Learning in A Changing World: Non-Bayesian Restless Multi-Armed Bandit

Haoyang Liu, Keqin Liu, Qing Zhao
University of California, Davis, CA 95616
{liu, kqliu, qzhao}@ucdavis.edu

Abstract

We consider the restless multi-armed bandit (RMAB) problem with unknown dynamics. In this problem, at each time, a player chooses K out of N $(N > K)$ arms to play. The state of each arm determines the reward when the arm is played and transits according to Markovian rules no matter the arm is engaged or passive. The Markovian dynamics of the arms are unknown to the player. The objective is to maximize the long-term reward by designing an optimal arm selection policy. The performance of a policy is measured by regret, defined as the reward loss with respect to the case where the player knows which K arms are the most rewarding and always plays these K best arms. We construct a policy, referred to as Restless Upper Confidence Bound (RUCB), that achieves a regret with logarithmic order of time when an arbitrary nontrivial bound on certain system parameters is known. When no knowledge about the system is available, we extend the RUCB policy to achieve a regret arbitrarily close to the logarithmic order. In both cases, the system achieves the maximum mean reward offered by the K best arms. Potential applications of these results include cognitive radio networks, opportunistic communications in unknown fading environments, and financial investment.

Index Terms

Restless multi-armed bandit, non-Bayesian formulation, regret, logarithmic order

I. Introduction

The Restless Multi-Armed Bandit (RMAB) problem is a generalization of the classic Multi-Armed Bandit (MAB) problem. In the classic MAB, there are N independent arms and a single

[0]This work was supported by the Army Research Office under Grant W911NF-08-1-0467 and by the National Science Foundation under Grant CCF-0830685.

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Learning in A Changing World: Non-Bayesian Restless Multi-Armed Bandit	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Department of Electrical and Computer Engineering,Davis,CA,95616**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We consider the restless multi-armed bandit (RMAB) problem with unknown dynamics. In this problem, at each time, a player chooses K out of N (N > K) arms to play. The state of each arm determines the reward when the arm is played and transits according to Markovian rules no matter the arm is engaged or passive. The Markovian dynamics of the arms are unknown to the player. The objective is to maximize the long-term reward by designing an optimal arm selection policy. The performance of a policy is measured by regret, defined as the reward loss with respect to the case where the player knows which K arms are the most rewarding and always plays these K best arms. We construct a policy, referred to as Restless Upper Confidence Bound (RUCB), that achieves a regret with logarithmic order of time when an arbitrary nontrivial bound on certain system parameters is known. When no knowledge about the system is available, we extend the RUCB policy to achieve a regret arbitrarily close to the logarithmic order. In both cases, the system achieves the maximum mean reward offered by the K best arms. Potential applications of these results include cognitive radio networks, opportunistic communications in unknown fading environments, and financial investment.

15. SUBJECT TERMS
Restless multi-armed bandit, non-Bayesian formulation, regret, logarithmic order

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**17**	

player. At each time, the player chooses one arm to play and receives certain amount of reward. The reward (*i.e.*, the state) of each arm evolves as an i.i.d. process over successive plays. The reward distribution of each arm is unknown to the player. The objective is to maximize the long-term reward by designing an optimal arm selection policy. This problem involves the well-known dilemma between exploitation and exploration. For exploitation, the player tends to select the arm suggested by past reward observations as the best. For exploration, the player selects an arm to learn its reward statistics. Under the non-Bayesian formulation, the performance measure of an arm selection policy is given by regret, defined as the reward loss compared with the optimal performance in the ideal scenario of a known reward model [1]. Note that in the ideal scenario, the player will always play the arm with the highest mean reward. The essence of the problem is to identify the best arm without engaging other inferior arms too often.

In 1985, Lai and Robbins showed that the minimum regret grows with time in a logarithmic order [1]. A policy was further constructed to achieve the minimum regret (both the logarithmic order and the best leading constant) [1]. In 1987, Anantharam *et al.* extended Lai and Robbins's results to accommodate multiple simultaneous plays [2] and Markovian reward model where the reward of each arm evolves as an unknown Markov process over successive plays and remains frozen when the arm is not played (the so-called rested Markovian reward model) [3]. For both extensions, the minimum regret growth rate has been shown to be logarithmic [2], [3]. There are also several simpler index policies that achieve logarithmic regret for the classic MAB under an i.i.d. reward model [4], [5]. In particular, the index policy—referred to as Upper Confidence Bound 1 (UCB1)—proposed in [5] achieves the logarithmic regret with a uniform bound on the leading constant over time. In [6], UCB1 was extended to the rested Markovian reward model adopted in [3].

A. Restless Multi-Armed Bandit with Unknown Dynamics

Different from the classic MAB, in an RMAB, the state of each arm can change (according to an unknown Markovian rule) even when the arm is *not* played. The unknown state transition matrix when the arm is played can be different from that when it is not played. We consider the general case where K $(K < N)$ arms are simultaneously played at each time. Even with a known model, the RMAB problem has been shown to be P-SPACE hard in general [7].

In this paper, we address the RMAB problem with unknown Markovian dynamics. Similar to

the classic MAB, we measure the performance of a policy by regret, defined as the reward loss compared to the case when the player knows which K arms are most rewarding and always plays the K best arms. We show that for RMAB, logarithmic regret can also be achieved as in the classic MAB. Specifically, we construct a policy that achieves logarithmic regret when an arbitrary nontrivial bound on certain system parameters is known. When no knowledge about the system is available, we show that a variation of the policy achieves a regret arbitrarily close to logarithmic order, *i.e.*, the regret has order $f(t)\log(t)$ for any increasing function $f(t)$ with $f(t) \to \infty$ as time $t \to \infty$. In both cases, the proposed policy achieves the maximum mean reward offered by the K best arms.

