Restless Hidden Markov Bandits with Linear Rewards

Michal Yemini, Amir Leshem, Anelia Somekh-Baruch

1 Department of Electrical Engineering
Stanford University, USA
michaly@stanford.edu
2 Faculty of Engineering
Bar Ilan University, Ramat Gan, Israel
leshema@biu.ac.il, anelia.somekhbaruch@gmail.com

Abstract
This paper presents an algorithm and regret analysis for the restless hidden Markov bandit problem with linear rewards. In this problem the reward received by the decision maker is a random linear function which depends on the arm selected and a hidden state. In contrast to previous works on Markovian bandits, we do not assume that the decision maker receives information regarding the state of the system, but has to infer it based on its actions and the received reward. Surprisingly, we can still maintain logarithmic regret in the case of polyhedral action set. Furthermore, the regret does not depend on the number of extreme points in the action space.

Introduction

Preliminaries

This work considers a setup in which at each time instant $t$, a decision maker chooses an arm $b_t \in B \subset \mathbb{N}$ to pull and an action $a_t \in \mathcal{A} \subset \mathbb{R}^N$ and gets a reward that depends linearly on a random function of the system state $s_t \in S \subset \mathbb{N}$ and the chosen arm $b_t \in B$ in the following way:

$$r_t(b_t, a_t) = \langle a_t, \theta(b_t, s_t) \rangle,
\tag{1}$$

where $\langle x, y \rangle$ is the inner product between $x$ and $y$. We assume that the set $\mathcal{A}$ is compact and that the sets $S$ and $B$ are finite. We assume that the process $(s_t)_{t=1,2,...}$ is a time-homogeneous and irreducible and aperiodic Markov chain over a finite state space $S$. Additionally, for each $s \in S$ and $b \in B$, $\theta(b, s)$ is a random function with a range $\Theta_{b,s} \subset \mathbb{R}^N$. We refer to this model as the restless hidden Markov bandit model with a linear reward. In this work we consider a setup in which a decision maker only knows the sets $\mathcal{A}$, $S$ and $\Theta_{b,s}$ for all $(b, s) \in \mathcal{B} \times S$ but not the transition probabilities of the Markov process $(s_t)_{t=1,2,...}$, the probability distribution of the random function $\theta(b, s)$, nor their realizations at time $t$. This model captures, for example, communication networks in which a user chooses a single frequency band for transmission, out of several possibilities such as 2.4 GHz, 3.6 GHz or 5 GHz. Each of these frequency bands includes several sub-bands. After choosing a frequency band for transmission, the user chooses a frequency allocation over the sub-frequency bands included in the chosen band, subject to a total spectrum and power utilization constraint over all sub-bands. Upon making the choices, the user does not know what the current quality of the chosen frequency bands is, but only knows the communication rates (or delays) of previous transmissions. The choice of the group of frequency bands can be thought of as choosing an arm, and the frequency allocation over the sub-bands can be considered as choosing an action vector.

Discussion

The restless hidden Markov bandit model with linear rewards described in $(1)$ is related to several learning models, among them are Markov bandits and Markov decision processes, stochastic linear bandits, restless bandits models, and the partially observed Markov decision process. In Markov bandit models the reward is generated by each arm independently of other arms, and changes over time according to a Markov process that progresses over time when an arm is played. The first analysis of the expected regret of a policy for the Markov bandit problem compared the expected reward of the policy to that of the arm with the best expected reward that was found based on the stationary distribution of the Markov chain, see Anantharam, Varaiya, and Walrand (1987). An instantaneous in time minimization regret approach which depends on the transition probabilities of the Markov chain, and not only its stationary distribution, is analyzed in the Markov decision process (MDP) literature which considers a Markov process that controls the state of a system, it is assumed that this state is known to the decision maker upon choosing an action to play. Several works analyze MDPs, among them are Filippi, Cappe, and Garivier (2011); Tekin and Liu (2012); Auer and Ortner (2007) which attain a logarithmic regret that depends on the Markov chain parameters and the size action space, assuming that the Markov chain is a unichain. Additionally, the general case that includes weakly communicating Markov chains is investigated in Bartlett and Tewari (2009); Jaksch, Ortner, and Auer (2010); Qian et al. (2018); Fruit, Pirotta, and Lazaric (2018) and Ortner (2018) in which a scheme that achieves an $O(1/\sqrt{T})$ regret that depends on the MDP parameters is considered. It was also proven in Fruit, Pirotta, and Lazaric (2018) that is not possible to achieve a logarithmic regret that depends on the MDP parameters.
We note that our model is different from Meshram, Gopalan, Wang et al. (2019) in which the process that governs the wards is also related to stochastic linear bandits studied, for linear bandit process since the state of the system evolves estimating the reward distribution for each action. We note however, that since the structure of the Urteaga and Wiggins (2018) in which the reward distributions for the multi-armed contextual bandit model considered in considers an information acquisition and sequential belief independent objects and uses an approximated Gittins index Evans (2001, 2003); Javidi (2016), where Krishnamurthy related POMDP models are discussed in Krishnamurthy and Wahlberg (2009) in which a decision maker aims to improving the belief of the unknown state of the system. Other its desire to increase the immediate reward with the benefitsof Kaelbling, Littman, and Cassandra (1998); Krishnamurthy and Manjunath (2017) since our setup assumes that the partially observed Markov decision process (POMDP) Kaelbling, Littman, and Cassandra (1998); Krishnamurthy and Evans (2003) considers a tracking problem with independent objects and uses an approximated Gittins index approach for finding policies. The work of Javidi (2016) considers an information acquisition and sequential belief refinement with a finite number of possible actions. Finally, our model is also related to the Gaussian mixture models for the multi-armed contextual bandit model considered in Urteaga and Wiggins (2018) in which the reward distributions are approximated using nonparametric Gaussian mixture models. We note however, that since the structure of the reward function is known to be linear in our model, we estimate the probability distribution of the system instead of estimating the reward distribution for each action.

Contributions: This work differs from the aforementioned works in several aspects: First, it is not a classical stochastic linear bandit process since the state of the system evolves over time according to a Markov process. However, it is not a classical Markov bandit model nor a restless one since the states are not directly observed or given to the decision maker. Additionally, the action space is uncountable. Interestingly, we prove that the cardinality of the action space or its set of extreme points does not affect the expected regret. Moreover, the scheme we propose divides the estimation of the probability distribution which controls the system evolution into two parts; the first part estimates the transition matrix of the Markov chain common to all arms and the second estimates the probability distributions of the unknown parameter \( \theta \), which depends on the system state and the arm. Our numerical results demonstrate the merits of our proposed scheme, namely the significant reduction of the expected regret of the decision maker.

