

Learning in A Changing World: Restless Multi-Armed Bandit with Unknown Dynamics

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 which a player chooses one out of N arms to play at each time. The reward state of each arm transits according to an unknown Markovian rule when it is played and evolves according to an arbitrary unknown random process when it is passive. The performance of an arm selection policy is measured by regret, defined as the reward loss with respect to the case where the player knows which arm is the most rewarding and always plays the best arm. We construct a policy with an interleaving exploration and exploitation epoch structure that achieves a regret with logarithmic order when arbitrary (but nontrivial) bounds on certain system parameters are known. When no knowledge about the system is available, we show that the proposed policy achieves a regret arbitrarily close to the logarithmic order. We further extend the problem to a decentralized setting where multiple distributed players share the arms without information exchange. Under both an exogenous restless model and an endogenous restless model, we show that a decentralized extension of the proposed policy preserves the logarithmic regret order as in the centralized setting. The results apply to adaptive learning in various dynamic systems and communication networks, as well as financial investment.

I. INTRODUCTION

A. Multi-Armed Bandit with *i.i.d.* and Rested Markovian Reward Models

In the classic multi-armed bandit (MAB) with an *i.i.d.* reward model, there are N independent arms and a single player. Each arm, when played, offers an *i.i.d.* random reward drawn from

⁰This work was supported by the Army Research Office under Grant W911NF-08-1-0467 and by the Army Research Lab under Grant W911NF-09-D-0006. Part of these results were presented at ITA, San Diego, CA, January 2011 and IEEE ICASSP, Prague, Czech Republic, May 2011.

a distribution with unknown mean. At each time, the player chooses one arm to play, aiming to maximize the total expected reward in the long run. This problem involves the well-known tradeoff between exploitation and exploration, where the player faces the conflicting objectives of playing the arm with the best reward history and playing a less explored arm to learn its reward statistics.

A commonly used performance measure of an arm selection policy is the so-called *regret* or *the cost of learning*, defined as the reward loss with respect to the case with a known reward model. It is clear that under a known reward model, the player should always play the arm with the largest reward mean. The essence of the problem is thus to identify the best arm without engaging other arms too often. Any policy with a sublinear growth rate of regret achieves the same maximum average reward (given by the largest reward mean) as in the known model case. However, the slower the regret growth rate, the faster the convergence to this maximum average reward, indicating a more effective learning ability of the policy.

In 1985, Lai and Robbins showed that regret grows at least at a logarithmic order with time, and an optimal policy was explicitly constructed to achieve the minimum regret growth rate for several reward distributions including Bernoulli, Poisson, Gaussian, Laplace [1]. Several other policies have been developed under different assumptions on the reward distribution [2], [3]. In particular, an index policy, referred to as Upper Confidence Bound 1 (UCB1) proposed by Auer *et al.* in [3], achieves logarithmic regret for any reward distributions with finite support. In [4], Liu and Zhao proposed a policy that achieves the optimal logarithmic regret order for a more general class of reward distributions and sublinear regret orders for heavy-tailed reward distributions.

In 1987, Anantharam *et al.* extended Lai and Robbin's results to a Markovian reward model where the reward state of each arm evolves as an unknown Markov process over successive plays and remains frozen when the arm is passive (the so-called *rested* Markovian reward model) [5]. In [6], Tekin and Liu extended the UCB1 policy proposed in [3] to the rested Markovian reward model.

B. Restless Multi-Armed Bandit with Unknown Dynamics

In this paper, we consider Restless Multi-Armed Bandit (RMAB), a generalization of the classic MAB. In contrast to the rested Markovian reward model, in RMAB, the state of each arm

continues to evolve even when it is not played. More specifically, the state of each arm changes according to an unknown Markovian transition rule when the arm is played and according to an arbitrary unknown random process when the arm is not played. We consider both the centralized (or equivalently, the single-player) setting and the decentralized setting with multiple distributed players.

1) *Centralized Setting*: A centralized setting where M players share their observations and make arm selections jointly is equivalent to a single player who chooses and plays M arms simultaneously. Without loss of generality, we focus on $M = 1$. Extensions to $M > 1$ are straightforward and can be found in [7]. The performance measure regret is similarly defined: it is the reward loss compared to the case when the player knows which arm is the most rewarding and always plays the best arm.

Compared to the i.i.d. and the rested Markovian reward models, the restless nature of arm state evolution requires that each arm be played consecutively for a period of time in order to learn its Markovian reward statistics. The length of each segment of consecutive plays needs to be carefully controlled to avoid spending too much time on a bad arm. At the same time, we experience a transient each time we switch out and then back to an arm, which leads to potential reward loss compared to the steady-state behavior of this arm. Thus, the frequency of arm switching needs to be carefully bounded.

To balance these factors, we construct a policy based on a deterministic sequencing of exploration and exploitation (DSEE) with an epoch structure. Specifically, the proposed policy partitions the time horizon into interleaving exploration and exploitation epochs with carefully controlled epoch lengths. During an exploration epoch, the player partitions the epoch into N contiguous segments, one for playing each of the N arms to learn their reward statistics. During an exploitation epoch, the player plays the arm with the largest sample mean (*i.e.*, average reward per play) calculated from the observations obtained so far. The lengths of both the exploration and the exploitation epochs grow geometrically. The number of arm switchings are thus at the logarithmic order with time. The tradeoff between exploration and exploitation is balanced by choosing the cardinality of the sequence of exploration epochs. Specifically, we show that with an $O(\log t)$ cardinality of the exploration epochs, sufficiently accurate learning of the arm ranks can be achieved when arbitrary (but nontrivial) bounds on certain system parameters are known, and the DSEE policy offers a logarithmic regret order. When no knowledge about the system

is available, we can increase the cardinality of the exploration epochs by an arbitrarily small order and achieve a regret arbitrarily close to the logarithmic order, *i.e.*, the regret has order $f(t) \log t$ for any increasing divergent function $f(t)$. In both cases, the proposed policy achieves the maximum average reward offered by the best arm.

We point out that the definition of regret here, while similar to that used for the classic MAB, is a weaker version of its counterpart. In the classic MAB with either i.i.d. or rested Markovian rewards, the optimal policy under a known model is indeed to stay with the best arm in terms of the reward mean¹. For RMAB, however, the optimal policy under a known model is no longer given by staying with the arm with the largest reward mean. Unfortunately, even under known Markovian dynamics, RMAB has been shown to be P-SPACE hard [8]. In this paper, we adopt a weaker definition of regret. First introduced in [9], weak regret measures the performance of a policy against a “partially-informed” genie who knows only which arm has the largest reward mean 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 an unknown model can be obtained is still open for exploration (see more discussions in Sec. I-C on related work).

2) *Decentralized Setting*: In the decentralized setting, there are M distributed players. At each time, a player chooses one arm to play based on its local observations without information exchange with other players. Collisions occur when multiple players choose the same arm and result in reward loss. The objective here is a decentralized policy to minimize the regret growth rate where regret is defined as the performance loss with respect to the ideal case where the players know the M best arms and are perfectly orthogonalized among these M best arms through centralized scheduling.

We consider two types of restless reward models: the exogenous restless model and the endogenous restless model. In the former, the system itself is rested: the state of an arm does not change when the arm is not engaged. However, from each individual player’s perspective, arms are restless due to actions of other players that are unobservable and uncontrollable. Under the endogenous restless model, the state of an arm evolves according to an arbitrary unknown

¹Under the rested Markovian reward model, staying with the best arm (in terms of the steady-state reward mean) is optimal up to a loss of $O(1)$ term resulting from the transient effect of the initial state which may not be the stationary distribution [5]. This $O(1)$ term, however, does not affect the order of the regret.

random process even when the arm is not played. Under both restless models, we extend the proposed DSEE policy to a decentralized policy that achieves the same logarithmic regret order as in the centralized scheduling. We emphasize that the logarithmic regret order is achieved under a complete decentralization among players. Players do not need to have synchronized global timing; each player can construct the exploration and exploitation epoch sequences according to its own local time.

We point out that the result under the exogenous restless model is stronger than that under the endogenous restless model in the sense that the regret is indeed defined with respect to the optimal policy under a known reward model and centralized scheduling. This is possible due to the inherent *rested* nature of the systems which makes any orthogonal sharing of the M best arms optimal (up to an $O(1)$ term) under a known reward model.