Referred to as the Restless Upper Confidence Bound (RUCB), the proposed policy borrows the basic index form of the UCB-1 policy developed in [5] for the classic MAB under i.i.d. reward models. To handle the restless nature of the problem, the basic structure of the proposed RUCB policy is fundamentally different from that of UCB-1. Specifically, the basic structure of RUCB consists of interleaving exploitation and exploration epochs with carefully controlled lengths to bound the frequency of arm switching and balance the tradeoff between exploitation and exploration. Another novelty of this paper is a general technique in choosing policy parameters whose value may have to depend on the range of certain system parameters. We show that by letting these policy parameters grow with time (rather than fixed *a priori*), one can get around with the dependency of the policy parameters on system parameters and achieve a regret order arbitrarily close to logarithmic without any knowledge about the system.

We point out that the definition of regret adopted in this paper, while similar to that used for the classic MAB, is a weaker version of its counterpart in the classic MAB. In the classic MAB with either i.i.d. or rested Markovian reward, the optimal policy under known model is to stay with the best arm in terms of the reward mean. For RMAB, however, the optimal policy under known model is no longer given by staying with the arm with the highest mean reward. Defining the regret in terms of this optimal policy would require that a general RMAB with known model be solved and optimal performance analyzed before the regret under unknown model can be approached. Unfortunately, RMAB under known model itself is intractable in general [7]. In this paper, we adopt a weaker definition of regret where the performance is compared with a "partially-informed" genie who knows only which k arms have the highest mean reward instead of the complete system dynamics. This definition of regret leads to a tractable problem, but at

the same time, weaker results. Whether stronger results for a general RMAB under unknown model can be obtained is still open for exploration (see more discussions in Sec. I-C on related work).

B. Applications

The restless multi-armed bandit problem has a broad range of applications. For example, in a cognitive radio network, a secondary user searches among several channels for idle slots that are temporarily unused by primary users. The state of each channel (busy or idle) can be modeled as a two-state Markov chain. At each time, a secondary user chooses one channel to sense and subsequently transmit if the channel is found in the idle state. The objective of the secondary user is to maximize the long-term throughput by designing an optimal channel selection policy without knowing the traffic dynamics of the primary users.

Consider opportunistic transmission over multiple wireless channels with unknown Markovian fading. In each slot, a user senses the fading realization of a selected channel and chooses its transmission power or date rate accordingly. The reward can model energy efficiency (for fixed-rate transmission) or throughput. The objective is to design the optimal channel selection policies under unknown fading dynamics.

Another potential application is financial investment, where a Venture Capital (VC) selects one company to invest at each year. The state (*e.g.*, annual profit) of each company evolves as a Markov chain with the transition matrix depending on whether the company is invested or not. The objective of the VC is to maximize the long-run profit by designing the optimal investment strategy without knowing the market dynamics *a priori*.

The proposed policy for RMAB also provides a basic building block for constructing decentralized policies for MAB with multiple distributed players under a Markovian reward model [8] (Decentralized MAB was first formulated and solved under an i.i.d. reward model in [9]). In the decentralized MAB with Markovian reward, multiple distributed players select arms to play and collide when they select the same arm. Arms are rested, *i.e.*, they do not change states when they are not played. However, from each player's point of view, each arm is restless since its state can be changed by other players. Applying the RUCB policy proposed here to the decentralized rested multi-armed bandit problem leads to the optimal logarithmic order of the regret [8].

C. Related Work

This paper is among the few first attempts on RMAB under uknown models. There are two parallel independent investigations reported in [10] and [11]. In [10], Tekin and Liu adopted the same definition of regret as used in this paper and proposed a policy that achieves logarithmic (weak) regret when certain knowledge about the system parameters is available [10]. The policy proposed in [10] also uses the index form of UCB-1 given in [5], but the structure is different from RUCB proposed in this paper. In [11], a stronger definition of regret is adopted, where regret is defined as reward loss with respect to the optimal performance in the ideal scenario of known reward model. However, the problem can only be solved for a special class of RMAB. Specifically, when arms are governed by stochastically identical two-state Markov chains, a policy was constructed in [11] to achieve a regret with an order arbitrarily close to logarithmic.

The RMAB with known reward model has been extensively studied in the literature. In [12], Whittle proposed a heuristic index policy that generalizes Gittins optimal index policy for the classic MAB with known reward model [13]. Weber showed that Whittle index policy is asymptotically optimal (as the number of arms goes to infinity) under certain conditions [14]. In the finite regime, the optimality of Whittle index policy has been shown for certain special families of RMAB (see, for example, [15]).

II. Problem Formulation

In the RMAB problem, we have one player and N independent arms. At each time, the player can choose K $(K < N)$ arms to play (we focus on $K = 1$ for the simplicity of presentation). Each arm, when played (activated), offers certain amount of reward that models the current state of the arm. Let $s_j(t)$ denote the state of arm j at time t. No matter an arm is played or not, the state of the arm changes according to a Markovian rule. In general, the transition matrices in the active mode and the passive mode are not necessarily the same. The player does not know the transition matrices of the arms. The objective is to choose one arm to play at each time in order to maximize the expected total reward collected in the long run.