Model Formulation
This section defines the restless hidden Markov bandit problem with a linear reward in more detail. We consider the setup that is stated in [1]. The process \((s_t)_{t=1,2,...}\) is a finite space \(S\) Markov chain with a transition matrix \(P_S\) and a unique stationary distribution \(\mu_S\). We denote the transition probability between state \(\tilde{s}\) and \(\bar{s}\) in \(S\) by \(P_S(\tilde{s},\bar{s})\). Let \(\mathcal{B}\) be the set of arms that the decision maker can choose from and let the action space \(\mathcal{A}\) be an \(N\)-dimensional compact and convex polytope. The set \(\mathcal{A}\) represents the possible resource allocations to \(N\) processes. We denote the set of extreme points of \(\mathcal{A}\) by \(\mathcal{V}\). We also assume that the set \(\Theta_{h,m}\) is finite for every \((b, s)\in\mathcal{B}\times\mathcal{S}\) and that \(|\mathcal{V}| \gg |\Theta_{h,m}|\). Additionally, \(\Theta_{h,m} \cap \Theta_{h,m} = \emptyset\) for every \(\tilde{s} \neq \bar{s}\). Finally, we denote by \(P_{\Theta_{h,m}}(\theta)\) the probability of the random vector \(\theta \in \Theta_{h,m}\).

At each time \(t\) the decision maker receives a reward \(r_t = \langle a_t, \theta(b_t,s_t) \rangle\). Upon receiving this reward the decision maker chooses an arm \(b_{t+1}\) to pull and an action choice \(a_{t+1}\), given the arm choices, actions and rewards of previous times, \(1, \ldots, t\), and the sets \(\mathcal{B}, \mathcal{A}, \mathcal{S}\) and \(\Theta_{h,m}, (b, s) \in \mathcal{B}\times\mathcal{S}\).

We define the regret of the hidden restless Markov bandit model with linear rewards with respect to the expected reward of the restless Markov bandit model with linear rewards defined next. This model assumes that the decision maker perfectly knows in advance all the parameters of the model as well as the identity of the previous state.

Definition 1 (The Restless Markov Bandit Model with Linear Rewards) In the restless Markov bandit model with linear rewards a decision maker knows in advance the transition matrix \(P_S\) and the probability distributions \(P_{\Theta_{h,m}}, (b, s) \in \mathcal{B}\times\mathcal{S}\) as well as the sets \(\mathcal{A}, \mathcal{S}\) and \(\Theta_{h,m}, (b, s) \in \mathcal{B}\times\mathcal{S}\).

At each time \(t\) the decision maker receives a reward \(r_t = \langle a_t, \theta(b_t,s_t) \rangle\) and the identity of the state \(s_t\). Upon receiving this reward the decision maker chooses an arm \(b_{t+1}\) given the actions, rewards and states of previous times, \(1, \ldots, t\), and the sets \(\mathcal{A}, \mathcal{S}\) and \(\Theta_{h,m}, (b, s) \in \mathcal{B}\times\mathcal{S}\).

Definition 2 (The Average Reward of a Policy for the Restless Markov Bandit Model with Linear Rewards) A policy for the restless Markov bandit model with linear rewards is defined as a mapping \(\pi : \mathcal{S} \to \mathcal{B}\times\mathcal{A}\). The average reward of an action policy \(\pi\) is defined as

\[
\rho(\pi) = \sum_{\tilde{s}, \bar{s} \in \mathcal{S}} \mu_S(\tilde{s}) P_S(\tilde{s}, \bar{s}) \sum_{\theta \in \Theta_{h,m}, \tilde{s}} P_{\Theta_{h,m}}(\theta) \langle a_\pi(\tilde{s}), \theta \rangle. \tag{2}
\]
Definition 3 (Regret Definition for the Restless Hidden Markov Bandit Model with Linear Rewards). Denote by \( \pi^* \) the policy that maximizes \( \mathbb{E}[R(T)] \) for the restless hidden Markov bandit model with linear rewards. We define the regret of the restless hidden Markov bandit model with linear rewards as

\[
R(T) = T \rho(\pi^*) - \sum_{t=1}^{T} r_t(b_t, a_t).
\]

That is, we define the regret to be relative to the optimal policy for the scenario in which the decision maker is in possession of the previous state, the Markov chain transition matrix and the probability distributions of \( \theta \).

Notation: We denote by \( B(c, r) \) the \( n \)-dimensional ball with center \( c \in \mathbb{R}^n \) and radius \( r \).

Upper Confidence Bound Reinforcement Learning for the Restless Hidden Markov Bandit Model with Linear Rewards

This section presents Algorithm 1 and establishes its expected regret for the restless hidden Markov bandit model with linear rewards. Algorithm 1 uses two types of upper confidence bounds, the first assists in estimating the transition probabilities of the Markov chain and is not arm dependent, the other assists in estimating the probability distributions \( P_{\Theta_{b,s}} \), which are arm dependent. Additionally, Algorithm 1 recovers at each time \( t \) the identity of the previous state \( s_{t-1} \) with probability 1, this recovery is \( \epsilon \)-optimal in the sense that for each \( \epsilon > 0 \) we can find a recovery scheme with an expected regret smaller than \( \epsilon \). Denote,

\[
\text{conf}_S(t, s) \triangleq \min \left\{ 1, \frac{\log(4(t - 1)^\alpha |S|^2)}{2N_t(s)} \right\},
\]

\[
\text{conf}_\Theta(t, b, s) \triangleq \min \left\{ 1, \frac{\log(4(t - 1)^\alpha |\Theta_{b,s}| |B||S|^2)}{2N_t(b, s)} \right\}.
\]

To evaluate the expected regret of Algorithm 1 we observe that (5) follows directly from the Chernoff-Hoeffding inequality

\[
\Pr \left( |\hat{P}_{\Theta_{b,s}}(\theta) - P_{\Theta_{b,s}}(\theta)| > \text{conf}_\Theta(t, b, s) \right) \leq \frac{(t - 1)^{-\alpha}}{2|\Theta_{b,s}| |B||S|^2}.
\]

Using these inequalities we define the confidence intervals

\[
|\hat{P}_S(\tilde{s}, \tilde{s}) - P_S(\tilde{s}, \tilde{s})| \leq \text{conf}_S(t, s),
\]

\[
|\hat{P}_{\Theta_{b,s}}(\theta) - P_{\Theta_{b,s}}(\theta)| \leq \text{conf}_\Theta(t, b, s)
\]

of length \( \text{conf}_S(t, s) \) and \( \text{conf}_\Theta(t, b, s) \), respectively.

Before we upper bound the expected regret for Algorithm 1 we define the following notations. Denote \( T_M = \max_{\tilde{s}, \tilde{s} \in S} E(T_{\tilde{s}, \tilde{s}}) \) where \( T_{\tilde{s}, \tilde{s}} \) is the passage time of first arriving at state \( \tilde{s} \) when starting from state \( \tilde{s} \), and let \( T_S = (\min_{\tilde{s}, \tilde{s} \in S} P_S(\tilde{s}, \tilde{s}) > 0)^{-1} \). Additionally, denote \( r_{\max} = \max_{a, \tilde{a} \in \mathcal{A}, \theta \in \Theta_{b,s}} \{ (a, \theta) - (\tilde{a}, \theta) \} \), and let \( C_{\max} = \max_{\tilde{s} \in S} |\Theta_{b,s}| \).

