Regret Bounds for Thompson Sampling in Restless Bandit Problems

Young Hun Jung [1] Ambuj Tewari [1]

Abstract

Restless bandit problems are instances of non-stationary multi-armed bandits. There are plenty of results from the optimization perspective, which aims to efficiently find a near-optimal policy when system parameters are known, but the learning perspective where the parameters are unknown is rarely investigated. In this paper, we analyze the performance of Thompson sampling in restless bandits with unknown parameters. We consider a general policy map to define our competitor and prove an $\tilde{\mathcal{O}}(\sqrt{T})$ Bayesian regret bound. Our competitor is flexible enough to represent various benchmarks including the best fixed action policy, the optimal policy, the Whittle index policy, or the myopic policy. The empirical results also support our theoretical findings.

1. Introduction

Restless bandits (Whittle, 1988) are a variant of multi-armed bandit (MAB) problems (Robbins, 1952). Unlike the classical MABs, the arms have non-stationary reward distributions. Specifically, we will focus on the class of restless bandits whose arms change their states based on Markov chains. Restless bandits are also distinguished from *rested bandits* where only the active arms evolve and the passive arms remain frozen. We will assume that each arm changes according to two different Markov chains depending on whether it is played or not. Because of their extra complexity, restless bandits can model more practical problems such as dynamic channel access (Liu et al., 2011; 2013) or online recommendation system (Meshram et al., 2017).

Due to the arms' non-stationary nature, playing the same set of arms for every round usually does not produce the optimal performance. This makes the optimal policy highly non-trivial, and Papadimitriou & Tsitsiklis (1999) show that it is generally PSPACE hard to identify the optimal policy for restless bandits. As a consequence, many researchers have been devoted to find an efficient way to approximate the optimal policy (Liu & Zhao, 2010; Meshram et al., 2018). This line of work primarily focuses on the *optimization* perspective in that the system parameters are already known.

Since the true system parameters are unavailable in many cases, the *learning* perspective for the restless bandits is necessary. Due to the learner's additional uncertainty, however, analyzing a learning algorithm in restless bandits is significantly challenging. Liu et al. (2011; 2013) and Tekin & Liu (2012) prove $\mathcal{O}(\log T)$ bounds for confidence bound based algorithms, but their competitor always selects a fixed set of actions, which is known to be weak. Dai et al. (2011; 2014) show $\mathcal{O}(\log T)$ bound against the optimal policy, but their assumptions on the underlying model is very limited. Ortner et al. (2012) prove $\tilde{\mathcal{O}}(\sqrt{T})$ bound in the general restless bandits, but their algorithm is infeasible in general.

On a different path, Osband et al. (2013) study Thompson sampling in the fully observable Markov decision process (MDP) and show the Bayesian regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$. Unfortunately, this result is not applicable in our setting as ours is partially observable due to bandit feedback. Following Ortner et al. (2012), it is possible to transform our setting to the fully observable case, but then we end up having exponentially many states, which restrict the practical utility of existing results.

In this work, we analyze Thompson sampling in restless bandits where the system resets every episode of a fixed length and the rewards are binary. We directly tackle the partial observability and achieve a meaningful regret bound, which also matches the result in the classical MAB. We are not the first to analyze Thompson sampling in restless bandits, and Meshram et al. (2016) study this type of algorithm as well, but their regret analysis remains in the one-armed-case with a fixed reward of not pulling the arm. They explicitly mention that regret analysis of Thompson sampling in the multi-armed case is an interesting open question.

2. Problem Setting

We begin by introducing our setting. There are K arms, and the algorithm selects N arms every round. We denote the learner's action at time t by a binary vector $A_t \in \{0,1\}^K$

[1] Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA. Correspondence to: Young Hun Jung <yhjung@umich.edu>, Ambuj Tewari <tewaria@umich.edu>.

Real-world Sequential Decision Making workshop at ICML 2019.
Copyright 2019 by the author(s).

where $||A_t||_1 = N$. We call the selected arms as *active* and the rest as *passive*. We assume each arm k has binary states, $\{0, 1\}$, which evolve as a Markov chain with transition matrix either P_k^{active} or P_k^{passive}, depending on whether the learner pulled the arm or not.