C. Related Work on RMAB

RMAB with unknown dynamics has not been studied in the literature except two parallel independent investigations reported in [10] and [11], both consider only a single player. In [10], Tekin and Liu considered the same problem and adopted the same definition of regret as in this paper. They proposed a policy that achieves logarithmic (weak) regret when certain knowledge about the system parameters is available [10]. Referred to as regenerative cycle algorithm (RCA), the policy proposed in [10] is based on the UCB1 policy proposed in [3] for the i.i.d. reward model. The basic idea of RCA is to play an arm consecutively for a random number of times determined by a regenerative cycle of a particular state and arms are selected based on the UCB1 index calculated from observations obtained only inside the regenerative cycles (observations obtained outside the regenerative cycles are not used in learning). The i.i.d. nature of the regenerative cycles reduces the problem to the classic MAB under the i.i.d. reward model. The DSEE policy proposed in this paper, however, has a deterministic epoch structure, and all observations are used in learning. As shown in the simulation examples in Sec. IV, DSEE can offer better performance than RCA since RCA may have to discard a large number of observations from learning before the chosen arm enters a regenerative cycle defined by a particular pilot state. Note that when the arm reward state space is large or when the chosen pilot state has a small stationary probability, it may take a long time for the arm to hit the pilot state, and since the transition probabilities are unknown, it is difficult to choose the pilot state for a

smaller hitting time. In [11], a strict definition of regret was adopted (*i.e.*, the reward loss with respect to the optimal performance in the ideal scenario with a known reward model). However, the problem can only be solved for a special class of RMAB with 2 or 3 arms governed by stochastically identical two-state Markov chains. For this special RMAB, the problem is tractable due to the semi-universal structure of the optimal policy of the corresponding RMAB with known dynamics established in [12], [13]. By exploiting the simple structure of the optimal policy under known Markovian dynamics, Dai *et al.* showed in [11] that a regret with an order arbitrarily close to logarithmic can be achieved for this special RMAB.

There are also several recent development on decentralized MAB with multiple players under the i.i.d. reward model. In [14], Liu and Zhao proposed a Time Division Fair Sharing (TDFS) framework which leads to a family of decentralized fair policies that achieve logarithmic regret order under general reward distributions and observation models [14]. Under a Bernoulli reward model, decentralized MAB was also addressed in [15], [16], where the single-player policy UCB1 was extended to the multi-player setting.

The basic idea of deterministic sequencing of exploration and exploitation was first proposed in [4] under the i.i.d. reward model. To handle the restless reward model, we introduce the epoch structure with epoch lengths carefully chosen to achieve the logarithmic regret order. The regret analysis also requires different techniques as compared to the i.i.d. case. Furthermore, the extension to the decentralized setting where different players are not required to synchronize in their epoch structures is highly nontrivial.

The results presented in this paper and the related work discussed above are developed within the non-Bayesian framework of MAB in which the unknowns in the reward models are treated as deterministic quantities and the design objective is universally (over all possible values of the unknowns) good policies. The other line of development is within the Bayesian framework in which the unknowns are modeled as random variables with known prior distributions and the design objective is policies with good average performance (averaged over the prior distributions of the unknowns). By treating the posterior probabilistic knowledge (updated from the prior distribution using past observations) about the unknowns as the system state, Bellman in 1956 abstracted and generalized the classic Bayesian MAB to a special class of Markov decision processes [17]. The long-standing Bayesian MAB was solved by Gittins in 1970s where he established the optimality of an index policy—the so-called Gittins index policy [18]. In 1988,

Whittle generalized the classic Bayesian MAB to the restless MAB (with known Markovian dynamics) and proposed an index policy based on a Lagrangian relaxation [19]. Weber and Weiss in 1990 showed that Whittle index policy is asymptotically optimal under certain conditions [20], [21]. In the finite regime, the strong performance of Whittle index policy has been demonstrated in numerous examples (see, e.g., [22]–[25]).

D. Applications

The restless multi-armed bandit problem has a broad range of potential applications. For example, in a cognitive radio network with dynamic spectrum access [26], 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 with unknown dynamics. At each time, a secondary user chooses one channel to sense and subsequently transmit if the channel is found to be idle. 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. The decentralized formulation under the endogenous restless model applies to a network of distributed secondary users.

The results obtained in this paper also apply to opportunistic communication in an unknown fading environment. Specifically, each user senses the fading realization of a selected channel and chooses its transmission power or data rate accordingly. The reward can be defined to capture energy efficiency (for fixed-rate transmission) or throughput. The objective is to design the optimal channel selection policies under unknown fading dynamics. Similar problems under known fading models have been considered in [27]–[29].

Another potential application is financial investment, where a Venture Capital (VC) selects one company to invest 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 [30]. 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 case with multiple VCs may fit into the decentralized formulation under the exogenous restless model.

E. Notations and Organization

For two positive integers k and l , define $k \oslash l \triangleq ((k-1) \bmod l) + 1$, which is an integer taking values from $1, 2, \dots, l$.

The rest of the paper is organized as follows. In Sec. II we consider the single-player setting. We propose the DSEE policy and establish its logarithmic regret order. In Sec. III, we consider the decentralized setting with multiple distributed players. We present several simulation examples in Sec. IV to compare the performance of DSEE with the policy proposed in [6]. Sec. V concludes this paper.

II. THE CENTRALIZED SETTING

In this section, we consider the centralized, or equivalently, the single-player setting. We first present the problem formulation and the definition of regret and then propose the DSEE policy and establish its logarithmic regret order.

A. Problem Formulation

In the centralized setting, we have one player and N independent arms. At each time, the player chooses one arm to play (the extension to the general case of simultaneous multiple plays is straightforward and can be found in [7]). Each arm, when played, offers certain amount of reward that defines the current state of the arm. Let $s_j(t)$ and \mathcal{S}_j denote, respectively, the state of arm j at time t and the state space of arm j . When arm j is played, its state changes according to an unknown Markovian rule with P_j as the transition matrix. The transition matrixes are assumed to be irreducible, aperiodic, and reversible. States of passive arms transit according to an arbitrary unknown random process. Let $\vec{\pi}_j = \{\pi_j(s)\}_{s \in \mathcal{S}_j}$ denote the stationary distribution of arm j under P_j . The stationary reward mean μ_j is given by $\mu_j = \sum_{s \in \mathcal{S}_j} s \pi_j(s)$. Let σ be a permutation of $\{1, \dots, N\}$ such that

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

Let μ^* denote $\mu_{\sigma(1)}$.

A policy Φ is a rule that specifies an arm to play based on the observation history. Let $t_j(n)$ denote the time index of the n th 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 under Φ by time t is given by

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

The performance of a policy Φ is measured by regret $r_\Phi(t)$ defined as the reward loss with respect to the best possible single-arm policy:

$$r_\Phi(t) = t\mu_{\sigma(1)} - \mathbb{E}_\Phi [R(t)] + O(1), \quad (2)$$

where the $O(1)$ constant term is caused by the transient effect of playing the best arm when its initial state is not given by the stationary distribution, \mathbb{E}_Φ denotes the expectation with respect to the random process induced by policy Φ . The objective is to minimize the growth rate of the regret with time t . Note that the constant term does not affect the order of the regret and will be omitted in the regret analysis in subsequent sections.

B. DSEE with An Epoch Structure

Compared to the i.i.d. and the rested Markovian reward models, the restless nature of arm state evolution requires that each arm be played consecutively for a period of time in order to learn its Markovian reward statistics and to approach the steady state. The length of each segment of consecutive plays needs to be carefully controlled: it should be short enough to avoid spending too much time on a bad arm and, at the same time, long enough to limit the transient effect. To balance these factors, we construct a policy based on DSEE with an epoch structure. As illustrated in Fig. 1, the proposed policy partitions the time horizon into interleaving exploration and exploitation epochs with geometrically growing epoch lengths. In the exploitation epochs, the player computes the sample mean (*i.e.*, average reward per play) of each arm based on the observations obtained so far and plays the arm with the largest sample mean, which can be considered as the current estimated best arm. In the exploration epochs, the player aims to learn the reward statistics 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 illustrated in Fig. 1, in the n th exploration epoch, the player plays every arm 4^{n-1} times. In the n th exploitation epoch with length $2 \times 4^{n-1}$, the player plays the arm with the largest sample mean (denoted as arm a^*) determined at the beginning of this epoch. At the end of each

epoch, whether to start an exploitation epoch or an exploration epoch is determined by whether sufficiently many (specifically, $D \log t$ as given in (3) in Fig. 2) observations have been obtained from every arm in the exploration epochs. This condition ensures that only logarithmically many plays are spent in the exploration epochs, which is necessary for achieving the logarithmic regret order. 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 reward statistics. A complete description of the proposed policy is given in Fig. 2.

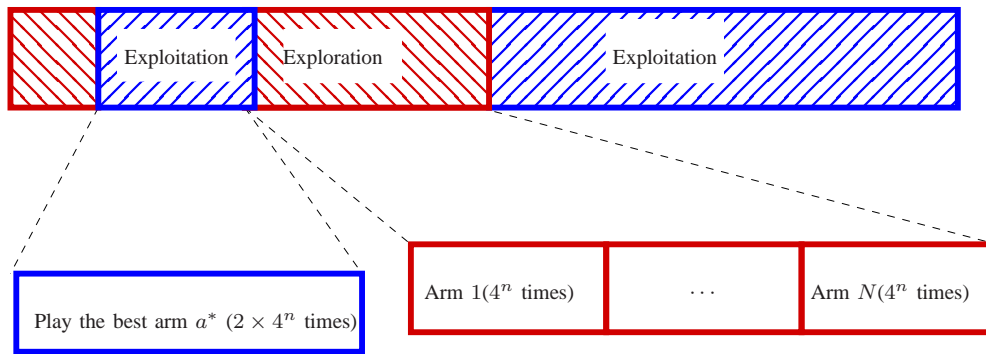


Fig. 1. The epoch structure with geometrically growing epoch lengths.

