2+S^2)SNH(T_1+k_T-\tau_1-1).
\end{aligned}$$

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 sampled transition functions and the true transition functions, the sampled reward functions 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:

$$\hat{P}^i(s' \mid s) = \frac{\epsilon_1+(1-\epsilon_1)\phi^i_{s,s'}(t)}{S\epsilon_1+(1-\epsilon_1)\left\|\phi^i_{s,\cdot}(t)\right\|_1}, \quad \hat{R}^i(r \mid s) = \frac{\epsilon_2+(1-\epsilon_2)\psi^i_{s,r}(t)}{S\epsilon_2+(1-\epsilon_2)\left\|\psi^i_{s,\cdot}(t)\right\|_1}. \tag{28}$$

where $\phi^i_{s,s'}(t)$ is the parameters in the posterior distribution of P^i at time t, $\psi^i_{s,r}(t)$ is the parameters in the posterior distribution of R^i at time t. For each arm, it can be pulled or not. When it is pulled, we define the action a is 1 and the action a is 0 when it is not pulled. Then we define the pseudo count $N^i_{t_k}(s,a)$ of the state-action pair (s,a) before the episode k for arm i as

$$\begin{aligned}
N^i_{t_k}(s,1) &= \left\|\psi^i_{s,\cdot}(t_k)\right\|_1-\left\|\psi^i_{s,\cdot}(0)\right\|_1, \\
N^i_{t_k}(s,0) &= (\sum_{j=1}^{k-1}T_j)-N^i_{t_k}(s,1).
\end{aligned}$$

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 B.3 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

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

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

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

Note that the following results are all focused on one arm. Define P_*^i is the true transition function for arm i, P_k^i is the sampled transition function for arm i. We can extend the results on a arm to the aggregated large MDP based on Lemma D.3.

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

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

where $P_*^i\left(s' \mid z_t\right)$ is the true transition function, $P_k^i\left(s' \mid z_t\right)$ is the sampled reward function and $\hat{P}_k^i\left(s' \mid z_t\right)$ is the posterior mean. The second inequality above in due to triangle inequality. Let $\mathcal{M}_k^i$ be the set of plausible MDPs in episode k with reward function $R^i\left(r \mid z\right)$ and transition function $P^i\left(s' \mid z\right)$ satisfying,

$$\sum_{s'\in\mathcal{S}}\left|P^i\left(s' \mid z\right) - \hat{P}_k^i\left(s' \mid z\right)\right| \le \beta_k^i(z), \quad \sum_{r\in\mathcal{R}}\left|R^i\left(r \mid z\right) - \hat{R}_k^i\left(r \mid z\right)\right| \le \beta_k^i(z),$$

where $\beta_k^i(s,a) := \sqrt{\frac{14S\log(2Nt_kT)}{\max\left\{1, N_{t_k}^i(s,a)\right\}}}$ is chosen conservatively (Auer et al., 2008) so that $\mathcal{M}_k^i$ contains both P_*^i and P_k^i, R_*^i and R_k^i with high probability. P_*^i and R_*^i are the true parameters as we defined in section 4.1. Note that $\beta_k^i(z)$ is the confidence set with $\delta = 1/t_k$.

Then we can obtain,

$$\begin{aligned}
&\sum_{s'\in\mathcal{S}}\left|P_*^i\left(s' \mid z_t\right) - \hat{P}_k^i\left(s' \mid z_t\right)\right| + \sum_{s'\in\mathcal{S}}\left|P_k^i\left(s' \mid z_t\right) - \hat{P}_k^i\left(s' \mid z_t\right)\right| \\
&\le 2\beta_k^i\left(z_t\right) + 2\left(\mathbb{I}_{\{P_*^i\notin B_k\}} + \mathbb{I}_{\left\{P_k^i\notin B_k\right\}}\right).
\end{aligned} \tag{29}$$