At round t, pulling an arm k incurs a binary reward $X_{t,k}$, which is the arm's current state. As we are in the bandit setting, the learner only observes the rewards of active arms, which we denote by X_{t,A_t}, and does not observe the passive arms' rewards nor their states. This feature makes our setting to be a *partially observable Markov decision process*, or POMDP. We denote the history of the learner's actions and rewards up to time t by $\mathcal{H}_t = (A_1, X_{1,A_1}, \cdots, A_t, X_{t,A_t})$.

We assume the system resets every episode of length L, which is also known to the learner. This means that at the beginning of each episode, the states of the arms are drawn from an initial distribution. The entire time horizon is denoted by T, and for simplicity, we assume it is a multiple of L, or $T = mL$.

2.1. Bayesian Regret and Competitor Policy

Let $\theta \in \Theta$ denote the entire parameters of the system. It includes transition matrices P^{active} and P^{passive}, and an initial distribution of each arm's state. The learner does not have the knowledge of these parameters at the beginning.

In order to define a regret, we need a competitor policy, or a benchmark. We first define a class of deterministic policies and policy mappings.

Definition 1. *A deterministic policy π takes time index and history $(t, \mathcal{H}_{t-1})$ as an input and outputs a fixed action $A_t = \pi(t, \mathcal{H}_{t-1})$. A deterministic policy mapping μ takes system parameters θ as an input and outputs a deterministic policy $\pi = \mu(\theta)$.*

We fix a deterministic policy mapping μ and let our algorithm compete against a deterministic policy $\pi^\star = \mu(\theta^\star)$, where $\theta^\star$ represents the true system parameters, which are unknown to the learner.

We keep our competitor policy abstract mainly because we are in the non-stationary setting. Unlike the classical (stationary) MABs, pulling the same set of arms with the largest expected rewards is not necessarily optimal. Moreover, it is in general PSPACE hard to compute the optimal policy when $\theta^\star$ is given. Regarding these statements, we refer the readers to the book by Gittins et al. (1989). As a consequence, researchers have identified conditions that the myopic policy is optimal (Ahmad et al., 2009) or proven that an efficient index-based policy has a reasonable performance against the optimal policy (Liu & Zhao, 2010).

We observe that most of proposed policies including the optimal policy, the myopic policy, or the index-based policy are deterministic. Therefore, researchers can plug in whatever competitor policy of their choice, and our regret bound will apply as long as the chosen policy mapping is deterministic.

Before defining the regret, we introduce a *value function*

$$V_{\pi,i}^{\theta}(\mathcal{H}) = \mathbb{E}_{\theta,\pi}[\sum_{j=i}^{L} A_j \cdot X_j | \mathcal{H}].$$

This is the expected reward of running a policy π from round i to L where the system parameter is θ and the starting history is $\mathcal{H}$. Note that the competitor policy $\pi^\star$ obtains $V_{\pi^\star,1}^{\theta^\star}(\emptyset)$ rewards per episode in expectation. Thus, the regret can be written as

$$R(T; \theta^\star) = m V_{\pi^\star,1}^{\theta^\star}(\emptyset) - \mathbb{E}_{\theta^\star} \sum_{t=1}^{T} A_t \cdot X_t. \tag{1}$$

We are particularly interested in bounding the following *Bayesian regret*,

$$BR(T) = \mathbb{E}_{\theta^\star \sim Q} R(T; \theta^\star),$$

where Q is a prior distribution over the set of system parameters Θ. The prior is assumed to be known to the learner.

3. Algorithm

Our algorithm is an instance of *Thompson sampling*, first proposed by Thompson (1933). At the beginning of episode l, the algorithm draws system parameters θ_l from posterior and plays $\pi_l = \mu(\theta_l)$ throughout the episode. Once an episode is over, it updates posterior based on additional observations. Algorithm 1 describes the steps.

We want to point out that the history $\mathcal{H}$ fulfills two different purposes. One is to update posterior Q_l, and the other is as an input to a policy π. For the latter, however, we don't need the entire history as the arms reset every episode. That is why we set $\mathcal{H}_0 = \emptyset$ (step 6) and feed $\mathcal{H}_{t-1}$ to π_l (step 8). Furthermore, as we assume that the arms evolve based on Markov chains, the history $\mathcal{H}_{t-1}$ can be summarized as

$$(r_1, n_1, \cdots, r_K, n_K), \tag{2}$$

which means that the arm k is played n_k rounds ago and r_k is the observed reward in that round. If an arm k is never played in the episode, then n_k becomes t and r_k becomes the expected reward from the initial distribution.