C. Regret Analysis

In this section, we show that the proposed policy achieves a logarithmic regret order. This is given in the following theorem.

Theorem 1: Assume that $\{P_i\}_{i=1}^N$ are finite state, irreducible, aperiodic and reversible. All the reward states are non-negative. Let ϵ_i be the second largest eigenvalue of P_i . Define $\epsilon_{\min} = \min_{1 \leq i \leq N} \epsilon_i$, $\pi_{\min} = \min_{1 \leq i \leq N, s \in \mathcal{S}_i} \pi_i(s)$, $r_{\max} = \max_{1 \leq i \leq N} \sum_{s \in \mathcal{S}_i} s$, $|\mathcal{S}|_{\max} = \max_{1 \leq i \leq N} |\mathcal{S}_i|$, $A_{\max} = \max_i (\min_{s \in \mathcal{S}_i} \pi_s^i)^{-1} \sum_{s \in \mathcal{S}_i} s$, and $L = \frac{30r_{\max}^2}{(3-2\sqrt{2})\epsilon_{\min}}$. Assume that the best arm has a distinct reward mean². Set the policy parameters D to satisfy the following condition:

$$D \geq \frac{4L}{(\mu_{\sigma(1)} - \mu_{\sigma(2)})^2}, \quad (4)$$

²The extension to the general case is straightforward

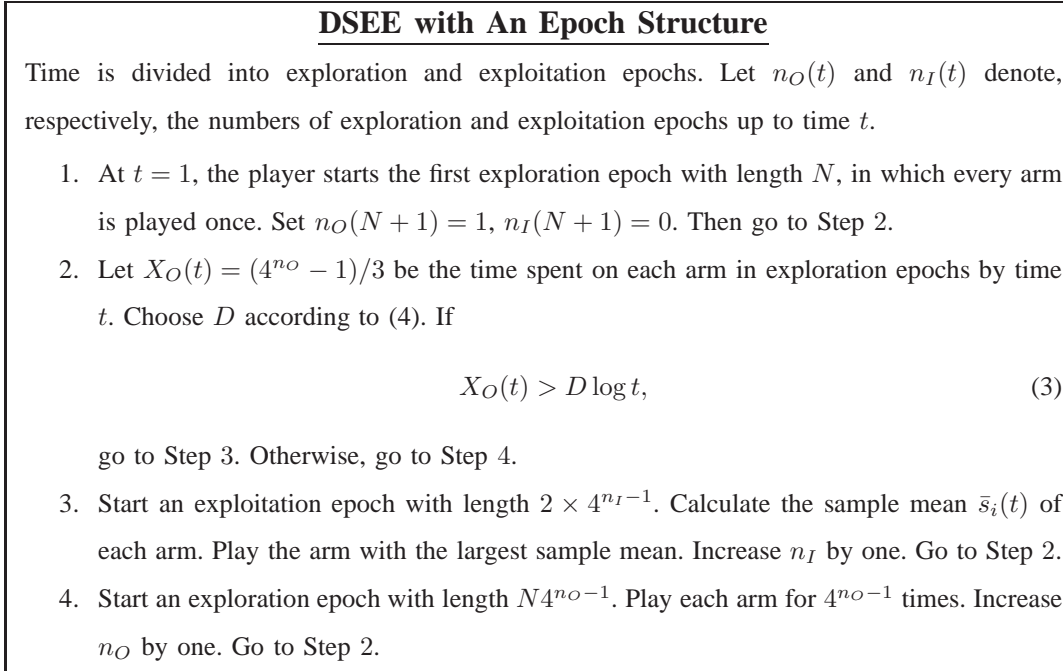


Fig. 2. DSEE with an epoch structure for RMAB.

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

$$r_\Phi(t) \leq C_1 \lceil \log_4 \left(\frac{3}{2}(t - N) + 1 \right) \rceil + C_2 [4(3D \log t + 1) - 1] + NA_{\max} (\lceil \log_4 (3D \log t + 1) \rceil + 1), \quad (5)$$

where

$$C_1 = \left(A_{\max} + 3 \sum_{j=2}^N \frac{\mu_{\sigma(1)} - \mu_{\sigma(j)}}{\pi_{\min}} \sum_{k=1, j} \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k \sqrt{L}}{10 \sum_{s \in S_k} s} \right) |S_k| \right), \quad (6)$$

$$C_2 = \frac{1}{3} \left(N\mu_{\sigma(1)} - \sum_{i=1}^N \mu_{\sigma(i)} \right). \quad (7)$$

Proof: See Appendix A for details. ■

Regret at any time t can be similarly bounded as given in (5) (see [7] for details). In the proposed policy, to ensure the logarithmic regret order, the policy parameter D needs to be chosen appropriately. This requires an arbitrary (but nontrivial) bound on r_{\max} , ϵ_{\min} , and $\mu_{\sigma(1)} - \mu_{\sigma(2)}$. In the case where no knowledge about the system is available, D can be chosen to increase with time rather than set *a priori* to achieve a regret order arbitrarily close to logarithmic. This is formally stated in the following theorem.

Theorem 2: Assume that $\{P_i\}_{i=1}^N$ are finite state, irreducible, aperiodic and reversible. All the reward states are non-negative. For any increasing sequence $f(t)$ ($f(t) \rightarrow \infty$ as $t \rightarrow \infty$), if $D(t) = f(t)$, then

$$r_\Phi(t) \sim O(f(t) \log t). \quad (8)$$

Proof: See Appendix B for details. ■

III. THE DECENTRALIZED SETTING

A. Problem Formulation

In the decentralized setting, there are M players and N independent arms ($N > M$). At each time, each player chooses one arm to play based on its local observations. As in the single player case, the reward state of arm j changes according to a Markovian rule when played, and the same set of notations are adopted. For the state transition of a passive arm, we consider two models: the endogenous restless model and the exogenous restless model. In the former, the arm evolves according to an arbitrary unknown random process even when it is not played. In the latter, the system itself is rested. From each individual player's perspective, however, arms are restless due to actions of other players that are unobservable and uncontrollable. The players do not know the arm dynamics and do not communicate with each other. Collisions occur when multiple players select the same arm to play. Different collision models can be adopted, where the players in conflict can share the reward or no one receives any reward. We consider here the latter; the extension to the former is straightforward (see [7]). In this case, the total reward under a policy Φ by time t is given by

$$R(t) = \sum_{j=1}^N \sum_{n=1}^{T_j(t)} s_j(t_j(n)) \mathbb{I}_j(t_j(n)), \quad (9)$$

where $\mathbb{I}_j(t_j(n)) = 1$ if arm j is played by one and only one player at time $t_j(n)$, and $\mathbb{I}_j(t_j(n)) = 0$ otherwise.

Under both restless models, regret $r_\Phi(t)$ is defined as the reward loss with respect to the ideal scenario of a perfect orthogonalization of the M players over the M best arms. We thus have

$$r_\Phi(t) = t \sum_{i=1}^M \mu_{\sigma(i)} - \mathbb{E}_\Phi R(t) + O(1), \quad (10)$$

where the $O(1)$ constant term comes from the transient effect of the M best arms (similar to the single-player setting). Note that under the exogenous restless model, this definition of regret is strict in the sense that $t \sum_{i=1}^M \mu_{\sigma(i)} + O(1)$ is indeed the maximal expected reward achievable under a known model of the arm dynamics.

B. Decentralized DSEE Policy

For the ease of presentation, we first assume that the players are synchronized according to a global time. Since the epoch structure of DSEE is deterministic, global timing ensures synchronized exploration and exploitation among players. We further assume that the players have pre-agreement on the time offset for sharing the arms, determined based on, for example, the players' ID. We will show in Sec. III-D that this requirement on global timing and pre-agreement can be eliminated to achieve a complete decentralization.

The decentralized DSEE has a similar epoch structure. In the exploration epochs (with the n th one having length $N \times 4^{n-1}$), the players play all N arms in a round-robin fashion with different offsets determined in the pre-agreement. In the exploitation epochs, each player calculates the sample mean of every arm based on its own local observations and plays the arms with the M largest sample mean in a round-robin fashion with a certain offset. Note that even though the players have different time-sharing offsets, collisions occur during exploitation epochs since the players may arrive at different sets and ranks of the M arms due to the randomness in their local observations. Each of these M arms is played $2 \times 4^{n-1}$ times. The n th exploitation epoch thus has length $2M \times 4^{n-1}$. A detailed description of the decentralized DSEE policy is given in Fig. 3.

C. Regret Analysis

In this section, we show that the decentralized DSEE policy achieves the same logarithmic regret order as in the centralized setting.

