Regime Switching Bandits

Xiang Zhou [*†] **Yi Xiong** [‡] **Ningyuan Chen** [§] **Xuefeng Gao** [¶]

Abstract

We study a multi-armed bandit problem where the rewards exhibit regime switching. Specifically, the distributions of the random rewards generated from all arms are modulated by a common underlying state modeled as a finite-state Markov chain. The agent does not observe the underlying state and has to learn the transition matrix and the reward distributions. We propose a learning algorithm for this problem, building on spectral method-of-moments estimations for hidden Markov models, belief error control in partially observable Markov decision processes and upper-confidence-bound methods for online learning. We also establish an upper bound $O(T^{2/3}\sqrt{\log T})$ for the proposed learning algorithm where T is the learning horizon. Finally, we conduct proof-of-concept experiments to illustrate the performance of the learning algorithm.

1 Introduction

The multi-armed bandit (MAB) problem is a popular model for sequential decision making with unknown information: the decision maker makes decisions repeatedly among I different options, or arms. After each decision she receives a random reward having an *unknown* probability distribution that depends on the chosen arm. The objective is to maximize the expected total reward over a finite horizon of T periods. The MAB problem has been extensively studied in various fields and applications including Internet advertising, dynamic pricing, recommender systems, clinical trials and medicine [12, 13, 43]. In the classical MAB problem, it is typically assumed that the random reward of each arm is i.i.d. (independently and identically distributed) over time and independent of the rewards from other arms. However, these assumptions do not necessarily hold in practice [10]. To address the drawback, a growing body of literature studies MAB problems with non-stationary rewards to capture temporal changes in the reward distributions in applications, see e.g. [11, 16, 20].

In this paper, we study a non-stationary MAB model with Markovian regime-switching rewards. We assume that the random rewards associated with all the arms are modulated by a *common unobserved* state (or regime) $\{M_t : t = 1, 2, \ldots\}$ modeled as a finite-state discrete-time Markov chain. This chain makes a transition at each period regardless of which arm is pulled and its transition probabilities are independent of the action chosen. Given $M_t = m$, the reward of arm i is i.i.d., whose distribution is denoted $Q(\cdot|m, i)$. Such structural change of the environment is usually referred to as regime switching in finance [33]. The agent doesn't observe or control the underlying state M_t, and has to learn the transition probability matrix P of $\{M_t\}$ as well as the distribution of reward of each arm $Q(\cdot|m, i)$, based on the observed historical rewards. The goal of the agent is to design a learning policy that decides which arm to pull in each period to minimize the expected regret over T periods.

[*]The first two authors (Xiang Zhou and Yi Xiong) have equal contribution.

[†]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong; 1911606962@qq.com

[‡]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong; yxiong@se.cuhk.edu.hk

[§]The Rotman School of Management, University of Toronto; ningyuan.chen@utoronto.ca

[¶]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong; xfgao@se.cuhk.edu.hk

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

The regime-switching models are widely used in various industries. For example, in finance, the sudden changes of market environments are usually modeled as hidden Markov chains. In revenue management and marketing, a firm may face a shift in consumer demand due to undetected changes in sentiment or competition. In such cases, when the agents (traders or firms) take actions (trading financial assets with different strategies or setting prices), they need to learn the reward and the underlying state at the same time. Our setup is designed to tackle such problems.

Our Contribution. Our study features novel designs in three regards: we propose a new bandit problem formulation, develop a new learning algorithm, and prove regret bounds with new techniques.

In terms of problem formulation, online learning with unobserved states has attracted some attention recently [8, 19]. We consider the strongest oracle among the studies, who knows P and $Q(\cdot|m, i)$, but doesn't observe the hidden state M_t. The oracle thus faces a partially observable Markov decision process (POMDP) [26] (with a long-run average reward objective). By reformulating it as a Markov decision process (MDP) with a continuous belief space (i.e., a distribution over hidden states), the oracle then solves the optimal policy (mapping belief states to actions) using the Bellman equation. Having sublinear regret benchmarked against the strong oracle, our algorithm has better theoretical performance than others with weaker oracles such as the best fixed arm [19] or memoryless policies (the action only depends on the current observation) [8].

In terms of algorithmic design, we propose a learning algorithm (see Algorithm 2) with two key ingredients. First, it builds on the recent advance on the estimation of the parameters of hidden Markov models (HMMs) using spectral method-of-moments methods [3, 2, 8]. It benefits from the theoretical finite-sample bound of spectral estimators, while the finite-sample guarantees of other alternatives such as maximum likelihood estimators remain an open problem [30]. Second, it builds on the well-known "upper confidence bound" (UCB) method in reinforcement learning [7, 25]. There are two difficulties here as the oracle uses the optimal (belief-based) policy of the POMDP. First, the spectral method can not use the non i.i.d. samples generated from the belief-based policy due to the complex history dependency. Second, the belief of the hidden state is subject to the estimation error. Hence, we divide the horizon into nested exploration and exploitation phases. We use spectral estimators in the exploration phase to gauge the estimation error of P and $Q(\cdot|m, i)$. We use the UCB method to control the regret in the exploitation phase. Different from other learning problems, we re-calibrate the belief at the beginning of each exploitation phase based on the parameters estimated in the most recent exploration phase using previous exploratory samples.

In terms of technical analysis, we establish a regret bound of $O(T^{2/3}\sqrt{\log(T)})$ for our proposed learning algorithm where T is the learning horizon. Our regret analysis draws inspirations from [25, 35] for learning MDPs and undiscounted reinforcement learning problems, but the analysis differs significantly from theirs since there are two main technical challenges in our problem.

First, to control the regret, we need to control of the error of the belief state, which itself is not directly observed and needs to be estimated. This is in stark contrast to learning MDPs [25, 35] with observed states. Specifically, we need to bound the estimation error of belief states by the estimation errors of the model parameters. In addition, since the belief state is un-observable, we also need to bound the error in the belief transition kernel, which measures the distance between the belief transition kernel under the optimistic belief MDP model (from the UCB component of our algorithm) at each episode and the belief transition kernel under the true model. These bounds are not trivial since the transition kernel of the belief state depends on the model parameters in a complex way via Bayesian updating. We overcome the difficulties by building on [18] and a delicate analysis of the belief transition kernel to control the errors in the estimations of belief states and the belief transitions.

Second, to establish regret bound, we need an explicit bound for the span of the bias function (also referred as the relative value function) for the belief MDP which has a continuous state space. Such a bound is often critical in the regret analysis of undiscounted reinforcement learning of continuous MDP, but it is either *taken as an assumption* [39] or proved under Hölder continuity assumptions that do not hold for the belief transitions in our setting [35, 27]. We overcome this challenge and bound the bias span by developing a novel approach, which could be of independent interest for learning continuous-state MDPs. Specifically, we bound the bias span by bounding the Lipschitz module of the bias function for our infinite-horizon undiscounted problem (with a long-run average reward objective). To achieve this, we rely on a non-trivial application of the results in [23] which provide general tools for proving Lipschitz continuity of value functions in *finite-horizon discounted* MDPs. One key step is to bound the Lipschitz module of the belief transitions using the Kantorovich

metric, which allows us to establish a bound on the Lipschitz module of the value function of the finite-horizon discounted problem *uniformly over the discounting factors*. Exploiting the connection with the infinite-horizon undiscounted problem then yields an explicit bound on the bias span for our problem. We also mention that our bound on the bias span is unrelated to the diameter of the POMDP discussed in [8]. The diameter in [8] is only for observation-based policies, not for belief-state based policies we consider. That also explains why we need a new approach to bound the bias span.

Related Work. Studies on non-stationary/switching MAB investigate the problem when the rewards of all arms may change over time [5, 20, 10, 6]. Our formulation can be regarded as non-stationary MAB with a special structure. They consider an even stronger oracle than ours, the best arm in each period. However, the total number of changes or the changing budget have to be sublinear in T to achieve sublinear regret. In our formulation, the total number of transitions of the underlying state is linear in T, and the algorithms proposed in these papers fail even considering our oracle (See Section 5). Other studies focus on linear changing budget with some structure such as seasonality [15], which is not present in this paper.

Our work is also related to the restless Markov bandit problem [21, 36, 44] in which the state of each arm evolves according to independent Markov chains. In contrast, our regime-switching MAB model assumes a *common* underlying Markov chain so that the rewards of all arms are correlated, and the underlying state is unobservable to the agent. In addition, our work is related to MAB studies where rewards of all arms depend on a common unknown parameter or a latent random variable, see, e.g., [4, 22, 28, 32, 34]. Our model differs from them in that the common latent state variable follows a dynamic stochastic process which introduces difficulties in algorithm design and analysis.

Two papers have similar settings to ours. [19] studies MAB problems whose rewards are modulated by an unobserved Markov chain and the transition matrix may depend on the action. However, their oracle is the best fixed arm when defining the regret, which is much weaker than the optimal policy of the POMDP (the performance gap is linear between the two oracles). Therefore, their algorithm is expected to have linear regret when using our oracle. [8] proposes an algorithm based on spectral estimators for learning POMDPs. Their algorithm will generally suffer linear regret in our problem setting. This is because we consider a stronger oracle which is the optimal POMDP policy, whereas their oracle is the optimal memoryless policy, i.e., a policy that only depends on the current reward observation instead of using all historical observations to form the belief of the underlying state. It is known that memoryless policies are in general not optimal and hence the gap between their oracle and our oracle can be linear. Note that considering the memoryless policy allows [8] to circumvent the introduction of the belief entirely. Indeed, the dynamics under the memoryless policy can be viewed as a finite-state (modified) HMM and spectral estimators can be applied. Instead, because we consider the optimal belief-based policy, such reduction is not available. Our algorithm and regret analysis hinge on the interaction between the estimation of the belief and the spectral estimators. Our algorithm needs separate exploration to apply the spectral methods and uses the belief-based policy for exploitation. For the regret analysis, unlike [8], we have to carefully control the belief error and bound the span of the bias function from the optimistic belief MDP in each episode. The comparison of our study with related papers is summarized in Table 1. Note that we only present the regret in terms of T.

Papers	Oracle	Changing Budget	Regret
[5]	Best fixed action	Linear	$\tilde{O}(\sqrt{T})$
[20]	Best action in each period	Finite	$\tilde{O}(\sqrt{T})$
[10]	Best action in each period	Sublinear	$\tilde{O}(T^{2/3})$
[8]	Optimal memoryless policy	Linear	$\tilde{O}(\sqrt{T})$
This paper	Optimal POMDP policy	Linear	$\tilde{O}(T^{2/3})$

Table 1: Comparison of our study with some related literature.

2 MAB with Markovian Regime Switching

2.1 Problem Formulation

Consider the MAB problem with arms $\mathcal{I} := \{1, \ldots, I\}$. There is a Markov chain $\{M_t\}$ with states $\mathcal{M} := \{1, 2, \ldots, M\}$ and transition probability matrix $P \in \mathbb{R}^{M \times M}$. In period $t = 1, 2, \ldots$, if the state of the Markov chain $M_t = m$ and the agent chooses arm $I_t = i$, then the reward in that period

is R_t with discrete finite support, and its distribution is denoted by $Q(\cdot|m,i) := \mathbb{P}(R_t \in \cdot | M_t = m, I_t = i)$, with $\mu_{m,i} := \mathbb{E}[R_t | M_t = m, I_t = i]$. We use $\boldsymbol{\mu} := (\mu_{m,i}) \in \mathbb{R}^{M \times I}$ to denote the mean reward matrix. The agent knows M and I, but has no knowledge about the underlying state M_t (also referred to as the regime), the transition matrix P or the reward distribution $Q(\cdot|m,i)$. The goal is to design a learning policy that is adapted to the filtration generated by the observed rewards to decide which arm to pull in each period to maximize the expected cumulative reward over T periods where T is unknown in advance.

If an oracle knows P, $Q(\cdot|m,i)$ and the underlying state M_t, then the problem becomes trivial as s/he would select $I_t^* = \text{argmax}_{i \in \mathcal{I}} \mu_{M_t,i}$ in period t. If we benchmark a learning policy against the oracle, then the regret must be linear in T, because the oracle always observes M_t while the agent cannot predict the transition based on the history. Whenever a transition occurs, there is non-vanishing regret incurred. Since the number of transitions during $[0,T]$ is linear in T, the total regret is of the same order. Since comparing to the oracle knowing M_t is uninformative, we consider a weaker oracle who knows P, $Q(\cdot|m,i)$, but not M_t. In this case, the oracle solves a POMDP since the states M_t are unobservable and the optimal policy maps belief states (a distribution over the hidden state) to actions. The total expected reward of the POMDP scales linearly in T, and asymptotically the reward per period converges to a constant denoted by ρ^* under the optimal belief-based policy. See Section 2.2.

For a learning policy π, we denote by R_t^π the reward received under the learning policy π (which does not know $P, Q(\cdot|m,i)$ initially) in period t. We follow the literature (see, e.g., [25, 36, 1]) and define its total regret after T periods by

$$\mathcal{R}_T := T\rho^* - \sum_{t=1}^{T} R_t^\pi. \tag{1}$$

Our goal is to design a learning algorithm with theoretical guarantees including high probability and expectation bounds (sublinear in T) on the total regret.

Without loss of generality we consider Bernoulli rewards with mean $\mu_{m,i} \in (0,1)$ for all m, i. Hence $\mu_{m,i}$ characterizes the distribution $Q(\cdot|m,i)$. Our analysis holds generally for random rewards with discrete finite support. In addition, we impose the following assumptions.

Assumption 1. *The transition matrix P of the Markov chain $\{M_t\}$ is invertible.*

Assumption 2. *The mean reward matrix $\boldsymbol{\mu} = (\mu_{m,i})$ has full row rank.*

Assumption 3. *The smallest element of the transition matrix $\epsilon := \min_{i,j \in \mathcal{M}} P_{ij} > 0$.*

The first two assumptions are required for finite-sample guarantees of spectral estimators for HMMs [3, 2]. The third assumption is needed to control the belief error of the hidden states by the state-of-art methods, see [18] and our Proposition 3. We next reformulate the POMDP as a belief MDP.

2.2 Reduction of POMDP to Belief MDP

To present our learning algorithm and analyze the regret, we first investigate the POMDP problem faced by the oracle where parameters P, $\boldsymbol{\mu}$ (equivalently Q for Bernoulli rewards) are known with unobserved states M_t. Based on the historical observed history, the oracle forms a belief of the underlying state. The belief can be encoded by a M-dimension vector $b_t = (b_t(1), \ldots, b_t(M)) \in \mathcal{B}$: $b_t(m) := \mathbb{P}(M_t = m | I_1, \cdots, I_{t-1}, R_1, \cdots, R_{t-1})$, where $\mathcal{B} := \left\{ b \in \mathbb{R}_+^M : \sum_{m=1}^M b(m) = 1 \right\}$. It is well known that the POMDP of the oracle can be seen as a MDP built on a (continuous) belief state space $\mathcal{B}$, and here a policy is a mapping from belief states to actions, see [26].

