ONLINE RESTLESS BANDITS WITH UNOBSERVED STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the online restless bandit problem, where each arm evolves according to a Markov chain independently, and the reward of pulling an arm depends on both the current state of the corresponding Markov chain and the action. The agent (decision maker) does not know the transition kernels and reward functions, and cannot observe the states of arms all the time. The goal is to sequentially choose which arms to pull so as to maximize the expected cumulative rewards collected.

In this paper, we propose TSEETC, a learning algorithm based on Thompson Sampling with Episodic Explore-Then-Commit. The algorithm proceeds in episodes of increasing length and each episode is divided into exploration and exploitation phases. In the exploration phase in each episode, action-reward samples are collected in a round-robin way and then used to update the posterior as 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 $\tilde{O}(\sqrt{T})$ for TSEETC, where $T$ is the time horizon. This is the first bound that is close to the lower bound of restless bandits, especially in an unobserved state setting. We show through simulations that TSEETC outperforms existing algorithms in regret.

1 INTRODUCTION

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

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

One common way to handle the unknown parameters but with observed states is to use the optimism in the face of uncertainty (OFU) principle (Liu et al., 2010; 2011; Ortner et al., 2012; Wang et al., 2020; Xiong et al., 2022a). These methods sometimes do not perform close to the optimal, because
of the baseline they consider, such as pulling the fixed arms (Liu et al., 2010; 2011), which is not optimal in RMAB problem. Ortner et al. (2012) derives the lower bound \( \tilde{O}(\sqrt{T}) \) for RMAB problem. However, it is not clear whether there is an efficient computational method to find the optimistic model in the confidence region (Lakshmanan et al., 2015). Another way to estimate the unknown parameters is Thompson Sampling (TS) method (Jung & Tewari, 2019; Jung et al., 2019; Jahromi et al., 2022; Hong et al., 2022). TS algorithm is more computationally efficient since it only needs to solve the sampled instance, without the need to solve all instances that lie within the confident sets as OFU-based algorithms (Ouyang et al., 2017). What’s more, empirical studies suggest that TS algorithms outperform OFU-based algorithms in bandit and Markov decision process (MDP) problems (Scott, 2010; Chapelle & Li, 2011; Osband & Van Roy, 2017).

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

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

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

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

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

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

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

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

**Unknown states.** There are some works that consider the states of the pulled arm are observed (Mate et al., 2020; Liu & Zhao, 2010; Wang et al., 2020; Jung & Tewari, 2019). Mate et al. (2020) and Liu & Zhao (2010) assume the unobserved states but with known parameters. Wang et al. (2020) constructs an offline instance and give the regret bound $\tilde{O}(T^{2/3})$. Jung & Tewari (2019) considers the episodic RMAB problems and the regret bound $\tilde{O}(\sqrt{T})$ 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 $O(T^{2/3})$ and there is a gap between the lower bound $O(\sqrt{T})$ (Ortner et al., 2012). Jahromi et al. (2022) considers the POMDP setting and propose the pseudo counts to store the state-action pairs. Their learning algorithm is based on Ouyang et al. (2017) and the regret bound is also $O(T^{2/3})$. And their algorithm is not programmable due to the pseudo counts is conditioned on the true counts which is uncountable.

3 PROBLEM SETTING

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

The whole game is divided into $T$ time steps. The initial state $s^i_0$ 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 $r_t \in R$ with probability $R^{a_t}(r_t \mid s_{a_t}^t)$.
that only the pulled arm has the reward feedback. His decision on which arm \( a_t \) to pull is based on the observed history \( \mathcal{H}_t = [a_1, r_1, a_2, r_2, \ldots, a_{t-1}, r_{t-1}] \). Note that the states of the arms are never observable, even after pulling. Each arm \( i \) makes a state transition independently according to the associated \( P_i \), whether it is pulled or not. This process continues until the end of the game. The goal of the agent is to maximize the total expected reward.

We use \( \theta \) to denote the unknown \( P_i \) and \( R^i \) for \( i \in [N] \) collectively. Since the true states are unobservable, the agent maintains a belief state \( \Delta^i_t = \{b^i_s(\theta), s \in S^i\} \in \Delta^S_i \), for each arm \( i \), where

\[
b^i_s(\theta) := \mathbb{P}(s_t = s | H_t, \theta),
\]

and \( \Delta^S_i := \{b \in \mathbb{R}^{S^i} : \sum_{s \in S^i} b(s) = 1\} \) is the probability simplex in \( \mathbb{R}^{S^i} \). Note that \( b^i_s(\theta) \) depends on the unknown model parameter \( \theta \), which itself has to be learned by the agent. For a given \( \theta \), the overall belief state \( b_t = (b^1_t, b^2_t, \ldots, b^N_t) \) is a sufficient statistic for \( H_t \) (Smallwood & Sondik, 1973), so the agent can base his decision at time \( t \) on \( b_t \) only. Let \( \Delta_b := \Delta^S_1 \times \cdots \times \Delta^S_N \).

A deterministic stationary policy \( \pi : \Delta_b \to [N] \) maps a belief state to an action. The long-term average reward of a policy \( \pi \) is defined as

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

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

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

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

\[
J(\theta) + v(b, \theta) = \max_a \left\{ r(b, a) + \sum_r P(r | b, a, \theta) v(b', \theta) \right\},
\]

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

We consider the Bayesian setting for the parameters. The parameters \( \theta^\ast \) 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 \( \pi \) by its regret, defined as the expected gap between the cumulative reward of an offline oracle and that of \( \pi \), where the oracle is the optimal policy with the full knowledge of \( \theta^\ast \), but unknown states. The offline oracle is similar to Zhou et al. (2021), which is stronger than those considered in Azizzadenesheli et al. (2016) and Fiez et al. (2018). We focus on the Bayesian regret of policy \( \pi \) (Ouyang et al., 2017; Jung & Tewari, 2019) as follows,

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

The above expectation is with respect to the prior distribution about \( \theta^\ast \), the randomness in state transitions and the random reward.

## 4 The TSEETC Algorithm

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

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

$$b^i_{t+1}(s', \theta) = \frac{\sum_s b^i_t(s, \theta) R^i_t(s' \mid s) P^i(s' \mid s)}{\sum_s b^i_t(s, \theta) R^i_t(s' \mid s)}, \quad (4)$$

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

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

$$b^i_{t+1}(s', \theta) = \sum_s b^i_t(s, \theta) P^i(s' \mid s). \quad (5)$$

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

4.2 Mixture of Dirichlet Distribution

In this section, we derive the estimation method for unknown $P^i$ and $R^i$ based on the mixture of Dirichlet distribution. The Dirichlet distribution is parameterized by a count vector, $\phi = (\phi_1, \ldots, \phi_k)$, where $\phi_i \geq 0$, such that the density of probability distribution $p = (p_1, \ldots, p_k)$ is defined as $f(p \mid \phi) \propto \prod_{i=1}^k p_i^{\phi_i-1}$ (Ghavamzadeh et al., 2015).

