Switching Latent Bandits

Anonymous authors
Paper under double-blind review

Abstract

We consider a Latent Bandit problem where the latent state keeps changing in time according to an underlying Markov chain, and every state is represented by a specific Bandit instance. At each step, the agent chooses an arm and observes a random reward but is unaware of which MAB he is currently pulling. As typical in Latent Bandits, we assume to know the reward distribution of the arms of all the Bandit instances. Within this setting, our goal is to learn the transition matrix determined by the Markov process. We propose a technique to tackle this estimation problem that results in solving a least-square problem obtained by exploiting the knowledge of the reward distributions and the properties of Markov chains. We prove the consistency of the estimation procedure, and we make a theoretical comparison with standard Spectral Decomposition techniques. We then discuss the dependency of the problem on the number of arms and present an offline method that chooses the best subset of possible arms that can be used for the estimation of the transition model. We ultimately introduce the SL-EC algorithm based on an Explore then Commit strategy that uses the proposed approach to estimate the transition model during the exploration phase. This algorithm achieves a regret of the order $\widetilde{\mathcal{O}}(T^{2/3})$ when compared against an oracle that builds a belief representation of the current state using the knowledge of both the observation and transition model and optimizes the expected instantaneous reward at each step. Finally, we illustrate the effectiveness of the approach and compare it with state-of-the-art algorithms for non-stationary bandits and with a modified technique based on spectral decomposition.

1 Introduction

The Multi-Armed Bandit (MAB) (Lattimore & Szepesvári, 2020) framework is a well-known model used for sequential decision-making with little or no information. This framework has been successfully applied in a large number of fields, such as recommender systems, advertising, and networking. In the general MAB formulation, a learner sequentially selects an action among a finite set of different ones. The choice over the arm to select is made by properly balancing the exploration-exploitation trade-off with the goal of maximizing the expected total reward over a horizon T and guaranteeing the *no-regret* property, thus meaning that the loss incurred by not knowing the best arm is increasing sublinearly over time. Standard MAB literature requires the payoff of the available actions to be stationary (i.e., rewards come from a fixed distribution) in order to design efficient no-regret algorithms.

However, in many real-life applications, the stationarity assumption may not necessarily hold as data may be subjected to changes over time. In some applications, it is also possible to identify different data distributions, each one corresponding to a specific working regime that can be modeled as a MAB. In cases of large availability of historical data appearing in the form of past user interactions, it is possible to learn *offline* the observation models associated with the different arms for each working regime. Exploiting the knowledge of observation models leads to many advantages over the *fully online exploration* setting where no prior information is available at the beginning, and a massive number of interactions is required to learn the observation models associated with each working regime. It is often the case that the underlying working regime (state) cannot be directly observed and the non-stationarity of the process leads to regime changes over time. By knowing how these regimes are characterized, it may be possible to learn the dynamics of the changes by repeatedly interacting with the evolving environment. Inferring the underlying state accelerates the adaptation of the agent to the environment, thus leading to improved performances over time.

Learning observation models independently and before transition models may be a possible choice when there is little availability of computational resources. Indeed, we will show in the following that spectral decomposition (SD) techniques (Anandkumar et al., 2014), which are used to learn jointly the observation and the transition model, typically require a large number of samples and involve computationally intensive operations. Other scenarios where we can assume that the observation models are already known are those where the models are learned offline from samples generated by simulators. Once these models are deployed in a environment that is characterized by changes, the dynamics can be learned by interacting with the environment. We can consider, for example, the problem of resource allocation such as the electricity allocation in a specific residential area. This problem can be modeled as a Bandit where each arm represents a specific allocation configuration, while the rewards represent the extent to which the allocation has been optimal. Obviously, the optimality of the allocation depends on the state of the system, which may be conditioned by several factors such as environmental conditions, community trends, and seasonality.

Another possible scenario that suits our setting is the one of *Transfer Learning*, where partial knowledge of the system (in our case, the observation model) can be used in a context with different dynamics (and a new transition model needs to be learned). In the scenario previously mentioned, we can consider using the same observation models in a new residential area, with a structure analog to the first one (thus justifying the use of the same observation model) but located in a different place, with potentially different weather conditions and inhabitants having different behaviors.

Assuming the existence of a finite set of discrete latent states is a relevant choice when approaching the modeling of complex real-life problems characterized by different and recurrent working regimes. These regimes can be typically observed in domains such as the financial market and online advertising, typically marked by high volatility and specific seasonality patterns (M. et al., 2022; Heston & Sadka, 2008; Guo et al., 2021). Introducing a more practical example, in the stock exchange market where different models are available, typically one for each regime, it is relevant to choose the best stock to exchange based on the unknown market condition. The different regimes may be identified through the availability of past data by either considering some seasonality patterns or specific indicators of the market conditions using some domain knowledge. In this case, inferring the current state of the market, associating a duration with it, and predicting future transitions allow for fairer decisions and higher outcomes.

Past works focused on this state identification problem under the assumption of knowing the conditional observation models (Maillard & Mannor, 2014; Zhou & Brunskill, 2016) and defined theoretically optimal UCB algorithms. Follow-up work by Hong et al. (2020a) provided more practical Thompson Sampling algorithms, also considering the problem of model misspecification and came up with an analysis on the Bayes regret.

The works cited above assume that the latent state does not change during the interaction process: once the real state is identified, the agent can act optimally. Differently, we embrace a more realistic scenario and assume that the latent state can change through time. In accordance with the latent bandits setting, we assume that the learning agent is aware of the observation models of the arms conditioned on each latent state. A setting similar to ours has been considered also in Hong et al. (2020b), the key difference is that they assume to have either full or partial knowledge of both the observation model and the transition model. We instead focus on the problem of learning the transition model given the knowledge of the observation models and maximizing the cumulative reward over T interaction steps.

More specifically, our problem is modeled by assuming the existence of a finite set $\mathbb{S}$ of different MABs all sharing the same set of finite arms $\mathbb{I}$, each generating rewards (observations) in a finite set $\mathbb{V}$. Each state $s \in \mathbb{S}$ represents a different instance of a MAB. At each time step t, there is a transition from latent state s_{t-1} to the new latent state s_t according to the transition matrix governing the process. The action a_t selected in t will thus generate a reward conditioned on the latent state s_t.

Contributions and Outline We introduce the Related Works in Section 2 and the Preliminaries in Section 3. After that, we define the formulation of the problem that considers known Bandit instances that switch through time according to an underlying Markov process. The information about the reward distributions of the bandit instances is encoded into a suitable observation matrix, while the transition

model of the chain needs to be estimated. The learning objective of the agent is to maximize at each instant the expected instantaneous reward given the estimated belief over the current Bandit. After this part, we introduce the main assumptions that hold in our setting, motivate the reasons behind them, and show how they can be relaxed for the estimation of the transition model.

Section 5.1 presents the estimation procedure of the transition model that uses samples collected using a round-robin procedure for selecting arms. Then, we propose an offline arm selection strategy that chooses a subset of the available arms for the estimation approach, with the objective of promoting diversity between observation distributions induced by the arms in order to enhance the identifiability capabilities. In Section 5.2, we detail the SL-EC algorithm that employs an Explore then Commit approach and uses the proposed estimation procedure for learning the transition model during the exploration phase.

Finally, Section 7 shows numerical simulations on synthetic and semi-synthetic data. We further provide additional experiments that highlight the difference in performance between our estimation procedure and a technique based on SD approaches. We complement the comparison with SD approaches on the theoretical side in Appendix C.

2 Related Works

Non-stationary Bandits Non-stationary behaviors are closer to real-world scenarios, and this has induced a vast interest in the scientific community leading to the formulation of different methods that consider either abruptly changing environments (Garivier & Moulines, 2011), smoothly changing environments (Trovò et al., 2020), or settings with a bounded variation of the rewards (Besbes et al., 2014). It is known that when rewards may arbitrarily change over time, the problem of Non-Stationary Bandits is intractable, meaning that only trivial bounds can be derived on the dynamic pseudo-regret. That is the main reason why in the literature, there is a large focus on non-stationary settings enjoying some specific structure in order to design algorithms with better guarantees. Non-stationary MAB approaches typically include both passive methods in which arm selection is mainly driven by the most recent feedback (Auer et al., 2019; Besbes et al., 2014; Trovò et al., 2020) and active methods where a change detection layer is used to actively perceive a drift in the rewards and to discard old information (Liu et al., 2017; Cao et al., 2018). Works such as Garivier & Moulines (2011) provide a $\mathcal{O}(\sqrt{T})$ regret guarantee under the assumption of knowing the number of abrupt changes. Other works, such as Besbes et al. (2014), employ a fixed budget to bound the total variation of expected rewards over the time horizon. They are able to provide a near-optimal frequentist algorithm with pseudo-regret $\mathcal{O}(T^{2/3})$ and a distribution-independent lower bound. All the above methods are not suited for environments that switch between different regimes as they do not keep in memory past interactions but rather tend to forget or discard the past.

A particular type of non-stationary Bandit problem related to our work includes the *restless Markov* setting (Ortner et al., 2014; Slivkins & Upfal, 2008) where each arm is associated with a different Markov process and the state of each arm evolves independently of the learner's actions. Differently, Fiez et al. (2018) investigate MAB problems with rewards determined by an unobserved Markov chain where the transition to the next state depends on the action selected at each time step, while Zhou et al. (2021) focus on MAB problems where the state transition dynamics evolves independently of the chosen action. This last work has many similarities with our setting. The main difference lies in the fact that they do not assume to know the conditional reward models and learn them jointly with the transition matrix. They make use of SD techniques (Anandkumar et al., 2014) and use this tool in a regret minimization algorithm achieving a $\mathcal{O}(T^{2/3})$ regret bound. Their setting is more complex than ours but involves additional assumptions, like the invertibility of the transition matrix that defines the chain. Furthermore, spectral methods need a vast amount of samples in order to provide reasonable estimation errors and can hardly be used in large problems. A detailed discussion on the differences between the estimation procedure used in Zhou et al. (2021) and ours is presented in Appendix C.

Latent Bandits More similar lines of work are related to bandit studies where latent variables determine the distribution of rewards (Maillard & Mannor, 2014; Zhou & Brunskill, 2016). In these works, the unobserved state is fixed across different rounds, and the conditional rewards depend on the latent state. Maillard

& Mannor (2014) developed UCB algorithms without context, considering the two different cases in which the conditional rewards are either known or need to be estimated. This line of work has been extended to the contextual bandit case in Zhou & Brunskill (2016) where there is an offline procedure to learn the policies and a selection strategy to use them online. Hong et al. (2020a) proposed a TS procedure in the contextual case that updates a prior probability over the set of states in order to give a higher probability to the real latent state. A non-stationary variant of this setting is proposed in Hong et al. (2020b) where the latent states are assumed to change according to an underlying Markov chain. They develop TS algorithms under different cases when both the reward and transition models are completely known and when partial information about them is available. For the partial information case, they provide an algorithm based on particle filters, which will be used for comparison in the experimental section. Differently from Hong et al. (2020b), we do not assume any prior information about the transition matrix, and we learn it through interactions with the environment using the information about the reward models.

Another interesting work associated with latent bandits is the one from Kwon et al. (2022) where, differently from previously cited works, they assume an episodic setting with a fixed horizon H. At the beginning of each episode, a specific MAB instance is sampled from a fixed mixing distribution, and the agent interacts with the sampled MAB until the end of the episode without being aware of the MAB she is interacting with. The goal is to learn both the mixture weights and the reward distributions associated with each MAB. The relevant difference with our work is that they consider an episodic setting, while we consider a continuous one. Another main difference is that they provide results in terms of sample complexity needed in order to learn a near-optimal policy, not taking into account the suffered regret.

3 Preliminaries

In the following, we will present the main elements that are useful to understand what will follow. We will denote with $\Delta(\mathbb{X})$ the simplex of a finite space $\mathbb{X}$, and we will use the bold symbol $\boldsymbol{P}$ to denote the transition matrix and the probabilities associated with a Markov chain (see Section 3.2).

3.1 Multi-Armed Bandits

A I-armed stochastic bandit (Lattimore & Szepesvári, 2020) is a collection of distributions $\nu = (\Pr(\cdot|a) \in \Delta(\mathbb{V}) : a \in \mathbb{I})$ where $\mathbb{I}$ is the set of available actions with cardinality I and $\mathbb{V}$ is a finite set with cardinality V of possible rewards. A learning agent sequentially interacts with the environment over T rounds. For each round $t \in \{1, \dots, T\}$, the learner chooses an action $a_t \in \mathbb{I}$ and the environment gives as output a reward $r_t \in \mathbb{V}$. The goal of the learner is to maximize the sum of cumulative rewards $\sum_{t=1}^{T} r_t$, which is a random quantity that depends on the stochasticity of both the environment and the choice of the agent's actions. In general, the performance of a bandit algorithm is measured using the notion of regret, which is defined as the deficit suffered by the learning agent with respect to the optimal policy. The regret of a policy θ on a bandit instance is defined as:

$$\mathcal{R}_T(\theta) = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \tag{1}$$

where μ^* defines the maximum expected reward among the available arms, while the expectation is taken with respect to the policy θ and the stochasticity of the environment.

3.2 Markov Chains

A Markov Chain (MC) (Feller, 1968) is defined by a tuple $\mathcal{M} := (\mathbb{S}, \boldsymbol{P}, \boldsymbol{\nu})$, where $\mathbb{S}$ is a finite state space ($|\mathbb{S}| = S$), $\boldsymbol{P} : \mathbb{S} \to \Delta(\mathbb{S})$ is the transition model, such that $\boldsymbol{P}(s, s')$ denotes the probability of reaching state $s' \in \mathbb{S}$ when being in state $s \in \mathbb{S}$, and $\boldsymbol{\nu}$ is the initial state distribution. We will denote with $\boldsymbol{P}$ the stochastic matrix of size $S \times S$ representing the transition model.

A Markov chain is said to be *ergodic* if its associated transition matrix consists of a single recurrent class containing all states (Puterman, 1994). Ergodic Markov chains satisfy the properties of being *irreducibile*,

thus meaning that it is possible to reach any state from any other state with positive probability in a finite number of steps and *aperiodic*, meaning that the chain does not follow a regular, repeating pattern in its transitions.

We state the following result for ergodic Markov chains.

Proposition 3.1. *Let $\boldsymbol{P}$ be the transition matrix of an ergodic Markov Chain and $\boldsymbol{\nu}$ an arbitrary probability vector. Then:*

$$\lim_{n\to\infty} \boldsymbol{\nu}\boldsymbol{P}^n = \boldsymbol{\pi},$$

where $\boldsymbol{P}^n$ represents the transition kernel induced after n steps, $\boldsymbol{\pi}$ is the unique stationary distribution of the chain, and the components of the vector $\boldsymbol{\pi}$ are all strictly positive.

