Online Learning in Opportunistic Spectrum Access: A Restless Bandit Approach

Cem Tekin, Mingyan Liu

Department of Electrical Engineering and Computer Science

University of Michigan, Ann Arbor, Michigan, 48109-2122

Email: {cmtkn, mingyan}@umich.edu

Abstract—We consider an opportunistic spectrum access (OSA) problem where the time-varying condition of each channel (e.g., as a result of random fading or certain primary users' activities) is modeled as an arbitrary finite-state Markov chain. At each instance of time, a (secondary) user probes a channel and collects a certain reward as a function of the state of the channel (e.g., good channel condition results in higher data rate for the user). Each channel has potentially different state space and statistics, both unknown to the user, who tries to learn which one is the best as it goes and maximizes its usage of the best channel. The objective is to construct a good online learning algorithm so as to minimize the difference between the user's performance in total rewards and that of using the best channel (on average) had it known which one is the best from a priori knowledge of the channel statistics (also known as the regret). This is a classic exploration and exploitation problem and results abound when the reward processes are assumed to be *iid*. Compared to prior work, the biggest difference is that in our case the reward process is assumed to be Markovian, of which *iid* is a special case. In addition, the reward processes are *restless* in that the channel conditions will continue to evolve independent of the user's actions. This leads to a restless bandit problem, for which there exists little result on either algorithms or performance bounds in this learning context to the best of our knowledge. In this paper we introduce an algorithm that utilizes regenerative cycles of a Markov chain and computes a sample-mean based index policy, and show that under mild conditions on the state transition probabilities of the Markov chains this algorithm achieves logarithmic regret uniformly over time, and that this regret bound is also optimal. We numerically examine the performance of this algorithm along with a few other learning algorithms in the case of an OSA problem with Gilbert-Elliot channel models, and discuss how this algorithm may be further improved (in terms of its constant) and how this result may lead to similar bounds for other algorithms.

I. Introduction

In this paper we study the following opportunistic spectrum access (OSA) problem. A (secondary) user has access to a set of K channels, each of time-varying conditions as a result of random fading and/or certain primary users' activities. Each channel is thus modeled as an arbitrary finite-state discrete-time Markov chain. At each time step, the secondary user (simply referred to as *the user* for the rest of the paper for there is no ambiguity) probes a channel to find out its condition, and is allowed to use the channel in a way consistent with its condition. For instance, good channel conditions result in higher data rates or lower power for the user and so on. This is modeled as a reward collected by the user, the reward being a function of the state of the channel or the Markov chain.

Channels have potentially different state spaces and statistics, both unknown to the user. The user will thus try to learn which one is the best and maximizes its usage of the best channel. Within this context, the player's performance is typically measured by the notion of *regret*. It is defined as the difference between the expected reward that can be gained by an "infeasible" or ideal policy, i.e., a policy that requires either a priori knowledge of some or all statistics of the arms or hindsight information, and the expected reward of the player's policy. The most commonly used infeasible policy is the *best single action* policy, that is optimal among all policies that continue to play the same arm. An ideal policy could play for instance the arm that has the highest expected reward (which requires statistical information but not hindsight). This type of regret is sometimes also referred to as the *weak regret*, see e.g., work by Auer et al. [1]. In this paper we will only focus on this definition of regret.

The above can be cast as a single player multiarmed bandit problem, where the reward of each channel (also referred to as an *arm* in the bandit problem literature) is generated by a Markov chain with unknown statistics. Furthermore, it is a *restless* bandit problem because the state of each Markov chain evolves independent of the action of the user (whether the channel is probed or not); by contrast, in a classic multiarmed bandit problem the state of a Markov chain only evolves when it is acted upon and stays frozen otherwise (also referred to as *rested*). The restless nature of the Markov chains follows naturally from the fact that channel conditions are governed by external factors like random fading, shadowing, and primary user activity.

In the remainder of this paper a channel will also be referred to as an *arm*, the user as the *player*, and probing a channel as *playing or selecting an arm*. This problem is a typical example of the tradeoff between *exploration* and *exploitation*. On the one hand, the player needs to sufficiently explore all arms so as to discover with accuracy the best arm and avoid getting stuck playing an inferior one erroneously believed to be the best. On the other hand, he needs to avoid spending too much time sampling the arms and collecting statistics and not playing the best arm often enough to get a high return.

In most prior work on the class of multiarmed bandit problems, originally proposed by Robbins [2], the rewards are assumed to be independently drawn from a fixed (but unknown) distribution. Its worth noting that with this iid

assumption on the reward process, whether an arm is rested or restless is inconsequential for the following reasons. Since the rewards are independently drawn each time, whether an unselected arm remains still or continues to change does not affect the reward the arm produces the next time it is played whenever that may be. This is clearly not the case with Markovian rewards. In the rested case, since the state is frozen when an arm is not played, the state in which we next observe the arm is *independent* of how much time elapses before we play the arm again. In the restless case, the state of an arm continues to evolve, thus the state in which we next observe it is now *dependent* on the amount of time that elapses between two plays of the same arm. This makes the problem significantly more difficult.

To the best of our knowledge, there has been no study of the restless bandits in this learning context, either in terms of algorithms or performance bounds. Here lies the main contribution of the present study. In this paper we give the first result on the existence of order-optimal policies for the above restless bandit problem. Specifically, we introduce an algorithm that utilizes regenerative cycles of a Markov chain and computes a sample-mean based index policy, and show that under mild conditions on the state transition probabilities this algorithm achieves logarithmic regret uniformly over time.

Below we briefly summarize the most relevant results in the literature. Lai and Robbins in [3] model rewards as single-parameter univariate densities and give a lower bound on the regret and construct policies that achieve this lower bound which are called *asymptotically efficient* policies. This result is extended by Anantharam et al. in [4] to the case of playing more than one arm at a time. Using a similar approach Anantharam et al. in [5] develops index policies that are asymptotically efficient for arms with rewards driven by finite, irreducible, aperiodic and rested Markov chains with identical state spaces and single-parameter families of stochastic transition matrices. Agrawal in [6] considers sample mean based index policies for the iid model that achieve $O(\log n)$ regret, where n is the total number of plays. Auer et al. in [7] also proposes sample mean based index policies for iid rewards with bounded support; these are derived from [6], but are simpler than the those in [6] and are not restricted to a specific family of distributions. These policies achieve logarithmic regret uniformly over time rather than asymptotically in time, but have bigger constant than that in [3]. In [8] it is shown that the index policy in [7] is order optimal for Markovian rewards drawn from *rested* arms but not restricted to single-parameter families, under some assumptions on the transition probabilities.

