Online Learning of Rested and Restless Bandits

Cem Tekin, *Student Member, IEEE*, and Mingyan Liu, *Senior Member, IEEE*

***Abstract*—In this paper, we study the online learning problem involving *rested* and *restless* bandits, in both a centralized and a decentralized setting. In a centralized setting, the system consists of a single player/user and a set of K finite-state discrete-time Markov chains (*arms*) with unknown state spaces (rewards) and statistics. The objective of the player is to decide in each step which M of the K arms to play over a sequence of trials so as to maximize its long-term reward. In a decentralized setting, multiple uncoordinated players each makes its own decision on which arm to play in a step, and if two or more players select the same arm simultaneously, a collision results and none of the players selecting that arm gets a reward. The objective of each player again is to maximize its long-term reward. We first show that logarithmic regret algorithms exist both for the centralized rested and restless bandit problems. For the decentralized setting, we propose an algorithm with logarithmic regret with respect to the optimal centralized arm allocation. Numerical results and extensive discussion are also provided to highlight insights obtained from this study.**

***Index Terms*—Exploration–exploitation tradeoff, multiarmed bandits, online learning, opportunistic spectrum access (OSA), regret, restless bandits.**

I. Introduction

In this paper, we study the online learning problem involving *rested* and *restless* bandits in both a centralized and a decentralized setting. In the centralized setting, 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 player. The objective of the player 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. 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). In the decentralized setting, multiple uncoordinated players each makes its own decision on which arm to play in a step, and if two or more players select the same arm simultaneously, a collision results and none of the players selecting that arm gets a reward. The objective of each player again is to maximize its long-term reward. We consider two cases, one with *rested* arms where the state of a Markov chain stays frozen unless it is 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 aforementioned 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 this 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 decentralized, multiplayer version of the aforementioned problem is also easily understood: here, multiple users compete for the set of K channels, and simultaneous access to the same channel results in a collision and reduced rewards. In this case, not only do the users have to find channels with good conditions, but must also try to avoid collision in order to maximize their rewards.

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*. In the centralized case, 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. In the decentralized case, the difference is calculated with respect to the centralized policy with the aforementioned properties. 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 IX.

Manuscript received February 17, 2011; revised November 03, 2011; accepted February 27, 2012. Date of publication May 16, 2012; date of current version July 10, 2012. This paper was presented in part at the 2010 Allerton Conference on Communication, Control, and Computing and in part at the 2011 IEEE International Conference on Computer Communications.

The authors are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48105 USA (e-mail: cmtkn@eecs.umich.edu; mingyan@eecs.umich.edu).

Communicated by N. Cesa-Bianchi, Associate Editor for Pattern Recognition, Statistical Learning, and Inference.

Digital Object Identifier 10.1109/TIT.2012.2198613

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 (centralized) bandit problems, originally proposed by Robbins [2], the rewards are assumed to be independently drawn from a fixed (but unknown) distribution. It is worth noting that with this i.i.d. 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.

In this paper, we first study the centralized rested bandit problem with Markovian rewards. Specifically, we show that a player using the UCB1 algorithm [3] achieves logarithmic regret for rested bandits. 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 with a single play or multiple plays. 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. We then consider the decentralized multiplayer restless bandit problem, where there are M uncoordinated players, each selects a single arm at every time step. We prove that a multiplayer extension to RCA leads to polylogarithmic regret with respect to the optimal centralized allocation for this problem.

The remainder of this paper is organized as follows. Related work is discussed in Section II. In Section III, we present the problem formulation. In Section IV, we introduce and analyze the rested bandit problem with a single play and multiple plays. In Sections V and VI, we analyze the restless bandit problem with a single play and multiple plays, respectively, by utilizing regenerative cycles. We consider the decentralized multiplayer restless bandit problem and analyze its regret in Section VII. In Section VIII, we numerically examine the performance of our algorithms in the case of an OSA problem with the Gilbert–Elliot channel model. In Section IX, we discuss possible improvements and compare our algorithm to existing literature. Section X concludes this paper.

II. Related Work

In the following, we briefly summarize the most relevant results in the literature. Lai and Robbins in [4] 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 [5] to the case of playing more than one arm at a time. Using a similar approach, Anantharam *et al.* in [6] develop 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 [7] considers sample mean based index policies for the i.i.d. model that achieve $O(\log n)$ regret, where n is the total number of plays. Auer *et al.* in [3] also propose sample mean based index policies for i.i.d. rewards with bounded support; these are derived from [7], but are simpler than those in [7] and are not restricted to a specific family of distributions. These policies achieve logarithmic regret uniformly over time rather than asymptotically in time, but in general have bigger constant than that in [4]. In [8], an index policy, KL-UCB, which is uniformly better than UCB [3], and its variants is proposed. Moreover, it is shown to be asymptotically optimal for Bernoulli rewards. An extension of the multiarmed bandit problem to linear optimization is considered in [9]. In [10], we show that the index policy in [3] is order optimal for Markovian rewards drawn from rested arms but not restricted to single-parameter families, under some assumptions on the transition probabilities. Later, we prove a logarithmic weak regret bound for the restless bandit problem in [11]. Parallel to the work presented here, in [12], an algorithm is constructed that achieves logarithmic regret for the restless bandit problem. The mechanism behind this algorithm, however, is quite different from ours; this difference is discussed in more detail in Section IX.

Work in decentralized multiplayer setting includes [13]–[15], which all consider the i.i.d. reward case, and where players selecting the same arms experience collision according to a certain collision model. Specifically, in [13] and [14], a logarithmic lower bound on the regret is derived, and algorithms with logarithmic regret are proposed. This is done through a time-division fair sharing scheme in [13], while in [14], players randomize to settle to orthogonal arms. In [15], the authors prove a logarithmic regret bound for a combinatorial bandit problem.

It is also worth mentioning 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 [16] and the optimal policy is proved to be an index policy that at each time plays the arm with highest Gittins' index. Whittle introduced the restless version of the bandit problem in [17]. 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 [18] for an OSA problem formulated as a restless bandit problem with each channel modeled as a two-state Markov chain (the Gilbert–Elliot model).

III. Problem Formulation and Preliminaries

Consider K arms (or channels) indexed by the set $\mathcal{K} = \{1, 2, \dots, K\}$. The ith arm is modeled as a discrete-time, irreducible, and aperiodic Markov chain with a 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 $P^i = \{p_{xy}^i,\ x, y \in S^i\}$ denote the transition probability matrix of the ith arm, and $\pi^i = \{\pi_x^i,\ x \in S^i\}$ 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 Section I, 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 apply 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

$$(p^i)'_{xy} = (\pi_y^i p_{yx}^i)/\pi_x^i,\ \forall x, y \in S^i$$

and $\hat{P}^i = (P^i)'P$ denotes the *multiplicative symmetrization* of P^i. We will assume that the P^i's 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 is 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 queuing 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

$$\mu^i = \sum_{x \in S^i} r_x^i \pi_x^i. \tag{1}$$

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 $t,\ t = 1, 2, \ldots$. We will also frequently refer to the time interval $(t-1, t]$ as time slot t. In the centralized model, 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 $\mu^1 \geq \mu^2 \geq \cdots \geq \mu^M > \mu^{M+1} \geq \cdots \geq \mu^K$. 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 Mth 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 mean different data rates. Possible relaxation of this condition is given in Section IX.

We will refer to the set of arms $\{1, 2, \ldots, M\}$ 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, $\alpha(t) \in A^\alpha(t)$ be an arm selected by policy α at $t,\ t = 1, 2, \ldots$, and $x_\alpha(t)$ be 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. Let $T^{\alpha,i}(t)$ be the number of times arm i is played by policy α at the end of time t, and $\bar{r}^i(T^{\alpha,i}(t))$ be the sample mean of the rewards observed from the first $T^{\alpha,i}(t)$ plays of arm i. When the policy used is clear from the context, we will suppress the superscript α from the aforementioned expressions.

The following result is due to Lezaud [19] that bounds the probability of a large deviation from the stationary distribution.

Lemma 1 [Theorem 3.3 From [19]]: 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 following notations are frequently used throughout this paper: $\beta = \sum_{t=1}^{\infty} 1/t^2$, $\pi_{\min}^i = \min_{x \in S^i} \pi_x^i$, $\pi_{\min} = \min_{i \in \mathcal{K}} \pi_{\min}^i$, $r_{\max} = \max_{x \in S^i,\, i \in \mathcal{K}} r_x^i$, $S_{\max} = \max_{i \in \mathcal{K}} |S^i|$, $\hat{\pi}_{\max} = \max_{x \in S^i,\, i \in \mathcal{K}} \{\pi_x^i, 1 - \pi_x^i\}$, $\epsilon_{\min} = \min_{i \in \mathcal{K}} \epsilon^i$, 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 $\Omega_{\max}^i = \max_{x,y \in S^i} \Omega_{x,y}^i$, where $\Omega_{x,y}^i$ is the mean hitting time of state y given the initial state x for arm i for P^i.

The following is a condition we will need on arms for most of the results in this paper.

[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 is in the passive mode. More is discussed in Section IX.

```
The Upper Confidence Bound (UCB) Algorithm:
1: Initialize: Play each arm once in the first $K$ slots
2: while $t \ge K$ do
3:     $\bar{r}^i(T^i(t)) = \frac{r^i(1)+r^i(2)+\ldots+r^i(T^i(t))}{T^i(t)}$, $\forall i$
4:     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:     $t := t+1$
6:     play the arm with the highest index (ties broken randomly), update $r^i(t)$ and $T^i(t), \forall i \in \mathcal{K}$.
7: end while
```

Fig. 1. Pseudocode for the UCB algorithm.

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

In the next few sections, we present algorithms for the rested and restless bandit problems with a single play and multiple plays, respectively, and analyze their regret.

IV. Analysis of the Rested Bandit Problem

In this section, we show that there exists an algorithm that achieves logarithmic regret uniformly over time for the rested bandit problem. We will start with the *single-play* scenario, where the player selects a single arm at each time step, thus $M = 1$. The algorithm we consider is called *the upper confidence bound* (UCB), which is a slight modification of UCB1 from [3] with an unspecified exploration constant L instead of fixing it at 2. The idea of modifying the exploration constant is also used in [20] under a very different setting where the idea is to exploit variance estimation in a multiarmed bandit problem. Throughout our discussion, we will consider a horizon of n time slots.

As shown in Fig. 1, UCB selects the arm with the highest index at each time step and updates the indices according to the rewards observed. The index given on line 4 of Fig. 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.

To upper bound the regret of the aforementioned 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.

Lemma 2: Assume that all arms are finite-state, irreducible, aperiodic, rested Markov chains. Then, using UCB, we have

$$\left| R(n) - \left(n\mu^1 - \sum_{i=1}^{K} \mu^i E[T^i(n)] \right) \right| \le C_{\mathbf{S},\mathbf{P},\mathbf{r}} \tag{3}$$

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

$$E[T^i(n)] \le 1 + \frac{4L\ln n}{(\mu^1-\mu^i)^2} + \frac{(|S^i|+|S^1|)\beta}{\pi_{\min}}.$$

Proof: see Appendix B. ■

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

$$R(n) \le 4L\ln n \sum_{i>1} \frac{1}{(\mu^1-\mu^i)} + \sum_{i>1}(\mu^1-\mu^i)\left(1+C_{i,1}\right) + C_{\mathbf{S},\mathbf{P},\mathbf{r}}$$

where $C_{i,1} = \frac{(|S^i|+|S^1|)\beta}{\pi_{\min}}$.

Proof:

$$\begin{aligned} R(n) &\le n\mu^1 - \sum_{i=1}^{K}\mu^i E[T^i(n)] + C_{\mathbf{S},\mathbf{P},\mathbf{r}} && (4)\\ &\le \sum_{i>1}(\mu^1-\mu^i)E[T^i(n)] + C_{\mathbf{S},\mathbf{P},\mathbf{r}}\\ &\le \sum_{i>1}(\mu^1-\mu^i)\left(1+\frac{4L\ln n}{(\mu^1-\mu^i)^2}+\frac{(|S^i|+|S^1|)\beta}{\pi_{\min}}\right)\\ &\quad + C_{\mathbf{S},\mathbf{P},\mathbf{r}} && (5)\\ &= 4L\ln n\sum_{i>1}\frac{1}{(\mu^1-\mu^i)} + \sum_{i>1}(\mu^1-\mu^i)\left(1+C_{i,1}\right)\\ &\quad + C_{\mathbf{S},\mathbf{P},\mathbf{r}} \end{aligned}$$

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

The aforementioned theorem says that provided that L satisfies the stated sufficient condition, UCB 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 IX.

We next extend the aforementioned results to the case where the player selects M arms at each time step. The multiple-play extension to UCB1, referred to as UCB-M below, is straightforward: initially, each arm is played M times in the first K slots (M arms in each slot, in arbitrary order); subsequently, at each time slot, the algorithm plays M of the K arms with the highest current indices. 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 $1, 2, \ldots, M$, 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. Under this presentation, let $T^{i,j}(t)$ be the total number of times (slots) player j played arm i up to the end of slot t.

Proofs of the following lemmas are not given since they are similar to the proofs of the lemmas presented earlier in this section.

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

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

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.

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

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

Theorem 2: 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

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

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

Proof:

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

Thus

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

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

V. Restless Bandit Problem: Single Play

In this and the next section, we study the restless bandit problem, in the single-play and the multiple-play cases, respectively. While the multiple-play case is more general, the analysis in the single-play case is more intuitive to illustrate with less cumbersome notations.