Since the stationary distribution of the Markov chain is unique, it follows that there is only one eigenvalue with unitary value ($\lambda_{\max} = 1$). Let's define the set of ordered moduli of the eigenvalues of transition matrix $\boldsymbol{P}$ as $(|\lambda_i|)_{i=1}^{S}$. By denoting $|\lambda_{\max}| = |\lambda_1|$, we have the following relation:

$$|\lambda_1| = 1 > |\lambda_2| \geq \cdots \geq |\lambda_S|,$$

where the inequality between $|\lambda_1|$ and $|\lambda_2|$ is strict for ergodic chains. The quantity $1 - |\lambda_2|$ is defined as the spectral gap of the Markov chain and controls the rate of convergence of the chain towards its stationary distribution (Krishnamurthy, 2016). In what will follow, we will use the symbol λ to denote the modulus of the second largest eigenvalue, such that $\lambda = |\lambda_2|$.

4 Switching Latent Bandits

We consider to have a finite set $\mathbb{S} := \{s_1, \ldots, s_S\}$ of $S = |\mathbb{S}|$ different MAB instances. Each MAB is characterized by a finite set of discrete arms $\mathbb{I} := \{a_1, \ldots, a_I\}$ with cardinality $I = |\mathbb{I}|$ and a finite set of possible rewards $\mathbb{V} = \{r_1, \ldots, r_V\}$ with cardinality $V = |\mathbb{V}|$. Whenever an arm $a \in \mathbb{I}$ is pulled, a corresponding reward $r \in \mathbb{V}$ is generated by the environment. We consider each reward $r \in \mathbb{V}$ bounded for simplicity in the range $[0, 1]$. All the considered MABs share the same sets of arms $\mathbb{I}$ and rewards $\mathbb{V}$. The distribution of rewards $\Pr(\cdot|s, a)$ conditioned on MAB instance s and action a is categorical[1]. In particular, we assume to know the parameters characterizing these distributions and to store this information into matrix $\boldsymbol{O} \in \mathbb{R}^{IV\times S}$, which we call action observation matrix. Each row of this matrix encodes a specific action-reward pair $(a, r) \in \mathbb{I} \times \mathbb{V}$. Then, for any pair $(a, r) \in \mathbb{I} \times \mathbb{V}$ and any state $s \in \mathbb{S}$, we have:

$$\boldsymbol{O}\big((a, r), s\big) = \Pr(r|s, a), \tag{2}$$

where $\Pr(r|s, a)$ represents the probability value of observing reward r while pulling action a from MAB s. At each step t, only one MAB $s_t \in \mathbb{S}$ is active, and it determines the reward r_t that is received when the agent pulls action a_t. The choice over the active MAB is determined by an underlying Markov chain with transition matrix $\boldsymbol{P} \in \mathbb{R}^{S\times S}$. More precisely, the probability over the next active MAB s_{t+1} is determined by the distribution $\boldsymbol{P}(s_t, \cdot) \in \Delta(\mathbb{S})$ and is thus independent of the chosen action a_t. The setting we consider assumes that the agent is not able to observe the active MAB at each step, and the objective is to learn the transition matrix $\boldsymbol{P}$ characterizing the underlying process while knowing the observation model $\boldsymbol{O}$.

Learning objective As already seen, the agent does not observe the sequence of MAB instances, but by deriving an estimate of the transition matrix $\boldsymbol{P}$, a belief representation over the current active MAB $s \in \mathbb{S}$ can be defined. In the following, we will report the update rule of the belief vector $\mathbf{b}_t \in \Delta(\mathbb{S})$ under the knowledge of the observation model $\boldsymbol{O}$ and the transition model $\boldsymbol{P}$. The update of the belief derives from the typical correction and update step of the Bayes rule, where the correction step adjusts the current belief $\mathbf{b}_t$ using the reward r_t obtained by pulling arm a_t and the prediction step computes the new belief $\mathbf{b}_{t+1}$ simulating a transition step of the chain. More formally, for each element $\mathbf{b}_{t+1}(s)$ of the belief vector $\mathbf{b}_{t+1}$, the update step is as follows:

$$\mathbf{b}_{t+1}(s) = \frac{\sum_{s'\in\mathbb{S}} \mathbf{b}_t(s')\boldsymbol{O}\big((a_t, r_t), s'\big)\boldsymbol{P}(s', s)}{\sum_{s''\in\mathbb{S}} \mathbf{b}_t(s'')\boldsymbol{O}\big((a_t, r_t), s''\big)}. \tag{3}$$

[1]In Section 6, we will see how this formulation can be extended to continuous distributions.

After having defined the update rule of the belief vector $\mathbf{b}_t$, we introduce, for each action $a \in \mathbb{I}$, vector $\boldsymbol{\mu}(a) \in \mathbb{R}^S$ where element $\boldsymbol{\mu}(a, s)$ referred to state $s \in \mathbb{S}$ contains the expected reward obtained when pulling arm a while being in state s. More formally:

$$\boldsymbol{\mu}(a, s) = \sum_{r \in \mathbb{V}} r\boldsymbol{O}\big((a, r), s\big). \tag{4}$$

Given the belief $\mathbf{b}_t$ over the states, the objective of the agent is to pull the action that maximizes the instantaneous expected reward such that:

$$a_t = \arg\max_{a \in \mathbb{I}} \sum_{s \in \mathbb{S}} \boldsymbol{\mu}(a, s)\mathbf{b}_t(s) = \arg\max_{a \in \mathbb{I}} \langle \boldsymbol{\mu}(a), \mathbf{b}_t \rangle, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product between the two vectors.
From the considerations reported above, we are now ready to formulate the notion of regret we try to minimize:

$$\mathfrak{R}_T = \sum_{t=1}^{T} \max_{a \in \mathbb{I}} \langle \boldsymbol{\mu}(a), \mathbf{b}_t \rangle - \max_{a \in \mathbb{I}} \langle \boldsymbol{\mu}(a), \widehat{\mathbf{b}}_t \rangle \tag{6}$$

where $\mathbf{b}_t$ and $\widehat{\mathbf{b}}_t$ denote the belief vectors updated using the real transition matrix $\boldsymbol{P}$ and the estimated one $\widehat{\boldsymbol{P}}$ respectively. Here, we used symbol $\mathfrak{R}$ to characterize the regret defined in Equation 6, in order to discriminate from the standard notion of regret $\mathcal{R}$ introduced in Section 3.1.

4.1 Assumptions

We need now to introduce some assumptions that should hold in our setting:

Assumption 4.1. *The smallest element of the transition matrix defining the Markov chain is* $\epsilon := \min_{s,s' \in \mathbb{S}} \boldsymbol{P}(s, s') > 0$.

This assumption ensures a non-null probability of transitioning from any state to any other in one step. It is possible to show that under this assumption, the induced Markov chain is ergodic, thus guaranteeing the presence of a unique stationary distribution, as shown in Proposition 3.1. Under the ergodic condition, the chain is able to reach its stationary distribution $\boldsymbol{\pi}$ geometrically fast, regardless of its initial distribution Krishnamurthy (2016). Our assumption on the minimum entry is not a necessary condition for the two aforementioned motivations but a sufficient one. However, we require this condition to bound the error between the belief computed using the real transition matrix and an estimated one. This result is presented in Proposition D.4 and builds on the original result present in De Castro et al. (2017). This one-step reachability assumption is always present in works dealing with partial observability that show results in terms of regret in non-episodic scenarios. Notably, it has been used in similar works such as Zhou et al. (2021); Jiang et al. (2023); Mattila et al. (2020) and also employed in the more complex POMDP setting (Xiong et al., 2022). Works not using this assumption either do not need it since they use a less powerful class of policies such as memoryless ones[2] (Azizzadenesheli et al., 2016) or they directly impose an error of the estimated belief that adequately decreases with the number of collected samples (Jafarnia Jahromi et al., 2022).

Assumption 4.2. *The action observation matrix* $\boldsymbol{O} \in \mathbb{R}^{IV \times S}$ *is full column rank.*

This second assumption, instead, is related to the identifiability of the parameters of the problem and has been largely used in works using spectral decomposition techniques (Zhou et al., 2021; Azizzadenesheli et al., 2016; Hsu et al., 2012). A robust version of this assumption, called weakly-revealing[3] is also present in other works involving learning parameters in POMDPs (Liu et al., 2022; Jin et al., 2020). In the following, we will see that this is a necessary condition in order to recover matrix $\boldsymbol{P}$. Indeed, we will see that the error of the estimation procedure has an inverse dependency on the minimum singular value $\sigma_{\min}(\boldsymbol{O})$ of the action observation matrix $\boldsymbol{O}$ and through Assumption 4.2, we implicitly require that $\sigma_{\min}(\boldsymbol{O}) > 0$.

[2] A memoryless policy defines the action to choose only based on the last observation seen. For this reason, it does not require a notion of belief over the states.

[3] The α-weakly revealing assumption defines a lower bound α to the minimum singular value of the observation matrix $\boldsymbol{O}$, such that $\sigma_{\min}(\boldsymbol{O}) \geq \alpha$.

5 Proposed Approach

As clarified in the previous section, our goal is to minimize the regret formulated in Equation 6. To reach this objective, we need to define a good estimate $\widehat{\boldsymbol{P}}$ of the transition matrix that in turn results in a more accurate update of the belief vector $\widehat{\mathbf{b}}_t$. We will now show how the transition model can be learned by exploiting the knowledge of the observation model $\boldsymbol{O}$ (Section 5.1) and we will present the SL-EC algorithm that makes use of the presented estimation approach in order to minimize the regret (Section 5.2).

5.1 Transition Model Estimation

The Markov chain estimation procedure presented in this section holds under weaker assumptions than those presented in Section 4.1. In particular, we relax the one-step reachability assumption (Assumption 4.1) and we only require the ergodicity of the transition matrix $\boldsymbol{P}$.

Stationary Distribution of Consecutive States We start with a consideration about the transition matrix that defines the chain. Building on Proposition 3.1, an ergodic chain admits a unique stationary distribution $\boldsymbol{\pi}$. By definition, this distribution satisfies the equation $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$.
From the uniqueness of this distribution, it can be easily shown that there exists as well a unique stationary distribution on consecutive states that we represent with a matrix $\boldsymbol{W} \in \Delta(\mathbb{S}^2)$ having dimension $S \times S$.
It can be formally obtained as follows. Let $\boldsymbol{\Pi} = diag(\boldsymbol{\pi})$ be a diagonal matrix of size $S \times S$ having values of the stationary distribution $\boldsymbol{\pi}$ along its diagonal. Matrix $\boldsymbol{W}$ of the stationary distribution of consecutive states can be computed as:

$$\boldsymbol{W} = \boldsymbol{\Pi}\boldsymbol{P},$$

that is obtained by multiplying each row of the transition matrix $\boldsymbol{P}$ by the associated probability value of the stationary distribution. The reverse procedure that allows retrieving matrix $\boldsymbol{P}$ from $\boldsymbol{W}$ is defined by the following equation:

$$\boldsymbol{P}(s, s') = \frac{\boldsymbol{W}(s, s')}{\sum_{s'' \in \mathbb{S}} \boldsymbol{W}(s, s'')}, \tag{7}$$

which shows that the rows of matrix $\boldsymbol{P}$ are obtained by normalizing the rows of matrix $\boldsymbol{W}$ such that they sum to 1, as required for stochastic matrices.
The next paragraph shows how the matrix $\boldsymbol{W}$ of stationary distribution of consecutive states relates to the stationary distribution of consecutive rewards.

Stationary Observation-State Relation Let's choose an arm $a \in \mathbb{I}$: we will denote with $\boldsymbol{d}_a(\cdot) \in \Delta(\mathbb{V})$ the stationary distribution of rewards conditioned on pulling action a when the chain has mixed[4]. Vector $\boldsymbol{d}_a$ has dimension V and its elements are characterized as follows:

$$\boldsymbol{d}_a(r) = \sum_{s \in \mathbb{S}} \boldsymbol{O}\big((a, r), s\big)\boldsymbol{\pi}(s), \qquad \forall\, r \in \mathbb{V}, \tag{8}$$

where we recall that $\boldsymbol{\pi}(s)$ represents the probability of state s taken from the stationary distribution of the chain and $\boldsymbol{O}\big((a, r), s\big)$ represents the probability of observing reward r while pulling action a and being in state s. A similar rationale can be extended to consecutive rewards $(r, r') \in \mathbb{V}^2$ conditioned on pulling a couple of consecutive actions $(a, a') \in \mathbb{I}^2$. We denote with $\boldsymbol{d}_{a,a'}(\cdot) \in \Delta(\mathbb{V}^2)$ the distribution over consecutive rewards conditioned on pulling the pair of arms (a, a'). We represent it with a vector of size V^2 and define it as follows:

$$\boldsymbol{d}_{a,a'}((r, r')) = \sum_{s, s' \in \mathbb{S}^2} \boldsymbol{O}\big((a, r), s\big)\boldsymbol{O}\big((a', r'), s'\big)\boldsymbol{W}(s, s'), \qquad \forall\, (r, r') \in \mathbb{V}^2, \tag{9}$$

[4]The distribution of states in a mixed chain corresponds by definition to its stationary distribution $\boldsymbol{\pi}$.

where we recall that matrix $\boldsymbol{W} \in \Delta(\mathbb{S}^2)$ represents the stationary distribution of consecutive states.
By considering the different vectors of type $\boldsymbol{d}_{a,a'}$, we define vector:

$$\boldsymbol{d} = \left(\boldsymbol{d}_{a,a'}\right)_{(a,a')\in\mathbb{I}^2} \tag{10}$$

where the term on the right denotes a concatenation of vectors $\boldsymbol{d}_{a,a'}$ for all $(a, a') \in \mathbb{I}^2$ and the resulting vector $\boldsymbol{d}$ has size I^2V^2.
We define now a new matrix $\boldsymbol{A} \in \mathbb{R}^{I^2V^2\times S^2}$ to which we will refer as reference matrix. It extends the information contained in the action observation matrix $\boldsymbol{O}$ considering consecutive pairs of elements and it is characterized as follows:

$$\boldsymbol{A} = \boldsymbol{O} \otimes \boldsymbol{O}, \tag{11}$$

where symbol $\otimes$ refers to the Kronecker product (Loan, 2000). Since we assume knowledge of the observation model $\boldsymbol{O}$, we can directly compute the reference matrix by applying the Kronecker operator.
As a last step before presenting the main result, we transform matrix $\boldsymbol{W}$ and vectorized[5] it to obtain vector $\boldsymbol{w} \in \Delta(\mathbb{S}^2)$ having dimension S^2. By using the quantities just defined, we can finally reformulate Equation 9 so that it can be extended to all pairs of actions. Using vector notation, we have:

$$\boldsymbol{d} = \boldsymbol{A}\boldsymbol{w}. \tag{12}$$

Basically, this equation relates the stationary probability distribution of consecutive observations with the stationary probability distribution of consecutive latent states. The next paragraph shows how to obtain an estimate $\widehat{\boldsymbol{d}}$ of vector $\boldsymbol{d}$, from which, by using Equation 12, it can then be computed an estimate $\widehat{\boldsymbol{w}}$ of the stationary distribution of consecutive states.

