Learning in a Changing World: Restless Multiarmed Bandit With Unknown Dynamics

Haoyang Liu, Keqin Liu, *Member, IEEE*, and Qing Zhao, *Fellow, IEEE*

***Abstract*—We consider the restless multiarmed bandit 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. 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.**

***Index Terms*—Distributed learning, online learning, regret, restless multiarmed bandit (RMAB).**

I. Introduction

A. Multiarmed Bandit With i.i.d. and Rested Markovian Reward Models

In the classic multiarmed 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 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, and 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.* [3], achieves logarithmic regret for any reward distributions with bounded 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 MAB With Unknown Dynamics

In this paper, we consider restless multiarmed 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

Manuscript received December 09, 2011; revised July 04, 2012; accepted October 17, 2012. Date of publication November 27, 2012; date of current version February 12, 2013. This work was supported in part by the Army Research Office under Grant W911NF-12-1-0271 and in part by the Army Research Lab under Grant W911NF-09-D-0006. This paper was presented in part at the Information Theory and Applications Workshop, San Diego, CA, January 2011, and in part at the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing.

The authors are with the Department of Electrical and Computer Engineering. University of California, Davis, CA 95616 USA (e-mail: liu@ucdavis.edu; kqliu@ucdavis.edu; qzhao@ucdavis.edu).

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2012.2230215

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 period 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 is 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 mean.[1] 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 Section 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 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 epoch structures; 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 a couple of parallel independent investigations reported in [10]–[13], all of them 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 ex-

[1] Under 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.

amples in Section V, 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. Furthermore, 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 two or three arms governed by stochastically identical two-state Markov chains. For this special RMAB, the problem is tractable due to the semiuniversal structure of the optimal policy of the corresponding RMAB with known dynamics established in [14] and [15]. 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. A similar special RMAB was also considered in [12] where the underlying two-state Markov chains are not necessarily identical but assumed to be positively correlated. This RMAB under known transition probabilities was considered in [16] and a $(2+\epsilon)$-approximation policy was proposed under the average reward criterion. In [12], Tekin and Liu addressed this RMAB under unknown dynamics and developed a learning algorithm with logarithmic regret with regret defined as the reward loss with respect to the approximation policy developed in [16]. In [13], Tekin and Liu considered a class of RMAB where the underlying partially observable Markov chains evolve independently of the actions. Strict regret was considered by assuming a finite horizon with fixed and known horizon length. Under certain implicit and rather complex conditions on the value function and the optimal policy for the known model case, a learning algorithm with logarithmic regret was proposed based on the idea of estimating the transition probabilities similar to that used in [17].

There are also several recent development on decentralized MAB with multiple players under the i.i.d. reward model. In [18], Liu and Zhao proposed a time-division fair sharing framework which leads to a family of decentralized fair policies that achieve logarithmic regret order under general reward distributions and observation models [18]. Under a Bernoulli reward model, decentralized MAB was also addressed in [19] and [20], where the single-player policy UCB1 was extended to the multiplayer setting. In [21], Tekin and Liu addressed decentralized learning under general interference functions and the i.i.d. reward model. Under the general interference functions, the total reward obtained by multiple players playing the same arm may be larger than that offered by the arm when played by a single player. As a consequence, orthogonalizing the players over the best arms may not be the optimal allocation. In [22], Kalathil *et al.* considered the case where arm ranks may be different across players. They proposed a decentralized policy that achieves $O(\log^3 t)$ regret under the i.i.d. reward model.

The basic idea of DSEE has been considered in [4] and [23] under the i.i.d. reward model. In [23], Cesa-Bianchi and Fischer studied MAB problems under the so-called fixed mean reward model which is more general than the i.i.d. model. The proposed policy operates in rounds with exponentially growing lengths. In each round, arms are first played one by one for a certain number of times (which grows linearly with the number of rounds). The arm with the best reward history is then played till the end of this round. It is shown in [23] that this policy achieves regret of the form $a + b\log t + c\log^2 t$ for any finite t. In [4], Liu and Zhao addressed the design of the cardinality of the exploration sequence to handle general reward distributions, including heavy-tailed distributions. 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 [24]. 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 [25]. 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 [26]. Weber and Weiss in 1990 showed that Whittle index policy is asymptotically optimal under certain conditions [27], [28]. In the finite regime, the strong performance of Whittle index policy has been demonstrated in numerous examples (see, e.g., [29]–[32]).

