Online Learning of Rested and Restless Bandits

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

In this paper we study the online learning problem involving *rested* and *restless* multiarmed bandits with multiple plays. The system consists of a single player/user and a set of K finite-state discrete-time Markov chains (*arms*) with unknown state spaces and statistics. At each time step the player can play M, $M \leq K$, arms. The objective of the user is to decide for each step which M of the K arms to play over a sequence of trials so as to maximize its long term reward. The restless multiarmed bandit is particularly relevant to the application of opportunistic spectrum access (OSA), where a (secondary) user has access to a set of K channels, each of time-varying condition as a result of random fading and/or certain primary users' activities.

We first show that a logarithmic regret algorithm exists for the *rested* multiarmed bandit problem. We then construct an algorithm for the restless bandit problem which utilizes regenerative cycles of a Markov chain and computes a sample mean based index policy. We 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.

I. Introduction

In this paper we study the online learning problem involving *rested* and *restless* multiarmed bandits with multiple plays. The system consists of a single player/user and a set of K finite-state discrete-time Markov chains (also referred to as *arms*) with unknown state spaces and statistics. At each time step the player can play M, $M \leq K$, arms. Each arm played generates a reward depending on the state the arm is in when played. The state of an arm is only observed when it is played, and otherwise unknown to the user. The objective of the user is to decide for each step which M of the K arms to play over a

Preliminary versions of this work appeared in Allerton 2010 and Infocom 2011.

sequence of trials so as to maximize its long term reward. To do so it must use all its past actions and observations to essentially learn the quality of each arm (e.g., their expected rewards). We consider two cases, one with *rested* arms where the state of a Markov chain stays frozen unless it's played, the other with *restless* arms where the state of a Markov chain may continue to evolve (accordingly to a possibly different law) regardless of the player's actions.

The above problem is motivated by the following opportunistic spectrum access (OSA) problem. A (secondary) user has access to a set of K channels, each of time-varying condition as a result of random fading and/or certain primary users' activities. The condition of a channel is assumed to evolve as a 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) senses or probes M of the K channels to find out their condition, and is allowed to use the channels in a way consistent with their conditions. For instance, good channel conditions result in higher data rates or lower power for the user and so on. In some cases channel conditions are simply characterized as being available and unavailable, and the user is allowed to use all channels sensed to be available. 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.

The restless bandit model is particularly relevant to this application because the state of each Markov chain evolves independently of the action of the user. 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 *player*, and probing a channel as *playing or selecting an arm*.

Within this context, the user'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 user'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. Discussion on possibly stronger regret measures is given in Section VI.

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 set of best arms and avoid getting stuck playing an inferior one erroneously believed to be in the set of best arms.

On the other hand, the player needs to avoid spending too much time sampling the arms and collecting statistics and not playing the best arms 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. It's 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.

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 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] we showed 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. Parallel to the work presented here, in [9] an algorithm was constructed that achieves logarithmic regret for the restless bandit problem. The mechanism behind this algorithm however is quite different from what's presented here; this difference is discussed in more detail in Section VI.

Other works such as [10], [11], [12] consider the iid reward case in a decentralized multiplayer 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 of the arms 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 [13] and the optimal policy is proved to be an index policy which at each time plays the arm with highest Gittins' index. Whittle introduced the restless version of the bandit problem in [14]. 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 [15] for an OSA problem formulated as a restless bandit problem with each channel modeled as a two-state Markov chain (the Gilbert-Elliot model).

In this paper we first study the rested bandit problem with Markovian rewards. Specifically, we show that a straightforward extension of the UCB1 algorithm [7] to the multiple play case (UCB1 was originally designed for the case of a single play: $M = 1$) results in logarithmic regret for restless bandits with Markovian rewards. We then use the key difference between rested and restless bandits to construct a regenerative cycle algorithm (RCA) that produces logarithmic regret for the restless bandit problem. The construction of this algorithm allows us to use the proof of the rested problem as a natural stepping stone, and simplifies the presentation of the main conceptual idea.

The work presented in this paper extends our previous results [8], [16] on single play to multiple plays ($M \geq 1$). Note that this single player model with multiple plays at each time step is conceptually equivalent to the centralized (coordinated) learning by multiple players, each playing a single arm at each time step. Indeed our proof takes this latter point of view for ease of exposition, and our results on logarithmic regret equally applies to both cases.

The remainder of this paper is organized as follows. In Section II we present the problem formulation. In Section III we analyze a sample mean based algorithm for the rested bandit problem. In Section IV we propose an algorithm based on regenerative cycles that employs sample mean based indices and analyze its regret. In Section V we numerically examine the performance of this algorithm in the case of an OSA problem with Gilbert-Elliot channel model. In Section VI we discuss possible improvements and compare our algorithm to other algorithms. Section VII concludes the paper.

II. Problem Formulation and Preliminaries

Consider K arms (or channels) indexed by the set $\mathcal{K} = \{1, 2, \ldots, K\}$. 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_x^i denote the reward obtained from state x of arm i, $x \in S^i$; this reward is in general different for different states. Let $P^i = \left\{p_{xy}^i, x, y \in S^i\right\}$ denote

the transition probability matrix of the i-th arm, and $\boldsymbol{\pi}^i = \{\pi^i_x, x \in S^i\}$ the stationary distribution of P^i.

We assume the arms (the Markov chains) are mutually independent. In subsequent sections we will consider the rested and the restless cases separately. As mentioned in the introduction, the state of a rested arm changes according to P^i only when it is played and remains frozen otherwise. By contrast, the state of a restless arm changes according to P^i regardless of the user's actions. All the assumptions in this section applies to both types of arms. We note that the rested model is a special case of the restless model, but our development under the restless model follows the rested model[1].

Let $(P^i)'$ denote the *adjoint* of P^i on $l_2(\pi)$ where

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

and $\hat{P}^i = (P^i)'P$ denotes the *multiplicative symmetrization* of P^i. We will assume that the P^is are such that $\hat{P}^i$s are irreducible. To give a sense of how weak or strong this assumption is, we first note that this is a weaker condition than assuming the Markov chains to be reversible. In addition, we note that one condition that guarantees the $\hat{P}^i$s are irreducible is $p_{xx} > 0, \forall x \in S^i, \forall i$. This assumption thus holds naturally for our main motivating application, as it's possible for channel condition to remain the same over a single time step (especially if the unit is sufficiently small). It also holds for a very large class of Markov chains and applications in general. Consider for instance a queueing system scenario where an arm denotes a server and the Markov chain models its queue length, in which it is possible for the queue length to remain the same over one time unit.

The mean reward of arm i, denoted by μ^i, is the expected reward of arm i under its stationary distribution:

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

Consistent with the discrete time Markov chain model, we will assume that the player's actions occur in discrete time steps. Time is indexed by t, $t = 1, 2, \cdots$. We will also frequently refer to the time interval $(t-1, t]$ as time slot t. The player plays M of the K arms at each time step.

Throughout the analysis we will make the additional assumption that the mean reward of arm M is strictly greater than the mean reward of arm $M+1$, i.e., we have $\mu^1 \geq \mu^2 \geq \cdots \geq \mu^M > \mu^{M+1} \geq$

[1]In general a restless arm may be given by two transition probability matrices, an active one (P^i) and a passive one (Q^i). The first describes the state evolution when it is played and the second the state evolution when it is not played. When an arm models channel variation, P^i and Q^i are in general assumed to be the same as the channel variation is uncontrolled. In the context of online learning we shall see that the selection of Q^i is irrelevant; indeed the arm does not even have to be Markovian when it's in the passive mode. More is discussed in Section VI.

$\cdots \geq \mu^K$. For rested arms this assumption simplifies the presentation and is not necessary, i.e., results will hold for $\mu^M \geq \mu^{M+1}$. However, for restless arms the strict inequality between μ^M and μ^{M+1} is needed because otherwise there can be a large number of arm switchings between the M-th and the $(M+1)$-th arms (possibly more than logarithmic). Strict inequality will prevent this from happening. We note that this assumption is not in general restrictive; in our motivating application distinct channel conditions typically means different data rates. Possible relaxation of this condition is given in Section VI.

We will refer to the set of arms $\{1, 2, \cdots, M\}$ as the M-best arms and say that each arm in this set is *optimal* while referring to the set $\{M+1, M+2, \cdots, K\}$ as the M-worst arms and say that each arm in this set is *suboptimal*.

For a policy α we define its regret $R^\alpha(n)$ as the difference between the expected total reward that can be obtained by only playing the M-best arms and the expected total reward obtained by policy α up to time n. Let $A^\alpha(t)$ denote the set of arms selected by policy α at t, $t = 1, 2, \cdots$, and $x_\alpha(t)$ the state of arm $\alpha(t) \in A^\alpha(t)$ at time t. Then we have

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

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 result is due to Lezaud [17] that bounds the probability of a large deviation from the stationary distribution.

Lemma 1: [Theorem 3.3 from [17]] 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 $f : S \to \mathbb{R}$ 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 result is due to Anantharam et al., which can be found in [5].