Theorem 1. The expected regret of Algorithm 1 is

\[
O \left( |B||S|T_M r_{\max} \log \frac{4T^\alpha C_{\max} |B||S|^2}{\Delta^2} + |B|^2 |S|^2 T_M r_{\max} \log_2 \left( \frac{T}{|S| |B|^2} + 1 \right) \right),
\]

where \( \Delta > 0 \) is the maximal \( \delta \) such that if \( |\hat{P}_S(\tilde{s}, \tilde{s}) - P_S(\tilde{s}, \tilde{s})| \leq \delta \) for all \( \tilde{s}, \tilde{s} \in S, \) and \( |\hat{P}_{\Theta_{b,s}}(\theta) - P_{\Theta_{b,s}}(\theta)| \leq \delta \) for all \( (b, s) \in B \times S \) then \( \mathcal{A}^* \{ \hat{P}_S, \{ \hat{P}_{\Theta_{b,s}} \} | b, s \in B \times S, \tilde{s} \} = \mathcal{A}^* \{ P_S, \{ P_{\Theta_{b,s}} \} | b, s \in B \times S, \tilde{s} \} \) for every \( \tilde{s} \in S \).

Theorem 1 will be proved in the next section. Before proving it we discuss Algorithm 1 and consider several adaptations. We note that the term (7) does not depend on the cardinality of the set \( V \). We also note that it follows from Theorem 1.2 in Kontorovich and Ramanan (2008) and Theorem 1.1 in Lezaud (1998) that using the estimation for the transition matrix directly instead of using its confidence interval as in Algorithm 1 yields the same upper-bound (7) for the regret. This simplifies the problem since \( \hat{P}(\tilde{s}, \tilde{s}) \) is replaced with \( \hat{P}(\tilde{s}, \tilde{s}) \). Furthermore, we remark that we can remove the term \( T_S \) from (7) if we use confidence intervals for the joint probability distribution \( P(\tilde{s}, \tilde{s}, \theta) \) instead of using separate sets of confidence intervals for estimating the transition matrix of the Markov chain and for the probability distributions of \( \theta \). However, our numerical results show that this may be suboptimal since in this case we can no longer estimate the transition matrix of the Markov chain jointly over all arms.

The Motivation for Algorithm 1

Let \( P_S \) be a transition matrix of a Markov chain with state set \( S \), let \( B \) be a finite set and let \( P_{\Theta_{b,s}}(b, s) \in B \times S \) be probability distributions. Denote

\[
(\tilde{b}, \tilde{a}) \in \arg \max_{b, a \in B, a \in \mathcal{A}} \left\{ \sum_{\tilde{s} \in S} P_S(\tilde{s}, \tilde{s}) \sum_{\theta \in \Theta_{b,s}} P_{\Theta_{b,s}}(\theta) \langle a, \theta \rangle \right\}.
\]

Lemma 1. For every transition matrix \( P_S \) and every collection of probability distributions \( P_{\Theta_{b,s}} \) there exists \( \delta > 0 \) such that if \( |\hat{P}_S(\tilde{s}, \tilde{s}) - P_S(\tilde{s}, \tilde{s})| \leq \delta \) for all \( \tilde{s}, \tilde{s} \in S, \) and \( |\hat{P}_{\Theta_{b,s}}(\theta) - P_{\Theta_{b,s}}(\theta)| \leq \delta \) for all \( (b, s) \in B \times S \) we have that \( \mathcal{A}^*(\hat{P}_S, \{ \hat{P}_{\Theta_{b,s}} \} | b, s \in B \times S, \tilde{s}) = \mathcal{A}^* \{ P_S, \{ P_{\Theta_{b,s}} \} | b, s \in B \times S, \tilde{s} \} \) for every \( \tilde{s} \in S \).

Proof. Recall that \( V \) is the set of extreme points of the polytope \( \mathcal{A} \). Since the set \( \mathcal{A} \) is a bounded and convex polytope, the optimal actions of \( \sum_{\tilde{s} \in S} P_S(\tilde{s}, \tilde{s}) \sum_{\theta \in \Theta_{b,s}} P_{\Theta_{b,s}}(\theta) \langle a, \theta \rangle \) are in the set \( V \) for every choice of arm \( b \in B \) and state \( \tilde{s} \in S \).
Algorithm 1:

1. **Notations:**\( \epsilon_t = \epsilon \left( 10 \cdot t^r \cdot \max_{\theta \in \mathcal{L}|_{\theta|_{t}}} \|\theta\|_1 \right)^{-1}, \quad \forall t \in \mathbb{N}, \)
2. \(\text{conf}_S(t, s) \triangleq \min \left\{ 1, \frac{\log(4(t-1)^\epsilon |\mathcal{S}|^2)}{2N(t,s)} \right\}, \quad \forall s \in \mathcal{S}, t \in \mathbb{N},\)
3. \(\text{conf}_\theta(t, b, s) \triangleq \min \left\{ 1, \frac{\log(4(t-1)^\epsilon |\theta_{b,s}| |\mathcal{S}|^2)}{2N(t,b,s)} \right\}, \quad \forall b \in \mathcal{B}, s \in \mathcal{S}, t \in \mathbb{N};\)
4. **Data:** \(\mathcal{A}, \mathcal{S}, \mathcal{B}, \Theta_{b,s} \forall (b,s) \in \mathcal{B} \times \mathcal{S}, \alpha > 2, \epsilon > 0, \alpha_\epsilon > 1;\)
5. Set \(\delta_{-1} = s \) for some \(s \in \mathcal{S};\)
6. Set \(N_0(\delta, \delta) = N_0(\delta) = 0 \) \(\forall \delta, \delta \in \mathcal{S};\)
7. Set \(N_0(b, s) = 0 \) \(\forall (b, s) \in \mathcal{B} \times \mathcal{S};\)
8. Set \(\text{conf}_S(0, b, s) = \text{conf}_\theta(1, b, s) = 1 \) \(\forall b \in \mathcal{B}, s \in \mathcal{S};\)
9. Set \(t = 0;\)
10. **for** \(\text{round } k = 0, 1 \ldots \) **do**
11. \hspace{1em} **Initialize round** \(k:\)
12. \hspace{2em} 1. Set \(t_k = t;\)
13. \hspace{2em} 2. For every \(\delta, \delta \in \mathcal{S}\) such that \(N_k(\delta) > 0\) set \(P_{t_k}(\delta, \delta) = \frac{N_k(\delta, \delta)}{N_k(\delta)}\). Otherwise, set \(P_{t_k}(\delta, \delta) = |\mathcal{S}|^{-1};\)
14. \hspace{2em} 3. For every \(b \in \mathcal{B}, s \in \mathcal{S}\) and \(\theta \in \Theta_{b,s}\) such that \(N_k(b, s) > 0\) set \(P_{t_k,\theta}(\delta) = \frac{N_k(b, s, \theta)}{N_k(b, s)}\). Otherwise, set \(P_{t_k,\theta}(\delta) = |\Theta_{b,s}|^{-1};\)
15. \hspace{2em} 4. Calculate the policy \((b_{t_k}^*(\delta), a_{t_k}^*(\delta))\) for every \(\delta \in \mathcal{S}\), where
16. \hspace{2em} \begin{align*}
(b_{t_k}^*(\delta), a_{t_k}^*(\delta)) &= \arg \max_{b \in \mathcal{B}, a \in \mathcal{A}} \left\{ \max_{\theta \in \Theta_{b,s}} \left\{ \sum_{\delta \in \mathcal{S}} P_{t_k}(\delta, \delta) \sum_{\theta \in \Theta_{b,s}} P_{t_k,\theta}(\delta) \langle a, \theta \rangle \right\} \right. \\
&\quad \left. \text{s.t.: } |P_{t_k}(\delta, \delta) - P_{t_k,\theta}(\delta)| \leq \text{conf}_S(t_k, \delta), \forall \delta \in \mathcal{S} \\
&\quad \quad |P_{t_k,\theta}(\delta) - P_{t_k,\theta}(\delta)| \leq \text{conf}_\theta(t_k, b, \delta), \forall \delta \in \mathcal{S} \\
&\quad \quad P_{t_k}(\delta, \delta) \geq 0, \sum_{\delta \in \mathcal{S}} P_{t_k}(\delta, \delta) = 1, \forall \delta \in \mathcal{S} \\
&\quad \quad \sum_{\theta \in \Theta_{b,s}} P_{t_k,\theta}(\delta) = 1, \forall b \in \mathcal{B}, \delta \in \mathcal{S}, \theta \in \Theta_{b,s} \hspace{1em} (8) \right. \\
\end{align*}
17. **Execute round** \(k:\)
18. \hspace{2em} **while**
19. \hspace{3em} • \(\text{conf}_S(t, s) > \text{conf}_S(t_k, s)/2\) for every \(s \in \mathcal{S},\) **and**
20. \hspace{3em} • \(\text{conf}_\theta(t, b, s) > \text{conf}_\theta(t_k, b, s)/2\) for every \(b \in \mathcal{B}, s \in \mathcal{S}\)
21. **do**
22. \hspace{3em} 1. Choose \(b_t = b_{t_k}^*(\delta_{t-1});\)
23. \hspace{3em} 2. Choose \(a_t\) randomly from the set \(B(a_{t_k}^*(\delta_{t-1}), \epsilon_t) \cap \mathcal{A};\)
24. \hspace{3em} 3. Play the pair \((b_t, a_t)\) and observe the reward \(r_t;\)
25. \hspace{3em} 4. Recover system states: set \(\hat{\delta}_t \in \bigcup_{s \in \mathcal{S}} \Theta_{b_t,s}\) to be a solution of \(r_t = \langle a_t, \hat{\delta}_t \rangle\) and set \(\hat{\delta}_t \in \mathcal{S}\) to be such that \(\hat{\delta}_t \in \Theta_{b_t,\hat{\delta}_t};\)
26. **Update:**
27. \hspace{4em} • Set \(N_{t+1}(s) = N_t(s) + 1_{\{s = \delta_{t-1}\}} 1_{\{t \geq 1\}};\)
28. \hspace{4em} • Set \(N_{t+1}(\delta, \delta) = N_t(\delta) + 1_{\{\delta = \delta_{t-1}\}} 1_{\{t \geq 1\}};\)
29. \hspace{4em} • Set \(N_{t+1}((b,s) = (b_t,s_t)) = N_t((b,s) = (b_t,s_t)) 1_{\{t \geq 1\}};\)
30. **end**
Denote \( \mathcal{P}_{\Theta} = \{ \mathcal{P}_{\Theta_{i}} \}_{(b,a) \in \mathcal{B} \times \mathcal{S}} \) and let \( g(b,a,\tilde{s}) = \sum_{s \in \mathcal{S}} \mathcal{P}_{S}(s,\tilde{s}) \sum_{\theta \in \Theta} \mathcal{P}_{\Theta_{i}}(\theta) (a,\theta) \). Since the set of arms \( \mathcal{B} \) is finite and bounded, and the set of states is finite as well, we have that
\[
\min_{b \in \mathcal{B}} \min_{a \in \mathcal{A}} \left[ g(b,a^*,\tilde{s}) - g(b,a,\tilde{s}) \right] > 0. \tag{10}
\]
Thus, there exists \( \delta > 0 \) that fulfills the optimality condition of Lemma\(^1\).

Lemma\(^1\) motivates the development of Algorithm\(^1\), which utilizes upper confidence bounds to establishing an exploration-exploitation trade-off. Interestingly, we show that we can recover the vector \( \theta \) that was generated while forfeiting a negligible amount of reward. Moreover, instead of estimating the reward function for every action in \( \mathcal{A} \) (or equivalently \( V \)) we estimate the probabilities of generating \( \theta \). This has a significant effect on the regret since we assumed that \( |V| \gg |\Theta_{b,s}| \) for all \( (b,s) \in \mathcal{B} \times \mathcal{S} \), such is the case for example when \( \mathcal{A} \) is an \( N \)-dimensional cube, in this case the cardinality of \( V \) is exponential in the dimension \( N \).

**Proof of Theorem\(^1\)**
The expected regret of Algorithm\(^1\) comprises the following events:

1. Regret caused by recovering the previous state.
2. Regret caused by suboptimal rounds in which the confidence intervals are larger than \( \Delta \).
3. Regret caused by failure of the confidence intervals.
4. Regret caused by the deviation of the initial distribution from the stationary distribution of the Markov chain \( P_{S} \).

Next, we show that the expected regret caused by each of these events is no greater than \( \Delta \).

**Regret Caused by Error in State Recovery**

In the restless hidden Markov bandit model the identity of the previous state is not available to the decision maker, thus the decision maker should balance minimizing the expected regret of the current time and learning the current state of the Markov chain. Suppose that the decision maker knows \( s_{t-1} \), it then chooses at time \( t \) the action \( a_{t} = a_{t}^{*}(s_{t-1}) \) and arm \( b_{t} = b_{t}(s_{t-1}) \), calculated in \( \mathcal{P}_{S} \), and receives a reward \( r_{t} = (a_{t}^{*}, \theta_{t}) \). Denote \( \Theta_{b_{t}} = \bigcup_{s \in \mathcal{S}} \Theta_{b_{t},s} \). We distinguish between two cases: 1) \( \theta_{t} \) is the unique solution of \( r_{t} = (a_{t}^{*}, \theta_{t}) \) in \( \Theta_{b_{t}} \). 2) There are multiple solutions to the linear equation \( r_{t} = (a_{t}, \theta) \) in \( \Theta_{b_{t}} \). In the first case, upon receiving the reward \( r_{t} \), the decision maker can fully recover the vector \( \theta_{t} \) and thus also the system state \( s_{t} \). The decision maker can then use this information to maximize the expected reward for the next play. In the second case, after receiving the reward the decision maker cannot distinguish between the different vectors that solve the equation \( r_{t} = (a_{t}, \theta) \) in \( \Theta_{b_{t}} \).

We overcome this uncertainty by choosing an action \( a_{t} \in \mathcal{A} \) instead of \( a_{t}^{*} \) such that the following conditions hold:

\( A_{1} \) \( a_{t} \in B(a_{t}^{*}, \epsilon_{t}) \cap \mathcal{A} \), for some choice of \( \epsilon_{t} > 0 \).

\( A_{2} \) \( (a_{t}, \hat{\theta}) = (a_{t}, \theta_{t}) \) for \( \hat{\theta}, \theta_{t} \in \Theta_{b_{t}} \) if and only if \( \hat{\theta} = \theta_{t} \).