4. Regret Bound

In this section, we prove that the Bayesian regret of Algorithm 1 is at most $\tilde{\mathcal{O}}(\sqrt{T})$. The following lemma decomposes the regret.

Algorithm 1 Thompson Sampling in Restless Bandits

```
Input prior $Q$, episode length $L$, policy mapping $\mu$
Initialize posterior $Q_1 = Q$, history $\mathcal{H} = \emptyset$
for episodes $l = 1, \cdots, m$ do
    Draw parameters $\theta_l \sim Q_l$
    Compute the policy $\pi_l = \mu(\theta_l)$
    Set $\mathcal{H}_0 = \emptyset$
    for $t = 1, \cdots, L$ do
        Select $N$ active arms $A_t = \pi_l(t, \mathcal{H}_{t-1})$
        Observe rewards $X_{t,A_t}$ and update $\mathcal{H}_t$
    end for
    Append $\mathcal{H}_L$ to $\mathcal{H}$
    Update posterior distribution $Q_{l+1}$ using $\mathcal{H}$
end for
```

Lemma 2. *The Bayesian regret of Algorithm 1 can be decomposed as below*

$$\begin{aligned}BR(T) &= \mathbb{E}_{\theta^\star \sim Q} \sum_{l=1}^{m} \mathbb{E}_{\theta_l \sim Q_l}[V^{\theta^\star}_{\pi^\star,1}(\emptyset) - V^{\theta^\star}_{\pi_l,1}(\emptyset)] \\ &= \mathbb{E}_{\theta^\star \sim Q} \sum_{l=1}^{m} \mathbb{E}_{\theta_l \sim Q_l}[V^{\theta_l}_{\pi_l,1}(\emptyset) - V^{\theta^\star}_{\pi_l,1}(\emptyset)].\end{aligned}$$

Proof. The first line is a simple rewriting of (1) based on Algorithm 1. Observe that given the history up to time $(l-1)L$, the distributions of $\theta^\star$ and θ_l are same as Q_l. Furthermore, the mapping from θ to $V^{\theta}_{\mu(\theta),1}(\emptyset)$ is deterministic. Using the observation by Russo & Van Roy (2014) and the tower rule, we get

$$\mathbb{E}_{\theta^\star \sim Q} \sum_{l=1}^{m} \mathbb{E}_{\theta_l \sim Q_l} V^{\theta^\star}_{\pi^\star,1}(\emptyset) = \mathbb{E}_{\theta^\star \sim Q} \sum_{l=1}^{m} \mathbb{E}_{\theta_l \sim Q_l} V^{\theta_l}_{\pi_l,1}(\emptyset),$$

which leads to the second line of the lemma. □

Next, we define the *Bellman operator*

$$\mathcal{T}^{\theta}_{\pi} V(\mathcal{H}_{t-1}) = \mathbb{E}_{\theta,\pi}[A_t \cdot X_t + V(\mathcal{H}_t)|\mathcal{H}_{t-1}].$$

It is easy to check that $V^{\theta}_{\pi,i} = \mathcal{T}^{\theta}_{\pi} V^{\theta}_{\pi,i+1}$. The next lemma further decomposes the regret.

Lemma 3. *Fix $\theta^\star$ and θ_l, and let $\mathcal{H}_0 = \emptyset$. Then we have*

$$\begin{aligned}&V^{\theta_l}_{\pi_l,1}(\mathcal{H}_0) - V^{\theta^\star}_{\pi_l,1}(\mathcal{H}_0) \\ &\qquad = \mathbb{E}_{\theta^\star,\pi_l} \sum_{t=1}^{L} (\mathcal{T}^{\theta_l}_{\pi_l} - \mathcal{T}^{\theta^\star}_{\pi_l}) V^{\theta_l}_{\pi_l,t+1}(\mathcal{H}_{t-1}).\end{aligned}$$