Other works such as [9], [10], [11] consider the iid reward case in a multiuser setting; players selecting the same arms experience collision according to a certain collision model. We would like to mention another class of multiarmed bandit problems in which the statistics about the problem are known a priori and the state is observed perfectly; these are thus optimization problems rather than learning problems. The rested case is considered by Gittins [12] and the optimal policy is proved to be an index policy which at each time plays the arm with highest Gittins' index, while Whittle introduced the restless bandit problem in [13]. The restless bandit problem does not have a known general solution though special cases may be solved. For instance, a myopic policy is shown to be optimal when channels are identical and bursty in [14] for an OSA problem formulated as a restless bandit problem with each channel modeled as a two-state Markov chain (the Gilbert-Elliot model).

The remainder of this paper is organized as follows. In Section II we formulate the single player restless bandit problem. In Section III we introduce an algorithm based on regenerative cycles that employs sample-mean based indices. The regret of this algorithm is analyzed and shown to be optimal in Section IV. In Section V we numerically examine its performance along with a few other learning algorithms in the case of an OSA problem with Gilbert-Elliot channel models, and discuss how this algorithm may be further improved (in terms of its constant) and how this result may lead to similar bounds for other algorithms. Finally, Section VI concludes the paper.

II. PROBLEM FORMULATION AND PRELIMINARIES

Consider K arms (or channels) indexed by $i = 1, 2, \cdots, K$. The ith arm is modeled as a discrete-time, irreducible and aperiodic Markov chain with finite state space S^i. There is a stationary and positive reward associated with each state of each arm. Let r^i_x denote the reward obtained from state x of arm i, $x \in S^i$; this reward is in general different for different states. Let $P^i = \left\{ p^i_{xy}, x, y \in S^i \right\}$ denote the transition probability matrix and $\boldsymbol{\pi}^i = \{ \pi^i_x, x \in S^i \}$ the stationary distribution of arm i.

Let $(P^i)'$ denote the *adjoint* of P^i on $l_2(\pi)$, and let $\hat{P}^i = (P^i)' P$ denote the *multiplicative symmetrization* of P^i, where

$$(p^i)'_{xy} = (\pi^i_y p^i_{yx})/\pi^i_x, \; \forall x, y \in S^i.$$

We will assume that the P^is are such that $\hat{P}^i$s are irreducible. To give a sense of how strong this assumption is, we note that one condition that guarantees that the $\hat{P}^i$s are irreducible is $p_{xx} > 0, \forall x \in S^i, \forall i$. For the application under consideration, this condition means that there is always positive probability for a channel to remain in the same state over one unit of time, which appears to be a natural and benign assumption[1].

We assume the arms (or Markov chains) are mutually independent and are restless, i.e., their states will continue to evolve regardless of the user's actions. The mean reward of arm i, denoted by μ^i, is the expected reward of arm i under its stationary distribution:

$$\mu^i = \sum_{x \in S^i} r^i_x \pi^i_x . \tag{1}$$

---

[1]Alternatively we could adopt a stronger assumption that the Markov chains are aperiodic and reversible (note that aperiodicity and reversibility implies that the multiplicative symmetrization of P^i is irreducible), in which case the same order results can be obtained with a different constant if we use a different large deviation bound from [15] instead of Lemma 1.

For convenience, we will use * in the superscript to denote the arm with the highest mean. For instance, $\mu^* = \max_{1 \leq i \leq K} \mu^i$, and so on. We assume that the arm with the highest mean is unique.

Consistent with the discrete-time Markov chain model, we will assume that the user's actions occur also in discrete time. For a policy α we define its regret $R^\alpha(n)$ as the difference between the expected total reward that can be obtained by playing the arm with the highest mean and the expected total reward obtained from using policy α up to time n. Always playing the arm with the highest mean reward is referred to as the *best single-action policy*, and this arm will also be referred to as the *optimal* arm; accordingly the others will be referred to as *suboptimal* arms.

Let $\alpha(t)$ denote the arm selected by policy α at t, $t = 1, 2, \cdots$, and $x_\alpha(t)$ the state of arm $\alpha(t)$ at time t. Then we have

$$R^\alpha(n) = n\mu^* - E^\alpha \left[\sum_{t=1}^n r_{x_\alpha(t)}^{\alpha(t)} \right] . \tag{2}$$

The objective is to examine how the regret $R^\alpha(n)$ behaves as a function of n for a given policy α and to construct a policy whose regret is order-optimal, through appropriate bounding. As we will show and as is commonly done, the key to bounding $R^\alpha(n)$ is to bound the expected number of plays of any suboptimal arm.

Our analysis utilizes the following known results on Markov chains; the proofs are not reproduced here for brevity. The first is a result by Lezaud [16] that bounds the probability of a large deviation from the stationary distribution.

Lemma 1: [Theorem 3.3 from [16]] Consider a finite-state, irreducible Markov chain $\{X_t\}_{t \geq 1}$ with state space S, matrix of transition probabilities P, an initial distribution $\mathbf{q}$ and stationary distribution π. Let $N_\mathbf{q} = \left\| \left(\frac{q_x}{\pi_x}, x \in S \right) \right\|_2$. Let $\hat{P} = P'P$ be the multiplicative symmetrization of P where P' is the adjoint of P on $l_2(\pi)$. Let $\epsilon = 1 - \lambda_2$, where λ_2 is the second largest eigenvalue of the matrix $\hat{P}$. ϵ will be referred to as the eigenvalue gap of $\hat{P}$. Let $f : S \to \mathcal{R}$ be such that $\sum_{y \in S} \pi_y f(y) = 0$, $\|f\|_\infty \leq 1$ and $0 < \|f\|_2^2 \leq 1$. If $\hat{P}$ is irreducible, then for any positive integer n and all $0 < \gamma \leq 1$

$$P\left(\frac{\sum_{t=1}^n f(X_t)}{n} \geq \gamma \right) \leq N_q \exp\left[-\frac{n\gamma^2 \epsilon}{28} \right] .$$

The second is a result by Bremaud, which can be found in [17].

Lemma 2: If $\{X_n\}_{n \geq 0}$ is a positive recurrent homogeneous Markov chain with state space S, stationary distribution π and τ is a stopping time that is finite almost surely for which $X_\tau = x$ then for all $y \in S$

$$E\left[\sum_{t=0}^{\tau-1} I(X_t = y) | X_0 = x \right] = E[\tau | X_0 = x]\pi_y .$$

In the next two sections we first present a policy, referred to as the *regenerative cycle algorithm*, and then analyze its regret.

III. REGENERATIVE CYCLE ALGORITHM (RCA)

In this section we present an algorithm called the *regenerative cycle algorithm* (RCA), and prove in the next section that this algorithm guarantees a logarithmic growth of the regret uniformly over time under mild assumptions on the state transition probabilities and the rewards.