D. Applications

The RMAB problem has a broad range of potential applications. For example, in a cognitive radio network with dynamic spectrum access [33], 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 [34]–[36].

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 [37]. 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 $k \oslash l \overset{\triangle}{=} ((k-1) \mod l) + 1$, which is an integer taking values from $1, 2, \ldots, l$.

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

II. 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 finite 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 $\{1, \ldots, N\}$ such that

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

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 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 under Φ by time t is given by

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

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:

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

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; $\mathbb{E}_\Phi$ denotes the expectation with respect to the random process induced by policy Φ. The objective is to minimize the growth rate of the regret 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 nth exploration epoch, the player plays every arm 4^{n-1} times. In the nth 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

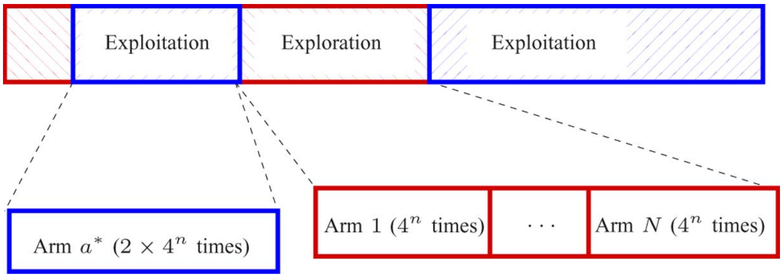

Fig. 1. Epoch structure with geometrically growing epoch lengths. (Arm a^*: the arm with the largest sample mean).

DSEE with An Epoch Structure

Time is divided into exploration and exploitation epochs. Let $n_O(t)$ and $n_I(t)$ denote, respectively, the numbers of exploration and exploitation epochs up to time t.

1. At $t = 1$, the player starts the first exploration epoch with length N, in which every arm is played once. Set $n_O(N+1) = 1$, $n_I(N+1) = 0$. Then go to Step 2.
2. Let $X_O(t) = (4^{n_O} - 1)/3$ be the time spent on each arm in exploration epochs by time t. Choose D according to (4). If
$$X_O(t) > D \log t, \tag{3}$$
go to Step 3. Otherwise, go to Step 4.
3. Start an exploitation epoch with length 2×4^{n_I}. Calculate the sample mean $\bar{s}_i(t)$ of each arm. Play the arm with the largest sample mean. Increase n_I by one. Go to Step 2.
4. Start an exploration epoch with length $N4^{n_O}$. Play each arm for 4^{n_O} times. Increase n_O by one. Go to Step 2.

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

learning the arm reward statistics. A complete description of the proposed policy is given in Fig. 2.

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

$$D \ge \frac{4L}{(\mu_{\sigma(1)} - \mu_{\sigma(2)})^2}. \tag{4}$$

The regret of DSEE at any time t can be upper bounded by

$$\begin{aligned} r_\Phi(t) \le\, & C_1\lceil \log_4(\frac{3}{2}(t-N)+1)\rceil + C_2[4(3D\log t + 1) - 1] \\ & + NA_{\max}(\lfloor \log_4(3D\log t + 1)\rfloor + 1)) \end{aligned} \tag{5}$$

where

$$\begin{aligned} C_1 = & \left(3\sum_{k=1,j}\left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k\sqrt{L}}{10\sum_{s\in S_k} s}\right)|\mathcal{S}_k|\right. \\ & \left.\times \sum_{j=2}^{N}\frac{\mu_{\sigma(1)} - \mu_{\sigma(j)}}{\pi_{\min}}\right) + A_{\max} \end{aligned} \tag{6}$$

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

Proof: See Appendix A for details. ■

We point out that the lower bound on D in (4) is sufficient but may not be necessary for the logarithmic regret order. As suggested in our simulation studies (see Section V), the value of D can be much smaller than that dictated by (4). Similar remarks hold for the values of D in Theorem 2 and of L_1 in Theorem 4.