Theorem 3: Under the same notations and definitions as in Theorem 1, assume that different

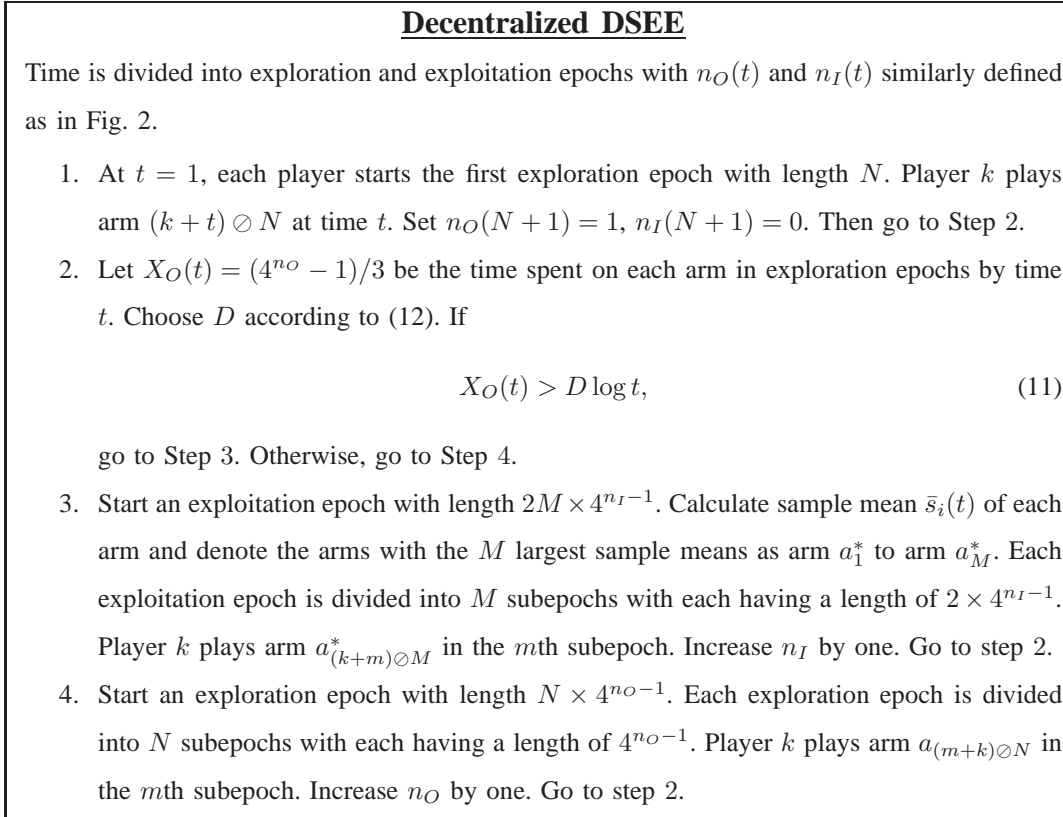


Fig. 3. Decentralized DSEE policy for RMAB.

arms have different mean values³. Set the policy parameter D to satisfy the following condition:

$$D \geq \frac{4L}{(\min_{j \leq M} (\mu_{\sigma(j)} - \mu_{\sigma(j+1)}))^2}. \quad (12)$$

Under the exogenous restless model, the regret of the decentralized DSEE at the end of any epoch can be upper bounded by

$$r_{\Phi}(t) \leq C_1 \lceil \log_4 \left(\frac{3t}{2M} + 1 \right) \rceil + C_2 (\lceil \log_4 (3D \log t + 1) \rceil + 1) + C_3 [4(3D \log t + 1) - 1], \quad (13)$$

³This assumption can be relaxed when the players determine the round-robin order of the arms based on pre-agreed arm indexes rather than the estimated arm rank. This assumption is only for simplicity of the presentation.

where

$$C_1 = \begin{cases} \sum_{m=1}^M \mu_m \frac{3M}{\pi_{\min}} \sum_{j=1}^M \sum_{i=1, i \neq j}^N \sum_{k=i, j} \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k \sqrt{L}}{10 \sum_{s \in S_k} s} \right) |\mathcal{S}_k| + M^2 A_{\max}, & \text{Endogenous restless model} \\ \sum_{m=1}^M \mu_m \frac{3M}{\pi_{\min}} \sum_{j=1}^M \sum_{i=1, i \neq j}^N \sum_{k=i, j} \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k \sqrt{L}}{10 \sum_{s \in S_k} s} \right) |\mathcal{S}_k|, & \text{Exogenous restless model} \end{cases} \quad (14)$$

$$C_2 = \begin{cases} NMA_{\max}, & \text{Endogenous restless model} \\ 0, & \text{Exogenous restless model} \end{cases} \quad (15)$$

$$C_3 = \frac{1}{3} \left(N \sum_{i=1}^M \mu_{\sigma(i)} - M \sum_{i=1}^N \mu_{\sigma(i)} \right). \quad (16)$$

Proof: See Appendix C for details. ■

Achieving the logarithmic regret order requires an arbitrary (but nontrivial) bound on r_{\max} , $\min_{j \leq M} (\mu_{\sigma(j)} - \mu_{\sigma(j+1)})$, and ϵ_{\min} . Similarly to the single-player case, D can be chosen to increase with time to achieve a regret order arbitrarily close to logarithmic as stated below.

Theorem 4: Under the same notations and definitions as in Theorem 1, assume that different arms have different mean values. For any increasing sequence $f(t)$ ($f(t) \rightarrow \infty$ as $t \rightarrow \infty$), if $D(t)$ is chosen such that $D(t) = f(t)$, then under both the endogenous and exogenous restless models,

$$r_{\Phi}(t) \sim O(f(t) \log t). \quad (17)$$

Proof: See Appendix D for details. ■

D. In the Absence of Global Synchronization and Pre-agreement

In this section, we show that the requirement on global synchronization and pre-agreement can be eliminated while maintaining the logarithmic order of the policy. As a result, players can join the system at different times.

Without global timing and pre-agreement, each player has its own exploration and exploitation epoch timing. The epoch structure of each player's local policy is similar to that given in Fig. 3. The only difference is that in each exploitation epoch, instead of playing the top M arms (in terms of sample mean) in a round-robin fashion, the player randomly and uniformly chooses one of them to play. When a collision occurs during the exploitation epoch, the player makes another

random and uniform selection among the top M arms. As shown in the proof of Theorem 5, this simple adjustment based on collisions achieves efficient sharing among all players without global synchronization and pre-agreement. Note that during an exploration epoch, the player plays all N arms in a round-robin fashion without reacting to collisions. Since the players still observe the reward state of the chosen arm, collisions affect only the immediate reward but not the learning ability of each player. As a consequence, collisions during a player's exploration epochs will not affect the logarithmic regret order since the total length of exploration epochs is at the logarithmic order. The key to establishing the logarithmic regret order in the absence of global synchronization and pre-agreement is to show that collisions during each player's exploitation epochs are properly bounded and efficient sharing can be achieved.

Theorem 5: Under the same notations and definitions as in Theorem 1, Decentralized DSEE without global synchronization and pre-agreement achieves logarithmic regret order.

Proof: See Appendix E for details. ■

The assumption that the arm reward state is still observed when collisions occur holds in many applications. For example, in the applications of dynamic spectrum access and opportunistic communications under unknown fading, each user first senses the state (busy/idle or the fading condition) of the chosen channel before a potential transmission. Channel states are always observed regardless of collisions. The problem is much more complex when collisions are unobservable and each player involved in a collision only observes its own local reward (which does not reflect the reward state of the chosen arm). In this case, collisions result in corrupted measurements that cannot be easily screened out, and learning from these corrupted measurements may lead to misidentified arm rank. How to achieve the logarithmic regret order without global timing and pre-agreement in this case is still an open problem.

IV. SIMULATION RESULTS

In this section, we study the performance of DSEE as compared to the RCA policy proposed in [10]. The first example is in the context of cognitive radio networks. We consider that a secondary user searches for idle channels unused by the primary network. Assume that the spectrum consists of N independent channels. The state—busy (0) or idle (1)—of each channel (say, channel n) evolves as a Markov chain with transition probabilities $\{p_{ij}^n\}$ $i, j \in \{0, 1\}$. At each time, the secondary user selects a channel to sense and choose the transmission power

according to the channel state. The reward obtained from a transmission over channel n in state i is given by r_i^n . We use the same set of parameters chosen in [6] (given in the caption of Fig. 4). We observe from Fig. 4 that RCA initially outperforms DSEE for a short period, but DSEE offers significantly better performance as time goes, and the regret offered by RCA does not seem to converge to the logarithmic order in a horizon of length 10^4 . We also note that while the condition on the policy parameter D given in (4) is sufficient for the logarithmic regret order, it is not necessary. Fig. 4 clearly shows the convergence to the logarithmic regret order for a small value of D , which leads to better finite-time performance.