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 $t_{k+1} - 1 - (t_k + \tau_1) \le T_{k_T} - \tau_1$, then we can obtain,

$$\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1}^{t_{k+1}-1}\beta_k^i\left(z_t\right) \le \sum_{k=1}^{k_T}\sum_{t=1}^{T_{k_T}-\tau_1}\beta_k^i\left(z_t\right). \tag{30}$$

Note that $\sum_{s'\in\mathcal{S}}\left|P_*^i\left(s' \mid z_t\right) - \hat{P}_k^i\left(s' \mid z_t\right)\right| \le 2$ is always true. And with our assumption $\tau_1 \le \frac{T_1+k_T-1}{2}$, it is easy to show that when $N_{t_k}^i \ge T_{k_T} - \tau_1$, $\beta_k^i\left(z_t\right) \le 2$ holds. Then we can obtain,

$$\sum_{k=1}^{k_T}\sum_{t=1}^{T_{k_T}-\tau_1}\min\{2,\beta_k^i(z_t)\} \le \sum_{k=1}^{k_T}\sum_{t=1}^{T_{k_T}-\tau_1} 2\mathbb{I}(N_{t_k}^i < T_{k_T}-\tau_1) + \sum_{k=1}^{k_T}\sum_{t=1}^{T_{k_T}-\tau_1}\mathbb{I}(N_{t_k}^i \ge T_{k_T}-\tau_1)\sqrt{\frac{14S\log(2Nt_kT)}{\max\left(1,N_{t_k}^i(z_t)\right)}}. \tag{31}$$

Consider the first part in (31). Obviously, the maximum of $N_{t_k}^i$ is $T_{k_T}-\tau_1$. Because there are totally SA state-action pairs, therefore, the first part in equation (31) can be bounded as, $\sum_{k=1}^{k_T}\sum_{t=1}^{T_{k_T}-\tau_1} 2\mathbb{I}(N_{t_k}^i < T_{k_T}-\tau_1) \le 2(T_{k_T}-\tau_1)SA$. Due to $T_{k_T} = T_1 + k_T - 1$ and Lemma B.6, 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 31. Denote the $N_t^i(s,a)$ is the count of (s,a) before time t(not including t). Due to we just consider the exploration phase in each episode, then $N_t^i(s,a)$ can be calculated as follows,

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

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

In the second part in (31), when $N_{t_k}^i \ge T_{k_T}-\tau_1$, based on our assumption $\tau_1 \le \frac{T_1+k_T-1}{2}$, we can get,

$$\tau_1 \le \frac{T_1+k_T-1}{2},$$
$$2\tau_1 \le T_1 + k_T - 1 = T_{k_T}.$$

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

$$N_t^i(s,a) \le N_{t_k}^i(s,a) + \tau_1 \le 2N_{t_k}^i(s,a).$$

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

$$\begin{aligned}\sum_{k=1}^{k_T}\sum_{t=1}^{T_{k_T}-\tau_1}\beta_k^i(z_t) &\le \sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{14S\log(2Nt_kT)}{\max\left(1,N_{t_k}^i(z_t)\right)}} \\ &\le \sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\sqrt{\frac{14S\log(2NT^2)}{\max\left(1,N_{t_k}^i(z_t)\right)}} \\ &= \sum_{t=1}^{T}\sqrt{\frac{28S\log(2NT^2)}{\max\left(1,N_t^i(z_t)\right)}} \\ &\le \sqrt{56S\log(2NT)}\sum_{t=1}^{T}\frac{1}{\sqrt{\max\left(1,N_t^i(z_t)\right)}}.\end{aligned} \tag{32}$$

where the second inequality in (32) is due to $t_k \le T$ for all episodes and the first equality is due to $N_t^i(s,a) \le 2N_{t_k}^i(s,a)$.