It is clear that the first condition can be fulfilled. We prove that the second condition can be fulfilled as well, simultaneously with the first condition. Let \( \mathcal{D}(b_{t}) = \bigcup_{\hat{\theta}, \theta_{t} \in \Theta_{b_{t}}} \{ a \in \mathcal{A} : \langle a, \hat{\theta} \rangle = \langle a, \theta_{t} \rangle \} \). \( \mathcal{D}(b_{t}) \) is contained in the union of \( \Theta_{b_{t}} \) \( |\Theta_{b_{t}}| - 1 \)/2 hyperplanes of dimension \( N - 1 \) whereas the set \( B(a_{t}^{*}, \epsilon_{t}) \cap \mathcal{A} \) is \( N \)-dimensional. Therefore the intersection of \( \mathcal{D}(b_{t}) \) with \( B(a_{t}^{*}, \epsilon_{t}) \cap \mathcal{A} \) has measure \( 0 \). Thus, the random choice of the action \( a_{t} \) from the set \( B(a_{t}^{*}, \epsilon_{t}) \cap \mathcal{A} \) fulfills condition \( A_{2} \) with probability one. For each such a vector \( a_{t} \), we have that \( |\langle a_{t}, \hat{\theta} \rangle - \langle a_{t}, \theta_{t} \rangle| = |\langle a_{t}, \theta_{t} \rangle - \langle a_{t}, \hat{\theta} \rangle| \leq \epsilon_{t} \cdot \max_{\theta \in \Theta_{b_{t}}} \| \theta \|_{1} \leq \epsilon_{t} \cdot \max_{\theta \in \Theta_{b_{t}}} \| \theta \|_{1} \).

Finally, we choose \( \epsilon_{t} = \frac{\epsilon}{\gamma \cdot r^{t} \cdot \max_{\theta \in \Theta_{b_{t}}} \| \theta \|_{1}} \) for all \( t \) where \( \gamma \) is a constant bigger than the finite sum \( \sum_{t=0}^{\infty} r^{t} \). Since \( \sum_{t=0}^{\infty} \epsilon_{t} \cdot \max_{\theta \in \Theta_{b_{t}}} \| \theta \|_{1} < \epsilon \), the expected regret caused by the state recovery process is smaller than \( \epsilon \) with probability one. Thus, hereafter we assume that the previous state \( s_{t-1} \) is known to the decision maker when the choice of the arm and action at time \( t \) are made.

**Regret Caused by Suboptimal Rounds**

Next we bound the expected regret caused by suboptimal rounds in which the lengths of the confidence intervals are greater than \( \Delta \). To analyze this expected regret we first present the following propositions.

**Proposition 1.** Let \( t_{0} \) be the starting time of round \( k \). For every \( s \in \mathcal{S} \) and \( t, t_{0} > 0 \) such that \( t > t_{0} \), if \( \text{conf}_s(t, s) \leq \frac{1}{2} \text{conf}_s(t_{0}, s) \), then \( N_{t}(s) > 4N_{t_{0}}(s) \). Additionally, for every \( s \in \mathcal{S}, b \in \mathcal{B} \) and \( t, t_{0} > 0 \), if \( \text{conf}_b(t, b, s) \leq \frac{1}{2} \text{conf}_b(t_{0}, b, s) \), then \( N_{t}(s, b) > 4N_{t_{0}}(s, b) \).

**Proposition 2.** In \( T \) time instants there are at most \( |\mathcal{S}| \cdot |\mathcal{B}| \cdot \log_2 \left( \frac{T}{|\mathcal{S}| \cdot |\mathcal{B}|} + 1 \right) \) rounds.

**Proposition 3.** If \( N_{t}(s) > \frac{\log(4(1-\epsilon)^{\epsilon}|\mathcal{S}|^{2})}{\Delta^{2}} \) then the confidence interval for \( s \) is smaller than \( \Delta \). Further, if \( N_{t}(b, s) > \frac{\log(4(1-\epsilon)^{\epsilon}|\Theta_{b,s}| \cdot |\mathcal{S}|)}{2 \Delta^{2}} \) then the confidence interval for \( (b, s) \) is smaller than \( \Delta \).

Suppose that \( k \) is a suboptimal round, then at least one of the following two error events occurs:

1. There exist \( \tilde{s}, \bar{s} \in \mathcal{S} \) such that \( |\hat{P}_{t_{k}}(\tilde{s}, \tilde{s}) - P_{s}(\tilde{s}, \tilde{s})| > \Delta \)

2. Suppose that the policy for round \( k \) chooses the arm \( b \) whenever state \( \bar{s} \) is observed, then there exist \( s \in \mathcal{S} \), an arm \( b \) and \( \theta \in \Theta_{b,s} \) such that \( |\hat{P}_{t_{k}}(\bar{s}, \bar{s}) - P_{s,b}(\bar{s}, \bar{s})| > \Delta \)

The expected regret that is caused by the first error event is upper-bounded by the term
\[
2c r_{\max} |\mathcal{S}| T \log \left( \frac{4T^{c}}{|\mathcal{S}| \cdot |\mathcal{B}|^{2}} \right) \Delta^{2} + 2c r_{\max} T \log_2 \left( \frac{T}{|\mathcal{S}| \cdot |\mathcal{B}|} + 1 \right) + r_{\max} |\mathcal{S}| \tag{11}
\]
where \( c \) is a constant satisfying \( c < 11 \). This is a direct result of the analysis presented in [Auer and Ortner (2007)] and Propositions\(^{13}\).
The expected regret caused by the second event can be upper bounded as follows. Denote by $S_b$ the set of states which upon observing, the decision maker plays the arm $b$. Suppose that there is $\theta \in \Theta_{b,s}$ such that $|\hat{P}_{\theta,\theta_{s,s}}(\theta) - P_{\theta,\theta_{s,s}}(\theta)| > \Delta$ for given $b$ and $s$ such that $S_b$ is not empty. Let $n(b,s)$ be the number of such rounds and let $\tau_1(b,s), \ldots, \tau_{n(b,s)}$ be their respective lengths. We next upper bound the expected value of the term $\sum_{i=1}^{n(b,s)} \tau_i(b,s)$ by dividing each suboptimal round $i$ into $\lceil \frac{\tau_i(b,s)}{2T_{\|S\|}} \rceil$ sub-intervals. By the Markov inequality the probability of reaching a state in $S_b$, playing the arm $b$, and then immediately reaching the state $s$ is, at least $\frac{1}{4}$, for each of these sub-intervals. Thus, by the Chernoff-Hoeffding inequality we have that:

$$\Pr \left( N(b,s,m) \geq \frac{m}{2} - \sqrt{\frac{m \log T}{2}} \right) \geq 1 - \frac{1}{T} \quad (12)$$

where $N(b,s,m)$ is the number of times where the arm $b$ was chosen and then the state $s$ was immediately observed in $m$ sub-intervals.

By Proposition 3 we have that $N_T(b,s) < 2\alpha$, since $N(b,s,m) \leq N_T(b,s)$ it follows that:

$$\sum_{i=1}^{n(b,s)} \frac{\tau_i(b,s)}{2T_{\|S\|}} \leq \frac{1}{\alpha} \log \left( \frac{4\alpha c \theta_{b,s} |\|B\|||S\|} {2\alpha} \right)$$

for some constant $\alpha < 11$ with probability $1 - \frac{1}{T}$. It follows that

$$\sum_{i=1}^{n(b,s)} \tau_i(b,s) \leq 2T_MT_{|S|} c \log \left( \frac{4\alpha c \theta_{b,s} |\|B\|||S\|} {2\alpha} \right) + 2T_MT_S n(b,s)$$

$$\leq 2T_MT_{|S|} c \log \left( \frac{4\alpha c \theta_{b,s} |\|B\|||S\|} {2\alpha} \right)$$

$$+ 2T_MT_S \log \left( \frac{T}{|\|S\|||B\|} + 1 \right) + 1 \right), \quad (13)$$

with probability $1 - \frac{1}{T}$, where the last inequality follows by Proposition 2.