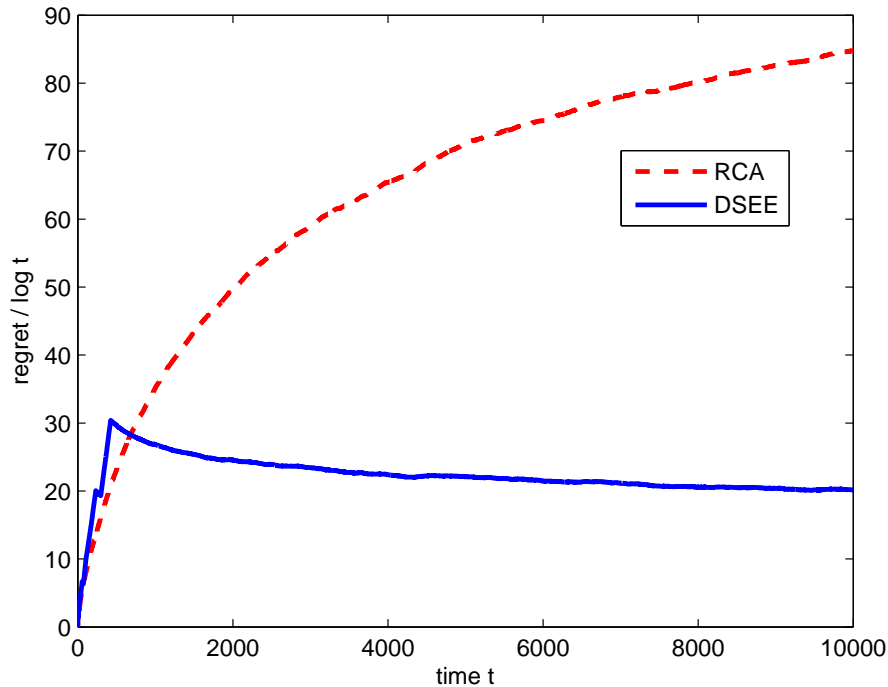


Fig. 4. Regret for DSEE and RCA, $p_{01} = [0.1, 0.1, 0.5, 0.1, 0.1]$, $p_{10} = [0.2, 0.3, 0.1, 0.4, 0.5]$, $r_1 = [1, 1, 1, 1, 1]$, $r_0 = [0.1, 0.1, 0.1, 0.1, 0.1]$, $D = 10$, $L = 10$, 100 Monte Carlo runs.

In the next example, we consider a case with a relatively large reward state space. We consider a case with 5 arms, each having 20 states. Rewards from each state for arm 2 to arm 5 is $[1, 2, \dots, 20]$. Rewards from each state for arm 1 is $1.5 \times [1, 2, \dots, 20]$ (to make it a better arm than the rest). Transition probabilities of all arms were generated randomly and can be found in [7]. The stationary distributions of all arms are close to uniform, which avoids the most

negative effect of randomly chosen pilot states in RCA. The values of D in DSEE and L in RCA were chosen to be the minimum as long as the ratio of the regret to $\log t$ converges to a constant with a reasonable time horizon. We again observe a better performance from DSEE as shown in Fig. 5.

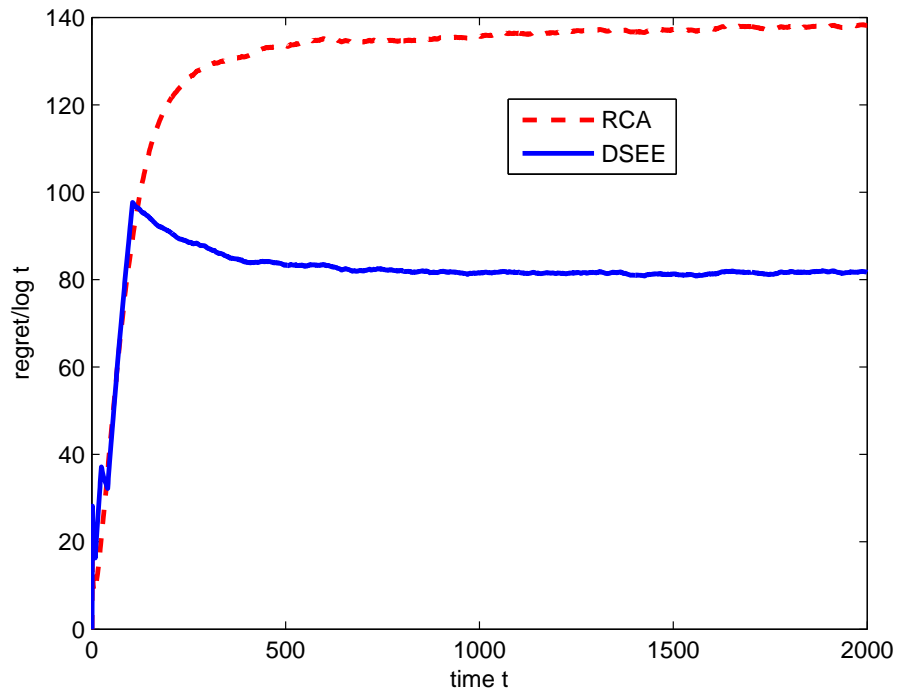


Fig. 5. Regret for DSEE and RCA with 5 arms, 20 states, $L = 20$, $D = 1.8$, 1000 Monte Carlo runs.

The better performance of DSEE over RCA may come from the fact that DSEE learns from all observations while RCA only uses observations within the regenerative cycles in learning. When the arm reward state space is large or the randomly chosen pilot state that defines the regenerative cycle has a small stationary probability, RCA may have to discard a large number of observations from learning.

V. CONCLUSION

In this paper, we studied the restless multi-armed bandit (RMAB) problem with unknown dynamics under both centralized (single-player) and decentralized settings. We developed a policy based on a deterministic sequencing of exploration and exploitation with geometrically growing

epochs that achieves the logarithmic regret order. In particular, in the decentralized setting with multiple distributed players, the proposed policy achieves a complete decentralization for both the exogenous and endogenous restless models.

APPENDIX A. PROOF OF THEOREM 1

We first rewrite the definition of regret as

$$r_{\Phi}(t) = t\mu_{\sigma(1)} - \mathbb{E}_{\Phi} R(t) \quad (18)$$

$$= \sum_{i=1}^N [\mu_i \mathbb{E}[T_i(t)] - \mathbb{E}[\sum_{n=1}^{T_i(t)} s_i(t_i(n))]] + \left[t\mu_{\sigma(1)} - \sum_{i=1}^N \mu_i \mathbb{E}[T_i(t)] \right]. \quad (19)$$

To show that the regret has a logarithmic order, it is sufficient to show that the two terms in (19) have logarithmic orders. The first term in (19) can be considered as the regret caused by transient effect. The second term can be considered as the regret caused by engaging a bad arm. First, we bound the regret caused by transient effect based on the following lemma.

Lemma 1 [5]: Consider an irreducible, aperiodic Markov chain with state space \mathcal{S} , 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 an 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 started the $(n+1)$ th exploration epoch, we have

$$\frac{1}{3}(4^n - 1) < D \log t, \quad (20)$$

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

$$n_O(t) \leq \lfloor \log_4(3D \log t + 1) \rfloor + 1. \quad (21)$$

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

$$n_I(t) \leq \lceil \log_4(\frac{3}{2}(t - N) + 1) \rceil. \quad (22)$$

Hence an logarithmic upper bound of the first term in (19) is

$$\sum_{i=1}^N [\mu_i \mathbb{E}[T_i(t)] - \mathbb{E}[\sum_{n=1}^{T_i(t)} s_i(t_i(n))]] \leq (\lceil \log_4(\frac{3}{2}(t - N) + 1) \rceil + N(\lfloor \log_4(3D \log t + 1) \rfloor + 1)) A_{\max}. \quad (23)$$

Next we show that the second term of (19) has a logarithmic order by bounding the total time spent on the bad arms. We first bound the time spent on the bad arms during the exploration epochs. Let $T_O(t)$ denote the time spent on each arm in the exploration epochs by time t . By (21), we have

$$T_O(t) \leq \frac{1}{3} [4(3D \log t + 1) - 1]. \quad (24)$$

Thus regret caused by playing bad arms in the exploration epochs is

$$\frac{1}{3} [4(3D \log t + 1) - 1] \left(N \mu_{\sigma(1)} - \sum_{i=1}^N \mu_{\sigma(i)} \right). \quad (25)$$

Next, we bound the time spent on the bad arms during the exploitation epochs. Let t_n denote the starting point of the n th exploitation epoch. Let $\Pr[i, j, n]$ denote the probability that arm i has a larger sample mean than arm j at t_n when arm j is the best arm, *i.e.*, $\Pr[i, j, n]$ is the probability of making a mistake in the n th exploitation epoch. Let w_i and w_j denote, respectively, the number of plays on arm i and arm j by t_n . Let $C_{t,w} = \sqrt{(L \log t/w)}$. We have

$$\Pr[i, j, n] \leq \Pr[\bar{s}_i(t_n) \geq \bar{s}_j(t_n)] \quad (26)$$

$$\begin{aligned} &\leq \Pr[\bar{s}_j(t_n) \leq \mu_j - C_{t_n, w_j}] + \Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n, w_i}] \\ &\quad + \Pr[\mu_j < \mu_i + C_{t_n, w_i} + C_{t_n, w_j}] \end{aligned} \quad (27)$$

$$\leq \Pr[\bar{s}_j(t_n) \leq \mu_j - C_{t_n, w_j}] + \Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n, w_i}], \quad (28)$$

where (28) follows from the fact that $w_i \geq D \log t_n$ and $w_j \geq D \log t_n$ and the condition on D given in (4).