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

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

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

$$3\sum_z \sqrt{N_{T+1}^i(z)} \leq 3\sqrt{SN\sum_z N_{T+1}^i(z)} = 3\sqrt{SNT}. \tag{33}$$

With (32) and (33) we get

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

Then we can bound the (30) as follows,

$$\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1} \beta_k^i(z_t) \leq 24S\sqrt{NT\log(NT)} + 2SA(T_1 + k_T - \tau_1 - 1). \tag{34}$$

Choose the $\delta = 1/T$ in Lemma D.4, and based by Lemma 5.4, we obtain that

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

Then we can obtain,

$$2\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T} T_k\left(\mathbb{I}_{\{\theta^* \notin B_k\}} + \mathbb{I}_{\{\theta_k \notin B_k\}}\right)\right] \leq \frac{4}{15}\sum_{k=1}^{\infty} t_k^{-6} \leq \frac{4}{15}\sum_{k=1}^{\infty} k^{-6} \leq 1. \tag{35}$$

Therefore we obtain

$$2H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T} T_k\left(\mathbb{I}_{\{\theta^* \notin B_k\}} + \mathbb{I}_{\{\theta_k \notin B_k\}}\right)\right] \leq H. \tag{36}$$

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

$$\begin{aligned}&\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_{s'\in\mathcal{S}}\left(P_*^i(s' \mid z_t) - P_k^i(s' \mid z_t)\right)v(s', \theta_k)\right)\right] \\ &\leq H + 4SNH(T_1 + k_T - \tau_1 - 1) + 48HS\sqrt{NT\log(NT)}.\end{aligned} \tag{37}$$

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 D.3 and our assumption that all arms have the same state space S, we can obtain

$$\begin{aligned}\sum_{s_{t+1}} \left|P(s_{t+1} \mid s_t, \theta^*) - P(s_{t+1} \mid s_t, \theta_k)\right| &\leq \sum_i^N \left\|P_*^i\left(s_{t+1}^i \mid s_t^i\right) - P_k^i\left(s_{t+1}^i \mid s_t^i\right)\right\|_1 \\ &\leq N\left\|P_*^i\left(s_{t+1}^i \mid s_t^i\right) - P_k^i\left(s_{t+1}^i \mid s_t^i\right)\right\|_1.\end{aligned} \tag{38}$$

Therefore, we can bound the R_3^0 as follows:

$$R_3^0 \leq NH + 4SN^2H(T_1 + k_T - \tau_1 - 1) + 48SNH\sqrt{NT\log(NT)}. \tag{39}$$

Bounding R_3^1. Based on the Proposition D.2, we know that

$$\left\|b_t^* - \hat{b}_t\right\|_1 \leq L_1\|R^* - R_k\|_1 + L_2 \max_s \|P^*(s,:) - P_k(s,:)\|_2 .$$

Note that the elements in the true transition matrix P^* and the sampled matrix P_k are between the interval $(0,1)$. Then based on the facts about the norm, we know that

$$\max_s \|P^*(s,:) - P_k(s,:)\|_2 \leq \|P^* - P_k\|_1 .$$

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

$$\left\|b_t^* - \hat{b}_t\right\|_1 \leq L_1\|R^* - R_k\|_1 + L_2\|P^* - P_k\|_1. \tag{40}$$

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 (34) and (35), we can bound the R_3^1 as follows:

$$\begin{aligned}
R_3^1 &\leq H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}(L_1\|R^* - R_k\|_1 + L_2\|P^* - P_k\|_1)\right] \\
&\leq (L_1 + L_2N)H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k}^{t_{k+1}-1}\left(2\beta_k^i(z_t) + 2\left(\mathbb{I}_{\{P^*\notin B_k\}} + \mathbb{I}_{\{P_k\notin B_k\}}\right)\right)\right] \\
&\leq 48(L_1 + L_2N)SH\sqrt{NT\log(NT)} + (L_1 + L_2N)H \\
&\quad + 4(L_1 + L_2N)SNH(T_1 + k_T - \tau_1 - 1).
\end{aligned} \tag{41}$$