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

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

$$g_t(P^i) \propto \sum_{\bar{s}_t \in S_t^i} g_0(P^i) w(\bar{s}_t) \prod_{s,s'} (P^i(s' \mid s))^N^i_{s,s'}(\bar{s}_t)^{+ \phi^i_{s,s'} - 1}, \quad (6)$$

$$g_t(R^i) \propto \sum_{\bar{s}_t \in S_t^i} g_0(R^i) w(\bar{s}_t) \prod_{s,r} (R^i(r \mid s))^N^i_{s,r}(\bar{s}_t)^{+ \psi^i_{s,r} - 1}. \quad (7)$$

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

This procedure is summarized in Algorithm 1.

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

1: Input: the history length $\tau_t$, the state space $S_t$, the belief history $\bar{b}^i_{\tau_t}$, $\bar{r}^i_{\tau_t}$, the initial parameters $\phi^i_{s,s'}, \psi^i_{s,r}$, for $s, s' \in S_t$, $r \in R$.
2: generate $S^i_{\tau_t}$ possible state sequences
3: calculate the weight $w(j) = \prod_{t=1}^{\tau_t} b^i_t(s, \theta), j \in S^i_{\tau_t}$
4: for $j$ in $1, \ldots, S^i_{\tau_t}$ do
5: count the occurrence 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 $\phi(\bar{j})$, $\psi^i_{s,r}$ as $\psi(\bar{j})$ for all $s, s' \in S_t, r \in R$
8: end for
9: update the mixture Dirichlet distribution

$$g_{\tau_t}(P^i) \propto \sum_{j=1}^{S^i_{\tau_t}} w(j)f(P^i \mid \phi(\bar{j})), g_{\tau_t}(R^i) \propto \sum_{j=1}^{S^i_{\tau_t}} w(j)f(R^i \mid \psi(\bar{j}))$$
4.3 Our Algorithm

Our algorithm, TSEETC, operates in episodes with different lengths. Each episode is split into exploration phase and exploitation phase. Denote the episode number is \( K_t \) and the first time in each episode is denoted as \( t_k \). We use \( T_k \) to denote the length of episode \( k \) and it can be determined as:
\[
T_k = T_1 + k - 1, \quad \text{where} \quad T_1 = \left\lfloor \frac{\sqrt{T+1}}{2} \right\rfloor.
\]

The length of exploration phase in each episode is fixed as \( \tau_1 \) which satisfies \( \tau_1 K_T = \mathcal{O}(\sqrt{T}) \) and \( \tau_1 \leq \frac{2}{1 + K_T - 1} \). With these notations, our whole algorithm is shown below.

Algorithm 2 Thompson Sampling with Episodic Explore-Then-Commit

1: Input: prior \( g_0(P), g_0(R) \), initial belief \( b_0 \), the first episode length \( T_1 \)
2: for episode \( k = 1, 2, \ldots, T \)
3: start the first time of episode \( k \), \( t_k := t \)
4: generate \( R(t_k) \sim g_{t_k}(R) \) and \( P(t_k) \sim g_{t_k}(P) \)
5: for \( t = t_k, t_k + 1, \ldots, t_k + \tau_1 \) do
6: pull the arm \( i \) for \( \tau_1/N \) times in a round robin way
7: receive the reward \( r_t \)
8: update the belief \( b^\tau_{t_k}(i) \) using \( R(t_k), P(t_k) \) based on equation 4
9: end for
10: update the belief \( b^\tau_{t_k}, j \in N \setminus \{i\} \) using \( P(t_k) \) based on equation 5
11: for \( i = 1, 2, \ldots, N \) do
12: input the obtained \( \bar{r}_{t_k}, \tilde{b}_{t_k} \) to Algorithm 1 to update the posterior distribution \( g_{t_k + \tau_1}(P) \), \( g_{t_k + \tau_1}(R) \)
13: end for
14: generate \( R(t_k + \tau_1) \sim g_{t_k + \tau_1}(P), P(t_k + \tau_1) \sim g_{t_k + \tau_1}(R) \)
15: for \( i \) in \( 0, 1, \ldots, N \) do
16: re-update the belief \( b^\tau_{t_k} \) from time 0 to \( t_k + \tau_1 \) based on \( R(t_k + \tau_1) \) and \( P(t_k + \tau_1) \)
17: end for
18: compute \( \pi^\tau_k(\cdot) = \text{Oracle}(\cdot, R(t_k + \tau_1), P(t_k + \tau_1)) \)
19: for \( t = t_k + \tau_1 + 1, \ldots, t_k + 1 \) do
20: apply action \( a_t = \pi^\tau_k(b_t) \)
21: observe new reward \( r_{t+1} \)
22: update the belief \( b_t \) of all arms based equation 4, equation 5
23: end for
24: end for

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

Next we enter into the exploitation phase. Firstly we derive the optimal policy \( \pi_k \) for the sampled parameter \( \theta_{t_k + \tau_1} \). Then we use policy \( \pi_k \) for the rest of the episode \( k \).

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

Remark 1. We use an Oracle to derive the optimal policy for the sampled parameters in Algorithm 2. The Oracle can be the Bellman equation for POMDP as we introduced in equation 2, or the
approximation methods (Pineau et al., 2003; Silver & Veness, 2010), etc. The approximation error is discussed in section 5.

5 PERFORMANCE ANALYSIS

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

5.1 REGRET BOUND AND DISCUSSIONS

Theorem 1. Suppose Assumption 1 holds and the Oracle returns the optimal policy. The Bayesian regret of TSEETC satisfies

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

where \( C_1 = L_1 + L_2MN + N + S^2, C_2 = r_{\max} + H \) are constants independent with time horizon \( T, L_1 = 4M (\frac{\epsilon}{c})^2 / \min \{R_{\min}^*, 1 - R_{\max}^*\}, L_2 = 4M(1 - \epsilon)^2/c + \sqrt{M}, M = S^N \), \( R_{\max}^* \) and \( R_{\min}^* \) are the maximum and minimum element of the functions \( R^* \) respectively, \( \tau_1 \) is the fixed exploration length in each episode, \( \Delta R \) is the biggest gap of the reward obtained at each two different time, \( H \) is the bounded span, \( r_{\max} \) is the maximum reward obtain each time.

Remark 2. The Theorem 1 shows that the regret of TSEETC is upper bound by \( \tilde{O}(\sqrt{T}) \). This is the first bound that matches the lower bound in restless bandit problem (Ortner et al., 2012) in such unobserved state setting. Although TSEETC looks similar to explore-then-commit (Lattimore & Szepesvári, 2020), a key novelty of TSEETC lies in using the approach of posterior sampling (Jung & Van Roy, 2017). In Jahromi et al. (2022), their pseudo count of state-action pair is always smaller than the true counts with some probability at any time. However, in our algorithm, the sampled parameter is more concentrated on true values with the posterior update. Therefore, our pseudo helps us obtain a tighter bound.