Next we bound the two quantities in (28). Consider first the second term $\Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n, w_i}] = \Pr[w_i \bar{s}_i(t_n) \geq w_i \mu_i + \sqrt{L w_i \log t_n}]$. Note that the total w_i plays on arm i consists of multiple contiguous segments of the Markov sample path, each in a different epoch. Let K denote the number of such segments. From the geometric growth of the epoch lengths, we

can see that the length of each segment is in the form of 2^{k_l} ($l = 1, \dots, K$) with k_l 's being distinct. Without loss of generality, let $k_1 < k_2 < \dots < k_K$. Note that w_i , K , and k_l 's are random variables. The derivation below holds for every realization of these random variables. Let $R_i(l)$ denote the total reward obtained during the l th segment. Notice that $w_i = \sum_{l=1}^K 2^{k_l}$ and $\sqrt{w_i} \geq \sum_{l=1}^K (\sqrt{2} - 1)\sqrt{2^{k_l}}$. We then have

$$\Pr \left[w_i \bar{s}_i(t_n) \geq w_i \mu_i + \sqrt{L w_i \log t_n} \right] \quad (29)$$

$$\leq \Pr \left[\sum_{l=1}^K R_i(l) \geq \mu_i \sum_{l=1}^K 2^{k_l} + \sqrt{L \log t_n} (\sqrt{2} - 1) \sum_{l=1}^K \sqrt{2^{k_l}} \right] \quad (30)$$

$$= \Pr \left[\sum_{l=1}^K R_i(l) - \mu_i \sum_{l=1}^K 2^{k_l} - \sqrt{L \log t_n} (\sqrt{2} - 1) \sum_{l=1}^K \sqrt{2^{k_l}} \geq 0 \right] \quad (31)$$

$$= \Pr \left[\sum_{l=1}^K \left(R_i(l) - \mu_i 2^{k_l} - \sqrt{L \log t_n} (\sqrt{2} - 1) \sqrt{2^{k_l}} \right) \geq 0 \right] \quad (32)$$

$$\leq \sum_{l=1}^K \Pr \left[R_i(l) - \mu_i 2^{k_l} - \sqrt{L \log t_n} (\sqrt{2} - 1) \sqrt{2^{k_l}} \geq 0 \right] \quad (33)$$

$$= \sum_{l=1}^K \Pr \left[R_i(l) - \mu_i 2^{k_l} \geq \sqrt{L \log t_n} (\sqrt{2} - 1) \sqrt{2^{k_l}} \right] \quad (34)$$

$$= \sum_{l=1}^K \Pr \left[\sum_{s \in \mathcal{S}_i} (s O_i^s(l) - s 2^{k_l - 1} \pi_s^i) \geq \sqrt{L \log t_n} (\sqrt{2} - 1) \sqrt{2^{k_l}} \right], \quad (35)$$

where $O_i^s(l)$ denote the number of occurrences of state s on arm i in the l th segment. The following Chernoff Bound will be used to bound (35).

Lemma 2 (Chernoff Bound, Theorem 2.1 in [31]): Consider a finite state, irreducible, aperiodic and reversible Markov chain with state space \mathcal{S} , transition probabilities P , and an initial distribution \mathbf{q} . Let $N_{\mathbf{q}} = \left| \left(\frac{q_x}{\pi_x} \right), x \in \mathcal{S} \right|_2$. Let ϵ be the eigenvalue gap given by $1 - \lambda_2$, where λ_2 is the second largest eigenvalue of the matrix P . Let $A \subset \mathcal{S}$ and $T_A(t)$ be the number of times that states in A are visited up to time t . Then for any $\gamma \geq 0$, we have

$$\Pr(T_A(t) - t\pi_A \geq \gamma) \leq \left(1 + \frac{\gamma\epsilon}{10t}\right) N_{\mathbf{q}} e^{-\gamma^2\epsilon/20t}. \quad (36)$$

Using Lemma 2, we have

$$\Pr \left[\sum_{s \in \mathcal{S}_i} (sO_i^s(l) - s2^{k_l-1}\pi_s^i) \geq \sqrt{L \log t_n}(\sqrt{2} - 1)\sqrt{2^{k_l}} \right] \quad (37)$$

$$= \Pr \left[\sum_{s \in \mathcal{S}_i} (sO_i^s(l) - s2^{k_l-1}\pi_s^i) \geq \sqrt{L \log t_n}(\sqrt{2} - 1)\sqrt{2^{k_l}} \left(\frac{\sum_{s \in \mathcal{S}_i} s}{\sum_{s \in \mathcal{S}_i} s} \right) \right] \quad (38)$$

$$= \Pr \left[\sum_{s \in \mathcal{S}_i} \left(sO_i^s(l) - s2^{k_l-1}\pi_s^i - \sqrt{L \log t_n}(\sqrt{2} - 1)\sqrt{2^{k_l}} \left(\frac{s}{\sum_{s \in \mathcal{S}_i} s} \right) \right) \geq 0 \right] \quad (39)$$

$$= \Pr \left[\sum_{s \in \mathcal{S}_i, s \neq 0} \left(sO_i^s(l) - s2^{k_l-1}\pi_s^i - \sqrt{L \log t_n}(\sqrt{2} - 1)\sqrt{2^{k_l}} \left(\frac{s}{\sum_{s \in \mathcal{S}_i} s} \right) \right) \geq 0 \right] \quad (40)$$

$$\leq \sum_{s \in \mathcal{S}_i, s \neq 0} \Pr \left[sO_i^s(l) - s2^{k_l-1}\pi_s^i - \sqrt{L \log t_n}(\sqrt{2} - 1)\sqrt{2^{k_l}} \left(\frac{s}{\sum_{s \in \mathcal{S}_i} s} \right) \geq 0 \right] \quad (41)$$

$$\leq \sum_{s \in \mathcal{S}_i, s \neq 0} \Pr \left[O_i^s(l) - 2^{k_l-1}\pi_s^i \geq \sqrt{L \log t_n}(\sqrt{2} - 1)\sqrt{2^{k_l}} \left(\frac{1}{\sum_{s \in \mathcal{S}_i} s} \right) \right] \quad (42)$$

$$\leq \sum_{s \in \mathcal{S}_i} \Pr \left[O_i^s(l) - 2^{k_l-1}\pi_s^i \geq \sqrt{L \log t_n}(\sqrt{2} - 1)\sqrt{2^{k_l}} \left(\frac{1}{\sum_{s \in \mathcal{S}_i} s} \right) \right] \quad (43)$$

$$= |\mathcal{S}_i| \left(1 + \frac{(\sqrt{2} - 1)\epsilon_i \sqrt{L \log t_n}}{10 \sum_{s \in \mathcal{S}_i} s} \frac{1}{\sqrt{2^{k_l}}} \right) N_{\mathbf{q}^i} t_n^{-((3-2\sqrt{2})L\epsilon_i)/(20(\sum_{s \in \mathcal{S}_i} s)^2))}. \quad (44)$$

Thus we have

$$\Pr \left[w_i \bar{s}_i(t_n) \geq w_i \mu_i + \sqrt{L w_i \log t_n} \right] \quad (45)$$

$$\leq K |\mathcal{S}_i| N_{\mathbf{q}^i} t_n^{-((3-2\sqrt{2})L\epsilon_i)/(20(\sum_{s \in \mathcal{S}_i} s)^2))} + |\mathcal{S}_i| \frac{\sqrt{2}\epsilon_i \sqrt{L \log t_n}}{10 \sum_{s \in \mathcal{S}_i} s} N_{\mathbf{q}^i} t_n^{-(3-2\sqrt{2})(L\epsilon_i)/(20(\sum_{s \in \mathcal{S}_i} s)^2))} \quad (46)$$

$$= \left(K + \frac{\sqrt{2}\epsilon_i \sqrt{L \log t_n}}{10 \sum_{s \in \mathcal{S}_i} s} \right) |\mathcal{S}_i| N_{\mathbf{q}^i} t_n^{-(3-2\sqrt{2})(L\epsilon_i)/(20(\sum_{s \in \mathcal{S}_i} s)^2))} \quad (47)$$

$$\leq \left(\frac{\log t_n}{\log 2} + \frac{\sqrt{2}\epsilon_i \sqrt{L \log t_n}}{10 \sum_{s \in \mathcal{S}_i} s} \right) |\mathcal{S}_i| N_{\mathbf{q}^i} t_n^{-(3-2\sqrt{2})(L\epsilon_i)/(20(\sum_{s \in \mathcal{S}_i} s)^2))} \quad (48)$$

$$\leq \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_i \sqrt{L}}{10 \sum_{s \in \mathcal{S}_i} s} \right) |\mathcal{S}_i| N_{\mathbf{q}^i} t_n^{1/2 - (3-2\sqrt{2})(L\epsilon_i)/(20(\sum_{s \in \mathcal{S}_i} s)^2))}, \quad (49)$$

where (48) follows from the fact $K \leq \log_2 t_n$. Since $L \geq \frac{30r_{\max}^2}{(3-2\sqrt{2})\epsilon_i}$, we arrive at

$$\Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n, w_i}] \leq \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_i \sqrt{L}}{10 \sum_{s \in \mathcal{S}_i} s} \right) |\mathcal{S}_i| N_{\mathbf{q}^i} t_n^{-1}. \quad (50)$$