Lemma 2: [Lemma 2.1 from [5]] Let Y be an irreducible aperiodic Markov chain with a state space S, transition probability matrix P, an initial distribution that is non-zero in all states, and a stationary distribution $\{\pi_x\}, \forall x \in S$. Let F_t be the σ-field generated by random variables $X_1, X_2, ..., X_t$ where X_t corresponds to the state of the chain at time t. Let G be a σ-field independent of $F = \vee_{t\geq 1} F_t$, the smallest σ-field containing $F_1, F_2, ...$ Let τ be a stopping time with respect to the increasing family of σ-fields $\{G \vee F_t, t \geq 1\}$. Define $N(x, \tau)$ such that

$$N(x,\tau) = \sum_{t=1}^{\tau} I(X_t = x).$$

Then $\forall \tau$ such that $E[\tau] < \infty$, we have

$$|E[N(x,\tau)] - \pi_x E[\tau]| \leq C_P, \tag{3}$$

where C_P is a constant that depends on P.

The third result is due to Bremaud, which can be found in [18].

Lemma 3: 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 $y \in S$

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

The following notations are frequently used throughout the paper: $\beta = \sum_{t=1}^{\infty} 1/t^2$, $\pi^i_{\min} = \min_{x\in S^i} \pi^i_x$, $\pi_{\min} = \min_{i\in\mathcal{K}} \pi^i_{\min}$, $r_{\max} = \max_{x\in S^i, i\in\mathcal{K}} r^i_x$, $S_{\max} = \max_{i\in\mathcal{K}} |S^i|$, $\hat{\pi}_{\max} = \max_{x\in S^i, i\in\mathcal{K}} \{\pi^i_x, 1-\pi^i_x\}$, $\epsilon_{\min} = \min_{i\in\mathcal{K}} \epsilon^i$, where ϵ^i is the eigenvalue gap (the difference between 1 and the second largest eigenvalue) of the multiplicative symmetrization of the transition probability matrix of the ith arm, and $\Omega^i_{\max} = \max_{x,y\in S^i} \Omega^i_{x,y}$, where $\Omega^i_{x,y}$ is the mean hitting time of state y given the initial state x for arm i for P^i.

In the next two sections we present algorithms for the rested and restless problems, referred to as the *upper confidence bound - multiple plays* (UCB-M) and the *regenerative cycle algorithm - multiple plays* (RCA-M), respectively, and analyze their regret.

III. Analysis of the Rested Bandit Problem with Multiple Plays

In this section we show that there exists an algorithm that achieves logarithmic regret uniformly over time for the rested bandit problem with Markovian reward and multiple plays. We present such

an algorithm, called *the upper confidence bound - multiple plays* (UCB-M), which is a straightforward extension of UCB1 from [7]. This algorithm plays M of the K arms with the highest indices with a modified exploration constant L instead of 2 in [7]. Throughout our discussion, we will consider a horizon of n time slots. For simplicity of presentation we will view a single player playing multiple arms at each time as multiple coordinated players each playing a single arm at each time. In other words we consider M players indexed by $1, 2, \cdots, M$, each playing a single arm at a time. Since in this case information is centralized, collision is completely avoided among the players, i.e., at each time step an arm will be played by at most one player.

Below we summarize a list of notations used in this section.

- $A(t)$: the set of arms played at time t (or in slot t).
- $T^i(t)$: total number of times (slots) arm i is played up to the end of slot t.
- $T^{i,j}(t)$: total number of times (slots) player j played arm i up to the end of slot t.
- $\bar{r}^i(T^i(t))$: sample mean of the rewards observed from the first $T^i(t)$ plays of arm i.

As shown in Figure 1, UCB-M selects M channels with the highest indices at each time step and updates the indices according to the rewards observed. The index given on line 4 of Figure 1 depends on the sample mean reward and an exploration term which reflects the relative uncertainty about the sample mean of an arm. We call L in the exploration term *the exploration constant*. The exploration term grows logarithmically when the arm is not played in order to guarantee that sufficient samples are taken from each arm to approximate the mean reward.

The Upper Confidence Bound - Multiple Plays (UCB-M):

1: Initialize: Play each arm M times in the first K slots
2: **while** $t \geq K$ **do**
3: $\quad \bar{r}^i(T^i(t)) = \frac{r^i(1)+r^i(2)+\ldots+r^i(T^i(t))}{T^i(t)},\ \forall i$
4: $\quad$ calculate index: $g^i_{t,T^i(t)} = \bar{r}^i(T^i(t)) + \sqrt{\frac{L \ln t}{T^i(t)}},\ \forall i$
5: $\quad t := t + 1$
6: $\quad$ play M arms with the highest indices, update $r^j(t)$ and $T^j(t)$.
7: **end while**

Fig. 1. pseudocode for the UCB-M algorithm.

To upper bound the regret of the above algorithm logarithmically, we proceed as follows. We begin by relating the regret to the expected number of plays of the arms and then show that each suboptimal arm is played at most logarithmically in expectation. These steps are illustrated in the following lemmas. Most of these lemmas are established under the following condition on the arms.

Condition 1: All arms are finite-state, irreducible, aperiodic Markov chains whose transition probability matrices have irreducible multiplicative symmetrizations and $r_x^i > 0$, $\forall i \in \mathcal{K}$, $\forall x \in S^i$.

Lemma 4: Assume that all arms are finite-state, irreducible, aperiodic, rested Markov chains. Then using UCB-M we have:

$$\left| R(n) - \left(n \sum_{j=1}^{M} \mu^j - \sum_{i=1}^{K} \mu^i E[T^i(n)] \right) \right| \leq C_{\mathbf{S},\mathbf{P},\mathbf{r}}, \tag{4}$$

where $C_{\mathbf{S},\mathbf{P},\mathbf{r}}$ is a constant that depends on the state spaces, rewards, and transition probabilities but not on time.

Proof: see Appendix A. ■

Lemma 5: Assume Condition 1 holds and all arms are rested. Under UCB-M with $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$, for any suboptimal arm i, we have

$$E[T^i(n)] \leq M + \frac{4L \ln n}{(\mu^M - \mu^i)^2} + \sum_{j=1}^{M} \frac{(|S^i| + |S^j|)\beta}{\pi_{\min}}$$

Proof: see Appendix C. ■

Theorem 1: Assume Condition 1 holds and all arms are rested. With constant $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$ the regret of UCB-M is upper bounded by

$$R(n) \leq 4L \ln n \sum_{i>M} \frac{(\mu^1 - \mu^i)}{(\mu^M - \mu^i)^2} + \sum_{i>M} (\mu^1 - \mu^i) \left(M + \sum_{j=1}^{M} C_{i,j} \right) + C_{\mathbf{S},\mathbf{P},\mathbf{r}}, \tag{5}$$

where $C_{i,j} = \frac{(|S^i| + |S^j|)\beta}{\pi_{\min}}$.

Proof:

$$\begin{aligned} n \sum_{j=1}^{M} \mu^j - \sum_{i=1}^{K} \mu^i E[T^i(n)] &= \sum_{j=1}^{M} \sum_{i=1}^{K} \mu^j E[T^{i,j}(n)] - \sum_{j=1}^{M} \sum_{i=1}^{K} \mu^i E[T^{i,j}(n)] \\ &= \sum_{j=1}^{M} \sum_{i>M} (\mu^j - \mu^i) E[T^{i,j}(n)] \leq \sum_{i>M} (\mu^1 - \mu^i) E[T^i(n)]. \end{aligned}$$

Thus,

$$\begin{aligned} R(n) &\leq n\sum_{j=1}^{M}\mu^j - \sum_{i=1}^{K}\mu^i E[T^i(n)] + C_{\mathbf{S},\mathbf{P},\mathbf{r}} \quad (6)\\ &\leq \sum_{i>M}(\mu^1-\mu^i)E[T^i(n)] + C_{\mathbf{S},\mathbf{P},\mathbf{r}} \\ &\leq \sum_{i>M}(\mu^1-\mu^i)\left(M + \frac{4L\ln n}{(\mu^M-\mu^i)^2} + \sum_{j=1}^{M}\frac{(|S^i|+|S^j|)\beta}{\pi_{\min}}\right) + C_{\mathbf{S},\mathbf{P},\mathbf{r}} \quad (7)\\ &= 4L\ln n\sum_{i>M}\frac{(\mu^1-\mu^i)}{(\mu^M-\mu^i)^2} + \sum_{i>M}(\mu^1-\mu^i)\left(M+\sum_{j=1}^{M}C_{i,j}\right) + C_{\mathbf{S},\mathbf{P},\mathbf{r}}, \end{aligned}$$

where (6) follows from Lemma 4 and (7) follows from Lemma 5. ■

The above theorem says that provided that L satisfies the stated sufficient condition, UCB-M results in logarithmic regret for the rested problem. This sufficient condition does require certain knowledge on the underlying Markov chains. This requirement may be removed if the value of L is adapted over time. More is discussed in Section VI.

IV. Analysis of the Restless Bandit Problem with Multiple Plays

In this section we study the restless bandit problem. We construct an algorithm called the *regenerative cycle algorithm - multiple plays* (RCA-M), and prove that this algorithm guarantees logarithmic regret uniformly over time under the same mild assumptions on the state transition probabilities as in the rested case. RCA-M is a multiple plays extension of RCA first introduced in [16]. Below we first present the key conceptual idea behind RCA-M, followed by a more detailed pseudocode. We then prove the logarithmic regret result.

As the name suggests, RCA-M operates in regenerative cycles. In essence RCA-M uses the observations from sample paths within regenerative cycles to estimate the sample mean of an arm in the form of an index similar to that used in UCB-M while discarding the rest of the observations (only for the computation of the index, but they are added to the total reward). Note that the rewards from the discarded observations are collected but are not used to make decisions. The reason behind such a construction has to do with the restless nature of the arms. 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 the arms are not played continuously, the sequence of observations from an arm which is not played consecutively does not correspond to a discrete time homogeneous Markov chain. While this certainly does not affect

our ability to collect rewards, it becomes hard to analyze the estimated quality (the index) of an arm calculated based on rewards collected this way.