Let $\mathcal{S}_j$ denote the state space of arm j. Each arm is assumed to have a finite state space. Different arms can have different state spaces. Let P_j denote the active transition matrix of arm j and Q_j the passive transition matrix. All transition matrices are assumed to be irreducible, aperiodic, and reversible. Let $\vec{\pi}_j = \{\pi_s^j\}_{s\in\mathcal{S}_j}$ denote the stationary distribution of arm j in the

active mode (*i.e.*, under P_j), where π_s^i is the stationary probability (under P_j) that arm j is in state s. The stationary mean reward μ_j is given by $\mu_j = \sum_{s\in\mathcal{S}_j} s\pi_s^j$. Let σ be a permutation of $\{1,\cdots,N\}$ such that

$$\mu_{\sigma(1)} \geq \mu_{\sigma(2)} \geq \mu_{\sigma(3)} \geq \cdots \geq \mu_{\sigma(N)}.$$

A policy Φ is a rule that specifies the arm to play based on the observation history. Let $t_j(n)$ denote the time index of the nth play on arm j, and $T_j(t)$ the total number of plays on arm j by time t. Notice that both $t_j(n)$ and $T_j(t)$ are random variables with distributions determined by the policy Φ. The total reward by time t is given by

$$R(t) = \sum_{j=1}^{N}\sum_{n=1}^{T_j(t)} s_j(t_j(n)). \tag{1}$$

As mentioned in Sec. I, the regret $r_\Phi(t)$ achieved by policy Φ is defined as the reward loss with respect to the case where the player knows which arm has the highest mean reward and always plays this best arm. We thus have

$$r_\Phi(t) = t\mu_{\sigma(1)} - \mathbb{E}_\Phi R(t), \tag{2}$$

where $\mathbb{E}_\Phi$ denotes the expectation with respect to the random process induced by policy Φ. The objective is to minimize the growth rate of the regret.

III. The RUCB Policy

The proposed policy RUCB is based on an epoch structure. We divide the time into disjoint epochs. There are two types of epochs: exploitation epochs and exploration epochs (see an illustration in Fig. 1). In the exploitation epochs, the player calculates indexes of all arms and play the arm with the highest index, which is believed to be the best arm. In the exploration epochs, the player obtains information of all arms by playing them equally many times. The purpose of the exploration epochs is to make decisions in the exploitation epochs sufficiently accurate. As shown in Fig. 1, in the nth exploration epoch, the player plays every arm 4^{n-1} times. At the beginning of the nth exploitation epoch the player calculates index for every arm (see (4) in Fig. 2) and selects the arm with the highest index (denoted as arm a^*). The player keeps playing arm a^* till the end of this epoch that has length $2\times 4^{n-1}$. How the two types of epochs interleave is detailed in Step 2 in Fig. 2. Specifically, whenever sufficiently many

($D \ln t$, see (3)) observations have been obtained from every arm in the exploration epochs, the player is ready to proceed with a new exploitation epoch. Otherwise, another exploration epoch is required to gain more information about each arm. It is also implied in (3) that only logarithmically many plays are spent in the exploration epochs, which is one of the key reasons for the logarithmic regret of RUCB. This also implies that the exploration epochs are much less frequent than the exploitation epochs. Though the exploration epochs can be understood as the "information gathering" phase, and the exploitation epochs as the "information utilization" phase, observations obtained in the exploitation epochs are also used in learning the arm dynamics. This can be seen in Step 3 in Fig. 2. In calculating the indexes using (4), observations from both the exploration and exploitation epochs are used. This is different from the policy in [10], which only uses part of the past observations in calculating indexes. A complete description of the proposed policy is given in Fig. 2.

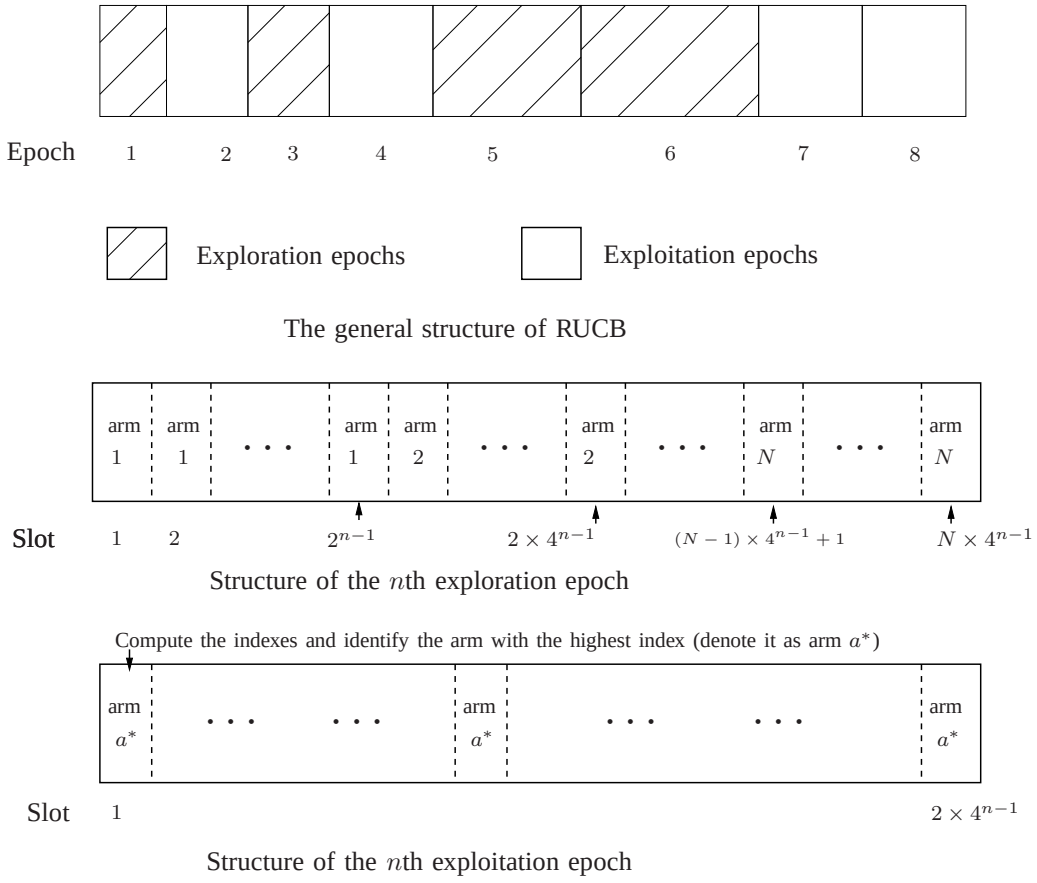

Fig. 1. Epoch structures of RUCB

RUCB

Time is divided into epochs. There are two types of epoch, exploration epoch and exploitation epoch. At the beginning of the nth exploitation epoch, we choose one arm to play for $2 \times 4^{n-1}$ many times. In the nth exploration epoch, we play every arm 4^{n-1} many times. Let $n_O(t)$ denote the number of exploration epochs played by time t and $n_I(t)$ the number of exploitation epochs played by time t.