We construct an algorithm called the *regenerative cycle algorithm*, 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. In the following, we first present the key conceptual idea behind RCA, followed by a more detailed pseudocode. We then prove the logarithmic regret result.

As the name suggests, RCA operates in regenerative cycles. In essence, RCA 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 while discarding the rest of the observations (only for the computation of the index; they contribute 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, the player maintains a block structure; a block consists of a certain number of slots. Within a block, a player plays the same arm continuously till a certain prespecified 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

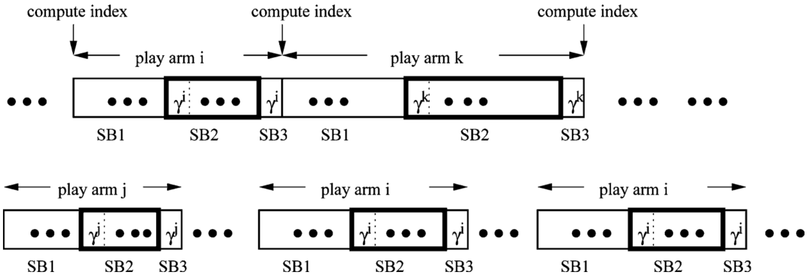

Fig. 2. Example realization of RCA.

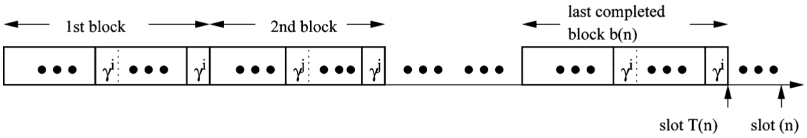

Fig. 3. Block structure of RCA.

the end of the block. 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. Fig. 2 shows an example sample path of the operation of RCA.

The key to the RCA 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.

1) γ^i: the state that determines the regenerative cycles for arm i.
2) $\tilde{\alpha}(b)$: the arm played in the bth block.
3) $b(n)$: the number of completed blocks up to time n.
4) $T(n)$: the time at the end of the last completed block (see Fig. 3).
5) $B^i(b)$: the total number of blocks within the first completed b blocks in which arm i is played.
6) $X_1^i(b)$: the vector of observed states from SB1 of the bth block in which arm i is played; this vector is empty if the first observed state is γ^i.
7) $X_2^i(b)$: the vector of observed states from SB2 of the bth block in which arm i is played;
8) $X^i(b)$: the vector of observed states from the bth block in which arm i is played. Thus, we have $X^i(b) = [X_1^i(b), X_2^i(b), \gamma^i]$.
9) $t(b)$: time at the end of block b.
10) $t_2(b)$: the number of time slots that lie within an SB2 of any completed block up to and including block b.
11) $T_2^i(t)$: the number of time slots arm i is played during SB2's when the number of time steps that lie within an SB2 is t.
12) $T^i(t)$: the number of time slots arm i is played by the end of time t.

The block structure along with some of the aforementioned definitions are presented in Fig. 3. 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'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

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

denotes the sample mean of the reward collected during SB2. It is also worth noting that under RCA, rewards are also collected during SB1's and SB3's. However, the computation of the indices only relies on SB2. The pseudocode of RCA is given in Fig. 4.

Proving the existence of a logarithmic upper bound on the regret for restless arms is a nontrivial 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 the regenerative property of the arms, all the observations from SB2's of an arm can be combined together and viewed as a sequence of continuous observations from a rested arm. Therefore, we can use a large deviation result to bound the expected number of times the index of a suboptimal arm exceeds the index of an optimal arm. Using this result, we show that the expected number of blocks in which a suboptimal arm is played is at most logarithmic in time. We then relate the expected number of blocks in which a suboptimal arm is played to the expected number of time slots in which a suboptimal arm is played using the positive recurrence property of the arms. Finally, we show that the regret due to arm switching is at most logarithmic, and the regret from the last, incomplete block is finite due to the positive recurrence property of the arms.

In the following, we first bound the expected number of plays from a suboptimal arm.

```
Regenerative Cycle Algorithm (RCA):
Initialize: $b = 1, t = 0, t_2 = 0, T_2^i = 0, r^i = 0, \forall i = 1, \cdots, K$
for $b \le K$ do
    play arm $b$; set $\gamma^b$ to be the first state observed
    $t := t+1$; $t_2 := t_2+1$; $T_2^b := T_2^b+1$; $r^b := r^b + r^b_{\gamma^i}$
    play arm $b$; denote observed state as $x$
    while $x \neq \gamma^b$ do
        $t := t+1$; $t_2 := t_2+1$; $T_2^b := T_2^b+1$; $r^b := r^b + r^b_x$
        play arm $b$; denote observed state as $x$
    end while
    $b := b+1$; $t := t+1$
end for
for $j = 1$ to $K$ do
    compute index $g^j := \frac{r^j}{T_2^j} + \sqrt{\frac{L \ln t_2}{T_2^j}}$
    $j++$
end for
$i := \arg\max_j g^j$
while (1) do
    play arm $i$; denote observed state as $x$
    while $x \neq \gamma^i$ do
        $t := t+1$
        play arm $i$; denote observed state as $x$
    end while
    $t := t+1$; $t_2 := t_2+1$; $T_2^i := T_2^i+1$; $r^i := r^i + r^i_x$
    play arm $i$; denote observed state as $x$
    while $x \neq \gamma^i$ do
        $t := t+1$; $t_2 := t_2+1$; $T_2^i := T_2^i+1$; $r^i := r^i + r^i_x$
        play arm $i$; denote observed state as $x$
    end while
    $b := b+1$; $t := t+1$
    for $j = 1$ to $K$ do
        compute index $g^j := \frac{r^j}{T_2^j} + \sqrt{\frac{L \ln t_2}{T_2^j}}$
        $j++$
    end for
    $i := \arg\max_j g^j$
end while
```

Fig. 4. Pseudocode of RCA.

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

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

where

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

$$D_i = \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + 1 \right).$$

Proof: See Appendix C. ■

We now state the main result of this section.

Theorem 3: 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 is upper bounded by

$$R(n) < 4L \ln n \sum_{i>1} \frac{1}{\mu^1 - \mu^i} \left(D_i + \frac{E_i}{\mu^1 - \mu^i} \right) + \sum_{i>1} C_{i,1} \left((\mu^1 - \mu^i) D_i + E_i \right) + F$$

where

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

$$D_i = \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + 1 \right)$$

$$E_i = \mu^i (1 + \Omega_{\max}^i) + \mu^1 \Omega_{\max}^1$$

$$F = \mu^1 \left(\frac{1}{\pi_{\min}} + \max_{i \in \{1, \ldots, K\}} \Omega_{\max}^i + 1 \right).$$

Proof: See Appendix D. ■

Theorem 3 suggests that 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 RCA that satisfies the stated condition. As in 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 IX.

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 used in the proof of Theorem 3 can be used to prove order optimality of combinations of RCA and other index policies. We comment more on this in Section IX.

VI. Restless Bandit Problem: Multiple Plays

In this section we extend the results of the previous section to the case of multiple plays. The multiple-play extension to the regenerative cycle algorithm will be referred to as the RCA-M. As in the rested case, even though our basic model is one of single player with multiple plays, our description is in the equivalent form of multiple coordinated players each with a single play.

As in RCA, RCA-M maintains the same block structure, where a player plays the same arm till it completes a regenerative cycle. 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 following analysis, 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. Fig. 5 shows an example sample path of the operation of RCA-M. The block structure of two players and the ordering of the blocks are shown.

The pseudocode of RCA-M is given in Fig. 6. The analysis is similar to that in Section V, with careful accounting of the expected number of blocks in which a suboptimal arm is played. The details can be found in the proof of Theorem 3.

Theorem 4: 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

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

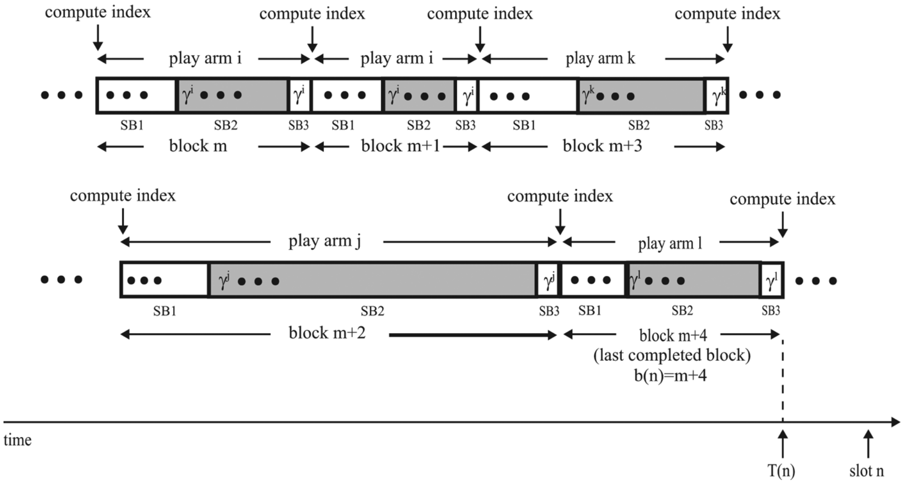

Fig. 5. Example realization of RCA-M with $M = 2$ for a period of n slots.

where

$$C_{i,j} = \frac{(|S^i| + |S^j|)\beta}{\pi_{\min}}, \quad \beta = \sum_{t=1}^{\infty} t^{-2}$$

$$D_i = \left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + 1\right)$$

$$E_i = \mu^i(1 + \Omega^i_{\max}) + \sum_{j=1}^{M} \mu^j \Omega^j_{\max}$$

$$F = \sum_{j=1}^{M} \mu^j \left(\frac{1}{\pi_{\min}} + \max_{i \in \mathcal{K}} \Omega^i_{\max} + 1\right).$$

Proof: See Appendix E. ■

VII. Decentralized Multiplayer Restless Bandit

In this section, we analyze the decentralized multiplayer restless bandit problem. In this case, there are M uncoordinated players. Each player must choose a single arm to play at each time step. A player is able to observe the state of the arm it selects, but if two or more players select the same arm simultaneously, then a *collision* results and no player involved receives any reward. Compared to the centralized restless bandit with multiple plays studied in Section VI, which was described from the point of view of multiple coordinated players, the key difference here is the possibility of collision which reduces a player's reward. Note however that a player's ability to observe state information remains unchanged from the previous case.

Within the context of our motivating application, this problem models a decentralized multiuser dynamic spectrum access scenario, where multiple users compete for a common set of channels. Each user performs channel sensing and data transmission tasks in each time slot. Sensing is done at the beginning of a slot; the user observes the quality of a selected channel. This is followed by data transmission in the same channel. The user receives feedback at the end of the slot (e.g., in the form of an acknowledgment) on whether the transmission is successful. If more than one user selects the same channel in the same slot, then a collision occurs and none of the users gets any reward.