Similarly, it can be shown that

$$\Pr[\bar{s}_j(t_n) \leq \mu_j - C_{t_n, w_j}] \leq \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_j \sqrt{L}}{10 \sum_{s \in S_j} s} \right) |\mathcal{S}_j| N_{\mathbf{q}^i} t_n^{-1}. \quad (51)$$

Thus

$$\Pr[i, j, n] \leq \left[\left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_j \sqrt{L}}{10 \sum_{s \in S_j} s} \right) |\mathcal{S}_j| + \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_i \sqrt{L}}{10 \sum_{s \in S_i} s} \right) |\mathcal{S}_i| \right] N_{\mathbf{q}^i} t_n^{-1}. \quad (52)$$

Thus the regret caused by engaging bad arms in the n th exploitation epoch is bounded by

$$4^{n-1} 2 \sum_{j=2}^N (\mu_{\sigma(1)} - \mu_{\sigma(j)}) \left[\left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_j \sqrt{L}}{10 \sum_{s \in S_j} s} \right) |\mathcal{S}_j| + \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_1 \sqrt{L}}{10 \sum_{s \in S_1} s} \right) |\mathcal{S}_1| \right] \frac{1}{\pi_{\min}} t_n^{-1}. \quad (53)$$

By (22) and $t_n \geq \frac{2}{3} 4^{n-1}$, the bound in (53) becomes

$$3 \lceil \log_4 \left(\frac{3}{2} (t - N) + 1 \right) \rceil \frac{1}{\pi_{\min}} \sum_{j=2}^N (\mu_{\sigma(1)} - \mu_{\sigma(j)}) \sum_{k=1, j} \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k \sqrt{L}}{10 \sum_{s \in S_k} s} \right) |\mathcal{S}_k|. \quad (54)$$

Combining (19) (23) (54), we arrive at the upper bound of regret given in (5).

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 Chernoff bound in [31] requires that all the observations used in calculating the sample means (\bar{s}_i and \bar{s}_j in (28)) 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 (28) is ensured by the carefully chosen (growing) lengths of the exploration and exploitation epochs.

APPENDIX B. PROOF OF THEOREM 2

Recall in Theorem 1, L and D are fixed *a priori*. Now we choose $L(t) \rightarrow \infty$ as $t \rightarrow \infty$ and $\frac{D(t)}{L(t)} \rightarrow \infty$ as $t \rightarrow \infty$. 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 of the regret is on a lower order or on the same order of $f(t) \log t$.

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

$$N \log_2(t/N + 1) A_{\max}. \quad (55)$$

Since $f(t) \rightarrow \infty$ as $t \rightarrow \infty$, we have

$$\lim_{t \rightarrow \infty} \frac{N \log_2(t/N + 1) \max_i A_i}{f(t) \log t} = 0. \quad (56)$$

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

$$\frac{1}{3} [4(3D(t) \log t + 1) - 1] \left(N \mu_{\sigma(1)} - \sum_{i=1}^N \mu_{\sigma(i)} \right). \quad (57)$$

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

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 $\frac{D(t)}{L(t)} \rightarrow \infty$ as $t \rightarrow \infty$, there exists a time t_1 such that $\forall t \geq t_1$, $D(t) \geq \frac{4L(t)}{(\mu_{\sigma(1)} - \mu_{\sigma(2)})^2}$. There also exists a time t_2 such that $\forall t \geq t_2$, $L(t) \geq \frac{70r_{\max}^2}{(3-2\sqrt{2})\epsilon_{\min}}$. The time spent on playing bad arms before $t_3 = \max(t_1, t_2)$ is at most t_3 , and the caused regret is at most $(\mu_{\sigma(1)})t_3$. After t_3 , the time spent on each bad arm i is upper bounded by (following similar reasoning from (27) to (54))

$$\frac{\pi^2}{2} \frac{|\mathcal{S}_i| + |\mathcal{S}_{\sigma(1)}|}{\pi_{\min}} \left(1 + \frac{\epsilon_{\max} \sqrt{L(t_5)}}{10s_{\min}} \right). \quad (58)$$

An upper bound for the corresponding regret is

$$\frac{\pi^2}{2} \sum_{j=2}^N (\mu_{\sigma(1)} - \mu_{\sigma(j)}) \sum_{k=1, j} \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k \sqrt{L}}{10 \sum_{s \in S_k} s} \right) |\mathcal{S}_k| \frac{1}{\pi_{\min}}, \quad (59)$$

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 or on the same order of $f(t) \log t$, the total regret is on the same order of $f(t) \log t$.

APPENDIX C. PROOF OF THEOREM 3

We first rewrite the definition of regret as

$$r_{\Phi}(t) = t \sum_{i=1}^M \mu_{\sigma(i)} - \mathbb{E}_{\Phi} R(t) \quad (60)$$

$$= \sum_{i=1}^N [\mu_i \mathbb{E}[T_i(t)] - \mathbb{E}[\sum_{n=1}^{T_i(t)} s_i(t_i(n))]] + \left[t \sum_{i=1}^M \mu_{\sigma(i)} - \sum_{i=1}^N \mu_i \mathbb{E}[T_i(t)] \right]. \quad (61)$$

Using Lemma 1 the first term in (61) can be bounded by the following constant under the endogenous restless model (it is zero under the exogenous model):

$$(M \lceil \log_4(\frac{3t}{2M} + 1) \rceil + N(\lfloor \log_4(3D \log t + 1) \rfloor + 1)) M A_{\max}, \quad (62)$$

which has a logarithmic order.

We are going to show that the second term in (60) has a logarithmic order. It will be verified by bounding regret in both exploitation and exploration epochs by logarithmic order.

The upper bound on $T_O(t)$ in (24) still holds and consequently the regret caused by engaging bad arms in the exploration epochs by time t is upper bounded by

$$\frac{1}{3} [4(3D \log t + 1) - 1] \left(N \sum_{i=1}^M \mu_{\sigma(i)} - M \sum_{i=1}^N \mu_{\sigma(i)} \right). \quad (63)$$

The second reason for regret in the second term of (60) is not playing the expected arms in the exploitation epochs. If in the m th subepoch player k plays the $(m+k) \odot M$ best arm, then every time the best M arms are played and there is no conflict. But arm $a_{(m+k) \odot M}^*$ may not be the $(m+k) \odot M$ best arm. Bounding the probabilities of mistakes can lead to an upper bound on the regret caused in the exploitation epochs.

We adopt the same notations in Appendix A. The upper bound on $\Pr[i, j, n]$ in (52) still holds. Since different subepochs in the exploitation epochs are symmetric, the expected regret in different subepochs are the same. In the first subepoch, player k aims at arm $\sigma(k)$. In the model where no player in conflict gets any reward, player k failing to identify arm $\sigma(k)$ in the first subepoch of the n th exploitation epoch can lead to a regret no more than $\sum_{m=1}^M 2\mu_m \times 4^{n-1}$. Thus an upper bound for regret in the n th exploitation epoch can be obtained as

$$4^{n-1} 2M t_n^{-1} \frac{1}{\pi_{\min}} \sum_{m=1}^M \mu_m \sum_{j=1}^M \sum_{i=1, i \neq j}^N \sum_{k=i, j} \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k \sqrt{L}}{10 \sum_{s \in S_k} s} \right) |\mathcal{S}_k|. \quad (64)$$

By time t , we have

$$n_I(t) \leq \lceil \log_4(\frac{3t}{2M} + 1) \rceil. \quad (65)$$

From the upper bound on the number of the exploitation epochs given in (22), and also the fact that $t_n \geq \frac{2}{3} 4^{n-1}$, we have the following upper bound on regret caused in the exploitation epochs by time t (Denoted by $r_{\Phi, I}(t)$):

$$r_{\Phi, I}(t) \leq 3M \lceil \log_4(\frac{3t}{2M} + 1) \rceil \sum_{m=1}^M \mu_m \frac{1}{\pi_{\min}} \sum_{j=1}^M \sum_{i=1, i \neq j}^N \sum_{k=i, j} \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k \sqrt{L}}{10 \sum_{s \in S_k} s} \right) |\mathcal{S}_k| \quad (66)$$

Combining (60) (62) (63) (66), we arrive at the upper bounds of regret given in (13).