However, if instead of the actual sample path of 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.

Under RCA-M each player maintains a block structure; a block consists of a certain number of slots. Recall that as mentioned earlier, even though our basic model is one of single-player multiple-play, our description is in the equivalent form of multiple coordinated players each with a single play. Within a block a player plays the same arm continuously till a certain pre-specified state (say γ^i) is observed. Upon this observation the arm enters a regenerative cycle and the player continues to play the same arm till state γ^i is observed for the second time, which denotes the end of the block. Since M arms are played (by M players) simultaneously in each slot, different blocks overlap in time. Multiple blocks may or may not start or end at the same time. In our analysis below blocks will be ordered; they are ordered according to their start time. If multiple blocks start at the same time then the ordering among them is randomly chosen.

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. Figure 2 shows an example sample path of the operation of RCA-M. The block structure of two players are shown in this example; the ordering of the blocks is also shown.

The key to the RCA-M algorithm is for each arm to single out only observations within SB2's in each block and virtually assemble them. Throughout our discussion, we will consider a horizon of n time slots. A list of notations used is summarized as follows:

- $A(t)$: the set of arms played at time t (or in time slot t).
- γ^i: the state that determines the regenerative cycles for arm i.
- $\alpha(b)$: the arm played in the b-th block.
- $b(n)$: the total number of completed blocks by all players up to time n.
- $T(n)$: the time at the end of the last completed block across all arms (see Figure 2).
- $T^i(n)$: the total number of times (slots) arm i is played up to the last completed block of arm i up

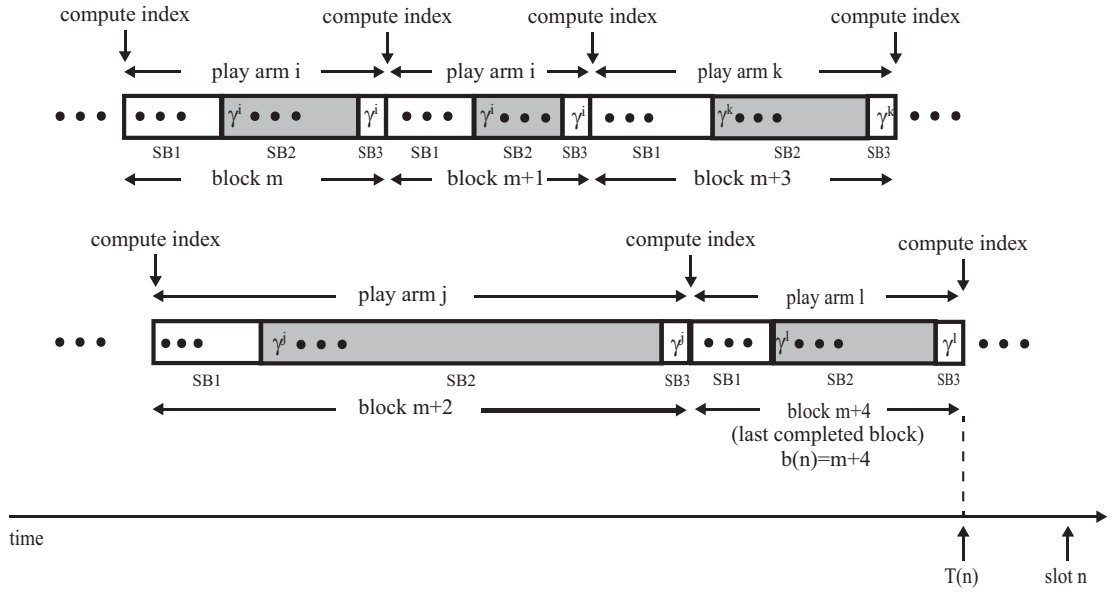

Fig. 2. Example realization of RCA-M with $M = 2$ for a period of n slots

to time $T(n)$.

- $T^{i,j}(n)$: the total number of times (slots) arm i is played by user j up to the last completed block of arm i up to time $T(n)$
- $B^i(b)$: the total number of blocks within the first completed b blocks in which arm i is played.
- $X_1^i(b)$: the vector of observed states from SB1 of the b-th block in which arm i is played; this vector is empty if the first observed state is γ^i.
- $X_2^i(b)$: the vector of observed states from SB2 of the b-th block in which arm i is played;
- $X^i(b)$: the vector of observed states from the b-th block in which arm i is played. Thus we have $X^i(b) = [X_1^i(b), X_2^i(b), \gamma^i]$.
- $t(b)$: time at the end of block b;
- $T^i(t(b))$: the total number of time slots arm i is played up to the last completed block of arm i within time $t(b)$.
- $t_2(b)$: the total number of time slots that lie within at least one SB2 in a completed block of any arm up to and including block b.
- $r^i(t)$: the reward from arm i upon its t-th play, counting only those plays during an SB2.
- $T_2^i(t_2(b))$: the total number of time slots arm i is played during SB2's up to and including block b.
- $O(b)$: the set of arms that are *free* to be selected by some player i upon its completion of the b-th

block; these are arms that are currently not being played by other players (during time slot $t(b)$), and the arms whose blocks are completed at time $t(b)$.

RCA-M computes and updates the value of an *index* g^i for each arm i in the set $O(b)$ at the end of block b based on the total reward obtained from arm i during all SB2's as follows:

$$g^i_{t_2(b),T_2^i(t_2(b))} = \bar{r}^i(T_2^i(t_2(b))) + \sqrt{\frac{L \ln t_2(b)}{T_2^i(t_2(b))}}, \tag{8}$$

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 SB2. Note that this is the same way the index is computed under UCB-M if we only consider SB2's. Its also worth noting that under RCA-M rewards are also collected during SB1's and SB3's. However, the computation of the indices only relies on SB2. The pseudocode of RCA-M is given in Figure 3.

Due to the regenerative nature of the Markov chains, the rewards used in the computation of the index of an arm can be viewed as rewards from a rested arm with the same transition matrix as the active transition matrix of the restless arm. However, to prove the existence of a logarithmic upper bound on the regret for restless arms remains a non-trivial task since the blocks may be arbitrarily long and the frequency of arm selection depends on the length of the blocks.

In the analysis that follows, we first show that the expected number of blocks in which a suboptimal arm is played is at most logarithmic by applying the result in Lemma 7 that compares the indices of arms in slots where an arm is selected. Using this result we then show that the expected number of blocks in which a suboptimal arm is played is at most logarithmic in time. Using irreducibility of the arms the expected block length is finite, thus the number of time slots in which a suboptimal arm is played is finite. Finally, we show that the regret due to arm switching is at most logarithmic.

We bound the expected number of plays from a suboptimal arm.

Lemma 6: Assume Condition 1 holds and all arms are restless. Under RCA-M with a constant $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$, we have

$$\sum_{i>M} (\mu^1 - \mu^i) E[T^i(n)] \leq 4L \sum_{i>M} \frac{(\mu^1 - \mu^i) D_i \ln n}{(\mu^M - \mu^i)^2} + \sum_{i>M} (\mu^1 - \mu^i) D_i \left(1 + M \sum_{j=1}^{M} C_{i,j} \right) ,$$

The Regenerative Cycle Algorithm - Multiple Plays (RCA-M):

Initialize: $b=1, t=0, t_2=0, T_2^i=0, r^i=0, I_{SB2}^i=0, I_{IN}^i=1, \forall i=1,\cdots,K,\ A=\emptyset$
  //I_{IN}^i indicates whether arm i has been played at least once
  //I_{SB2}^i indicates whether arm i is in an SB2 sub-block
while (1) **do**
  for $i=1$ to K **do**
    if $I_{IN}^i=1$ and $|A|<M$ **then**
      $A \leftarrow A \cup \{i\}$ //arms never played is given priority to ensure all arms are sampled initially
    end if
  end for
  if $|A|<M$ **then**
    Add to A the set $\left\{i : g^i \text{ is one of the } M-|A| \text{ largest among } \{g^k, k\in\{1,\cdots,K\}-A\}\right\}$
      //for arms that have been played at least once, those with the largest indices are selected
  end if
  for $i \in A$ **do**
    play arm i; denote state observed by x^i
    if $I_{IN}^i=1$ **then**
      $\gamma^i=x^i$, $T_2^i:=T_2^i+1$, $r^i:=r^i+r_{x^i}^i$, $I_{IN}^i=0$, $I_{SB2}^i=1$
        //the first observed state becomes the regenerative state; the arm enters SB2
    else if $x^i\neq\gamma^i$ and $I_{SB2}^i=1$ **then**
      $T_2^i:=T_2^i+1$, $r^i:=r^i+r_{x^i}^i$
    else if $x^i=\gamma^i$ and $I_{SB2}^i=0$ **then**
      $T_2^i:=T_2^i+1$, $r^i:=r^i+r_{x^i}^i$, $I_{SB2}^i=1$
    else if $x^i=\gamma^i$ and $I_{SB2}^i=1$ **then**
      $r^i:=r^i+r_{x^i}^i$, $I_{SB2}^i=0$, $A\leftarrow A-\{i\}$
    end if
  end for
  $t:=t+1$, $t_2:=t_2+\min\left\{1,\sum_{i\in S} I_{SB2}^i\right\}$ //t_2 is only accumulated if at least one arm is in SB2
  for $i=1$ to K **do**
    $g^i=\frac{r^i}{T_2^i}+\sqrt{\frac{L\ln t_2}{T_2^i}}$
  end for
end while

Fig. 3. Pseudocode of RCA-M

where

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

Proof: see Appendix E. ■

We now state the main result of this section.