Finally, denote $C_{\theta_{b,s}} = \max_{\theta \in \Theta_{b,s}} \{ \theta_{b,s} \}$, then the expected regret is:

$$|\|B\|||S\| r_{\max} T \frac{1}{T} + 2|\|B\|||S\| T_MT_{|S|} r_{\max} c \log \left( \frac{4\alpha c \theta_{b,s} |\|B\|||S\|} {2\alpha} \right)$$

$$+ 2|\|B\|^2 |S| T_MT_S r_{\max} \left( \log \left( \frac{T}{|\|S\|||B\|} + 1 \right) + 1 \right). \quad (14)$$

### Regret Caused by Failure of the Confidence Intervals

Next we upper bound the expected regret caused by the failure of the confidence intervals, i.e., the probability distributions that we estimate are outside the confidence intervals.

Recall that $t_k$ is the starting time of round $k$; by (9), the probability that one of the confidence intervals fails in round $k$ is:

$$|\|S\|^2 \frac{(t_k - 1)^\alpha}{2|\|S\|^2} + \sum_{(b,s) \in \mathcal{B} \times \mathcal{S}} \theta_{b,s} \frac{(t_k - 1)^\alpha}{2\theta_{b,s} |\|B\|||S\|} = (t_k - 1)^\alpha.$$
Setup 1: $\mathcal{A} = \{0, 1\}^2$, $\mathcal{B} = \{1, 2\}$, and $\mathcal{S} = \{1, 2\}$. Transition probability: $P_S = \begin{bmatrix} 0.4 & 0.3 \\ 0.5 & 0.5 \end{bmatrix}$. $|\Theta_{b,s}| = 2$ for every $b \in \mathcal{B}, s \in \mathcal{S}$. The vectors $\theta \in \Theta_{b,s}$ were drawn uniformly from the set $\{-7, -6, \ldots, 10\}^2$. We also set the values of the following probability distributions $P_{\theta_{b,s,1}} = (0.4, 0.6)$, $P_{\theta_{b,s,2}} = (0.7, 0.3)$, $P_{\theta_{b,s,3}} = (0.5, 0.5)$.

Setup 2: $\mathcal{A} = \{0, 1\}^5$, $\mathcal{B} = \{1, 2, 3\}$, and $\mathcal{S} = \{1, 2, 3\}$. Transition probability: $P_S = \begin{bmatrix} 0.4 & 0.6 & 0.25 & 0.25 & 0.3 \\ 0.2 & 0.5 & 0.25 & 0.3 & 0.3 \\ 0.25 & 0.5 & 0.25 & 0.3 & 0.3 \\ 0.25 & 0.5 & 0.25 & 0.3 & 0.3 \\ 0.25 & 0.5 & 0.25 & 0.3 & 0.3 \end{bmatrix}$. $|\Theta_{b,s}| = 2$ for every $b \in \mathcal{B}, s \in \mathcal{S}$. The vectors $\theta \in \Theta_{b,s}$ were drawn uniformly from the set $\{-7, -6, \ldots, 10\}^5$. We also set the values of the following probability distributions:

- $P_{\theta_{b,s,1}} = (0.4, 0.6)$,
- $P_{\theta_{b,s,2}} = (0.7, 0.3)$,
- $P_{\theta_{b,s,3}} = (0.5, 0.5)$,
- $P_{\theta_{b,s,4}} = (0.45, 0.55)$,
- $P_{\theta_{b,s,5}} = (0.1, 0.9)$,
- $P_{\theta_{b,s,6}} = (0.32, 0.68)$.

We also set the following values $\epsilon = 0.5$, $\alpha = 2.5$, $\alpha_\epsilon = 1.5$, $\gamma = 1$. We ran a Monte Carlo simulation with 100 realizations of the sets $\Theta_{b,s}$, for each such realization we generated 20 realizations of the state sequence, and their respective $\theta$ given the choice of arm $b$. Finally we set $T = 10^6$.

Figures 1a and 1b depict the average regret of each of the schemes that we mentioned at the beginning of this section, that is, Algorithm 1 with no maximization over $P_S$ when solving the problem (8), Algorithm 2 an adaptation of Algorithm 1 with confidence intervals for $P(\hat{s}, b, \hat{s}, \theta)$, and an adaptation of Algorithm 1 with confidence intervals for $P(\hat{s}, b, a, \hat{s}, \theta)$, see Auer and Ortner (2007).

Figures 1a and 1b show that Algorithm 1, with or without maximization over $P_S$, outperforms both the aforementioned possible schemes. This leads to the conclusion that separating the estimation of the probability distributions into two groups, one that is common to all arms (the transition matrix), and one that depends on the identity of the arm played (the probability distribution of $\theta$) decreases the regret. Additionally, we note that utilizing the information regarding the reward function significantly decreases the regret, in our model it removes the dependency on the cardinality of the action set that may be large. Additionally, Figures 1a and 1b confirm that our state recovery scheme is indeed correct. Finally, Figures 1a and 1b show that setting $\hat{P}_S = \hat{P}_S$ instead of maximizing over $P_S$ in Algorithm 1 did not increase the regret, this follows from Theorem 1.2 in Kontorovich and Ramanan (2008) and Theorem 1.1 in Lezaud (1998) that bound the probability of deviating from the true transition probability.

Conclusion

This work presented the restless hidden Markov bandit model with linear rewards in which the action of a decision maker does not affect the Markov process that governs the state of the system. Additionally, the system state is not revealed to the decision maker, but rather it is estimated from the previous actions and arms played and their respective rewards. We showed that by increasing the regret by an arbitrarily small value (independent of $T$) the decision maker can learn the state of the system. Furthermore, we also developed an algorithm that takes advantage of the linear structure of the reward function and yields logarithmic regret that does not depend on the size of the action space (which can be exponential with the number of dimensions). This is a significant improvement to a naive implementation of existing algorithm for Markov decision processes and restless Markovian bandits.
Supplementary Material

Proving of Propositions 1-3

Proof of Proposition 1. Recall that $t_k$ is the starting time of round $k$ and that $t > t_k$. We separate the proof for the cases of $\text{conf}(t_k, \emptyset) = 1$ and $\text{conf}(t_k, \emptyset) < 1$.

Suppose that $\text{conf}(t_k, s) < 1$ and that $\text{conf}(t, s) \leq \frac{1}{2} \text{conf}(t_k, s)$. Then

$$\sqrt{\frac{\log (4(t-1)^t |S|^2)}{2N_t(s)}} \leq \frac{1}{2} \sqrt{\frac{\log (4(t_k-1)^t |S|^2)}{2N_t(s)}}$$

$$\iff \frac{\log (4(t-1)^t |S|^2)}{N_t(s)} \leq \frac{1}{4} \frac{\log (4(t_k-1)^t |S|^2)}{N_t(s)}$$

$$\iff 4 \cdot \frac{\log (4(t-1)^t |S|^2)}{\log (4(t_k-1)^t |S|^2)} \leq \frac{N_t(s)}{N_t(s)}$$

Since $\frac{\log (4(t-1)^t |S|^2)}{\log (4(t-1)^t |S|^2)} > 1$ we have that $\frac{N_t(s)}{N_t(s)} \geq 4$.