As the name suggests, this algorithm operates on regenerative cycles. In essence what the algorithm does is to construct sample paths of each arm solely using those observed within regenerative cycles while discarding the rest in its estimation of the quality of an arm (in the form of an *index*). The reason behind such a construction has to do with the restless nature of the arms. As noted in the introduction, since each arm continues to evolve according to the Markov chain regardless of the user's action, the probability distribution of the reward we get by playing an arm is a function of the amount of time that has elapsed since the last time we played the same arm. Since we play one arm at a time, the arms become coupled (in terms of the probability distributions of the rewards). While this certainly does not affect our ability to collect rewards, it makes it extremely hard to analyze the estimated quality (or the index) of an arm calculated based on rewards collected this way.

However, if instead of the actual sample path of all observations from an arm, we limit ourselves to a sample path constructed (or rather stitched together) using only the observations from regenerative cycles, then this sample path essentially has the same statistics as the original Markov chain due to the renewal property and one can now use the sample mean of the rewards from the regenerative sample paths to approximate the mean reward under stationary distribution.

Figure 1 illustrates one possible realization of this algorithm. As shown, RCA operates in blocks. Within a block, the algorithm plays the same arm in each time slot (arm i in the first block in this example) till a certain pre-specified state (say γ^i) is observed. Upon this observation we enter a regenerative cycle and continue to play till the same state γ^i is observed a second time. This marks the end of the block labeled "play arm i". At the end of each block, the algorithm computes an index for all arms and selects the one with the highest index to play in the next block (arm k shown in the figure). It follows that the block length is a random variable.

For the purpose of index computation and subsequent analysis, each block is further broken into three sub-blocks (SBs). SB1 consists of all time slots from the beginning of the block to right before the first visit to γ^i; SB2 includes all time slots from the first visit to γ^i up to but excluding the second visit to state γ^i; SB3 consists of a single time slot with the second visit to γ^i. These are also shown in Figure 1. The key to the algorithm is for each arm to single out only observations within SB2's in each block and virtually assemble them (these are highlighted with thick lines). Because of the
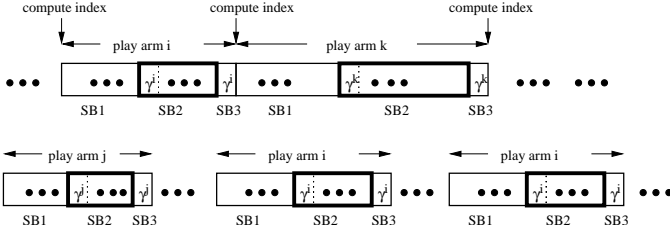
Fig. 1. Example realization of RCA

regenerative nature of the Markov chain, once put together, the resulting sample path has exactly the same statistics as given by the transition probability matrix P^i; this results in a tractable problem.

Throughout our discussion, we will consider a horizon of n time slots. A list of notations used is summarized as follows; some are also marked on Figure 2 for convenience:
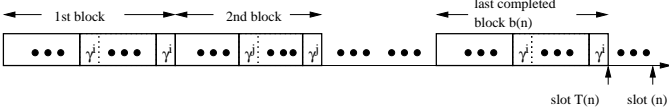

Fig. 2. Running RCA over a period of n slots

- γ^i: state that determine the regenerative cycles for arm i.
- $\alpha(b)$: the arm played in block b.
- $b(n)$: total number of completed blocks up to time n.
- $T(n)$: time at the end of the last completed block.
- $T^i(n)$: total number of times (slots) arm i is played up to time $T(n)$.
- $B^i(b)$: total number of blocks within the first b blocks in which arm i is played.
- $X_1^i(j)$: vector of observed states from SB1 of the jth block in which arm i is played; it is empty if the first observed state is γ^i.
- $X_2^i(j)$: vector of observed states from SB2 of the jth block in which arm i is played;
- $X^i(j)$: vector of observed states from the jth block in which arm i is played. Thus we have $X^i(j) = [X_1^i(j), X_2^i(j), \gamma^i]$.
- $t(b)$: time at the end of block b; $t(b) = \sum_{i=1}^{K} \sum_{j=1}^{B^i(b)} |X^i(j)|$.
- $T^i(t(b))$: total number of time slots arm i is played up to time $t(b)$. Thus $T^i(t(b)) = \sum_{j=1}^{B^i(b)} |X^i(j)|$. Also note that $T^i(t(b(n))) = T^i(n)$.
- $t_2(b)$: total number of time slots spent in SB2 up to block b. Thus $t_2(b) = \sum_{i=1}^{K} \sum_{j=1}^{B^i(b)} |X_2^i(j)|$.
- $r^i(k)$: the reward from arm i when it's played for the k-th time, counting only those times played during an SB2.
- $T_2^i(t_2(b))$: total number of time slots arm i is played during SB2 up to block b. Thus $T_2^i(t_2(b)) = \sum_{j=1}^{B^i(b)} |X_2^i(j)|$.

RCA computes and updates the value of an *index* g^i for each arm i at the end of block b, based on the total reward obtained from arm i during all SB2 as follows:

Regenerative Cycle Algorithm (RCA):
1: Initialize: $b = 1, t = 0, t_2 = 0, T_2^i = 0, r^i = 0, \forall i = 1, \cdots, K$
2: **for** $b \leq K$ **do**
3:　play arm b; set γ^b to be the first state observed
4:　$t := t+1; t_2 := t_2+1; T_2^b := T_2^b+1; r^b := r^b + r_{\gamma^i}^b$
5:　play arm b; denote observed state as x
6:　**while** $x \neq \gamma^b$ **do**
7:　　$t := t+1; t_2 := t_2+1; T_2^b := T_2^b+1; r^b := r^b + r_x^b$
8:　　play arm b; denote observed state as x
9:　**end while**
10:　$b := b+1; t := t+1$
11: **end for**
12: **for** $j = 1$ to K **do**
13:　compute index $g^j := \frac{r^j}{T_2^j} + \sqrt{\frac{L \ln t_2}{T_2^j}}$
14:　$j++$
15: **end for**
16: $i := \arg\max_j g^j$
17: **while** (1) **do**
18:　play arm i; denote observed state as x
19:　**while** $x \neq \gamma^i$ **do**
20:　　$t := t+1$
21:　　play arm i; denote observed state as x
22:　**end while**
23:　$t := t+1; t_2 := t_2+1; T_2^i := T_2^i+1; r^i := r^i + r_x^i$
24:　play arm i; denote observed state as x
25:　**while** $x \neq \gamma^i$ **do**
26:　　$t := t+1; t_2 := t_2+1; T_2^i := T_2^i+1; r^i := r^i + r_x^i$
27:　　play arm i; denote observed state as x
28:　**end while**
29:　$b := b+1; t := t+1$
30:　**for** $j = 1$ to K **do**
31:　　compute index $g^j := \frac{r^j}{T_2^j} + \sqrt{\frac{L \ln t_2}{T_2^j}}$
32:　　$j++$
33:　**end for**
34:　$i := \arg\max_j g^j$
35: **end while**