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

5.2 PROOF SKETCH

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

\[ R_T = \mathbb{E}_{\theta_*} \left[ \sum_{k=1}^{k_T} \sum_{t_k}^{t_k+\tau_1} J(\theta^*) - \tau_1 \right] + \mathbb{E}_{\theta_*} \left[ \sum_{k=1}^{k_T} \sum_{t_k+\tau_1}^{t_k+\tau_1+1} J(\theta^*) - \tau_1 \right] \quad \text{(8)} \]

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

\[ \text{Regret (A)} \leq \mathbb{E}_{\theta_*} \left[ \sum_{k=1}^{k_T} \tau_1 \Delta R \right] \leq \tau_1 \Delta R k_T \quad \text{(9)} \]
where \( \Delta R = r_{\text{max}} - r_{\text{min}} \) is the biggest gap of the reward received at each two different times. The regret in equation 9 is related with the episode number \( k_T \), which can be bounded in Lemma 2.

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

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

\[
J(\theta_k) + v(\hat{b}_t, \theta_k) = r(\hat{b}_t, a_t) + \sum_r P(r | \hat{b}_t, a_t, \theta_k)v(b'_t, \theta_k).
\]

(10)

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

\[
\text{Regret(B)} = R_1 + R_2 + R_3 + R_4
\]

(11)

where each term is defined as follows:

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

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

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

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

**Bounding R1.** One key property of Posterior Sampling algorithms is that for given the history \( \mathcal{H}_{tk} \), the true parameter \( \theta^* \) and sampled \( \theta_k \) are identically distributed at the time \( t_k \). Due to the length \( T_k \) determined and independent with \( \theta_k \), then \( R_1 \) is zero thanks to Lemma 3.

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

\[
\mathbb{E} [g(\theta^*) | \mathcal{H}_{tk}] = \mathbb{E} [g(\theta_k) | \mathcal{H}_{tk}].
\]

**Bounding R2.** The regret \( R_2 \) is the telescopic sum of value function and can be bounded as \( R_2 \leq HK_T \). It solely depends on the episode number and the upper bound \( H \) of span function. As a result, \( R_2 \) reduce to a finite bound over the number of episodes \( k_T \), which can be bounded in Lemma 2.

**Bounding R3 and R4.** The regret terms \( R_3 \) and \( R_4 \) is related with estimation error about \( \theta \). Thus we should bound the parameters’ error especially in our unobserved state setting. Note that when the states are observed, it is easy to count the state-action pairs and then the confidence interval can be also constructed such as in Wang et al. (2020); Xiong et al. (2022a). Recall the definition of \( \phi, \psi \), we can define the empirical estimation of \( \hat{P}(s' | s) \) and \( \hat{R}(r | s) \) for arm \( i \) at time \( t \) as follows:

\[
\hat{P}_t(s' | s)(t) = \frac{\phi_{s,s'}^i(t)}{\|\phi_{s,s'}^i(t)\|_1}, \quad \hat{R}_t(r | s)(t) = \frac{\psi_{s,r}^i(t)}{\|\psi_{s,r}^i(t)\|_1}
\]

(12)

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

\[
N_{tk}^i(s, a) = \|\psi_{s,a}^i(t_k)\|_1 - \|\psi_{s,a}^i(0)\|_1
\]

(13)

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

\[
\mathcal{M}_k := \left\{ P : \sum_{s' \in S} |P(s' | z) - \hat{P}_k^i(s' | z)| \leq \beta_k(z), R : \sum_{r \in R} |R(r | z) - \hat{R}_k^i(r | z)| \leq \beta_k(z) \right\}
\]
where $\beta_k(s, a) := \sqrt{\frac{14S \log(2SAk)}{\max\{1, N'_{ak}(s, a)\}}}$ is chosen conservatively (Auer et al., 2008) so that $M_k$ contains both $M^*$ and $M_k$ with high probability. Specially, for the unobserved state setting, the belief error under different parameters is upper bounded by the gap between the estimators as in Proposition 1. Then the core of the proofs lies in deriving a high-probability confidence set with our pseudo counts and show that the estimated error accumulated to $T$ for each arm is bounded by $\sqrt{T}$. Then with the error bound for each arm, we can achieve the final regret bound for the MDP aggregated by all arms. With $C_1 = L_1 + L_2 M N + N + S^2$, We show the final results here and the detailed proof in Appendix B.3,B.4.

**Lemma 4.** $R_3$ satisfies the following bound

$$R_3 \leq 48C_1 SH \sqrt{AT \log AT} + 4C_1 SAH \sqrt{T} + C_1 H.$$  

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

$$R_4 \leq 48C_1 S r_{max} \sqrt{AT \log(AT)} + 4C_1 SA r_{max} \sqrt{T} + C_1 r_{max}.$$  

6 Numerical Experiments

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

![Figure 1: The cumulative regret](image1.png)

![Figure 2: The log-log regret](image2.png)

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

7 Conclusion

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


Under review as a conference paper at ICLR 2023


A  TABLE OF NOTATIONS

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>The length of horizon</td>
</tr>
<tr>
<td>$K_T$</td>
<td>The episode number of time $T$</td>
</tr>
<tr>
<td>$t_k$</td>
<td>The episode length of episode $k$</td>
</tr>
<tr>
<td>$\tau_1$</td>
<td>The fixed exploration length in each episode</td>
</tr>
<tr>
<td>$P^i$</td>
<td>The transition functions for arm $i$</td>
</tr>
<tr>
<td>$R^i$</td>
<td>The reward function for arm $i$</td>
</tr>
<tr>
<td>$P_k$</td>
<td>The sampled transition function for aggregated MDP</td>
</tr>
<tr>
<td>$R_k$</td>
<td>The sampled reward function for aggregated MDP</td>
</tr>
<tr>
<td>$r_t$</td>
<td>The reward obtained at time $t$</td>
</tr>
<tr>
<td>$b_i^t(s, \theta)$</td>
<td>The belief state for being in state $s$ at time $t$ for arm $i$ with parameter $\theta$</td>
</tr>
<tr>
<td>$b_t$</td>
<td>The belief of all arms at time $t$ with parameter $\theta_k$</td>
</tr>
<tr>
<td>$b_t^*$</td>
<td>The belief of all arms at time $t$ with parameter $\theta^*$</td>
</tr>
<tr>
<td>$a_t$</td>
<td>The action at time $t$</td>
</tr>
<tr>
<td>$J(\theta_k)$</td>
<td>The optimal long term average reward with parameter $\theta_k$</td>
</tr>
<tr>
<td>$r_{max}$</td>
<td>The maximum reward obtained each time</td>
</tr>
<tr>
<td>$r_{min}$</td>
<td>The minimum reward obtained each time</td>
</tr>
<tr>
<td>$\Delta R$</td>
<td>The biggest gap of the obtained reward</td>
</tr>
</tbody>
</table>

B  PROOF OF THEOREM 1

Recall that our goal is to minimize the regret:

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

$r_t$ depends on the state $s_t$ and $a_t$. Thus $r_t$ can be written as $r(s_t, a_t)$. Due to $\mathbb{E}_{\theta^*} [r(s_t, a_t) \mid H_{t-1}] = r(b_t^*, a_t)$ for any $t$, we have,

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

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

$$ R_T = \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t_k}^{t_k+\tau_1} (J(\theta^*) - r(b_t^*, a_t)) + \sum_{k=1}^{K_T} \sum_{t_k+\tau_1+1}^{t_k+\tau_2} (J(\theta^*) - r(b_t^*, a_t)) \right], \quad (16) $$

where $\tau_1$ is the exploration length for each episode. $\tau_1$ 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.