Theorem 2: Assume Condition 1 holds and all arms are restless. With constant $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2/\epsilon_{\min}$

the regret of RCA-M is upper bounded by

$$\begin{aligned} R(n) &< 4L\ln n \sum_{i>M} \frac{1}{(\mu^M-\mu^i)^2}\left((\mu^1-\mu^i)D_i+E_i\right) \\ &+ \sum_{i>M}\left((\mu^1-\mu^i)D_i+E_i\right)\left(1+M\sum_{j=1}^{M}C_{i,j}\right)+F \end{aligned}$$

where

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

Proof: see Appendix F. ■

Theorem 2 suggests that given minimal information about the arms such as an upper bound for $S^2_{\max}r^2_{\max}\hat{\pi}^2_{\max}/\epsilon_{\min}$ the player can guarantee logarithmic regret by choosing an L in RCA-M that satisfies the stated condition. As the rested case, this requirement on L can be completely removed if the value of L is adapted over time; more is discussed in Section VI.

We conjecture that the order optimality of RCA-M 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-M, the observations used to calculate the indices can be in effect treated as coming from rested arms. Thus an approach similar to the one used in the proof of Theorem 2 can be used to prove order optimality of combinations of RCA-M and other index policies.

V. An Example for OSA: Gilbert-Elliot Channel Model

In this section we simulate RCA-M under the commonly used Gilbert-Elliot channel model where each channel has two states, *good* and *bad* (or 1, 0, respectively). We assume that channel state transitions are caused by primary user activity, therefore the problem reduces to the OSA problem. For any channel i, $r^i_1=1$, $r^i_0=0.1$. We simulate RCA-M in four environments with different state transition probabilities. We compute the normalized regret values, i.e., the regret per single play $R(n)/M$ by averaging the results of 100 runs.

The state transition probabilities are given in Table I and the mean rewards of the channels under these state transition probabilities are given in Table II. The four environment, denoted as S1, S2, S3 and S4, respectively, are summarized as follows. In S1 channels are bursty with mean rewards not close to each other; in S2 channels are non-bursty with mean rewards not close to each other; in S3 there are bursty and non-bursty channels with mean rewards not close to each other; and in S4 there are bursty and non-bursty channels with mean rewards close to each other.

In Figures 4, 6, 8, 10, we observe the normalized regret of RCA-M for the minimum values of L such that the logarithmic bound hold. However, comparing with Figures 5, 7, 9, 11 we see that the normalized regret is smaller for $L = 1$. Therefore the condition on L we have for the logarithmic bound, while sufficient, does not appear necessary. We also observe that for the Gilbert-Elliot channel model the regret can be smaller when L is set to a value smaller than $112S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2/\epsilon_{\min}$.

channel	1	2	3	4	5	6	7	8	9	10
S1, p_{01}	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.05
S1, p_{10}	0.08	0.07	0.08	0.07	0.08	0.07	0.02	0.01	0.02	0.01
S2, p_{01}	0.1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
S2, p_{10}	0.9	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
S3, p_{01}	0.01	0.1	0.02	0.3	0.04	0.5	0.06	0.7	0.08	0.9
S3, p_{10}	0.09	0.9	0.08	0.7	0.06	0.5	0.04	0.3	0.02	0.1
S4, p_{01}	0.02	0.04	0.04	0.5	0.06	0.05	0.7	0.8	0.9	0.9
S4, p_{10}	0.03	0.03	0.04	0.4	0.05	0.06	0.6	0.7	0.8	0.9

TABLE I
TRANSITION PROBABILITIES

channel	1	2	3	4	5	6	7	8	9	10
S1	0.20	0.21	0.28	0.30	0.35	0.37	0.70	0.82	0.74	0.85
S2	0.19	0.19	0.28	0.37	0.46	0.55	0.64	0.73	0.82	0.91
S3	0.19	0.19	0.28	0.37	0.46	0.55	0.64	0.73	0.82	0.91
S4	0.460	0.614	0.550	0.600	0.591	0.509	0.585	0.580	0.577	0.550

TABLE II
MEAN REWARDS

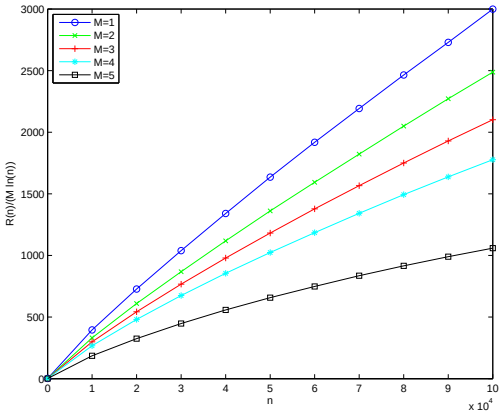

Fig. 4. Normalized regret under S1, $L = 7200$

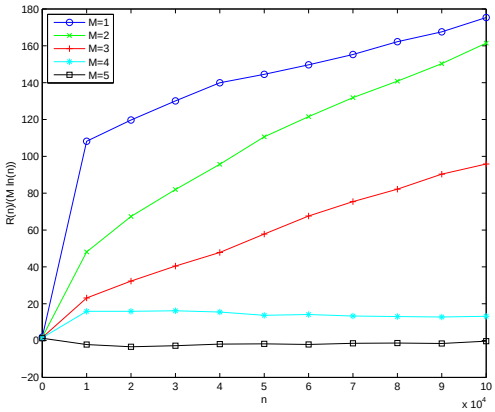

Fig. 5. Normalized regret under S1, $L = 1$

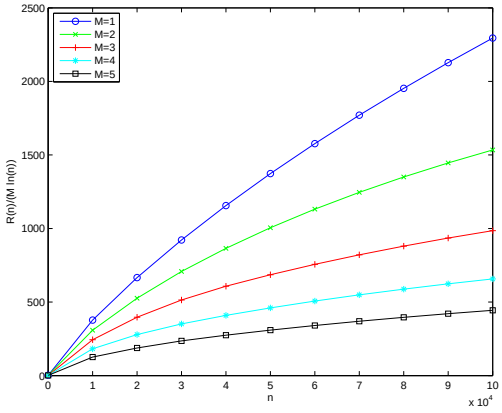

Fig. 6. Normalized regret under S2, $L = 360$

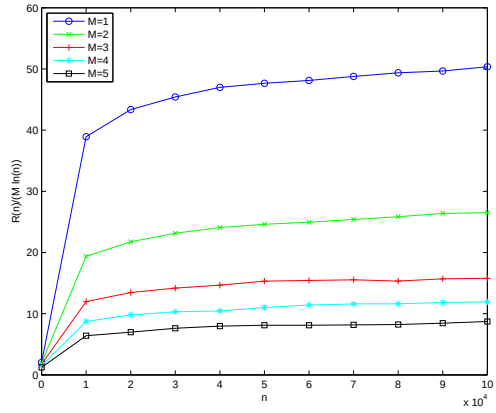

Fig. 7. Normalized regret under S2, $L = 1$

VI. Discussion

In this section we discuss how the performance of RCA-M may be improved (in terms of the constants and not in order), and possible relaxation and extensions.

A. Applicability, Performance Improvement, and Relaxation

We note that the same logarithmic bound derived in this paper holds for the general restless bandit where the state evolution is given by two matrices: the active and passive transition probability matrices (P^i and Q^i respectively for arm i), which are potentially different. The addition of a different Q^i does not affect the analysis because the reward to the player from an arm is determined only by the active

transition probability matrix and the first state after a discontinuity in playing the arm. Since the number of plays from any suboptimal arm is logarithmic and the expected hitting time of any state is finite the regret due to Q^i is at most logarithmic. We further note that for the same reason the arm may not even follow a Markovian rule in the passive state, and the same logarithmic bound will continue to hold.

The regenerative state for an arm under RCA-M is chosen based on the random initial observation. This means that RCA-M may happen upon a state with long recurrence time which will result in long SB1 and SB2 sub-blocks. We propose the following modification: RCA-M records all observations from all arms. Let $k_i(s,t)$ be the total number of observations from arm i up to time t that are *excluded* from the computation of the index of arm i when the regenerative state is s. Recall that the index of an arm is computed based on observations from regenerative cycles; this implies that $k_i(s,t)$ is the total number of slots in SB1's when the regenerative state is s. Let t_n be the time at the end of the n-th block. If the arm to be played in the n-th block is i then the regenerative state is set to $\gamma^i(n) = \arg\min_{s\in S^i} k_i(s,t_{n-1})$. The idea behind this modification is to estimate the state with the smallest recurrence time and choose the regenerative cycles according to this state. With this modification the number of observations that does not contribute to the index computation and the probability of choosing a suboptimal arm can be minimized over time.

It's also worth noting that the selection of the regenerative state γ^i in each block in general can be arbitrary: within the same SB2, we can start and end in different states. As long as we guarantee that two successive SB2's end and start with the same state, we will have a continuous sample path for which our analysis in Section IV holds.