Fig. 3. Pseudocode of RCA

$$g_{t_2(b),T_2^i(t_2(b))}^i = \bar{r}^i(T_2^i(t_2(b))) + \sqrt{\frac{L \ln t_2(b)}{T_2^i(t_2(b))}}, \qquad (3)$$

where L is a constant, and

$$\bar{r}^i(T_2^i(t_2(b)) = \frac{r^i(1) + r^i(2) + ... + r^i(T_2^i(t_2(b)))}{T_2^i(t_2(b))}$$

denotes the sample mean of the reward collected during an SB2: $X_2^i(1), X_2^i(2), \cdots, X_2^i(B^i(b))$ (this is arm i's total reward over the total number of times it's played). The second term in the index computation serves the purpose of *exploration*: the relative uncertainty of the mean reward of an arm grows as the arm is not played. This index definition is similar to that proposed in [7], but computed only over SB2s. RCA is formally given in Figure 3. In this description the

algorithm continues indefinitely, but can obviously be stopped at anytime that some desired horizon is reached.

Its worth noting that RCA also collects reward during SB1 and SB3. However, the computation of the indices only relies on SB2. The reason becomes clearer in the next section where we analyze its regret and show that it grows at most logarithmically in n.

IV. REGRET ANALYSIS OF RCA

We begin by bounding the expected number of plays of a suboptimal arm.

Theorem 1: Assume all arms are finite-state, irreducible, aperiodic Markov chains whose transition probability matrices have irreducible multiplicative symmetrizations and assume all rewards are positive. Let $\pi_{\min}^i = \min_{x \in S^i} \pi_x^i$, $\pi_{\min} = \min_{1 \le i \le K} \pi_{\min}^i$, $r_{\max} = \max_{x \in S^i, 1 \le i \le K} r_x^i$, $S_{\max} = \max_{1 \le i \le K} |S^i|$, $\hat{\pi}_{\max} = \max_{x \in S^i, 1 \le i \le K} \{\pi_x^i, 1 - \pi_x^i\}$, $\epsilon_{\min} = \min_{1 \le i \le K} \epsilon^i$, $M_{\max}^i = \max_{x, y \in S^i, x \ne y} M_{x,y}^i$, where ϵ^i is the eigenvalue gap of the multiplicative symmetrization of the transition probability matrix of the ith arm and $M_{x,y}^i$ is the mean hitting time of state y starting from an initial state x for the ith arm. Then for a player using RCA with a constant $L \ge 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$ in (3), we have

$$\sum_{i: \mu^i < \mu^*} (\mu^* - \mu^i) E[T^i(n)]$$
$$\le 4L \sum_{i: \mu^i < \mu^*} \frac{D_i \ln n}{(\mu^* - \mu^i)} + \sum_{i: \mu^i < \mu^*} (\mu^* - \mu^i) D_i C_i ,$$

where

$$C_i = \left(1 + \frac{(|S^i| + |S^*|)\beta}{\pi_{\min}}\right), \quad \beta = \sum_{t=1}^{\infty} t^{-2}$$
$$D_i = \left(\frac{1}{\pi_{\min}^i} + M_{\max}^i + 1\right).$$

Proof: Throughout the proof all quantities pertain to RCA, which will be denoted by α and suppressed from the superscript whenever there is no ambiguity. Let $c_{t,s} = \sqrt{L \ln t / s}$, and let l be any positive integer. Then,

$$B^i(b) = 1 + \sum_{m=K+1}^{b} I(\alpha(m) = i)$$
$$\le l + \sum_{m=K+1}^{b} I(\alpha(m) = i, B^i(m-1) \ge l)$$
$$\le l + \sum_{m=K+1}^{b} I(g_{t_2(m-1), T_2^*(t_2(m-1))}^*$$
$$\le g_{t_2(m-1), T_2^i(t_2(m-1))}^i, B^i(m-1) \ge l)$$
$$\le l + \sum_{m=K+1}^{b} I \left(\min_{1 \le s \le t_2(m-1)} g_{t_2(m-1), s}^*\right.$$
$$\left.\le \max_{t_2(l) \le s_i \le t_2(m-1)} g_{t_2(m-1), s_i}^i\right)$$

$$\le l + \sum_{m=K+1}^{b} \sum_{s=1}^{t_2(m-1)} \sum_{s_i = t_2(l)}^{t_2(m-1)} I(g_{t_2(m), s}^* \le g_{t_2(m), s_i}^i) \quad (4)$$
$$\le l + \sum_{t=1}^{t_2(b)} \sum_{s=1}^{t-1} \sum_{s_i=l}^{t-1} I(g_{t,s}^* \le g_{t,s_i}^i) \quad (5)$$

where as given in (3), $g_{t,s}^i = \bar{r}^i(s) + c_{t,s}$. The inequality in (5) follows from the fact that the outer sum in (5) is over time while the outer sum in (4) is over blocks and each block lasts at least two time slots.

We now show that $g_{t,s}^* \le g_{t,s_i}^i$ implies that at least one of the following holds:

$$\bar{r}^*(s) \le \mu^* - c_{t,s} \quad (6)$$
$$\bar{r}^i(s_i) \ge \mu^i + c_{t,s_i} \quad (7)$$
$$\mu^* < \mu^i + 2c_{t,s_i}. \quad (8)$$

This is because if none of the above holds, then we must have

$$g_{t,s}^* = \bar{r}^*(s) + c_{t,s} > \mu^* \ge \mu^i + 2c_{t,si} > \bar{r}^i(s_i) + c_{t,s_i} = g_{t,s_i}^i,$$

which contradicts $g_{t,s}^* \le g_{t,s_i}^i$.

If we choose $s_i \ge 4L \ln(t_2(b))/(\mu^* - \mu^i)^2$, then $2c_{t,s_i} \le \mu^* - \mu^i$ for $t \le t_2(b)$, which means (8) is false, and therefore at least one of (6) and (7) is true with this choice of s_i. We next take $l = \left\lceil \frac{4L \ln t_2(b)}{(\mu^* - \mu^i)^2} \right\rceil$, and proceed from (5). Taking expectation on both sides and relaxing the outer sum in (5) from $t_2(b)$ to ∞,

$$E[B^i(b)] \le \left\lceil \frac{4L \ln t_2(b)}{(\mu^* - \mu^i)^2} \right\rceil$$
$$+ \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i = \left\lceil \frac{4L \ln t_2(b)}{(\mu^* - \mu^i)^2} \right\rceil}^{t-1} P(\bar{r}^*(s) \le \mu^* - c_{t,s})$$
$$+ \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i = \left\lceil \frac{4L \ln t_2(b)}{(\mu^* - \mu^i)^2} \right\rceil}^{t-1} P(\bar{r}^i(s_i) \ge \mu^i + c_{t,s_i}).$$