APPENDIX D. PROOF OF THEOREM 4

We set $L(t) \rightarrow \infty$ as $t \rightarrow \infty$ and $\frac{D(t)}{L(t)} \rightarrow \infty$ as $t \rightarrow \infty$. The regret has three parts: the transient effect of arms, the regret caused by playing bad arms in the exploration epochs, and the regret caused by mistakes in the exploitation epochs. It will be shown that each part of the regret is on a lower order or at the same order of $f(t) \log t$. The transient effect of arms is the same as in Theorem 3. Thus it is upper bounded by a constant under the exogenous restless model and on the order of $\log t$ under the endogenous restless model. Thus it is on a lower order than $f(t) \log t$.

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

$$\frac{1}{3} [4(3D(t) \log t + 1) - 1] \left(N \sum_{i=1}^M \mu_{\sigma(i)} - M \sum_{i=1}^N \mu_{\sigma(i)} \right). \quad (67)$$

Since $D(t) = f(t)$, regret in (67) is on the same order $f(t) \log t$.

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 $\frac{D(t)}{L(t)} \rightarrow \infty$ as $t \rightarrow \infty$, there exists a time t_1 such that $\forall t \geq t_1$, $D(t) \geq \frac{4L(t)}{(\min_{j \leq M} (\mu_{\sigma(j)} - \mu_{\sigma(j+1)}))^2}$. There also exists a time t_2 such that $\forall t \geq t_2$, $L(t) \geq \frac{70r_{\max}^2}{(3-2\sqrt{2})\epsilon_{\min}}$. The time spent on playing bad arms before $t_3 = \max(t_1, t_2)$ is at most t_3 , and the time spent on playing bad arms after t_3 is also bounded by a finite constant, which can be found in a similar manner in Appendix B. Thus the regret caused by mistakes in the exploitation epochs is on a lower order than $f(t) \log t$.

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

APPENDIX E. PROOF OF THEOREM 5

At each time, regret incurs if one of the following three events happens: (i) at least one player incorrectly identifies the set of M best arms in the exploitation sequence, (ii) at least one player is exploring, (iii) at least a collision occurs among the players. In the following, we will bound the expected number of the occurrences of these three events by the logarithmic order with time.

We first consider events (i) and (ii). Define a singular slot as the time slot in which either (i) or (ii) occurs. Based on the previous theorems, the local expected number of learning mistakes at each player is bounded by the logarithmic order with time. Furthermore, the cardinality of the local exploration sequence at each player is also bounded by the logarithmic order with time.

We thus have that the expected number of singular slots is bounded by the logarithmic order with time, *i.e.*, the expected number of the occurrences of events (i) and (ii) is bounded by the logarithmic order with time.

To prove the theorem, it remains to show that the expected number of collisions in all non-singular slots is also bounded by the logarithmic order with time. Consider the contiguous period consisting of all slots between two successive singular slots. During this period, all players correctly identify the M best arms and a collision occurs if and only if at least two players choose the same arm. Due to the randomized arm selection after each collision, it is clear that, in this period, the expected number of collisions before all players are orthogonalized into the M best arms is bounded by a constant uniform over time. Since the expected number of such periods has the same order as the expected number of singular slots, the expected number of such periods is bounded by the logarithmic order with time. The expected number of collisions over all such periods is thus bounded by the logarithmic order with time, *i.e.*, the expected number of collisions in all non-singular slots is bounded by the logarithmic order with time. We thus proved the theorem.

REFERENCES

- [1] T. Lai and H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4C22, 1985.
- [2] 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.
- [3] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, 47, 235-256, 2002.
- [4] K. Liu and Q. Zhao "Deterministic Sequencing of Exploration and Exploitation for Multi-Armed Bandit Problems," *Proc. of Allerton Conference on Communications, Control, and Computing*, Sep., 2011.
- [5] 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.
- [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] H. Liu, K. Liu, Q. Zhao, "Learning in A Changing World: Non-Bayesian Restless Multi-Armed Bandit," Technical Report, Oct. 2010. Available at <http://arxiv.org/abs/1011.4969>
- [8] C. Papadimitriou, J. Tsitsiklis, "The Complexity of Optimal Queuing Network Control," *Mathematics of Operations Research*, vol. 24, no. 2, pp. 293-305, May, 1999.

- [9] P. Auer, N. Cesa-Bianchi, Y. Freund, R.E. Schapire “The nonstochastic multiarmed bandit problem,” *SIAM Journal on Computing*, vol. 32, pp. 48C77, 2002.
- [10] C. Tekin, M. Liu, “Online Learning in Opportunistic Spectrum Access: A Restless Bandit Approach,” *Proc. of International Conference on Computer Communications (INFOCOM)*, April 2011, Shanghai, China.
- [11] W. Dai, Y. Gai, B. Krishnamachari, Q. Zhao “The Non-Bayesian Restless Multi-armed Bandit: A Case Of Near-Logarithmic Regret,” *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May, 2011.
- [12] Q. Zhao, B. Krishnamachari, and K. Liu, “On Myopic Sensing for Multi-Channel Opportunistic Access: Structure, Optimality, and Performance,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431-5440, Dec., 2008.
- [13] S.H. Ahmad, M. Liu, T. Javidi, Q. Zhao, B. Krishnamachari “Optimality of Myopic Sensing in Multi-Channel Opportunistic Access,” *IEEE Transactions on Information Theory*, vol. 55, No. 9, pp. 4040-4050, Sep., 2009.
- [14] K. Liu, Q. Zhao, “Distributed Learning in Multi-Armed Bandit with Multiple Players,” *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667-5681, Nov., 2010.
- [15] A. Anandkumar, N. Michael, A.K. Tang, A. Swami “Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret,” *IEEE JSAC on Advances in Cognitive Radio Networking and Communications*, vol. 29, no. 4, pp. 731-745, Mar., 2011.
- [16] Y. Gai and B. Krishnamachari, “Decentralized Online Learning Algorithms for Opportunistic Spectrum Access,” *IEEE Global Communications Conference (GLOBECOM 2011)*, Houston, USA, Dec., 2011.
- [17] R. Bellman, “A Problem in the Sequential Design of Experiments,” *Sankhya*, vol. 16, pp. 221-229, 1956.
- [18] J. Gittins, “Bandit Processes and Dynamic Allocation Indices,” *Journal of the Royal Statistical Society*, vol. 41, no. 2, pp. 148177, 1979.
- [19] P. Whittle, “Restless Bandits: Activity Allocation in a Changing World,” *J. Appl. Probab.*, vol. 25, pp. 287-298, 1988.
- [20] R. R.Weber and G.Weiss, “On an Index Policy for Restless Bandits,” *J. Appl. Probab.*, vol.27, no.3, pp. 637-648, Sep., 1990.
- [21] R. R. Weber and G. Weiss, “Addendum to On an Index Policy for Restless Bandits,” *Adv. Appl. Prob.*, vol. 23, no. 2, pp. 429-430, Jun., 1991.
- [22] K. D. Glazebrook, H. M. Mitchell, “An Index Policy for a Stochastic Scheduling Model with Improving/ Deteriorating Jobs,” *Naval Research Logistics (NRL)*, vol. 49, pp. 706-721, Mar., 2002.
- [23] P. S. Ansell, K. D. Glazebrook, J.E. Nino-Mora, and M. O’Keeffe, “Whittles Index Policy for a Multi-Class Queuing System with Convex Holding Costs,” *Math. Meth. Operat. Res.*, vol. 57, pp. 21-39, 2003.
- [24] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride, “Some Indexable Families of Restless Bandit Problems,” *Advances in Applied Probability*, vol. 38, pp. 643-672, 2006.
- [25] 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.
- [26] Q. Zhao and B.M. Sadler “A Survey of Dynamic Spectrum Access,” *IEEE Signal Processing Magazine*, vol. 24, no. 3, pp. 79-89, May. 2007.
- [27] M. Agarwal, V. S. Borkar, A. Karandikar, “Structural Properties of Optimal Transmission Policies Over a Randomly Varying Channel,” *IEEE Transactions on Wireless Communications*, vol. 53, no. 6, pp. 1476-1491, Jul. 2008.
- [28] S. Ali, V. Krishnamurthy, V. Leung, “ Optimal and Approximate Mobility Assisted Opportunistic Scheduling in Cellular Data Networks,” *IEEE Transactions Mobile Computing*, vol. 6, no. 6, pp. 633-648, Jun. 2007
- [29] L. Johnston and Vikram. Krishnamurthy, “Opportunistic File Transfer over a Fading Channel - A POMDP Search Theory

Formulation with Optimal Threshold Policies,” *IEEE Transactions Wireless Communications*, vol. 5, no. 2, pp. 394-405, Feb. 2006.

[30] M. Sorensen “Learning By Investing: Evidence from Venture Capital,” *Proc. of American Finance Association Annual Meeting*, Feb. 2008.

[31] D. Gillman, “A Chernoff Bound for Random Walks on Expander Graphs,” *Proc. 34th IEEE Symp. on Foundations of Computer Science (FOCS93)*, vol. *SIAM J. Comp.*, vol. 27, no. 4, 1998.