We next introduce a few notations that facilitate the analysis. For notation simplicity, we write $c(m,i) := \mu_{m,i}$. Given belief $b \in \mathcal{B}$, the expected reward of arm i is $\bar{c}(b,i) := \sum_{m=1}^M c(m,i) b(m)$. In period t, given belief $b_t = b, I_t = i, R_t = r$, by Bayes' theorem, the belief b_{t+1} is updated by $b_{t+1} = H_{\boldsymbol{\mu},P}(b,i,r)$, where the m-th entry is $b_{t+1}(m) = \frac{\sum_{m'} P(m',m) \cdot (\boldsymbol{\mu}_{m',i})^r (1-\boldsymbol{\mu}_{m',i})^{1-r} \cdot b_t(m')}{\sum_{m''} (\boldsymbol{\mu}_{m'',i})^r (1-\boldsymbol{\mu}_{m'',i})^{1-r} \cdot b_t(m'')}$. It is obvious that the forward function H depends on the transition matrix P and the reward matrix $\boldsymbol{\mu}$. We can also define the transition probability of the belief state conditional on the arm pulled: $\bar{T}(\cdot|b,i) := \mathbb{P}(b_{t+1} \in \cdot | b, i)$, where b_{t+1} is random due to the random reward.

The long-run average reward of the infinite-horizon belief MDP following policy π given the initial belief b can be written as $\rho_b^\pi := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}[\sum_{t=1}^T R_t^\pi | b_1 = b]$. The optimal (belief-based) policy maximizes ρ_b^π for a given b. One can show that $\sup_\pi \rho_b^\pi$ is independent of the initial belief b (Proposition 8.2.1 [38]). Therefore, we use $\rho^* := \sup_\pi \rho_b^\pi$ to denote the optimal long-run average reward. Under this belief MDP formulation, for all $b \in \mathcal{B}$, the Bellman equation states that

$$\rho^* + v(b) = \max_{i\in\mathcal{I}} \left[\bar{c}(b,i) + \int_{\mathcal{B}} \bar{T}(db'|b,i) v(b') \right], \tag{2}$$

where $v : \mathcal{B} \mapsto \mathbb{R}$ is the bias function. It can be shown (see the proof of Proposition 2 in Appendix) that under our assumptions, ρ^* and $v(b)$ are well defined and there exists a stationary deterministic optimal policy π^* which maps a belief state to an arm to pull (an action that maximizes the right side of (2)). Finding the optimal policy for the POMDP model is computationally intractable in general [37, 31]. Therefore, various approximate methods have been proposed to solve (2) and to find the optimal policy for the belief MDP, see e.g. [46, 41, 42]. In this work, we do not focus on this planning problem for a known model, and we assume the access to an optimization oracle that solve (2) and returns the optimal average reward ρ^* and the optimal stationary policy for a given known model.

3 The SEEU Algorithm

This section describes our learning algorithm for the regime switching MAB model: the Spectral Exploration and Exploitation with UCB (SEEU) algorithm. To device a learning policy for the POMDP with unknown $\boldsymbol{\mu}$ and P, one needs a procedure to estimate those quantities from observed rewards. [3, 2] propose the so-called spectral estimator for the unknown parameters in HMMs. However, the algorithm is not directly applicable to ours, because there is no decision making involved in HMMs . To use the spectral estimator, we divide the learning horizon T into nested "exploration" and "exploitation" phases. In the exploration phase, we randomly select an arm in each period. This transforms the system into a HMM so that we can apply the spectral method to estimate $\boldsymbol{\mu}$ and P from the observed rewards in the phase. In the exploitation phase, based on the estimators obtained from the exploration phase, we use a UCB-type policy to further narrow down the optimal belief-based policy in the POMDP introduced in Section 2.2.

3.1 Spectral Estimator

We introduce the spectral estimator [3, 2], and adapt it to our setting. To simplify the notation, suppose the exploration phase starts from period 1 until period n, with realized arms $\{i_1, \dots, i_n\}$, and realized rewards $\{r_1, \dots, r_n\}$ sampled from Bernoulli distributions. Recall I is the cardinality of the arm set $\mathcal{I}$, then one can create a one-to-one mapping from a pair (i, r) into a scalar $s \in \{1, 2, ..., 2I\}$. Therefore, the pair can be expressed as a vector $y \in \{0,1\}^{2I}$ such that in each period t, y_t satisfies $\mathbb{1}_{\{y_t = e_s\}} = \mathbb{1}_{\{r_t = r, i_t = i\}}$, where e_s is a basis vector with its s-th element being one and zero otherwise. Let $A \in \mathbb{R}^{2I\times M}$ be the observation probability matrix conditional on the state: $A(s,m) = \mathbb{P}(R_t = r, I_t = i | M_t = m)$. It can be shown that A satisfies $\mathbb{E}[y_t | M_t = m] = Ae_m$, and $\mathbb{E}[y_{t+1} | M_t = m] = AP^T e_m$. Write $\otimes$ for the tensor product. For three consecutive observations y_{t-1}, y_t, y_{t+1}, define

$$\widetilde{y}_{t-1} := \mathbb{E}[y_{t+1} \otimes y_t]\mathbb{E}[y_{t-1} \otimes y_t]^{-1} y_{t-1}, \qquad \widetilde{y}_t := \mathbb{E}[y_{t+1} \otimes y_{t-1}]\mathbb{E}[y_t \otimes y_{t-1}]^{-1} y_t,$$
$$M_2 := \mathbb{E}[\widetilde{y}_{t-1} \otimes \widetilde{y}_t], \qquad M_3 := \mathbb{E}[\widetilde{y}_{t-1} \otimes \widetilde{y}_t \otimes y_{t+1}].$$

From the observations $\{y_1, y_2, \dots, y_n\}$, we may construct the estimations $\hat{M}_2$ and $\hat{M}_3$ for M_2 and M_3 respectively, and apply the tensor decomposition to obtain the estimator $\hat{\boldsymbol{\mu}}$ for the unknown mean reward matrix and $\hat{P}$ for the transition matrix. This procedure is summarized in Algorithm 1. We use $\hat{A}_m$ (respectively $\hat{B}_m$) to denote the m-th column vector of $\hat{A}$ (respectively $\hat{B}$). In addition, we have the following result ([8]) and it provides the confidence regions of the estimators in Algorithm 1.

Proposition 1. *Under Assumptions 1 and 2, for any $\delta \in (0,1)$ and any initial distribution, there exists N_0 such that when $n \geq N_0$, with probability $1-\delta$, the estimated $\hat{\boldsymbol{\mu}}$ and $\hat{P}$ by Algorithm 1*

Algorithm 1 Spectral estimation of $(\boldsymbol{\mu}, P)$ from the observations from the exploration phase [2, 3, 8].

Input: sample size n, $\{y_1, y_2, \ldots, y_n\}$ created from the rewards $\{r_1, \ldots, r_n\}$ and arms $\{i_1, \ldots, i_n\}$
Output: The estimation $\hat{\boldsymbol{\mu}}, \hat{P}$

1: For $i, j \in \{-1, 0, 1\}$: compute $\hat{W}_{i,j} = \frac{1}{N-2}\sum_{t=2}^{N-1} y_{t+i} \otimes y_{t+j}$.
2: For $t = 2, \ldots, n-1$: compute $\hat{y}_{t-1} := \hat{W}_{1,0}(\hat{W}_{-1,0})^{-1} y_{t-1}$, $\hat{y}_t := \hat{W}_{1,-1}(\hat{W}_{0,-1})^{-1} y_t$.
3: Compute $\hat{M}_2 := \frac{1}{N-2}\sum_{t=2}^{N-1} \hat{y}_{t-1} \otimes \hat{y}_t$, $\hat{M}_3 := \frac{1}{N-2}\sum_{t=2}^{N-1} \hat{y}_{t-1} \otimes \hat{y}_t \otimes y_{t+1}$.
4: Apply tensor decomposition ([2]):
$\hat{B} =$ **TensorDecomposition**$(\hat{M}_2, \hat{M}_3)$.
5: Compute $\hat{A}_m = \hat{W}_{-1,0}(\hat{W}_{1,0})^{\dagger}\hat{B}_m$ for each $m \in \mathcal{M}$.
6: Return mth row vector $(\hat{\boldsymbol{\mu}})^m$ of $\hat{\boldsymbol{\mu}}$ from $\hat{A}_m$.
7: Return $\hat{P} = (\hat{A}^{\dagger}\hat{B})^{\top}$ ($\dagger$ represents the pseudoinverse of a matrix)

satisfy

$$||(\boldsymbol{\mu})^m - (\hat{\boldsymbol{\mu}})^m||_2 \leq C_1\sqrt{\frac{\log(6\frac{S^2+S}{\delta})}{n}}, \quad m \in \mathcal{M},$$
$$||P - \hat{P}||_2 \leq C_2\sqrt{\frac{\log(6\frac{S^2+S}{\delta})}{n}}. \tag{3}$$

where $(\boldsymbol{\mu})^m$ and $(\hat{\boldsymbol{\mu}})^m$ are the m-th row vectors of $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$, respectively. Here, $S = 2I$, and C_1, C_2 are constants independent of n.

The expressions of constants N_0, C_1, C_2 are given in Section B in the appendix. Note that parameters $\boldsymbol{\mu}, P$ are identifiable up to permutations of the hidden states [8].

3.2 The SEEU Algorithm

The SEEU algorithm proceeds in episodes of increasing length. As mentioned before, each episode is divided into exploration and exploitation phases. In episode k, it starts with the exploration phase that lasts for a fixed number of periods τ_1, and the algorithm uniformly randomly chooses an arm and observes the rewards. After the exploration phase, the algorithm applies Algorithm 1 to (re-)estimate $\boldsymbol{\mu}$ and P. Moreover, it constructs a confidence interval based on Proposition 1 with a confidence level $1 - \delta_k$, where $\delta_k := \delta/k^3$ is a vanishing sequence. Then the algorithm enters the exploitation phase. Its length is proportional to $\sqrt{k}$. In the exploitation phase, it conducts UCB-type learning: the arm is pulled according to a policy that corresponds to the optimistic estimator of $\boldsymbol{\mu}$ and P inside the confidence interval. The detailed steps are listed in Algorithm 2. Note we use the UCB component in the exploitation phase instead of point estimators of $\boldsymbol{\mu}$ and P for the ease of analysis.

3.3 Discussions on the SEEU Algorithm

Computations. For given parameters $(\boldsymbol{\mu}, P)$, we need to compute the optimal average reward $\rho^*(\boldsymbol{\mu}, P)$ that depends on the parameters (Step 8 in Algorithm 2). Various computational and approximation methods have been proposed to tackle this planning problem for belief MDPs as mentioned in Section 2.2. In addition, we need to find out the optimistic POMDP in the confidence region $\mathcal{C}_k(\delta_k)$ with the best average reward. In general it is not clear whether there is an efficient computational method to find the optimistic plausible POMDP model in the confidence region when the unknown parameters are high-dimensional. The extended value iteration method in [25] does not work in our POMDP setting. This is because we cannot separately find $\boldsymbol{\mu}$ and P as $\bar{c}$ and $\bar{T}$ in (2) both depend on $\boldsymbol{\mu}$, and we cannot write the inner optimization in the extended value iteration to find optimistic P as a linear programming over the convex polytope $\mathcal{C}_k(\delta_k)$ since $\bar{T}$ is a nonlinear function of $\boldsymbol{\mu}$ and P. This issue is also present in recent studies on learning continuous-state MDPs with the upper confidence bound approach, see e.g. [27] for further discussions. For low dimensional models, one can discretize $\mathcal{C}_k(\delta_k)$ into grids and calculate the corresponding optimal average reward ρ^* at each grid point so as to find (approximately) the optimistic model $(\boldsymbol{\mu}_k, P_k)$. The discretization error does affect the regret, although it can be controlled arbitrarily well with sufficient computational capacity.

Algorithm 2 The SEEU Algorithm

Input: Initial belief b_1, precision δ, exploration parameter τ_1, exploitation parameter τ_2

1: **for** $k = 1, 2, 3, \ldots$ **do**
2: Set the start time of episode k, $t_k := t$
3: **for** $t = t_k, t_k + 1, \ldots, t_k + \tau_1$ **do**
4: Uniformly randomly select an arm: $\mathbb{P}(I_t = i) = \frac{1}{I}$
5: **end for**
6: Input the realized actions and rewards in all previous exploration phases $\hat{\mathcal{I}}_k := \{i_{t_1:t_1+\tau_1}, \cdots, i_{t_k:t_k+\tau_1}\}$ and $\hat{\mathcal{R}}_k := \{r_{t_1:t_1+\tau_1}, \cdots, r_{t_k:t_k+\tau_1}\}$ to Algorithm 1 to compute $\hat{\boldsymbol{\mu}}_k, \hat{P}_k =$ **Spectral Estimation**$(\hat{\mathcal{I}}_k, \hat{\mathcal{R}}_k)$
7: Compute the confidence interval $\mathcal{C}_k(\delta_k)$ from (3) using the confidence level $1 - \delta_k = 1 - \frac{\delta}{k^3}$ such that $\mathbb{P}\{(\boldsymbol{\mu}, P) \in \mathcal{C}_k(\delta_k)\} \geq 1 - \delta_k$
8: Find the optimistic POMDP in the confidence interval
$$(\boldsymbol{\mu}_k, P_k) = \operatorname{argmax}_{(\boldsymbol{\mu}, P) \in \mathcal{C}(\delta_k)} \rho^*(\boldsymbol{\mu}, P)$$
9: **for** $t = 1, 2, \ldots, t_k + \tau_1$ **do**
10: Update belief b_t^k to $b_{t+1}^k = H_{\boldsymbol{\mu}_k, P_k}(b_t^k, i_t, r_t)$ under the new parameters $(\boldsymbol{\mu}_k, P_k)$
11: **end for**
12: **for** $t = t_k + \tau_1 + 1, \ldots, t_k + \tau_1 + \tau_2\sqrt{k}$ **do**
13: Execute the optimal policy $\pi^{(k)}$ by solving the Bellman equation (2) under parameters $(\boldsymbol{\mu}_k, P_k)$: $i_t = \pi^{(k)}(b_t^k)$
14: Observe reward r_t
15: Update the belief at $t + 1$ by $b_{t+1}^k = H_{\boldsymbol{\mu}_k, P_k}(b_t^k, i_t, r_t)$
16: **end for**
17: **end for**

Below we discuss its impact on regret. The discretization kicks in when we want to find the optimistic POMDP in the confidence region, whose gain is denoted as ρ^k in episode k. In practice, we have to discretize the confidence region into grid points and find the optimistic POMDP among the finite set. Suppose one can obtain an approximate optimistic model with error ϵ_k, that is, suppose we can find a model with the gain $\tilde{\rho}_k \geq \rho^k - \epsilon_k$. Then we can infer from formula (22) in the appendix that the extra regret incurred due to the discretization error is given by $\sum_{k=1}^{K} \sum_{t \in E_k} \epsilon_k = \tau_2 \sum_{k=1}^{K} \sqrt{k}\epsilon_k$, where E_k denotes the exploitation phase in episode k. One can show that the order of K, the number of episodes up to time T, is $T^{2/3}$. Hence if the discretization error can be controlled at $\epsilon_k = c/\sqrt{k}$, then the extra regret is simply $c\tau_2 K$, which is $O(T^{2/3})$. On the other hand, if the discretization error ϵ_k is a constant $c > 0$ in all the exploitation phases, then the extra regret incurred is of the order $\sum_{k=1}^{K} \sqrt{k}c \approx c \cdot K^{3/2}$, which is of order cT. There is a trade-off between the additional computational complexity due to discretization and the regret bound. For instance, to control the discretization error at $\epsilon_k = c/\sqrt{k}$, the computational cost is higher compared with the case $\epsilon_k = c$. In general, it remains open to find efficient methods to solve the optimistic POMDP approximately in the high-dimensional setting. In our regret analysis below, we do not take into account approximation errors arising from the computational aspects discussed above, as in [35, 27].