Consider an initial distribution $\mathbf{q}^i$ for the ith arm. We have:

$$N_{\mathbf{q}^i} = \left\| \left(\frac{q_y^i}{\pi_y^i}, y \in S^i\right) \right\|_2 \le \sum_{y \in S^i} \left\| \frac{q_y^i}{\pi_y^i} \right\|_2 \le \frac{1}{\pi_{\min}},$$

where the first inequality follows from the Minkowski inequality. Let $n_y^i(t)$ denote the number of times state y of arm i is observed during all SB2s up to the tth play. Then,

$$P(\bar{r}^i(s_i) \ge \mu^i + c_{t,s_i})$$
$$= P\left(\sum_{y \in S^i} r_y^i n_y^i(s_i) \ge s_i \sum_{y \in S^i} r_y^i \pi_y^i + s_i c_{t,s_i}\right)$$
$$= P\left(\sum_{y \in S^i} (r_y^i n_y^i(s_i) - r_y^i s_i \pi_y^i) \ge s_i c_{t,s_i}\right)$$

$$= P(\sum_{y \in S^i} (-r_y^i n_y^i(s_i) + r_y^i s_i \pi_y^i) \leq -s_i c_{t,s_i}) . \quad (9)$$

Consider a sample path ω and the events

$$A = \left\{ \omega : \sum_{y \in S^i} (-r_y^i n_y^i(s_i)(\omega) + r_y^i s_i \pi_y^i) \leq -s_i c_{t,s_i} \right\},$$

$$B = \bigcup_{y \in S^i} \left\{ \omega : -r_y^i n_y^i(s_i)(\omega) + r_y^i s_i \pi_y^i \leq -\frac{s_i c_{t,s_i}}{|S^i|} \right\}.$$

If $\omega \notin B$ then,

$$-r_y^i n_y^i(s_i)(\omega) + r_y^i s_i \pi_y^i > -\frac{s_i c_{t,s_i}}{|S^i|}, \ \forall y \in S^i$$

$$\Rightarrow \quad \sum_{y \in S^i} (-r_y^i n_y^i(s_i)(\omega) + r_y^i s_i \pi_y^i) > -s_i c_{t,s_i}$$

Thus $\omega \notin A$, and $P(A) \leq P(B)$. Then continuing from (9):

$$P(\bar{r}^i(s_i) \geq \mu^i + c_{t,s_i})$$

$$\leq \sum_{y \in S^i} P\left(-r_y^i n_y^i(s_i) + r_y^i s_i \pi_y^i \leq -\frac{s_i c_{t,s_i}}{|S^i|} \right)$$

$$= \sum_{y \in S^i} P\left(r_y^i n_y^i(s_i) - r_y^i s_i \pi_y^i \geq \frac{s_i c_{t,s_i}}{|S^i|} \right)$$

$$\leq \sum_{y \in S^i} N_{q^i} t^{-\frac{L\epsilon^i}{28(|S^i|r_y^i \hat{\pi}_y^i)^2}} \quad (10)$$

$$\leq \frac{|S^i|}{\pi_{\min}} t^{-\frac{L\epsilon_{\min}}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}}, \quad (11)$$

where (10) follows from letting

$$\gamma = \frac{c_{t,s_i}}{|S^i|r_y^i \hat{\pi}_y^i}, \quad \hat{\pi}_y^i = \max\left\{\pi_y^i, 1 - \pi_y^i\right\}$$

$$f(X_t^i) = \frac{I(X_t^i = y) - \pi_y^i}{\hat{\pi}_y^i},$$

and using Lemma 1 (note $\hat{P}^i$ is irreducible), which gives

$$P\left(n_y^i(s_i) - s_i \pi_y^i \geq \frac{s_i c_{t,s_i}}{|S^i|r_y^i} \right)$$

$$= P\left(\frac{\sum_{t=1}^{s_i} I(X_t^i = y) - s_i \pi_y^i}{\hat{\pi}_y^i s_i} \geq \frac{c_{t,s_i}}{|S^i|r_y^i \hat{\pi}_y^i} \right)$$

$$\leq N_{q^i} t^{-\frac{L\epsilon^i}{28(|S^i|r_y^i \hat{\pi}_y^i)^2}} \quad (12)$$

We note that for $\gamma > 1$ the deviation probability is zero so the bound still holds.

Similarly, we have

$$P(\bar{r}^*(s) \leq \mu^* - c_{t,s})$$

$$= P(\sum_{y \in S^*} r_y^*(n_y^*(s) - s\pi_y^*) \leq -sc_{t,s})$$

$$\leq \sum_{y \in S^*} P(r_y^* n_y^*(s) - r_y^* s\pi_y^* \leq -\frac{sc_{t,s}}{|S^*|})$$

$$= \sum_{y \in S^*} P(r_y^*(s - \sum_{x \neq y} n_x^*(s)) - r_y^* s(1 - \sum_{x \neq y} \pi_x^*) \leq -\frac{sc_{t,s}}{|S^*|})$$

$$= \sum_{y \in S^*} P(r_y^* \sum_{x \neq y} n_x^*(s) - r_y^* s \sum_{x \neq y} \pi_x^* \geq \frac{sc_{t,s}}{|S^*|})$$

$$\leq \sum_{y \in S^*} N_{q^*} t^{-\frac{L\epsilon^*}{28(|S^*|r_y^* \hat{\pi}_y^*)^2}} \quad (13)$$

$$\leq \frac{|S^*|}{\pi_{\min}} t^{-\frac{L\epsilon_{\min}}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}} \quad (14)$$

where (13) again follows from Lemma 1. Since

$$\frac{|S^i| + |S^*|}{\pi_{\min}} \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=1}^{t-1} t^{-\frac{L\epsilon_{\min}}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}}$$

$$= \frac{|S^i| + |S^*|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-\frac{L\epsilon_{\min} - 56 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}}$$

$$\leq \frac{|S^i| + |S^*|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}, \quad (15)$$

from (11) and (14), given $b(n) = b$ we have

$$E[B^i(b(n))|b(n) = b] \leq \left\lceil \frac{4L \ln t_2(b)}{(\mu^* - \mu^i)^2} \right\rceil + \frac{(|S^i| + |S^*|)\beta}{\pi_{\min}},$$

for all suboptimal arms. The inequality in (15) follows from the assumption $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$. Therefore,

$$E[B^i(b(n))] \leq \frac{4L \ln n}{(\mu^* - \mu^i)^2} + 1 + \frac{(|S^i| + |S^*|)\beta}{\pi_{\min}}, \quad (16)$$

since $n \geq t_2(b(n))$ almost surely.