Bounding R_3^2. Based on (34) and (35), we can bound R_3^2 as follows,

$$\begin{aligned}
R_3^2 &= S^2H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\|R^*(\cdot\mid s) - R_k(\cdot\mid s)\|_1\right] \\
&\leq S^2H\mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t=t_k+\tau_1+1}^{t_{k+1}-1}\left(2\beta_k^i(z_t) + 2\left(\mathbb{I}_{\{R^*\notin B_k\}} + \mathbb{I}_{\{R_k\notin B_k\}}\right)\right)\right] \\
&\leq HS^2 + 4S^3NH(T_1 + k_T - \tau_1 - 1) + 48HS^3\sqrt{NT\log(NT)}.
\end{aligned} \tag{42}$$

Combine the bound in (39), (41) and (42), we bound the term R_3 as follows:

$$\begin{aligned}
R_3 &\leq 48(L_1 + L_2N)SH\sqrt{NT\log(NT)} + 4(L_1 + L_2N)SNH(T_1 + k_T - \tau_1 - 1) \\
&\quad + (L_1 + L_2N)H + NH + 4SN^2H(T_1 + k_T - \tau_1 - 1) + 48SNH\sqrt{NT\log(NT)} \\
&\quad + HS^2 + 4S^3NH(T_1 + k_T - \tau_1 - 1) + 48HS^3\sqrt{NT\log(NT)} \\
&= 48(L_1 + L_2N + N + S^2)SH\sqrt{NT\log(NT)} + (L_1 + L_2N + N + S^2)H \\
&\quad + 4(L_1 + L_2N + N^2 + S^2)SNH(T_1 + k_T - \tau_1 - 1).
\end{aligned} \tag{43}$$

□

B.4. Bound R_4

Lemma B.5. *R_4 satisfies the following bound*

$$\begin{aligned}
R_4 &\leq 48(L_1 + L_2N + N + S^2)Sr_{max}\sqrt{NT\log(NT)} + (L_1 + L_2N + N + S^2)r_{max} \\
&\quad + 4(L_1 + L_2N + N + S^2)SAr_{max}(T_1 + k_T - \tau_1 - 1).
\end{aligned}$$

Proof. We can rewrite the R_4 as follows:

$$\begin{aligned} R_4 &= \mathbb{E}_{\theta^*}\left[\sum_{k=1}^{k_T}\sum_{t_k+\tau_1+1}^{t_{k+1}-1}\left(\sum_s r_k\left(s,a_t\right)\hat{b}_t(s) - \sum_s r^*\left(s,a_t\right)b_t^*(s)\right)\right] \\ &\leq \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s,a_t\right)\hat{b}_t(s) - \sum_s r_k\left(s,a_t\right)b_t^*(s) + \sum_s r_k\left(s,a_t\right)b_t^*(s) - \sum_s r^*\left(s,a_t\right)b_t^*(s)\right)\right] \end{aligned} \tag{44}$$

where $r_k\left(s,a_t\right) = \sum_r r R_k^{a_t}(r \mid s)$ is the expect reward conditioned on the state s of pulled arm and a_t, when the reward function is $R_k^{a_t}$. And $r^*\left(s,a_t\right) = \sum_r r R_*^{a_t}(r \mid s)$ is the expect reward conditioned on the state s and a_t,with the true reward function $R_*^{a_t}$. The (44) 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,

$$\begin{aligned} R_4^0 &= \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s,a_t\right)\hat{b}_t(s) - \sum_s r_k\left(s,a_t\right)b_t^*(s)\right)\right] \\ &= \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s,a_t\right)\left(\hat{b}_t(s) - b_t^*(s)\right)\right)\right] \\ &\leq r_{max}\mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s \left|\hat{b}_t(s) - b_t^*(s)\right|\right)\right] \end{aligned} \tag{45}$$

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

For R_4^1,

$$\begin{aligned} R_4^1 &= \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s r_k\left(s,a_t\right)b_t^*(s) - \sum_s r^*\left(s,a_t\right)b_t^*(s)\right)\right] \\ &= \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s \left[r_k\left(s,a_t\right) - r^*\left(s,a_t\right)\right]b_t^*(s)\right)\right] \\ &\leq \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s \left|r_k\left(s,a_t\right) - r^*\left(s,a_t\right)\right|\right)\right] \\ &\leq \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\sum_s\sum_r r\left|R_k^{a_t}\left(r \mid s\right) - R_*^{a_t}\left(r \mid s\right)\right|\right)\right] \\ &\leq Sr_{max}\mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(\left\|R_k^{a_t} - R_*^{a_t}\right\|_1\right)\right] \end{aligned} \tag{46}$$

where the first inequality in 46 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 $\sum_r \left[R_k^{a_t}\left(r \mid s\right) - R_*^{a_t}\left(r \mid s\right)\right] \leq \left\|R_k^{a_t} - R_*^{a_t}\right\|_1$.