Transition Model Estimation We will now see how to concretely compute an estimate of $\boldsymbol{w}$ using Equation 12. Going back to vectors $\boldsymbol{d}_{a,a'}$, we can build an estimate $\widehat{\boldsymbol{d}}_{a,a'}$ for each pair of actions $(a, a') \in \mathbb{I}^2$. For this purpose, let's take a pair of action (a, a') and repeatedly pull it. We can count the number of occurrences of each pair of observed rewards $(r, r') \in \mathbb{V}^2$ and store this information in a suitable count vector $\boldsymbol{n}_{a,a'}$ of size V^2. We can easily derive an estimate of vector $\boldsymbol{d}_{a,a'}$ as follows:

$$\widehat{\boldsymbol{d}}_{a,a'} = \frac{\boldsymbol{n}_{a,a'}}{N}, \tag{13}$$

where N represents the number of times the pair of consecutive arms (a, a') has been pulled.
We propose an estimation procedure that pulls each pair of arms $(a, a') \in \mathbb{I}^2$ in a round-robin fashion and stores the observed pairs of rewards in the corresponding vector count $\boldsymbol{n}_{a,a'}$. The choice of a round-robin approach allows for showing some interesting properties in the theoretical analysis, as will be shown later in Section 6. By executing N different rounds, thus meaning that each pair of arms is pulled exactly N times and by exploiting the knowledge of the reference matrix $\boldsymbol{A}$, we can derive:

$$\widehat{\boldsymbol{w}} = \boldsymbol{A}^\dagger\widehat{\boldsymbol{d}} = \boldsymbol{A}^\dagger\frac{\boldsymbol{n}}{N} \tag{14}$$

where $\boldsymbol{A}^\dagger$ is the Moore–Penrose inverse of reference matrix $\boldsymbol{A}$, while vectors $\widehat{\boldsymbol{d}}$ and $\boldsymbol{n}$ are obtained by concatenating the different vectors $\widehat{\boldsymbol{d}}_{a,a'}$ and $\boldsymbol{n}_{a,a'}$ analogously to what done in Equation 10. The second equality is derived from extending Equation 13 to the concatenated vectors. The stated equation shows that the estimation procedure involves solving a simple least-square problem, which can be done in a computationally efficient way.
Once an estimate $\widehat{\boldsymbol{w}}$ is computed, the corresponding matrix $\widehat{\boldsymbol{W}}$ can be obtained by reverting the vectorization operation and eventually an estimate $\widehat{\boldsymbol{P}}$ of the transition model is computed using Equation 7.
The pseudocode of the presented estimation procedure is detailed in Algorithm 1.

[5]The vectorization operation used here creates a new vector $\boldsymbol{w}$ by concatenating each row of matrix $\boldsymbol{W}$.

Algorithm 1: Estimation Procedure

Input: Action Observation matrix $\boldsymbol{O}$, number of rounds N

Build Reference matrix $\boldsymbol{A}$ using Equation 11
Initialize vector of counts $\boldsymbol{n}_{a,a'}$ with zeroes for all $(a, a') \in \mathbb{I}^2$
$k = 0$
while $k < N$ **do**
    $t = k * I^2$
    foreach $(a, a') \in \mathbb{I}^2$ **do**
        Pull arm $a_t = a$
        Observe reward $r_t = r$
        Pull arm $a_{t+1} = a'$
        Observe reward $r_{t+1} = r'$
        $\boldsymbol{n}_{a,a'}(r, r') = \boldsymbol{n}_{a,a'}(r, r') + 1$
        $t = t + 2$
Compute $\widehat{\boldsymbol{d}}_{a,a'}$ for all $(a, a') \in \mathbb{I}^2$ using Equation 13
Obtain $\widehat{\boldsymbol{d}}$ by concatenating all the different $\widehat{\boldsymbol{d}}_{a,a'}$ (as done in 10)
Estimate $\widehat{\boldsymbol{w}}$ from Equation 14
Reshape vector $\widehat{\boldsymbol{w}}$ to obtain matrix $\widehat{\boldsymbol{W}}$
Compute $\widehat{\boldsymbol{P}}$ using Equation 7

5.1.1 Arm Selection Strategy

In Algorithm 1, we propose a simple approach for choosing the arms to pull. Each pair of arms is indeed pulled the same number of times during the exploration phase by using a deterministic approach. However, it can be shown that the estimation procedure proposed in Section 5.1 can be extended to a more flexible arm selection policy. We may randomize the arm choice by assigning non-uniform probabilities to each pair of arms. In principle, this aspect allows exploiting the knowledge of the known reward distribution of each arm, for example, giving at the beginning a higher probability to the pairs of arms that are more rewarding. For example, this arm selection policy may be beneficial if we plug our estimation approach into an iterative two-phase exploration and exploitation algorithm, as the one used in Zhou et al. (2021).

Offline arm selection In problems with a large number of available arms, a round-robin approach among all possible pairs of arms may be detrimental as it considers all arms equivalently. There may be cases where some actions are less useful for state identification. The extreme case is an action that induces the same observation distribution for all the Bandit instances. Indeed, pulling that action will not provide any additional information on the current MAB and the effect will only be to slow down the estimation procedure. In general, actions that induce *similar* observation distributions for all the MABs will provide *less* information with respect to actions that induce highly different distributions for all the MABs.

A more convenient approach, in this case, would be to select a subset of different arms to be used during the exploration phase. Intuitively, the arm selection procedure tends to promote diversity among arms conditioned on the latent states, with the objective of increasing the identifiability capabilities deriving from the actions. It turns out that we are able to get an understanding of the information loss we suffer by selecting specific arms, given the knowledge of the action observation matrix $\boldsymbol{O}$. In particular, in Section 6 devoted to the theoretical analysis, we will see that the quality of the estimation highly depends on the minimum singular value $\sigma_{\min}(\boldsymbol{O})$ of the action observation matrix $\boldsymbol{O}$. We can thus use this value to drive the choice of the best subset of arms to use.

In particular, by fixing a number $J < I$ of arms to use among those available, the choice over the best subset of size J can be done as follows. We consider all the possible subsets of arms of size J, and for each of these subsets, we derive a reduced action observation matrix $\boldsymbol{G}$ of size $JV \times S$ that is obtained by simply removing from the original matrix $\boldsymbol{O}$ all the rows associated to the actions not belonging to the considered subset of

Algorithm 2: SL-EC Algorithm

Input: Observation model $\boldsymbol{O}$, Exploration horizon T_0, Total horizon T

1 Define number of rounds $N = T_0/2I^2$
2 $\widehat{\boldsymbol{P}} = EstimationProcedure(\boldsymbol{O}, N)$ (Algorithm 1)
3 $\mathbf{b}_0 = Uniform()$
4 Compute $\widehat{\mathbf{b}}_{T_0}$ using samples collected during Algorithm 1
5 $t \leftarrow T_0$
6 **while** $t \leq T$ **do**
7 $\quad a_t = \arg\max_{a \in \mathbb{I}} \langle \boldsymbol{\mu}(a), \widehat{\mathbf{b}}_t \rangle$
8 $\quad$ Observe reward r_t
9 $\quad \widehat{\mathbf{b}}_{t+1} = UpdateBelief(\widehat{\mathbf{b}}_t, a_t, r_t)$ (Equation 3)
10 $\quad t = t + 1$

arms. Having defined a new action observation matrix for each generated subset, a good candidate subset of arms is the one yielding the reduced action observation matrix $\boldsymbol{G}$ with the highest $\sigma_{\min}(\boldsymbol{G})$. Understandably, this approach implies that the reduced action observation matrix $\boldsymbol{G}$ derived from the subset of selected arms should be full-column rank, thus satisfying Assumption 4.2.

5.2 SL-EC Algorithm

Having established an estimation procedure for the transition matrix $\widehat{\boldsymbol{P}}$, we will now provide an algorithm that makes use of this approach in a regret minimization framework.

We consider a finite horizon T for our problem. We propose an algorithm called *Switching Latent Explore then Commit* (SL-EC) that proceeds using an EC approach where the exploration phase is devoted to finding the best estimation of the transition matrix $\widehat{\boldsymbol{P}}$, while during the exploitation phase, we maximize the instantaneous expected reward following the formulation provided in Equation 5. The exploration phase lasts for T_0 episodes, where T_0 is optimized w.r.t. the total horizon T, as will be seen in Section 6. The pseudocode of the SL-EC Algorithm is presented in Algorithm 2.

Basically, the exploration phase pulls each pair of arms in a round-robin fashion and uses the estimation procedure presented in Algorithm 1. When the exploration phase is over, an estimation of the transition matrix $\widehat{\boldsymbol{P}}$ is computed. After that, a belief vector $\mathbf{b}_0$ is initialized by assigning a uniform probability to all the states (Line 3), and it is updated using Equation 3 and the estimated $\widehat{\boldsymbol{P}}$, considering the history of samples collected from the beginning up to T_0 (Line 4). Finally, the exploitation phase starts, as described in the pseudocode of the algorithm.

6 Theoretical Analysis

Having defined the estimation procedure of the transition model in Section 5.1 and having introduced the SL-EC algorithm, we will now provide theoretical guarantees for them.

We start with a concentration bound on the transition matrix $\widehat{\boldsymbol{P}}$ computed from the estimation procedure in Algorithm 1. As already highlighted, this estimation procedure only requires the ergodicity of the chain, thus relaxing Assumption 4.1.

Lemma 6.1. *Suppose Assumption 4.2 holds and suppose that the Markov chain with transition matrix $\boldsymbol{P}$ is ergodic, such that $\pi_{\min} := \min_{s \in \mathbb{S}} \boldsymbol{\pi}(s) > 0$ with $\boldsymbol{\pi} \in \Delta(\mathbb{S})$ being the stationary distribution of the chain. By assuming that the chain starts from its stationary distribution $\boldsymbol{\pi}$ and by pulling each pair of arms in a round-robin fashion for N rounds and using the estimation procedure reported in Algorithm 1, we have that*

with probability at least $1-\delta$ *the estimation error of the transition matrix* $\boldsymbol{P}$ *will be:*

$$\|\boldsymbol{P}-\widehat{\boldsymbol{P}}\|_F \leq \frac{2I}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}}\sqrt{\frac{2S(1+\log(I^2/\delta))}{(1-\lambda^{2I^2})N}}, \tag{15}$$

where $\|\cdot\|_F$ represents the Frobenius norm (Golub & Van Loan, 1996), $\sigma_{\min}(\boldsymbol{O})$ represents the minimum singular value of the action observation matrix $\boldsymbol{O}$, and λ represents the modulus of the second highest eigenvalue of matrix $\boldsymbol{P}$. As reported in the statement of the Lemma, N denotes the number of times each pair of arms is pulled, thus meaning that the stated error guarantee holds when interacting with the environment for a total number of $2I^2N$ steps, where the I^2 term arises from the total number of pairs of arms while the constant value 2 accounts for considering pairs of arms.

As a last remark, we note the Lemma 6.1 assumes that the chain starts from its stationary distribution. It can be shown that similar estimation guarantees can be obtained when the starting distribution is not the invariant one. Indeed, since the chain is ergodic, it converges to its stationary distribution geometrically fast (Krishnamurthy, 2016). It can be shown that, by discarding a first subset of samples and using the successive ones for estimation, the estimation error would incorporate a further logarithmic dependency on the number of rounds N. We will not focus on this aspect, but more details on this matter can be found in Fan et al. (2021).

Here, we will provide a sketch of the proof of the presented Lemma. A more detailed version of this proof is reported in Appendix B.

Sketch of the proof. The proof of Lemma 6.1 builds on two principal results. The former comprises a relation that links the estimation error of the matrix $\boldsymbol{P}$ with the estimation error of matrix $\boldsymbol{W}$, while the latter is a concentration bound on the estimated $\widehat{\boldsymbol{W}}$ from the true one $\boldsymbol{W}$. Concerning the first result, we can say that:

$$\|\boldsymbol{P}-\widehat{\boldsymbol{P}}\|_F \leq \frac{2\sqrt{S}\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\|_F}{\pi_{\min}}. \tag{P.1}$$

This result follows from a sequence of algebraic manipulations and makes use of Lemma D.1 appearing in Appendix D.

We now need to define a bound on $\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\|_F$. In order to bound this quantity, we resort to the vectorized versions $\boldsymbol{w}$ and $\widehat{\boldsymbol{w}}$ of the two matrices and use the result $\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\|_F = \|\boldsymbol{w}-\widehat{\boldsymbol{w}}\|_2$. We proceed as follows:

$$\begin{aligned}\|\boldsymbol{w}-\widehat{\boldsymbol{w}}_{T_0}\|_2 =& \left\|\boldsymbol{A}^\dagger(\boldsymbol{d}-\widehat{\boldsymbol{d}})\right\|_2 \\ \leq& \|\boldsymbol{A}^\dagger\|_2\|\boldsymbol{d}-\widehat{\boldsymbol{d}}\|_2 \\ =& \frac{1}{\sigma_{\min}(\boldsymbol{A})}\|\boldsymbol{d}-\widehat{\boldsymbol{d}}\|_2 = \frac{1}{\sigma^2_{\min}(\boldsymbol{O})}\|\boldsymbol{d}-\widehat{\boldsymbol{d}}\|_2,\end{aligned} \tag{P.2}$$

where the first equality follows from Equation 14. In the inequality instead, we used the consistency property for the spectral norm of matrix $\boldsymbol{A}^\dagger$, while in the last equality we used a property of the Kronecker product for which it holds that:

$$\sigma_{\min}(\boldsymbol{A}) = \sigma_{\min}(\boldsymbol{O})\sigma_{\min}(\boldsymbol{O}) = \sigma^2_{\min}(\boldsymbol{O}).$$

Let's now consider the estimation error of each vector $\boldsymbol{d}_{a,a'}$ that represents the stationary distribution over consecutive rewards conditioned on pulling the pair of arms (a,a'). From Equation 10, we know that by concatenating each of these vectors, we obtain the quantity $\boldsymbol{d}$. Thus, by definition, we have:

$$\|\boldsymbol{d}-\widehat{\boldsymbol{d}}\|_2 = \sqrt{\sum_{(a,a')\in\mathbb{I}^2}\|\boldsymbol{d}_{a,a'}-\widehat{\boldsymbol{d}}_{a,a'}\|_2^2}. \tag{P.3}$$

The estimation error of each $\boldsymbol{d}_{a,a'}$ can be bounded by using a result shown in Proposition D.2 and inspired by the work of Hsu et al. (2012). It bounds the estimation error of categorical distributions when the observed samples derive from a Markov chain. With probability at least $1-\delta/I^2$ we have that:

$$\|\boldsymbol{d}_{a,a'} - \widehat{\boldsymbol{d}}_{a,a'}\|_2 \leq \sqrt{\Big(\frac{1+\lambda^{2I^2}}{1-\lambda^{2I^2}}\Big)\frac{1+\log(I^2/\delta)}{N}}.$$

The exponential term $2I^2$ that appears to the modulus of the second highest eigenvalue λ has been introduced thanks to the adoption of the round-robin procedure for the choice of combinations of arms. Notably, each pair of arms is pulled every $2I^2$ steps of the Markov Process, thus resulting in a faster mixing of the chain. For more details, please refer to Appendix B.