Note that all the quantities in computing the indices and the probabilities in above come from the intervals $X_2^i(1), X_2^i(2), \cdots \forall i \in \{1, \cdots, K\}$. Since these intervals begin with state γ^i and end with a return to γ^i, by the strong Markov property the process at these stopping times has the same distribution as the original process. Moreover by connecting these intervals together we form a continuous sample path which can be viewed as a sample path generated by a Markov chain with a transition matrix identical to the original arm. This is the reason why we can apply Lezaud's bound to this Markov chain.

The total number of plays of arm i at the end of block $b(n)$ is equal to the total number of plays of arm i during the regenerative cycles of visiting state γ^i plus the total number of plays before entering the regenerative cycles plus one more play resulting from the last play of the block which is state γ^i. This gives:

$$E[T^i(n)] \leq \left(\frac{1}{\pi_{\min}^i} + M_{\max}^i + 1 \right) E[B^i(b(n))] .$$

Thus,

$$\sum_{i:\mu^i < \mu^*} (\mu^* - \mu^i)E[T^i(n)]$$

$$\leq 4L \sum_{i:\mu^i < \mu^*} \frac{D_i \ln n}{(\mu^* - \mu^i)} + \sum_{i:\mu^i < \mu^*} (\mu^* - \mu^i)C_i D_i. (17)$$

$\blacksquare$

We now state the main theorem of this paper.

Theorem 2: Assume all arms are finite-state, irreducible, aperiodic Markov chains whose transition probability matrices have irreducible multiplicative symmetrizations and assume all rewards are positive. Let $\pi^i_{\min} = \min_{x \in S^i} \pi^i_x$, $\pi_{\min} = \min_{1 \leq i \leq K} \pi^i_{\min}$, $r_{\max} = \max_{x \in S^i, 1 \leq i \leq K} r^i_x$, $S_{\max} = \max_{1 \leq i \leq K} |S^i|$, $\hat{\pi}_{\max} = \max_{x \in S^i, 1 \leq i \leq K} \left\{ \pi^i_x, 1 - \pi^i_x \right\}$, $\epsilon_{\min} = \min_{1 \leq i \leq K} \epsilon^i$, $M^i_{\max} = \max_{x,y \in S^i, x \neq y} M^i_{x,y}$, where ϵ^i is the eigenvalue gap of the multiplicative symmetrization of the transition probability matrix of the ith arm and $M^i_{x,y}$ is the mean hitting time of state y starting from an initial state x for the ith arm. Then using a constant $L \geq 112 S^2_{\max} r^2_{\max} \hat{\pi}^2_{\max} / \epsilon_{\min}$, the regret of RCA can be upper bounded uniformly over time by the following, $\forall n$:

$$
\begin{aligned}
R^{RCA}(n) \;<\;& 4L \ln n \sum_{i:\mu^i < \mu^*} \frac{1}{\mu^* - \mu^i} \left(D_i + \frac{E_i}{\mu^* - \mu^i} \right) \\
& + \sum_{i:\mu^i < \mu^*} C_i \left((\mu^* - \mu^i) D_i + E_i \right) + F
\end{aligned}
$$

where

$$
\begin{aligned}
C_i &= \left(1 + \frac{(|S^i| + |S^*|)\beta}{\pi_{\min}} \right), \quad \beta = \sum_{t=1}^{\infty} t^{-2} \\
D_i &= \left(\frac{1}{\pi^i_{\min}} + M^i_{\max} + 1 \right), \\
E_i &= \mu^i (1 + M^i_{\max}) + \mu^* M^*_{\max}, \\
F &= \mu^* \left(\frac{1}{\pi_{\min}} + \max_{i \in \{1,\dots,K\}} M^i_{\max} + 1 \right).
\end{aligned}
$$

Proof: Assume that the states which determine the regenerative sample paths are given *a priori* by $\gamma = [\gamma^1, \cdots, \gamma^K]$. We denote the expectations with respect to RCA given γ as E_γ. First we rewrite the regret in the following form:

$$
\begin{aligned}
R_\gamma(n) =\;& \mu^* E_\gamma[T(n)] - E_\gamma\Big[\sum_{t=1}^{T(n)} r^{\alpha(t)}_{x_{\alpha(t)}}\Big] \\
& + \mu^* E_\gamma[n - T(n)] - E_\gamma\Big[\sum_{t=T(n)+1}^{n} r^{\alpha(t)}_{x_{\alpha(t)}}\Big] \\
=\;& \left\{ \mu^* E_\gamma[T(n)] - \sum_{i=1}^{K} \mu^i E_\gamma\left[T^i(n)\right] \right\} - Z_\gamma(n) \\
& + \mu^* E_\gamma[n - T(n)] - E_\gamma\Big[\sum_{t=T(n)+1}^{n} r^{\alpha(t)}_{x_{\alpha(t)}}\Big]. \quad (18)
\end{aligned}
$$

where for notational convenience, we have used

$$
Z_\gamma(n) = E_\gamma\left[\sum_{t=1}^{T(n)} r^{\alpha(t)}_{x_{\alpha(t)}} \right] - \sum_{i=1}^{K} \mu^i E_\gamma\left[T^i(n)\right].
$$

We can bound the first difference in (18) logarithmically using Theorem 1, so it remains to bound $Z_\gamma(n)$ and the last

difference. We have

$$
\begin{aligned}
Z_\gamma(n) \geq\;& \sum_{y \in S^*} r^*_y E_\gamma\left[\sum_{j=1}^{B^*(b(n))} \sum_{X^*_t \in X^*(j)} I(X^*_t = y) \right] \\
& + \sum_{i:\mu^i < \mu^*} \sum_{y \in S^i} r^i_y E_\gamma\left[\sum_{j=1}^{B^i(b(n))} \sum_{X^i_t \in X^i_2(j)} I(X^i_t = y) \right] \quad (19) \\
& - \mu^* E_\gamma\left[T^*(n)\right] \\
& - \sum_{i:\mu^i < \mu^*} \mu^i \left(\frac{1}{\pi^i_{\gamma^i}} + M^i_{\max} + 1 \right) E_\gamma\left[B^i(b(n))\right],
\end{aligned}
$$

where the inequality comes from counting only the rewards obtained during the SB2s for all suboptimal arms. Applying Lemma 2 to (19) we get

$$
E_\gamma\left[\sum_{j=1}^{B^i(b(n))} \sum_{X^i_t \in X^i_2(j)} I(X^i_t = y) \right] = \frac{\pi^i_y}{\pi^i_{\gamma^i}} E_\gamma\left[B^i(b(n))\right].
$$