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

```
Initialize: $b = 1, t = 0, t_2 = 0, T_2^i = 0, r^i = 0, I^i_{SB2} = 0, I^i_{IN} = 1, \forall i = 1, \cdots, K$, $A = \emptyset$
    //$I^i_{IN}$ indicates whether arm $i$ has been played at least once
  //$I^i_{SB2}$ indicates whether arm $i$ is in an SB2 sub-block
while (1) do
  for $i = 1$ to $K$ do
    if $I^i_{IN} = 1$ 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 $\{i : g^i$ is one of the $M - |A|$ largest among $\{g^k, k \in \{1, \cdots, K\} - A\}\}$
      //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 $I^i_{IN} = 1$ then
      $\gamma^i = x^i, T_2^i := T_2^i + 1, r^i := r^i + r^i_{x^i}, I^i_{IN} = 0, I^i_{SB2} = 1$
        //the first observed state becomes the regenerative state; the arm enters SB2
    else if $x^i \neq \gamma^i$ and $I^i_{SB2} = 1$ then
      $T_2^i := T_2^i + 1, r^i := r^i + r^i_{x^i}$
    else if $x^i = \gamma^i$ and $I^i_{SB2} = 0$ then
      $T_2^i := T_2^i + 1, r^i := r^i + r^i_{x^i}, I^i_{SB2} = 1$
    else if $x^i = \gamma^i$ and $I^i_{SB2} = 1$ then
      $r^i := r^i + r^i_{x^i}, I^i_{SB2} = 0, A \leftarrow A - \{i\}$
    end if
  end for
  $t := t + 1, t_2 := t_2 + \min\{1, \sum_{i \in S} I^i_{SB2}\}$    //$t_2$ 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. 6. Pseudocode of RCA-M.

The algorithm we construct and analyze in this section is a decentralized extension to RCA-M and will be referred to as the decentralized regenerative cycle algorithm or DRCA. This algorithm works similarly as RCA-M, using the same block structure. However, since players are uncoordinated, each player keeps its own locally computed indices for all arms, and they may vary from player to player. As earlier, a player continues to play the same arm till it completes a block, upon which it updates the indices for the arms using state observations from SB2's. Within this completed block, it may experience collision in any of the time slots; for these slots, it does not receive any reward. At the end of a block, if the player did not experience a collision in the last slot of the block, it continues to play the arm with the same rank in the next block after the index update. If it did experience a collision, then the player updates the indices for the arms, and then randomly selects an arm within the top M arms, based on the indices it currently has for all the arms, to play in the next block.

We see that compared to RCA-M, the main difference in DRCA is the randomization upon completion of a block. This is because if all players choose the arm with the highest index, then collision will be high even if players do not have exactly the same local indices; this in turn leads to large regret. Letting a player randomize among its M highest-ranked arms can help alleviate this problem, and aims to eventually orthogonalize the M users in their choice of arms. This is the same idea used in [14]. The difference is that in [14], the randomization is done each time a collision occurs under an i.i.d. reward model, whereas in our case, the randomization is done at the end of a completed block and is therefore less frequent as block lengths are random. The reason for this is because with the Markovian reward model, index updates can only be done after a regenerative cycle; switching before a block is completed will waste the state observations made within that incomplete block.

In the remainder of this section, we show that using the aforementioned algorithm, the regret summed over all players with respect to the optimal centralized (coordinated) solution, where M players always play the M-best arms, is polylogarithmic in time. Our analysis follows a similar approach as in [14], adapted to blocks rather than time slots and with a number of technical differences. In particular, the proof of Lemma 9 is significantly different because a single block of some player may collide with multiple blocks of other players; thus, we need to consider the actions of players jointly.

Let $Y^{k,j}(b)$ be the sample mean of the rewards inferred from state observations (not the actual rewards received since in this case reward is zero when there is collision) by player j during its bth block in which it plays arm k. Without loss of generality, in this section, we assume that $r_x^i \leq 1,\ \forall x \in S^i,\ i \in \mathcal{K}$. Let $B^{k,j}(b)$ be the number of blocks in which arm k is played by j at the end of its bth block. Let $b_j(n)$ be the number of j's completed blocks up to time n. Then, the index of arm k computed (and perceived) by player j at the end of its bth completed block is given by

$$g^{k,j}(b) = \frac{\sum_{v=1}^{B^{k,j}(b)} Y^{k,j}(v)}{B^{k,j}(b)} + \sqrt{\frac{2\ln b}{B^{k,j}(b)}}. \quad (9)$$

The difference between the index given in (8) and (9) is that the exploration term in (9) depends on the number of blocks completed by a player, while in (8) it depends on the number of time steps spent in SB2's of a player.

Let $J(n)$ be the number of slots involving collisions in the M optimal arms in the first n slots, and let $T^{k,j}(n)$ denote the number of slots player j plays arm k up to time n. Then, from Proposition 1 in [14], we have

$$R(n) \leq \mu^1 \left(\sum_{j=1}^{M} \sum_{k=M+1}^{K} E[T^{k,j}(n)] + E[J(n)] \right). \quad (10)$$

This result relates the regret to the amount of loss due to collision in the optimal arms, and the plays in the suboptimal arms.

Lemma 7: Under DRCA, for any player j and any suboptimal arm k, we have

$$E[B^{k,j}(b_j(n))] \leq \frac{8\ln n}{(\mu^M - \mu^k)^2} + 1 + M\beta.$$

Proof: See Appendix F. ■

The next lemma shows that provided all players have the correct ordering of arms, the expected number of blocks needed to reach an orthogonal configuration by randomization at the end of blocks is finite.

Lemma 8: Given all players have the correct ordering of the arms and do not change this ordering anymore, the expected number of blocks needed summed over all players to reach an orthogonal configuration is bounded above by

$$O_B = M\left[\binom{2M-1}{M} + 1\right].$$

Proof: The proof is similar to the proof of Lemma 2 in [14], by performing randomization at the end of each block instead of at every time step. ■

Let $B'(n)$ be the number of completed blocks up to n, in which at least one of the top M estimated ranks of the arms at some player is wrong. Let $p_{xy}^k(t)$ be the t step transition probability from state x to y of arm k. Since all arms are ergodic, there exists $N > 0$ such that $p_{xy}^k(N) > 0$, for all $k \in \mathcal{K},\ x, y \in S^k$. We now bound the expectation of $B'(n)$.

Lemma 9: Under DRCA, we have

$$E[B'(n)] < M\left[2N(M-1)\left(1 + \frac{1}{\lambda}(\ln n + 1)\right) + 1\right] \times \sum_{a=1}^{M} \sum_{c=a+1}^{K} \left(\frac{8\ln n}{(\mu^a - \mu^c)^2} + 1 + \beta\right)$$

where N is the minimum integer such that $p_{xy}^k(N) > 0$ for all $k \in \mathcal{K},\ x, y \in S^k$, $\lambda = \ln\left(\frac{1}{1-p^*}\right)$, and $p^* = \min_{k\in\mathcal{K},\ x,y\in S^k} p_{xy}^k(N)$.

Proof: See Appendix G. ■

Next, we show that the expected number of collisions in the optimal arms is at most $\log^2(.)$ in time. Let $H(n)$ be the number of completed blocks in which some collision occurred in the optimal arms up to time n.

Lemma 10: Under DRCA, we have

$$E[H(n)] \leq O_B E[B'(n)].$$

Proof: See Appendix H. ■

Combining all the aforementioned lemmas and using the fact that the expected block length is finite, we have the following result.

Theorem 5: When all players use DRCA, we have

$$\begin{aligned} R(n) \leq \mu^1 D_{\max} \Bigg[& \sum_{j=1}^{M} \sum_{k=M+1}^{K} \left(\frac{8 \ln n}{(\mu^M - \mu^k)^2} + M\beta + 2 \right) \\ & + O_B M \left[2N(M-1) \left(1 + \frac{1}{\lambda} (\ln n + 1) \right) + 1 \right] \\ & \times \sum_{a=1}^{M} \sum_{c=a+1}^{K} \left(\frac{8 \ln n}{(\mu^a - \mu^c)^2} + 1 + \beta \right) \Bigg] \end{aligned}$$

where

$$D_{\max} = \frac{1}{\pi_{\min}} + \Omega_{\max} + 1, \quad \Omega_{\max} = \max_{k \in \mathcal{K}} \Omega_{\max}^k$$
$$\pi_{\min} = \min_{k \in \mathcal{K}} \pi_{\min}^k.$$

Proof: Since the expected length of each block is at most $D_{\max}$ and the expected number of time steps between current time n and the time of the last completed block is at most the expected block length, we have $E[T^{k,j}(n)] \leq D_{\max}(E[B^{k,j}(b_j(n))] + 1)$ and $E[J(n)] \leq D_{\max}(E[H(n)] + 1)$. The result follows from substituting these into (10) and using results of Lemmas 7 and 10. ■

It is worth mentioning that our proof of the polylogarithmic regret upper bound in this section is based on the regenerative cycles but does not rely on a large deviation bound for Markov chains as we have done in the previous sections. The main idea is that the sample mean rewards observed within regenerative cycles with the same regenerative state form an i.i.d. random process; our results are easier to prove by exploiting the i.i.d. structure. The same method can be used in the previous sections as well by choosing a constant regenerative state for each arm. Moreover, under this method, we no longer need the assumption that $p_{xx}^k > 0$ for any $k \in \mathcal{K},\ x \in S^k$. Indeed, with this method, the same results can be derived for arbitrary non-Markovian discrete-time renewal processes with finite mean cycle time and bounded rewards. However, we note that the previous method based on the large deviation bound for Markov chains is still of importance because it works when the regenerative states are adapted over time. In this case, the cycles are no longer i.i.d. and the expected average reward in a cycle is not necessarily the mean reward of an arm. We give applications where there is a need to change the regenerative state of an arm over time in Section IX.

TABLE I
TRANSITION PROBABILITIES OF ALL CHANNELS

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

VIII. EXAMPLES IN OSA USING THE GILBERT–ELLIOT CHANNEL MODEL