Dependence on the unknown parameters. When computing the confidence region in Step 7 of Algorithm 2, the agent needs the information of the constants C_1 and C_2 in Proposition 1. These constants depend on a few "primitives" that can be hard to know, for example, the mixing rate of the underlying Markov chain. However we only need upper bounds for C_1 and C_2 for the theoretical guarantee, and hence a rough and conservative estimate would be sufficient. Such dependence on some unknown parameters is common in learning problems, and one remedy is to dedicate the beginning of the horizon to estimate the unknown parameters, which typically doesn't increase the rate of the regret. Alternatively, C_1 and C_2 can be replaced by parameters that are tuned by hand. See Remark 3 of [8] for a further discussion on this issue.

4 Regret Bound

This section presents our main theoretical results on the regret bound for the SEEU algorithm. We first state two technical results that are important in proving the regret bounds.

Proposition 2 (Uniform bound on the bias span)**.** *If the belief MDP satisfies Assumption 3, then for (ρ, v) satisfying the Bellman equation* (2)*, we have the span of the bias function span$(v) := \max_{b\in\mathcal{B}} v(b) - \min_{b\in\mathcal{B}} v(b)$ is bounded by $D(\epsilon)$, where*

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

Recall v_k is the bias function for the optimistic belief MDP in episode k. Proposition 2 guarantees that $span(v_k)$ is bounded by $D = D(\epsilon/2)$ uniformly in k, because Assumption 3 can be satisfied (with ϵ replaced by $\epsilon/2$) by the optimistic MDPs when T is sufficiently large due to Proposition 1.

Proposition 3 (Controlling the belief error)**.** *Suppose Assumption 3 holds. Given $(\hat{\boldsymbol{\mu}}, \hat{P})$, an estimator of the true model parameters $(\boldsymbol{\mu}, P)$. For an arbitrary reward-action sequence $\{r_{1:t}, i_{1:t}\}_{t\geq 1}$, let $\hat{b}_t$ and b_t be the corresponding beliefs in period t under $(\hat{\boldsymbol{\mu}}, \hat{P})$ and $(\boldsymbol{\mu}, P)$ respectively. Then there exists constants L_1, L_2 such that*

$$||\hat{b}_t - b_t||_1 \leq L_1||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||_1 + L_2||\hat{P} - P||_F, \tag{4}$$

where $L_1 = \frac{4M(1-\epsilon)^2}{\epsilon^2 \min\{\boldsymbol{\mu}_{\min}, 1-\boldsymbol{\mu}_{\max}\}}$, $L_2 = \frac{4M(1-\epsilon)^2}{\epsilon^3} + \sqrt{M}$, $||\cdot||_F$ is the Frobenius norm, $\boldsymbol{\mu}_{\max}$ and $\boldsymbol{\mu}_{\min}$ are the maximum and minimum element of the matrix $\boldsymbol{\mu}$ respectively.

We now state our first main result. The proof is given in Appendix E.

Theorem 1. *Suppose Assumptions 1 to 3 hold. Fix the parameter τ_1 in Algorithm 2 to be sufficiently large. Then there exist constants T_0, C which are independent of T, such that when $T > T_0$, with probability $1 - \frac{7}{2}\delta$, the regret of Algorithm 2 satisfies*

$$\mathcal{R}_T \leq CT^{2/3}\sqrt{\log\left(\frac{9(S+1)}{\delta}T\right)} + T_0\rho^*,$$

where $S = 2I$ and ρ^* denotes the optimal long-run average reward under the true model.

The constant T_0 measures the number of periods needed for the sample size in the exploration phases to exceed N_0 arising in Proposition 1. The constant C has the following expression:

$$C = 3\sqrt{2}\left[\left(D + 1 + \left(1 + \frac{D(1-\alpha)}{2}\right)L_1\right)M^{3/2}C_1 + \left(1 + \frac{D(1-\alpha)}{2}\right)L_2 M^{1/2}C_2\right]\tau_2^{1/3}\tau_1^{-1/2}$$

$$+ 3\tau_2^{-2/3}(\tau_1\rho^* + D) + (D+1)\sqrt{2\ln(\frac{1}{\delta})}.$$

Here, M is the number of hidden states, C_1, C_2 are given in Proposition 1, α, D is given in Proposition 2, and L_1, L_2 are given in Proposition 3. One can verify that $C = \tilde{O}(\epsilon^{-4}C_1M^{5/2} + \epsilon^{-5}C_2M^{3/2})$. Here we require τ_1 to be large so as to apply Proposition 1 from the first episode to simplify the presentation. Theoretically, this requirement is not essential and can be removed without affecting the order of the regret. For τ_2, it can be an arbitrary positive integer in our regret analysis.

Remark 2 (Effect of lengths of exploration and exploitation phases on regret order)**.** *In Algorithm 2, we use an exploration phase of length τ_1 and a exploitation phase of length $\tau_2\sqrt{k}$ in episode k. This in the end leads to $O(T^{2/3})$ of the regret as the total length of exploration (and the total number of episodes) is $\Theta(T^{2/3})$. The intuition of such a choice is the following. After k episodes, the total length of exploration is $\tau_1 k$, and hence the estimation error of the model parameters and also the belief states is of order $1/\sqrt{k}$. As a result, we needs to choose an exploitation phase with length $\tau_2\sqrt{k}$ to balance this so that the regret incurred per episode is controlled. One might wonder whether one can simply choose a longer exploitation phase of length $\tau_2 k^\alpha$ with $\alpha \geq 1/2$, and then the regret order will be reduced. This is not possible, and the intuition is as follows. The total number of episodes K satisfies $\sum_{k=1}^K(\tau_1 + \tau_2 k^\alpha) = T$, which yields $K \sim T^{\frac{1}{\alpha+1}}$. The total regret incurred then is, ignoring the logarithmic factors, on the order of $\sum_{k=1}^K(\tau_1 + \frac{1}{\sqrt{k}}\cdot\tau_2 k^\alpha)$, which is $T^{\frac{\alpha+1/2}{\alpha+1}}$. One can immediately see that the choice of $\alpha = 1/2$ is actually optimal, leading to a regret of order $T^{2/3}$. The impact of the hyper-parameters τ_1 and τ_2 on the regret will be studied numerically in Section 5.*

From Theorem 1, we can choose an appropriate $\delta = \frac{9(S+1)}{T}$ and readily obtain the following expectation bound on the regret. The proof is omitted.

Theorem 3. *Under the same assumptions as Theorem 1, the regret of Algorithm 2 satisfies*

$$\mathbb{E}[\mathcal{R}_T] \leq CT^{2/3}\sqrt{2\log T} + (T_0 + 32(S+1))\rho^*.$$

Remark 4 (Lower bound)**.** *For the lower bound of the regret, consider the I problem instances (equal to the number of arms): In instance i, let $\mu_{m,i} = 0.5 + \bar{\epsilon}$ for all m for a small positive constant $\bar{\epsilon}$, and let $\mu_{m,j} = 0.5$ for all m and $j \neq i$. Such structure makes sure that the oracle policy simply pulls one arm without the need to infer the state. Since the problem reduces to the classic MAB, the regret is at least $O(\sqrt{IT})$ in this case. Note that the setup of the instances may violate Assumption 2, but this can be easily fixed by introducing an arbitrarily small perturbation to $\boldsymbol{\mu}$. The gap between the upper and lower bounds is probably caused by the split of exploration/exploitation phases in our algorithm. In the exploration phase, arms have to be pulled purely randomly in order to estimate the parameters. Such naive exploration without any maximization may lead to unnecessary regret. In fact, it resembles the structure of the explore-then-commit algorithm (Chapter 6 of [29]) for the classic MAB problem, which turns out to have the suboptimal regret $O(T^{2/3})$ as well. With special structures, such as a common mean for all states, special-purpose algorithms can be designed to achieve the optimal rate $O(T^{1/2})$. Unfortunately, our algorithm doesn't adapt to the special structure. In fact, the spectral estimator no longer works because of the degeneracy as it cannot differentiate between states from the observations. One natural direction that may potentially improve our algorithm and close the gap is to integrate the exploration and exploitation. We may follow the design similar to UCRL2 in [25]: in each episode, we use the parameters estimated in previous episodes and use the optimistic policy in the confidence region. The observations are then used to update the estimation. This can lower the cost for exploration because the naive exploration is replaced by a near-optimal policy when the confidence region is small. However, it requires the spectral estimator to work with non i.i.d. samples generated by the optimistic belief-based policy. The design of the spectral estimator makes it hard to handle non i.i.d. observations with complex history dependency. This is also why [8] focused on observation-based policies in order to apply the spectral estimator. We may replace the spectral estimator by other estimators with better theoretical properties. The likelihood-based estimators may be a promising candidate [45], but they have not been fully extended to POMDPs yet. In summary, we need better estimators for POMDP/HMMs (which is a dynamic research area itself) with finite-sample guarantees to improve the upper bound. Nevertheless, we are not aware of other algorithms that can achieve sublinear regret in our setting.*

Remark 5 (General reward distribution)**.** *Our model requires the estimation of the reward distribution instead of just the mean to estimate the belief. For discrete rewards taking O possible values, the regret bound holds with $S = OI$ instead of $2I$ for Bernoulli rewards. For continuous reward distributions, it might be possible to combine the non-parametric HMM inference method in [18] with our algorithm design to obtain regret bounds, and we leave it for future work.*

5 Numerical Experiment

In this section, we present proof-of-concept experiments. Note that large-scale POMDPs with long-run average objectives (the oracle in our problem) are computationally difficult to solve, i.e. it is challenging to find the optimal belief-based policy [14]. On the other hand, while there can be many hidden states in general, often only two or three states are important to model in several application areas, e.g. "bull market" and "bear market" in finance [17]. Hence we focus on small-scale experiments, following some recent literature on reinforcement learning for POMDPs [8, 24].

As a representative example, we consider a 2-hidden-state, 2-arm setting with $P = \begin{bmatrix} 1/3 & 2/3 \\ 3/4 & 1/4 \end{bmatrix}$ and $\boldsymbol{\mu} = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.6 \end{bmatrix}$, where the random reward follows a Bernoulli distribution. We compare our algorithm with (1) ϵ-greedy ($\epsilon = 0.1$), non-stationary bandits algorithms including (2) Sliding-Window UCB (SW-UCB) [20] with tuned window size, (3) Exp3.S [5] with $L = T$ (the hyperparameter L is the number of changes in their algorithm), and (4) the optimal memoryless policy for POMDPs discussed in [8]. To implement (4) under our model setting, we assume the reward μ and the transition matrix P are known and search for the optimal stochastic memoryless policy that maps the reward and action in the last period to a discrete distribution over two arms. We also run an algorithm/oracle

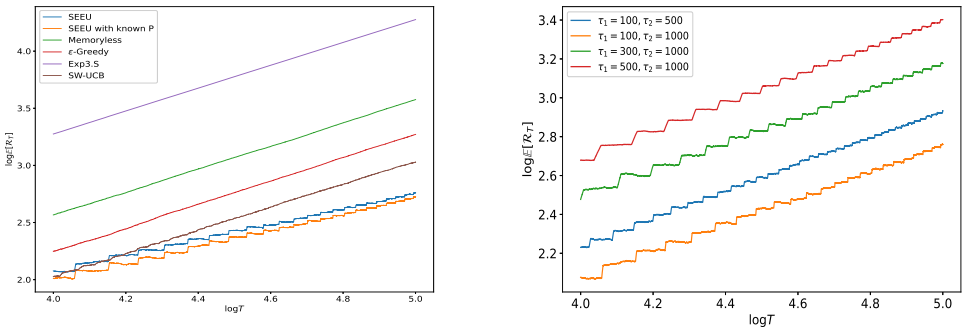

Figure 1: (a) Regret comparison of different algorithms; (b) Effect of (τ_1, τ_2) on the regret of SEEU.

that implements SEEU with known P but unknown μ. This can allow us to better understand the value of knowing the information of P. In Figure 1(a), we plot the average regret versus T of different algorithms in log-log scale, where the number of runs for each algorithm is 500. The numerical experiments are conducted on a PC with 3.10 GHz Intel Processor and 16 GB of RAM. We observe that the slopes of all algorithms except for two SEEU implementations are close to one, suggesting that they incur linear regrets. This is expected, because these algorithms don't take into account the hidden states. On the other hand, the slope of SEEU is close to $2/3$. This is consistent with our theoretical result (Theorem 3). Similar observations are made on other small-scale examples. This demonstrates the effectiveness of our SEEU algorithm, particularly when the horizon length T is relatively large.

We also briefly discuss the impact of parameters τ_1 and τ_2 on the performance of the SEEU algorithm. For the example above, we calculate the average regret for several pairs of parameters (τ_1, τ_2). It can be seen that the choices of these parameters do not affect the order $O(T^{2/3})$ of the regret (the slope). See Figure 1(b) for an illustration.