By combining the last obtained bound with P.2 and P.3 and using a union bound for the estimation error of all vectors of type $\boldsymbol{d}_{a,a'}$, we have that with probability at least $1-\delta$:

$$\begin{aligned}\|\boldsymbol{w} - \widehat{\boldsymbol{w}}\|_2 \leq& \frac{1}{\sigma^2_{\min}(\boldsymbol{O})}\sqrt{\Big(\frac{1+\lambda^{2I^2}}{1-\lambda^{2I^2}}\Big)\frac{I^2(1+\log(I^2/\delta))}{N}} \\ \leq& \frac{I}{\sigma^2_{\min}(\boldsymbol{O})}\sqrt{\frac{2\big(1+\log(I^2/\delta)\big)}{(1-\lambda^{2I^2})N}}.\end{aligned}$$

Ultimately, by putting together the bound in P.1 with the one just obtained, we are able to obtain the final result stated in the Lemma. □

Having established the results on the estimation matrix $\boldsymbol{P}$, we can now provide regret guarantees for Algorithm 2. We recall that the oracle we use is aware of both the observation model $\boldsymbol{O}$ and the transition model $\boldsymbol{P}$ but does not observe the hidden state. As shown in the definition of the regret in Equation 6, it builds a belief over the states, using the formulation defined in Equation 3 and selects the arm that maximizes the expected instantaneous reward. The derived regret upper bound is provided in the following:

Theorem 6.1. *Suppose Assumptions 4.1 and 4.2 hold and suppose that the Markov chain with transition matrix $\boldsymbol{P}$ starts from its stationary distribution $\boldsymbol{\pi} \in \Delta(\mathbb{S})$. By considering a finite horizon T, there exists a constant T_0, with $T > T_0$, such that with probability at least $1-\delta$, the regret of the SL-EC Algorithm satisfies:*

$$\mathfrak{R}(T) \leq 2\left(\frac{2LI^2}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}}\sqrt{\frac{S(1+\log(I^2/\delta))}{(1-\lambda^{2I^2})}}\cdot T\right)^{2/3}, \tag{16}$$

where $L = \frac{4S(1-\epsilon)^2}{\epsilon^3} + \sqrt{S}$ is a constant that is used to bound the error in the estimated belief (more details in Proposition D.4 in Appendix D). The presented regret has an order of $\mathcal{O}(T^{2/3})$ w.r.t. the horizon T, as common when using an Explore then Commit algorithm. A detailed proof of this theorem can be found in Appendix B. The presented bound on the regret can be achieved by appropriately choosing the exploration horizon T_0. More specifically, we set it as follows:

$$T_0 = \left(\frac{2LTI^2}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}}\sqrt{\frac{S(1+\log(I^2/\delta))}{(1-\lambda^{2I^2})}}\right)^{2/3}. \tag{17}$$

To compute T_0, we need to have information about the minimum value of the stationary distribution $\pi_{\min}$ and about the modulus of the second highest eigenvalue λ. If they are not available, a slightly different version of the bound can be derived so that T_0 can be optimized by only requiring the knowledge of ϵ from Assumption 4.1. More details are reported in Section B.3 of Appendix B.

Dependency on the Number of Arms By analyzing the results on the bound of the regret, we can observe that it scales with $I^{4/3}$. This may seem concerning especially when dealing with problems involving

a high number of arms. However, we have already observed in Section 5.1.1 that when the number of arms is large, it is possible to select a subset of arms that allows solving the problem. In particular, the best subset $\mathbb{J}$ of size J we can select is the one minimizing the term $\frac{J^2}{\sigma^2_{\min}(\boldsymbol{G}_{\mathbb{J}})}$, with $\boldsymbol{G}_{\mathbb{J}}$ being the reduced action observation matrix obtained from the choice of the arms in $\mathbb{J}$. It is indeed likely that when $I \gg S$, some arms contain redundant information and can be easily discarded for the estimation procedure.

Continuous Reward Distributions The presented setting tackles the case of discrete observations. It appears that handling continuous reward distributions within this framework is not feasible and this is true if we apply our approach as is. However, we can discretize the observation distributions and consider the discretized distribution as a categorical one. The process of discretization involves dividing the continuous observation distributions into a predetermined number U of distinct consecutive intervals. Each interval is assigned a probability value that represents the likelihood of a sample drawn from the continuous distribution and belonging to that interval. Throughout this discretization procedure, we can define an action observation matrix of dimension $IU \times S$ and then apply Algorithm 1. More details on this aspect can be found in Appendix E.

7 Numerical Simulations

In this section, we provide numerical simulations on synthetic and semi-synthetic data based on the MovieLens 1M (Harper & Konstan, 2015) dataset, demonstrating the effectiveness of the proposed Markov chain estimation procedure. Specifically, we show the efficiency of the offline arm selection procedure described in Section 5.1.1 and conduct a comparison between our SL-EC Algorithm and several baselines in non-stationary settings. In Section 7.3, we provide additional experiments that highlight the performance difference between our approach and a modified technique based on Spectral Decomposition.

7.1 Estimation Error of Transition Matrix

The first set of experiments is devoted to showing the error incurred by the estimation procedure of the transition matrix in relation to the number of samples considered and the set of actions used for estimation. The left side of Figure 1 illustrates the estimation error of the transition matrix given different instances of Switching Bandits with an increasing number of states. In particular, we fix the number of total actions $I = 10$ and number of observations $V = 10$ and consider three instances with $S = 5$, $S = 10$ and $S = 15$ number of states. As expected, we can see that as the number of states increases, the problem becomes more complex, and more samples are needed to improve the estimation. Figure 1 reports the Frobenius norm $\|\cdot\|_F$ of the error between the true and the estimated transition matrix. We can notice that the estimation procedure is particularly efficient leading to low error values even with a limited number of samples, as can be observed from the steep error drop appearing in the first part of the plot.

The right plot in Figure 1, instead, shows the estimation error obtained by using a different subset of arms. As mentioned in previous sections, it is not always beneficial to use all the available actions during the estimation procedure, but selecting a subset of actions may be preferable. Furthermore, we show that by selecting specific subsets of arms, we can improve the estimation w.r.t using other subsets. For this experiment, we consider $J = 3$ arms among the $I = 8$ available for a Switching MAB instance with $S = 5$ states. We then identify the optimal subset of arms of size J and initiate the estimation process using the selected subset. In order to find the best one, we generate all matrices of type $\boldsymbol{G}$, as described in Section 5.1.1 and choose the matrix with the highest $\sigma_{\min}(\boldsymbol{G})$. The subset of arms generating that matrix will be used for estimation. The estimation error of the best subset of arms is represented in the plot with the red line, while we represent in green the estimation error of the subset having the lowest $\sigma_{\min}(\boldsymbol{G})$. The figure clearly exhibits the performance difference between the two choices, thereby validating our claims. Additional details about the characteristics of the matrices used in the experiments are provided in Appendix A.

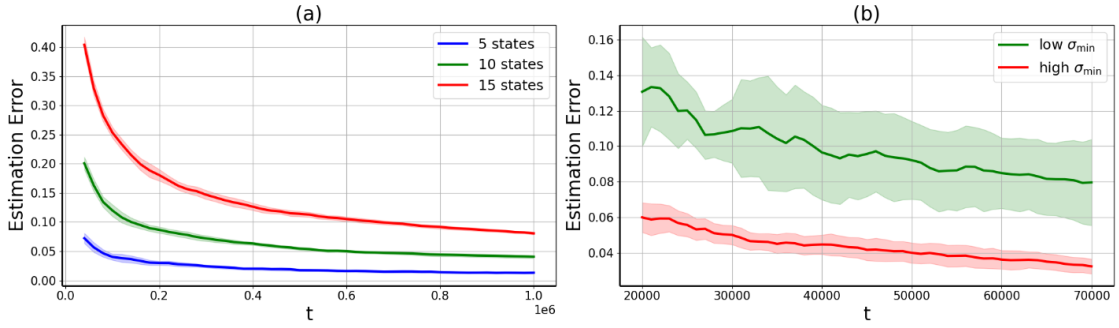

Figure 1: (a) Difference between the estimated and real transition matrix with an increasing number of samples. The metric used is $\|\cdot\|_F$ (10 runs, 95% c.i.), (b) Difference between real and estimated transition matrix using two different subsets of arms of size $J = 3$ arms from the 8 available on a problem with 5 states. The metric used is $\|\cdot\|_F$ (10 runs, 95% c.i.).

7.2 Algorithms Comparisons

In this second set of experiments, we compare the regret suffered by our SL-EC approach with other algorithms specifically designed for non-stationary environments. Following the recent work of Zhou et al. (2021), we consider the subsequent baseline algorithms: the simple *ϵ-greedy* heuristics, a sliding-window algorithm such as *SW-UCB* (Garivier & Moulines, 2011) that is generally able to deal with non-stationary settings and the *Exp3.S* (Auer et al., 2002) algorithm. The parameters for all the baseline algorithms have been properly tuned according to the settings considered. It is worth noting that, unlike our SL-EC algorithm, the baselines do not assume knowledge of the observation model or the underlying Markov chain. In contrast, our approach utilizes the observation model to estimate the transition matrix and to update the belief over the current state. Additionally, we compare our approach with a particle filter algorithm proposed in Hong et al. (2020b) about non-stationary Latent Bandits. They consider two settings: one with complete knowledge of both the observation and transition models and another that incorporates priors on the parameters of the models to account for uncertainty. We compare against a mixture of these two settings by providing their algorithm with full information about the observation model (as it is for our case) and an informative prior about the true transition model. The comparison is made in terms of the empirical cumulative regret $\widehat{\mathfrak{R}}(t)$ averaged over multiple independent runs.

7.2.1 Synthetic Experiments

These experiments have been conducted on various problem configurations with different numbers of states S, actions I, and observations V. The regret results for one configuration are shown in Figure 2(a). From the figure, it is clear that most of the baseline algorithms display a linear time dependence for the regret. This is expected since these algorithms do not take into account the underlying Markov chain that governs the process. The particle filter algorithm, despite being given a good initial prior on the transition model, does not achieve the same performance of SL-EC in the long run. Conversely, we can notice a quite different behavior for our algorithm that, in line with an Explore then Commit approach, initially accumulates a large regret and then experiences a drastic slope change when the exploitation phase begins. The regret shown in each plot is the average over all the runs. For further information regarding the generation of the transition and observation models, as well as the hyperparameters used for the baseline algorithms, we refer the reader to Appendix A.

As a remark, our algorithm outperforms the others when the absolute spectral gap $1 - \lambda$ of the chain has values closer to 1. Indeed, if this is not the case, simple exploration heuristics such as ϵ-greedy would lead to comparable performance. A clear example is when the transition matrix $\boldsymbol{P}$ defining the chain assigns equal probability to all transitions. In this scenario, all states can be considered independent and identically distributed, and we get no advantage from the knowledge of the matrix $\boldsymbol{P}$ over the use of an algorithm such as ϵ-greedy.

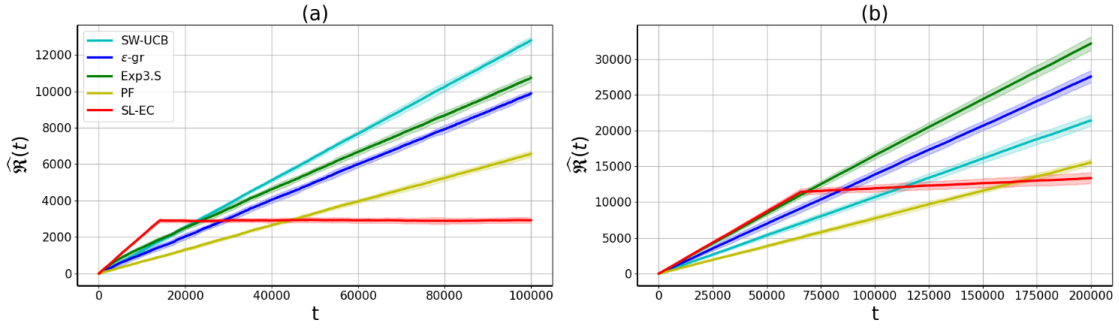

Figure 2: Plots of regret comparing the SL-EC Algorithm with some non-stationary bandit algorithms using: (a) synthetic data with parameters $S = 3$ states, $I = 4$ actions and $V = 5$ observations (5 runs, 95% c.i.); (b) data from MovieLens assuming $S = 5$ states, $I = 18$ actions and $V = 5$ observations. (5 runs, 95% c.i.).

7.2.2 MovieLens Experiments

We also perform some experiments on semi-synthetic data based on MovieLens 1M (Harper & Konstan, 2015), a well-known collaborative filtering dataset where users rate different movies each belonging to a specific set of genres. We adopt a procedure similar to the one used in Hong et al. (2020b). The dataset is initially filtered to include only users who rated at least 100 movies and the movies that have been rated by at least 100 users. After that, we combine the available information in order to obtain a table where each row contains the mean of the ratings for each observed genre for each user (user-genre-rating table). If the user didn't observe any movie belonging to a specific genre, the cell is empty. From the obtained matrix, we select 70% of all ratings as a training dataset and use the remaining 30% as a test set. The sparse matrices so obtained are completed using least-squares matrix completion (Mnih & Salakhutdinov, 2007) using rank 10 and leading to a low prediction error.