In this section, we give numerical results for the algorithms we proposed under the Gilbert–Elliot channel model in which each channel has two states, *good* and *bad* (or 1, 0, respectively). For any channel i, the rewards are given by $r_1^i = 1$, $r_0^i = 0.1$. We consider four OSA scenarios, denoted S1–S4, each consisting of ten channels with different state transition probabilities. The state transition probabilities and mean rewards of the channels in each scenario are given in Tables I and II, respectively. The four scenarios are intended to capture the following differences. In S1, channels are bursty with mean rewards not close to each other; in S2, channels are nonbursty with mean rewards not close to each other; in S3, there are bursty and nonbursty channels with mean rewards not close to each other; and in S4, there are bursty and nonbursty channels with mean rewards close to each other. All simulations are done for a time horizon $n = 10^5$, and averaged over 100 random runs. Initial states of the channels are drawn from their stationary distributions. For each algorithm that requires a regenerative state, the regenerative state of an arm for a player is set to be the first state the player observes from that arm, and is kept fixed throughout a single run.

We first compute the normalized regret values, i.e., the regret per play $R(n)/M$, for RCA-M. In Figs. 7, 9, 11, and 13, we observe the normalized regret of RCA-M for the minimum values of L such that the logarithmic regret bound holds. However, comparing with Figs. 8, 10, 12, and 14, we see that the normalized regret is smaller for $L = 1$. Therefore, it appears that the condition on L we have for the logarithmic bound, while sufficient, may not be necessary.

We next compute the regret of UCB with single play under the OSA model. We note that our theoretical regret bound for UCB is for rested channels but the numerical results are given for a special case of restless channels. Results in Fig. 15 show that when $L = 1$, for S1, S3, and S4, UCB has negative regret, which means that it performs better than the best single action policy, while for S2, it has a positive regret, which is also greater than the regret of RCA with single play under S2 with $L = 1$. In Fig. 16, we see the regret of UCB for larger values of L. As expected, the regret of UCB increases with L due to the increase in explorations. However, comparing the regret of UCB with that of RCA under the same value of L, we see that UCB outperforms RCA for all scenarios considered here. These results imply that although there is no theoretical bounds for the regret of UCB, its performance is comparable to RCA under the

TABLE II
MEAN REWARDS OF ALL CHANNELS

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

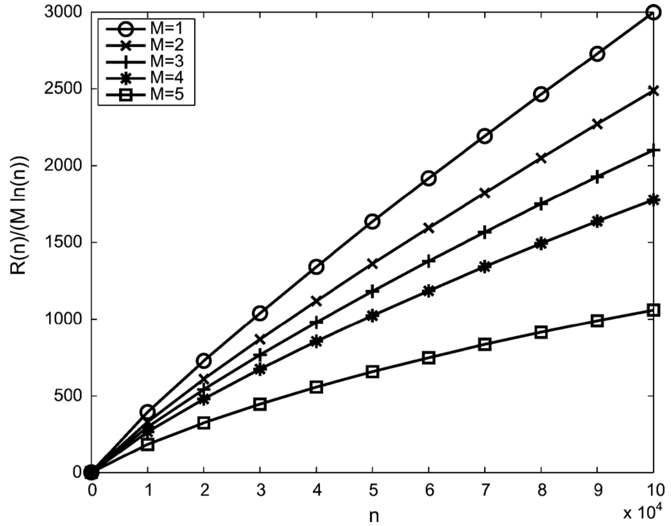

Fig. 7. Normalized regret of RCA-M: S1, $L = 7200$.

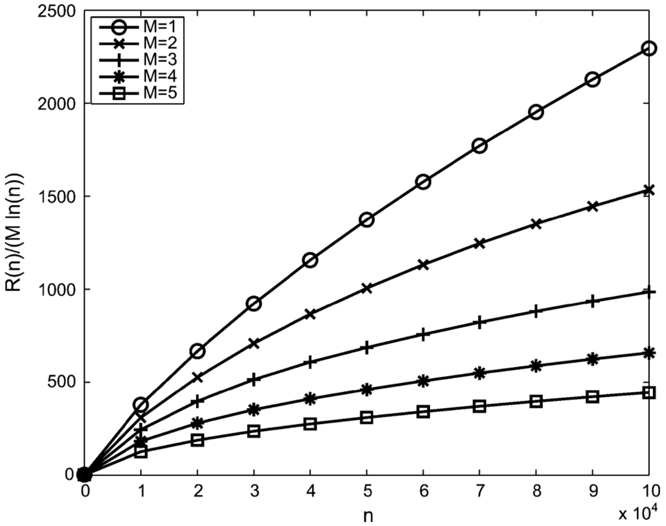

Fig. 9. Normalized regret of RCA-M: S2, $L = 360$.

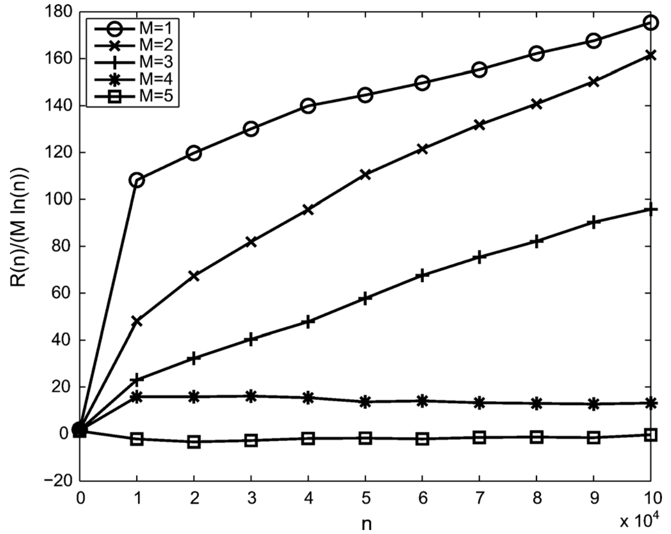

Fig. 8. Normalized regret of RCA-M: S1, $L = 1$.

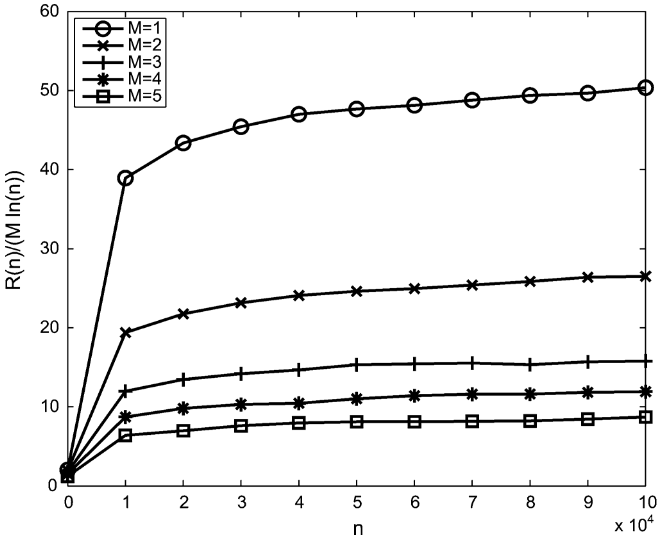

Fig. 10. Normalized regret of RCA-M: S2, $L = 1$.

presented setting. This is because 1) RCA has a smaller update rate due to the random length of the regenerative cycles; thus, it takes longer to use the latest observations in arm selection, and 2) even though there is no guarantee that UCB produces accurate estimates on the mean rewards, the simple structure of the problem helps UCB keep track of the shorter term (not the stationary) quality of each arm.

We also compute the regret of RCA with the index given in (9), where the exploration term is the ratio of the number of completed blocks b to the number of completed blocks of arm i up to b. This approach reduces the problem to an i.i.d. one, where the average reward in each block can be seen as a random reward drawn from an i.i.d. arm. We can then exploit the well-known result for the i.i.d. problem [3] which says that setting $L = 2$ is enough to get a logarithmic regret bound. The regret for single play under different scenarios is given in Fig. 17. Comparing them with their counterparts using RCA with an L such that the logarithmic regret bound holds, we observe that the modified index results in better performance. This is because L is smaller, and the exploration is more *balanced* in a way that the growth

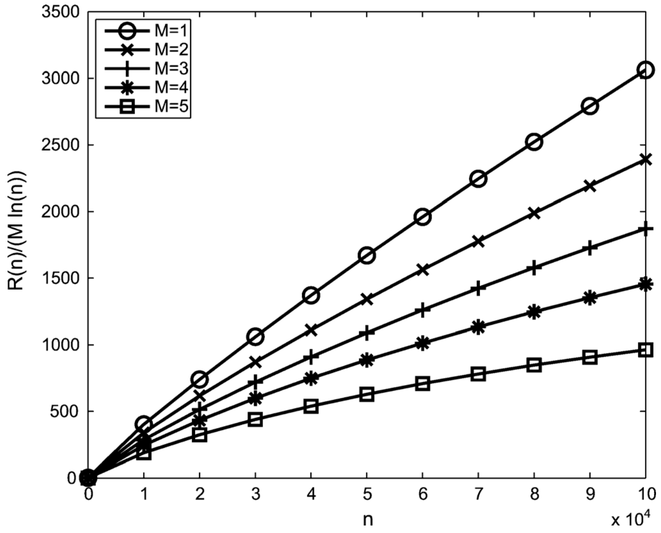

Fig. 11. Normalized regret of RCA-M: S3, $L = 3600$.

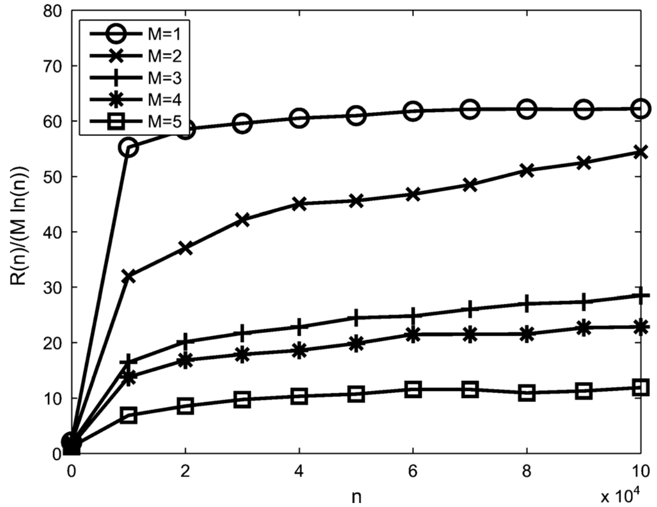

Fig. 12. Normalized regret of RCA-M: S3, $L = 1$.

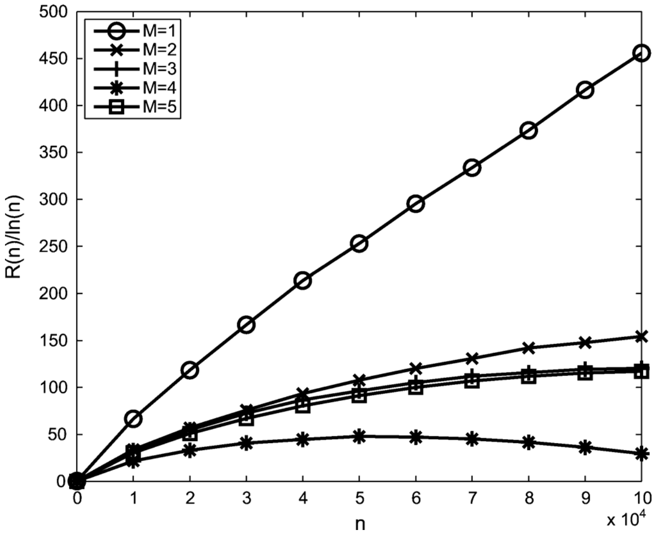

Fig. 13. Normalized regret of RCA-M: S4, $L = 7200$.

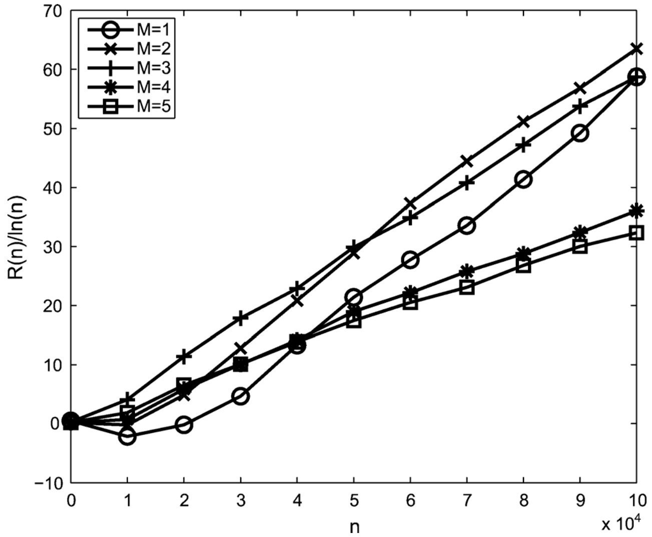

Fig. 14. Normalized regret RCA-M: S4, $L = 1$.

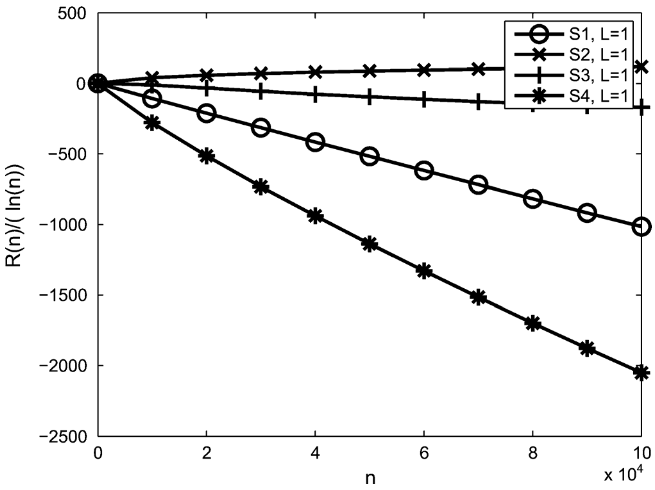

Fig. 15. Regret of UCB, $M = 1$.

of the exploration term does not depend on the randomness of the block lengths.

Finally, we present the regret of DRCA with two users in Fig. 18. The results are similar to that of RCA with the index given in (9), but with a larger regret due to collisions.

IX. Discussion

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

A. Applicability and Performance Improvement

We note that the same logarithmic bound derived in this paper holds for the general restless bandit problem independent of the state transition law of an arm when it is not played. Indeed, the state transitions of an arm when it is not played can even be adversarial. This is 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 is

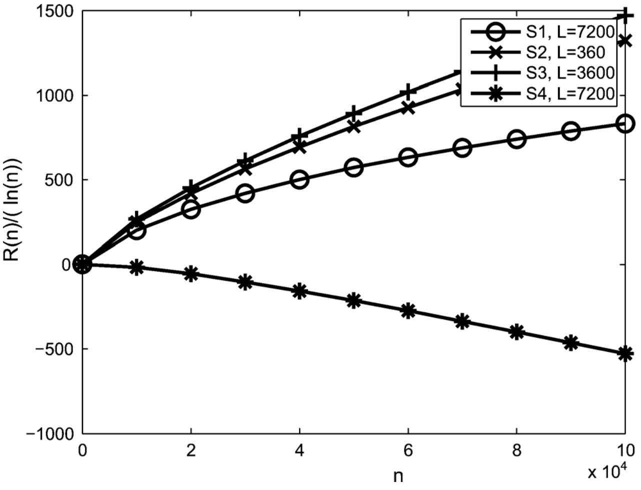

Fig. 16. Regret of UCB, $M = 1$.

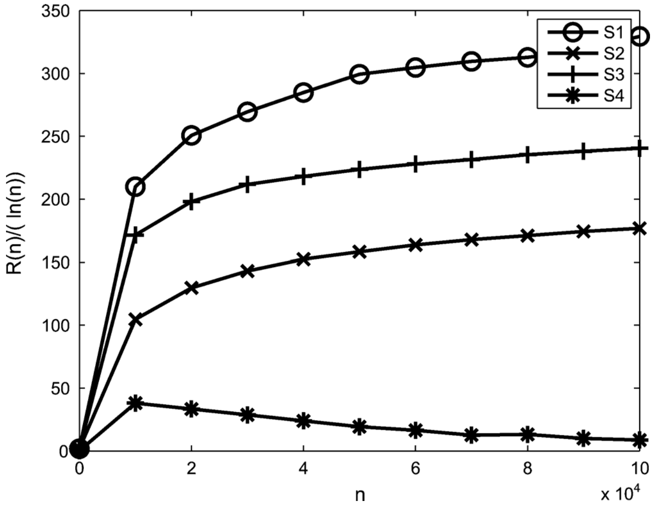

Fig. 17. Regret of RCA with modified index.

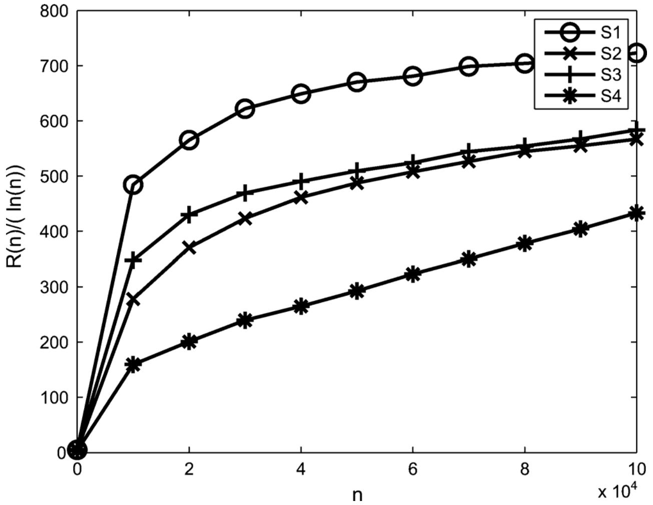

Fig. 18. Regret of DRCA with two users.

at most logarithmic independent of the first observed state of a block.

The regenerative state for an arm under RCA is chosen based on the random initial observation. It is 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 V holds.

It is possible that RCA may happen upon a state with long recurrence time which results in long SB1 and SB2 subblocks. Consider now the following modification: RCA 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 nth block. If the arm to be played in the nth 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.

B. Universality of the Block Structure

We note that any index policy used under the i.i.d. reward model can be used in the restless bandit problem with a Markovian reward model by exploiting the regenerative cycles. This is because the normalized rewards collected in each regenerative cycle of the same arm can be seen as i.i.d. samples from that arm whose expectation is equal to the mean reward of that arm. Thus, any upper bound for the expected number of times an arm is played in an i.i.d. problem will hold for the expected number of blocks an arm is played for the restless bandit problem under the block structure proposed in RCA. Specifically, we have shown via numerical results in Section VIII that if RCA is used with the index given in (9), logarithmic regret is achieved assuming that the regenerative state for each arm is kept fixed and the rewards are in the unit interval $[0,1]$. We do not provide a technical analysis here since the details are included in the analysis of the i.i.d. model [3] and our analysis of RCA. Instead, we illustrate the generality of the block structure by using the KL-UCB algorithm proposed in [8] for i.i.d. rewards inside our block structure. KL-UCB is shown to outperform most of the other index policies for i.i.d. rewards including UCB. For simplicity, we only consider single play, i.e., $M = 1$.

Lemma 11: Assume that Condition 1 holds and $r_x^i \leq 1,\ \forall i \in \mathcal{K},\ \forall x \in S^i$. Then, using KL-UCB in the regenerative block under RCA, we have for any suboptimal arm i

$$\limsup_{n\to\infty} \frac{E\left[B^i(b(n))\right]}{\log b(n)} \leq \frac{1}{d(\mu^i, \mu^1)}$$

where

$$d(p,q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q}).$$

Proof: The normalized reward during a block (sum of the rewards collected during an SB2 divided by the length of the SB2) forms an i.i.d. process with support in $[0,1]$. Thus, the result follows from [8, Th. 2]. ■

We see as earlier, bounding the expected number of blocks a suboptimal arm is played is a key step in bounding the regret. The main result is given in the following theorem.

Theorem 6: Assume that Condition 1 holds and $r_x^i \leq 1$, $\forall i \in \mathcal{K}$, $\forall x \in S^i$. Then, using KL-UCB in the regenerative block under RCA, we have

$$\limsup_{n\to\infty} \frac{R(n)}{\log n} \leq \sum_{i>1} \left(\frac{1}{d(\mu^i,\mu^1)}\right) ((\mu^1-\mu^i)D_i + E_i)$$

where

$$\begin{aligned} D_i &= \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + 1\right) \\ E_i &= \mu^i(1+\Omega_{\max}^i) + \mu^1\Omega_{\max}^1. \end{aligned}$$

Proof: The result follows from Lemma 11 and using steps similar to the proof of Theorem 3. ■

C. Extension to Random State Rewards

So far we considered the case where each state $x \in S^i$ corresponds to a deterministic reward r_x^i. An interesting extension is to consider the case where each state $x \in S^i$ has a random reward with a fixed i.i.d. distribution. That is, after observing the state x, a player receives reward r_x^i which is drawn from a probability distribution F_x^i. In our application context, this may correspond to a situation where the player/user observes the received SNR which gives the probability of correct reception, but not the actual bit error rate. When the distribution (or its expectation) F_x^i, $\forall x \in S^i$, $i \in \mathcal{K}$ is known to the player, the player can use the expectation of the distribution of each state instead of the actual observed rewards to update the indices. In doing so, logarithmic regret will be achieved by using RCA.