Proof. Using the relation $V^{\theta}_{\pi,i} = \mathcal{T}^{\theta}_{\pi} V^{\theta}_{\pi,i+1}$, we may write

$$\begin{aligned}&V^{\theta_l}_{\pi_l,1}(\mathcal{H}_0) - V^{\theta^\star}_{\pi_l,1}(\mathcal{H}_0) \\ &= (\mathcal{T}^{\theta_l}_{\pi_l} V^{\theta_l}_{\pi_l,2} - \mathcal{T}^{\theta^\star}_{\pi_l} V^{\theta^\star}_{\pi_l,2})(\mathcal{H}_0) \\ &= (\mathcal{T}^{\theta_l}_{\pi_l} - \mathcal{T}^{\theta^\star}_{\pi_l}) V^{\theta_l}_{\pi_l,2}(\mathcal{H}_0) + \mathcal{T}^{\theta^\star}_{\pi_l} (V^{\theta_l}_{\pi_l,2} - V^{\theta^\star}_{\pi_l,2})(\mathcal{H}_0).\end{aligned}$$

The second term can be written as

$$\mathbb{E}_{\theta^\star,\pi_l}[(V^{\theta_l}_{\pi_l,2} - V^{\theta^\star}_{\pi_l,2})(\mathcal{H}_1)|\mathcal{H}_0],$$

and we can repeat this L times to obtain the equation. □

We also record the following technical result.

Lemma 4. *Let $a_i, b_i \in [0,1]$ for $i \in [k]$ and $|a_i - b_i| \le \Delta_i$. Then we can show*

$$\begin{aligned}\sum_{x \in \{0,1\}^k} |\prod_i a_i^{x_i}(1-a_i)^{1-x_i} - \prod_i b_i^{x_i}(1-b_i)^{1-x_i}| \\ \le 2\sum_{j=1}^{k} \Delta_j.\end{aligned} \tag{3}$$

Proof. Fix x. For simplicity, let $c_i = a_i^{x_i}(1-a_i)^{1-x_i}$ and $d_i = b_i^{x_i}(1-b_i)^{1-x_i}$. Since x_i is either 0 or 1, we have $|c_i - d_i| = |a_i - b_i| \le \Delta_i$. Then we write

$$\begin{aligned}|\prod_{i=1}^{k} c_i - \prod_{i=1}^{k} d_i| &\le (\prod_{i=1}^{k-1} c_i)|c_k - d_k| + |\prod_{i=1}^{k-1} c_i - \prod_{i=1}^{k-1} d_i|d_k \\ &\le (\prod_{i=1}^{k-1} c_i)\Delta_k + |\prod_{i=1}^{k-1} c_i - \prod_{i=1}^{k-1} d_i|d_k \\ &\le \cdots \\ &\le \sum_{j=1}^{k} (\prod_{i=1}^{j-1} c_i)\Delta_j(\prod_{i=j+1}^{k} d_i).\end{aligned}$$

Summing the last term over x finishes the argument. □

Now we are ready to prove our main theorem.

Theorem 5. (Thompson Sampling, Regret Bound) *The Bayesian regret of Algorithm 1 satisfies the following bound*

$$BR(T) = \mathcal{O}(\sqrt{KL^3N^3T\log T}).$$

Remark. *If the system is classical stationary MABs, then it corresponds to the case $L = 1, N = 1$, and our result reproduces the result of $\mathcal{O}(\sqrt{KT\log T})$ (Lattimore & Szepesvári, 2018, Chp. 36). Furthermore, when $N > \frac{K}{2}$, we can think of the problem as choosing the passive arms, and the smaller bound with N replaced by $K - N$ would apply.*

Proof. We fix an episode l and analyze the regret in this episode. Let $t_l = (l-1)L$ so that the episode starts at time $t_l + 1$. Define

$$N_l(k, r, n) = \sum_{t=1}^{t_l} \mathbb{1}\{A_{t,k} = 1, r_k = r, n_k = n\}.$$

It counts the number of rounds where the arm k was chosen by the learner with history $r_k = r$ and $n_k = n$ (see (2) for definition). Note that

$$k \in [K], r \in \{0, 1, \rho(k)\}, \text{ and } n \in [L],$$

where $\rho(k)$ is the initial success rate of arm k. This implies that there are $3KL$ possible tuples of (k, r, n).