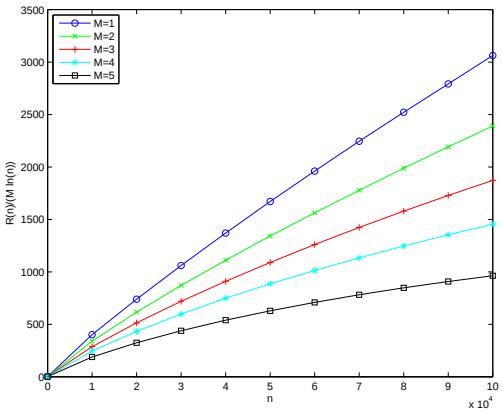

Fig. 8. Normalized regret under S3, $L = 3600$

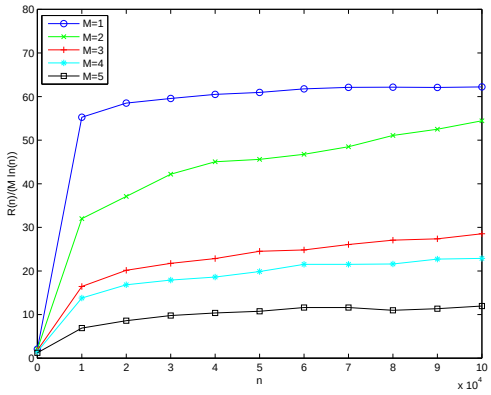

Fig. 9. Normalized regret under S3, $L = 1$

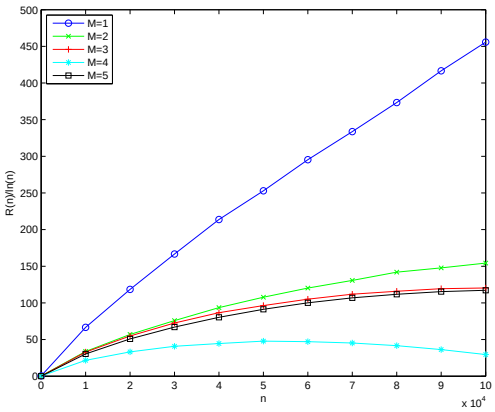

Fig. 10. Normalized regret under S4, $L = 7200$

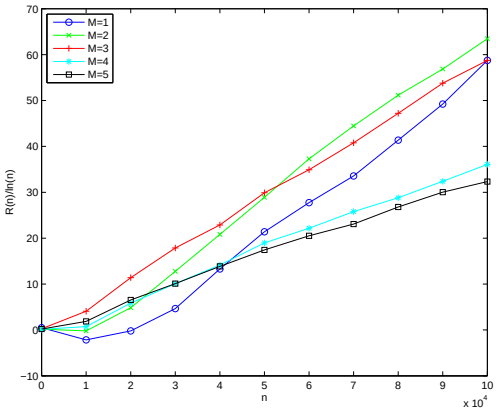

Fig. 11. Normalized regret under S4, $L = 1$

B. Relaxation of Certain Conditions

We have noted in Section V that the condition on L while sufficient does not appear necessary for the logarithmic regret bound to hold. Indeed our examples show that smaller regret can be achieved by setting $L = 1$. Note that this condition on L originates from the large deviation bound by Lezaud given in Lemma 1. This condition can be relaxed if we use a tighter large deviation bound.

We further note that even if no information is available on the underlying Markov chains to derive this sufficient condition on L, an $o(log(n)f(n))$ regret is achievable by letting L grow slowly with time where $f(n)$ is any increasing sequence. Such approach has been used in other settings and algorithms, see e.g., [11], [9].

We have noted earlier that the strict inequality $\mu^M > \mu^{M+1}$ is required for the restless multiarmed bandit problem because in order to have logarithmic regret, we can have no more than a logarithmic number of discontinuities from the optimal arms. When $\mu^M = \mu^{M+1}$ the rankings of the indices of arms M and $M + 1$ can oscillate indefinitely resulting in a large number of discontinuities. Below we briefly discuss how to resolve this issue if indeed $\mu^M = \mu^{M+1}$. Consider adding a threshold ϵ to the algorithm such that a new arm will be selected instead of an arm currently being played only if the index of that arm is at least ϵ larger than the index of the currently played arm which has the smallest index among all currently played arms. Then given that ϵ is sufficiently small (with respect to the differences of mean rewards) indefinite switching between the M-th and the $M + 1$-th arms can be avoided. However, further analysis is needed to verify that this approach will result in logarithmic regret.

C. Definition of Regret

We have used the weak regret measure throughout this paper, which compares the learning strategy with the best single-action strategy. When the statistics are known a priori, it is clear that in general the best one can do is not a single-action policy (in principle one can drive such a policy using dynamic programming). Ideally one could try to adopt a regret measure with respect to this optimal policy. However, such an optimal policy in the restless case is not known in general [14], [19], which makes the comparison intractable, except for some very limited cases when such a policy happens to be known [15], [20].

D. Extensions to A Decentralized Multiplayer Setting and Comparison with Similar Work

As mentioned in the introduction, there has been a number of recent studies extending single player algorithms to multi-player settings where collisions are possible [21], [11]. Within this context we note that RCA-M in its currently form does not extend in a straightforward way to a decentralized multi-player setting. It remains an interesting subject of future study. A recent work [9] considers the same restless multiarmed bandit problem studied in the present paper. They achieve logarithmic regret by using exploration and exploitation blocks that grow geometrically with time. The construction in [9] is very different from ours, but is amenable to multi-player extension [21] due to the constant, though growing, nature of the block length which can be synchronized among players.

It is interesting to note that the essence behind our approach RCA-M is to reduce a restless bandit problem to a rested bandit problem; this done by sampling in a way to construct a continuous sample path, which then allows us to use the same set of large deviation bounds over this reconstructed, entire sample path. By contrast, the method introduced in [9] applies large deviation bounds to individual segments (blocks) of the observed sample path (which is not a continuous sample path representative of the underlying Markov chain because the chain is restless); this necessitates the need to precisely control the length and the number of these blocks, i.e., they must grow in length over time. Another difference is that under our scheme, the exploration and exploitation are done simultaneously and implicitly through the use of the index, whereas under the scheme in [9], the two are done separately and explicitly through two different types of blocks.

VII. Conclusion

In this paper we considered the rested and restless multiarmed bandit problem with Markovian rewards and multiple plays. We showed that a simple extension to UCB1 produces logarithmic regret uniformly

over time. We then constructed an algorithm RCA-M that utilizes regenerative cycles of a Markov chain to compute a sample mean based index policy. The sampling approach reduces a restless bandit problem to the rested version, and we showed that under mild conditions on the state transition probabilities of the Markov chains this algorithm achieves logarithmic regret uniformly over time for the restless bandit problem, and that this regret bound is also optimal. We numerically examine the performance of this algorithm in the case of an OSA problem with the Gilbert-Elliot channel model.

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," in *Allerton Conference*, 2010.

[9] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Non-bayesian restless multi-armed bandit," *Techinal Report, UC Davis*, October 2010.

[10] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," vol. 58, no. 11, pp. 5667–5681, November 2010.

[11] A. Anandkumar, N. Michael, A. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," 2010.

[12] 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.

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

[14] P. Whittle, "Restless bandits: : Activity allocation in a changing world," *A Celebration of Applied Probability, ed. J. Gani, Journal of applied probability*, vol. 25A, pp. 287–298, 1988.

[15] 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.

[16] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," *IEEE INFOCOM*, April 2011.

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

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

[19] J. T. C. Papadimitriou, "The complexity of optimal queuing network control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293–305, May 1999.

[20] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao, "The non-bayesian restless multi-armed bandit: A case of near-logarithmic regret," *Proc. of Internanional Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011.

[21] H. Liu, K. Liu, and Q. Zhao, "Learning and sharing in a changing world: Non-bayesian restless bandit with multiple players," *Proc. of Information Theory and Applications Workshop (ITA)*, January 2011.

APPENDIX A
PROOF OF LEMMA 4

Let $X^{i,j}(t)$ be the state observed from the tth play of arm i by player j and $T^{i,j}(n)$ be the total number of times player j played arm i up to and including time n. Then we have,

$$\begin{aligned}
& \left| R(n) - \left(n \sum_{j=1}^{M} \mu^j - \sum_{i=1}^{K} \mu^i E[T^i(n)] \right) \right| \\
= & \left| E\left[\sum_{j=1}^{M} \sum_{i=1}^{K} \sum_{x \in S^i} r_x^i \sum_{t=1}^{T^{i,j}(n)} I(X^{i,j}(t) = x) \right] - \sum_{j=1}^{M} \sum_{i=1}^{K} \sum_{x \in S^i} r_x^i \pi_x^i E[T^{i,j}(n)] \right| \\
= & \left| \sum_{j=1}^{M} \sum_{i=1}^{K} \sum_{x \in S^i} r_x^i (E[N^j(x, T^{i,j}(n))] - \pi_x^i E[T^{i,j}(n)]) \right| \\
\leq & \sum_{j=1}^{M} \sum_{i=1}^{K} \sum_{x \in S^i} r_x^i C_{P^i} = C_{\mathbf{S},\mathbf{P},\mathbf{r}} \qquad (9)
\end{aligned}$$

where

$$N^j(x, T^{i,j}(n)) = \sum_{t=1}^{T^{i,j}(n)} I(X^{i,j}(t) = x),$$

and (9) follows from Lemma 2 using the fact that $T^{i,j}(n)$ is a stopping time with respect to the σ-field generated by the arms played up to time n.