1. At $t = 1$, we start the first exploration epoch, in which every arm is played once. We set $n_O(N+1) = 1$, $n_I(N+1) = 0$. Then go to Step 2.
2. Let $X_1(t) = (4^{n_O(t)} - 1)/3$ be the time spent on each arm in exploration epochs by time t. Choose D according to (5)(6). If
$$X_1(t) > D \ln t, \tag{3}$$
go to Step 3 (start an exploitation epoch). Otherwise, go to Step 4 (start an exploration epoch).
3. Calculate indexes $d_{i,t}$ for all arms using the formula below:
$$d_{i,t} = \bar{s}_i(t) + \sqrt{\frac{L \ln t}{T_i(t)}}, \tag{4}$$
where t is the current time, $\bar{s}_i(t)$ is the sample mean from arm i by time t, L is chosen according to (5), and $T_i(t)$ is the number of times we have played arm i by time t. Then choose the arm with the highest index and play it for $2 \times 4^{(n_I-1)}$ slots. Increase n_I by one. Go to step 2.
4. Play each arm for $4^{(n_O-1)}$ slots. Increase n_O by one. Go to Step 2.

Fig. 2. RUCB policy

IV. The Logarithmic Regret of RUCB

In this section, we show that the regret achieved by the RUCB policy has a logarithmic order. This is given in the following theorem.

Theorem 1: Assume all arms are modeled as finite state, irreducible, aperiodic, and reversible Markov chains. All the states (rewards) are positive. Let $\pi_{\min} = \min_{s \in \mathcal{S}_i, 1 \le i \le N} \pi_s^i$, $\epsilon_{\max} = \max_{1 \le i \le N} \epsilon_i$, $\epsilon_{\min} = \min_{1 \le i \le N} \epsilon_i$, $s_{\max} = \max_{s \in \mathcal{S}_i, 1 \le i \le N} s$, $s_{\min} = \min_{s \in \mathcal{S}_i, 1 \le i \le N} s$, and $|\mathcal{S}|_{\max} = \max_{1 \le i \le N} |\mathcal{S}_i|$ where ϵ_i is the second largest eigenvalue of P_i. Let $M < N$ denote the number of optimal arms. Set the policy parameters L and D to satisfy the following conditions:

$$L \geq \frac{1}{\epsilon_{\min}} \left(4 \frac{20 s_{\max}^2 |\mathcal{S}|_{\max}^2}{(3 - 2\sqrt{2})} + 10 s_{\max}^2\right), \tag{5}$$

$$D \geq \frac{4L}{(\mu^* - \mu_{\sigma(M+1)})^2}. \tag{6}$$

The regret of RUCB at the end of any epoch can be upper bounded by

$$\begin{aligned} r_\Phi(t) \leq & (\lceil \log_4(\frac{3}{2}(t-M)+1)\rceil) \max_i A_i \\ & + N(\lfloor \log_4(3D\ln t + 1)\rfloor + 1) \max_i A_i \\ & + \sum_i (\mu^* - \mu_i)(\lceil \log_4(\frac{3}{2}(t-M)+1)\rceil 3\frac{|\mathcal{S}_i| + |\mathcal{S}^*|}{\pi_{\min}}(1 + \frac{\epsilon_{\max}\sqrt{L}}{10 s_{\min}}) \\ & + \sum_i (\mu^* - \mu_i)\frac{1}{3}[4(3D\ln t + 1) - 1]), \end{aligned} \tag{7}$$

where $A_i = (\min_{s\in\mathcal{S}_i} \pi_s^i)^{-1} \sum_{s\in\mathcal{S}_i} s$.

Proof: See Appendix A for details. ■

In RUCB, to ensure logarithmic regret order, the policy parameters L and D need to be chosen appropriately. This requires an arbitrary (nontrivial) bound on $s_{\max}^2$, $|\mathcal{S}|_{\max}$, $\epsilon_{\min}$, and $\mu^* - \mu_{\sigma(M+1)}$. In the case where these bounds are unavailable, D and L can be chosen to increase with time to achieve a regret order arbitrarily close to logarithmic order. This is formally stated in the following theorem.

Theorem 2: Assume all arms are modeled as finite state, irreducible, aperiodic, and reversible Markov chains. For any increasing sequence $f(t)$ ($f(t) \to \infty$ as $t \to \infty$), if $L(t)$ and $D(t)$ are chosen such that $L(t) \to \infty$ as $t \to \infty$,$\frac{f(t)}{D(t)} \to \infty$ as $t \to \infty$, and $\frac{D(t)}{L(t)} \to \infty$ as $t \to \infty$, then we have

$$r_\Phi(t) \sim o(f(t)\log(t)). \tag{8}$$

Proof: See Appendix B for details.

■

V. Conclusion

In this paper, we considered the non-Bayesian restless multi-armed bandit problem. We adopted the definition of regret from the classic MAB and developed a policy that achieves logarithmic regret when an arbitrary (nontrivial) bound on certain system parameters is known. When no knowledge about the system is available, we extend the RUCB policy to achieve a regret with an order arbitrarily close to logarithmic.