Let $\omega^\theta(k, r, n)$ denote the conditional probability of $X_k = 1$ given a history (r, n) and system parameters θ. Also let $\hat{\omega}(k, r, n)$ denote the empirical mean of this quantity (using $N_l(k, r, n)$ past observations and set the estimate to 0 if $N_l(k, r, n) = 0$). Then define

$$\Theta_l = \{\theta \mid \forall k, r, n, \ |(\hat{\omega} - \omega^\theta)(k, r, n)| < \sqrt{\frac{2\log(1/\delta)}{1 \vee N_l(k, r, n)}}\}.$$

Since $\hat{\omega}(k, r, n)$ is $\mathcal{H}_{t_l}$-measurable, so is this set. Using the Hoeffding inequality, one can show

$$\mathbb{P}(\theta^\star \notin \Theta_l) = \mathbb{P}(\theta_l \notin \Theta_l) \leq 3\delta KL.$$

We now turn our attention to the following Bellman operator

$$\mathcal{T}^\theta_{\pi_l} V^{\theta_l}_{\pi_l,t}(\mathcal{H}_{t-1}) = \mathbb{E}_{\theta,\pi_l}[A_{t_l+t} \cdot X_{t_l+t} + V^{\theta_l}_{\pi_l,t}(\mathcal{H}_t) | \mathcal{H}_{t-1}].$$

Since π_l is deterministic, so is A_{t_l+t} given $\mathcal{H}_{t-1}$ and π_l. Let $(k_1, \ldots, k_N)$ be the active arms at time $t_l + t$ and write $\omega^\theta(k_i, r_{k_i}, n_{k_i}) = \omega_{\theta,i}$. Then we can rewrite

$$\begin{aligned}
&\mathcal{T}^\theta_{\pi_l} V^{\theta_l}_{\pi_l,t}(\mathcal{H}_{t-1}) \\
&\quad = \sum_{i=1}^N \omega_{\theta,i} + \sum_{x \in \{0,1\}^N} P^\theta_x V^{\theta_l}_{\pi_l,t}(\mathcal{H}_{t-1} \cup (A_{t_l+t}, x)),
\end{aligned} \tag{4}$$

where $P^\theta_x = \prod_{i=1}^N \omega^{x_i}_{\theta,i}(1 - \omega_{\theta,i})^{1-x_i}$. Under the event that $\theta^\star, \theta_l \in \Theta_l$, we have

$$|\omega_{\theta_l,i} - \omega_{\theta^\star,i}| < 1 \wedge \sqrt{\frac{8\log(1/\delta)}{1 \vee N_l(k_i, r_{k_i}, n_{k_i})}} =: \Delta_i(t_l + t),$$

where the dependence on $t_l + t$ comes from the mapping from i to k_i. Lemma 4 provides

$$\sum_{x \in \{0,1\}^N} |P^{\theta_l}_x - P^{\theta^\star}_x| \leq 2\sum_{i=1}^N \Delta_i(t_l + t). \tag{5}$$

From (4), (5), and the fact that $|V^\theta_{\pi,t}| \leq LN$, we obtain given $\mathcal{H}_{t-1}$ and the event $\theta^\star, \theta_l \in \Theta_l$,

$$\begin{aligned}
|(\mathcal{T}^{\theta^\star}_{\pi_l} - \mathcal{T}^{\theta_l}_{\pi_l}) V^{\theta_l}_{\pi_l,t}(\mathcal{H}_{t-1})| &\leq (2LN + 1)\sum_{i=1}^N \Delta_i(t_l + t) \\
&\leq 3LN \sum_{i=1}^N \Delta_i(t_l + t).
\end{aligned}$$

The above inequality holds whenever $\theta^\star, \theta_l \in \Theta_l$. When $\theta^\star \notin \Theta_l$ or $\theta_l \notin \Theta_l$, which happens with probability less than $6\delta KL$, then we have a trivial bound $|V^{\theta_l}_{\pi_l,1}(\emptyset) - V^{\theta^\star}_{\pi_l,1}(\emptyset)| \leq LN$. Therefore, we can deduce

$$\begin{aligned}
|V^{\theta_l}_{\pi_l,1}(\emptyset) - V^{\theta^\star}_{\pi_l,1}(\emptyset)| &\leq 6\delta KL^2 N \\
&\quad + 3LN\mathbb{1}(E_l)\mathbb{E}_{\theta^\star,\pi_l}\sum_{t=1}^L \sum_{i=1}^N \Delta_i(t_l + t),
\end{aligned}$$

where E_l denotes the event $\theta^\star, \theta_l \in \Theta_l$.