Now, if $\text{conf}(t_k, s) = 1$, then $\text{conf}(t, s) < \frac{1}{2}$. Thus, if $N_t(s) = 0$ then $N_t(s) \geq 4N_t(s)$. Else, if $N_t(s) > 0$ then

$$\sqrt{\frac{\log (4(t-1)^t |S|^2)}{2N_t(s)}} \leq \frac{1}{2} \sqrt{\frac{\log (4(t_k-1)^t |S|^2)}{2N_t(s)}}$$

and we concluded above that in this case $N_t(s) \geq 4$.

The proof of the second part of the proposition is similar.

Proof of Proposition 2. First, note that $N_t(s) \geq N_t(b, s)$ for every $s \in S$, $b \in B$. Thus, by Proposition 1, for each round $k$ the shortest possible length of this round is four times the value of $\min_{b, s} \{N_t(b, s)\}$. It follows that the number of rounds can be upper-bounded by $|S||B|\rho_{\text{max}}$ where $\rho_{\text{max}}$ is the smallest positive integer such that $T = |S||B|\sum_{i=1}^{\rho_{\text{max}}} 4^i$. It follows that $\rho_{\text{max}}$ is the smallest positive integer greater than $\log_4 \left( 1 + \frac{3T}{|S||B|} \right)$. Now, since

$$\log_4 \left( 1 + \frac{3T}{|S||B|} \right) = \frac{1}{2} \log_2 \left( 1 + \frac{3T}{|S||B|} \right)$$

$$\leq \log_2 \left( 1 + \frac{T}{|S||B|} \right) + 1$$

we have that the number of rounds is upper bounded by $|S||B| \log_2 \left( 1 + \frac{T}{|S||B|} \right) + 1$.

Proof of Proposition 3. This is a direct result of the definitions: $\text{conf}_S(t, s) \triangleq \min \left\{ 1, \sqrt{\frac{\log (4(t-1)^t |S|^2)}{2N_t(s)}} \right\}$ and $\text{conf}_\emptyset(t, b, s) \triangleq \min \left\{ 1, \sqrt{\frac{\log (4(t-1)^t |\emptyset| B|S|^2)}{2N_t(b, s)}} \right\}$.

If $N_t(s) > \frac{\log(4(t-1)^t |S|^2)}{2\Delta^2}$, then

$$\text{conf}_S(t, s) = \min \left\{ 1, \sqrt{\frac{\log (4(t-1)^t |S|^2)}{2N_t(s)}} \right\}$$

$$\leq \min \left\{ 1, \frac{\log(4(t-1)^t |S|^2)}{2\Delta^2} \right\}$$

$$= \min \{1, \Delta \} = \Delta.$$
Finally, by the union bound over $\tilde{s}$ we have that the expected regret caused by suboptimal rounds in which the estimation of the transition probability is inaccurate is upper bounded by:

$$2c r_{\max} |S|^2 T M \frac{\log (4(T-1)^{\alpha} |S|^2)}{\Delta^2} + 2 r_{\max} T M |S|^2 |B| \left[ \log_2 \left( \frac{T}{|S||B|} + 1 \right) + 1 \right] + r_{\max} |S|.$$  

(23)

□

**Proving Lemma 2**

*Proof of Lemma 2.* Since $T \rho(\pi^*) = \sum_{t=1}^{T} E[r_t(b_t^*, a_t^*)]$, we lower bound the term $\sum_{t=1}^{T} E[r_t(b_t^*, a_t^*)]$. Recall that $(b_t^*(\tilde{s}), a_t^*(\tilde{s})) = \pi^*(\tilde{s})$. Since $\mu_S$ is the stationary distribution of the Markov chain $P_S$ we have that

$$T \rho(\pi^*)$$

$$= \sum_{t=1}^{T} \sum_{\tilde{s}, \tilde{s} \in S} \mu_S(\tilde{s}) P_S^{t-1}(\tilde{s}, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle$$

$$= \sum_{\tilde{s} \in S} \mu_S(\tilde{s}) \sum_{t=1}^{T} \sum_{\tilde{s}, \tilde{s} \in S} P_S^{t-1}(\tilde{s}, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle. \quad (24)$$

Thus, there exists $s \in S$ such that

$$\sum_{t=1}^{T} \sum_{\tilde{s}, \tilde{s} \in S} P_S^{t-1}(s, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle \geq T \rho(\pi^*), \quad (25)$$

Thus, for every $1 \leq \tau \leq T$,

$$\sum_{t=\tau+1}^{T} \sum_{\tilde{s}, \tilde{s} \in S} P_S^{t-1}(s, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle$$

$$\geq T \rho(\pi^*) - \sum_{t=1}^{\tau} \sum_{s, \tilde{s} \in S} P_S^{t-1}(s, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle. \quad (26)$$

Now, let $t_s$ be the first occurrence of state $s$ that fulfills (25), then for every $\tilde{s} \in S$ we have that

$$E \sum_{t=1}^{T} \sum_{\tilde{s}, \tilde{s} \in S} P_S^{t-1}(\tilde{s}, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle$$

$$E_{t_s} \left[ E \sum_{t=1}^{T} \sum_{\tilde{s}, \tilde{s} \in S} P_S^{t-1}(\tilde{s}, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle | t_s \right]$$

$$= E_{t_s} \sum_{t=1}^{t_s} \sum_{\tilde{s}, \tilde{s} \in S} P_S^{t-1}(s, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle$$

$$+ E_{t_s} \left[ E \sum_{t=t_s+1}^{T} \sum_{\tilde{s}, \tilde{s} \in S} P_S^{t-1}(s, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle | t_s \right]$$

$$\geq E_{t_s} \sum_{t=1}^{t_s} \sum_{\tilde{s}, \tilde{s} \in S} P_S^{t-1}(s, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle$$

$$+ T \rho(\pi^*) - E_{t_s} \sum_{t=1}^{t_s} \sum_{\tilde{s}, \tilde{s} \in S} P_S^{t-1}(s, \tilde{s}) P_S(\tilde{s}, \tilde{s})$$

$$\cdot \sum_{\theta \in \Theta_{b_t^*, \tilde{s}}} P_{b_t^*, \tilde{s}}(\theta) \langle a_{\pi^*}(\tilde{s}), \theta \rangle$$

$$\geq T \rho(\pi^*) - E(t_s) r_{\max}. \quad (27)$$

where the inequality (a) follows from (26) and the inequality (b) follows from the notation $r_{\max} = \max_{a, \tilde{a} \in \mathcal{A}, \theta, \tilde{\theta} \in [1]_{|b_t^*, \tilde{a}_t^*, \tilde{s}}, \Theta_{b_t^*, \tilde{s}}} \{ (a, \theta) - (\tilde{a}, \tilde{\theta}) \}$ that appears before Theorem 1 in the main manuscript.