Having defined the appropriate rank, we use the predictions on the empty cells of the original *user-genre rating* matrix to fill the entire table. We define a switching bandit instance by using the notion of a *superuser* inspired by Hong et al. (2020b). We use k-means to cluster users using the rows of the user-genre-rating matrix. The users belonging to the same cluster define a superuser that embeds a set of users with similar tastes. The information about the users belonging to the same clusters is then combined and used to generate categorical distributions on the rating, given each superuser and each possible genre (our actions). We choose $k = 5$ for the number of superusers as it is the one that yields clusters with more similar dimensions, and we use $I = 18$ for the actions since it represents the number of identified genres. The number of observations $V = 5$ corresponds to the 5 possible ratings that a movie can get. The transition matrix that governs the dynamics with which superusers alternate is defined by giving higher probabilities to transitions to similar states and also giving higher weights to self-loops in order to avoid too frequent changes. The interaction goes as follows. At each step, a new superuser s_t is sampled based on s_{t-1} and the transition model. The agent chooses an action a_t corresponding to a genre to propose and gets a rating that is sampled from the categorical distribution defined by vector $\boldsymbol{O}\big((a_t, \cdot), s_t\big)$.

As for the synthetic case, our algorithm is compared to other baselines. From Figure 2(b), we can see that our SL-EC still outperforms the other baselines in the considered horizon. However, we highlight that our goal is not to beat the baselines since the comparison is not fair as most of them do not take into account the underlying Markov process, but we aim to show the difference w.r.t. other algorithms belonging to state of the art. More details about the experiments on Movielens are reported in Appendix A.

7.3 Numerical Comparisons with a Modified Spectral Decomposition Technique

The focus of this last set of experiments is to show the difference between a modified Spectral Decomposition (SD) technique and our estimation approach detailed in Algorithm 1. Among the various applications, SD techniques are typically used for learning with Hidden Markov Models (HMM) where no information about

Table 1: Comparison with Nearly Deterministic Models

2 States	3K samples	6K samples	9K samples	15K samples
SD O	0.0493 (0.0097)	0.0379 (0.0103)	0.0335 (0.0097)	0.0259 (0.0081)
SD T	0.0342 (0.0185)	**0.0189** (0.0097)	0.0149 (0.0032)	0.0101 (0.007)
Alg. 1	**0.0234** (0.015)	0.02 (0.0203)	**0.0119** (0.009)	**0.008** (0.0032)
3 States	150K samples	300K samples	600K samples	900K samples
SD O	0.0165 (0.0044)	0.0113 (0.0036)	0.0097 (0.0033)	0.0085 (0.0018)
SD T	0.1547 (0.0517)	0.154 (0.0532)	0.1544 (0.0534)	0.1541 (0.0532)
Alg. 1	**0.0066** (0.0026)	**0.0046** (0.0012)	**0.0037** (0.0018)	**0.0031** (0.0008)
5 States	150K samples	300K samples	600K samples	900K samples
SD O	0.0681 (0.0178)	0.0513 (0.0111)	0.0354 (0.0127)	0.0283 (0.0082)
SD T	0.2409 (0.0633)	0.2484 (0.0584)	0.243 (0.0603)	0.2407 (0.0601)
Alg. 1	**0.0283** (0.0054)	**0.0195** (0.0036)	**0.0137** (0.0033)	**0.0115** (0.0034)

the observation and transition model is provided. Zhou et al. (2021) make use of these techniques to get an estimation of both the observation and the transition model. It is important to highlight that SD methods are hardly used in practice because of their computational and sample complexity. Indeed, both the related works of Zhou et al. (2021) and Azizzadenesheli et al. (2016) include only proof-of-concept experiments with 2 hidden states and 2 possible actions.

To make the comparison fairer, we consider a modified SD technique that is provided with information about the observation model in order to help the estimation process, as will be briefly explained. The original SD technique to which we refer follows the procedures highlighted in Anandkumar et al. (2014) for HMM and makes use of the Robust Tensor Power (RTP) method for orthogonal tensor decomposition. In typical SD techniques, data is collected by sampling an action at each time step and updating the computed statistics with the observed realization. With the presented modified SD technique, at each step, we do not simply update the statistics with the observation obtained when pulling the arm, but we give information about the observation distribution for all the available arms, with this information being conditioned on the underlying current state. In this way, it is like pulling all the arms at each step and receiving full information about their associated reward distributions, given the underlying state.

We perform various experiments by fixing the number of arms ($I = 20$) and the number of possible rewards ($V = 5$) for each arm and by changing the number of states. Each experiment is performed over 10 different runs, where for each run a transition and observation model are generated. For each experiment, our estimation procedure uses 3 arms among the 20 available, which are selected using the offline arms selection strategy. The transition and observation matrices are created in two different ways: the former focuses on nearly-deterministic matrices (Table 1), while the latter considers more stochasticity for both of them (Table 2).

The results of the experiments are structured in the following way. Each of the two Tables contains mini-tables representing sets of experiments characterized by different number of states. By fixing the number of states for the experiments, each mini-table shows three rows: the first one (indicated with *SD O*) contains the Frobenius norm of the estimation error of the observation matrix with the modified SD technique, the second row (indicated with *SD T*) represents the Frobenius norm of the estimation error of the transition matrix with the modified SD technique, while the third row represents the Frobenius norm of the estimation error of the transition matrix computed with Algorithm 1. For each experiment, we report the mean error over the 10 runs and one standard deviation between parenthesis. The modified SD technique clearly enhances the accuracy of estimating the observation model compared to standard SD approaches: this aspect is evident from the relatively low estimation errors observed in the *SD O* rows. We present this information to illustrate that the comparison between our estimation procedure and SD approaches is now more fair due to the modified SD technique employed. Having clarified this aspect, we focus on the estimation error of the transition model between the two different methods: this information is indeed separated from *SD O* by a dashed line. We show the experiments with lower estimation errors in bold.

Table 2: Comparison with Higher Model Stochasticity

2 States	150K samples	210K samples	270K samples
SD O	0.1500 (0.2639)	0.1411 (0.2741)	0.1455 (0.2665)
SD T	0.1488 (0.1536)	0.1699 (0.1742)	0.1576 (0.1702)
Alg. 1	**0.0145** (0.0175)	**0.0145** (0.0134)	**0.0125** (0.0103)
3 States	300K samples	600K samples	900K samples
SD O	0.2987 (0.2128)	0.3078 (0.2177)	0.2594 (0.2309)
SD T	0.3916 (0.2804)	0.4425 (0.2637)	0.4187 (0.2728)
Alg. 1	**0.0077** (0.003)	**0.0063** (0.0023)	**0.0052** (0.002)

The results of this first set of experiments are reported in Table 1. As already anticipated, both the observation and transition matrices are almost deterministic, hence having high probability on a specific observation/state and low probabilities for all the others. For transition matrices, the highest probability is assigned to the probability of staying in the same state. Near-determinism is defined to simplify the problem by making states more distinguishable. By inspecting the results, it is clear that Algorithm 1 outperforms the modified SD technique in almost all the scenarios. Comparable results are only achieved in the experiment with 2 states.

Table 2 reports instead the experimental results obtained using both transition and observation matrices with less peaked distributions, thus higher stochasticity. The discrepancy between our approach and the modified SD technique is more evident in this scenario. This aspect can be justified by the theoretical comparison reported in Appendix C, where it can be observed that, compared to our estimation approach, SD techniques have a dependency of higher order on the minimum singular values of both the observation and the transition models. Thus, when the observation matrix is more stochastic, its $\sigma_{\min}(\boldsymbol{O})$ typically decreases, and this aspect results in a higher estimation error. Indeed, it can be noticed that the estimation error is significant and the number of samples required to perform this set of experiments is much higher than that used for the nearly-deterministic case. Experiments involving a higher number of states instead were not able to reach convergence with a number of samples of the order 10^5 and, by trying to increase this quantity, there were memory space problems with the used hardware (Intel i7-11th and 16G RAM).

Again, we would like to emphasize that SD techniques are explicitly meant to work in a different setting, intrinsically more complex, where no information about either the transition or the observation model is provided. However, with this set of experiments, we wanted to show that if, instead, we have knowledge about the observation model, directly using this information in the SD techniques does not lead to performances comparable to our approach.

8 Discussion and Conclusions

This paper studies a Latent Bandit problem with latent states changing in time according to an underlying unknown Markov chain. Each state is represented by a different Bandit instance that is unobserved by the agent. As common in the latent Bandit literature, we assumed to know the observation model relating each MAB to the reward distribution of its actions, and by using some mild assumptions, we presented a novel estimation technique using the information derived from consecutive pulls of pairs of arms. As far as we know, we are the first to present an estimation procedure of this type aiming at directly estimating the stationary distribution $\boldsymbol{w}$ of consecutive states. The approach is easy to use and does not require specific hyperparameters to be set. We provided an offline arm selection that selects the best subsets of arms to speed up the estimation process. We analyzed the dependence of the parameters on the complexity of the problem, and we showed how our estimation approach can be extended to handle models with continuous observation distributions. We used the presented technique in our SL-EC algorithm that uses an Explore then Commit approach and for which we proved a $\mathcal{O}(T^{2/3})$ regret bound. The experimental evaluation confirmed our theoretical findings showing advantages over some baseline algorithms designed for non-stationary MABs and showing good estimation performances even in scenarios with larger problems. Furthermore, we compared

our approach both empirically and theoretically (Appendix C) with SD techniques, taking into account the differences between the two procedures.

We identify different future research directions for the presented work, such as designing new algorithms that are able to exploit the flexibility in the exploration policy determined by the defined procedure, allegedly in an optimistic way. It may also be interesting to deepen the understanding of this problem when dealing with continuous reward models, trying to design optimal ways to discretize them in order to reach faster estimation performances. We could also consider the extension to the continuous state space setting (e.g., linearMDPs): among the main challenges in this scenario, we consider the adoption of a different representation for the reference matrix that would otherwise not be computable with infinite states and the redefinition of the stationary distribution over consecutive states. In such a case, it might be beneficial to estimate the feature functions directly by means of which the linear MDP is defined. Finally, it might be worth considering a contextual version of the proposed setting. According to the assumptions made, for example, whether the context is discrete or continuous or whether it is related or not to the latent state, this aspect may bring another dimension to the observation space. Redefining the reference matrix by taking this feature into account will likely lead to more informative components and help with the estimation process.

References

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, jan 2014.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Peter Auer, Pratik Gajane, and Ronald Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 138–158, 2019.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Anima Anandkumar. Reinforcement learning of pomdps using spectral methods. In *Annual Conference Computational Learning Theory*, 2016.

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

Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie. Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Yohann De Castro, Élisabeth Gassiat, and Sylvain Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden markov models. *IEEE Transactions on Information Theory*, 63(8):4758–4777, 2017. doi: 10.1109/TIT.2017.2696959.

Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's inequality for general markov chains and its applications to statistical learning. *Journal of Machine Learning Research*, 22(139):1–35, 2021.

William Feller. *An Introduction to Probability Theory and its Applications Vol. I.* Wiley, 1968.

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

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory*, pp. 174–188, 2011.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

Jiaxing Guo, Qian Sang, and Niklas Karlsson. Adaptive seasonality estimation for campaign optimization in online advertising. In *2021 American Control Conference (ACC)*, pp. 1450–1455, 2021.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Steven L. Heston and Ronnie Sadka. Seasonality in the cross-section of stock returns. *Journal of Financial Economics*, 87(2):418–445, 2008. ISSN 0304-405X.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, and Craig Boutilier. Latent bandits revisited. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13423–13433. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9b7c8d13e4b2f08895fb7bcead930b46-Paper.pdf.

Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, Amr Ahmed, Mohammad Ghavamzadeh, and Craig Boutilier. Non-stationary latent bandits. *CoRR*, abs/2012.00386, 2020b.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012. ISSN 0022-0000. JCSS Special Issue: Cloud Computing 2011.

Mehdi Jafarnia Jahromi, Rahul Jain, and Ashutosh Nayyar. Online learning for unknown partially observable mdps. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

Bowen Jiang, Bo Jiang, Jian Li, Tao Lin, Xinbing Wang, and Chenghu Zhou. Online restless bandits with unobserved states. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Chi Jin, Sham M. Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Vikram Krishnamurthy. *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambridge University Press, 2016. doi: 10.1017/CBO9781316471104.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Tractable optimality in episodic latent mabs. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23634–23645. Curran Associates, Inc., 2022.

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

Fang Liu, Joohyung Lee, and Ness B. Shroff. A change-detection based framework for piecewise-stationary multi-armed bandit problem. In *AAAI Conference on Artificial Intelligence*, 2017.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary?, 2022.

Charles F.Van Loan. The ubiquitous kronecker product. *Journal of Computational and Applied Mathematics*, 123(1):85–100, 2000. ISSN 0377-0427. doi: https://doi.org/10.1016/S0377-0427(00)00393-9. Numerical Analysis 2000. Vol. III: Linear Algebra.

Krishnadas M., K.P. Harikrishnan, and G. Ambika. Recurrence measures and transitions in stock market dynamics. *Physica A: Statistical Mechanics and its Applications*, 608:128240, 2022. ISSN 0378-4371.

Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. *31st International Conference on Machine Learning, ICML 2014*, 1, 05 2014.

Robert Mattila, Cristian Rojas, Eric Moulines, Vikram Krishnamurthy, and Bo Wahlberg. Fast and consistent learning of hidden Markov models by incorporating non-consecutive correlations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6785–6796. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/mattila20a.html.

Colin McDiarmid. *On the method of bounded differences*, pp. 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/d7322ed717dedf1eb4e6e52a37ea7bcd-Paper.pdf.

Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless markov bandits. *Theoretical Computer Science*, 558:62–76, 2014. ISSN 0304-3975. Algorithmic Learning Theory.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley; Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

Giorgia Ramponi, Amarildo Likmeta, Alberto Maria Metelli, Andrea Tirinzoni, and Marcello Restelli. Truly batch model-free inverse reinforcement learning about multiple intentions. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, 2020.

Aleksandrs Slivkins and Eli Upfal. Adapting to a changing environment: the brownian restless bandits. In *Annual Conference Computational Learning Theory*, 2008.

Francesco Trovò, Stefano Paladino, Marcello Restelli, and Nicola Gatti. Sliding-window thompson sampling for non-stationary settings. *Journal of Artificial Intelligence Research*, 68:311–364, 05 2020. doi: 10.1613/jair.1.11407.