A more complex case is when the reward distribution of each state is unknown to the player but has a bounded support. Then, to estimate the quality of each arm, playing logarithmic number of blocks from each arm may not be sufficient because there may be cases where the number of samples from some state of a sufficiently sampled arm may not be enough to accurately estimate the quality of that state. This may result in an inaccurate estimate of the expected reward received during a regenerative cycle. To avoid this, we can use arbitrary regenerative states discussed in Section IX-A, and modify RCA as follows: at the end of each block in which arm i is played, we record the least sampled state x of arm i up to that point. Whenever arm i is played in a block, the state x is then used as the regenerative state to terminate that block. This guarantees that x is sampled at least once during that block. Of course, to preserve the regenerative property, the player needs to set the first state of its next SB2 to x in the next block it plays arm i. This way *fairness* between the states of each arm is guaranteed. By logarithmically playing each arm, the player can guarantee logarithmic number of samples taken from each state; thus, the sample mean estimate of the expected reward of each state will be accurate. Then, the player can use the sample mean of the rewards for each state in calculating the index with RCA to obtain good performance. In order to have theoretical results, we will need to use two large deviation bounds: one for the sample mean estimates of the rewards of each state of each arm, and the other for bounding the deviation of the index from the expected reward of an arm. The detailed analysis is omitted for brevity.

D. Relaxation of Certain Conditions

As observed in Section VIII, the condition on L, while sufficient, does not appear necessary for the logarithmic regret bound to hold. Indeed, our examples will 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. If we use an alternative bound, e.g., the large deviation bound in [10], then $L \geq 90S_{\max}^2 r_{\max}^2/\epsilon_{\min}$ will be sufficient, and our theoretical results will hold for smaller L, provided that $\hat{\pi}_{\max}^2 \geq 90/112$ and the arms are reversible Markov chains.

We further note that even if no information is available on the underlying Markov chains to derive this sufficient condition on L, $o(log(n)f(n))$ 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., [12] and [14]).

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. In the following, 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 Mth and the $M+1$th arms can be avoided. However, further analysis is needed to verify that this approach will result in logarithmic regret.

E. 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 policy is not a single-action policy (in principle, one can derive such a policy using dynamic programming). Ideally, one could try to adopt a stronger regret measure with respect to this optimal policy. Under some conditions on the structure of the optimal policy, we have proposed a learning algorithm with logarithmic regret with respect to the optimal policy in [21]. However, in general, such an optimal policy is PSPACE-hard even to approximate in the restless case (see, e.g., [17], [22]), which makes the comparison intractable, except for some very limited cases when such a policy happens to be known (see, e.g., [18] and [23]) or special cases when approximation algorithms with guaranteed performance are known (see, e.g., [24] and [25]).

F. Comparison With Similar Work

A recent work [12] considers the same restless multiarmed bandit problem studied in this paper. They achieve logarithmic regret by using exploration and exploitation blocks that grow geometrically with time. The construction in [12] is very different from ours. The essence behind our approach RCA-M is to reduce a restless bandit problem to a rested bandit problem; this is 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 [12] applies large deviation bounds to individual segments 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 segments, 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 [12], the two are done separately and explicitly through two different types of blocks.

X. Conclusion

In this paper, we considered the rested and restless bandit problems with Markovian rewards and multiple plays in both a centralized and a decentralized setting. We showed that a simple extension to UCB1 produces logarithmic regret uniformly over time for the centralized rested bandit problem. 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 centralized restless bandit problem. For the decentralized multiplayer restless bandit problem, we introduced the DRCA algorithm and proved that polylogarithmic regret is achievable under the collision model where no player gets a reward when there is more than one player using the same arm. We numerically examined the performance of the RCA in the case of an OSA problem with the Gilbert–Elliot channel model and compared it with the naive UCB algorithm. Finally, we discussed possible extensions and improvements.

Appendix A
Proof of Lemma 2

We first state the following lemma which will be used to prove Lemma 2.

Lemma 12: [Lemma 2.1 From [6]]: Let Y be an irreducible aperiodic Markov chain with a state space S, transition probability matrix P, an initial distribution that is nonzero in all states, and a stationary distribution $\{\pi_x\}$, $\forall x \in S$. Let F_t be the σ-field generated by random variables $X_1, X_2, \ldots, X_t$ 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 $F_1, F_2, \ldots$. 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

$$|E[N(x,\tau)] - \pi_x E[\tau]| \leq C_P \tag{11}$$

where C_P is a constant that depends on P.

Let $X^i(t)$ be the state observed from the tth play of arm i. We have

$$\begin{aligned}
&\left| R(n) - \left(n\mu^1 - \sum_{i=1}^{K} \mu^i E[T^i(n)] \right) \right| \\
&= \left| E\left[\sum_{i=1}^{K} \sum_{x\in S^i} r_x^i \sum_{t=1}^{T^i(n)} I(X^i(t) = x) \right] \right. \\
&\quad \left. - \sum_{i=1}^{K} \sum_{x\in S^i} r_x^i \pi_x^i E[T^i(n)] \right| \\
&= \left| \sum_{i=1}^{K} \sum_{x\in S^i} r_x^i (E[N^i(x, T^i(n))] - \pi_x^i E[T^i(n)]) \right| \\
&\leq \sum_{i=1}^{K} \sum_{x\in S^i} r_x^i C_{P^i} = C_{\mathbf{S},\mathbf{P},\mathbf{r}}
\end{aligned} \tag{12}$$

where

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

and (12) follows from Lemma 12 using the fact that $T^i(n)$ is a stopping time with respect to the σ-field generated by the arms played up to time n.

Appendix B
Proof of Lemma 3

We first state and prove the following lemma which will be used to prove Lemma 3.

Lemma 13: 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 with constant $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2/\epsilon_{\min}$, for any suboptimal arm i, we have

$$E\left[\sum_{t=1}^{n} \sum_{w=1}^{t-1} \sum_{w_i=l}^{t-1} I(g_{t,w}^1 \leq g_{t,w_i}^i) \right] \leq \frac{|S^i| + |S^1|}{\pi_{\min}} \beta \tag{13}$$

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

Proof: First, we show that for any suboptimal arm i, we have that $g^1_{t,w} \leq g^i_{t,w_i}$ implies at least one of the following holds:

$$\bar{r}^1(w) \leq \mu^1 - c_{t,w} \tag{14}$$
$$\bar{r}^i(w_i) \geq \mu^i + c_{t,w_i} \tag{15}$$
$$\mu^1 < \mu^i + 2c_{t,w_i}. \tag{16}$$

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

$$\begin{aligned} g^1_{t,w} &= \bar{r}^1(w) + c_{t,w} > \mu^1 \geq \mu^i + 2c_{t,w_i} \\ &> \bar{r}^i(w_i) + c_{t,w_i} = g^i_{t,w_i} \end{aligned}$$

which contradicts $g^1_{t,w} \leq g^i_{t,w_i}$.

If we choose $w_i \geq 4L\ln n/(\mu^1 - \mu^i)^2$, then

$$2c_{t,w_i} = 2\sqrt{\frac{L\ln t}{w_i}} \leq 2\sqrt{\frac{L\ln t(\mu^1-\mu^i)^2}{4L\ln n}} \leq \mu^1 - \mu^i$$

for $t \leq n$, which means (16) is false, and therefore at least one of (14) and (15) is true with this choice of w_i. Let $l = \left\lceil \frac{4L\ln n}{(\mu^1-\mu^i)^2} \right\rceil$. Then, we have

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

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

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

where the first inequality follows from the Minkowski inequality. Let $n^i_y(t)$ denote the number of times state y of arm i is observed up to and including the tth play of arm i

$$\begin{aligned} &P(\bar{r}^i(w_i) \geq \mu^i + c_{t,w_i}) \\ &= P\left(\sum_{y\in S^i} r^i_y n^i_y(w_i) \geq w_i \sum_{y\in S^i} r^i_y \pi^i_y + w_i c_{t,w_i}\right) \\ &= P\left(\sum_{y\in S^i} (r^i_y n^i_y(w_i) - w_i r^i_y \pi^i_y) \geq w_i c_{t,w_i}\right) \\ &= P\left(\sum_{y\in S^i} (-r^i_y n^i_y(w_i) + w_i r^i_y \pi^i_y) \leq -w_i c_{t,w_i}\right). \end{aligned} \tag{17}$$

Consider a sample path ω and the events

$$A = \left\{\omega : \sum_{y\in S^i}(-r^i_y n^i_y(w_i)(\omega) + w_i r^i_y \pi^i_y) \leq -w_i c_{t,w_i}\right\}$$
$$B = \bigcup_{y\in S^i}\left\{\omega : -r^i_y n^i_y(w_i)(\omega) + w_i r^i_y \pi^i_y \leq -\frac{w_i c_{t,w_i}}{|S^i|}\right\}.$$

If $\omega \notin B$, then

$$\begin{aligned} &-r^i_y n^i_y(w_i)(\omega) + w_i r^i_y \pi^i_y > -\frac{w_i c_{t,w_i}}{|S^i|}, \ \forall y \in S^i \\ &\Rightarrow \sum_{y\in S^i}(-r^i_y n^i_y(w_i)(\omega) + w_i r^i_y \pi^i_y) > -w_i c_{t,w_i}. \end{aligned}$$

Thus, $\omega \notin A$; therefore, $P(A) \leq P(B)$. Then, continuing from (17)

$$\begin{aligned} &P(\bar{r}^i(w_i) \geq \mu^i + c_{t,w_i}) \\ &\leq \sum_{y\in S^i} P\left(-r^i_y n^i_y(w_i) + w_i r^i_y \pi^i_y \leq -\frac{w_i c_{t,w_i}}{|S^i|}\right) \\ &= \sum_{y\in S^i} P\left(r^i_y n^i_y(w_i) - w_i r^i_y \pi^i_y \geq \frac{w_i c_{t,w_i}}{|S^i|}\right) \\ &= \sum_{y\in S^i} P\left(n^i_y(w_i) - w_i \pi^i_y \geq \frac{w_i c_{t,w_i}}{|S^i| r^i_y}\right) \\ &= \sum_{y\in S^i} P\left(\frac{\sum_{t=1}^{w_i} I(X^i_t = y) - w_i \pi^i_y}{\hat{\pi}^i_y w_i} \geq \frac{c_{t,w_i}}{|S^i| r^i_y \hat{\pi}^i_y}\right) \\ &\leq \sum_{y\in S^i} N_{q^i} t^{-\frac{L\epsilon^i}{28(|S^i| r^i_y \hat{\pi}^i_y)^2}} \qquad (18) \\ &\leq \frac{|S^i|}{\pi_{\min}} t^{-\frac{L\epsilon_{\min}}{28 S^2_{\max} r^2_{\max} \hat{\pi}^2_{\max}}} \qquad (19) \end{aligned}$$

where (18) follows from Lemma 1 by letting

$$\gamma = \frac{c_{t,w_i}}{|S^i| r^i_y \hat{\pi}^i_y}, \quad f(X^i_t) = \frac{I(X^i_t = y) - \pi^i_y}{\hat{\pi}^i_y}$$

and recalling $\hat{\pi}^i_y = \max\{\pi^i_y, 1 - \pi^i_y\}$ (note $\hat{P}^i$ is irreducible). Similarly, we have