Rearrange terms and noting $\mu^* = \sum_y r^*_y \pi^*_y$,

$$
Z_\gamma(n) \geq R^*(n) - \sum_{i:\mu^i < \mu^*} \mu^i (M^i_{\max} + 1) E_\gamma\left[B^i(b(n))\right] \quad (20)
$$

where

$$
\begin{aligned}
R^*(n) =\;& \sum_{y \in S^*} r^*_y E_\gamma\left[\sum_{j=1}^{B^*(b(n))} \sum_{X^*_t \in X^*(j)} I(X^*_t = y) \right] \\
& - \sum_{y \in S^*} r^*_y \pi^*_y E_\gamma\left[T^*(n)\right].
\end{aligned}
$$

Consider now $R^*(n)$. Since all suboptimal arms are played at most logarithmically, the number of time steps in which the best arm is not played is at most logarithmic. It follows that the number of discontinuities between plays of the best arm is at most logarithmic. Suppose we combine successive blocks in which the best arm is played, and denote by $\bar{X}^*(j)$ the j-th combined block. Let $\bar{b}^*$ denote the total number of combined blocks up to block b. Each $\bar{X}^*$ thus consists of two sub-blocks: $\bar{X}^*_1$ that contains the states visited from beginning of $\bar{X}^*$ (empty if the first state is γ^*) to the state right before hitting γ^*, and sub-block $\bar{X}^*_2$ that contains the rest of $\bar{X}^*$ (a random number of regenerative cycles).

Since a block $\bar{X}^*$ starts after discontinuity in playing the best arm, $\bar{b}^*(n)$ is less than or equal to total number of completed blocks in which the best arm is not played up to time n. Thus

$$
E_\gamma[\bar{b}^*(n)] \leq \sum_{i:\mu^i < \mu^*} E_\gamma[B^i(b(n))]. \quad (21)
$$

We rewrite $R^*(n)$ in the following from:

$$
\begin{aligned}
R^*(n) =\;& \sum_{y \in S^*} r^*_y E_\gamma\left[\sum_{j=1}^{\bar{b}^*(n)} \sum_{X^*_t \in \bar{X}^*_2(j)} I(X^*_t = y) \right] \quad (22) \\
& - \sum_{y \in S^*} r^*_y \pi^*_y E_\gamma\left[\sum_{j=1}^{\bar{b}^*(n)} |\bar{X}^*_2(j)| \right] \quad (23)
\end{aligned}
$$

$$+ \sum_{y \in S^*} r_y^* E_\gamma \left[\sum_{j=1}^{\bar{b}^*(n)} \sum_{X_t^* \in \bar{X}_1^*(j)} I(X_t^* = y) \right] \quad (24)$$

$$- \sum_{y \in S^*} r_y^* \pi_y^* E_\gamma \left[\sum_{j=1}^{\bar{b}^*(n)} |\bar{X}_1^*(j)| \right] \quad (25)$$

$$> 0 - \mu^* M_{\max}^* \sum_{i:\mu^i < \mu^*} E_\gamma[B^i(b(n))] \quad (26)$$

where the last inequality is obtained by noting the difference between (22) and (23) is zero by Lemma 2, using positivity of rewards to lower bound (24) by 0, and (21) to upper bound (25). Combine this with (16) and (20) we can thus obtain a logarithmic upper bound on $-Z_\gamma(n)$. Finally, we have

$$\mu^* E_\gamma[n - T(n)] - E_\gamma \left[\sum_{t=T(n)+1}^{n} r_{x_{\alpha(t)}}^{\alpha(t)} \right]$$

$$\leq \mu^* \left(\frac{1}{\pi_{\min}} + \max_{i \in \{1,\dots,K\}} M_{\max}^i + 1 \right) . \quad (27)$$

Therefore we have obtained the stated logarithmic bound for (18). Note that this bound does not depend on γ, and therefore is also an upper bound for $R(n)$, completing the proof. ∎

Therefore, given minimal information about the arms such as an upper bound for $S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$ the player can guarantee logarithmic regret by choosing an L in the RCA algorithm that satisfies the condition in Theorem 2.

We end this section by noting that the logarithmic bound in n is also *order optimal* for this restless bandit problem, i.e., no better bound than $\ln n$ is possible (however a better constant is possible). This follows from the fact that the *rested* bandit problem is a special case of the *restless* problem and in [5] it is shown that the best order is logarithmic for the *rested* problem. Moreover, we conjecture that the order optimality of RCA holds when it is used with any index policy that is order optimal for the rested bandit problem. Because of the use of regenerative cycles in RCA, the observations used to calculate the indices can be in effect treated as coming from rested arms. Thus an approach similar to the one in Theorem 1 can be used to prove order optimality.

V. AN EXAMPLE: GILBERT-ELLIOT CHANNEL MODEL

In this section we simulate RCA and two other algorithms under the commonly used Gilbert-Elliot channel model, where each channel has two states, *good* and *bad* (or $1, 0$, respectively). The first algorithm is the upper confidence bound (UCB1) algorithm from [7]. In [8] we have proved that it has a logarithmic regret in the case of Markovian rewards when all arms are rested, by replacing the constant 2 in the index calculation of UCB1 with L and using a result from [15]. Using Lezaud's bound as we have done in the present paper it can be shown that this modified UCB1 algorithm, shown in Figure 4, has a logarithmic regret for $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$ for the rested bandit problem.

Upper Confidence Bound (UCB1):
1: Initialize: $n = 1$
2: **for** $n \leq K$ **do**
3:   play arm n; $n := n + 1$.
4: **end for**
5: **while** $n > K$ **do**
6:   $\bar{r}^i(T^i(n)) = \frac{r^i(1) + r^i(2) + \dots + r^i(T^i(n))}{T^i(n)}$, $\forall i$
7:   $g_{n,T^i(n)}^i = \bar{r}^i(T^i(n)) + \sqrt{\frac{L \ln n}{T^i(n)}}$, $\forall i$
8:   play arm j, such that $j = \arg\max_i g_{n,T^i(n)}^i$, update $r^j(n)$ and $T^j(n)$.
9:   $n := n + 1$
10: **end while**

Fig. 4. The UCB1 algorithm.

The second algorithm is an online randomized algorithm proposed in [1], referred to as the Exp3 algorithm and shown in Figure 5. The main distinction of Exp3 is that it is a randomized algorithm: given all past observations the algorithm's current action is the outcome of a random variable. Randomization is helpful when rewards from arms are determined by an adversary rather than a stochastic process. This is the context in which Exp3 is introduced and studied in [1].

Exp3:
1: Initialize: select parameter $a \in (0,1)$ and set weights $w^i(1) = 1, \forall i \in \{1, 2, \cdots, K\}$
2: **while** (1) **do**
3:   at time n compute the probabilities $p^i(n) = (1 - a)\frac{w^i(n)}{\sum_{j=1}^{K} w^j(n)} + \frac{a}{K}$, $\forall i$.
4:   take a random sample of the random variable $X(n)$ with pmf: $P(X(n) = i) = p^i(n)$; denote the outcome by $\alpha(n)$.
5:   play arm $\alpha(n)$, and get reward $r^{\alpha(n)}$.
6:   **if** $\alpha(n) = i$ **then**
7:     set weight $w^i(n+1) = w^i(n) \exp(\frac{ar^i(n)}{Kp^i(n)})$
8:   **else**
9:     $w^i(n+1) = w^i(n)$
10:   **end if**
11: **end while**

Fig. 5. The Exp3 algorithm.