APPENDIX A. PROOF OF THEOREM 1

We first rewrite the definition of regret as

$$r_\Phi(t) = t\mu^* - \mathbb{E}_\Phi R(t) = \sum_{i=1}^{N}[\mu_i \mathbb{E}[T_i(t)] - \mathbb{E}[\sum_{n=1}^{T_i(t)} s_i(t_i(n))]] + \sum_{i=1}^{N}(\mu^* - \mu_i)\mathbb{E}[T_i(t)]. \quad (9)$$

To show that the regret has a logarithmic order, it is sufficient to show that both terms in (9) have logarithmic orders. The first term in (9) can be understood as the regret caused by arm switching. The second term can be understood as the regret caused by engaging a bad arm. First, we bound the regret caused by arm switching based on the following lemma.

Lemma 1 [3]: Consider an irreducible, aperiodic Markov chain with state space $\mathcal{S}$, matrix of transition probabilities P, an initial distribution $\vec{q}$ which is positive in all states, and stationary distribution $\vec{\pi}$ (π_s is the stationary probability of state s). The state (reward) at time t is denoted by $s(t)$. Let μ denote the mean reward. If we play the chain for an arbitrary time T, then there exists a value $A_P \leq (\min_{s\in\mathcal{S}} \pi_s)^{-1} \sum_{s\in\mathcal{S}} s$ such that $\mathbb{E}[\sum_{t=1}^{T} s(t) - \mu T] \leq A_P$.

Lemma 1 shows that if the player continues to play one arm for time T, the difference between the expected reward and $T\mu$ can be bounded by a constant that is independent of T. This constant is an upper bound for the regret caused by each arm switching. If there are only logarithmically many arm switchings as times goes, the regret caused by arm switching has a logarithmic order. An upper bound on the number of arm switchings is shown below. It is developed by bounding the numbers of the exploration epochs and the exploitation epochs respectively.

For the exploration epochs, by time t, if the player has began to play the $(n+1)$th exploration epoch, we have

$$\frac{1}{3}(4^n - 1) < D\ln t, \quad (10)$$

where $\frac{1}{3}(4^n - 1)$ is the time spent on each arm in the first n exploration epochs.

Consequently the number of the exploration epochs can be bounded by

$$n_O(t) \leq \lfloor \log_4(3D\ln t + 1) \rfloor + 1. \quad (11)$$

By time t, at most $(t - N)$ time slots have been spent on the exploitation epochs. Thus

$$n_I(t) \leq \lceil \log_4(\frac{3}{2}(t - N) + 1) \rceil. \quad (12)$$

Hence an logarithmic upper bound of the first term in (9) is

$$\sum_{i=1}^{N}[\mu_i\mathbb{E}[T_i(t)] - \mathbb{E}[\sum_{n=1}^{T_i(t)} s_i(t_i(n))]] \leq (\lceil\log_4(\frac{3}{2}(t-N)+1)\rceil + N(\lfloor\log_4(3D\ln t+1)\rfloor+1))\max_i A_i, \quad (13)$$

where $A_i = (\min_{s\in\mathcal{S}_i}\pi_s^i)^{-1}\sum_{s\in\mathcal{S}_i} s$.

Next we show that the second term of (9) has a logarithmic order. The approach here is to show that for every bad arm i, $\mathbb{E}[T_i(t)]$ has a logarithmic order. Let $T_{i,O}(t)$ denote the time spent on arm i in the exploration epochs by time t. Let $T_{i,I}(t)$ denote the time spent on arm i in the exploitation epochs by time t. So we have

$$T_i(t) = T_{i,O}(t) + T_{i,I}(t), \quad (14)$$

We will show that both $\mathbb{E}[T_{i,O}(t)]$ and $\mathbb{E}[T_{i,I}(t)]$ have logarithmic orders.

The logarithmic order of $\mathbb{E}[T_{i,O}(t)]$ follows directly from (11), *i.e.*,

$$T_{i,O}(t) \leq \frac{1}{3}[4(3D\ln t+1)-1]. \quad (15)$$

The logarithmic order of $\mathbb{E}[T_{i,I}(t)]$ is established by bounding $\Pr[i,n]$, the probability that arm i is played in the nth exploitation epoch.

Recall that if arm i is selected in the nth exploitation epoch, it will be played for $2\times 4^{(n-1)}$ times. From the upper bound on the number of the exploitation epochs given in (12), we thus have

$$\mathbb{E}[T_{i,I}(t)] \leq \sum_{n=1}^{\lceil\log_4(\frac{3}{2}(t-M)+1)\rceil} 2\times 4^{n-1}\Pr[i,n] \quad (16)$$

$$\leq \sum_{n=1}^{\lceil\log_4(\frac{3}{2}(t-M)+1)\rceil} 3t_n\Pr[i,n], \quad (17)$$

where t_n denote the starting time of the nth exploitation epoch and (17) follows from the fact that $t_n \geq \frac{2}{3}4^{n-1}$. Notice that (17) has only logarithmically many terms, if each term can be bounded by a fixed constant, *i.e.*, if $\Pr[i,n]$ has an order of t_n^{-1}, then the sum has a logarithmic order.

Let $C_{t,w} = \sqrt{(L\ln t/w)}$ denote the second part of the RUCB index. If arm i is played in the nth exploitation epoch, then

$$\exists w < t_n, w_i < t_n, \quad \text{such that} \quad \bar{s}^*(t_n) + C_{t_n,w} \leq \bar{s}_i(t_n) + C_{t_n,w_i}. \quad (18)$$

We thus have

$$\begin{aligned}
\Pr[i,n] &\leq \sum_{w=1}^{t_n-1} \sum_{w_i=D\ln t_n}^{t_n-1} \Pr[\bar{s}^*(t_n) + C_{t_n,w} \leq \bar{s}_i(t_n) + C_{t_n,w_i}] && (19)\\
&\leq \sum_{w=1}^{t_n-1} \sum_{w_i=D\ln t_n}^{t_n-1} (\Pr[\bar{s}^*(t_n) \leq \mu^* - C_{t_n,w}] + \Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n,w_i}] \\
&\qquad + \Pr[\mu^* < \mu_i + 2C_{t_n,w_i}]) && (20)\\
&\leq \sum_{w=1}^{t_n-1} \sum_{w_i=D\ln t_n}^{t_n-1} (\Pr[\bar{s}^*(t_n) \leq \mu^* - C_{t_n,w}] + \Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n,w_i}]), && (21)
\end{aligned}$$

where (21) follows from the fact that $w_i \geq D\ln t_n$.