$$\begin{aligned} &P\left(\bar{r}^1(w) \leq \mu^1 - c_{t,w}\right) \\ &= P\left(\sum_{y\in S^1} r^1_y(n^1_y(w) - w\pi^1_y) \leq -wc_{t,w}\right) \\ &\leq \sum_{y\in S^1} P\left(r^1_y n^1_y(w) - w r^1_y \pi^1_y \leq -\frac{wc_{t,w}}{|S^1|}\right) \\ &= \sum_{y\in S^1} P\left(r^1_y(w - \sum_{x\neq y} n^1_x(w))\right. \\ &\left. - w r^1_y(1 - \sum_{x\neq y}\pi^1_x) \leq -\frac{wc_{t,w}}{|S^j|}\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{y\in S^1} P\left(r_y^1 \sum_{x\neq y} n_x^1(w) - w r_y^1 \sum_{x\neq y} \pi_x^1 \geq \frac{w c_{t,w}}{|S^1|} \right)\\
&\leq \sum_{y\in S^1} N_{q^1} t^{-\frac{L\epsilon^1}{28(|S^1| r_y^1 \hat{\pi}_y^1)^2}} \qquad (20)\\
&\leq \frac{|S^1|}{\pi_{\min}} t^{-\frac{L\epsilon_{\min}}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}} \qquad (21)
\end{aligned}
$$

where (20) again follows from Lemma 1. The result then follows from combining (19) and (21)

$$
\begin{aligned}
&E\left[\sum_{t=1}^{n}\sum_{w=1}^{t-1}\sum_{w_i=l}^{t-1} I(g_{t,w}^1 \leq g_{t,w_i}^i)\right]\\
&\leq \frac{|S^i|+|S^1|}{\pi_{\min}} \sum_{t=1}^{\infty}\sum_{w=1}^{t-1}\sum_{w_i=1}^{t-1} t^{-\frac{L\epsilon_{\min}}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}}\\
&= \frac{|S^i|+|S^1|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-\frac{L\epsilon_{\min}-56 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}{28 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2}}\\
&\leq \frac{|S^i|+|S^1|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}. \qquad (22)
\end{aligned}
$$

■

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

$$
\begin{aligned}
T^i(n) &= 1 + \sum_{t=K+1}^{n} I(\alpha(t)=i)\\
&\leq l + \sum_{t=K+1}^{n} I(\alpha(t)=i, T^i(t-1)\geq l). \qquad (23)
\end{aligned}
$$

Consider the event

$$E = \left\{ g_{t,T^1(t)}^1 \leq g_{t,T^i(t)}^i \right\}.$$

For a sample path $\omega \in E^C$, we have $\alpha(t) \neq i$. Therefore, $\{\omega : \alpha(t)=i\} \subset E$ and

$$
\begin{aligned}
&I(\alpha(t)=i,\ T^i(t-1)\geq l) \leq I(\omega\in E,\ T^i(t-1)\geq l)\\
&= I(g_{t,T^1(t)}^1 \leq g_{t,T^i(t)}^i,\ T^i(t-1)\geq l).
\end{aligned}
$$

Therefore, continuing from (23)

$$
\begin{aligned}
T^i(n) &\leq l + \sum_{t=K+1}^{n} I(g_{t,T^1(t)}^1 \leq g_{t,T^i(t)}^i, T^i(t-1)\geq l)\\
&\leq l + \sum_{t=K+1}^{n} I\left(\min_{1\leq w<t} g_{t,w}^1 \leq \max_{l\leq w_i<t} g_{t,w_i}^i\right)\\
&\leq l + \sum_{t=K+1}^{n}\sum_{w=1}^{t-1}\sum_{w_i=l}^{t-1} I(g_{t,w}^1 \leq g_{t,w_i}^i)\\
&\leq l + \sum_{t=1}^{n}\sum_{w=1}^{t-1}\sum_{w_i=l}^{t-1} I(g_{t,w}^1 \leq g_{t,w_i}^i).
\end{aligned}
$$

Using Lemma 13 with $l = \left\lceil \frac{4L\ln n}{(\mu^1-\mu^i)^2} \right\rceil$, we have for any suboptimal arm

$$E[T^i(n)] \leq 1 + \frac{4L\ln n}{(\mu^1-\mu^i)^2} + \frac{(|S^i|+|S^1|)\beta}{\pi_{\min}}. \qquad (24)$$

Appendix C
Proof of Lemma 6

We first state and prove the following lemma which will be used to prove Lemma 6.

Lemma 14: Assume Condition 1 holds and all arms are restless. Let $g_{t,w}^i = \bar{r}^i(w) + c_{t,w}$, $c_{t,w} = \sqrt{L\ln t/w}$. Under RCA with constant $L \geq 112 S_{\max}^2 r_{\max}^2 \hat{\pi}_{\max}^2/\epsilon_{\min}$, for any suboptimal arm i, we have

$$E\left[\sum_{t=1}^{t_2(b)}\sum_{w=1}^{t-1}\sum_{w_i=l}^{t-1} I(g_{t,w}^j \leq g_{t,w_i}^i)\right] \leq \frac{|S^i|+|S^1|}{\pi_{\min}}\beta \qquad (25)$$

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

Proof: Note that all the quantities in computing the indices in (25) comes from the intervals $X_2^i(1), X_2^i(2), \ldots \forall i \in \{1,\ldots,K\}$. 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 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. Therefore, we can proceed in exactly the same way as the proof of Lemma 13. If we choose $s_i \geq 4L\ln(n)/(\mu^1-\mu^i)^2$, then for $t \leq t_2(b) = n' \leq n$, and for any suboptimal arm i

$$2c_{t,s_i} = 2\sqrt{\frac{L\ln(t)}{s_i}} \leq 2\sqrt{\frac{L\ln(t)(\mu^1-\mu^i)^2}{4L\ln(n)}} \leq \mu^1-\mu^i.$$

The result follows from letting $l = \left\lceil \frac{4L\ln n}{(\mu^1-\mu^i)^2} \right\rceil$ and using Lemma 13. ■

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

$$
\begin{aligned}
B^i(b) &= 1 + \sum_{m=K+1}^{b} I(\tilde{\alpha}(m)=i)\\
&\leq l + \sum_{m=K+1}^{b} I(\tilde{\alpha}(m)=i, B^i(m-1)\geq l)\\
&\leq l + \sum_{m=K+1}^{b} I\Big(g_{t_2(m-1),T_2^1(t_2(m-1))}^1\\
&\qquad \leq g_{t_2(m-1),T_2^i(t_2(m-1))}^i, B^i(m-1)\geq l\Big)\\
&\leq l + \sum_{m=K+1}^{b} I\Big(\min_{1\leq w\leq t_2(m-1)} g_{t_2(m-1),w}^1
\end{aligned}
$$

$$\leq \max_{t_2(l) \leq w_i \leq t_2(m-1)} g^i_{t_2(m-1),w_i} \Bigg)$$

$$\leq l + \sum_{m=K+1}^{b} \sum_{w=1}^{t_2(m-1)} \sum_{w_i=t_2(l)}^{t_2(m-1)} I(g^1_{t_2(m),w} \leq g^i_{t_2(m),w_i}) \quad (26)$$

$$\leq l + \sum_{t=1}^{t_2(b)} \sum_{w=1}^{t-1} \sum_{w_i=l}^{t-1} I(g^1_{t,w} \leq g^i_{t,w_i}) \quad (27)$$

where as given in (8), $g^i_{t,w} = \bar{r}^i(w) + c_{t,w}$. The inequality in (27) follows from the fact that the outer sum in (27) is over time, while the outer sum in (26) is over blocks and each block lasts at least two time slots.

From this point on, we use Lemma 14 to get

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

for all suboptimal arms. Therefore

$$E[B^i(b(n))] \leq \frac{4L \ln n}{(\mu^1 - \mu^i)^2} + C_{i,1} \quad (28)$$

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

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

APPENDIX D
PROOF OF THEOREM 3

We first state the following lemma which will be used to prove Theorem 3.

Lemma 15: 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 $y \in S$

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

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

$$R_\gamma(n)$$
$$= \mu^1 E_\gamma[T(n)] - E_\gamma[\sum_{t=1}^{T(n)} r^{\alpha(t)}_{x_{\alpha(t)}}]$$
$$+ \mu^1 E_\gamma[n - T(n)] - E_\gamma[\sum_{t=T(n)+1}^{n} r^{\alpha(t)}_{x_{\alpha(t)}}]$$
$$= \left\{ \mu^1 E_\gamma[T(n)] - \sum_{i=1}^{K} \mu^i E_\gamma \left[T^i(T(n)) \right] \right\} - Z_\gamma(n)$$
$$+ \mu^1 E_\gamma[n - T(n)] - E_\gamma[\sum_{t=T(n)+1}^{n} r^{\alpha(t)}_{x_{\alpha(t)}}] \quad (29)$$

where for notational convenience, we have used

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

We can bound the first difference in (29) logarithmically using Lemma 6, so it remains to bound $Z_\gamma(n)$ and the last difference. We have

$$Z_\gamma(n) \geq \sum_{y \in S^1} r^1_y E_\gamma \left[\sum_{j=1}^{B^1(b(n))} \sum_{X^1_t \in X^1(j)} I(X^1_t = y) \right]$$
$$+ \sum_{i:\mu^i < \mu^1} \sum_{y \in S^i} r^i_y E_\gamma \left[\sum_{j=1}^{B^i(b(n))} \sum_{X^i_t \in X^i_2(j)} I(X^i_t = y) \right] \quad (30)$$
$$- \mu^1 E_\gamma \left[T^1(T(n)) \right]$$
$$- \sum_{i>1} \mu^i \left(\frac{1}{\pi^i_{\gamma^i}} + \Omega^i_{\max} + 1 \right) E_\gamma \left[B^i(b(n)) \right]$$

where the inequality comes from counting only the rewards obtained during the SB2s for all suboptimal arms. Applying Lemma 15 to (30), 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 $\mu^1 = \sum_y r^1_y \pi^1_y$

$$Z_\gamma(n) \geq R^1(n) - \sum_{i:\mu^i < \mu^1} \mu^i (\Omega^i_{\max} + 1) E_\gamma \left[B^i(b(n)) \right] \quad (31)$$

where

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

Consider now $R^1(n)$. 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 $\bar{X}^1(j)$ the jth combined block. Let $\bar{b}^1$ denote the total number of combined blocks up to block b. Each $\bar{X}^1$ thus consists of two SBs: $\bar{X}^1_1$ that contains the states visited from beginning of $\bar{X}^1$ (empty if the first state is γ^1) to the state right before hitting γ^1, and SB $\bar{X}^1_2$ that contains the rest of $\bar{X}^1$ (a random number of regenerative cycles).

Since a block $\bar{X}^1$ starts after discontinuity in playing the best arm, $\bar{b}^1(n)$ is less than or equal to total number of completed blocks in which the best arm is not played up to time n. Thus

$$E_\gamma[\bar{b}^1(n)] \le \sum_{i>1} E_\gamma[B^i(b(n))]. \tag{32}$$

We rewrite $R^1(n)$ in the following form:

$$\begin{aligned} R^1(n) = &\sum_{y\in S^1} r^1_y E_\gamma \left[\sum_{j=1}^{\bar{b}^1(n)} \sum_{X^1_t \in \bar{X}^1_2(j)} I(X^1_t = y) \right] && (33)\\ &- \sum_{y\in S^1} r^1_y \pi^1_y E_\gamma \left[\sum_{j=1}^{\bar{b}^1(n)} |\bar{X}^1_2(j)| \right] && (34)\\ &+ \sum_{y\in S^1} r^1_y E_\gamma \left[\sum_{j=1}^{\bar{b}^1(n)} \sum_{X^1_t \in \bar{X}^1_1(j)} I(X^1_t = y) \right] && (35)\\ &- \sum_{y\in S^1} r^1_y \pi^1_y E_\gamma \left[\sum_{j=1}^{\bar{b}^1(n)} |\bar{X}^1_1(j)| \right] && (36)\\ &> 0 - \mu^1 \Omega^1_{\max} \sum_{i>1} E_\gamma[B^i(b(n))] && (37) \end{aligned}$$

where the last inequality is obtained by noting the difference between (33) and (34) is zero by Lemma 15, using positivity of rewards to lower bound (35) by 0, and (32) to upper bound (36). Combining this with (28) and (31), we can thus obtain a logarithmic upper bound on $-Z_\gamma(n)$. Finally, we have

$$\begin{aligned} &\mu^1 E_\gamma[n - T(n)] - E_\gamma[\sum_{t=T(n)+1}^{n} r^{\alpha(t)}_{x_{\alpha(t)}}] \\ &\le \mu^1 \left(\frac{1}{\pi_{\min}} + \max_{i\in\{1,\ldots,K\}} \Omega^i_{\max} + 1 \right). \end{aligned} \tag{38}$$

Therefore, we have obtained the stated logarithmic bound for (29). Note that this bound does not depend on γ, and therefore is also an upper bound for $R(n)$, completing the proof.

APPENDIX E
PROOF OF THEOREM 4

A list of notations used in the proof (in addition to the ones used in Section V) is summarized as follows.

1) $T^{i,j}(t)$: 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.
2) $O(b)$: the set of arms that are *free* to be selected by some player i upon its completion of the bth 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 $t(b)$.

Before proving Theorem 3, we state the following lemmas which will be used to prove Theorem 3.

Lemma 16: 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 $L \ge 112 S^2_{\max} r^2_{\max} \hat{\pi}^2_{\max}/\epsilon_{\min}$, for any suboptimal arm i and optimal arm j, we have

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

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

Proof: Result is obtained by following steps similar to the proof of Lemma 14. ■

Lemma 17: Assume Condition 1 holds and all arms are restless. Under RCA-M with a constant $L \ge 112 S^2_{\max} r^2_{\max} \hat{\pi}^2_{\max}/\epsilon_{\min}$, we have

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

where