APPENDIX B

Lemma 7: Assume Condition 1 holds and all arms are rested. Let $g_{t,s}^i = \bar{r}^i(s) + c_{t,s}$, $c_{t,s} = \sqrt{L \ln t / s}$. Under UCB-M with constant $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2 / \epsilon_{\min}$, for any suboptimal arm i and optimal arm j

we have

$$E\left[\sum_{t=1}^{n}\sum_{w=1}^{t-1}\sum_{w_i=l}^{t-1} I(g_{t,w}^{j} \le g_{t,w_i}^{i})\right] \le \frac{|S^i|+|S^j|}{\pi_{\min}}\beta, \tag{10}$$

where $l = \left\lceil \frac{4L\ln n}{(\mu^M-\mu^i)^2} \right\rceil$ and $\beta = \sum_{t=1}^{\infty} t^{-2}$.

Proof: First, we show that for any suboptimal arm i and optimal arm j, we have that $g_{t,w}^{j} \le g_{t,w_i}^{i}$ implies at least one of the following holds:

$$\bar{r}^j(w) \le \mu^j - c_{t,w} \tag{11}$$

$$\bar{r}^i(w_i) \ge \mu^i + c_{t,w_i} \tag{12}$$

$$\mu^j < \mu^i + 2c_{t,w_i}. \tag{13}$$

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

$$g_{t,w}^{j} = \bar{r}^j(w) + c_{t,w} > \mu^j \ge \mu^i + 2c_{t,w_i} > \bar{r}^i(w_i) + c_{t,w_i} = g_{t,w_i}^{i},$$

which contradicts $g_{t,w}^{j} \le g_{t,w_i}^{i}$.

If we choose $w_i \ge 4L\ln n/(\mu^M-\mu^i)^2$, then

$$2c_{t,w_i} = 2\sqrt{\frac{L\ln t}{w_i}} \le 2\sqrt{\frac{L\ln t(\mu^M-\mu^i)^2}{4L\ln n}} \le \mu^j - \mu^i \text{ for } t \le n,$$

which means (13) is false, and therefore at least one of (11) and (12) is true with this choice of w_i. Let $l = \left\lceil \frac{4L\ln n}{(\mu^M-\mu^i)^2} \right\rceil$. Then we have,

$$\begin{aligned} E\left[\sum_{t=1}^{n}\sum_{w=1}^{t-1}\sum_{w_i=l}^{t-1} I(g_{t,w}^{j} \le g_{t,w_i}^{i})\right] &\le \sum_{t=1}^{n}\sum_{w=1}^{t-1}\sum_{w_i=\left\lceil \frac{4L\ln n}{(\mu^M-\mu^i)^2}\right\rceil}^{t-1} \left(P(\bar{r}^j(w) \le \mu^j - c_{t,w}) + P(\bar{r}^i(w_i) \ge \mu^i + c_{t,w_i})\right) \\ &\le \sum_{t=1}^{\infty}\sum_{w=1}^{t-1}\sum_{w_i=\left\lceil \frac{4L\ln n}{(\mu^M-\mu^i)^2}\right\rceil}^{t-1} \left(P(\bar{r}^j(w) \le \mu^j - c_{t,w}) + P(\bar{r}^i(w_i) \ge \mu^i + c_{t,w_i})\right) . \end{aligned}$$

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 up to and including the t-th play of arm i.

$$\begin{aligned} & P(\bar{r}^i(w_i) \geq \mu^i + c_{t,w_i}) \\ = \; & P\left(\sum_{y\in S^i} r_y^i n_y^i(w_i) \geq w_i \sum_{y\in S^i} r_y^i \pi_y^i + w_i c_{t,w_i}\right) \\ = \; & P\left(\sum_{y\in S^i} (r_y^i n_y^i(w_i) - w_i r_y^i \pi_y^i) \geq w_i c_{t,w_i}\right) \\ = \; & P\left(\sum_{y\in S^i} (-r_y^i n_y^i(w_i) + w_i r_y^i \pi_y^i) \leq -w_i c_{t,w_i}\right) \; . \end{aligned} \tag{14}$$

Consider a sample path ω and the events

$$\begin{aligned} A &= \left\{\omega : \sum_{y\in S^i} (-r_y^i n_y^i(w_i)(\omega) + w_i r_y^i \pi_y^i) \leq -w_i c_{t,w_i}\right\} \; , \\ B &= \bigcup_{y\in S^i} \left\{\omega : -r_y^i n_y^i(w_i)(\omega) + w_i r_y^i \pi_y^i \leq -\frac{w_i c_{t,w_i}}{|S^i|}\right\} \; . \end{aligned}$$

If $\omega \notin B$, then

$$\begin{aligned} & -r_y^i n_y^i(w_i)(\omega) + w_i r_y^i \pi_y^i > -\frac{w_i c_{t,w_i}}{|S^i|}, \; \forall y \in S^i \\ \Rightarrow \quad & \sum_{y\in S^i} (-r_y^i n_y^i(w_i)(\omega) + w_i r_y^i \pi_y^i) > -w_i c_{t,w_i} \; . \end{aligned}$$

Thus $\omega \notin A$, therefore $P(A) \leq P(B)$. Then continuing from (14):

$$\begin{aligned} & P(\bar{r}^i(w_i) \geq \mu^i + c_{t,w_i}) \\ \leq \; & \sum_{y\in S^i} P\left(-r_y^i n_y^i(w_i) + w_i r_y^i \pi_y^i \leq -\frac{w_i c_{t,w_i}}{|S^i|}\right) \\ = \; & \sum_{y\in S^i} P\left(r_y^i n_y^i(w_i) - w_i r_y^i \pi_y^i \geq \frac{w_i c_{t,w_i}}{|S^i|}\right) \\ = \; & P\left(n_y^i(w_i) - w_i \pi_y^i \geq \frac{w_i c_{t,w_i}}{|S^i| r_y^i}\right) \end{aligned}$$

$$
\begin{aligned}
&= P\left(\frac{\sum_{t=1}^{w_i} I(X_t^i = y) - w_i \pi_y^i}{\hat{\pi}_y^i w_i} \geq \frac{c_{t,w_i}}{|S^i| r_y^i \hat{\pi}_y^i}\right) \\
&\leq \sum_{y \in S^i} N_{q^i} t^{-\frac{L\epsilon^i}{28(|S^i| r_y^i \hat{\pi}_y^i)^2}} \qquad (15) \\
&\leq \frac{|S^i|}{\pi_{\min}} t^{-\frac{L\epsilon_{\min}}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}}, \qquad (16)
\end{aligned}
$$

where (15) follows from Lemma 1 by letting

$$
\gamma = \frac{c_{t,w_i}}{|S^i| r_y^i \hat{\pi}_y^i}, \quad f(X_t^i) = \frac{I(X_t^i = y) - \pi_y^i}{\hat{\pi}_y^i},
$$

and recalling $\hat{\pi}_y^i = \max\{\pi_y^i, 1 - \pi_y^i\}$ (note $\hat{P}^i$ is irreducible).

Similarly, we have

$$
\begin{aligned}
& P\left(\bar{r}^j(w) \leq \mu^j - c_{t,w}\right) \\
&= P\left(\sum_{y \in S^j} r_y^j (n_y^j(w) - w\pi_y^j) \leq -w c_{t,w}\right) \\
&\leq \sum_{y \in S^j} P\left(r_y^j n_y^j(w) - w r_y^j \pi_y^j \leq -\frac{w c_{t,w}}{|S^j|}\right) \\
&= \sum_{y \in S^j} P\left(r_y^j (w - \sum_{x \neq y} n_x^j(w)) - w r_y^j (1 - \sum_{x \neq y} \pi_x^j) \leq -\frac{w c_{t,w}}{|S^j|}\right) \\
&= \sum_{y \in S^j} P\left(r_y^j \sum_{x \neq y} n_x^j(w) - w r_y^j \sum_{x \neq y} \pi_x^j \geq \frac{w c_{t,w}}{|S^j|}\right) \\
&\leq \sum_{y \in S^j} N_{q^j} t^{-\frac{L\epsilon^j}{28(|S^j| r_y^j \hat{\pi}_y^j)^2}} \qquad (17) \\
&\leq \frac{|S^j|}{\pi_{\min}} t^{-\frac{L\epsilon_{\min}}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}} \qquad (18)
\end{aligned}
$$

where (17) again follows from Lemma 1. The result then follows from combining (16) and (18):

$$
\begin{aligned}
E\left[\sum_{t=1}^{n} \sum_{w=1}^{t-1} \sum_{w_i=l}^{t-1} I(g_{t,w}^j \leq g_{t,w_i}^i)\right] &\leq \frac{|S^i| + |S^j|}{\pi_{\min}} \sum_{t=1}^{\infty} \sum_{w=1}^{t-1} \sum_{w_i=1}^{t-1} t^{-\frac{L\epsilon_{\min}}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}} \\
&= \frac{|S^i| + |S^j|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-\frac{L\epsilon_{\min} - 56 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}} \\
&\leq \frac{|S^i| + |S^j|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}. \qquad (19)
\end{aligned}
$$

■

APPENDIX C
PROOF OF LEMMA 5

Let l be any positive integer and consider a suboptimal arm i. Then,

$$T^i(n) = M + \sum_{t=K+1}^{n} I(i \in A(t)) \leq M - 1 + l + \sum_{t=K+1}^{n} I(i \in A(t), T^i(t-1) \geq l) \ . \tag{20}$$