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

$$ \text{Regret (B)} = \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t_k+\tau_1+1}^{t_k+\tau_2} (J(\theta^*) - r(b_t^*, a_t)) \right]. $$

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:

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

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

$$ \text{Regret (A)} \leq \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \tau_1 \Delta R \right] \leq \tau_1 \Delta R k_T. $$
Regret (A) is related with the episode number $k_T$ obviously, which is bounded in Lemma 11. Next we should bound the term Regret (B).

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

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

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

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

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

where

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

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

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

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

Next we bound the four parts one by one.

### B.1 Bound $R_1$

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

**Proof.** Recall that:

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

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

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

therefore, the part $R_1$ is 0. \hfill $\Box$

### B.2 Bound $R_2$

**Lemma 7.** $R_2$ satisfies the following bound

$$R_2 \leq HK_T,$$

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 $t$.

$$R_2 = \mathbb{E}_{\theta^*} \sum_{k=1}^{k_T} \left[ \sum_{t_k + \tau_1 + 1}^{t_{k+1}-1} \left( v(\hat{b}_{t+1}, \theta_k) - v(\hat{b}_t, \theta_k) \right) \right]. \quad (19)$$
We consider the whole sum in episode \( k \), then the \( R_2 \) can be rewrite as:

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

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

\[
R_2 \leq H K_T.
\]

\[\Box\]

### B.3 Bound \( R_3 \)

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

#### B.3.1 Rewrite \( R_3 \)

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

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

where \( P_k \) is the sampled transition functions in episode \( k \), \( R_k \) is the sampled reward functions in episode \( k \), \( \hat{b}_t^* \) 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 \( R_3 = \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k+\tau_t+1}^{t_k+1} \left( \sum_r P(r \mid \hat{b}_t, a_t, \theta_k) v(b', \theta_k) - v(\hat{b}_{t+1}, \theta_k) \right) \right] \).

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

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

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

\[
P(r \mid b_t^*, a_t, \theta^*) = \sum_{s' \in s_t + 1} R^* (r \mid s') P(s_{t+1} = s' \mid \mathcal{H}_t, \theta^*)
\]

\[
= \sum_{s' \in s_t + 1} R^* (r \mid s') \sum_{s} P^* (s_{t+1} = s' \mid s_t = s, \mathcal{H}_t, a_t, \theta^*) P(s_t = s \mid \mathcal{H}_t, \theta^*)
\]

\[
= \sum_{s} \sum_{s'} b_t^*(s) P^* (s' \mid s) R^* (r \mid s'),
\]

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,

\[
R_3 = \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k+\tau_t+1}^{t_k+1} \left( \sum_{r \in R} (P(r \mid \hat{b}_t, a_t, \theta_k) - P(r \mid b_t^*, a_t, \theta^*) v(b', \theta_k) \right) \right].
\]
Based on equation 21, we get

$$R_3 = E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \sum_s \hat{b}_k(s) P_k(s' | s) \right) \right]$$

$$- E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R^*(r | s') \sum_s b_k^*(s) P^*(s' | s) \right) \right]$$

$$= E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \sum_s \hat{b}_k(s) P_k(s' | s) \right) \right]$$

$$- E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \sum_s b_k^*(s) P^*(s' | s) \right) \right]$$

$$+ E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \sum_s b_k^*(s) P^*(s' | s) \right) \right]$$

$$- E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R^*(r | s') \sum_s b_k^*(s) P^*(s' | s) \right) \right].$$

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

Define

$$R'_3 = E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \left[ \sum_s \hat{b}_k(s) P_k(s' | s) - \sum_s b_k^*(s) P^*(s' | s) \right] \right) \right],$$

$$R''_3 = E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) [R_k(r | s') - R^*(r | s')] \sum_s b_k^*(s) P^*(s' | s) \right) \right].$$

Bounding $R'_3$. The part $R'_3$ can be bounded as Jahromi et al. (2022).

$$R'_3 = E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \left[ \sum_s \hat{b}_k(s) P_k(s' | s) - \sum_s b_k^*(s) P^*(s' | s) \right] \right) \right]$$

$$= R'_{3}(0) + R'_{3}(1),$$

where

$$R'_{3}(0) = E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \sum_s \hat{b}_k(s) P_k(s' | s) \right) \right]$$

$$- E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \sum_s b_k^*(s) P_k(s' | s) \right) \right]$$

$$R'_{3}(1) = E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \sum_s b_k^*(s) P_k(s' | s) \right) \right]$$

$$- E_{\theta^*} \left[ \sum_{k=1}^{t_T} \sum_{t=t_k+1}^{t_{k+1}-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r | s') \sum_s b_k^*(s) P^*(s' | s) \right) \right].$$
For $R'_3(0)$, because $\sum_r R_k(r \mid s') = 1$, $\sum_{s'} P_k(s' \mid s) = 1$, $v(b', \theta_k) \leq H$, we have

$$R'_3(0) = \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+1}^{t_k+1-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r \mid s') \sum_s \hat{b}_t(s) P_k(s' \mid s) \right) \right]$$

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

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

$$\leq \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+1}^{t_k+1-1} \left( \sum_s |\hat{b}_t(s) - \hat{b}^*_t(s)| \right) \right]$$

$$= \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+1}^{t_k+1-1} \left( \|\hat{b}_t(s) - \hat{b}^*_t(s)\|_1 \right) \right],$$

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

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

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

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

Substitute equation 23, equation 24 into $R'_3(1)$, we have

$$R'_3(1) = \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+1}^{t_k+1-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r \mid s') \left( P_k(s' \mid s) - P^*(s' \mid s) \right) \right) \right]$$

$$\leq \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+1}^{t_k+1-1} \left( \sum_r \sum_{s'} v(b', \theta_k) R_k(r \mid s') \left| P_k(s' \mid s) - P^*(s' \mid s) \right| \right) \right]$$

$$\leq \mathbb{H} \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+1}^{t_k+1-1} \left( \sum_{s'} \left| P_k(s' \mid s) - P^*(s' \mid s) \right| \right) \right]$$

$$\leq \mathbb{H} \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+1}^{t_k+1-1} \left( \| P_k - P^* \|_1 \right) \right],$$

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 $v(b', \theta_k) \leq H$ and $\sum_r R_k(r \mid s') = 1$.