III. 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

$$R(t) = \sum_{j=1}^{N}\sum_{n=1}^{T_j(t)} s_j(t_j(n))\mathbb{I}_j(t_j(n)) \tag{8}$$

[2] Note that the reversibility assumption can be relaxed to the irreducible multiplicative symmetrization by using Theorem 3.3 in [38] instead of Lemma 2.

[3] The extension to the general case is straightforward.

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

$$r_\Phi(t) = t\sum_{i=1}^{M} \mu_{\sigma(i)} - \mathbb{E}_\Phi R(t) + O(1) \tag{9}$$

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.

```
Decentralized DSEE
Time is divided into exploration and exploitation epochs with $n_O(t)$ and $n_I(t)$ similarly defined as in Fig. 2.
  1. At $t = 1$, each player starts the first exploration epoch with length $N$. Player $k$ plays arm $(k + t) \oslash N$ at time $t$. Set $n_O(N+1) = 1$, $n_I(N+1) = 0$. Then go to Step 2.
  2. Let $X_O(t) = (4^{n_O} - 1)/3$ be the time spent on each arm in exploration epochs by time $t$. Choose $D$ according to (11). If
         $X_O(t) > D\log t,$        (10)
     go to Step 3. Otherwise, go to Step 4.
  3. Start an exploitation epoch with length $2M \times 4^{n_I}$. Calculate sample mean $\bar{s}_i(t)$ of each arm and denote the arms with the $M$ largest sample means as arm $a_1^*$ to arm $a_M^*$. Each exploitation epoch is divided into $M$ subepochs with each having a length of $2 \times 4^{n_I}$. Player $k$ plays arm $a^*_{(k+m)\oslash M}$ in the $m$th subepoch. Increase $n_I$ by one. Go to step 2.
  4. Start an exploration epoch with length $N \times 4^{n_O}$. Each exploration epoch is divided into $N$ subepochs with each having a length of $4^{n_O}$. Player $k$ plays arm $a_{(m+k)\oslash N}$ in the $m$th subepoch. Increase $n_O$ by one. Go to step 2.
```

Fig. 3. Decentralized DSEE policy for RMAB.

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 preagreement on the time offset for sharing the arms, determined based on, for example, the players' ID. We will show in Section III-D that this requirement on global timing and preagreement can be eliminated to achieve a complete decentralization.

The decentralized DSEE has a similar epoch structure. In the exploration epochs (with the nth 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 preagreement. In the exploitation epochs, each players 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 nth 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 2: Under the same notations and definitions as in Theorem 1, assume that different arms have different mean values.[4] Set the policy parameter D to satisfy the following condition:

$$D \geq \frac{4L}{(\min_{j\leq M}(\mu_{\sigma(j)} - \mu_{\sigma(j+1)}))^2}. \tag{11}$$

The regret of the decentralized DSEE at any time t can be upper bounded by

$$\begin{aligned} r_\Phi(t) \leq\ & C_1\lceil \log_4(\frac{3t}{2M} + 1)\rceil \\ & + C_2(\lfloor \log_4(3D\log t + 1)\rfloor + 1) \\ & + C_3[4(3D\log t + 1) - 1] \end{aligned} \tag{12}$$