Based on the (41), we can bound the R_4^0,

$$\begin{aligned} R_4^0 \leq\ & 48(L_1 + L_2N)Sr_{max}\sqrt{NT\log(NT)} + (L_1 + L_2N)r_{max} \\ & + 4(L_1 + L_2N)SNr_{max}(T_1 + k_T - \tau_1 - 1). \end{aligned}$$

Note that for any reward function $R(r \mid z)$ in confidence set $\mathcal{M}_k$, the reward function satisfies,

$$\sum_{r\in\mathcal{R}} \left| R(r \mid z) - \hat{R}_k^i(r \mid z) \right| \leq \beta_k^i(z)$$

Then based on (42), we get

$$R_4^1 \leq 48S^2 r_{max}\sqrt{NT\log(NT)} + 2S^2 N r_{max}(T_1 + k_T - \tau_1 - 1) + S r_{max}.$$

Then we can obtain the final bound:

$$\begin{aligned} R_4 &\leq 48(L_1 + L_2N + S)Sr_{max}\sqrt{NT\log(NT)} + 4(L_1 + L_2N + S)SNr_{max}(T_1 + k_T - \tau_1 - 1) \\ &\quad + (L_1 + L_2N + S)r_{max} \\ &\leq 48(L_1 + L_2N + N + S^2)Sr_{max}\sqrt{NT\log(NT)} + (L_1 + L_2N + N + S^2)r_{max} \\ &\quad + 4(L_1 + L_2N + N + S^2)SNr_{max}(T_1 + k_T - \tau_1 - 1) \end{aligned}$$

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

B.5. The total regret

Next we bound the episode number.

Lemma B.6. *(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,

$$\begin{aligned} T &= T_1 + T_2 + ... + T_{k_T} \\ &= T_1 + (T_1 + 1) + ... + (T_1 + k_T - 1) \\ &= k_T T_1 + (1 + 2 + ... + k_T - 1) \\ &= k_T T_1 + \frac{k_T(k_T - 1)}{2}. \end{aligned} \tag{47}$$

Therefore,

$$k_T^2 + (2T_1 - 1)k_T - 2T = 0. \tag{48}$$

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 $C_1 = L_1 + L_2N + N^2 + S^2$, $C_2 = H + r_{max}$ and $C_3 = T_1 + k_T - \tau_1 - 1$, then we can get the final regret:

$$\begin{aligned} R_T &= \text{Regret(A)} + R_1 + R_2 + R_3 + R_4 \\ &\leq \tau_1 \Delta R k_T + H k_T + 48C_1 SH\sqrt{NT\log(NT)} + 4C_1C_3SAH + C_1H \\ &\quad + 48C_1 Sr_{max}\sqrt{NT\log(NT)} + 4C_1C_3SAr_{max} + C_1 r_{max} \\ &\leq (\tau_1\Delta R + H)\sqrt{T} + 48C_1S(H + r_{max})\sqrt{NT\log(NT)} \\ &\quad + 4C_1SA(r_{max} + H)\sqrt{T} + C_1(H + r_{max}) \\ &= 48C_1C_2S\sqrt{NT\log(NT)} + (\tau_1\Delta R + H + 4C_1C_2SN)\sqrt{T} + C_1C_2. \end{aligned}$$

Thus, we get the final Theorem.