Therefore we obtain the final results,

$$R'_3 \leq H \mathbb{E} \left[ \sum_{k=1}^{K_T} \sum_{t=k+1}^{t_k+1-1} \| P^* - P_k \|_1 \right] + H \mathbb{E} \left[ \sum_{k=1}^{K_T} \sum_{t=k+1}^{t_k+1-1} \| b^*_t - \hat{b}_t \|_1 \right].$$
Bounding $R''_3$. For part $R''_3$, note that for any fixed $s'$, $\sum_s b_t^i(s) P^*(s' | s) \leq S$, therefore we can bound $R''_3$ as follows,

\[
R''_3 = \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k + \tau_t + 1}^{t_k + \tau_{t+1}} \left( \sum_r \sum_{s'} v(b', \theta_k) \left[ R_k(r | s') - R^*(r | s') \right] \sum_s b_t^i(s) P^*(s' | s) \right) \right]
\]
\[
\leq S H \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k + \tau_t + 1}^{t_k + \tau_{t+1}} \left( \sum_{s'} \sum_r \left[ R_k(r | s') - R^*(r | s') \right] \right) \right]
\]
\[
\leq S H \mathbb{E}_{\theta^*} \sum_{k=1}^{K_T} \sum_{t=t_k + \tau_t + 1}^{t_k + \tau_{t+1}} S \| R_k - R^* \|_1
\]
\[
\leq S^2 H \mathbb{E}_{\theta^*} \sum_{k=1}^{K_T} \sum_{t=t_k + \tau_t + 1}^{t_k + \tau_{t+1}} \| R_k - R^* \|_1,
\]

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

B.3.2 Bound $R_3$

Lemma 9. $R_3$ satisfies the following bound

\[
R_3 \leq 48 (L_1 + L_2 S N^N + N + S^2) S H \sqrt{AT \log AT} + (L_1 + L_2 S N^N + N + S^2) H \\
+ 4 (L_1 + L_2 S N^N + N + S^2) S AH (T_1 + K_T - \tau_t - 1).
\]

Proof. Recall that the $R_3$ is as follows:

\[
R_3 = \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k + \tau_t + 1}^{t_k + \tau_{t+1}} \left( \sum_r P[r | \hat{b}_t, \alpha_t, \theta_k] v(b', \theta_k) - v(\hat{b}_{t+1}, \theta_k) \right) \right].
\]

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

\[
\bar{\psi}^i(s' | s)(t) = \frac{\phi_{s', s}(t)}{\| \phi^i_{s}(t) \|_1}, \quad \bar{R}^i(r | s)(t) = \frac{\psi_{s, r}(t)}{\| \psi^i_{s}(t) \|_1},
\]

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

\[
N_{ik}^i(s, a) = \| \psi^i_{s}(t_k) \|_1 - \| \psi^i_{s}(0) \|_1.
\]

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

\[
R_3 = \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k + \tau_t + 1}^{t_k + \tau_{t+1}} \left( \sum_r P[r | \hat{b}_t, \alpha_t, \theta_k] v(b', \theta_k) - v(\hat{b}_{t+1}, \theta_k) \right) \right]
\]
\[
= \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k + \tau_t + 1}^{t_k + \tau_{t+1}} \left( \sum_r \left( P(r | \hat{b}_t, \alpha_t, \theta_k) - P(r | b_t^i, \alpha_t, \theta_k) \right) v(b', \theta_k) \right) \right]
\]
\[
\leq R^0_3 + R^1_3 + R^2_3
\]
where
\[ R_3^0 = H \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+\tau_t+1}^{t_{k+1}-1} ||P^* - P_k||_1 \right], \]
\[ R_3^1 = H \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+\tau_t+1}^{t_{k+1}-1} ||b_t^* - \hat{b}_t||_1 \right], \]
\[ R_3^2 = S^2 \mathbb{H}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=k+\tau_t+1}^{t_{k+1}-1} ||R^* - R_k||_1 \right]. \]

Note that the following results are all focused on one arm. Define \( P^i \) is the true transition function for arm \( i \), \( \hat{P}^i_k \) is the sampled transition functions for arm \( i \). We can extend the results on a arm to the aggregated large MDP based on the Lemma 12.

**Bounding** \( R_3^0 \). Since \( 0 \leq v(b', \theta_k) \leq H \) from our assumption, each term in the inner summation is bounded by
\[
\sum_{s' \in S} \left| (P^i(s' \mid z_i) - \hat{P}^i_k(s' \mid z_i)) v(s', \theta_k) \right|
\leq H \sum_{s' \in S} \left| P^i(s' \mid z_i) - \hat{P}^i_k(s' \mid z_i) \right|
\leq H \sum_{s' \in S} \left| P^i(s' \mid z_i) - \hat{P}^i_k(s' \mid z_i) \right| + H \sum_{s' \in S} \left| \hat{P}^i_k(s' \mid z_i) - \hat{P}^i_k(s' \mid z_i) \right|
\]
where \( P^i(s' \mid z_i) \) is the true transition function, \( \hat{P}^i_k(s' \mid z_i) \) is the sampled reward function and \( \hat{P}^i_k(s' \mid z_i) \) is the empirical estimation. The second inequality above in due to triangle inequality.

Note that for any \( M_k \) in confidence set \( \mathcal{M}_k \), for all state-action pairs, the sampled \( P \) and \( R \) satisfy,
\[
\mathcal{M}_k = \left\{ P : \sum_{s' \in S} \left| P(s' \mid z) - \hat{P}^i_k(s' \mid z) \right| \leq \beta_k(z), R : \sum_{r \in R} \left| R(r \mid z) - \hat{R}_k(r \mid z) \right| \leq \beta_k(z) \right\}
\]
where \( \beta_k(z) = \sqrt{\frac{14S \log(2At_kT)}{\max(1,N^i_k(z))}} \) (Auer et al., 2008) and \( N^i_k(s,a) = ||\psi^i_k,(t_k)||_1 - ||\psi^i_k,(0)||_1 \).

Note that \( \beta_k(z) \) is the confidence set with \( \delta = 1/t_k \).
\[
\sum_{s' \in S} \left| P^i(s' \mid z_i) - \hat{P}^i_k(s' \mid z_i) \right| + \sum_{s' \in S} \left| \hat{P}^i_k(s' \mid z_i) - \hat{P}^i_k(s' \mid z_i) \right| 
\leq 2\beta_k(z) + 2 \left( I_{\{P^i \notin B_k\}} + I_{\{\hat{P}^i_k \notin B_k\}} \right).
\]

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 all the episode length is not bigger than the last episode, that is \( t_{k+1} - 1 - (t_k + \tau_t) \leq T_{K_T} - \tau_1 \), then we can obtain,
\[
\sum_{k=1}^{K_T} \sum_{t=t_k+\tau_t}^{t_{k+1}-1} \beta_k(z_t) \leq \sum_{k=1}^{K_T} \sum_{t=1}^{T_{K_T} - \tau_1} \beta_k(z_t).
\]