where

$$C_1 = \begin{cases} \left(\sum_{m=1}^{M} \mu_m \frac{3M}{\pi_{\min}} \sum_{j=1}^{M} \sum_{i=1,i\neq j}^{N} \sum_{k=i,j}\right. \\ \left.\left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k\sqrt{L}}{10\sum_{s\in S_k} s}\right)|\mathcal{S}_k|\right) + M^2 A_{\max}, \\ \qquad\qquad \text{Endogenous restless model} \\ \left(\sum_{m=1}^{M} \mu_m \frac{3M}{\pi_{\min}} \sum_{j=1}^{M} \sum_{i=1,i\neq j}^{N} \sum_{k=i,j}\right. \\ \left.\left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_k\sqrt{L}}{10\sum_{s\in S_k} s}\right)|\mathcal{S}_k|\right), \\ \qquad\qquad \text{Exogenous restless model} \end{cases} \tag{13}$$

$$C_2 = \begin{cases} NMA_{\max}, & \text{Endogenous restless model} \\ 0, & \text{Exogenous restless model} \end{cases} \tag{14}$$

$$C_3 = \frac{1}{3}\left(N\sum_{i=1}^{M} \mu_{\sigma(i)} - M\sum_{i=1}^{N} \mu_{\sigma(i)}\right). \tag{15}$$

Proof: See Appendix B for details. ■

[4] This assumption can be easily relaxed when the players determine the round-robin order of the arms based on preagreed arm labels rather than the estimated arm rank.

D. Eliminating Global Synchronization and Preagreement

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

Without global synchronization and preagreement, 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 3, this simple adjustment based on collisions achieves efficient sharing among all players without global synchronization and preagreement. 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 preagreement is to show that collisions during each player's exploitation epochs are properly bounded and efficient sharing can be achieved.

Theorem 3: Under the same notations and definitions as in Theorem 2, the decentralized DSEE without global synchronization and preagreement achieves logarithmic regret order.

Proof: See Appendix C 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 preagreement in this case is still an open problem.

```
RUCB
1. From $t = 1$ to $t = N$, the player plays each arm once.
2. Choose $L_1$ according to (17). Choose the arm with the
   highest index
       $\bar{s}_i(t) + \sqrt{\frac{L_1 \ln t}{T_i(t)}}.$        (16)
   If this is the $k$th time that arm $i$ is selected, the player plays
   arm $i$ for $2^{k-1}$ times.
```

Fig. 4. RUCB.

```
Decentralized RUCB
1. From $t = 1$ to $t = N$, player $j$ plays each arm once.
2. Choose $L_1$ according to (17). First find the $j$ arms with the
   highest indices
       $\bar{s}_i(t) + \sqrt{\frac{L_1 \ln t}{T_i(t)}}.$        (21)
   Then among the $j$ arms, choose the one with the lowest
   index
       $\bar{s}_i(t) - \sqrt{\frac{L_1 \ln t}{T_i(t)}}.$        (22)
   If this is the $k$th time that arm $i$ is selected, the player plays
   arm $i$ for $2^{k-1}$ times.
```

Fig. 5. Decentralized RUCB policy for RMAB.

IV. Relaxing Requirement on System Knowledge

We can see from Theorems 1 and 2 that the proposed policies require certain knowledge on the reward model to achieve the logarithmic regret order. This is also the case for the classic policies [1]–[3] under the i.i.d. reward model and the parallel work [10] for RMAB. In particular, the policies proposed in [1] and [2] require the knowledge of the distribution type (e.g., Gaussian or Laplacian) of each arm. The policies proposed in [3] require that the reward distributions have bounded support with a known support range. The policy proposed in [10] requires the same set of knowledge on the system parameters as the policies proposed in this paper except a positive lower bound on the difference between the best and the second best arms. In this section, we focus on how to relax the requirement on system knowledge in policy design.

A. Restless UCB

We first propose a policy that combines the geometrically growing epoch structure of DSEE with the UCB index. Referred to as restless UCB (RUCB), this policy achieves the logarithmic regret order without the knowledge on the difference between the best and the second best arms as required by DSEE. It thus requires the same set of knowledge as the policy proposed in [10]. Details of RUCB are given in Fig. 4.

The regret performance of RUCB in the centralized setting is given in the following theorem.