Yi Xiong, Ningyuan Chen, Xuefeng Gao, and Xiang Zhou. Sublinear regret for learning pomdps, 2022.

Li Zhou and Emma Brunskill. Latent contextual bandits and their application to personalized recommendations for new users. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pp. 3646–3653. AAAI Press, 2016. ISBN 9781577357704.

Xiang Zhou, Yi Xiong, Ningyuan Chen, and Xuefeng Gao. Regime switching bandits. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

A Additional Simulation Details

In this section, we are going to give some details on the characteristics of the experiments reported in the main paper.

A.1 Estimation Error of Transition Matrix

For the experiments related to the estimation error of the transition matrix in Figure 1, we generated a set of transition and observation matrices with the following characteristics.

- for the plot on the left, we fixed the number $I = 10$ of possible actions and $V = 10$ of finite observations. We then consider the estimation procedure for problems of different sizes with respectively $S = 5$, $S = 10$ and $S = 15$ number of states;
- for the plot on the right, the considered estimated problem has $S = 5$ states, $I = 8$ possible actions, $V = 10$ finite observations.

Starting from the presented parameters, the transition and observation matrices have been generated as follows. An initial version of transition and observation matrices is generated with random elements and, subsequently:

- regarding the transition matrix, we add a tuned diagonal matrix to the initial random version and then normalize. In this way, we give more probability on self transitions;
- regarding the observation matrix, for each pair of states and actions, we choose a specific reward that will be drawn with higher probability, in order to avoid having too much stochasticity in the reward distributions.

The scheme just presented is also used for the generation of matrices in the experiments showing the regret of the different algorithms.
For the experiments in the plot on the right, let's denote with $\boldsymbol{G}_g$ and $\boldsymbol{G}_r$ the reduced action observation matrices containing the subset of arms for the green and the red lines respectively. Their values are $\sigma_{\min}(\boldsymbol{G}_g) \approx 0.14$ and $\sigma_{\min}(\boldsymbol{G}_r) \approx 0.27$.

A.2 Algorithms Comparisons

A.2.1 Synthetic Experiments

For this set of experiments, the parameters used for the generation of the transition and observation matrix are $S = 3$ states, $I = 4$ possible actions and $V = 5$ finite observations. The generation of the matrices is not completely random and follows the same procedure explained in the previous paragraph for the experiment on the matrix estimation error. For the specific experiments considered, we adopted scaled values for the exploration horizon T_0 w.r.t. the result derived from the theory. However, despite a reduced number of samples, the estimation still presents good performances. For the plots shown in the main paper, the hyperparameters used are $\epsilon = 0.05$ for the ϵ-greedy approach, a value of $L_w = 1000$ for the sliding-window UCB algorithm, and the suggested value $1/T$ for the α parameter in the Exp3.S algorithm. For the particle filter algorithm, we used 100 different particles and a resampling threshold of 25 for the *Effective Sample Size*.

A.2.2 Movielens Experiments

In this section, we give more details about the experiments on the Movielens 1M dataset. As detailed in the main body of the paper the transition matrix is constructed taking into account the similarity between superusers. In order to do that we use the cosine similarity to define the initial transition matrix. After

that, we add a diagonal matrix to the previously obtained matrix in order to give higher values to self-loop probabilities. The final transition matrix is obtained by normalizing along each row. The self-loop probabilities are close to the value 0.98 thus representing a more realistic scenario with regimes not switching too often. The rewards are obtained by scaling the ratings obtained from a single movie in the range $[0, 1]$. For the plots shown in the main paper, the hyperparameters used are $\epsilon = 0.05$ for the ϵ-greedy approach, $L_w = 300$ for the sliding-window UCB algorithm, and the suggested value $1/T$ for the α parameter in the Exp3.S algorithm. For the particle filter algorithm, we used 300 different particles and a resampling threshold of 50 for the *Effective Sample Size*. Our SL-EC algorithm has been run using an offline arm selection procedure for choosing the arms to use during the exploration phase. The number of selected arms has been fixed to $J = 5$.

B Theoretical Results

In this Section, we will provide the proofs of Lemma 6.1 (Section B.1) and Theorem 6.1 (Section B.2) presented in the main paper. Finally, Section B.3 shows how to compute the exploration horizon T_0 when only the minimum value of the transition matrix ϵ is known.

B.1 Proof of Lemma 6.1

We will start by reporting Lemma 6.1 of the main paper and its proof.

Lemma 6.1. *Suppose Assumption 4.2 holds and suppose that the Markov chain with transition matrix $\boldsymbol{P}$ is ergodic, such that $\pi_{\min} := \min_{s\in\mathbb{S}} \boldsymbol{\pi}(s) > 0$ with $\boldsymbol{\pi} \in \Delta(\mathbb{S})$ being the stationary distribution of the chain. By assuming that the chain starts from its stationary distribution $\boldsymbol{\pi}$ and by pulling each pair of arms in a round-robin fashion for N rounds and using the estimation procedure reported in Algorithm 1, we have that with probability at least $1-\delta$ the estimation error of the transition matrix $\boldsymbol{P}$ will be:*

$$\|\boldsymbol{P} - \widehat{\boldsymbol{P}}\|_F \leq \frac{2I}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}} \sqrt{\frac{2S(1+\log(I^2/\delta))}{(1-\lambda^{2I^2})N}}, \tag{15}$$

Proof. The proof of the presented bound can be decomposed into two main parts. On one side, we can define the bound of the estimation error of the matrix $\boldsymbol{W}$ and, secondly, the error of the transition matrix $\boldsymbol{P}$ that derives from the estimation error of $\boldsymbol{W}$. We first will tackle this last part:

$$\begin{aligned}
\|\boldsymbol{P} - \widehat{\boldsymbol{P}}\|_F &= \sqrt{\sum_{(s,s')\in\mathbb{S}^2} (\boldsymbol{P}(s,s') - \widehat{\boldsymbol{P}}(s,s'))^2} = \sqrt{\sum_{s\in\mathbb{S}} \|\boldsymbol{P}(s,\cdot) - \widehat{\boldsymbol{P}}(s,\cdot)\|_2^2} \\
&= \sqrt{\sum_{s\in\mathbb{S}} \left\| \frac{\boldsymbol{W}(s,\cdot)}{\|\boldsymbol{W}(s,\cdot)\|_1} - \frac{\widehat{\boldsymbol{W}}(s,\cdot)}{\|\widehat{\boldsymbol{W}}(s,\cdot)\|_1} \right\|_2^2} \\
&\leq \sqrt{\sum_{s\in\mathbb{S}} \left\| \frac{\boldsymbol{W}(s,\cdot)}{\|\boldsymbol{W}(s,\cdot)\|_2} - \frac{\widehat{\boldsymbol{W}}(s,\cdot)}{\|\widehat{\boldsymbol{W}}(s,\cdot)\|_2} \right\|_2^2} \qquad \text{(P.4)} \\
&\leq \sqrt{\sum_{s\in\mathbb{S}} \frac{4\|\boldsymbol{W}(s,\cdot) - \widehat{\boldsymbol{W}}(s,\cdot)\|_2^2}{\max\{\|\boldsymbol{W}(s,\cdot)\|_2, \|\widehat{\boldsymbol{W}}(s,\cdot)\|_2\}^2}} \qquad \text{(P.5)} \\
&\leq \sqrt{\sum_{s\in\mathbb{S}} \frac{4\|\boldsymbol{W}(s,\cdot) - \widehat{\boldsymbol{W}}(s,\cdot)\|_2^2}{\|\boldsymbol{W}(s,\cdot)\|_2^2}} \qquad \text{(P.6)}
\end{aligned}$$

$$\leq \sqrt{\sum_{s\in\mathbb{S}} \frac{4S\|\boldsymbol{W}(s,\cdot)-\widehat{\boldsymbol{W}}(s,\cdot)\|_2^2}{\pi_{\min}^2}} \tag{P.7}$$

$$= \sqrt{\frac{4S\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\|_F^2}{\pi_{\min}^2}} = \frac{2\sqrt{S}\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\|_F}{\pi_{\min}}. \tag{P.8}$$

We recall that $\boldsymbol{P}(s,\cdot) \in \Delta(\mathbb{S})$ denotes the distribution over the next state when starting from state $s \in \mathbb{S}$ and we denote with $\boldsymbol{W}(s,\cdot)$ the vector of dimension S representing the row of the matrix $\boldsymbol{W}$ of the stationary distribution of consecutive states. Line P.4 derives from the fact that $\|\boldsymbol{W}(s,\cdot)\|_1 \geq \|\boldsymbol{W}(s,\cdot)\|_2$, while line P.5 is obtained by using Lemma D.1. Line P.6 easily follows from

$$\max\{\|\boldsymbol{W}(s,\cdot)\|_2, \|\widehat{\boldsymbol{W}}(s,\cdot)\|_2\} \geq \|\boldsymbol{W}(s,\cdot)\|_2.$$

Line P.7 is instead derived from the following considerations. For any vector $\boldsymbol{W}(s,\cdot)$ of size S, it holds that:

$$\|\boldsymbol{W}(s,\cdot)\|_2^2 = \sum_{s'\in\mathbb{S}} \boldsymbol{W}(s,s')^2 \geq \frac{1}{S}\Big(\sum_{s'\in\mathbb{S}} \boldsymbol{W}(s,s')\Big)^2 = \frac{1}{S}\boldsymbol{\pi}(s)^2 \geq \frac{\pi_{\min}^2}{S},$$

where the first inequality in the expression above follows from the fact that $\sqrt{Y}\|\boldsymbol{y}\|_2 \geq \|\boldsymbol{y}\|_1 \ \forall \boldsymbol{y} \in \mathbb{R}^Y$. The second equality is instead derived from the definition of matrix $\boldsymbol{W}$, since the sum of the elements along the row associated with state s corresponds to the probability value $\boldsymbol{\pi}(s)$ of state s from the stationary distribution $\boldsymbol{\pi}$ induced by the chain. In the last inequality, we bound each of these probabilities by $\pi_{\min} = \min_{s\in\mathbb{S}} \boldsymbol{\pi}(s)$.
Finally, the first equality in line P.8 derives from $\sum_{s\in\mathbb{S}} \|\boldsymbol{W}(s,\cdot) - \widehat{\boldsymbol{W}}(s,\cdot)\|_2^2 = \|\boldsymbol{W} - \widehat{\boldsymbol{W}}\|_F^2$, which holds by the definition of the used quantities.

We will now derive the first part of the proof by defining a high probability bound on the estimation error of the stationary distribution over consecutive states represented by matrix $\boldsymbol{W}$. In order to do that, we use the relation $\|\boldsymbol{W}-\widehat{\boldsymbol{W}}\|_F = \|\boldsymbol{w}-\widehat{\boldsymbol{w}}\|_2$, where we recall that $\boldsymbol{w}$ is obtained by the vectorization of matrix $\boldsymbol{W}$. The bound is obtained assuming that each pair of arms is pulled in a round-robin fashion, as required by the algorithm. The derivation is as follows:

$$\begin{aligned}
\|\boldsymbol{w}-\widehat{\boldsymbol{w}}\|_2 &= \Big\|\boldsymbol{A}^\dagger(\boldsymbol{d}-\widehat{\boldsymbol{d}})\Big\|_2 \\
&\leq \|\boldsymbol{A}^\dagger\|_2\|\boldsymbol{d}-\widehat{\boldsymbol{d}}\|_2 && \text{(P.9)} \\
&= \frac{1}{\sigma_{\min}(\boldsymbol{A})}\|\boldsymbol{d}-\widehat{\boldsymbol{d}}\|_2 = \frac{1}{\sigma_{\min}^2(\boldsymbol{O})}\|\boldsymbol{d}-\widehat{\boldsymbol{d}}\|_2 && \text{(P.10)}
\end{aligned}$$

where vector $\widehat{\boldsymbol{d}}$ is obtained from the vector count $\boldsymbol{n}$ by assuming that each pair of arms has been pulled N times (refer to Equation 13). The first inequality is obtained by the consistency property of matrices: the first norm represents the spectral norm of matrix $\boldsymbol{A}^\dagger$, while the second is a $\|\cdot\|_2$ of a vector. The last equality is instead obtained from the properties of the Kronecker product for which it holds that:

$$\sigma_{\min}(\boldsymbol{A}) = \sigma_{\min}(\boldsymbol{O})\sigma_{\min}(\boldsymbol{O}) = \sigma_{\min}^2(\boldsymbol{O}). \tag{P.11}$$

Let's consider now the estimation error of the different vectors $\boldsymbol{d}_{a,a'}$ constituting vector $\boldsymbol{d}$, as defined in Equation 10. We have seen in the main paper that each vector $\boldsymbol{d}_{a,a'}$ denotes the distribution over consecutive rewards conditioned on pulling the consecutive arms (a,a') (see Equation 9).
Since the pairs of rewards are not i.i.d but depend on the underlying Markov process, we can use a suited concentration result for discrete distributions that appears in Proposition D.2.
In particular, from Proposition D.2, it holds that $\forall (a,a') \in \mathbb{I}^2$, with probability at least $1-\delta/I^2$:

$$\|\boldsymbol{d}_{a,a'} - \widehat{\boldsymbol{d}}_{a,a'}\|_2 \leq \sqrt{\Big(\frac{1+\lambda}{1-\lambda}\Big)\frac{1+\log(I^2/\delta)}{N}},$$

with N being the number of samples used for the estimation. Considering the round-robin procedure used in the exploration phase, the result just presented can be slightly improved. Indeed, being $\boldsymbol{P}$ the transition matrix governing

the chain and λ the modulus of its second largest eigenvalue, since samples used to estimate each $\widehat{\boldsymbol{d}}_{a,a'}$ are collected every $2I^2$ time instants, this leads to an effective transition matrix $\boldsymbol{P}^{2I^2}$, with associated λ^{2I^2} modulus of the second largest eigenvalue. This collection policy has thus the consequence of inducing less temporal dependence among samples. We can modify the previous result by having $\forall (a,a') \in \mathbb{I}^2$, with probability at least $1-\delta/I^2$:

$$\|\boldsymbol{d}_{a,a'} - \widehat{\boldsymbol{d}}_{a,a'}\|_2 \leq \sqrt{\Big(\frac{1+\lambda^{2I^2}}{1-\lambda^{2I^2}}\Big)\frac{1+\log(I^2/\delta)}{N}}. \tag{P.12}$$

We can now express the following relation that easily follows from the definition of the quantities involved:

$$\|\boldsymbol{d} - \widehat{\boldsymbol{d}}\|_2 = \sqrt{\sum_{(a,a')\in\mathbb{I}^2} \|\boldsymbol{d}_{a,a'} - \widehat{\boldsymbol{d}}_{a,a'}\|_2^2}. \tag{P.13}$$

By combining the results in P.10, P.12, P.13, and using a union bound, we have that with probability at least $1-\delta$:

$$\begin{aligned}\|\boldsymbol{w} - \widehat{\boldsymbol{w}}\|_2 \leq & \frac{1}{\sigma^2_{\min}(\boldsymbol{O})}\sqrt{\Big(\frac{1+\lambda^{2I^2}}{1-\lambda^{2I^2}}\Big)\frac{I^2(1+\log(I^2/\delta))}{N}} \\ \leq & \frac{I}{\sigma^2_{\min}(\boldsymbol{O})}\sqrt{\frac{2\big(1+\log(I^2/\delta)\big)}{(1-\lambda^{2I^2})N}}.\end{aligned} \tag{P.14}$$

We have now all the elements needed to show the final result. Putting together P.8 and P.14, we have that with probability at least $1-\delta$, it holds that:

$$\|\boldsymbol{P} - \widehat{\boldsymbol{P}}\|_F \leq \frac{2I}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}}\sqrt{\frac{2S(1+\log(I^2/\delta))}{(1-\lambda^{2I^2})N}},$$

which concludes the proof. □

B.2 Proof of Theorem 6.1

We are now ready to derive the main result related to the regret of the SL-EC Algorithm. We will report here Theorem 6.1 of the main paper.