Note that \( \sum_{s' \in S} \left| P^i(s' \mid z_i) - \hat{P}^i_k(s' \mid z_i) \right| \leq 2 \) is always true. And with our assumption \( \tau_1 \leq \frac{T_1 + K_T - 1}{2} \), it is easy to show that when \( N^i_{t_k} \geq T_{K_T} - \tau_1, \beta_k(z_t) \leq 2 \) holds. Then we can obtain,
\[
\sum_{k=1}^{K_T} \sum_{t=1}^{T_{K_T} - \tau_1} \min\{2, \beta_k(z_t)\} \leq \sum_{k=1}^{K_T} \sum_{t=1}^{T_{K_T} - \tau_1} 2N^i_{t_k} < T_{K_T} - \tau_1
\]
\[
+ \sum_{k=1}^{K_T} \sum_{t=1}^{T_{K_T} - \tau_1} I(N^i_{t_k} \geq T_{K_T} - \tau_1) \frac{14S \log (2At_kT)}{\max(1,N^i_{t_k}(z_t))},
\]
Consider the first part in equation 29. Obviously, the maximum of $N_{tk}^i$ is $T_{kT} - \tau_1$. Because there are totally $SA$ state-action pairs, therefore, the first part in equation equation 29 can be bounded as, $\sum_{k=1}^{K_T} \sum_{t=1}^{T_{kT} - \tau_1} 2\mathbb{I}(N_{tk}^i < T_{kT} - \tau_1) \leq 2(T_{kT} - \tau_1)SA$. Due to $T_{kT} = T_1 + K_T - 1$ and Lemma 11, we get, 

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

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

$$N_{tk}^i (s, a) = \left| \{ \tau < t, \tau \in [t_k, t_k + \tau_1], k \leq k(t) : (s^i_\tau, a^i_\tau) = (s, a) \} \right|,$$

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

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

$$\tau_1 \leq \frac{T_1 + K_T - 1}{2},$$

$$2\tau_1 \leq T_1 + K_T - 1 = T_{kT}.$$

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

$$N_{tk}^i (s, a) \leq N_{tk}^i (s, a) + \tau_1 \leq 2N_{tk}^i (s, a).$$

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

$$\sum_{k=1}^{K_T} \sum_{t=1}^{T_{kT} - \tau_1} \beta_k (z_t) \leq \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_k+1-1} \sqrt{\frac{14S \log (2AtTk)}{\max (1, N_{tk}^i (z_t))}}$$

$$\leq \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_k+1-1} \sqrt{\frac{14S \log (2AT^2)}{\max (1, N_{tk}^i (z_t))}}$$

$$= \sum_{t=1}^{T} \frac{28S \log (2AT^2)}{\max (1, N_{tk}^i (z_t))}$$

$$\leq \sqrt{56S \log (2AT)} \sum_{t=1}^{T} \frac{1}{\max (1, N_{tk}^i (z_t))}.$$ (30)

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

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

$$\sum_{t=1}^{T} \sqrt{\frac{1}{\max (1, N_{tk}^i (z_t))}} = \sum_{t=1}^{T} \frac{\mathbb{I}_{\{z_{t}=z\}}}{\max (1, N_{tk}^i (z_t))}$$

$$= \sum_{z} \left( \frac{\mathbb{I}_{\{N_{T+1}^i(z) > 0\}}}{\max (1, N_{tk}^i (z))} + \sum_{j=1}^{N_{T+1}^i(z) - 1} \frac{1}{\sqrt{j}} \right)$$

$$\leq \sum_{z} \left( \frac{\mathbb{I}_{\{N_{T+1}^i(z) > 0\}}}{\sqrt{N_{T+1}^i(z)}} + \frac{2}{\sqrt{N_{T+1}^i(z)}} \right) \leq 3 \sum_{z} \sqrt{N_{T+1}^i(z)}.$$

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

$$3 \sum_{z} \sqrt{N_{T+1}^i(z)} \leq 3 \sqrt{SA \sum_{z} N_{T+1}^i(z)} = 3\sqrt{SAT}. \quad (31)$$
With equation 30 and equation 31 we get
\[ 2H \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_k+1-1} \beta_k (z_t) \leq 6\sqrt{56}HS \sqrt{AT \log(\delta)} \leq 48\sqrt{HS \sqrt{AT \log(\delta)}}. \]

Then we can bound the equation 28 as follows,
\[ \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_k+1-1} \beta_k (z_t) \leq 24S \sqrt{AT \log(\delta)} + 2SA(T_1 + K_T - \tau_1 - 1). \tag{32} \]

Choose the \( \delta = 1/T \) in Lemma 13, and based by Lemma 3, we obtain that
\[ \mathbb{P}(P^*_k \notin B_k) = \mathbb{P}(P^*_k \notin B_k) \leq \frac{1}{15T \delta_k}. \]

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

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

Therefore, we can obtain the bound for one arm as follows,
\[ \mathbb{E}_{\theta^*} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k+\tau_1+1}^{t_k+1-1} \left( \sum_{s' \in S} \left( P^*_k (s' \mid z_t) - P^*_k (s' \mid z_t) \right) v (s', \theta_k) \right) \right] \leq H + 4SAH(T_1 + K_T - \tau_1 - 1) + 48\sqrt{AT \log(\delta)}. \tag{35} \]

Next we consider the state transition of all arms. Recall that the states of all arms at time \( t \) is \( s_t \). Because every arm evolves independently, then the transition probability from state \( s_t \) to state \( s_{t+1} \) is as follows,
\[ P (s_{t+1} \mid s_t, \theta^*) = \prod_{i=1}^{N} P^*_i (s^i_{t+1} \mid s^i_t), \]

where \( P^*_i \) is the true transition functions of arm \( i \). Based by the Lemma 12 and our assumption that all arms have the same state space \( S \), we can obtain
\[ \sum_{s_{t+1}} |P (s_{t+1} \mid s_t, \theta^*) - P (s_{t+1} \mid s_t, \theta_k)| \leq \sum_{i} \|P^*_i (s^i_{t+1} \mid s^i_t) - P^*_k (s^i_{t+1} \mid s^i_t)\|_1 \leq N \|P^*_i (s^i_{t+1} \mid s^i_t) - P^*_k (s^i_{t+1} \mid s^i_t)\|_1. \tag{36} \]

Therefore, we can bound the \( R^0_3 \) as follows:
\[ R^0_3 \leq NH + 4SANH(T_1 + K_T - \tau_1 - 1) + 48SNH \sqrt{AT \log(\delta)}. \tag{37} \]

**Bounding \( R^1_3 \).** Based on the Proposition 1, we know that
\[ \|b^*_t - \hat{b}_t\|_1 \leq L_1 \|R^* - R_k\|_1 + L_2 \|P* - P_k\|_F. \]

Based on the facts about the functions norm, we know that,
\[ \|P^* - P_k\|_F \leq SN \|P^* - P_k\|_1. \]
Therefore, we can bound the belief error at any time as follows:

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

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

\[
R_3^1 \leq H \mathbb{E}_{\theta} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_k+1-1} \left( L_1 \| R^* - R_k \|_1 + L_2 S^N \| P^* - P_k \|_1 \right) \right] \leq (L_1 + L_2 S^N N) H \mathbb{E}_{\theta} \left[ \sum_{k=1}^{K_T} \sum_{t=t_k}^{t_k+1-1} (2\beta_k(z_t) + 2 \left( \mathbb{I}_{\{R^* \notin B_k\}} + \mathbb{I}_{\{P^* \notin B_k\}} \right)) \right] \leq 48(L_1 + L_2 S^N N) S H \sqrt{AT \log AT} + (L_1 + L_2 S^N N) H 4(L_1 + L_2 S^N N) S A H (T_1 + K_T - \tau_1 - 1). \tag{39}
\]

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

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

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

\[
R_3 \leq 48(L_1 + L_2 S^N N) S H \sqrt{AT \log AT} + 4(L_1 + L_2 S^N N) S A H (T_1 + K_T - \tau_1 - 1) + (L_1 + L_2 S^N N) H + (L_1 + L_2 S^N N + N + S^2) S H \sqrt{AT \log AT} + (L_1 + L_2 S^N N + N + S^2) H + 4(L_1 + L_2 S^N N + N + S^2) S A H (T_1 + K_T - \tau_1 - 1). \tag{41}
\]

**B.4 Bound $R_4$**

**Lemma 10.** $R_4$ satisfies the following bound

\[
R_4 \leq 48(L_1 + L_2 S^N N + N + S^2) S r_{\max} \sqrt{AT \log(AT)} + (L_1 + L_2 S^N N + N + S^2) r_{\max} + 4(L_1 + L_2 S^N N + N + S^2) S A r_{\max} (T_1 + K_T - \tau_1 - 1). \]

**Proof.** We can rewrite the $R_4$ as follows:

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

where $r_k(s, a_t) = \sum_r r_k^a_t(r | s)$ 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 $r^*(s, a_t) = \sum_r \mathbb{E} R_k^a_t(r | s)$ is the expect reward
conditioned on the state $s$ and $a_t$, with the true reward function $R_t^n$. The equation 42 is due to the add the term $\sum s r_k(s, a_t) b_t^*(s)$ and subtract it.

Denote

$$R^0_4 = \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \left( \sum_s r_k(s, a_t) \hat{b}_t(s) - \sum_s r_k(s, a_t) b_t^*(s) \right) \right],$$

$$R^1_4 = \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \left( \sum_s r_k(s, a_t) b_t^*(s) - \sum_s r^*(s, a_t) b_t^*(s) \right) \right].$$

For $R^0_4$,

$$R^0_4 = \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \left( \sum_s r_k(s, a_t) \hat{b}_t(s) - \sum_s r_k(s, a_t) b_t^*(s) \right) \right] = \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \left( \sum_s r_k(s, a_t) (\hat{b}_t(s) - b_t^*(s)) \right) \right] \leq r_{\text{max}} \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \left( \sum_s |\hat{b}_t(s) - b_t^*(s)| \right) \right] \tag{43}$$

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

For $R^1_4$,

$$R^1_4 = \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \left( \sum_s r_k(s, a_t) b_t^*(s) - \sum_s r^*(s, a_t) b_t^*(s) \right) \right] = \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \left( \sum_s [r_k(s, a_t) - r^*(s, a_t)] b_t^*(s) \right) \right] \leq \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \left( \sum_s |r_k(s, a_t) - r^*(s, a_t)| \right) \right] \tag{44}$$

$$\leq \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \sum_s \sum_r |R_k^{a_t}(r \mid s) - R^a_t(r \mid s)| \right] \leq S r_{\text{max}} \mathbb{E}_{\theta^*} \left[ \sum_{t=1}^T \left( \|R_k^{a_t} - R^a_t\|_1 \right) \right] \tag{45}$$

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

Based on the equation 39, we can bound the $R^0_4$,

$$R^0_4 \leq 48(L_1 + L_2 S^N N) S r_{\text{max}} \sqrt{AT \log AT} + (L_1 + L_2 S^N N) r_{\text{max}} + 4(L_1 + L_2 S^N N) S A r_{\text{max}} (T_1 + K_T - \tau_1 - 1).$$

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

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

Then based on equation 40, we get

$$R^1_4 \leq 48 S^2 r_{\text{max}} \sqrt{AT \log AT} + 2 S^2 A r_{\text{max}} (T_1 + K_T - \tau_1 - 1) + S r_{\text{max}}.$$
Then we can obtain the final bound:
\[
R_4 \leq 48(L_1 + L_2 S^N N + S) S r_{max} \sqrt{AT \log(AT)} + 4(L_1 + L_2 S^N N + S) S A r_{max} (T_1 + K_T - \tau_1 - 1)
\]
\[
+ (L_1 + L_2 S^N N + S) r_{max}
\]
\[
\leq 48(L_1 + L_2 S^N N + N + S^2) S r_{max} \sqrt{AT \log(AT)} + (L_1 + L_2 S^N N + N + S^2) r_{max}
\]
\[
+ 4(L_1 + L_2 S^N N + N + S^2) S A r_{max} (T_1 + K_T - \tau_1 - 1)
\]
where the last inequality is due to \( S \leq N + S^2 \).

\[ \blacksquare \]

**B.5 The Total Regret**

Next we bound the episode number.

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

**Proof.** Note that the total horizon is \( T \). The length of episode \( k \) is \( T_k = T_1 + k - 1 \). Then we can get,
\[
T = T_1 + T_2 + ... + T_{k_T}
\]
\[
= T_1 + (T_1 + 1) + ... + (T_1 + K_T - 1)
\]
\[
= K_T T_1 + (1 + 2 + ... + K_T - 1)
\]
\[
= K_T T_1 + \frac{K_T(K_T - 1)}{2}.
\]
Therefore,
\[
K_T^2 + (2T_1 - 1)K_T - 2T = 0.
\]
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 \( C_1 = L_1 + L_2 S^N N + N + S^2 \), \( C_2 = H + r_{max} \) and \( C_3 = T_1 + K_T - \tau_1 - 1 \), then we can get the final regret:
\[
R_T = \text{Regret}(A) + R_1 + R_2 + R_3 + R_4
\]
\[
\leq \tau_1 \Delta R K_T + H K_T + 48C_1 S H \sqrt{AT \log(AT)} + 4C_1 C_3 S A H + C_1 H
\]
\[
+ 48C_1 S r_{max} \sqrt{AT \log(AT)} + 4C_1 C_3 S A r_{max} + C_1 r_{max}
\]
\[
\leq (\tau_1 \Delta R + H) \sqrt{T} + 48C_1 S (H + r_{max}) \sqrt{AT \log(AT)}
\]
\[
+ 4C_1 S A (r_{max} + H) \sqrt{T} + C_1 (H + r_{max})
\]
\[
= 48C_1 C_2 S \sqrt{AT \log(AT)} + (\tau_1 \Delta R + H + 4C_1 C_2 S A) \sqrt{T} + C_1 C_2.
\]
Thus, we get the final Theorem.