6 Conclusions, Limitations and Future Research

In this paper, we study a non-stationary MAB model with Markovian regime-switching rewards. We propose a learning algorithm that integrates spectral estimators for hidden Markov models and upper confidence methods from reinforcement learning. We also establish a regret bound of order of $O(T^{2/3}\sqrt{\log T})$ for the learning algorithm. As far as we know, this is the first algorithm with sublinear regret for MAB with unobservable regime switching.

The main limitation of our work is that the regret upper bound does not match the lower bound. It would be interesting to find out whether one can improve the regret upper bound $O(T^{2/3}\sqrt{\log T})$. The gap in the upper and lower bounds are likely to be filled if one can resolve either of the following two fundamental open questions: (1) show the spectral method can be applied to non i.i.d. samples generated from belief-based policies, and (2) establish finite-sample guarantees for other estimators (e.g. maximum likelihood estimators) of POMDP parameters with data generated from adaptive policies. We leave them for future research.

Acknowledgments and Disclosure of Funding

The authors thank the Area Chair and three anonymous referees for constructive comments and helpful suggestions. Xuefeng Gao acknowledges support from Hong Kong RGC GRF Grants 14201520 and 14201421. The research of Ningyuan Chen is partially supported by NSERC Discovery Grants RGPIN-2020-04038.

References

[1] S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *In Advances in Neural Information Processing Systems*, pages 1184–1194, 2017.

[2] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[3] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pages 33–1, 2012.

[4] O. Atan, C. Tekin, and M. van der Schaar. Global bandits. *IEEE transactions on neural networks and learning systems*, 29(12):5798–5811, 2018.

[5] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

[6] P. Auer, P. Gajane, and R. Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Conference on Learning Theory*, pages 138–158, 2019.

[7] P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56, 2007.

[8] K. Azizzadenesheli, A. Lazaric, and A. Anandkumar. Reinforcement learning of POMDPs using spectral methods. *arXiv preprint arXiv:1602.07764*, 2016.

[9] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[10] O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in neural information processing systems*, pages 199–207, 2014.

[11] O. Besbes, Y. Gur, and A. Zeevi. Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *Stochastic Systems*, 2019.

[12] D. Bouneffouf and I. Rish. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*, 2019.

[13] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[14] K. Chatterjee, R. Saona, and B. Ziliotto. The complexity of pomdps with long-run average objectives. *arXiv preprint arXiv:1904.13360*, 2019.

[15] N. Chen, C. Wang, and L. Wang. Learning and optimization with seasonal patterns. *Working paper*, 2020.

[16] W. C. Cheung, D. Simchi-Levi, and R. Zhu. Learning to optimize under non-stationarity. *arXiv preprint arXiv:1810.03024*, 2018.

[17] M. Dai, Q. Zhang, and Q. J. Zhu. Trend following trading under a regime switching model. *SIAM Journal on Financial Mathematics*, 1(1):780–810, 2010.

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

[19] T. Fiez, S. Sekar, and L. J. Ratliff. Multi-armed bandits for correlated markovian environments with smoothed reward feedback. *arXiv preprint arXiv:1803.04008*, 2019.

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

[21] S. Guha, K. Munagala, and P. Shi. Approximation algorithms for restless bandit problems. *Journal of the ACM (JACM)*, 58(1):3, 2010.

[22] S. Gupta, G. Joshi, and O. Yağan. Correlated multi-armed bandits with a latent random source. *arXiv preprint arXiv:1808.05904*, 2018.

[23] K. Hinderer. Lipschitz continuity of value functions in markovian decision processes. *Mathematical Methods of Operations Research*, 62(1):3–22, 2005.

[24] M. Igl, L. Zintgraf, T. A. Le, F. Wood, and S. Whiteson. Deep variational reinforcement learning for pomdps. *arXiv preprint arXiv:1806.02426*, 2018.

[25] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

[26] V. Krishnamurthy. *Partially observed Markov decision processes*. Cambridge University Press, 2016.

[27] K. Lakshmanan, R. Ortner, and D. Ryabko. Improved regret bounds for undiscounted continuous reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pages 524–532, 2015.

[28] T. Lattimore and R. Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pages 550–558, 2014.

[29] T. Lattimore and C. Szepesvári. *Bandit algorithms*. preprint, 2018.

[30] L. Lehéricy. Consistent order estimation for nonparametric hidden markov models. *Bernoulli*, 25(1):464–498, 2019.

[31] O. Madani, S. Hanks, and A. Condon. On the computability of infinite-horizon partially observable markov decision processes. In *AAAI98 Fall Symposium on Planning with POMDPs, Orlando, FL*. Citeseer, 1998.

[32] O. A. Maillard and S. Mannor. Latent bandits. In *International Conference on Machine Learning*, pages 136–144, 2014.

[33] R. S. Mamon and R. J. Elliott. *Hidden Markov models in finance*, volume 4. Springer, 2007.

[34] A. J. Mersereau, P. Rusmevichientong, and J. N. Tsitsiklis. A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802, 2009.

[35] R. Ortner and D. Ryabko. Online regret bounds for undiscounted continuous reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1763–1771, 2012.

[36] R. Ortner, D. Ryabko, P. Auer, and R. Munos. Regret bounds for restless markov bandits. *Theoretical Computer Science*, 558:62–76, 2014.

[37] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

[38] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, 2014.

[39] J. Qian, R. Fruit, M. Pirotta, and A. Lazaric. Exploration bonus for regret minimization in undiscounted discrete and continuous markov decision processes. *arXiv preprint arXiv:1812.04363*, 2018.

[40] S. M. Ross. Arbitrary state markovian decision processes. *The Annals of Mathematical Statistics*, 39(6):2118–2122, 1968.

[41] N. Saldi, S. Yüksel, and T. Linder. On the asymptotic optimality of finite approximations to markov decision processes with borel spaces. *Mathematics of Operations Research*, 42(4):945–978, 2017.

[42] H. Sharma, M. Jafarnia-Jahromi, and R. Jain. Approximate relative value learning for average-reward continuous state mdps. In *Uncertainty in Artificial Intelligence*, pages 956–964. PMLR, 2020.

[43] A. Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

[44] A. Slivkins and E. Upfal. Adapting to a changing environment: the brownian restless bandits. In *Conference on Learning Theory*, pages 343–354, 2008.

[45] F. Yang, S. Balakrishnan, and M. J. Wainwright. Statistical and computational guarantees for the baum-welch algorithm. *The Journal of Machine Learning Research*, 18(1):4528–4580, 2017.

[46] H. Yu and D. P. Bertsekas. Discretized approximations for pomdp with average cost. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 619–627. AUAI Press, 2004.

A Table of Notations

Notation	Description
T	The length of decision horizon
P	The transition matrix of underlying states
$\boldsymbol{\mu}$	The mean reward matrix
$\mathcal{M}$	The set of underlying states
$\mathcal{I}$	The set of arms
$\mathcal{R}$	The set of rewards
M_t	The underlying state at time t
I_t	The chosen arm at time t
R_t	The random reward at time t
$\mathcal{F}_t$	The history up to time t
ρ^*	The optimal long term average reward
$\mathcal{R}_T$	Regret during the total horizon
$\boldsymbol{b_t}$	The belief state at time t
$\mathcal{B}$	The set of belief states
$c(m,i)$	The reward function given the state and arm
$\bar{c}(b,i)$	The reward function w.r.t. belief state
Q	The reward distribution
H	The belief state forward kernel
$\bar{T}$	The transition kernel w.r.t. belief state
D	The uniform upper bound of span(v)
$\mathbb{E}^\pi$	Taken expectation respect to the true parameters $\boldsymbol{\mu}$ and P under policy π
$\mathbb{E}_k^\pi$	Taken expectation respect to estimated parameters $\boldsymbol{\mu}_k$ and P_k under policy π

Table 2: Summary of notations

B Constants in Proposition 1

We consider the action-reward pair (R_t, I_t) as our observation of the underlying state M_t. We encode the pair (r,i) into a variable $s \in \{1,2,...,2I\}$ through a suitable one-to-one mapping. We rewrite our observable random vector (R_t, I_t) as a random variable S_t. Hence we can define the following matrix $A_1, A_2, A_3 \in \mathbb{R}^{2I\times M}$, where

$$\begin{aligned} A_1(s,m) &= \mathbb{P}(S_{t-1}=s|M_t=m), \\ A_2(s,m) &= \mathbb{P}(S_t=s|M_t=m), \\ A_3(s,m) &= \mathbb{P}(S_{t+1}=s|M_t=m), \end{aligned}$$

for $s \in \{1,2,...,2I\}, k \in \mathcal{M} = \{1,2,...,M\}$. It follows from Lemma 5, Lemma 8 and Theorem 3 in [8] that the spectral estimators $\hat{\boldsymbol{\mu}}, \hat{P}$ have the following guarantee: pick any $\delta \in (0,1)$, when the number of samples n satisfies

$$n \geq N_0 := \left(\frac{G\frac{2\sqrt{2}+1}{1-\theta}}{\omega_{\min}\sigma^2}\right)^2 \log\left(\frac{2(S^2+S)}{\delta}\right) \max\left\{\frac{16\times M^{1/3}}{C_0^{2/3}\omega_{\min}^{1/3}}, \frac{2\sqrt{2}M}{C_0^2\omega_{\min}\sigma^2}, 1\right\},$$

then with probability $1-\delta$ we have

$$||(\boldsymbol{\mu})^m - (\hat{\boldsymbol{\mu}})^m||_2 \leq C_1\sqrt{\frac{\log(6\frac{S^2+S}{\delta})}{n}},$$

$$||P-\hat{P}||_2 \leq C_2\sqrt{\frac{\log(6\frac{S^2+S}{\delta})}{n}},$$

for $m \in \mathcal{M}$ up to permutation, with

$$C_1 = \frac{21}{\sigma_{1,-1}} I C_3,$$
$$C_2 = \frac{4}{\sigma_{\min}(A_2)} \left(\sqrt{M} + M \frac{21}{\sigma_{1,-1}} \right) C_3,$$
$$C_3 = 2G \frac{2\sqrt{2}+1}{(1-\theta)\omega_{\min}^{0.5}} \left(1 + \frac{8\sqrt{2}}{\omega_{\min}^2 \sigma^3} + \frac{256}{\omega_{\min}^2 \sigma^2} \right),$$

where $S = 2I$, C_0 is a numerical constant (see Theorem 16 in [8]), $\sigma_{1,-1}$ is the smallest nonzero singular value of the covariance matrix $\mathbb{E}[y_{t+1} \otimes y_{t-1}]$, and $\sigma = \min\{\sigma_{\min}(A_1), \sigma_{\min}(A_2), \sigma_{\min}(A_3)\}$, where $\sigma_{\min}(A_i)$ represents the smallest nonzero singular value of the matrix A_i, for $i = 1, 2, 3$. In addition, $\omega = (\omega(m))$ represents the stationary distribution of the underlying Markov chain $\{M_t\}$, and $\omega_{\min} = \min_m \omega(m) \geq \epsilon = \min_{i,j} P_{ij}$. Finally, G and θ are the mixing rate parameters such that

$$\sup_{m_1} ||f_{1\to t}(\cdot|m_1) - \omega||_{TV} \leq G\theta^{t-1},$$

where $f_{1\to t}(\cdot|m_1)$ denotes the probability distribution vector of the underlying state at time t, starting from the initial state m_1. Under Assumption 3, one can take $G = 2$ and have the (crude) bound $\theta \leq 1 - \epsilon$, see e.g. Theorems 2.7.2 and 2.7.4 in [26].

C Proof of Proposition 2

Proof. We first introduce a few notations. Let $V_\beta(b)$ be the (optimal) value function of the infinite-horizon discounted version of the POMDP (or belief MDP) with discount factor $\beta \in (0, 1)$ and initial belief state b. Define $v_\beta(b) := V_\beta(b) - V_\beta(s)$ for a fixed belief state s, where v_β is the bias function for the *discounted* problem. We also introduce ℓ_1 distance to the belief space $\mathcal{B}$: $\rho_b(b, b') := \|b - b'\|_1$. For any function $f : \mathcal{B} \mapsto \mathbb{R}$, define the Lipschitz module of a function f by

$$l_{\rho_\mathcal{B}}(f) := \sup_{b \neq b'} \frac{|f(b) - f(b')|}{\rho_b(b, b')}.$$

The main idea of the proof is as follows. To bound $\text{span}(v) := \max_{b\in\mathcal{B}} v(b) - \min_{b\in\mathcal{B}} v(b)$ where v is the bias function for our *undiscounted* problem, it suffices to bound the Lipschitz module of v due to the fact that $\sup_{b\neq b'} ||b - b'||_1 = 2$. Under our assumptions, it can be shown that the bias function v for the undiscounted problem satisfies the relation

$$v(b) = \lim_{\beta\to 1-} v_\beta(b), \quad \text{for } b \in \mathcal{B}. \tag{5}$$

Then applying Lemma 3.2(a) [23] yields

$$l_{\rho_\mathcal{B}}(v) \leq \lim_{\beta\to 1-} l_{\rho_\mathcal{B}}(v_\beta) = \lim_{\beta\to 1-} l_{\rho_\mathcal{B}}(V_\beta). \tag{6}$$

So it suffices to bound $\lim_{\beta\to 1-} l_{\rho_\mathcal{B}}(V_\beta)$. The bound for $l_{\rho_\mathcal{B}}(V_\beta)$ in turn implies that $(v_\beta)_\beta$ is a uniformly bounded equicontinuous family of functions, and hence validates (5) by Theorem 2 in [40].

We next proceed to bound $l_{\rho_\mathcal{B}}(V_\beta)$ and we will show that for any $\beta \in (0, 1)$ we have $l_{\rho_\mathcal{B}}(V_\beta) \leq \frac{\eta}{1-\gamma}$ for some constants $\eta > 0, \gamma \in (0, 1)$ that are both independent of β. To this end, we consider the finite horizon discounted belief MDP, and let $V_{n,\beta}$ be the optimal value function for the discounted problem with horizon length n and discount factor β. Since the reward is bounded, it is readily seen that $\lim_{n\to\infty} V_{n,\beta} = V_\beta$. Then Lemma 2.1(e) [23] suggests that

$$l_{\rho_b}(V_\beta) \leq \liminf_{n\to\infty} l_{\rho_b}(V_{n,\beta}). \tag{7}$$

Thus, to bound $l_{\rho_b}(V_\beta)$, it suffices to bound the Lipschitz module $l_{\rho_b}(V_{n,\beta})$. The strategy is to apply the results including Lemmas 3.2 and 3.4 in [23], but it requires a new analysis to verify the conditions there.