$$C_{i,j} = \frac{(|S^i| + |S^j|)\beta}{\pi_{\min}}, \quad \beta = \sum_{t=1}^{\infty} t^{-2}$$

$$D_i = \left(\frac{1}{\pi^i_{\min}} + \Omega^i_{\max} + 1 \right).$$

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

$$\begin{aligned} B^i(b) &= 1 + \sum_{m=K+1}^{b} I(\tilde{\alpha}(m) = i) \\ &\le l + \sum_{m=K+1}^{b} I(\tilde{\alpha}(m) = i, B^i(m-1) \ge l). \end{aligned} \tag{39}$$

Consider any sample path ω and the following sets:

$$\begin{aligned} E = \bigcup_{j=1}^{M} \Big\{ \omega : &\, g^j_{t_2(m-1), T^j_2(t_2(m-1))}(\omega) \\ &\le g^i_{t_2(m-1), T^i_2(t_2(m-1))}(\omega) \Big\} \end{aligned}$$

and

$$\begin{aligned} E^C = \bigcap_{j=1}^{M} \Big\{ \omega : &\, g^j_{t_2(m-1), T^j_2(t_2(m-1))}(\omega) \\ &> g^i_{t_2(m-1), T^i_2(t_2(m-1))}(\omega) \Big\}. \end{aligned}$$

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

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

Therefore, continuing from (39)

$$\begin{aligned} B^i(b) &\leq l + \sum_{j=1}^{M} \sum_{m=K+1}^{b} I\left(g^j_{t_2(m-1),T_2^j(t_2(m-1))}\right. \\ &\qquad \left.\leq g^i_{t_2(m-1),T_2^i(t_2(m-1))}, B^i(m-1) \geq l\right) \\ &\leq l + \sum_{j=1}^{M} \sum_{m=K+1}^{b} I\left(\min_{1 \leq w \leq t_2(m-1)} g^j_{t_2(m-1),w}\right. \\ &\qquad \left.\leq \max_{t_2(l) \leq w_i \leq t_2(m-1)} g^i_{t_2(m-1),w_i}\right) \\ &\leq l + \sum_{j=1}^{M} \sum_{m=K+1}^{b} \sum_{w=1}^{t_2(m-1)} \sum_{w_i=t_2(l)}^{t_2(m-1)} I(g^j_{t_2(m),w} \leq g^i_{t_2(m),w_i}) \end{aligned} \tag{40}$$

$$\leq l + M \sum_{j=1}^{M} \sum_{t=1}^{t_2(b)} \sum_{w=1}^{t-1} \sum_{w_i=l}^{t-1} I(g^j_{t,w} \leq g^i_{t,w_i}) \tag{41}$$

where $g^i_{t,w} = \bar{r}^i(w) + c_{t,w}$, and we have assumed that the index value of an arm remains the same between two updates.

The inequality in (41) follows from the facts that the second outer sum in (41) is over time, while the second outer sum in (40) 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 14 to get

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

for all suboptimal arms. Therefore

$$E[B^i(b(n))] \leq \frac{4L \ln n}{(\mu^M - \mu^i)^2} + 1 + M \sum_{j=1}^{M} C_{i,j} \tag{42}$$

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

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

Thus

$$\begin{aligned} \sum_{i>M} (\mu^1 - \mu^i) E[T^i(T(n))] &\leq 4L \sum_{i>M} \frac{(\mu^1 - \mu^i) D_i \ln n}{(\mu^M - \mu^i)^2} \\ &+ \sum_{i>M} (\mu^1 - \mu^i) D_i \left(1 + M \sum_{j=1}^{M} C_{i,j}\right). \end{aligned} \tag{43}$$

■

Now, we give the proof of Theorem 3. Assume that the states which determine the regenerative sample paths are given *a priori* by $\gamma = [\gamma^1, \ldots, \gamma^K]$. 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:

$$\begin{aligned} &R_\gamma(n) \\ &= \sum_{j=1}^{M} \mu^j E_\gamma[T(n)] - E_\gamma\left[\sum_{t=1}^{T(n)} \sum_{\alpha(t) \in A(t)} r^{\alpha(t)}_{x_{\alpha(t)}}\right] \\ &+ \sum_{j=1}^{M} \mu^j E_\gamma[n - T(n)] - E_\gamma\left[\sum_{t=T(n)+1}^{n} \sum_{\alpha(t) \in A(t)} r^{\alpha(t)}_{x_{\alpha(t)}}\right] \\ &= \left\{\sum_{j=1}^{M} \mu^j E_\gamma[T(n)] - \sum_{i=1}^{K} \mu^i E_\gamma[T^i(T(n))]\right\} - Z_\gamma(n) \qquad (44) \\ &+ \sum_{j=1}^{M} \mu^j E_\gamma[n - T(n)] - E_\gamma\left[\sum_{t=T(n)+1}^{n} \sum_{\alpha(t) \in A(t)} r^{\alpha(t)}_{x_{\alpha(t)}}\right] \qquad (45) \end{aligned}$$

where for notational convenience, we have used

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

We have

$$\begin{aligned} &\sum_{j=1}^{M} \mu^j E_\gamma[T(n)] - \sum_{i=1}^{K} \mu^i E_\gamma\left[T^i(T(n))\right] \\ &= \sum_{j=1}^{M} \sum_{i=1}^{K} \mu^j E_\gamma[T^{i,j}(T(n))] - \sum_{j=1}^{M} \sum_{i=1}^{K} \mu^i E_\gamma[T^{i,j}(T(n))] \\ &= \sum_{j=1}^{M} \sum_{i>M} (\mu^j - \mu^i) E_\gamma[T^{i,j}(T(n))] \\ &\leq \sum_{i>M} (\mu^1 - \mu^i) E_\gamma[T^i(T(n))]. \end{aligned} \tag{46}$$

Since we can bound (46), i.e., the difference in the brackets in (44) logarithmically using Lemma 17, it remains to bound $Z_\gamma(n)$ and the difference in (45). We have

$$\begin{aligned}
&Z_\gamma(n)\\
&\geq \sum_{i=1}^{M}\sum_{y\in S^i} r_y^i E_\gamma\left[\sum_{b=1}^{B^i(b(n))}\sum_{X_t^i\in X^i(b)} I(X_t^i=y)\right]\\
&+\sum_{i>M}\sum_{y\in S^i} r_y^i E_\gamma\left[\sum_{b=1}^{B^i(b(n))}\sum_{X_t^i\in X_2^i(b)} I(X_t^i=y)\right] \qquad (47)\\
&-\sum_{i=1}^{M}\mu^i E_\gamma\left[T^i(T(n))\right]\\
&-\sum_{i>M}\mu^i\left(\frac{1}{\pi_{\gamma^i}^i}+\Omega_{\max}^i+1\right)E_\gamma\left[B^i(b(n))\right]
\end{aligned}$$

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 17. Applying Lemma 15 to (47), we get

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

Rearranging terms, we get

$$Z_\gamma(n)\geq R^*(n)-\sum_{i>M}\mu^i(\Omega_{\max}^i+1)E_\gamma\left[B^i(b(n))\right] \qquad (48)$$

where

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

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 $i\in\{1,\ldots,M\}$, we combine *consecutive* blocks in which arm i is played into a single *combined* block, and denote by $\bar{X}^i(j)$ the jth 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 SBs: $\bar{X}_1^i$ 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 SB $\bar{X}_2^i$ 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 ith best arm, $\bar{b}^i(n)$ is less than or equal to the total number of discontinuities of play of the ith best arm up to time n. At the same time, the total number of discontinuities of play of the ith 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

$$E_\gamma[\bar{b}^i(n)]\leq\sum_{k>M}E_\gamma[B^k(b(n))]. \qquad (49)$$

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

$$\begin{aligned}
&R^*(n)\\
&=\sum_{i=1}^{M}\sum_{y\in S^i} r_y^i E_\gamma\left[\sum_{b=1}^{\bar{b}^i(n)}\sum_{X_t^i\in\bar{X}_2^i(b)} I(X_t^i=y)\right] \qquad (50)\\
&-\sum_{i=1}^{M}\sum_{y\in S^i} r_y^i\pi_y^i E_\gamma\left[\sum_{b=1}^{\bar{b}^i(n)}|\bar{X}_2^i(b)|\right] \qquad (51)\\
&+\sum_{i=1}^{M}\sum_{y\in S^i} r_y^i E_\gamma\left[\sum_{b=1}^{\bar{b}^i(n)}\sum_{X_t^i\in\bar{X}_1^i(b)} I(X_t^i=y)\right] \qquad (52)\\
&-\sum_{i=1}^{M}\sum_{y\in S^i} r_y^i\pi_y^i E_\gamma\left[\sum_{b=1}^{\bar{b}^i(n)}|\bar{X}_1^i(b)|\right] \qquad (53)\\
&>0-\sum_{i=1}^{M}\mu^i\Omega_{\max}^i\sum_{k>M}E_\gamma[B^k(b(n))] \qquad (54)
\end{aligned}$$

where the last inequality is obtained by noting the difference between (50) and (51) is zero by Lemma 15, using positivity of rewards to lower bound (52) by 0, and (49) to upper bound (53). Combining this with (42) and (48), we can obtain a logarithmic upper bound on $-Z_\gamma(n)$ by the following steps:

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

We also have

$$\begin{aligned}
&\sum_{j=1}^{M}\mu^j E_\gamma[n-T(n)]-E_\gamma\left[\sum_{t=T(n)+1}^{n}\sum_{\alpha(t)\in A(t)} r_{x_{\alpha(t)}}^{\alpha(t)}\right]\\
&\leq\sum_{j=1}^{M}\mu^j E_\gamma[n-T(n)]\\
&\quad=\sum_{j=1}^{M}\mu^j\left(\frac{1}{\pi_{\min}}+\max_{i\in\{1,\ldots,K\}}\Omega_{\max}^i+1\right). \qquad (55)
\end{aligned}$$

Finally, combining the aforementioned results as well as Lemma 17, we get

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

Therefore, we have obtained the stated logarithmic bound for (44). Note that this bound does not depend on γ, and therefore is also an upper bound for $R(n)$, completing the proof.

Appendix F
Proof of Lemma 7

Let $\tilde{\alpha}_j(b)$ be the arm selected by player j in its bth block. Assume that player j has completed the b'th block.

$$\begin{aligned}
&B^{i,j}(b')\\
&=1+\sum_{b=K+1}^{b'}I(\tilde{\alpha}_j(b)=i)\\
&\le l+\sum_{b=K+1}^{b'}I(\tilde{\alpha}_j(b)=i,B^{i,j}(b-1)\ge l)\\
&\le l+\sum_{b=K+1}^{b'}\sum_{k=1}^{M}I\left(g^{k,j}_{b-1,B^{k,j}(b-1)}\le g^{i,j}_{b-1,B^{i,j}(b-1)},\right.\\
&\qquad\left.B^{i,j}(b-1)\ge l\right)\\
&\le l+\sum_{k=1}^{M}\sum_{b=K+1}^{b'}I\left(\min_{0<s_k<b}g^{k,j}_{b-1,s_k}\le\max_{l\le s_i<b}g^{i,j}_{b-1,s_i}\right)\\
&\le l+\sum_{k=1}^{M}\sum_{b=1}^{b'}\sum_{s_k=1}^{b-1}\sum_{s_i=l}^{b-1}I\left(g^{k,j}_{b-1,s_k}\le g^{i,j}_{b-1,s_i}\right).
\end{aligned}\tag{56}$$

Then, proceeding from (56) the same way as in the proof of Lemma 13, but using a Chernoff–Hoeffding bound for i.i.d. process instead of the large deviation bound for a Markov chain, for $l=\left\lceil\frac{8\ln b'}{(\mu^M-\mu^i)^2}\right\rceil$, we have

$$E[B^{i,j}(b_j(n))|b_j(n)=b']\le\frac{8\ln b'}{(\mu^M-\mu^i)^2}+1+M\beta.$$

Thus, we have

$$E[B^{i,j}(b_j(n))]\le\frac{8\ln n}{(\mu^M-\mu^i)^2}+1+M\beta.$$

Appendix G
Proof of Lemma 9

The event that the index of any one of the optimal arms calculated by player j is in wrong order at v_jth block of player j is included in the event

$$E_j(v_j)=\bigcup_{a=1}^{M}\bigcup_{c=a+1}^{K}\{g^{a,j}_{v_j,B^{a,j}(v_j)}\le g^{c,j}_{v_j,B^{c,j}(v_j)}\}.$$

Let $\mathcal{B}_{i,j}(b)$ denote the set of blocks that player i is in, during the bth block of player j. The event that the index of any one of the optimal arms calculated by player $i\neq j$ is in wrong order during any interval at v_jth block of player j is included in the event

$$E_i(v_j)=\bigcup_{v_i\in\mathcal{B}_{i,j}(v_j)}\bigcup_{a=1}^{M}\bigcup_{c=a+1}^{K}\{g^{a,i}_{v_i,B^{a,i}(v_i)}\le g^{c,i}_{v_i,B^{c,i}(v_i)}\}.$$

The event that the index of any one of the optimal arms calculated by any player is in wrong order during any interval at v_jth block of player j is included in the event

$$\bigcup_{i=1}^{M}E_i(v_j).$$

Let $\tilde{B}^j(b_j(n))$ be the number of completed blocks of player j up to time n in which there is at least one player who has a wrong order for an index of some optimal arm during some part of a block of player j. Then

$$\tilde{B}^j(b_j(n))=\sum_{v_j=1}^{b_j(n)}I\left(\bigcup_{i=1}^{M}E_i(v_j)\right)\le\sum_{i=1}^{M}\sum_{v_j=1}^{b_j(n)}I(E_i(v_j)).$$

Using union bound, we have

$$\begin{aligned}
&\sum_{v_j=1}^{b_j(n)}I(E_j(v_j))\\
&\le\sum_{v_j=1}^{b_j(n)}\sum_{a=1}^{M}\sum_{c=a+1}^{K}I(g^{a,j}_{v_j,B^{a,j}(v_j)}\le g^{c,j}_{v_j,B^{c,j}(v_j)})
\end{aligned}\tag{57}$$

and

$$\sum_{v_j=1}^{b_j(n)} I(E_i(v_j)) \le \sum_{v_j=1}^{b_j(n)} \sum_{v_i \in \mathcal{B}_{i,j}(v_j)} \sum_{a=1}^{M} \sum_{c=a+1}^{K} I(g^{a,i}_{v_i,B^{a,i}(v_i)} \le g^{c,i}_{v_i,B^{c,i}(v_i)}). \quad (58)$$

Proceeding from (57) the same way as in the proof of Lemma 7, we have