**Theorem 2.** Suppose Assumptions 1 holds and the Oracle returns the optimal policy in each episode. The Bayesian regret of our algorithm satisfies
\[
R_T \leq 48C_1 C_2 S \sqrt{AT \log(AT)} + (\tau_1 \Delta R + H + 4C_1 C_2 S A) \sqrt{T} + C_1 C_2,
\]
where \( C_1 = L_1 + L_2 M N + N + S^2 \), \( C_2 = r_{max} + H \) are constants independent with time horizon \( T \), \( L_1 = 4M \left( \frac{1 - \epsilon}{\epsilon} \right)^2 / \min \{ R_{\min}^*, 1 - R_{\max}^* \} \), \( L_2 = 4M (1 - \epsilon)^2 / \epsilon^3 + \sqrt{M} \), \( M = S^N \), \( R_{\max}^* \) and \( R_{\min}^* \) are the maximum and minimum element of the functions \( R^* \) respectively, \( \tau_1 \) is the fixed exploration length in each episode, \( \Delta R \) is the biggest gap of the reward obtained at each two different time, \( H \) is the bounded span, \( r_{max} \) is the maximum reward obtain each time.

\[ \blacksquare \]
C Posterior distribution

Note that we assume the state transition is independent of the action for each arm. Denote the states visited history from time 0 till \(t\) of arm \(i\) as \(\bar{s}_i^t\) and the reward collected history is \(\bar{r}_i^t\). And \(N_{s,s'}^{i} (\bar{s}_i)\) is the occurrence time of state evolves from \(s\) to \(s'\) for arm \(i\) in the state history \(\bar{s}_i^t\). Hence, if the prior \(g (P_t (s, \cdot))\) is Dirichlet \((\phi_{s,s_1}^i, \ldots, \phi_{s,s_{s_t}}^i)\), then after the observation of history \(\bar{s}_i^t\), the posterior \(g (P_t (s, \cdot) | \bar{s}_i^t)\) is Dirichlet \((\phi_{s,s_1}^i + N_{s,s_1}^i (\bar{s}_i), \ldots, \phi_{s,s_{s_t}}^i + N_{s,s_{s_t}}^i (\bar{s}_i))\) (Ross et al., 2011). The posterior update for \(R_t (s, \cdot)\) is similar to \(P_t (s, \cdot)\), where \(N_{s,s'} (\bar{s}_i, \bar{a}_{t-1}) = \sum_{i=1}^{t-1} I_{(s,s')} (s_i, s_{t+1})\) is the number of times the transition \((s, s')\) appears in the history. Similarly, if the prior \(g (R_t (s, \cdot))\) is Dirichlet \((\psi_{s,r_1}^i, \ldots, \psi_{s,r_{r_t}}^i)\), then after the observation of reward history \(\bar{r}_i^t\) and \(\bar{s}_i^t\), the posterior \(g (R_t (s, \cdot) | \bar{r}_i^t, \bar{s}_i^t)\) is Dirichlet \((\psi_{s,r_1}^i + N_{s,r_1}^i (\bar{s}_i, \bar{r}_i), \ldots, \psi_{s,r_{r_t}}^i + N_{s,r_{r_t}}^i (\bar{s}_i, \bar{r}_i))\), and \(N_{s,r}^i\) is the number of times the observation \((s, r)\) appears in the history \((\bar{s}_i, \bar{r}_i)\).

Here we drop the arm index and consider a fixed arm. For the unknown transition functions, assume it follows the Dirichlet prior distribution. Let \(\bar{r}_t = (r_1, r_2, \ldots, r_t)\) be the history of observations of the agent up to time \(t\). Recall also that we denote \(\bar{s}_t = (s_1, s_2, \ldots, s_t)\) and \(\bar{a}_{t-1} = (a_1, a_2, \ldots, a_{t-1})\) the history of visited states and actions respectively. Next we show the details that how to update the posterior distribution for unknown \(P\).

\[
g (P | \bar{a}_{t-1}, \bar{r}_{t-1}) = \frac{P (\bar{r}_{t-1}, s_t | P, \bar{a}_{t-1}) g (P, \bar{a}_{t-1})}{\int P (\bar{r}_{t-1}, s_t | P, \bar{a}_{t-1}) g (P, \bar{a}_{t-1}) dP} = \frac{\sum_{\bar{s}_{t-1} \in S_t} P (\bar{r}_{t-1}, s_t | P, \bar{a}_{t-1}) g (P) \prod_{i=1}^{t} P (s_i | s_{i-1})}{\sum_{\bar{s}_{t-1} \in S_t} g (P) [\prod_{s,s'} P (s' | s)^{N_{s,s'} (\bar{s}_t)}]} = \frac{\sum_{\bar{s}_{t-1} \in S_t} g (P) \prod_{s,s'} P (s' | s)^{N_{s,s'} (\bar{s}_t)}}{\sum_{\bar{s}_{t-1} \in S_t} g (P) \prod_{s,s'} P (s' | s)^{N_{s,s'} (\bar{s}_t)}}
\]

Next we show the Bayesian approach to learning unknown \(P\) and \(R\) with the history \((\bar{a}_{t-1}, \bar{r}_t)\). Since the current state \(s_t\) of the agent at time \(t\) is unknown in the POMDP, we consider a joint posterior \(g (s_t, P, R | \bar{a}_{t-1}, \bar{r}_t)\) over \(s_t, P\), and \(R\) (Ross et al., 2011). By the laws of probability and Markovian assumption of the POMDP, we have:

\[
g (s_t, P, R | \bar{a}_{t-1}, \bar{r}_{t-1}) \propto P (\bar{r}_{t-1}, s_t | P, R, \bar{a}_{t-1}) g (P, R, \bar{a}_{t-1})
\]

\[
\propto \sum_{\bar{s}_{t-1} \in S_t} P (\bar{r}_{t-1}, s_t | P, R, \bar{a}_{t-1}) g (P, R)
\]

\[
\propto \sum_{\bar{s}_{t-1} \in S_t} g (s_0, P, R) \prod_{i=1}^{t} P (s_i | s_{i-1}) R (r_i | s_i)
\]

\[
\propto \sum_{\bar{s}_{t-1} \in S_t} g (s_0, P, R) \prod_{s,s'} (P (s' | s)^{N_{s,s'} (\bar{s}_t)}) \times \prod_{s,r} (R (r | s)^{N_{s,r} (\bar{s}_t, \bar{r}_t)})
\]

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

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

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

where

\[
L_1 = 4M \left( \frac{1 - \epsilon}{\epsilon} \right)^2 / \min \{R^*_{\text{min}}, 1 - R^*_{\text{max}} \}, \quad L_2 = 4M (1 - \epsilon)^2 / \epsilon^3 + \sqrt{M},
\]

\(M\) is the Cartesian product of each arms’ state space, \(R^*_{\text{max}}\) and \(R^*_{\text{min}}\) are the maximum and minimum element of the functions \(R^*\) respectively.

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

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

Lemma 13. (Lemma 17 in Auer et al. (2008)) For any \(t \geq 1\), the probability that the true MDP \(M\) is not contained in the set of plausible MDPs \(\mathcal{M}(t)\) at time \(t\) is at most \(\frac{4}{15t^6}\), that is

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