Theorem B.7. *Suppose Assumptions 3.1,3.2 hold and the Oracle returns the optimal policy in each episode. The Bayesian regret of our algorithm satisfies*

$$R_T \leq 48C_1C_2S\sqrt{NT\log(NT)} + (\tau_1\Delta R + H + 4C_1C_2SN)\sqrt{T} + C_1C_2,$$

where $C_1 = L_1 + L_2N + N^2 + S^2, C_2 = r_{max} + H$ are constants independent with time horizon T, $L_1 = \frac{4(1-\epsilon_1)^2}{N\epsilon_1^2\epsilon_2}, L_2 = \frac{4(1-\epsilon_1)^2}{\epsilon_1^3}$, ϵ_1 and ϵ_2 are the minimum elements of the functions P^* and R^*, respectively. τ_1 is the fixed exploration length in each episode, ΔR is is the gap between the maximum and the minimum rewards, H is the bounded span, r_{max} is the maximum reward obtain each time, N is the number of arms and S is the state size for each arm.

□

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 $s^i_{0:t}$ and the reward collected history is $r^i_{0:t}$. And the action history from time 0 to t is $a^i_{0:t}$. Denote $N^i_{s,s'}\left(s^i_{0:t}\right)$ as the occurence time of state evolves from s to s' for arm i in the state history $s^i_{0:t}$. Hence, if the prior $g\left(P_i(s,\cdot)\right)$ is Dirichlet $\left(\phi^i_{s,s_1},\ldots,\phi^i_{s,\mathcal{S}_i}\right)$, then after the observation of history $s^i_{0:t}$, the posterior $g\left(P_i(s,\cdot) \mid s^i_{0:t}\right)$ is Dirichlet $\left(\phi^i_{s,s_1} + N^i_{s,s_1}\left(s^i_{0:t}\right),\ldots,\phi^i_{s,\mathcal{S}_i} + N^i_{s,\mathcal{S}_i}\left(s^i_{0:t}\right)\right)$ (Ross et al., 2011).

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 $r^i_{0:t}$ and $s^i_{0:t}$, the posterior $g\left(R_i(s,\cdot) \mid r^i_{0:t}, s^i_{0:t}\right)$ is Dirichlet $\left(\psi^i_{s,r_1} + N^i_{s,r_1}\left(s^i_{0:t}, r^i_{0:t}\right),\ldots,\psi^i_{s,r_k} + N^i_{s,r_k}\left(s^i_{0:t}, r^i_{0:t}\right)\right)$, and $N^i_{s,r}$ is the number of times the observation (s,r) appears in the history $\left(s^i_{0:t}, r^i_{0:t}\right)$.

Here we drop the arm index and consider a fixed arm. For the unknown transition function, we assume its prior $g_0\left(P\right) = f(\frac{P-\epsilon_1\mathbf{1}}{1-S\epsilon_1} \mid \phi)$. We consider this special prior is due to the minimum elements of the transition matrix is bigger than ϵ_1. Next we show the details that how to update the posterior distribution for unknown P and omit the details of unknown reward function R.

$$\begin{aligned}
g\left(P \mid a_{0:t-1}, r_{0:t-1}\right) =& \frac{P\left(r_{0:t-1}, s_t \mid P, a_{0:t-1}\right) g\left(P, a_{0:t-1}\right)}{\int P\left(r_{0:t-1}, s_t \mid P, a_{0:t-1}\right) g\left(P, a_{0:t-1}\right) dP} \\
=& \frac{\sum_{s_{0:t-1}\in S^t} P\left(r_{0:t-1}, s_{0:t} \mid P, a_{0:t-1}\right) g(P)}{\int P\left(r_{0:t-1}, s_t \mid P, a_{0:t-1}\right) g\left(P, a_{0:t-1}\right) dP} \\
=& \frac{\sum_{s_{0:t-1}\in S^t} g\left(P\right) \prod_{i=1}^t P(s_i \mid s_{i-1})}{\int P\left(r_{0:t-1}, s_t \mid P, a_{0:t-1}\right) g\left(P, a_{0:t-1}\right) dP} \\
=& \frac{\sum_{s_{0:t-1}\in S^t} g\left(P\right) \left[\prod_{s,s'} (\frac{P(s'|s)-\epsilon_1}{1-\epsilon_1})^{N_{ss'}(s_{0:t})}\right]}{\int P\left(r_{0:t-1}, s_t \mid P, a_{0:t-1}\right) g\left(P, a_{0:t-1}\right) dP}.
\end{aligned},$$

where the last equality is due to the prior for unknown P is $g_0\left(P\right) = f(\frac{P-\epsilon_1\mathbf{1}}{1-S\epsilon_1} \mid \phi)$.