Consider

$$E = \bigcup_{j=1}^{M} \left\{ g^j_{t,T^j(t)} \leq g^i_{t,T^i(t)} \right\},$$

and

$$E^C = \bigcap_{j=1}^{M} \left\{ g^j_{t,T^j(t)} > g^i_{t,T^i(t)} \right\}.$$

If $w \in E^C$ then $i \notin A(t)$. Therefore $\{i \in A(t)\} \subset E$ and

$$\begin{aligned} I(i \in A(t),\ T^i(t-1) \geq l) &\leq I(\omega \in E,\ T^i(t-1) \geq l) \\ &\leq \sum_{j=1}^{M} I(g^j_{t,T^j(t)} \leq g^i_{t,T^i(t)},\ T^i(t-1) \geq l). \end{aligned}$$

Therefore continuing from (20),

$$\begin{aligned} T^i(n) &\leq M - 1 + l + \sum_{j=1}^{M} \sum_{t=K+1}^{n} I(g^j_{t,T^j(t)} \leq g^i_{t,T^i(t)}, T^i(t-1) \geq l) \\ &\leq M - 1 + l + \sum_{j=1}^{M} \sum_{t=K+1}^{n} I\left(\min_{1 \leq w \leq t} g^j_{t,w} \leq \max_{l \leq w_i \leq t} g^i_{t,w_i} \right) \\ &\leq M - 1 + l + \sum_{j=1}^{M} \sum_{t=K+1}^{n} \sum_{w=1}^{t-1} \sum_{w_i=l}^{t-1} I(g^j_{t,w} \leq g^i_{t,w_i}) \\ &\leq M - 1 + l + \sum_{j=1}^{M} \sum_{t=1}^{n} \sum_{w=1}^{t-1} \sum_{w_i=l}^{t-1} I(g^j_{t,w} \leq g^i_{t,w_i}). \end{aligned} \tag{21}$$

Using Lemma 7 with $l = \left\lceil \frac{4L\ln n}{(\mu^M - \mu^i)^2} \right\rceil$, we have for any suboptimal arm

$$E[T^i(n)] \leq M + \frac{4L\ln n}{(\mu^M - \mu^i)^2} + \sum_{j=1}^{M} \frac{(|S^i| + |S^j|)\beta}{\pi_{\min}}. \tag{22}$$

APPENDIX D

Lemma 8: Assume Condition 1 holds and all arms are restless. Let $g^i_{t,w} = \bar{r}^i(w) + c_{t,w}$, $c_{t,w} = \sqrt{L\ln t/w}$. Under RCA-M with constant $L \geq 112 S^2_{\max} r^2_{\max} \hat{\pi}^2_{\max}/\epsilon_{\min}$, for any suboptimal arm i and optimal arm j we have

$$E\left[\sum_{t=1}^{t_2(b)} \sum_{w=1}^{t-1} \sum_{w_i=l}^{t-1} I(g^j_{t,w} \leq g^i_{t,w_i})\right] \leq \frac{|S^i| + |S^j|}{\pi_{\min}}\beta, \tag{23}$$

where $l = \left\lceil \frac{4L\ln n}{(\mu^M - \mu^i)^2} \right\rceil$ and, $\beta = \sum_{t=1}^{\infty} t^{-2}$.

Proof: Note that all the quantities in computing the indices in (23) comes from the intervals $X^i_2(1), X^i_2(2), \cdots \forall i \in \{1, \cdots, K\}$. Since these intervals begin with state γ^i and end with a return to γ^i (but excluding the return visit to γ^i), by the strong Markov property the process at these stopping times have 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 an transition matrix identical to the original arm. Therefore we can proceed in exactly the same way as the proof of Lemma 7. If we choose $s_i \geq 4L\ln(n)/(\mu^M - \mu^i)^2$, then for $t \leq t_2(b) = n' \leq n$, and for any suboptimal arm i and optimal arm j,

$$2c_{t,s_i} = 2\sqrt{\frac{L\ln(t)}{s_i}} \leq 2\sqrt{\frac{L\ln(t)(\mu^M - \mu^i)^2}{4L\ln(n)}} \leq \mu^j - \mu^i.$$

The result follows from letting $l = \left\lceil \frac{4L\ln n}{(\mu^M - \mu^i)^2} \right\rceil$ and using Lemma 7. ■

APPENDIX E
PROOF OF LEMMA 6

Let $c_{t,w} = \sqrt{L\ln t/w}$, and let l be any positive integer. Then,

$$B^i(b) = 1 + \sum_{m=K+1}^{b} I(\alpha(m) = i) \leq l + \sum_{m=K+1}^{b} I(\alpha(m) = i, B^i(m-1) \geq l) \tag{24}$$

Consider any sample path ω and the following sets

$$E = \bigcup_{j=1}^{M} \left\{ \omega : g^j_{t_2(m-1),T_2^j(t_2(m-1))}(\omega) \leq g^i_{t_2(m-1),T_2^i(t_2(m-1))}(\omega) \right\},$$

and

$$E^C = \bigcap_{j=1}^{M} \left\{ \omega : g^j_{t_2(m-1),T_2^j(t_2(m-1))}(\omega) > g^i_{t_2(m-1),T_2^i(t_2(m-1))}(\omega) \right\}.$$

If $\omega \in E^C$ then $\alpha(m) \neq i$. Therefore $\{\omega : \alpha(m)(\omega) = i\} \subset E$ and

$$\begin{aligned} & I(\alpha(m) = i, B^i(m-1) \geq l) \leq I(\omega \in E, B^i(m-1) \geq l) \\ \leq & \sum_{j=1}^{M} I(g^j_{t_2(m-1),T_2^j(t_2(m-1))} \leq g^i_{t_2(m-1),T_2^i(t_2(m-1))}, B^i(m-1) \geq l) \ . \end{aligned}$$

Therefore continuing from (24),

$$\begin{aligned} B^i(b) &\leq l + \sum_{j=1}^{M} \sum_{m=K+1}^{b} I(g^j_{t_2(m-1),T_2^j(t_2(m-1))} \leq g^i_{t_2(m-1),T_2^i(t_2(m-1))}, B^i(m-1) \geq l) \\ &\leq l + \sum_{j=1}^{M} \sum_{m=K+1}^{b} I\left(\min_{1 \leq w \leq t_2(m-1)} g^j_{t_2(m-1),w} \leq \max_{t_2(l) \leq w_i \leq t_2(m-1)} g^i_{t_2(m-1),w_i} \right) \\ &\leq l + \sum_{j=1}^{M} \sum_{m=K+1}^{b} \sum_{w=1}^{t_2(m-1)} \sum_{w_i=t_2(l)}^{t_2(m-1)} I(g^j_{t_2(m),w} \leq g^i_{t_2(m),w_i}) \qquad (25) \\ &\leq l + M \sum_{j=1}^{M} \sum_{t=1}^{t_2(b)} \sum_{w=1}^{t-1} \sum_{w_i=l}^{t-1} I(g^j_{t,w} \leq g^i_{t,w_i}) \ , \qquad (26) \end{aligned}$$

where as given in (8), $g^i_{t,w} = \bar{r}^i(w) + c_{t,w}$, and we have assumed that the index value of an arm remains the same between two updates. The inequality in (26) follows from the facts that the second outer sum in (26) is over time while the second outer sum in (25) is over blocks, each block lasts at least two time slots and at most M blocks can be completed in each time step. From this point on we use Lemma 8 to get

$$E[B^i(b(n))|b(n) = b] \leq \left\lceil \frac{4L \ln t_2(b)}{(\mu^M - \mu^i)^2} \right\rceil + M \sum_{j=1}^{M} \frac{(|S^i| + |S^j|)\beta}{\pi_{\min}},$$

for all suboptimal arms. Therefore,

$$E[B^i(b(n))] \leq \frac{4L \ln n}{(\mu^M - \mu^i)^2} + 1 + M \sum_{j=1}^{M} C_{i,j}\beta, \qquad (27)$$

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

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:

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

Thus,

$$\sum_{i>M} (\mu^1 - \mu^i) E[T^i(n)] \tag{28}$$

$$\leq \ 4L \sum_{i>M} \frac{(\mu^1 - \mu^i) D_i \ln n}{(\mu^M - \mu^i)^2} + \sum_{i>M} (\mu^1 - \mu^i) D_i \left(1 + M \sum_{j=1}^{M} C_{i,j}\right). \tag{29}$$

APPENDIX F
PROOF OF THEOREM 2

Assume that the states which determine the regenerative sample paths are given *a priori* by $\gamma = [\gamma^1, \cdots, \gamma^K]$. This is to simplify the analysis by skipping the initialization stage of the algorithm and we will show that this choice does not affect the regret bound. We denote the expectations with respect to RCA-M given γ as E_γ. First we rewrite the regret in the following form:

$$\begin{aligned} R_\gamma(n) &= \sum_{j=1}^{M} \mu^j E_\gamma[T(n)] - E_\gamma\left[\sum_{t=1}^{T(n)} \sum_{\alpha(t)\in A(t)} r^{\alpha(t)}_{x_{\alpha(t)}}\right] + \sum_{j=1}^{M} \mu^j E_\gamma[n - T(n)] - E_\gamma\left[\sum_{t=T(n)+1}^{n} \sum_{\alpha(t)\in A(t)} r^{\alpha(t)}_{x_{\alpha(t)}}\right] \\ &= \left\{\sum_{j=1}^{M} \mu^j E_\gamma[T(n)] - \sum_{i=1}^{K} \mu^i E_\gamma\left[T^i(n)\right]\right\} - Z_\gamma(n) \qquad (30) \\ &+ \sum_{j=1}^{M} \mu^j E_\gamma[n - T(n)] - E_\gamma\left[\sum_{t=T(n)+1}^{n} \sum_{\alpha(t)\in A(t)} r^{\alpha(t)}_{x_{\alpha(t)}}\right], \qquad (31) \end{aligned}$$

where for notational convenience, we have used

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

We have

$$\begin{aligned}\sum_{j=1}^{M}\mu^j E_\gamma[T(n)] - \sum_{i=1}^{K}\mu^i E_\gamma\left[T^i(n)\right] &= \sum_{j=1}^{M}\sum_{i=1}^{K}\mu^j E_\gamma[T^{i,j}(n)] - \sum_{j=1}^{M}\sum_{i=1}^{K}\mu^i E_\gamma[T^{i,j}(n)] \\ &= \sum_{j=1}^{M}\sum_{i>M}(\mu^j-\mu^i)E_\gamma[T^{i,j}(n)] \\ &\leq \sum_{i>M}(\mu^1-\mu^i)E_\gamma[T^i(n)] \qquad (32)\end{aligned}$$