Next we bound $\Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n,w_i}]$ and $\Pr[\bar{s}^*(t_n) \leq \mu^* - C_{t_n,w}]$. The event $\bar{s}_i(t_n) \geq \mu_i + C_{t_n,w_i}$ is equivalent to

$$w_i\bar{s}_i(t_n) \geq w_i\mu_i + \sqrt{Lw_i\ln t_n}. \tag{22}$$

The inequality (22) is the event that the sample mean from multiple epochs for arm i is too high. This event implies that the sample mean from at least one epoch is significantly higher than the true mean. Notice that the tolerant deviation in (22) is of the form $\sqrt{Lw_i\ln t_n}$. It is convenient if the tolerant deviation for each epoch is of the form $C\sqrt{Lw\ln t_n}$, where w is the number of plays done on one arm in one epoch and C is a constant independent of w. In this way, the tolerant deviations for the sample mean in each epoch and in all the epochs are of similar forms. The possible values for the number of plays in the exploitation epochs are 2×4^n. The possible values of the numbers of plays done on an arm in the exploration epochs are 4^n. Consequently it can be assumed that the player has spent time w_i on arm i by playing the epochs with lengths of $2^{n_1-1}, 2^{n_2-1}, \cdots, 2^{n_K-1}$, with each n_j distinct. Thus $w_i = \sum_{j=1}^{K} 2^{n_j-1}$ and

$$\begin{aligned}
\sqrt{w_i} &= \sum_{j=1}^{K}\left[\sqrt{\sum_{k=1}^{j} 2^{n_k-1}} - \sqrt{\sum_{k=1}^{j-1} 2^{n_k-1}}\right]\\
&= \sum_{j=1}^{K}\left[\sqrt{\sum_{k=1}^{j-1} 2^{n_k-1} + 2^{n_j-1}} - \sqrt{\sum_{k=1}^{j-1} 2^{n_k-1}}\right]\\
&\geq \sum_{j=1}^{K}\left[\sqrt{2^{n_j-1}+2^{n_j-1}} - \sqrt{2^{n_j-1}}\right]\\
&= \sum_{j=1}^{K}(\sqrt{2}-1)\sqrt{2^{n_j-1}}. && (23)
\end{aligned}$$

The tolerant deviation for an continuous period of play with length 2^{n_j-1} is $(\sqrt{2}-1)\sqrt{L\ln t_n 2^{n_j-1}}$. Let $R_i(w)$ denote the reward gained from arm i in a period with length w. An upper bound on $\Pr[w_i\bar{s}_i(t_n) \geq w_i\mu_i + \sqrt{L\ln t_n w_i}]$ is derived below

$$\begin{aligned}
&\Pr \quad [w_i\bar{s}_i(t_n) \geq w_i\mu_i + \sqrt{L\ln t_n w_i}] \\
&\leq \quad \sum_{j=1}^{K} \Pr[R_i(2^{n_j-1}) \geq \mu_i \cdot 2^{n_j-1} + \sqrt{L\ln t_n}\Big[\sqrt{\sum_{k=1}^{j} 2^{n_k-1}} - \sqrt{\sum_{k=1}^{j-1} 2^{n_k-1}}\Big]] \\
&\leq \quad \sum_{j=1}^{K} \Pr[R_i(2^{n_j-1}) \geq \mu_i \cdot 2^{n_j-1} + (\sqrt{2}-1)\sqrt{2^{n_j-1}L\ln t_n}]. \qquad (24)
\end{aligned}$$

The probability $\Pr[R_i(2^{n_j-1}) \geq \mu_i \cdot 2^{n_j-1} + (\sqrt{2}-1)\sqrt{2^{n_j-1}L\ln t_n}]$ is for the event that the sum of reward during a period of time of length 2^{n_j-1} from arm i is significantly deviated from $\mu_i 2^{n_j-1}$. It can be written in terms of the numbers of occurrences of states. Specifically, let $O_s^i(w)$ denote the number of occurrences of state s from arm i in a period with length w, we have

$$\begin{aligned}
&\Pr \quad [R_i(2^{n_j-1}) \geq \mu_i \cdot 2^{n_j-1} + (\sqrt{2}-1)\sqrt{2^{n_j-1}L\ln t_n}] \\
&= \quad \Pr[\sum_{s\in\mathcal{S}_i}(-sO_s^i(2^{n_j-1}) + s2^{n_j-1}\pi_s^i) \leq -(\sqrt{2}-1)\sqrt{2^{n_j-1}L\ln t_n}]. \qquad (25)
\end{aligned}$$

The above equality leads to

$$\begin{aligned}
&R_i(2^{n_j-1}) \geq \mu_i \cdot 2^{n_j-1} + (\sqrt{2}-1)\sqrt{2^{n_j-1}L\ln t_n} \quad \text{implies that} \\
&-O_i^s(2^{n_j-1}) + 2^{n_j-1}\pi_s^i \leq -(\sqrt{2}-1)\sqrt{2^{n_j-1}L\ln t_n}/(s|\mathcal{S}_i|) \ \text{ for some } s\in\mathcal{S}_i. \qquad (26)
\end{aligned}$$

Thus the event that the sample mean is significantly deviated from the true mean implies that at least one state occurs much often than predicted by its stationary probability.

Lemma 2 below is used to bound the probability that a state occurs much often than predicted by its stationary probability.