Theorem 4: Under the same assumptions as in Theorem 1, set the policy parameter L_1 to satisfy the following condition:

$$L_1 \geq \frac{110 r_{\max}^2}{(3 - 2\sqrt{2})\epsilon_{\min}}. \tag{17}$$

The regret of RUCB at any time t can be upper bounded by

$$C_1 \log(t) + A_{\max}\left(\lceil \log_2\left(\frac{t}{N}\right)\rceil + 1\right) + C_2 \tag{18}$$

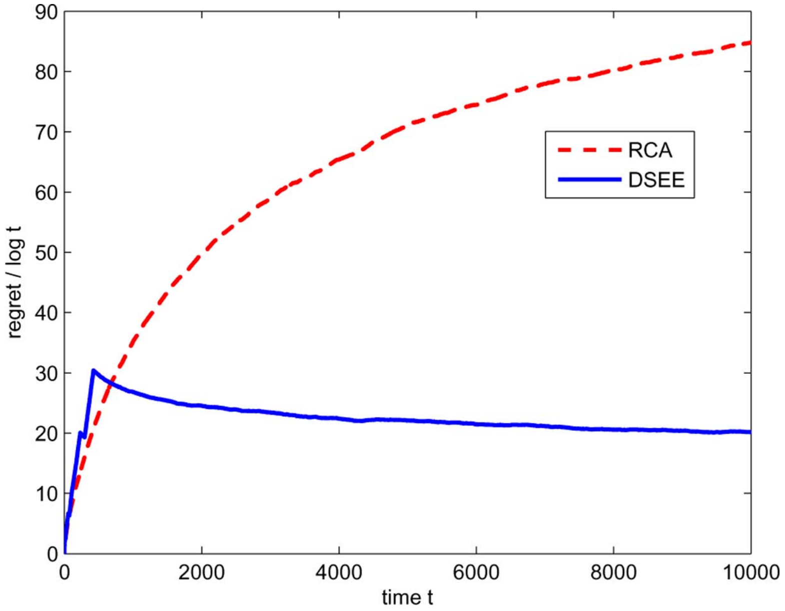

Fig. 6. Regret for DSEE and RCA, $p_{01} = [0.1, 0.1, 0.5, 0.1, 0.1]$, $p_{10} = [0.2, 0.3, 0.1, 0.4, 0.5]$, $r_1 = [1, 1, 1, 1, 1]$, $r_0 = [0.1, 0.1, 0.1, 0.1, 0.1]$, $D = 10$, $L = 10$, 100 Monte Carlo runs.

where

$$C_1 = \sum_{j\neq\sigma(1)} \frac{8L_1}{\mu_{\sigma(1)} - \mu_j} \tag{19}$$

$$C_2 = \left(\sum_{j\neq\sigma(1)} \frac{2\pi}{3}(\mu_{\sigma(1)} - \mu_j) \sum_{k=\sigma(1),j} \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_{\sigma(1)}\sqrt{L_1}}{10\sum_{s\in S_k} s} \right) |\mathcal{S}_k| \frac{1}{\pi_{\min}} \right). \tag{20}$$

Proof: See Appendix D for details. ■

Decentralized RUCB assumes preagreement among the M players i.e., player j targets at the jth best arm. We can use the idea in [20] to identify the jth ($j > 1$) best arm. Specifically, we first estimate the top j arms using the upper confidence bound index and then use the lower confidence bound index to find the worst arm among these j arms. A complete description is given in Fig. 5. It can be shown that player j can correctly identify the jth best arm except a logarithmic number of slots. The logarithmic regret order thus follows with a similar line of arguments as given in Appendix D.

B. DSEE

In this section, we show that by increasing the cardinality of the exploration sequence in DSEE, we can eliminate all the requirements on system knowledge with an arbitrarily small sacrifice to the regret order. The basic idea is to let the policy parameter D grow with time rather than set *a priori*. The result is given in the following theorem.