$$E\left[\sum_{v_j=1}^{b_j(n)} I(E_j(v_j))\right] \le \sum_{a=1}^{M} \sum_{c=a+1}^{K} \left(\frac{8\ln n}{(\mu^a - \mu^c)^2} + 1 + \beta\right). \quad (59)$$

In (58), for each block of player j, the second sum counts the number of blocks of player i which intersects with that block of player j. This is less than or equal to counting the number of blocks of j which intersects with a block of i for blocks $1, \ldots, b_i(n)+1$ of i. We consider block $b_i(n)+1$ of i because it may intersect with completed blocks of player j up to $b_j(n)$. Thus, we have

$$\sum_{v_j=1}^{b_j(n)} I(E_i(v_j)) \le \sum_{v_i=1}^{b_i(n)+1} \sum_{v_j \in \mathcal{B}_{j,i}(v_i)} \sum_{a=1}^{M} \sum_{c=a+1}^{K} I(g^{a,i}_{v_i,B^{a,i}(v_i)} \le g^{c,i}_{v_i,B^{c,i}(v_i)})$$

with probability 1. Taking the conditional expectation, we get

$$\begin{aligned} &E\left[\sum_{v_j=1}^{b_j(n)} I(E_i(v_j)) \middle| |\mathcal{B}_{j,i}(1)| = n_1, \ldots \right. \\ &\left. |\mathcal{B}_{j,i}(b_i(n)+1)| = n_{b_i(n)+1}\right] \\ &= E\left[\sum_{v_i=1}^{b_i(n)+1} \sum_{a=1}^{M} \sum_{c=a+1}^{K} n_{v_i} I(g^{a,i}_{v_i,B^{a,i}(v_i)} \le g^{c,i}_{v_i,B^{c,i}(v_i)})\right] \\ &\le \max_{v_i=1:b_i(n)+1} E\left[\sum_{v_i=1,a=1,c=a+1}^{b_i(n)+1,M,K} I(g^{a,i}_{v_i,B^{a,i}(v_i)} \le g^{c,i}_{v_i,B^{c,i}(v_i)})\right]. \end{aligned}$$

Using the aforementioned result and following the same approach as in (59), we have

$$\begin{aligned} E\left[\sum_{v_j=1}^{b_j(n)} I(E_i(v_j))\right] &\le E\left[\max_{v_i=1:b_i(n)+1} |\mathcal{B}_{j,i}(v_i)|\right] \\ &\times \sum_{a=1}^{M} \sum_{c=a+1}^{K} \left(\frac{8\ln n}{(\mu^a - \mu^c)^2} + 1 + \beta\right). \quad (60) \end{aligned}$$

The next step is to bound $E\left[\max_{v_i=1:b_i(n)+1} |\mathcal{B}_{j,i}(v_i)|\right]$. Let $l_i(v_i)$ be the length of the v_ith block of player i. Clearly, we have $|\mathcal{B}_{j,i}(v_i)| \le l_i(v_i)$ with probability 1. Therefore, $E\left[\max_{v_i=1:b_i(n)+1} |\mathcal{B}_{j,i}(v_i)|\right] \le E\left[\max_{v_i=1:b_i(n)+1} l_i(v_i)\right]$. Note that the random variables $l_i(v_i), v_i = 1 : b_i(n)+1$ are independent due to Markov property but not necessarily identically distributed since player i might play different arms at different blocks.

Let $p^k_{xy}(t)$ denote the t step transition probability from state x to y of arm k. Since all arms are ergodic, there exists $N > 0$ such that $p^k_{xy}(N) > 0$, for all $k \in \mathcal{K}$, $x, y \in S^k$. Let $p^* = \min_{k\in\mathcal{K},\, x,y\in S^k} p^k_{xy}(N)$. We define a geometric random variable $l_{\max}$ with distribution $P(l_{\max} = 2Nz) = (1-p^*)^{z-1}p^*$, $z = 1, 2, \ldots$. It is easy to see that $P(l_i(v_i) \le z) \ge P(l_{\max} \le z)$, $z = 1, 2, \ldots$. Consider an i.i.d. set of random variables $\{l_{\max}(1), \ldots, l_{\max}(b_i(n)+1)\}$ where each $l_{\max}(v)$ has the same distribution as $l_{\max}$. Since $l_i(.)$ and $l_{\max}(.)$ are nonnegative random variables, we have

$$\begin{aligned} &E\left[\max_{v_i=1:b_i(n)+1} l_i(v_i) \middle| b_i(n) = b\right] \\ &= \sum_{z=0}^{\infty} P\left(\max_{v_i=1:b+1} l_i(v_i) > z\right) \\ &= \sum_{z=0}^{\infty} \left(1 - \prod_{v_i=1}^{b+1} P(l_i(v_i) \le z)\right) \\ &\le \sum_{z=0}^{\infty} \left(1 - \prod_{v_i=1}^{b+1} P(l_{\max}(v_i) \le z)\right) \\ &= E\left[\max_{v_i=1:b_i(n)+1} l_{\max}(v_i) \middle| b_i(n) = b\right]. \end{aligned}$$

Finally

$$\begin{aligned} &E\left[\max_{v_i=1:b_i(n)+1} l_{\max}(v_i) \middle| b_i(n) = b\right] \\ &= \sum_{z=0}^{\infty} \left(1 - P(l_{\max} \le z)^{b+1}\right) \\ &= 2N \sum_{z=0}^{\infty} \left(1 - P(l_{\max} \le 2Nz)^{b+1}\right) \\ &< 2N\left(1 + \frac{1}{\lambda}\sum_{l=1}^{b+1}\frac{1}{l}\right) \quad (61) \\ &\le 2N\left(1 + \frac{1}{\lambda}(\ln n + 1)\right) \quad (62) \end{aligned}$$

where $\lambda = \ln\left(\frac{1}{1-p^*}\right)$, (61) follows from [26, eq. (4)], and (62) follows from $b_i(n) + 1 \le \log n$ with probability 1. Using the aforementioned results on (60), we get

$$\begin{aligned} E\left[\sum_{v_j=1}^{b_j(n)} I(E_i(v_j))\right] &< 2N\left(1 + \frac{1}{\lambda}(\ln n + 1)\right) \\ &\times \sum_{a=1}^{M} \sum_{c=a+1}^{K} \left(\frac{8\ln n}{(\mu^a - \mu^c)^2} + 1 + \beta\right). \quad (63) \end{aligned}$$

Using (59) and (63), we have

$$\begin{aligned} E[\tilde{B}^j(b_j(n))] &\leq E\left[\sum_{i=1}^{M}\sum_{v_j=1}^{b_j(n)} I(E_i(v_j))\right] \\ &< \left[2N(M-1)\left(1+\frac{1}{\lambda}(\ln n+1)\right)+1\right] \\ &\quad \times \sum_{a=1}^{M}\sum_{c=a+1}^{K}\left(\frac{8\ln n}{(\mu^a-\mu^c)^2}+1+\beta\right). \end{aligned}$$

Thus, we have

$$\begin{aligned} E[B'(n)] < M &\left[2N(M-1)\left(1+\frac{1}{\lambda}(\ln n+1)\right)+1\right] \\ &\times \sum_{a=1}^{M}\sum_{c=a+1}^{K}\left(\frac{8\ln n}{(\mu^a-\mu^c)^2}+1+\beta\right). \end{aligned} \quad (64)$$

Appendix H
Proof of Lemma 10

Let b be a block in which all players know the correct order of the M-best channels and $b-1$ be a block in which there exists at least one player whose order of indices for M-best channels are different than the order of the mean rewards. We call such an event a transition from a bad state to a good state. Then, by Lemma 7, the expected number of blocks needed to settle to an orthogonal configuration after block b is bounded by O_B. Since the expected number of such transitions is $E[B'(n)]$, we have $E[H(n)] \leq O_B E[B'(n)]$.

References

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

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

[3] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, 2002.

[4] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, pp. 4–22, 1985.

[5] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part I: I.I.D. rewards," *IEEE Trans. Autom. Control*, vol. 32, no. 11, pp. 968–976, Nov. 1987.

[6] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part II: Markovian rewards," *IEEE Trans. Autom. Control*, vol. 32, no. 11, pp. 977–982, Nov. 1987.

[7] R. Agrawal, "Sample mean based index policies with $O(\log(n))$ regret for the multi-armed bandit problem," *Adv. Appl. Probabil.*, vol. 27, no. 4, pp. 1054–1078, Dec. 1995.

[8] A. Garivier and O. Cappe, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. JMLR Workshop Conf.*, 2011, vol. 19, pp. 359–376.

[9] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proc. 21st Annu. Conf. Learn. Theory*, Jul. 2008, pp. 355–366.

[10] C. Tekin and M. Liu, "Online algorithms for the multi-armed bandit problem with Markovian rewards," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2010, pp. 1675–1682.

[11] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," in *Proc. 30th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2011, pp. 2462–2470.

[12] H. Liu, K. Liu, and Q. Zhao, Learning in a changing world: Non-Bayesian restless multi-armed bandit Univ. California Davis, Davis, 2010.

[13] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.

[14] A. Anandkumar, N. Michael, A. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 731–745, Apr. 2011.

[15] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *IEEE Symp. Dyn. Spectrum Access Netw. (DySPAN)*, Apr. 2010, pp. 1–9.

[16] J. Gittins, "Bandit processes and dynamic allocation indices," *J. Roy. Statist. Soc.*, vol. 41, no. 2, pp. 148–177, 1979.

[17] P. Whittle, , J. Gani, Ed., "Restless bandits: Activity allocation in a changing world," in *A Celebration of Applied Probability*. Sheffield, U.K.: Applied Probability Trust, 1988, vol. 25A, pp. 287–298.

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

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

[20] J. Y. Audibert and R. M. Szepesvári, "Exploration-exploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Comput. Sci.*, vol. 410, no. 19, pp. 1876–1902, 2009.

[21] C. Tekin and M. Liu, "Adaptive learning of uncontrolled restless bandits with logarithmic regret," in *Proc. 49th Annu. Allerton Conf. Commun., Control, Comput.*, Sep. 2011, pp. 983–990.

[22] C. Papadimitriou and J. Tsitsiklis, "The complexity of optimal queuing network control," *Math. Oper. Res.*, vol. 24, no. 2, pp. 293–305, May 1999.

[23] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao, "The non-Bayesian restless multi-armed bandit: A case of near-logarithmic regret," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 2940–2943.

[24] S. Guha, K. Mungala, and P. Shi, "Approximation algorithms for restless bandit problems," in *Proc. 20th ACM-SIAM Symp. Discr. Algorithms*, 2009, pp. 28–37.

[25] C. Tekin and M. Liu, "Approximately optimal adaptive learning in opportunistic spectrum access," presented at the presented at the IEEE INFOCOM, Orlando, FL, Mar. 2012.

[26] B. Eisenberg, "On the expectation of the maximum of i.i.d. geometric random variables," *Statist. Probab. Lett.*, vol. 78, pp. 135–143, 2008.

Cem Tekin (S'11) received his B.Sc. degree in electrical and electronics engineering in 2008 from the Middle East Technical University, Ankara, Turkey. He received his M.S.E. in electrical engineering: systems in 2010 and his M.S. in mathematics in 2011 both from the University of Michigan, Ann Arbor, Michigan. He is currently a Ph.D. candidate in the Department of Electrical Engineering and Computer Science, University of Michigan. His research interests include online learning algorithms, stochastic optimization, multi armed bandit problems, multiuser systems, game theory.

Mingyan Liu (M'00–SM'11) received her B.Sc. degree in electrical engineering in 1995 from the Nanjing University of Aero. and Astro., Nanjing, China, M.Sc. degree in systems engineering and Ph.D. Degree in electrical engineering from the University of Maryland, College Park, in 1997 and 2000, respectively. She joined the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, in September 2000, where she is currently an Associate Professor. Her research interests are in optimal resource allocation, performance modeling and analysis, and energy efficient design of wireless, mobile ad hoc, and sensor networks. She is the recipient of the 2002 NSF CAREER Award, the University of Michigan Elizabeth C. Crosby Research Award in 2003, and the 2010 EECS Department Outstanding Achievement Award. She serves on the editorial board of IEEE/ACM Trans. Networking, IEEE Trans. Mobile Computing, and ACM Trans. Sensor Networks.