Combining this with Lemma 2 and Lemma 3, we get

$$\begin{aligned}
BR(T) &\leq 6\delta mKL^2 N \\
&+ \mathbb{E}_{\theta^\star \sim Q} 3LN \sum_{l=1}^m \mathbb{1}(E_l)\mathbb{E}_{\theta^\star,\pi_l}\sum_{t=1}^L \sum_{i=1}^N \Delta_i(t_l + t).
\end{aligned} \tag{6}$$

We further analyze the summation to finish the argument. Recall that for this summation, we have $\theta^\star, \theta_l \in \Theta_l$. We shorten $N_l(k_i, r_{k_i}, n_{k_i})$ to N_l for simplicity. We have

$$\begin{aligned}
&\sum_{l=1}^m \sum_{t=1}^L \sum_{i=1}^N \Delta_i(t_l + t) \\
&\quad \leq \sum \mathbb{1}\{N_l \leq L\} + \Delta_i \mathbb{1}\{N_l > L\} \qquad (7)\\
&\quad \leq 6KL^2 + \sum \mathbb{1}\{N_l > L\}\sqrt{\frac{8\log(1/\delta)}{N_l}},
\end{aligned}$$

where the second inequality holds because there are $3KL$ possible tuples of (k, r, n) and a tuple can contribute at most $2L$ to the first summation.

We can bound the second term as follows

$$\begin{aligned}
&\sum_{l=1}^m \sum_{t=1}^L \sum_{i=1}^N \mathbb{1}\{N_l > L\}\sqrt{\frac{1}{N_l}} \\
&\quad = \sum_{l=1}^m \sum_{(k,r,n)} \mathbb{1}\{N_l > L\}(N_{l+1} - N_l)\sqrt{\frac{1}{N_l}} \\
&\quad \leq \sum_{l=1}^m \sum_{(k,r,n)} (N_{l+1} - N_l)\sqrt{\frac{2}{N_{l+1}}} \qquad (8)\\
&\quad \leq \sqrt{8} \sum_{(k,r,n)} \sqrt{N_{m+1}(k, r, n)} \\
&\quad \leq \sqrt{24KLNT}.
\end{aligned}$$

For the first inequality, we use $N_{l+1} \leq N_l + L \leq 2N_l$. The second inequality holds due to the integral trick. Finally, the last inequality holds by the Cauchy-Schwartz inequality along with the fact that $\sum_{(k,r,n)} N_{m+1}(k, r, n) = NT$.

Combining (6), (7), and (8), we get

$$BR(T) = \mathcal{O}(\delta KNLT + KL^3 N + \sqrt{KL^3 N^3 T \log(1/\delta)}).$$

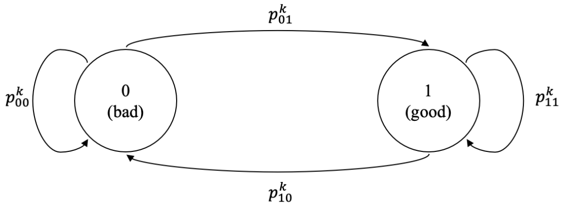

Figure 1. The Gilbert-Elliott channel model

Since NT is a trivial upper bound of $BR(T)$, we may ignore the KL^3N term. Setting $\delta = \frac{1}{T}$ completes the proof. □

5. Experiments

We empirically examine our model in a practical case.

5.1. Model Design

We particularly investigate the Gilbert-Elliott channel model, which is studied by Liu & Zhao (2010) in a restless bandit perspective. This model can be broadly used in communication systems such as cognitive radio networks, downlink scheduling in cellular systems, opportunistic transmission over fading channels, and resource-constrained jamming and anti-jamming.

Each arm k has two parameters p_{01}^k and p_{11}^k, which determine the transition matrix. We assume $P^{\text{active}} = P^{\text{passive}}$ and each arm's transition matrix is independent on the learner's action. There are only two states, *good* and *bad*, and the reward of playing an arm is 1 if its state is good and 0 otherwise. We assume the initial distribution of arm k follows the stationary distribution. In other words, its initial state is good with probability $\omega_k = \frac{p_{01}^k}{p_{01}^k + 1 - p_{11}^k}$.