Theorem 6.1. *Suppose Assumptions 4.1 and 4.2 hold and suppose that the Markov chain with transition matrix $\boldsymbol{P}$ starts from its stationary distribution $\boldsymbol{\pi} \in \Delta(\mathbb{S})$. By considering a finite horizon T, there exists a constant T_0, with $T > T_0$, such that with probability at least $1-\delta$, the regret of the SL-EC Algorithm satisfies:*

$$\mathfrak{R}(T) \leq 2\left(\frac{2LI^2}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}}\sqrt{\frac{S(1+\log(I^2/\delta))}{(1-\lambda^{2I^2})}}\cdot T\right)^{2/3}, \tag{16}$$

Proof. The proof of the regret of the SL-EC Algorithm makes use of some of the results previously derived and it can be divided into the regret from the exploration and regret from the exploitation phase.
Considering an exploration phase of length T_0, the regret initially suffered can be trivially bounded as:

$$\mathfrak{R}_{1:T_0} = \sum_{t=1}^{T_0} \max_{a\in\mathbb{I}} \langle \boldsymbol{\mu}(a), \mathbf{b}_t\rangle - \mathrm{r}_t \leq \sum_{i=1}^{T_0} 1 = T_0, \tag{P.15}$$

where vector $\boldsymbol{\mu}(a)$ has dimension S and its elements are defined in Equation 4.
For the exploitation phase, we compute a belief vector $\widehat{\mathbf{b}}_t$ at each step according to Equation 5 using the estimate $\widehat{\boldsymbol{P}}$ of the transition matrix. The belief vector is initialized uniformly over the states and updated starting from the initial samples up to those collected at the end of the exploration phase. The analysis of the regret in this part is as

follows:

$$\begin{aligned}
\mathfrak{R}_{T_0:T} &= \sum_{t=T_0+1}^{T} \max_{a\in\mathbb{I}} \langle \boldsymbol{\mu}(a), \mathbf{b}_t\rangle - \max_{a\in\mathbb{I}} \langle \boldsymbol{\mu}(a), \widehat{\mathbf{b}}_t\rangle \\
&\leq \sum_{t=T_0+1}^{T} \max_{a\in\mathbb{I}} |\langle \boldsymbol{\mu}(a), \mathbf{b}_t - \widehat{\mathbf{b}}_t\rangle| \\
&\leq \sum_{t=T_0+1}^{T} \|\boldsymbol{\mu}(a)\|_\infty \|\mathbf{b}_t - \widehat{\mathbf{b}}_t\|_1 \\
&\leq \sum_{t=T_0+1}^{T} \|\mathbf{b}_t - \widehat{\mathbf{b}}_t\|_1 \\
&\leq \sum_{t=T_0+1}^{T} L\|\boldsymbol{P} - \widehat{\boldsymbol{P}}_{T_0}\|_F \\
&\leq \frac{4LTI^2}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}} \sqrt{\frac{S(1+\log(I^2/\delta))}{(1-\lambda^{2I^2})T_0}},
\end{aligned} \tag{P.16}$$

where in the second inequality we applied Hölder's inequality with norms ∞ and 1, while the third inequality is obtained from $\|\boldsymbol{\mu}(a)\|_\infty \leq 1 \; \forall a \in \mathbb{I}$. The fourth inequality is obtained by applying Proposition D.4, while the last inequality uses the concentration derived in Lemma 6.1, considering that the number of rounds is $N = T_0/(2I^2)$ if the exploration horizon has length T_0.

Combining together the regrets of the two phases derived in Equation P.15 and in Equation P.16 we have:

$$\mathfrak{R}(T) \leq T_0 + \frac{4LTI^2}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}} \sqrt{\frac{S(1+\log(I^2/\delta))}{(1-\lambda^{2I^2})T_0}}. \tag{P.17}$$

We can now optimize this bound w.r.t. the exploration length T_0 by vanishing the derivative of the right-hand side of Equation P.17. What we get is the following term:

$$T_0 = \left(\frac{2LTI^2}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}} \sqrt{\frac{S(1+\log(I^2/\delta))}{1-\lambda^{2I^2}}} \right)^{2/3}.$$

By substituting this value of T_0 into Equation P.17, we get the result of the Theorem. □

B.3 Optimization of T_0

In order to be able to compute T_0, we need to have information about the minimum value of the stationary distribution $\pi_{\min}$ and about the modulus of the second highest eigenvalue λ. If they are not available, a slightly different version of the bound can be derived in order to compute T_0 by only using the knowledge of ϵ from Assumption 4.1.

Since Theorem 6.1 requires Assumption 4.1 to hold, which is stronger than the ergodicity of the chain, we can further characterize the result obtained in Lemma 6.1. In particular, we can lower bound the minimum probability of the stationary distribution $\pi_{\min}$ with the minimum quantity ϵ appearing in the transition model $\boldsymbol{P}$, such that:

$$\pi_{\min} \geq \epsilon. \tag{18}$$

A second step involves instead characterizing the value of λ as a function of ϵ. To do that, we resort to a quantity $\rho(\cdot)$ known as Dobrushin coefficient (Krishnamurthy, 2016) that defines the rate of convergence of ergodic chains towards their stationary distribution. Given a stochastic matrix $\boldsymbol{T}$ representing the dynamics of a Markov chain defined on a finite state space $\mathbb{X}$ and given two probability vectors in $\nu, \bar{\nu} \in \Delta(\mathbb{X})$, the Dobrushin coefficient satisfies the following relation:

$$\|\boldsymbol{T}^\top \nu - \boldsymbol{T}^\top \bar{\nu}\|_1 \leq \rho(\boldsymbol{T})\|\nu - \bar{\nu}\|_1,$$

where $\boldsymbol{T}^\top$ represents the transpose of matrix $\boldsymbol{T}$. The inequality above says that the one-step evolution of two probability vectors induced by the same transition matrix $\boldsymbol{T}$ can be bounded by the quantity on the right where scalar $\rho(\boldsymbol{T}) \in [0, 1]$ represents the Dobrushin coefficient. For ergodic chains, this coefficient is always strictly smaller than 1. Hence, by iteratively applying the inequality, it is possible to ensure geometric convergence of the initial distance.
Among the properties of the Dobrushin coefficient (Krishnamurthy, 2016), we have that it is an upper bound to the modulus of the second largest eigenvalue λ, thus:

$$\lambda(\boldsymbol{T}) \leq \rho(\boldsymbol{T}), \tag{19}$$

where here we use $\lambda(\boldsymbol{T})$ to denote the modulus of the second largest eigenvalue of matrix $\boldsymbol{T}$. Furthermore, the Dobrushin coefficient corresponds to the following quantity:

$$\rho(\boldsymbol{T}) = 1 - \min_{i,j \in \mathbb{X}} \sum_{l \in \mathbb{X}} \min\{\boldsymbol{T}(i,l), \boldsymbol{T}(j,l)\}. \tag{20}$$

For a Markov chain with transition matrix $\boldsymbol{P}$ satisfying Assumption 4.1, we can set an upper bound to the Dobrushin coefficient as follows:

$$\rho(\boldsymbol{P}) = 1 - \min_{i,j \in \mathbb{S}} \sum_{l \in \mathbb{S}} \min\{\boldsymbol{P}(i,l), \boldsymbol{P}(j,l)\} \tag{21}$$

$$\leq 1 - \sum_{l \in \mathbb{S}} \epsilon = 1 - S\epsilon \tag{22}$$

where the inequality derives indeed from Assumption 4.1.
Using now 19, we have that:

$$\lambda(\boldsymbol{P}) \leq \rho(\boldsymbol{P}) \leq 1 - S\epsilon. \tag{23}$$

The bound on the regret of Theorem 6.1 can now be rewritten as:

$$\mathfrak{R}(T) \leq T_0 + \frac{4LTI^2}{\sigma^2_{\min}(\boldsymbol{O})\epsilon} \sqrt{\frac{S(1 + \log(I^2/\delta))}{(1 - (1 - S\epsilon)^{2I^2})T_0}}, \tag{24}$$

where we have used both the results in 18 and in 23. With this expression of the bound, the value of T_0 minimizing the formulation is:

$$T_0 = \left(\frac{2LTI^2}{\sigma^2_{\min}(\boldsymbol{O})\epsilon} \sqrt{\frac{S(1 + \log(I^2/\delta))}{1 - (1 - S\epsilon)^{2I^2}}} \right)^{2/3}.$$

which can now be obtained without requiring the knowledge of ether $\pi_{\min}$ or λ.

C Comparison with Spectral Decomposition technique used in (Zhou et al., 2021)

We devote this section to the comparison of our estimation approach with spectral decomposition techniques (Anandkumar et al., 2014). In particular, we will focus on the comparison with the spectral procedure adopted in the work of Zhou et al. (2021) which faces a similar problem to the one we consider. We start by highlighting the main differences between their work and ours:

- they consider learning both the transition and the observation models, while we assume to know the latter.
- they have a further assumption compared to ours as they require the invertibility of the transition matrix $\boldsymbol{P}$.

- they assume to have access to an optimization oracle that returns the optimal policy for a given known model. Differently, our oracle optimizes the best instantaneous expected reward given the belief on the MABs at each timestep computed using the real transition and observation matrices.

The authors propose the SEEU (Spectral Exploration and Exploitation with UCB) algorithm that alternates between exploration phases used to make parameter estimation and exploitation phases where the actions are pulled according to the computed optimistic policy. During the exploration phases, they use standard spectral decomposition methods to estimate both the observation and the transition models. The guarantees they provide require both Assumption 4.1 and 4.2 and they further require the invertibility of the transition matrix, which is needed for the application of SD approaches. The algorithm they devise reaches $\mathcal{O}(T^{2/3})$ regret, disregarding logarithmic terms.

To define a comparison with Spectral Decomposition techniques, we need to introduce some quantities that will be helpful in what will follow. We will report some results appearing in Appendix B of Zhou et al. (2021) on spectral decomposition techniques, based on the work of Anandkumar et al. (2014).
In particular, they introduce the following three matrices $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3 \in \mathbb{R}^{IV \times S}$ where each row is associated to a couple $(a, r) \in \mathbb{I} \times \mathbb{V}$ and each column is associated to a state $s \in \mathbb{S}$. They are defined as follows:

$$\begin{aligned}
\boldsymbol{B}_1\big((a,r),s\big) &= \Pr(a_{t-1} = a, r_{t-1} = r | s_t = s) \\
\boldsymbol{B}_2\big((a,r),s\big) &= \Pr(a_{t} = a, r_{t} = r | s_t = s) \\
\boldsymbol{B}_3\big((a,r),s\big) &= \Pr(a_{t+1} = a, r_{t+1} = r | s_t = s)
\end{aligned}$$

for any $(a, r) \in \mathbb{I} \times \mathbb{V}$ and for any state $s \in \mathbb{S}$.
We can now present the bound on the error of the estimated transition matrix. From Anandkumar et al. (2014), it can be shown that with a sufficient number of samples N and with probability at least $1 - \delta$, spectral approaches ensure that:

$$\|\boldsymbol{P} - \widehat{\boldsymbol{P}}\|_F \leq C_2 \sqrt{\frac{\log\left(6\frac{I^2V^2+IV}{\delta}\right)}{N}}, \tag{25}$$

with

$$\begin{aligned}
C_2 &= \frac{4}{\sigma_{\min}(\boldsymbol{B}_2)}\Big(S + S^{3/2} * \frac{21}{\sigma_{1,-1}}\Big) C_3 \\
C_3 &= 2G\frac{2\sqrt{2}+1}{(1-\theta)\sqrt{\pi_{\min}}}\Big(1 + \frac{8\sqrt{2}}{\pi_{\min}^2 \sigma^3} + \frac{256}{\pi_{\min}^3 \sigma^2}\Big)
\end{aligned}$$

where $\sigma_{1,-1}$ is the smallest nonzero singular value of a covariance matrix computed during the estimation process (see Section 3.1 in Zhou et al. (2021)) and $\sigma = \min\{\sigma_{\min}(\boldsymbol{B}_1), \sigma_{\min}(\boldsymbol{B}_2), \sigma_{\min}(\boldsymbol{B}_3)\}$, where $\sigma_{\min}(\boldsymbol{B}_i)$ represents the smallest nonzero singular value of the matrix $\boldsymbol{B}_i$, for $i = 1, 2, 3$. $\boldsymbol{\pi}$ represents the stationary distribution of the underlying chain and $\pi_{\min} := \min_s \boldsymbol{\pi}(s) \geq \epsilon$. Finally, θ and G are some mixing rate parameters. Under Assumption 4.1, it is possible to show that we can take $G = 2$ thus having $\theta \leq 1 - \epsilon$.
The term C_2 that we report here shows a further $\sqrt{S}$ term with respect to the C_2 term reported in Zhou et al. (2021). This is due to the fact that the bound reported in Equation 25 uses the Frobenious norm of the difference of the matrices, while Zhou et al. (2021) report the bound in terms of the spectral norm $\|\cdot\|_2$. The additional $\sqrt{S}$ of the C_2 term reported here is indeed due to the conversion between the two norms, for which it holds that $\|\boldsymbol{P} - \widehat{\boldsymbol{P}}\|_F \leq \sqrt{S}\|\boldsymbol{P} - \widehat{\boldsymbol{P}}\|_2$.