To proceed, standard dynamic programming theory states that $V_{n,\beta}(b) = J_1(b)$, and $J_1(b)$ can be computed by the backward recursion:

$$\begin{aligned} J_n(b_n) &= \bar{c}(b_n, i_n), \\ J_t(b_t) &= \max_{i_t \in \mathcal{I}} \left\{ \bar{c}(b_t, i_t) + \beta \int_{\mathcal{B}} J_{t+1}(b_{t+1}) \bar{T}(db_{t+1}|b_t, i_t) \right\}, \quad 1 \le t < n, \end{aligned} \tag{8}$$

where $\bar{T}$ is the (action-dependent) one-step transition law of the belief state, and $J_t(b_t)$ are finite for each t. More generally, for a given sequence of actions $i_{1:n}$, the n-step transition kernel for the belief state is define by

$$\bar{T}^{(n)}(\mathbf{A}|b, i_{1:n}) \coloneqq \mathbb{P}(b_n \in \mathbf{A}|b_1 = b, i_{1:n}), \quad \mathbf{A} \subset \mathcal{B}. \tag{9}$$

To use the results in [23], we need to study the Lipschitz property of this multi-step transition kernel as we will see later. Following [23], we introduce the Lipschitz module for a transition kernel $\phi(b, db')$ on belief states. Let $K_{\rho_{\mathcal{B}}}(\nu, \theta)$ be the Kantorovich metric of two probability measure ν, θ defined on $\mathcal{B}$:

$$K_{\rho_{\mathcal{B}}}(\nu, \theta) \coloneqq \sup_f \left\{ \left| \int_{\mathcal{B}} f(b)\nu(db) - \int_{\mathcal{B}} f(b)\theta(db) \right|, f \in \mathrm{Lip}_1(\rho_{\mathcal{B}}) \right\},$$

where $\mathrm{Lip}_1(\rho_{\mathcal{B}})$ is the set of functions on $\mathcal{B}$ with Lipschitz module $l_{\rho_{\mathcal{B}}}(f) \le 1$. Then the Lipschitz module of the transition kernel $l_{\rho_{\mathcal{B}}}(\phi)$ is defined as:

$$l_{\rho_{\mathcal{B}}}(\phi) \coloneqq \sup_{b^1 \neq b^2} \frac{K_{\rho_{\mathcal{B}}}(\phi(b^1, db'), \phi(b^2, db'))}{\rho_{\mathcal{B}}(b^1, b^2)}.$$

The transition kernel ϕ is called Lipschitz continuous if $l_{\rho_{\mathcal{B}}}(\phi) < \infty$. To bound $l_{\rho_{\mathcal{B}}}(V_{n,\beta})$ and to apply the results in [23], the key technical result we need is the following lemma. We defer its proof to the end of this section. Recall that $\epsilon = \min_{i,j \in \mathcal{M}} P_{i,j} > 0$.

Lemma 1. *For $1 \le n < \infty$, the n-step belief state transition kernel $\bar{T}^{(n)}(\cdot|b, i_{1:n})$ in* (9) *is uniformly Lipschitz in $i_{1:n}$, and the Lipschitz module is bounded as follows:*

$$l_{\rho_{\mathcal{B}}}(\bar{T}^{(n)}) \le C_4 \alpha^n + C_5,$$

where $C_4 = \frac{2}{1-\alpha}$ and $C_5 = \frac{1}{2} + \frac{\alpha}{2}$ with $\alpha = 1 - \frac{\epsilon}{1-\epsilon} \in (0, 1)$. As a consequence, there exist constants $n_0 \in \mathbb{Z}^+$ and $\gamma < 1$ such that $l_{\rho_{\mathcal{B}}}(\bar{T}^{(n_0)}) < \gamma$ for any $i_{1:n}$. Here, we can take $n_0 = \lceil \log_\alpha \frac{1-C_5}{2C_4} \rceil$, and $\gamma = \frac{1}{2}(1 + C_5) = \frac{3+\alpha}{4}$.

With Lemma 1, we are now ready to bound $l_{\rho_{\mathcal{B}}}(V_{n,\beta})$. Consider $n = kn_0$ for some positive integer k. We can infer from the value iteration in (8) that

$$\begin{aligned} J_t(b_t) = \sup_{i_{t:t+n_0-1}} \Big\{ &\sum_{l=0}^{n_0-1} \beta^l \int_{\mathcal{B}} \bar{c}(b_{t+l}, i_{t+l-1}) \bar{T}^{(l)}(db_{t+l}|b_t, i_{t:t+l-1}) \\ &+ \beta^{n_0} \int_{\mathcal{B}} J_{t+n_0}(b_{t+n_0}) \bar{T}^{(n_0)}(db_{t+n_0}|b_t, i_{t:t+n_0-1}) \Big\}, \quad 1 \le t \le n - n_0 \end{aligned} \tag{10}$$

Bounding $\bar{c}$ in (10) by $r_{\max} = 1$ (the bound for rewards) and $\bar{T}^{(l)}$ by its Lipschitz module, we obtain the following inequality using Lemmas 3.2 and 3.4 in [23]:

$$l_{\rho_{\mathcal{B}}}(J_t) \le r_{\max} \cdot \sum_{l=0}^{n_0-1} \beta^t l_{\rho_{\mathcal{B}}}^{\mathcal{I}^l}(\bar{T}^{(l)}) + \beta^{n_0} \cdot l_{\rho_{\mathcal{B}}}^{\mathcal{I}^{n_0}}(\bar{T}^{(n_0)}) \cdot l_{\rho_{\mathcal{B}}}(J_{t+n_0}),$$

where $l_{\rho_{\mathcal{B}}}^{\mathcal{I}^l}(\bar{T}^{(l)})$ is the supremum of the Lipschitz module $l_{\rho_{\mathcal{B}}}(\bar{T}^{(l)})$ over actions:

$$l_{\rho_{\mathcal{B}}}^{\mathcal{I}^l}(\bar{T}^{(l)}) \coloneqq \sup_{i_{t:t+l-1}} \sup_{b_t \neq b_t'} \frac{K_{\rho_{\mathcal{B}}}(\bar{T}^{(l)}(db_{t+l}|b_t, i_{t:t+l-1}), \bar{T}^{(l)}(db_{t+l}|b_t', i_{t:t+l-1}))}{\rho_{\mathcal{B}}(b_t, b_t')}, \quad 0 \le l \le n_0.$$

Note that the value function in the last period $J_n(b_n) = \bar{c}(b_n, i_n)$ is uniformly Lipschitz in i_n with Lipschitz module $r_{\max} = 1$. Applying the last inequality iteratively for $n_i = 1 + in_0$ with $0 \leq i < k$ and by Lemma 1, we have

$$\begin{aligned} l_{\rho_{\mathcal{B}}}(J_{n_i}) &\leq \sum_{t=0}^{n_0-1} \beta^t l_{\rho_{\mathcal{B}}}^{\mathcal{I}^t}(\bar{T}^{(t)}) + \beta^{n_0} \cdot \gamma \cdot l_{\rho_{\mathcal{B}}}(J_{n_{i+1}}) \\ &\leq \sum_{t=0}^{n_0-1} [C_4 \alpha^t + C_5] + \beta^{n_0} \cdot \gamma \cdot l_{\rho_{\mathcal{B}}}(J_{n_{i+1}}) \\ &\leq \eta + \beta^{n_0} \gamma \cdot l_{\rho_{\mathcal{B}}}(J_{n_{i+1}}), \end{aligned}$$

where

$$\eta = \frac{C_4}{1-\alpha} + C_5 n_0, \tag{11}$$

and C_4, C_5, n_0, α are given in Lemma 1. Iterating over i and using $l_{\rho_{\mathcal{B}}}(J_n) = l_{\rho_{\mathcal{B}}}(J_{kn_0}) = r_{\max}$, we obtain

$$l_{\rho_{\mathcal{B}}}(J_0) \leq \eta \cdot \frac{1 - (\beta^{n_0}\gamma)^k}{1 - \beta^{n_0}\gamma} + (\beta^{n_0}\gamma)^k \cdot r_{\max}.$$

Recall that for $n = kn_0$, $V_{n,\beta}(b) = V_{kn_0,\beta}(b) = J_0(b)$. Since $\beta < 1$ and $\gamma < 1$, we then get

$$\liminf_{k \to \infty} l_{\rho_{\mathcal{B}}}(V_{kn_0,\beta}) \leq \frac{\eta}{1-\gamma}.$$

Together with (6) and (7), we can deduce that for a belief MDP satisfying $\min_{i,j \in \mathcal{M}} P_{ij} = \epsilon > 0$, the span of the bias function is upper bounded by

$$span(v) \leq D(\epsilon) := \frac{2\eta(\epsilon)}{1 - \gamma(\epsilon)},$$

where with slight abuse of notations we use $\eta(\epsilon)$ (see (11)) and $\gamma(\epsilon)$ (see Lemma 1) to emphasize their dependency on ϵ. The proof is completed by simplifying the expression of $D(\epsilon)$. □

C.1 Proof of Lemma 1

Proof. Rewriting the Kantorovich metric, we have:

$$\begin{aligned} &K\{\bar{T}^{(n)}(db'|b^1, i_{1:n}), \bar{T}^{(n)}(db'|b^2, i_{1:n}\} \\ &= \sup_f \left\{ \left| \int f(b')\bar{T}^{(n)}(db'|b^1, i_{1:n}) - \int f(b')\bar{T}^{(n)}(db'|b^2, i_{1:n}) \right|, f \in \text{Lip}_1 \right\} \\ &= \sup_f \left\{ \left| \int f(b')\bar{T}^{(n)}(db'|b^1, i_{1:n}) - \int f(b')\bar{T}^{(n)}(db'|b^2, i_{1:n}) \right|, f \in \text{Lip}_1, ||f||_\infty \leq 1 \right\}. \end{aligned}$$

The last equality follows from the following argument. Note that the span of a function f with Lipschitz module 1 is bounded by $\text{Diam}(\mathcal{B})$ where $\text{Diam}(\mathcal{B}) := \sup_{b^1 \neq b^2} ||b^1 - b^2||_1 = 2$. So for any $f \in \text{Lip}_1$ we can find a constant c that $||f + c||_\infty \leq \text{Diam}(\mathcal{B})/2$. Moreover, let $\phi(f) = |\int f(b')\bar{T}^{(n)}(db'|b^1, i_{1:n}) - \int f(b')\bar{T}^{(n)}(db'|b^2, i_{1:n})|$, we know $\phi(f) = \phi(f + c)$ for any constant c. Without loss of generality, we can constrain $||f||_\infty \leq \text{Diam}(\mathcal{B})/2 \leq 1$.

We introduce a few notations to facilitate the presentation. We define the n-step reward kernel $\bar{Q}^{(n)}$, where $\bar{Q}^{(n)}(\prod_{t=1}^n dr_t|b, i_{1:n})$ is a probability measure on $\mathcal{R}^n$:

$$\bar{Q}^{(n)}(A_1 \times ... \times A_n|b, i_{1:n}) = \mathbb{P}((r_1, \ldots, r_n) \in A_1 \times ... \times A_n|b, i_{1:n}).$$

Given the initial belief b, we can define the n-step forward kernel $H^{(n)}$ as follows where b_{n+1} is the belief at time $n+1$:

$$b_{n+1} = H^{(n)}(b, i_{1:n}, r_{1:n}).$$

Then it is easy to see that the belief transition kernel $\bar{T}^{(n)}$ defined in (9) satisfies

$$\bar{T}^{(n)}(\mathbf{A}|b, i_{1:n}) = \int_{\mathcal{R}^n} \mathbb{1}_{\{H^{(n)}(b, i_{1:n}, r_{1:n}) \in \mathbf{A}\}} \bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t | b, i_{1:n}).$$

Then we can obtain:

$$\begin{aligned}
&\left| \int_{\mathcal{R}^n} f(b') \bar{T}^{(n)}(db'|b^1, i_{1:n}) - \int_{\mathcal{R}^n} f(b') \bar{T}^{(n)}(db'|b^2, i_{1:n}) \right| \\
&= \left| \int_{\mathcal{R}^n} f(H^{(n)}(b^1, i_{1:n}, r_{1:n})) \bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t | b^1, i_{1:n}) - \int_{\mathcal{R}^n} f(H^{(n)}(b^2, i_{1:n}, r_{1:n})) \bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t | b^2, i_{1:n}) \right| \\
&\leq \left| \int_{\mathcal{R}^n} f(H^{(n)}(b^1, i_{1:n}, r_{1:n})) \left(\bar{Q}^{(n)}(\prod_{i=0}^{n-1} dr_t | b^1, i_{1:n}) - \bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t | b^2, i_{1:n}) \right) \right| \\
&\quad + \left| \int_{\mathcal{R}^n} \left(f(H^{(n)}(b^1, i_{1:n}, r_{1:n})) - f(H^{(n)}(b^2, i_{1:n}, r_{1:n})) \right) \bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t | b^2, i_{1:n}) \right|. \qquad (12)
\end{aligned}$$

We first bound the second term in (12). We can infer from Theorem 3.7.1 in [26] and its proof that the impact of initial belief decays exponentially fast:

$$|H^{(n)}(b^1, i_{1:n}, r_{1:n}) - H^{(n)}(b^2, i_{1:n}, r_{1:n})| \leq C_4 \alpha^n ||b^1 - b^2||_1,$$

where constant $C_4 = \frac{2(1-\epsilon)}{\epsilon}$ and $\alpha = \frac{1-2\epsilon}{1-\epsilon} < 1$. So the second term of (12) can be bounded by

$$\begin{aligned}
&\left| \int_{\mathcal{R}^n} \left(f(H^{(n)}(b^1, i_{1:n}, r_{1:n})) - f(H^{(n)}(b^2, i_{1:n}, r_{1:n})) \right) \bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t | b^2, i_{1:n}) \right| \\
&\leq \left| \int_{\mathcal{R}^n} \left| H^{(n)}(b^1, i_{1:n}, r_{1:n}) - H^{(n)}(b^2, i_{1:n}, r_{1:n}) \right| \bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t | b^2, i_{1:n}) \right| \\
&\leq C_4 \alpha^n ||b^1 - b^2||_1, \qquad (13)
\end{aligned}$$

where the first inequality follows from $f \in \text{Lip}_1$.