Lemma 2 (Chernoff Bound, Theorem 2.1 in [16]): Consider a finite state, irreducible, aperiodic and reversible Markov chain with state space $\mathcal{S}$, matrix of transition probabilities P, and an initial distribution $\mathbf{q}$. Let $N_{\mathbf{q}} = |(\frac{q_x}{\pi_x}), x\in\mathcal{S}|_2$. Let $\epsilon = 1-\lambda_2$, where λ_2 is the second largest eigenvalue of the matrix P. ϵ will be referred to as the eigenvalue gap. Let $A\subset\mathcal{S}$. Let $T_A(t)$ be the number of times that states in the set A are visited up to time t. Then for any $\gamma\geq 0$, we have

$$\Pr(T_A(t) - t\pi_A \geq \gamma) \leq (1+\frac{\gamma\epsilon}{10t})N_{\mathbf{q}}e^{-\gamma^2\epsilon/20t}. \qquad (27)$$

Using Lemma 2, we have

$$\Pr[R_i(2^{n_j-1}) \geq \mu_i \cdot 2^{n_j-1} + (\sqrt{2}-1)\sqrt{2^{n_j-1}L\ln t_n}] \tag{28}$$

$$\leq \sum_{s\in\mathcal{S}_i} P[-O_i^s(2^{n_j-1}) + 2^{n_j-1}\pi_s^i \leq -(\sqrt{2}-1)\sqrt{2^{n_j-1}L\ln t_n}/(s|\mathcal{S}_i|)] \tag{29}$$

$$= \sum_{s\in\mathcal{S}_i} P[O_i^s(2^{n_j-1}) - 2^{n_j-1}\pi_s^i \geq (\sqrt{2}-1)\sqrt{2^{n_j-1}L\ln t_n}/(s|\mathcal{S}_i|)] \tag{30}$$

$$\leq \sum_{s\in\mathcal{S}_i}(1+\frac{\epsilon_i\sqrt{L\ln t_n/2^{n_j-1}}}{10s})N_{\mathbf{q}^i}t_n^{-(3-2\sqrt{2})(L\epsilon^i/(20(s)^2|\mathcal{S}_i|^2)))} \tag{31}$$

$$\leq \frac{|\mathcal{S}_i|}{\pi_{\min}}(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}})t_n^{-(3-2\sqrt{2})(\frac{L\epsilon_{\min}-10s_{\max}^2}{20s_{\max}^2|\mathcal{S}|_{\max}^2})}. \tag{32}$$

Since $L \geq \frac{1}{\epsilon_{\min}}(4\frac{20s_{\max}^2|\mathcal{S}|_{\max}^2}{(3-2\sqrt{2})} + 10s_{\max}^2)$ and $K < t_n$ in (23), we have

$$\Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n,w_i}] \leq \frac{|\mathcal{S}_i|}{\pi_{\min}}(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}})t_n^{-3}. \tag{33}$$

Similarly, it can be shown that

$$\Pr[\bar{s}^*(t_n) \leq \mu^* - C_{t_n,w}] \leq \frac{|\mathcal{S}^*|}{\pi_{\min}}(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}})t_n^{-3}. \tag{34}$$

So

$$\mathbb{E}[T_i^2(t)] \leq \lceil\log_4(\frac{3}{2}(t-M)+1)\rceil 3\frac{|\mathcal{S}_i|+|\mathcal{S}^*|}{\pi_{\min}}(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}}). \tag{35}$$

Combining (9) (13) (14) (15) (35), we can get the upper bound of regret:

$$\begin{aligned} r_\Phi(t) \leq\ & (\lceil\log_4(\frac{3}{2}(t-M)+1)\rceil)\max_i A_i \\ & + N(\lfloor\log_4(3D\ln t+1)\rfloor+1)\max_i A_i \\ & + \sum_i(\mu^*-\mu_i)(\lceil\log_4(\frac{3}{2}(t-M)+1)\rceil 3\frac{|\mathcal{S}_i|+|\mathcal{S}^*|}{\pi_{\min}}(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}}) \\ & + \sum_i(\mu^*-\mu_i)\frac{1}{3}[4(3D\ln t+1)-1]). \end{aligned} \tag{36}$$

We point out that the same Chernoff bound given in Lemma 2 is also used in [6] to handle the *rested* Markovian reward MAB problem. Note that the Cheroff bound in [16] requires that all the observations used in calculating the sample means ($\bar{s}_i$ and $\bar{s}^*$ in (21)) are from a continuously evolving Markov process. This condition is naturally satisfied in the rested MAB problem. However, for the restless MAB problem considered here, the sample means are calculated using

observations from multiple epochs, which are noncontiguous segments of the Markovian sample path. As detailed in the above proof, the desired bound on the probabilities of the events in (21) is ensured by the carefully chosen (growing) lengths of the exploration and exploitation epochs.

APPENDIX B. PROOF OF THEOREM 2

The choice of $L(t)$ and $D(t)$ implies that $D(t) \to \infty$ as $t \to \infty$. By the same reasoning in the proof of Theorem 1, the regret has three parts: The regret caused by arm switching, the regret caused by playing bad arms in the exploration epochs, and the regret caused by playing bad arms in the exploitation epochs. It will be shown that each part part of the regret is on a lower order than $f(t)\log(t)$.

The number of arm switchings is upper bounded by $N\log_2(t/N+1)$. So the regret caused by arm switching is upper bounded by

$$N\log_2(t/N+1)\max_i A_i, \tag{37}$$

where $A_i = (\min_{s\in\mathcal{S}_i}\pi_s^i)^{-1}\sum_{s\in\mathcal{S}_i} s$. Since $f(t)\to\infty$ as $t\to\infty$, we have

$$\lim_{t\to\infty}\frac{N\log_2(t/N+1)\max_i A_i}{f(t)\log(t)} = 0. \tag{38}$$

Thus the regret caused by arm switching is on a lower order than $f(t)\log(t)$.