Theorem 5: In both the centralized and decentralized setting, for any increasing sequence $f(t)$ ($f(t) \to \infty$ as $t \to \infty$), if we set the policy parameter D in DSEE as $D(t) = f(t)$, then

$$r_\Phi(t) \sim O(f(t) \log t). \tag{23}$$

Proof: See Appendix E for details. ■

V. 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 $\{p^n_{ij}\}$ $i, j \in \{0, 1\}$. 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^n_i. We use the same set of parameters chosen in [6] (given in the caption of Fig. 6). We observe from Fig. 6 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. 6 clearly shows the convergence to the logarithmic regret order for a small value of D, which leads to better finite-time performance.

In the next example, we consider a case with a relatively large reward state space. We consider a case with five arms, each having 20 states. Rewards from each state for arm 2 to arm 5 is $[1, 2, \ldots, 20]$. Rewards from each state for arm 1 is

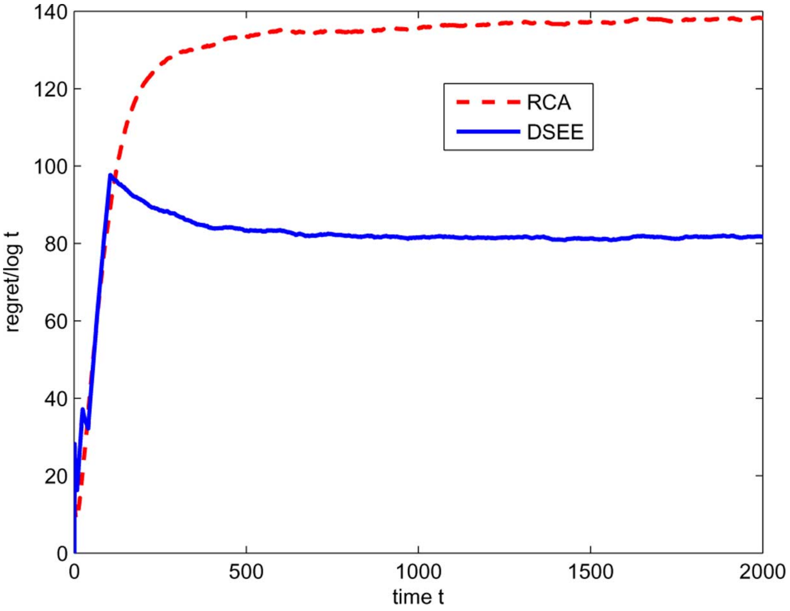

Fig. 7. Regret for DSEE and RCA with five arms, 20 states, $L = 20$, $D = 1.8$, 1000 Monte Carlo runs.

$1.5 \times [1, 2, \ldots, 20]$ (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. 7.

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.

VI. Conclusion

In this paper, we studied the RMAB problem with unknown dynamics under both centralized (single-player) and decentralized settings. We developed a policy based on a DSEE 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

$$r_\Phi(t) = t\mu_{\sigma(1)} - \mathbb{E}_\Phi R(t) \tag{24}$$

$$= \sum_{i=1}^{N} \left[\mu_i \mathbb{E}[T_i(t)] - \mathbb{E}[\sum_{n=1}^{T_i(t)} s_i(t_i(n))] \right] + \left[t\mu_{\sigma(1)} - \sum_{i=1}^{N} \mu_i \mathbb{E}[T_i(t)] \right]. \tag{25}$$

To show that the regret has a logarithmic order, it is sufficient to show that the two terms in (25) have logarithmic orders. The first term in (25) 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

$$\frac{1}{3}(4^n - 1) < D \log t \tag{26}$$

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

$$n_O(t) \leq \lfloor \log_4(3D\log t+1)\rfloor + 1. \tag{27}$$

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

$$n_I(t) \leq \lceil \log_4(\frac{3}{2}(t-N)+1)\rceil. \tag{28}$$

Hence, an logarithmic upper bound of the first term in (25) is

$$\begin{aligned}\sum_{i=1}^{N}[\mu_i\mathbb{E}[T_i(t)] &- \mathbb{E}[\sum_{n=1}^{T_i(t)} s_i(t_i(n))]] \\ &\leq A_{\max}\Bigg(\lceil \log_4(\frac{3}{2}(t-N)+1)\rceil \\ &\quad + N(\lfloor \log_4(3D\log t+1)\rfloor+1)\Bigg). \end{aligned} \tag{29}$$