It remains to bound the first term in (12). Recall that b defines the initial probability distribution M_1. The n steps observation kernel is

$$\bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t | b, i_{1:n}) = \sum_{m \in \mathcal{M}} \mathbb{P}(M_1 = m) \mathbb{P}(\prod_{t=1}^{n} dr_t | M_1 = m, i_{1:n}) = \sum_{m \in \mathcal{M}} b(m) \mathbb{P}(\prod_{t=1}^{n} dr_t | M_1 = m, i_{1:n}).$$

Define a vector $g \in \mathbb{R}^M$ as:

$$g(m) \coloneqq \int_{\mathcal{R}^n} f(H^{(n)}(b^1, i_{1:n}, r_{1:n})) \mathbb{P}(\prod_{t=1}^{n} dr_t | M_1 = m, i_{1:n}).$$

We can rewrite the first term of (12):

$$
\begin{aligned}
&\left|\int_{\mathcal{R}^n} f(H^{(n)}(b^1, i_{1:n}, r_{1:n}))\left(\bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t|b^1, i_{1:n}) - \bar{Q}^{(n)}(\prod_{t=1}^{n} dr_t|b^2, i_{1:n})\right)\right| \\
&= \left|\sum_{m=1}^{M}(b^1(m) - b^2(m))\int_{\mathcal{R}^n} f(H^{(n)}(b^1, i_{1:n}, r_{1:n}))\mathbb{P}(\prod_{t=1}^{n} dr_t|M_1 = m, i_{1:n})\right| \\
&= \left|\sum_{m=1}^{M}\left(b^1(m) - b^2(m)\right)g(m)\right| \\
&= \left|\sum_{m=1}^{M}\left(b^1(m) - b^2(m)\right)\left(g(m) - \frac{\max_m g(m) + \min_m g(m)}{2}\right)\right| \\
&\leq \left\|b^1 - b^2\right\|_1 \cdot \left\|g(m) - \frac{\max_m g(m) + \min_m g(m)}{2}\right\|_\infty \\
&= \left\|b^1 - b^2\right\|_1 \frac{1}{2}\left(\max_m g(m) - \min_m g(m)\right), \qquad (14)
\end{aligned}
$$

where the next to last equality follows from $\sum_{m=1}^{M}\left(b^1(m) - b^2(m)\right) = 0$.

Next we bound $\max_m g(m) - \min_m g(m)$. From the equation above, it is clear that the quantity $\frac{1}{2}\left(\max_m g(m) - \min_m g(m)\right) \leq 1$, because $||f||_\infty \leq 1$. However to prove Lemma 1, we need a sharper bound so that we can find a constant $C_5 < 1$ (that is independent of b^1, n and $i_{1:n}$) with

$$
\frac{1}{2}\left(\max_m g(m) - \min_m g(m)\right) \leq C_5 < 1. \qquad (15)
$$

Suppose (15) holds. Then on combining (12), (13) and (14), we obtain

$$
\left|\int_{\mathcal{B}} f(b')\bar{T}^{(n)}(db'|b^1, i_{1:n}) - \int_{\mathcal{B}} f(b')\bar{T}^{(n)}(db'|b^2, i_{1:n})\right| \leq C_4\alpha^n||b^1 - b^2||_1 + C_5||b^1 - b^2||_1.
$$

It then follows that the Kantorovich metric is bounded by

$$
K\left(\bar{T}^{(n)}(db'|b^1, i_{1:n}), \bar{T}^{(n)}(db'|b^2, i_{1:n})\right) \leq C_4\alpha^n||b^1 - b^2||_1 + C_5||b^1 - b^2||_1,
$$

where $C_4 = \frac{2(1-\epsilon)}{\epsilon}, \alpha = 1 - \frac{\epsilon}{1-\epsilon}$, and $\epsilon = \min_{m,m' \in \mathcal{M}} P_{m,m'} > 0$. So $\bar{T}^{(n)}$ is Lipschitz uniformly in actions, and its Lipschitz module can be bounded as follows:

$$
l_{\rho_{\mathcal{B}}}^{\mathcal{I}^n}(\bar{T}^{(n)}) := \sup_{i_{1:n}} \sup_{b^1 \neq b^2} \frac{K\left(\bar{T}^{(n)}(db'|b^1, i_{1:n}), \bar{T}^{(n)}(db'|b^2, i_{1:n})\right)}{\rho_{\mathcal{B}}(b^1, b^2)} \leq C_4\alpha^n + C_5.
$$

If we choose $n = n_0 := \lceil \log_\alpha \frac{1-C_5}{2C_4} \rceil$, so that $C_4\alpha^{n_0} + C_5 < \frac{1}{2}(1 + C_5) := \gamma < 1$, then we obtain the desired result $l_{\rho_{\mathcal{B}}}^{\mathcal{I}^{n_0}}(\bar{T}^{(n_0)}) < \gamma$.

It remains to prove (15). Since the set $\mathcal{M} = \{1, \ldots, M\}$ is finite, we pick $m^* \in \operatorname{argmin}_{m \in \mathcal{M}} g(m), \hat{m} \in \operatorname{argmax}_{m \in \mathcal{M}} g(m)$. We have

$$
\begin{aligned}
&\frac{1}{2}\left(\max_m g(m) - \min_m g(m)\right) \qquad (16) \\
&= \frac{1}{2}\sum_{r_{1:n} \in \mathcal{R}^n} f(H^{(n)}(b^1, i_{1:n}, r_{1:n}))\left(\mathbb{P}(r_{1:n}|M_1 = \hat{m}, i_{1:n}) - \mathbb{P}(r_{1:n}|M_1 = m^*, i_{1:n})\right) \\
&\leq \frac{1}{2}\sum_{r_{1:n} \in \mathcal{R}^n} \left|\mathbb{P}(r_{1:n}|M_1 = \hat{m}, i_{1:n}) - \mathbb{P}(r_{1:n}|M_1 = m^*, i_{1:n})\right|,
\end{aligned}
$$

where the inequality follows from Hölder's inequality with $||f||_\infty \leq 1$. We can compute

$$\begin{aligned}
&\mathbb{P}(r_{1:n}|M_1 = m_1, i_{1:n}) \\
&= \sum_{m_{2:n}\in\mathcal{M}^{n-1}} \mathbb{P}(r_{1:n}|M_1 = m_1, i_{1:n}, M_{2:n} = m_{2:n}) \cdot \mathbb{P}(M_{2:n} = m_{2:n}|M_1 = m_1, i_{1:n}) \\
&= \sum_{m_{2:n}\in\mathcal{M}^{n-1}} \mathbb{P}(r_{1:n}|i_{1:n}, m_{1:n}) \cdot \mathbb{P}(m_{2:n}|M_1 = m_1, i_{1:n}) \\
&= \sum_{m_{2:n}\in\mathcal{M}^{n-1}} \left(\prod_{t=1}^{n} \mathbb{P}(r_t|m_t, i_t)\right) \cdot \left(\prod_{t=1}^{n-1} \mathbb{P}(m_{t+1}|m_t, i_t)\right),
\end{aligned}$$

where the last equality holds due to the conditional independence. We can then infer that for any $\{r_{1:n}\}, \{i_{1:n}\}$,

$$\begin{aligned}
&\mathbb{P}(r_{1:n}|M_1 = m^*, i_{1:n}) \\
&= \sum_{m_{2:n}\in\mathcal{M}^{n-1}} \left(\prod_{t=2}^{n} \mathbb{P}(r_t|m_t, i_t)\right) \cdot \left(\prod_{t=1}^{n-1} \mathbb{P}(m_{t+1}|m_t, i_t)\right) \cdot P(m^*, m_2) \\
&\geq \sum_{m_{2:n}\in\mathcal{M}^{n-1}} \left(\prod_{t=1}^{n} \mathbb{P}(r_t|m_t, i_t)\right) \cdot \left(\prod_{t=2}^{n-1} \mathbb{P}(m_{t+1}|m_t, i_t)\right) \cdot P(\hat{m}, m_1) \cdot \frac{\epsilon}{1-\epsilon} \\
&= \mathbb{P}(r_{1:n}|M_1 = \hat{m}, i_{1:n}) \cdot \frac{\epsilon}{1-\epsilon}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&|\mathbb{P}(r_{1:n}|M_1 = \hat{m}, i_{1:n}) - \mathbb{P}(r_{1:n}|M_1 = m^*, i_{1:n})| \\
&\leq \max\left\{\left(1 - \frac{\epsilon}{1-\epsilon}\right)\mathbb{P}(r_{1:n}|M_1 = \hat{m}, i_{1:n}), \mathbb{P}(r_{1:n}|M_1 = m^*, i_{1:n})\right\} \\
&\leq \left(1 - \frac{\epsilon}{1-\epsilon}\right)\mathbb{P}(r_{1:n}|M_1 = \hat{m}, i_{1:n}) + \mathbb{P}(r_{1:n}|M_1 = m^*, i_{1:n}).
\end{aligned}$$

Then we can obtain from (16) that

$$\begin{aligned}
&\frac{1}{2}\left(\max_m g(m) - \min_m g(m)\right) \\
&\leq \frac{1}{2}\sum_{r_{1:n}\in\mathcal{R}^n}\left[\left(1 - \frac{\epsilon}{1-\epsilon}\right)\mathbb{P}(r_{1:n}|M_0 = \hat{m}, i_{1:n}) + \mathbb{P}(r_{1:n}|M_0 = m^*, i_{1:n})\right] \\
&= \alpha/2 + 1/2 := C_5 < 1,
\end{aligned}$$

where $\alpha = 1 - \frac{\epsilon}{1-\epsilon} \in (0, 1)$. The proof is complete. □

D Proof of Proposition 3

The proof of Proposition 3 largely follows the proof of Proposition 3 in [18], with minor changes to take into account the action sequence, so we omit details.

E Proof of Theorem 1

In this section we provide a complete proof of Theorem 1. We follow our prior simplication that random reward follows the Bernoulli distribution. Here we want to clarify two notations in advance. $\mathbb{E}^\pi$ means the expectation is taken respect to true mean reward matrix $\boldsymbol{\mu}$ and transition probabilities P, and $\mathbb{E}_k^\pi$ denotes that the underlying parameters are estimators $\boldsymbol{\mu}_k$ and P_k.

Recalling the definition of regret in (1), it can be rewritten as:

$$\mathcal{R}_T = \sum_{t=1}^{T}(\rho^* - R_t) = \sum_{t=1}^{T}(\rho^* - \mathbb{E}^\pi[R_t|\mathcal{F}_{t-1}]) + \sum_{t=1}^{T}(\mathbb{E}^\pi[R_t|\mathcal{F}_{t-1}] - R_t). \tag{17}$$

We first bound the second term of (17), i.e. the total bias between the conditional expectation of reward and the realization. Define a stochastic process $\{X_n, n = 0, \cdots, T\}$ as:

$$X_0 = 0, \quad X_t = \sum_{l=1}^{t} (\mathbb{E}^{\pi}[R_l|\mathcal{F}_{l-1}] - R_l),$$

then the second term in (17) is X_T. It is easy to see that X_t is a martingale. Moreover, due to the Bernoulli distribution of R_t,

$$|X_{t+1} - X_t| = |\mathbb{E}^{\pi}[R_{t+1}|\mathcal{F}_t] - R_{t+1}| \leq 1.$$

Applying the Azuma-Hoeffding inequality [9], we have

$$\mathbb{P}\left(\sum_{t=1}^{T} (\mathbb{E}^{\pi}[R_t|\mathcal{F}_{t-1}] - R_t) \geq \sqrt{2T \ln \frac{1}{\delta}}\right) \leq \delta. \tag{18}$$

Next we bound the first term of (17). Recall that the definition of belief state under the optimistic and true parameters: $b_t^k(m) = \mathbb{P}_{\boldsymbol{\mu}_k, P_k}(M_t = m|\mathcal{F}_{t-1})$ and $b_t(m) = \mathbb{P}(M_t = m|\mathcal{F}_{t-1})$, and the definition of reward functions with respect to the true belief state $\bar{c}(b_t, i) = \sum_{m=1}^{M} \mu_{m,i} b_t(m)$. We can also define the reward functions with respect to the optimistic belief state b_t^k as:

$$\bar{c}_k(b_t^k, i) = \sum_{m=1}^{M} (\boldsymbol{\mu}_k)_{m,i} b_t^k(m).$$

Because I_t is also adapted to $\mathcal{F}_{t-1}$, we have

$$\begin{aligned} &\mathbb{E}^{\pi}[\mu(M_t, I_t)|\mathcal{F}_{t-1}] = \bar{c}(b_t, I_t) = \langle (\boldsymbol{\mu})_{I_t}, b_t \rangle, \\ &\mathbb{E}_k^{\pi}[\mu_k(M_t, I_t)|\mathcal{F}_{t-1}] = \bar{c}_k(b_t^k, I_t) = \langle (\boldsymbol{\mu}_k)_{I_t}, b_t^k \rangle, \end{aligned} \tag{19}$$

where $\mu(M_t, I_t), \mu_k(M_t, I_t)$ are the M_t-th row I_t-th column element of matrix μ, μ_k respectively, and $(\boldsymbol{\mu}_k)_{I_t}$ and $(\boldsymbol{\mu})_{I_t}$ are the I_t-th column vector of the reward matrix $\boldsymbol{\mu}_k$ and $\boldsymbol{\mu}$, respectively.

Then we can rewrite the first term of (17):

$$\sum_{t=1}^{T} (\rho^* - \mathbb{E}^{\pi}[R_t|\mathcal{F}_{t-1}]) = \sum_{t=1}^{T} (\rho^* - \mathbb{E}^{\pi}[\mu(M_t, I_t)|\mathcal{F}_{t-1}]) = \sum_{t=1}^{T} (\rho^* - \bar{c}(b_t, I_t)), \tag{20}$$

where the first equation is due to the tower property and the fact that R_t and $\mathcal{F}_{t-1}$ are conditionally independent given M_t and I_t.

Let K be the number of total episodes. For each episode $k = 1, 2, \cdots, K$, let H_k, E_k be the exploration and exploitation phases, respectively. Then we can split equation (20) to the summation of the bias in these two phases as:

$$\sum_{k=1}^{K} \sum_{t \in H_k} (\rho^* - \bar{c}(b_t, I_t)) + \sum_{k=1}^{K} \sum_{t \in E_k} (\rho^* - \bar{c}(b_t, I_t)). \tag{21}$$

Moreover, we remark here that the length of the last exploitation phase $|E_K| = \min\{\tau_2\sqrt{K}, \max\{T - (K\tau_1 + \sum_{k=1}^{K-1} \tau_2\sqrt{k}), 0\}\}$, as it may end at period T.