The regret caused by playing bad arms in the exploration epochs is bounded by

$$\sum_i(\mu^*-\mu_i)\frac{N}{3}[4(3D(t)\ln t+1)-1]. \tag{39}$$

Since $\frac{f(t)}{D(t)}\to\infty$ as $t\to\infty$, we have

$$\lim_{t\to\infty}\frac{\sum_i(\mu^*-\mu_i)\frac{N}{3}[4(3D(t)\ln t+1)-1]}{f(t)\log(t)} = 0. \tag{40}$$

Thus the regret caused by playing bad arms in the exploration epochs is on a lower order than $f(t)\log(t)$.

For the regret caused by playing bad arms in the exploitation epochs, it is shown below that the time spent on a bad arm i can be bounded by a constant independent of t.

Since $\frac{D(t)}{L(t)}\to\infty$ as $t\to\infty$, there exists a time t_3 such that $\forall t\geq t_3$, $D(t)\geq\frac{4L(t)}{(\mu_{\sigma(1)}-\mu_{\sigma(2)})^2}$. There also exists a time t_4 such that $\forall t\geq t_4$, $L(t)\geq\frac{1}{\epsilon_{\min}}(6\frac{20s_{\max}^2|\mathcal{S}|_{\max}^2}{(3-2\sqrt{2})}+10s_{\max}^2)$. The time

spent on playing bad arms before $t_5 = \max(t_3, t_4)$ is at most t_5, and the caused regret is at most $(\mu^* - \mu_{\sigma(N)})t_5$. After t_5, the time spent on each bad arm i is upper bounded by:

$$3\frac{|\mathcal{S}_i| + |\mathcal{S}^*|}{\pi_{\min}}(1 + \frac{\epsilon_{\max}\sqrt{L(t_5)}}{10s_{\min}}). \tag{41}$$

An upper bound for the corresponding regret is

$$\sum_i (\mu^* - \mu_i)(3\frac{|\mathcal{S}_i| + |\mathcal{S}^*|}{\pi_{\min}})(1 + \frac{\epsilon_{\max}\sqrt{L(t_5)}}{10s_{\min}}). \tag{42}$$

So the regret caused by playing bad arms in the exploitation epochs is

$$(\mu^* - \mu_{\sigma(N)})t_5 + \sum_i (\mu^* - \mu_i)(3\frac{|\mathcal{S}_i| + |\mathcal{S}^*|}{\pi_{\min}})(1 + \frac{\epsilon_{\max}\sqrt{L(t_5)}}{10s_{\min}}), \tag{43}$$

which is a constant independent of time t. Thus the regret caused by playing bad arms in the exploration epochs is on a lower order than $f(t)\log(t)$.

Because each part of the regret is on a lower order than $f(t)\log(t)$, the total regret is also on a lower order than $f(t)\log(t)$.

REFERENCES

[1] T. Lai and H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, Vol. 6, No. 1, pp. 4C22, 1985.

[2] V. Anantharam, P. Varaiya, J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part I: I.I.D. Rewards," *IEEE Transaction on Automatic Control*, Vol. AC-32 ,No.11 , pp. 968-976, Nov., 1987.

[3] V. Anantharam, P. Varaiya, J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part II: Markovian Rewards," *IEEE Transaction on Automatic Control*, Vol. AC-32 ,No.11 ,pp. 977-982, Nov., 1987.

[4] R. Agrawal, "Sample Mean Based Index Policies With O(log n) Regret for the Multi-armed Bandit Problem," *Advances in Applied Probability*, Vol. 27, pp. 1054C1078, 1995.

[5] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, 47, 235-256, 2002.

[6] C. Tekin, M. Liu, "Online Algorithms for the Multi-Armed Bandit Problem With Markovian Rewards," *Proc. of Allerton Conference on Communications, Control, and Computing*, Sep., 2010.

[7] C. Papadimitriou, J. Tsitsiklis, "The Complexity of Optimal Queuing Network Control," *Mathematics of Operations Research*, Vol. 24, No. 2, pp. 293-305, May 1999.

[8] H. Liu, K. Liu, Q. Zhao, "Distributed Learning in Multi-Armed Bandit with Multiple Players - Markovian Rewards," Technical Report, Oct. 2010. Available at http://www.ece.ucdavis.edu/∼qzhao/Report.html.

[9] K. Liu, Q. Zhao, "Distributed Learning in Multi-Armed Bandit with Multiple Players," *Transations on Signal Processing*, Vol. 58, No. 11, pp. 5667-5681, Nov. 2010.

[10] C. Tekin, M. Liu, "Online Learning in Opportunistic Spectrum Access: A Restless Bandit Approach," Arxiv pre-print http://arxiv.org/abs/1010.0056, Oct. 2010.

[11] W. Dai, Y. Gai, B. Krishnamachari, Q. Zhao "The Non-Bayesian Restless Multi-armed Bandit: A Case Of Near-Logarithmic Regret," submitted to *ICASSP*, Oct., 2010.

[12] P. Whittle, "Restless Bandits: Activity Allocation in a Changing World," *Journal of Applied Probability*, Vol. 25, pp. 287-298, 1988.

[13] J. Gittins, "Bandit processes and dynamic allocation indicies (with discussion)," *J. R. Statist. Soc.*, B, 41, pp. 148-177, 1979

[14] R. Weber and G. Weiss, "On an Index Policy for Restless Bandits," *Journal of Applied Probability*, Vol. 27, No. 3, pp. 637-648, Sep., 1990.

[15] K. Liu and Q. Zhao "Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access," *IEEE Transactions on Information Theory*, Vol. 55, No. 11, pp. 5547-5567, Nov. 2010.

[16] D. Gillman, "A Chernoff Bound for Random Walks on Expander Graphs," *Proc. 34th IEEE Symp. on Foundatioins of Computer Science (FOCS93),vol. SIAM J. Comp.*,Vol. 27, No. 4, 1998.