Next, we show that the second term of (25) 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 (27), we have

$$T_O(t) \leq \frac{1}{3}[4(3D\log t+1)-1]. \tag{30}$$

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

$$\frac{1}{3}[4(3D\log t+1)-1]\left(N\mu_{\sigma(1)} - \sum_{i=1}^{N}\mu_{\sigma(i)}\right). \tag{31}$$

Next, we bound the time spent on the bad arms during the exploitation epochs. Let t_n denote the starting point of the nth 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 nth 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

$$\begin{aligned}\Pr[i,j,n] &= \Pr[\bar{s}_i(t_n) \geq \bar{s}_j(t_n)] \\ &\leq \Pr[\bar{s}_j(t_n) \leq \mu_j - C_{t_n,w_j}] \\ &\quad + \Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n,w_i}] \\ &\quad + \Pr[\mu_j < \mu_i + C_{t_n,w_i} + C_{t_n,w_j}] \quad (32)\\ &\leq \Pr[\bar{s}_j(t_n) \leq \mu_j - C_{t_n,w_j}] \\ &\quad + \Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n,w_i}] \quad (33)\end{aligned}$$

where (33) 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 (33). Consider first the second terms $\Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n,w_i}] = \Pr[w_i\bar{s}_i(t_n) \geq w_i\mu_i + \sqrt{Lw_i\log t_n}]$. 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} $(l=1,\ldots,K)$ with k_l's being distinct. Without loss of generality, let $k_1 < k_2 < \cdots < k_K$. 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 lth 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 where $O_i^s(l)$ denote the number of occurrences of state s on arm i in the lth segment. The following Chernoff Bound will be used to bound (35), given at the bottom of the following page.

Lemma 2 (Chernoff Bound [39, Th. 2.1]): 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 $N_{\mathbf{q}} = |(\frac{q_x}{\pi_x}), x \in \mathcal{S}|_2$. 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

$$\Pr(T_A(t) - t\pi_A \geq \gamma) \leq (1+\frac{\gamma\epsilon}{10t})N_{\mathbf{q}}e^{-\gamma^2\epsilon/20t}. \tag{37}$$

Using Lemma 2, we have (36), given at the bottom of the following page. Thus, we have

$$\begin{aligned}&\Pr\left[w_i\bar{s}_i(t_n) \geq w_i\mu_i + \sqrt{Lw_i\log t_n}\right] \\ &\leq K|\mathcal{S}_i|N_{\mathbf{q}^i}t_n^{-((3-2\sqrt{2})L\epsilon^i)/(20(\sum_{s\in S_i} s)^2))} \\ &\quad + |\mathcal{S}_i|\frac{\sqrt{2}\epsilon_i\sqrt{L\log t_n}}{10\sum_{s\in S_i} s}N_{\mathbf{q}^i}t_n^{-(3-2\sqrt{2})\frac{L\epsilon^i}{20(\sum_{s\in S_i} s)^2)}} \quad (38)\\ &= \left(K + \frac{\sqrt{2}\epsilon_i\sqrt{L\log t_n}}{10\sum_{s\in S_i} s}\right) \\ &\quad \times |\mathcal{S}_i|N_{\mathbf{q}^i}t_n^{-(3-2\sqrt{2})(L\epsilon^i/(20(\sum_{s\in S_i} s)^2))} \\ &\leq \left(\frac{\log t_n}{\log 2} + \frac{\sqrt{2}\epsilon_i\sqrt{L\log t_n}}{10\sum_{s\in S_i} s}\right) \\ &\quad \times |\mathcal{S}_i|N_{\mathbf{q}^i}t_n^{-(3-2\sqrt{2})(L\epsilon^i/(20(\sum_{s\in S_i} s)^2))} \quad (39)\\ &\leq \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_i\sqrt{L}}{10\sum_{s\in S_i} s}\right) \\ &\quad \times |\mathcal{S}_i|N_{\mathbf{q}^i}t_n^{1/2-(3-2\sqrt{2})(L\epsilon^i/(20(\sum_{s\in S_i} s)^2))} \quad (40)\end{aligned}$$