Step 1: Bounding the regret in exploration phases

The first term of (21) can be simply upper bounded by:

$$\sum_{k=1}^{K} \sum_{t \in H_k} (\rho^* - \bar{c}(b_t, I_t)) \leq \sum_{k=1}^{K} \sum_{t \in H_k} \rho^* = K\tau_1\rho^*. \tag{22}$$

Step 2: Bounding the regret in exploitation phases

We bound it by separating into "success" and "failure" events below. Recall that in episode k, we define the set of plausible POMDPs $\mathbb{G}_k(\delta_k)$, which is defined in terms of confidence regions $\mathcal{C}_k(\delta_k)$

around the estimated mean reward matrix $\boldsymbol{\mu}_k$ and the transition probabilities P_k. Then choose an optimistic POMDP $\widetilde{\mathbb{G}}_k \in \mathbb{G}_k(\delta_k)$ that has the optimal average reward among the plausible POMDPs and denote its corresponding reward matrix, value function, and the optimal average reward by $\boldsymbol{\mu}_k, v_k$ and ρ^k, respectively. Thus, we say a "success" event if and only if the set of plausible POMDPs $\mathbb{G}_k(\delta_k)$ contains the true POMDP $\mathbb{G}$. In the following proof, we omit the dependence on δ_k from $\mathbb{G}_k(\delta_k)$ for simplicity. From Algorithm 2, the confidence level of μ_k in episode k is $1-\delta_k$, we can obtain:

$$\mathbb{P}(\mathbb{G} \notin \mathbb{G}_k, \text{for some } k) \leq \sum_{k=1}^{K} \delta_k = \sum_{k=1}^{K} \frac{\delta}{k^3} \leq \frac{3}{2}\delta.$$

Thus, with probability at least $1-\frac{3}{2}\delta$, "success" events happen. It means $\rho^* \leq \rho^k$ for any k because ρ^k is the optimal average reward of the optimistic POMDP $\widetilde{\mathbb{G}}_k$ from the set $\mathbb{G}_k$. Then we can bound the regret of "success" events in exploitation phases as follows.

$$\begin{aligned}\sum_{k=1}^{K}\sum_{t\in E_k}(\rho^* - \bar{c}(b_t, I_t)) &\leq \sum_{k=1}^{K}\sum_{t\in E_k}(\rho^k - \bar{c}(b_t, I_t)) \\ &= \sum_{k=1}^{K}\sum_{t\in E_k}(\rho^k - \bar{c}_k(b_t^k, I_t)) + (\bar{c}_k(b_t^k, I_t) - \bar{c}(b_t, I_t)). \end{aligned} \tag{23}$$

To bound the first term of formula (23), we use the Bellman optimality equation for the optimistic belief MDP $\widetilde{\mathbb{G}}_k$ on the continuous belief state space $\mathcal{B}$:

$$\rho^k + v_k(b_t^k) = \bar{c}_k(b_t^k, I_t) + \int_{b_{t+1}^k \in \mathcal{B}} v_k(b_{t+1}^k)\bar{T}_k(db_{t+1}^k|b_t^k, I_t) = \bar{c}_k(b_t^k, I_t) + \langle \bar{T}_k(\cdot|b_t^k, I_t), v_k(\cdot)\rangle,$$

where $\bar{T}_k(\cdot|b_t^k, I_t) = \mathbb{P}_{\boldsymbol{\mu}_k, P_k}(b_{t+1} \in \cdot|b_t^k, I_t)$ means transition probability of the belief state conditional on pulled arm under estimated reward matrix $\boldsymbol{\mu}_k$ and transition matrix of underlying Markov chain P_k at time t.

Moreover, we note that if value function v_k satisfies the Bellman equation (2), then so is $v_k + c\mathbf{1}$. Thus, without loss of generality, we assume that v_k needs to satisfy $||v_k||_\infty \leq \text{span}(v_k)/2$. Then from Proposition 2, suppose the uniform bound of $\text{span}(v_k)$ is D, then we have:

$$||v_k||_\infty \leq \frac{1}{2}\text{span}(v_k) \leq \frac{D}{2}. \tag{24}$$

Thus the first term of (23) can be bounded by

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{t\in E_k}(\rho^k - \bar{c}_k(b_t^k, I_t)) \\ &= \sum_{k=1}^{K}\sum_{t\in E_k}(-v_k(b_t^k) + \langle \bar{T}_k(\cdot|b_t^k, I_t), v_k(\cdot)\rangle) \\ &= \sum_{k=1}^{K}\sum_{t\in E_k}(-v_k(b_t^k) + \langle \bar{T}(\cdot|b_t^k, I_t), v_k(\cdot)\rangle) + \langle \bar{T}_k(\cdot|b_t^k, I_t) - \bar{T}(\cdot|b_t^k, I_t), v_k(\cdot)\rangle, \end{aligned} \tag{25}$$

where recall that $\bar{T}_k(\cdot|b_t^k, I_t)$ and $\bar{T}(\cdot|b_t^k, I_t)$ in the second equality are the belief state transition probabilities under estimated and true parameters, respectively. And the last inequality is applying the Hölder's inequality.

For the first term of (25), we have

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{t\in E_k}(-v_k(b_t^k)+\langle\bar{T}(\cdot|b_t^k,I_t),v_k(\cdot)\rangle)\\
&=\sum_{k=1}^{K}\sum_{t\in E_k}(-v_k(b_t^k)+v_k(b_{t+1}^k))+(-v_k(b_{t+1}^k)+\langle\bar{T}(\cdot|b_t^k,I_t),v_k(\cdot)\rangle)\\
&=\sum_{k=1}^{K}v_k(b^k_{t_k+\tau_1+\tau_2\sqrt{k}})-v_k(b^k_{t_k+\tau_1+1})+\sum_{k=1}^{K}\sum_{t\in E_k}\mathbb{E}^{\pi}[v_k(b_{t+1}^k)|\mathcal{F}_t]-v_k(b_{t+1}^k).
\end{aligned}$$

where the first term in the last equality is due to the telescoping from $t_k+\tau_1+1$ to $t_k+\tau_1+\tau_2\sqrt{k}$, the start and end of the exploitation phase in episode k. The second term in the last equality is because:

$$\langle\bar{T}(\cdot|b_t^k,I_t),v_k(\cdot)\rangle=\int_{b_{t+1}^k\in\mathcal{B}}v_k(b_{t+1}^k)\bar{T}(db_{t+1}^k|b_t^k,I_t)=\mathbb{E}^{\pi}[v_k(b_{t+1}^k)|b_t^k]=\mathbb{E}^{\pi}[v_k(b_{t+1}^k)|\mathcal{F}_t].$$

Applying Proposition 2, we have

$$v_k(b^k_{t_k+\tau_1+\tau_2\sqrt{k}})-v_k(b^k_{t_k+\tau_1+1})\le D.$$

We also need the following result, the proof of which is deferred to the end of this section.

Proposition 4. *Let K be the number of total episodes up to time T. For each episode $k=1,\cdots,K$, let E_k be the index set of the kth exploitation phase and v_k be the value function of the optimistic POMDP at the kth exploitation phase. Then with probability at most δ,*

$$\sum_{k=1}^{K}\sum_{t\in E_k}\mathbb{E}^{\pi}[v_k(b_{t+1})|\mathcal{F}_t]-v_k(b_{t+1})\ge D\sqrt{2T\ln(\frac{1}{\delta})},$$

where the expectation $\mathbb{E}^{\pi}$ is taken respect to the true parameters $\boldsymbol{\mu}$ and P under policy π, and the filtration $\mathcal{F}_t$ is defined as $\mathcal{F}_t:=\sigma(\pi_1,R_1^{\pi},...,\pi_{t-1},R_{t-1}^{\pi})$.

Applying Proposition 4, with probability at least $1-\delta$, we have:

$$\sum_{k=1}^{K}\sum_{t\in E_k}\mathbb{E}^{\pi}[v_k(b_{t+1}^k)|\mathcal{F}_t]-v_k(b_{t+1}^k)\le D\sqrt{2T\ln(\frac{1}{\delta})}.$$

Thus, the first term of (25) can be upper bounded by:

$$KD+D\sqrt{2T\ln(\frac{1}{\delta})}. \tag{26}$$

For the second term of (25), we note that $\bar{T}(b_{t+1}^k|b_t^k,I_t)$ is zero except for two points where b_{t+1}^k are exactly the Bayesian updating after receiving an observation of Bernoulli reward r_t taking value 0 or 1. Thus, we have the following transition kernel:

$$\begin{aligned}
&\langle\bar{T}_k(\cdot|b_t^k,I_t)-\bar{T}(\cdot|b_t^k,I_t),v_k(\cdot)\rangle\\
&\le\left|\int_{\mathcal{B}}v_k(b')\bar{T}_k(db'|b_t^k,I_t)-\int_{\mathcal{B}}v_k(b')\bar{T}(db'|b_t^k,I_t)\right|\\
&=\left|\sum_{r_t\in\mathcal{R}}v_k\left(H_k\left(b_t^k,I_t,r_t\right)\right)\mathbb{P}_k\left(r_t|b_t^k,I_t\right)-\sum_{r_t\in\mathcal{R}}v_k\left(H\left(b_t^k,I_t,r_t\right)\right)\mathbb{P}\left(r_t|b_t^k,I_t\right)\right|\\
&\le\left|\sum_{r_t\in\mathcal{R}}v_k\left(H_k\left(b_t^k,I_t,r_t\right)\right)\cdot\left[\mathbb{P}_k\left(r_t|b_t^k,I_t\right)-\mathbb{P}\left(r_t|b_t^k,I_t\right)\right]\right|\\
&\qquad+\left|\sum_{r_t\in\mathcal{R}}\left[v_k\left(H_k\left(b_t^k,I_t,r_t\right)\right)-v_k\left(H\left(b_t^k,I_t,r_t\right)\right)\right]\cdot\mathbb{P}\left(r_t|b_t^k,I_t\right)\right|,
\end{aligned}\tag{27}$$

where we use H_k and H to denote the belief updating function under the optimistic model $(\boldsymbol{\mu}_k, P_k)$ and the true model $(\boldsymbol{\mu}, P)$, and we use $\mathbb{P}_k$ and $\mathbb{P}$ to denote the probability with respect to the optimistic model and true model respectively.

We bound the first term of (27) by

$$\begin{aligned}
&\left|\sum_{r_t\in\mathcal{R}} v_k\left(H_k\left(b_t^k, I_t, r_t\right)\right)\cdot\left[\mathbb{P}_k\left(r_t|b_t^k, I_t\right)-\mathbb{P}\left(r_t|b_t^k, I_t\right)\right]\right| \\
&\leq \left|v_k\left(H_k\left(b_t^k, I_t, r_t=1\right)\right)\cdot\left[\mathbb{P}_k\left(r_t=1|b_t^k, I_t\right)-\mathbb{P}\left(r_t=1|b_t^k, I_t\right)\right]\right| \\
&\qquad+\left|v_k\left(H_k\left(b_t^k, I_t, r_t=0\right)\right)\cdot\left[\mathbb{P}_k\left(r_t=0|b_t^k, I_t\right)-\mathbb{P}\left(r_t=0|b_t^k, I_t\right)\right]\right| \\
&= \left|v_k\left(H_k\left(b_t^k, I_t, 1\right)\right)\cdot\left[\langle(\boldsymbol{\mu}_k)_{I_t}, b_t^k\rangle-\langle(\boldsymbol{\mu})_{I_t}, b_t^k\rangle\right]\right| \\
&\qquad+\left|v_k\left(H_k\left(b_t^k, I_t, 0\right)\right)\cdot\left[1-\langle(\boldsymbol{\mu}_k)_{I_t}, b_t^k\rangle-\left(1-\langle(\boldsymbol{\mu})_{I_t}, b_t^k\rangle\right)\right]\right| \\
&\leq 2||v_k||_\infty\cdot\left|\langle(\boldsymbol{\mu}_k)_{I_t}, b_t^k\rangle-\langle(\boldsymbol{\mu})_{I_t}, b_t^k\rangle\right| \\
&\leq D\left|\langle(\boldsymbol{\mu}_k)_{I_t}, b_t^k\rangle-\langle(\boldsymbol{\mu})_{I_t}, b_t^k\rangle\right| \\
&\leq D||(\boldsymbol{\mu}_k)_{I_t}-(\boldsymbol{\mu})_{I_t}||_1\cdot||b_t^k||_\infty \\
&\leq D||(\boldsymbol{\mu}_k)_{I_t}-(\boldsymbol{\mu})_{I_t}||_1, \qquad (28)
\end{aligned}$$

where the first equality comes from $\mathbb{P}_k\left(r_t=1|b_t^k, I_t\right) = \sum_{m\in\mathcal{M}}\mathbb{P}_k(r_t = 1|m, I_t)b_t^k(m) = \langle(\boldsymbol{\mu})_{I_t}, b_t^k\rangle$ and $\mathbb{P}_k\left(r_t=0|b_t^k, I_t\right) = \sum_{m\in\mathcal{M}}\mathbb{P}_k(r_t = 0|m, I_t)b_t^k(m) = 1-\langle(\boldsymbol{\mu})_{I_t}, b_t^k\rangle$, and the third inequality is from (24), we have $||v_k||_\infty \leq \frac{D}{2}$.

We can bound the second term of (27) as follows:

$$\begin{aligned}
&\left|\sum_{r_t\in\mathcal{R}}\left[v_k\left(H_k\left(b_t^k, I_t, r_t\right)\right)-v_k\left(H\left(b_t^k, I_t, r_t\right)\right)\right]\cdot\mathbb{P}\left(r_t|b_t^k, I_t\right)\right| \\
&\leq \sum_{r_t\in\mathcal{R}}\left|v_k\left(H_k\left(b_t^k, I_t, r_t\right)\right)-v_k\left(H\left(b_t^k, I_t, r_t\right)\right)\right|\cdot\mathbb{P}\left(r_t|b_t^k, I_t\right) \\
&\leq \sum_{r_t\in\mathcal{R}}\frac{D}{2}\left|H_k\left(b_t^k, I_t, r_t\right)-H\left(b_t^k, I_t, r_t\right)\right|\cdot\mathbb{P}\left(r_t|b_t^k, I_t\right) \\
&\leq \sum_{r_t\in\mathcal{R}}\frac{D}{2}\left(L_1||\boldsymbol{\mu}-\boldsymbol{\mu}_k||_1+L_2||P-P_k||_F\right)\cdot(1-\alpha)\cdot\mathbb{P}\left(r_t|b_t^k, I_t\right) \\
&= \frac{D}{2}\left(L_1||\boldsymbol{\mu}-\boldsymbol{\mu}_k||_1+L_2||P-P_k||_F\right)\cdot(1-\alpha), \qquad (29)
\end{aligned}$$

where $\alpha = 1-\frac{\epsilon}{1-\epsilon}$. Here, the second inequality is implied from the proof of Proposition 2. The last inequality follows from a refined bound in (4) of Proposition 3. Specifically, the current upper bound in Proposition 3 is uniform in time, and it is crude for small t since the initial belief is the same with $b_1 = \hat{b}_1$. One can improve the bound by multiplying it with a factor of $1-\alpha^{t-1}$. This follows from Proposition 3 of [18] and its proof. Therefore, we can obtain the last inequality of (29), since here we consider the one-step update of the beliefs under two different sets of model parameters.