We simulate and compare the regret of these three algorithms averaged over 100 runs, under two scenarios, denoted S1 and S2, respectively. Each scenario involves 5 two-state channels with varying transition probabilities. The statistics and rewards used are given in Table V. Exp3 is run with two different values of a: $a_1 = 0.1$, $a_2 = \min\left\{1, \sqrt{\frac{K \ln K}{(e-1)N}}\right\}$ where $N = 10^5$ is the time horizon. All arms are assumed to be in stationary distribution at the beginning. $112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$ is equal to 9556 in S1 and 1037.2 in S2.

Results are shown in Figures 6 and 7, under scenarios S1

S1	p_{01}, p_{10}	r_0, r_1	S2	p_{01}, p_{10}	r_0, r_1
ch.1	0.01, 0.03	0.1, 1	ch.1	0.1, 0.2	0.1, 1
ch.2	0.04, 0.01	0.1, 1	ch.2	0.1, 0.3	0.1, 1
ch.3	0.03, 0.01	0.1, 1	ch.3	0.5, 0.1	0.1, 1
ch.4	0.02, 0.01	0.1, 1	ch.4	0.1, 0.4	0.1, 1
ch.5	0.01, 0.02	0.1, 1	ch.5	0.1, 0.5	0.1, 1

TABLE I

CHANNEL PARAMETERS
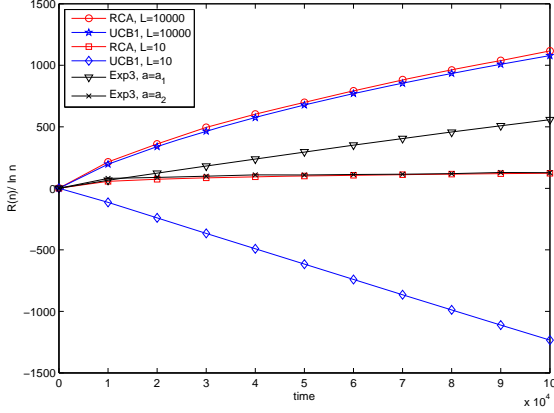

Fig. 6.   Regret under scenario S1

Fig. 7.   Regret under scenario S2

and S2, respectively. We make the following observations from this set of curves. Firstly, both RCA's and UCB1's performance improves when a smaller value of L is used. This suggests that the condition $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$ is sufficient but not in general necessary for the logarithmic regret to hold. Secondly, Exp3 shows good performance when a_2 is the constant choice, which utilizes the knowledge of time horizon. If the time horizon is not given, then Exp3 has a linear regret instead as was proven in [1]. Lastly, overall the performance of UCB1 is competitive compared to RCA, which has been shown to have logarithmic regret in the previous section. In particularly, in Figure 6 for $L = 10$ UCB1 outperforms RCA significantly. This is because in this case the channels are very bursty, thus updating the indices at every time step in UCB1 is a better option than waiting for regenerative cycles to occur in RCA, which can take a long time for an update to occur. These results suggest that there may exist logarithmic bounds for UCB1 as well. Furthermore, they suggest obvious ways to improve the performance of RCA. However, as discussed earlier due to the restless nature of the arms when the indices are updated constantly the problem becomes intractable. It remains an interesting future study to show such bounds for UCB1.

VI. CONCLUSION

We considered the OSA problem when the primary users' activities are modeled as generic finite-state Markov chains. This was formulated as a single-player restless bandit problem. We proposed an algorithm that updates the sample mean based indices using regenerative sample paths and showed that its regret can be upper bounded uniformly and logarithmically over time. This is the first results showing that log-regret is
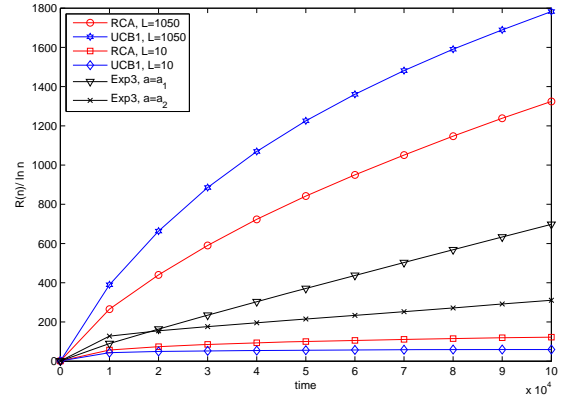
possible in a restless bandit learning problem. We numerically compare the performance of RCA with two other algorithms, UCB1 and Exp3, and conjectured that similar logarithmic bounds may exist for UCB1 as well.

REFERENCES

[1] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, pp. 48–77, 2002.

[2] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 55, pp. 527–535, 1952.

[3] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

[4] V. Anantharam, P. Varaiya, and J. . Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards," *IEEE Trans. Automat. Contr.*, pp. 968–975, November 1987.

[5] ——, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards," *IEEE Trans. Automat. Contr.*, pp. 977–982, November 1987.

[6] R. Agrawal, "Sample mean based index policies with o(log n) regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, December 1995.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, p. 235256, 2002.

[8] C. Tekin and M. Liu, "Online algortihms for the multi-armed bandit problem with markovian rewards, http://arxiv.org/abs/1007.2238."

[9] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players, http://arxiv.org/abs/0910.2065."

[10] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple players: Learning under competition," in *Proc. of IEEE INFOCOM*, March 2010.

[11] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: a combinatorial multi-armed bandit formulation," in *IEEE Symp. on Dynamic Spectrum Access Networks (DySPAN)*, April 2010.

[12] J. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society*, vol. 41, no. 2, pp. 148–177, 1979.

[13] P. Whittle, "Restless bandits," *J. Appl. Prob.*, pp. 301–313, 1988.

[14] S. H. A. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multi-channel opportunistic access," *IEEE Transactions on Information Theory*, vol. 55, no. 9, pp. 4040–4050, September 2009.

[15] D. Gillman, "A chernoff bound for random walks on expander graphs," *SIAM J. Comp.*, vol. 27, no. 4, p. 12031220, 1998.

[16] P. Lezaud, "Chernoff-type bound for finite markov chains," *Ann. Appl. Prob.*, vol. 8, pp. 849–867, 1998.

[17] P. Bremaud, *Markov Chains, Gibbs Fields, Monte Carlo Simulation and Queues.* Springer, 1998.