We can conclude the proof by following inequalities

$$T \rho(\pi^*) - \sum_{t=1}^{T} E[r_t(\pi^*(s_t-1))]$$

$$= T \rho(\pi^*) - \sum_{t=1}^{T} E[t_s \{ E[r_t(\pi^*(s_t-1))] | t_s \}]$$
where (a) follows by (27) and since we assume in Lemma 2 that \( (b^*_t, a^*_t) = \pi^*(s_{t-1}) \), and (b) follows by the notation \( T_M = \max_{s_i \in S} E(T_i | s_i) \) that appears before Theorem 1 in the main manuscript.

\[
(a) \leq T \rho(\pi^*) - [T \rho(\pi^*) - E(t_s)r_{max}] \\
(b) \leq r_{max} T_M, \tag{28}
\]

where we can rewrite (29) as

\[
E[R(T)] = E \left[ T \rho(\pi^*) - \sum_{t=1}^{T} r_t(b_t, a_t) \right]. \tag{29}
\]

where \( b_t, a_t \) are played according to Algorithm 1 and \( \pi^* \) is the optimal policy that maximizes (2) in the manuscript. Now, we can rewrite (29) as

\[
E[R(T)] = E \left[ T \rho(\pi^*) - \sum_{t=1}^{T} r_t(b_t^*, a_t^*) \right] \\
+ \sum_{t=1}^{T} r_t(b_t, a_t) - \sum_{t=1}^{T} r_t(b_t, a_t) \tag{30}
\]

where \( (b_t^*, a_t^*) \) denotes playing the optimal policy assuming that the decision maker knows the identity of the previous state and \( (b_t, a_t) \) is the arm and action choices when playing according to Algorithm 1.

Now, by the analysis of the regret caused by the deviation of the initial distribution from the stationary distribution of the Markov chain \( P_S \), we have that

\[
E \left[ T \rho(\pi^*) - \sum_{t=1}^{T} r_t(b_t^*, a_t^*) \right] \leq T \rho(\pi^*) - E \left[ \sum_{t=1}^{T} r_t(b_t^*, a_t^*) \right] \leq r_{max} T_M |S||B| \log_2 \left( \frac{T}{|S||B|} + 1 \right) + 1. \tag{31}
\]

Now, the term \( E \left[ \sum_{t=1}^{T} r_t(b_t^*, a_t^*) - \sum_{t=1}^{T} r_t(b_t, a_t) \right] \) depends on the three other regret events, that is,

- Regret caused by recovering the previous state.
- Regret caused by suboptimal rounds in which the confidence intervals are larger than \( \Delta \).
- Regret caused by failure of the confidence intervals.

We prove in the manuscript that the expected regret caused by first event is bounded, i.e., \( O(1) \), the expected regret of the second event is upper bounded by

\[
2T_M |S| c \log(4T^a |S|^2) \frac{2\Delta^2}{\Delta^2} + 2T_M |S|^2 |B| \log_2 \left( \frac{T}{|S||B|} + 1 \right) + 1 \\
+ r_{max} |S| + 2|B||S|T_M T_S r_{max} c \log(4T^a C \omega_{max} |B||S|) \frac{2\Delta^2}{\Delta^2} + 2|B|^2 |S|^2 T_M T_S r_{max} \log_2 \left( \frac{T}{|S||B|} + 1 \right) + 1 \\
+ |B||S| r_{max} T \frac{1}{T}. \tag{32}
\]

Additionally, we prove in the manuscript that the expected regret caused by third event is bounded, i.e., \( O(1) \).

This proves that the expected regret of Algorithm 1 is:

\[
O \left( |B||S| T_M T_S r_{max} \frac{\log(4T^a C \omega_{max} |B||S|)}{\Delta^2} + |B|^2 |S|^2 T_M T_S r_{max} \log_2 \left( \frac{T}{|S||B|} + 1 \right) \right). \tag{33}
\]

**References**

Anantharam, V.; Varaiya, P.; and Walrand, J. 1987. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part II: Markovian rewards. *IEEE Transactions on Automatic Control* 32(11):977–982.

Auer, P., and Ortner, R. 2007. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems* 19, 49–56. MIT Press.

Auer, P. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research (JMLR)* 3:397–422.

Bartlett, P. L., and Tewari, A. 2009. REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *UAI*.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory (COLT)*.

Filippi, S.; Cappe, O.; and Garivier, A. 2011. Optimally sensing a single channel without prior information: The tiling algorithm and regret bounds. *IEEE Journal of Selected Topics in Signal Processing* 5(1):68–76.

Fruit, R.; Pirotta, M.; and Lazaric, A. 2018. Near optimal exploration-exploitation in non-communicating Markov decision processes. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS’18, 2998–3008.
Hanawal, M. K.; Leshem, A.; and Saligrama, V. 2016. Efficient algorithms for linear polyhedral bandits. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4796–4800.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. J. Mach. Learn. Res. 11:1563–1600.

Javidi, T. 2016. Information acquisition and sequential belief refinement. In 2016 IEEE 55th Conference on Decision and Control (CDC), 7635–7654.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. Artif. Intell. 101:99–134.

Kontorovich, L. A., and Ramanan, K. 2008. Concentration inequalities for dependent random variables via the martingale method. Ann. Probab. 36(6):2126–2158.

Krishnamurthy, V., and Evans, R. J. 2001. Hidden Markov model multiarm bandits: a methodology for beam scheduling in multitarget tracking. IEEE Transactions on Signal Processing 49(12):2893–2908.

Krishnamurthy, V., and Evans, R. J. 2003. Correction to “Hidden Markov model multiarm bandits: a methodology for beam scheduling in multitarget tracking”. IEEE Transactions on Signal Processing 51(6):1662–1663.

Krishnamurthy, V., and Wahlberg, B. 2009. Partially observed Markov decision process multiarmed bandits—structural results. Mathematics of Operations Research 34(2):287–302.

Levin, D. A.; Peres, Y.; and Wilmer, E. L. 2009. Markov chains and mixing times. Providence, R.I. American Mathematical Society.

Lezaud, P. 1998. Chernoff-type bound for finite Markov chains. Ann. Appl. Probab. 8(3):849–867.

Meshram, R.; Gopalan, A.; and Manjunath, D. 2017. A hidden Markov restless multi-armed bandit model for playout recommendation systems. In Communication Systems and Networks, 335–362. Cham: Springer International Publishing.

Ortner, R.; Ryabko, D.; Auer, P.; and Munos, R. 2014. Regret bounds for restless Markov bandits. Theoretical Computer Science 558:62 – 76.

Ortner, R. 2018. Regret bounds for reinforcement learning via Markov chain concentration. CoRR abs/1808.01813.

Qian, J. Z.; Fruit, R.; Pirotta, M.; and Lazaric, A. 2018. Exploration bonus for regret minimization in undiscounted discrete and continuous Markov decision processes. CoRR abs/1812.04363.

Rusmevichientong, P., and Tsitsiklis, J. N. 2010. Linearly parameterized bandits. Mathematics of Operations Research 35(2):395–411.

Tekin, C., and Liu, M. 2012. Online learning of rested and restless bandits. IEEE Transactions on Information Theory 58(8):5588–5611.

Urteaga, I., and Wiggins, C. 2018. Nonparametric gaussian mixture models for the multi-armed contextual bandit. CoRR abs/1808.02932.

Wang, J.; Ren, X.; Mo, Y.; and Shi, L. 2019. Whittle index policy for dynamic multi-channel allocation in remote state estimation. IEEE Transactions on Automatic Control.