Therefore, from (28) and (29), we can obtain that the second term of equation (25):

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{t\in E_k}\langle\bar{T}_k(\cdot|b_t^k, I_t)-\bar{T}(\cdot|b_t^k, I_t), v_k(\cdot)\rangle \\
&\leq \sum_{k=1}^{K}\sum_{t\in E_k} D\left[||(\boldsymbol{\mu})_{I_t}-(\boldsymbol{\mu}_k)_{I_t}||_1+\frac{L_1(1-\alpha)}{2}||\boldsymbol{\mu}-\boldsymbol{\mu}_k||_1+\frac{L_2(1-\alpha)}{2}||P-P_k||_F\right]. \qquad (30)
\end{aligned}$$

Summing up (26) and (30), the first term of formula (23) can be bounded by

$$\sum_{k=1}^{K}\sum_{t\in E_k}(\rho^k-\bar{c}_k(b_t^k,I_t))\leq KD+D\sqrt{2T\ln(\frac{1}{\delta})}$$
$$+\sum_{k=1}^{K}\sum_{t\in E_k}D\left[||(\boldsymbol{\mu})_{I_t}-(\boldsymbol{\mu}_k)_{I_t}||_1+\frac{L_1(1-\alpha)}{2}||\boldsymbol{\mu}-\boldsymbol{\mu}_k||_1+\frac{L_2(1-\alpha)}{2}||P-P_k||_F\right]. \quad (31)$$

Next we proceed to bound the second term of (23). By (19), it can be rewritten as

$$\sum_{k=1}^{K}\sum_{t\in E_k}\bar{c}_k(b_t^k,I_t)-\bar{c}(b_t,I_t)$$
$$=\sum_{k=1}^{K}\sum_{t\in E_k}\langle(\boldsymbol{\mu}_k)_{I_t},b_t^k\rangle-\langle(\boldsymbol{\mu})_{I_t},b_t\rangle$$
$$=\sum_{k=1}^{K}\sum_{t\in E_k}\langle(\boldsymbol{\mu}_k)_{I_t},b_t^k\rangle-\langle(\boldsymbol{\mu})_{I_t},b_t^k\rangle+\langle(\boldsymbol{\mu})_{I_t},b_t^k\rangle-\langle(\boldsymbol{\mu})_{I_t},b_t\rangle,$$

then from the Hölder's inequality, we can further bounded the right hand side of above formula to

$$\sum_{k=1}^{K}\sum_{t\in E_k}||(\boldsymbol{\mu}_k)_{I_t}-(\boldsymbol{\mu})_{I_t}||_1||b_t^k||_\infty+||(\boldsymbol{\mu})_{I_t}||_\infty||b_t^k-b_t||_1$$
$$\leq\sum_{k=1}^{K}\sum_{t\in E_k}||(\boldsymbol{\mu}_k)_{I_t}-(\boldsymbol{\mu})_{I_t}||_1+||b_t^k-b_t||_1.$$

By Proposition 3, we have $||b_t^k-b_t||_1\leq L_1||\boldsymbol{\mu}-\boldsymbol{\mu}_k||_1+L_2||P-P_k||_F$. Combining the above expression with (31), we can see that with probability $1-\delta$, the regret incurred from "success" events, i.e., (23), can be bounded

$$\sum_{k=1}^{K}\sum_{t\in E_k}(\rho^*-\bar{c}(b_t,I_t))$$
$$\leq KD+D\sqrt{2T\ln(\frac{1}{\delta})}+\sum_{k=1}^{K}\sum_{t\in E_k}(D+1)||(\boldsymbol{\mu})_{I_t}-(\boldsymbol{\mu}_k)_{I_t}||_1$$
$$+\sum_{k=1}^{K}\sum_{t\in E_k}\left[\left(1+\frac{D(1-\alpha)}{2}\right)L_1||\boldsymbol{\mu}-\boldsymbol{\mu}_k||_1+\left(1+\frac{D(1-\alpha)}{2}\right)L_2||P-P_k||_F\right].$$

Let T_0 be the period that the number of samples collected in the exploration phases exceeds N_0, that is,

$$T_0:=\inf_{t\geq1}\{\sum_{n=1}^{t}\mathbb{1}_{(n\in H_k,\text{ for some }k)}\geq N_0\}.$$

If $T\geq T_0$ after episode k_0, then from Proposition 1, under the confidence level $\delta_k=\delta/k^3$, we have

$$||(\boldsymbol{\mu}_k)^m-(\boldsymbol{\mu})^m||_2\leq C_1\sqrt{\frac{\log(\frac{6(S^2+S)k^3}{\delta})}{\tau_1 k}},\quad m\in\mathcal{M},$$
$$||P_k-P||_2\leq C_2\sqrt{\frac{\log(\frac{6(S^2+S)k^3}{\delta})}{\tau_1 k}}.$$

Together with the fact that the vector norm and matrix norm satisfy

$$||(\boldsymbol{\mu}_k)_i - (\boldsymbol{\mu})_i||_1 \leq \sum_{m=1}^{M} ||(\boldsymbol{\mu}_k)^m - (\boldsymbol{\mu})^m||_1 \leq \sum_{m=1}^{M} \sqrt{M}||(\boldsymbol{\mu}_k)^m - (\boldsymbol{\mu})^m||_2, \quad i \in \mathcal{I},$$

$$||\boldsymbol{\mu}_k - \boldsymbol{\mu}||_1 = \max_i ||(\boldsymbol{\mu}_k)_i - (\boldsymbol{\mu})_i||_1 \leq \sum_{m=1}^{M} ||(\boldsymbol{\mu}_k)^m - (\boldsymbol{\mu})^m||_1 \leq \sum_{m=1}^{M} \sqrt{M}||(\boldsymbol{\mu}_k)^m - (\boldsymbol{\mu})^m||_2,$$

$$||P - P_k||_F \leq \sqrt{M}||P - P_k||_2,$$

we obtain with probability at least $1 - \frac{5}{2}\delta$, the regret in exploitation phase can be bounded by

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{t \in E_k} \rho^* - \bar{c}(b_t, I_t) \\
&\leq KD + D\sqrt{2T\ln(\frac{1}{\delta})} + \sum_{k=k_0}^{K} \tau_2\sqrt{k}\left(D + 1 + \left(1 + \frac{D(1-\alpha)}{2}\right)L_1\right) M\sqrt{M}C_1\sqrt{\frac{\log(\frac{6(S^2+S)k^3}{\delta})}{\tau_1 k}} \\
&\quad + \sum_{k=k_0}^{K} \tau_2\sqrt{k}\left(1 + \frac{D(1-\alpha)}{2}\right)L_2\sqrt{M}C_2\sqrt{\frac{\log(\frac{6(S^2+S))k^3}{\delta})}{\tau_1 k}} + T_0\rho^* \\
&\leq KD + D\sqrt{2T\ln(\frac{1}{\delta})} + T_0\rho^* \\
&\quad + K\tau_2\left[\left(D + 1 + \left(1 + \frac{D(1-\alpha)}{2}\right)L_1\right)M^{3/2}C_1 + \left(1 + \frac{D(1-\alpha)}{2}\right)L_2 M^{1/2}C_2\right]\sqrt{\frac{\log(\frac{6(S^2+S)K^3}{\delta})}{\tau_1}}.
\end{aligned} \tag{32}$$

Step 3: Summing up the regret

Combining (22) and (32), we can get that with probability at least $1 - \frac{5}{2}\delta$, the first term of regret (17) is bounded by

$$\begin{aligned}
&\sum_{t=1}^{T} \rho^* - \bar{c}(b_t, I_t) \\
&\leq (K\tau_1 + T_0)\rho^* + KD + D\sqrt{2T\ln(\frac{1}{\delta})} \\
&\quad + K\tau_2\left[\left(D + 1 + \left(1 + \frac{D(1-\alpha)}{2}\right)L_1\right)M^{3/2}C_1 + \left(1 + \frac{D(1-\alpha)}{2}\right)L_2 M^{1/2}C_2\right]\sqrt{\frac{\log(\frac{6(S^2+S)K^3}{\delta})}{\tau_1}}.
\end{aligned} \tag{33}$$

Finally, combining (18) and (33), we can see that with probability at least $1 - \frac{7}{2}\delta$, the regret presented in (17) can be bounded by

$$\begin{aligned}
\mathcal{R}_T &\leq (K\tau_1 + T_0)\rho^* + KD + D\sqrt{2T\ln(\frac{1}{\delta})} + \sqrt{2T\ln\frac{1}{\delta}} \\
&\quad + K\tau_2\left[\left(D + 1 + \left(1 + \frac{D(1-\alpha)}{2}\right)L_1\right)M^{3/2}C_1 + \left(1 + \frac{D(1-\alpha)}{2}\right)L_2 M^{1/2}C_2\right]\sqrt{\frac{\log(\frac{6(S^2+S)K^3}{\delta})}{\tau_1}}).
\end{aligned}$$

Note that

$$\sum_{k=1}^{K-1} \tau_1 + \tau_2\sqrt{k} \leq T \leq \sum_{k=1}^{K} \tau_1 + \tau_2\sqrt{k},$$

so the number of episodes K is bounded by $(\frac{T}{\tau_1+\tau_2})^{2/3} \leq K \leq 3(\frac{T}{\tau_2})^{2/3}$.
Thus, we have shown that with probability at least $1-\frac{7}{2}\delta$,

$$\mathcal{R}_T \leq CT^{2/3}\sqrt{\log\left(\frac{9(S+1)}{\delta}T\right)} + T_0\rho^*,$$

where $S = 2I$, and

$$C = 3\sqrt{2}\left[\left(D+1+\left(1+\frac{D(1-\alpha)}{2}\right)L_1\right)M^{3/2}C_1 + \left(1+\frac{D(1-\alpha)}{2}\right)L_2M^{1/2}C_2\right]\tau_2^{1/3}\tau_1^{-1/2}$$

$$+ 3\tau_2^{-2/3}(\tau_1\rho^* + D) + (D+1)\sqrt{2\ln(\frac{1}{\delta})}.$$

Here, M is the number of Markov chain states, where $L_1 = \frac{4M(1-\epsilon)^2}{\epsilon^2\min\{\boldsymbol{\mu}_{\min}, 1-\boldsymbol{\mu}_{\max}\}}$, $L_2 = \frac{4M(1-\epsilon)^2}{\epsilon^3} + \sqrt{M}$, with $\epsilon = \min_{1\leq i,j\leq M} P_{i,j}$, $\alpha = 1-\frac{\epsilon}{1-\epsilon}$, and $\boldsymbol{\mu}_{\max}$, $\boldsymbol{\mu}_{\min}$ are the maximum and minimum element of the matrix $\boldsymbol{\mu}$ respectively. The proof of Theorem 1 is therefore complete.

E.1 Proof of Proposition 4

Proof. For each episode $k = 1, 2, \cdots, K$, let E_k be the index set of the kth exploration phase, and $E = \cup_{k=1}^K E_k$ be the set of all exploitation periods among the horizon T and v_k be the value function of the optimistic POMDP at the kth exploitation phase. For an arbitrary time t, let $n = \sum_{i=1}^t \mathbb{1}_{i\in E}$, which means the number of exploitation periods up to time t. Define a stochastic process $\{Z_n, n \geq 0\}$:

$$Z_0 = 0,$$
$$Z_n = \sum_{j=1}^n \mathbb{E}^\pi[v_{k_j}(b_{t_j+1}^{k_j})|\mathcal{F}_{t_j}] - v_{k_j}(b_{t_j+1}^{k_j}),$$

where $k_j = \{k : j \in E_k\}$ and $t_j = \min\{t : \sum_{i=1}^t \mathbb{1}_{i\in E} = j\}$ mean the corresponding episode and period of jth exploitation, respectively.

We first show that $\{Z_n, n \geq 0\}$ is a martingale. Note that $\mathbb{E}^\pi[|Z_n|] \leq \sum_{j=1}^n \text{span}(v_{k_j}) \leq nD < TD < \infty$. It remains to show $\mathbb{E}^\pi[Z_n|\mathcal{F}_{n-1}] = Z_{n-1}$ holds, i.e., $\mathbb{E}^\pi[Z_n - Z_{n-1}|\mathcal{F}_{n-1}] = 0$. Note that

$$\mathbb{E}^\pi[Z_n - Z_{n-1}|\mathcal{F}_{n-1}] = \mathbb{E}^\pi[\mathbb{E}^\pi[v_{k_n}(b_{t_n+1}^{k_n})|\mathcal{F}_{t_n}] - v_{k_n}(b_{t_n+1}^{k_n})|\mathcal{F}_{n-1}] = 0,$$

where the last equality is due to $n-1 \leq n \leq t_n$ then applying the tower property.

Therefore, $\{Z_n, n \geq 0\}$ is a martingale for any given policy π. Moreover, by Proposition 2, we have

$$|Z_n - Z_{n-1}| = |\mathbb{E}^\pi[v_{k_n}(b_{t_n+1}^{k_n})|\mathcal{F}_{t_n}] - v_{k_n}(b_{t_n+1}^{k_n})| \leq \text{span}(v_{k_n}) \leq D.$$

Thus, $\{Z_n, n \geq 0\}$ is a martingale with bounded difference.

Let $\bar{N} = \sum_{i=1}^T \mathbb{1}_{i\in E_k}$ and apply the Azuma-Hoeffding inequality [9], we have

$$\mathbb{P}(Z_{\bar{N}} - Z_0 \geq \epsilon) \leq \exp\left(\frac{-\epsilon^2}{2\sum_{t=1}^{\bar{N}} D^2}\right).$$

Note that $\bar{N} \leq T$ and $Z_{\bar{N}} = \sum_{k=1}^K \sum_{t\in E_k} \mathbb{E}^\pi[v_k(b_{t+1})|\mathcal{F}_t] - v_k(b_{t+1})$. Thus, setting $\epsilon = D\sqrt{2T\ln(\frac{1}{\delta})}$, we can obtain

$$\mathbb{P}\left(\sum_{k=1}^K \sum_{t\in E_k} \mathbb{E}^\pi[v_k(b_{t+1})|\mathcal{F}_t] - v_k(b_{t+1}) \geq D\sqrt{2T\ln(\frac{1}{\delta})}\right) \leq \delta.$$

Hence we have completed the proof. □