Since we can bound (32), i.e. the difference in the brackets in (30) logarithmically using Lemma 6, it remains to bound $Z_\gamma(n)$ and the difference in (31). We have

$$\begin{aligned}Z_\gamma(n) \geq\ & \sum_{i=1}^{M}\sum_{y\in S^i} r_y^i E_\gamma\left[\sum_{b=1}^{B^i(b(n))}\sum_{X_t^i\in X^i(b)} I(X_t^i=y)\right] \\ & + \sum_{i>M}\sum_{y\in S^i} r_y^i E_\gamma\left[\sum_{b=1}^{B^i(b(n))}\sum_{X_t^i\in X_2^i(b)} I(X_t^i=y)\right] \qquad (33) \\ & - \sum_{i=1}^{M}\mu^i E_\gamma\left[T^i(n)\right] \\ & - \sum_{i>M}\mu^i\left(\frac{1}{\pi^i_{\gamma^i}}+\Omega^i_{\max}+1\right)E_\gamma\left[B^i(b(n))\right]\ ,\end{aligned}$$

where the inequality comes from counting only the rewards obtained during the SB2's for all suboptimal arms and the last part of the proof of Lemma 6. Applying Lemma 3 to (33) we get

$$E_\gamma\left[\sum_{b=1}^{B^i(b(n))}\sum_{X_t^i\in X_2^i(b)} I(X_t^i=y)\right] = \frac{\pi_y^i}{\pi^i_{\gamma^i}}E_\gamma\left[B^i(b(n))\right]\ .$$

Rearranging terms we get

$$Z_\gamma(n) \geq R^*(n) - \sum_{i>M}\mu^i(\Omega^i_{\max}+1)E_\gamma\left[B^i(b(n))\right] \qquad (34)$$

where

$$R^*(n) = \sum_{i=1}^{M}\sum_{y\in S^i} r_y^i E_\gamma\left[\sum_{b=1}^{B^i(b(n))}\sum_{X_t^i\in X^i(b)} I(X_t^i=y)\right] - \sum_{i=1}^{M}\sum_{y\in S^i} r_y^i\pi_y^i E_\gamma\left[T^i(n)\right].$$

Consider now $R^*(n)$. Since all suboptimal arms are played at most logarithmically, the total number of time slots in which an optimal arm is not played is at most logarithmic. It follows that the number of discontinuities between plays of any single optimal arm is at most logarithmic. For any optimal arm

$i \in \{1, \cdots, M\}$ we combine *consecutive* blocks in which arm i is played into a single *combined* block, and denote by $\bar{X}^i(j)$ the j-th combined block of arm i. Let $\bar{b}^i$ denote the total number of combined blocks for arm i up to block b. Each $\bar{X}^i$ thus consists of two sub-blocks: $\bar{X}^i_1$ that contains the states visited from the beginning of $\bar{X}^i$ (empty if the first state is γ^i) to the state right before hitting γ^i, and sub-block $\bar{X}^i_2$ that contains the rest of $\bar{X}^i$ (a random number of regenerative cycles).

Since a combined block $\bar{X}^i$ necessarily starts after certain discontinuity in playing the i-th best arm, $\bar{b}^i(n)$ is less than or equal to the total number of discontinuities of play of the i-th best arm up to time n. At the same time, the total number of discontinuities of play of the i-th best arm up to time n is less than or equal to the total number of blocks in which suboptimal arms are played up to time n. Thus

$$E_\gamma[\bar{b}^i(n)] \leq \sum_{k>M} E_\gamma[B^k(b(n))]. \tag{35}$$

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

$$\begin{aligned}
R^*(n) &= \sum_{i=1}^{M} \sum_{y \in S^i} r^i_y E_\gamma \left[\sum_{b=1}^{\bar{b}^i(n)} \sum_{X^i_t \in \bar{X}^i_2(b)} I(X^i_t = y) \right] && (36)\\
&\quad - \sum_{i=1}^{M} \sum_{y \in S^i} r^i_y \pi^i_y E_\gamma \left[\sum_{b=1}^{\bar{b}^i(n)} |\bar{X}^i_2(b)| \right] && (37)\\
&\quad + \sum_{i=1}^{M} \sum_{y \in S^i} r^i_y E_\gamma \left[\sum_{b=1}^{\bar{b}^i(n)} \sum_{X^i_t \in \bar{X}^i_1(b)} I(X^i_t = y) \right] && (38)\\
&\quad - \sum_{i=1}^{M} \sum_{y \in S^i} r^i_y \pi^i_y E_\gamma \left[\sum_{b=1}^{\bar{b}^i(n)} |\bar{X}^i_1(b)| \right] && (39)\\
&> 0 - \sum_{i=1}^{M} \mu^i \Omega^i_{\max} \sum_{k>M} E_\gamma[B^k(b(n))] && (40)
\end{aligned}$$

where the last inequality is obtained by noting the difference between (36) and (37) is zero by Lemma 3, using positivity of rewards to lower bound (38) by 0, and (35) to upper bound (39). Combining this

with (27) and (34) we can obtain a logarithmic upper bound on $-Z_\gamma(n)$ by the following steps:

$$\begin{aligned} -Z_\gamma(n) &\leq -R^*(n) + \sum_{i>M} \mu^i(\Omega^i_{\max}+1)E_\gamma\left[B^i(b(n))\right] \\ &\leq \sum_{i=1}^{M} \mu^i \Omega^i_{\max} \sum_{k>M} \left(\frac{4L\ln n}{(\mu^M-\mu^k)^2} + 1 + M\sum_{j=1}^{M} C_{k,j}\beta \right) \\ &+ \sum_{i>M} \mu^i(\Omega^i_{\max}+1) \left(\frac{4L\ln n}{(\mu^M-\mu^i)^2} + 1 + M\sum_{j=1}^{M} C_{k,i}\beta \right) \end{aligned} \tag{41}$$

We also have,

$$\begin{aligned} \sum_{j=1}^{M} \mu^j E_\gamma[n-T(n)] - E_\gamma[\sum_{t=T(n)+1}^{n} \sum_{\alpha(t)\in A(t)} r^{\alpha(t)}_{x_{\alpha(t)}}] &\leq \sum_{j=1}^{M} \mu^j E_\gamma[n-T(n)] \\ &= \sum_{j=1}^{M} \mu^j \left(\frac{1}{\pi_{\min}} + \max_{i\in\{1,\ldots,K\}} \Omega^i_{\max} + 1 \right). \end{aligned} \tag{42}$$

Finally, combining the above results as well as Lemma 6 we get

$$\begin{aligned} R_\gamma(n) &= \left\{ \sum_{j=1}^{M} \mu^j E_\gamma[T(n)] - \sum_{i=1}^{K} \mu^i E_\gamma\left[T^i(n)\right] \right\} - Z_\gamma(n) \\ &\quad + \sum_{j=1}^{M} \mu^j E_\gamma[n-T(n)] - E_\gamma\left[\sum_{t=T(n)+1}^{n} \sum_{\alpha(t)\in A(t)} r^{\alpha(t)}_{x_{\alpha(t)}} \right] \\ &\leq \sum_{i>M} (\mu^1-\mu^i) E_\gamma[T^i(n)] \\ &\quad + \sum_{i=1}^{M} \mu^i \Omega^i_{\max} \sum_{k>M} \left(\frac{4L\ln n}{(\mu^M-\mu^k)^2} + 1 + M\sum_{j=1}^{M} C_{k,j}\beta \right) \\ &\quad + \sum_{i>M} \mu^i(\Omega^i_{\max}+1) \left(\frac{4L\ln n}{(\mu^M-\mu^i)^2} + 1 + M\sum_{j=1}^{M} C_{k,i}\beta \right) \\ &\quad + \sum_{j=1}^{M} \mu^j \left(\frac{1}{\pi_{\min}} + \max_{i\in\{1,\ldots,K\}} \Omega^i_{\max} + 1 \right) \\ &= 4L\ln n \sum_{i>M} \frac{1}{(\mu^M-\mu^i)^2} \left((\mu^1-\mu^i)D_i + E_i \right) \\ &\quad + \sum_{i>M} \left((\mu^1-\mu^i)D_i + E_i \right) \left(1 + M\sum_{j=1}^{M} C_{i,j} \right) + F \end{aligned}$$

Therefore we have obtained the stated logarithmic bound for (30). Note that this bound does not depend on γ, and therefore is also an upper bound for $R(n)$, completing the proof.