We fix $K = 8$, $N = 3$, $L = 50$, and $m = 30$. We use Monte Carlo simulation with size 100 to approximate expectations. As we assume $K = 8$ and each arm has two parameters, there are 16 parameters. For these, we use the uniform prior over the support $\{0.1, 0.2, \cdots, 0.9\}$.

5.2. Competitors

As mentioned earlier, one distinguishable strength of our result is that various policy mappings can be used as a competitor. Here we test three different policies: the best fixed arm policy, the myopic policy, and the Whittle index policy. We emphasize again that these competitor policies know the system parameters while our algorithm does not.

The best fixed arm policy computes the stationary distribution $\omega_k = \frac{p_{01}^k}{p_{01}^k + 1 - p_{11}^k}$ for all k and pulls the arms with top N values. The myopic policy keeps updating the belief $\omega_k(t)$ for the arm k being in a good state and pulls the top N arms. Finally, the Whittle index policy computes the Whittle index of each arm and uses it to rank the arms. The Whittle index is proposed by Whittle (1988), and its computation in this particular setting can be found in Liu & Zhao (2010).

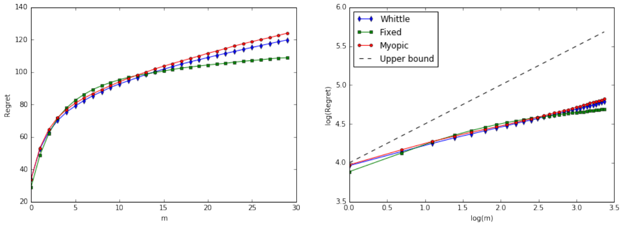

Figure 2. Bayesian regret of Thompson sampling versus episode (left) and its log-log plot (right)

One observation is that these three policies are reduced to the best fixed arm policy in the stationary case. However, the first two policies are known to be sub-optimal in general (Gittins et al., 1989). Liu & Zhao (2010) justify both theoretically and empirically the performance of the Whittle index policy on the Gilbert-Elliott channel model.

5.3. Results

The value functions $V_{\pi,1}^{\theta}(\emptyset)$ for the best fixed arm policy, the myopic policy, and the Whittle index policy are $105.4, 110.3$, and 111.4, respectively. If a competitor policy has a weak performance, then Thompson sampling also uses this weak policy mapping to get a policy π_l for the episode l. This implies that the regret does not necessarily become negative when the competitor policy is weak. Figure 2 shows the trend of regret as a function of episode indices. Regardless of the choice of policy mapping, the regret is sub-linear, and the slope of $\log$-$\log$ plot is less than 0.5, which agrees with Theorem 5.

References

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

Dai, W., Gai, Y., Krishnamachari, B., and Zhao, Q. The non-bayesian restless multi-armed bandit: A case of near-logarithmic regret. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2940–2943. IEEE, 2011.

Dai, W., Gai, Y., and Krishnamachari, B. Online learning for multi-channel opportunistic access over unknown markovian channels. In *IEEE International Conference* *on Sensing, Communication, and Networking (SECON)*, pp. 64–71. IEEE, 2014.

Gittins, J. C., Glazebrook, K. D., Weber, R., and Weber, R. *Multi-armed bandit allocation indices*, volume 25. Wiley Online Library, 1989.

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

Liu, H., Liu, K., and Zhao, Q. Logarithmic weak regret of non-bayesian restless multi-armed bandit. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1968–1971. IEEE, 2011.

Liu, H., Liu, K., and Zhao, Q. Learning in a changing world: Restless multiarmed bandit with unknown dynamics. *IEEE Transactions on Information Theory*, 59(3): 1902–1916, 2013.

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

Meshram, R., Gopalan, A., and Manjunath, D. Optimal recommendation to users that react: Online learning for a class of pomdps. In *IEEE 55th Conference on Decision and Control (CDC)*, pp. 7210–7215. IEEE, 2016.

Meshram, R., Gopalan, A., and Manjunath, D. Restless bandits that hide their hand and recommendation systems. In *IEEE International Conference on Communication Systems and Networks (COMSNETS)*, pp. 206–213. IEEE, 2017.

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

Ortner, R., Ryabko, D., Auer, P., and Munos, R. Regret bounds for restless markov bandits. In *International Conference on Algorithmic Learning Theory*, pp. 214–228. Springer, 2012.

Osband, I., Russo, D., and Van Roy, B. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

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

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Whittle, P. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.