Next we show the Bayesian approach to learning unknown P and R with the history $(a_{0:t-1}, r_{0:t})$. Since the current state s_t of the agent at time t is unobserved, we consider a joint posterior $g\left(s_t, P, R \mid a_{0:t-1}, r_{0:t}\right)$ over s_t, P, and R (Ross et al.,

2011). The most parts are similar to Ross et al. (2011), except for our special priors.

$$
\begin{aligned}
g\left(s_t, P, R \mid a_{0:t-1}, r_{0:t-1}\right) \propto & P\left(r_{0:t}, s_t \mid P, R, a_{0:t-1}\right) g\left(P, R, a_{0:t-1}\right) \\
\propto & \sum_{s_{0:t-1} \in S^t} P\left(r_{0:t}, s_{0:t} \mid P, R, a_{0:t-1}\right) g(P, R) \\
\propto & \sum_{s_{0:t-1} \in S^t} g\left(s_0, P, R\right) \prod_{i=1}^{t} P(s_i \mid s_{i-1}) R(r_i \mid s_i) \\
\propto & \sum_{s_{0:t-1} \in S^t} g\left(s_0, P, R\right) \left[\prod_{s,s'} \left(\frac{P(s' \mid s) - \epsilon_1}{1-\epsilon_1}\right)^{N_{ss'}(s_{0:t})}\right] \times \\
& \left[\prod_{s,r} \left(\frac{R(r \mid s) - \epsilon_2}{1-\epsilon_2}\right)^{N_{sr}(s_{0:t}, r_{0:t-1})}\right]
\end{aligned}
$$

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; $N_{ss'}\left(s_{0:t}\right)$ is the number of times the transition (s, s') appears in the history of state-action $(s_{0:t})$; and $N_{sr}\left(s_{0:t}, r_{0:t-1}\right)$ is the number of times the observation (s, r) appears in the history of state-rewards $(s_{0:t}, r_{0:t-1})$.

D. Technical Results

Proposition D.1. *(Uniform bound on the bias span (Zhou et al., 2021)). If the belief MDP satisfies Assumption 3.1,3.2, then for $(J(\theta), v(:,\theta))$ satisfying the Bellman equation (2), we have the span of the bias function* $\operatorname{span}(v,\theta) := \max_{b,\theta} v(b,\theta) - \min_{b,\theta} v(b,\theta)$ *is bounded by H, where*

$$
H := \frac{8\left(\frac{2}{(1-\alpha)^2} + (1+\alpha)\log_\alpha \frac{1-\alpha}{8}\right)}{1-\alpha}, \quad \text{with } \alpha = \frac{1-\epsilon_1}{1-\epsilon_1/2} \in (0,1)
$$

Proposition D.2. *(Controlling the belief error (Xiong et al., 2022e)). Suppose Assumption 3.1,3.2 hold. 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*

$$
\left\| b_t(\cdot, R^*, P^*) - \hat{b}_t(\cdot, R_k, P_k) \right\|_1 \leq L_1 \|R_k - R^*\|_1 + L_2 \max_s \|P^*(m,:) - P_k(m,:)\|_2 ,
$$

where $L_1 = \frac{4(1-\epsilon_1)^2}{N\epsilon_1^2\epsilon_2}, L_2 = \frac{4(1-\epsilon_1)^2}{\epsilon_1^3}$, ϵ_1 and ϵ_2 are the minimum elements of the functions P^* and R^*, respectively.

Lemma D.3. *(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 D.4. *(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*

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