We can simplify the expression of C_2 by reporting a new constant C_2' for which it can be easily shown that $C_2' \leq C_2$. It is defined as follows:

$$C_2' = \frac{16(2\sqrt{2}+1)}{\sigma_{\min}(\boldsymbol{B}_2)\sqrt{\pi_{\min}}}\Big(S + S^{3/2} * \frac{21}{\sigma_{1,-1}}\Big)\Big(1 + \frac{8\sqrt{2}}{\pi_{\min}^2 \sigma_{\min}^3(\boldsymbol{B}_2)} + \frac{256}{\pi_{\min}^3 \sigma_{\min}^2(\boldsymbol{B}_2)}\Big).$$

Having defined the quantity C_2' containing the dependency on the problem parameters of the spectral decomposition approach used in Zhou et al. (2021), we can now consider the guarantees of the estimation

procedure reported in Lemma 6.1.
Before proceeding with the comparison, we recall that Lemma 6.1 is expressed with respect to the number N of pulls of each pair of arms, while the total number of samples used corresponds to $N_{tot} = 2NI^2$. By substituting the expression $N = N_{tot}/(2I^2)$ into the bound, we get:

$$\|\boldsymbol{P} - \widehat{\boldsymbol{P}}\|_F \leq \frac{4I^2}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}} \sqrt{\frac{S(1+\log(I^2/\delta))}{(1-\lambda^{2I^2})N_{tot}}}.$$

Using this new result, we define our constant C_{our} as:

$$C_{our} = \frac{4I^2}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}} \sqrt{\frac{S}{1-\lambda^{2I^2}}}.$$

It is now relevant to note the similarities between matrix $\boldsymbol{B}_2$ and our action observation matrix $\boldsymbol{O}$. Given any state $s \in \mathbb{S}$ and any action-reward pair (a, r), we have that:

$$\begin{aligned}\boldsymbol{O}\big((a,r),s\big)\Pr(a_t = a|s_t = s) &= \Pr(r_t = r|s_t = s, a_t = a)\Pr(a_t = a|s_t = s)\\ &= \Pr(a_t = a, r_t = r|s_t = s)\\ &= \boldsymbol{B}_2\big((a,r),s\big),\end{aligned}$$

where the first equality follows by the definition of $\boldsymbol{O}$ in Equation 2. Furthermore, since the SEEU algorithm samples uniformly over the I actions during the exploration phase, for any $(a, r) \in \mathbb{I} \times \mathbb{V}$ and for any state $s \in \mathbb{S}$, we have that:

$$\boldsymbol{O}\big((a,r),s\big)\Pr(a_t = a|s_t = s) = \boldsymbol{O}\big((a,r),s\big)\frac{1}{I} = \boldsymbol{B}_2\big((a,r),s\big).$$

Considering the minimum singular values of these matrices, the stated result allows also to say:

$$\sigma_{\min}(\boldsymbol{O}) = \sigma_{\min}(\boldsymbol{B}_2)I. \tag{26}$$

We can now rewrite constant C_{our} as:

$$C_{our} = \frac{4I^2}{\sigma^2_{\min}(\boldsymbol{O})\pi_{\min}} \sqrt{\frac{S}{1-\lambda^{2I^2}}} = \frac{4}{\sigma^2_{\min}(\boldsymbol{B}_2)\pi_{\min}} \sqrt{\frac{S}{1-\lambda^{2I^2}}}$$

where the equality directly follows from 26.

We are now ready to compare the constant C_2' of SD approaches with constant C_{our} appearing in our approach[6]. C_2' has a dependency of order $-7/2$ with respect to $\pi_{\min}$ while C_{our} enjoys a dependency of order -1. By considering instead the number of states S, constant C_2' contains a term that scales with order $3/2$, while we have a dependency of order $1/2$. Finally, the dependency on the minimum singular value of the matrix $\boldsymbol{B}_2$ has order -4 in C_2' and order -2 in C_{our}[7]. Again, we recall that these considerations are made on C_2' which is a smaller value than the real one C_2 appearing in their bound.
From this analysis, we have shown that our approach enjoys better dependence with respect to SD approaches on all the problem parameters.

D Useful Lemmas and Deviation Inequalities

This section is devoted to the presentation of some results that are useful in understanding some proofs appearing in Appendix B.

Lemma D.1. *(Lemma A.1 in Ramponi et al. (2020)) Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ any pair of vectors, then it holds that:*

$$\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} \right\|_2 \leq \frac{2\|\boldsymbol{x} - \boldsymbol{y}\|_2}{\max\{\|\boldsymbol{x}\|_2, \|\boldsymbol{y}\|_2\}}$$

[6]We only consider the terms that can be directly compared, thus disregarding the term $\sqrt{1-\lambda^{2I^2}}$.
[7]The empirical impact of this dependency is evident in the numerical comparison in Section 7.3.

Proof. The presented result follows from a sequence of algebraic manipulations:

$$\begin{aligned}\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} \right\|_2 &= \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_2} \pm \frac{\boldsymbol{y}}{\|\boldsymbol{x}\|_2} \right\|_2 \\ &= \left\| \frac{\boldsymbol{x}-\boldsymbol{y}}{\|\boldsymbol{x}\|_2} - \frac{\boldsymbol{y}(\|\boldsymbol{y}\|_2 - \|\boldsymbol{x}\|_2)}{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2} \right\|_2 \\ &\leq \frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2}{\|\boldsymbol{x}\|_2} + \frac{|\|\boldsymbol{x}\|_2 - \|\boldsymbol{y}\|_2|}{\|\boldsymbol{x}\|_2} \\ &\leq 2\frac{\|\boldsymbol{x}-\boldsymbol{y}\|_2}{\|\boldsymbol{x}\|_2},\end{aligned}$$

where the triangular inequality has been applied in the third line and the reverse triangular inequality in the last one, i.e. $|\|\boldsymbol{x}\|_2 - \|\boldsymbol{y}\|_2| \leq \|\boldsymbol{x}-\boldsymbol{y}\|_2$. The result in the lemma can be derived by observing that, for symmetry reasons, the same derivation can be performed by getting $\|\boldsymbol{y}\|_2$. □

D.1 Concentration Results for Discrete Distributions

We present here some concentration bounds that can be derived from McDiarmid's inequality (McDiarmid, 1989). The first proposition considers samples that are i.i.d. while the second result considers samples coming from a Markov chain and is indeed used in the proof of Lemma 6.1.

Proposition D.1. *(Concentration for Discrete Distributions with Independent Samples (Proposition 19 (Hsu et al., 2012)) Let* z *be a discrete random variable that takes values in* $\{1, \dots, d\}$*, distributed according to* $\mathbf{q}$*. We write* $\mathbf{q}$ *as a vector* $\mathbf{q} = [Pr(z = j)]_{j=1}^{d}$*. Assume we have* N *i.i.d. samples, and that our empirical estimate of* $\mathbf{q}$ *is* $[\widehat{q}]_j = \sum_{i=1}^{N} \mathbb{1}[z_i = j]/N$*.*
We have that $\forall \epsilon > 0$ *:*

$$Pr\left(\|\widehat{\mathbf{q}} - \mathbf{q}\|_2 \geq \frac{1}{\sqrt{N}} + \epsilon \right) \leq \exp\left(-N\epsilon^2 \right)$$

The presented result can be also written for dependent samples coming from a Markov chain. For this case, it is possible to discount the number of samples based on the modulus of the second largest eigenvalue of the transition matrix $\boldsymbol{P}$ (as reported in Appendix A of Hsu et al. (2012)). Also, in Fan et al. (2021), they state that $\frac{1+\lambda}{1-\lambda}N$ Markov chain samples are needed to achieve the same accuracy with N independent samples in the naive Monte Carlo method, with λ being the modulus of the second largest eigenvalue of $\boldsymbol{P}$.

Proposition D.2. *(Concentration for Discrete Distributions with Samples coming from a Markov Chain) Let* z *be a discrete random variable that takes values in* $\{1, \dots, d\}$*, distributed according to* $\mathbf{q}$*. We write* $\mathbf{q}$ *as a vector* $\mathbf{q} = [Pr(z = j)]_{j=1}^{d}$*. Assume we have* N *samples coming from a Markov process having a transition matrix with the modulus of the second largest eigenvalue* λ*, and assume that the Markov chain starts from its stationary distribution* $\boldsymbol{\pi}$*. Our empirical estimate of* $\mathbf{q}$ *is* $[\widehat{q}]_j = \sum_{i=1}^{N} \mathbb{1}[z_i = j]/N$*.*
We have that $\forall \epsilon > 0$ *:*

$$Pr\left(\|\widehat{\mathbf{q}} - \mathbf{q}\|_2 \geq \sqrt{\left(\frac{1+\lambda}{1-\lambda}\right)\frac{1}{N}} + \epsilon \right) \leq \exp\left(-\frac{1-\lambda}{1+\lambda} \cdot N\epsilon^2 \right)$$

D.2 Bounds on the Error of the Estimated Belief

We present here the result appearing in Zhou et al. (2021) that controls the error in the estimated belief. They consider a setting with Bernoulli rewards $\mathbb{V} = \{0, 1\}$ which results in an action observation matrix $\boldsymbol{O} \in \mathbb{R}^{2I \times S}$. Since the rewards are Bernulli, the dimension of the action observation matrix can be halved, because half of the probabilities contained are complementary to the other half. In particular, we will use the reward matrix $\boldsymbol{\mu} \in \mathbb{R}^{I \times S}$ such that each of its elements is defined as in Equation 4 that we report here for simplicity.

$$\boldsymbol{\mu}(a, s) = \sum_{r \in \mathbb{V}} r\boldsymbol{O}\big((a, r), s\big) \qquad \forall\, a \in \mathbb{I},\ s \in \mathbb{S}.$$

Proposition D.3. *(Controlling the belief error (Zhou et al., 2021)) Assume to have a transition matrix $\boldsymbol{P}$ of size $S \times S$ with minimum entry $\epsilon > 0$ and reward matrix $\boldsymbol{\mu} \in \mathbb{R}^{I \times S}$. Let's assume to have an estimator $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{P}})$ of the true model parameters $(\boldsymbol{\mu}, \boldsymbol{P})$. For an arbitrary reward-action sequence $\{r_{1:t}, a_{i:t}\}_{t \geq 1}$, let $\widehat{\mathbf{b}}_t$ and $\mathbf{b}_t$ be the corresponding beliefs in period t under $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{P}})$ and $(\boldsymbol{\mu}, \boldsymbol{P})$ respectively. Then there exists constants L_1 and L_2 such that:*

$$\|\widehat{\mathbf{b}}_t - \mathbf{b}_t\|_1 \leq L_1 \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_1 + L_2 \|\widehat{\boldsymbol{P}} - \boldsymbol{P}\|_F,$$

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

We adapt the previous result to our setting with a discrete set of rewards and with a known observation matrix $\boldsymbol{O}$. It can be expressed as follows.

Proposition D.4. *(Controlling the belief error in Switching Latent Bandits) Assume to have a transition matrix $\boldsymbol{P}$ of size $S \times S$ with minimum entry $\epsilon > 0$. Given a model with parameters $(\boldsymbol{O}, \boldsymbol{P})$, we assume to know the observation model $\boldsymbol{O}$ and to have an estimation of the transition matrix $\widehat{\boldsymbol{P}}$. For an arbitrary reward-action sequence $\{r_{1:t}, a_{i:t}\}_{t \geq 1}$, let $\widehat{\mathbf{b}}_t$ and $\mathbf{b}_t$ be the corresponding beliefs in period t under $(\boldsymbol{O}, \widehat{\boldsymbol{P}})$ and $(\boldsymbol{O}, \boldsymbol{P})$ respectively. Then there exists a constant L such that:*

$$\|\widehat{\mathbf{b}}_t - \mathbf{b}_t\|_1 \leq L \|\widehat{\boldsymbol{P}} - \boldsymbol{P}\|_F,$$

where $L = \frac{4S(1-\epsilon)^2}{\epsilon^3} + \sqrt{S}$ and $\|\cdot\|_F$ is the Frobenius norm.

E Continuous Reward Distributions

In the main paper, we focused on the case where the set of possible observations is discrete and the distribution of the observations $\boldsymbol{O}\big((a, \cdot), s\big) \in \Delta(\mathbb{V})$ is categorical, for each latent state $s \in \mathbb{S}$ and action $a \in \mathbb{I}$. In this section, we show how we can extend the estimation procedure to also handle continuous reward distributions. Formally, if we consider having a Switching Latent Bandit problem with continuous rewards and a number S of bandits and a number I of actions available for each bandit, there will be IS potentially different continuous reward distributions $\Pr(\cdot|s, a)$ for each latent state $s \in \mathbb{S}$ and action $a \in \mathbb{I}$. If we assume to discretize each reward distribution into U consecutive intervals, we will have $U - 1$ splitting points. By considering the ordered set of splitting points and taking two consecutive splitting points u_h and u_k for which holds that $u_h < u_k$, we can define the interval $\mathcal{I}_{hk} = (u_h, u_k]$. The probability that a realization from a continuous distribution $\Pr(\cdot|s, a)$ falls within interval $\mathcal{I}_{hk}$ is defined as:

$$\Pr(r \in \mathcal{I}_{hk}|s, a)) = \int_{u_h}^{u_k} \Pr(dr|s, a)\, dr.$$

Of course, if we are able to exactly compute the integrals in the previous formulation we will not introduce any error in the discretization process. By applying the same procedure for all the U intervals identified, we can define the parameters of the new categorical distribution. This procedure is then applied to all the continuous probabilities $\Pr(\cdot|s, a)$ for each $s \in \mathbb{S}$ and $a \in \mathbb{I}$ using the same splitting points and we finally obtain a new action observation matrix of size $IU \times S$, which should of course satisfy Assumption 4.2.
From this point on, we can build the new reference matrix and we can proceed with the estimation procedure presented in Algorithm 1. Whenever a reward is observed during the estimation procedure, the count vector is updated by considering the interval to which the observed reward belongs.
It is an interesting problem to determine in this setting the number of suitable splits and the location of the split points that leads to an action observation matrix with higher $\sigma_{\min}(\boldsymbol{O})$.

Another issue arises when the environment comprises numerous but finite observations. In such scenarios, we can employ the inverse approach by clustering some observations, thereby reducing the scale of the problem. By selecting a number of clusters $C < V$, we can divide the observations into distinct groups. This allows us to utilize cluster-level probabilities (obtained by summing probabilities of the single observations) to construct a new action observation matrix and then proceed with the standard estimation procedure.