where (39) 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

$$\begin{aligned}&\Pr[\bar{s}_i(t_n) \geq \mu_i + C_{t_n,w_i}] \\ &\leq \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_i\sqrt{L}}{10\sum_{s\in S_i} s}\right)|\mathcal{S}_i|N_{\mathbf{q}^i}t_n^{-1}. \end{aligned} \tag{41}$$

Similarly, it can be shown that

$$\Pr[\bar{s}_j(t_n) \le \mu_j - C_{t_n,w_j}] \le \left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_j\sqrt{L}}{10\sum_{s\in S_j} s}\right)|\mathcal{S}_j| N_{\mathbf{q}^i} t_n^{-1}. \tag{42}$$

Thus

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

$$\begin{aligned}
&\Pr\left[w_i\bar{s}_i(t_n) \ge w_i\mu_i + \sqrt{Lw_i\log t_n}\right]\\
&\le \Pr\left[\sum_{l=1}^{K} R_i(l) \ge \mu_i\sum_{l=1}^{K} 2^{k_l} + \sqrt{L\log t_n}(\sqrt{2}-1)\sum_{l=1}^{K}\sqrt{2^{k_l}}\right]\\
&= \Pr\left[\sum_{l=1}^{K} R_i(l) - \mu_i\sum_{l=1}^{K} 2^{k_l} - \sqrt{L\log t_n}(\sqrt{2}-1)\sum_{l=1}^{K}\sqrt{2^{k_l}} \ge 0\right]\\
&= \Pr\left[\sum_{l=1}^{K}\left(R_i(l) - \mu_i 2^{k_l} - \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\right) \ge 0\right]\\
&\le \sum_{l=1}^{K}\Pr\left[R_i(l) - \mu_i 2^{k_l} - \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}} \ge 0\right]\\
&= \sum_{l=1}^{K}\Pr\left[R_i(l) - \mu_i 2^{k_l} \ge \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\right] \qquad (34)\\
&= \sum_{l=1}^{K}\Pr\left[\sum_{s\in\mathcal{S}_i}(sO_i^s(l) - s2^{k_l-1}\pi_s^i) \ge \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\right] \qquad (35)
\end{aligned}$$

$$\begin{aligned}
&\Pr\left[\sum_{s\in\mathcal{S}_i}(sO_i^s(l) - s2^{k_l-1}\pi_s^i) \ge \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\right]\\
&= \Pr\left[\sum_{s\in\mathcal{S}_i}(sO_i^s(l) - s2^{k_l-1}\pi_s^i) \ge \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\left(\frac{\sum_{\mathcal{S}_i} s}{\sum_{\mathcal{S}_i} s}\right)\right]\\
&= \Pr\left[\sum_{s\in\mathcal{S}_i}\left(sO_i^s(l) - s2^{k_l-1}\pi_s^i - \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\left(\frac{s}{\sum_{\mathcal{S}_i} s}\right)\right) \ge 0\right]\\
&= \Pr\left[\sum_{s\in\mathcal{S}_i, s\neq 0}\left(sO_i^s(l) - s2^{k_l-1}\pi_s^i - \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\left(\frac{s}{\sum_{\mathcal{S}_i} s}\right)\right) \ge 0\right]\\
&\le \sum_{s\in\mathcal{S}_i, s\neq 0}\Pr\left[sO_i^s(l) - s2^{k_l-1}\pi_s^i - \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\left(\frac{s}{\sum_{\mathcal{S}_i} s}\right) \ge 0\right]\\
&\le \sum_{s\in\mathcal{S}_i, s\neq 0}\Pr\left[O_i^s(l) - 2^{k_l-1}\pi_s^i \ge \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\left(\frac{1}{\sum_{\mathcal{S}_i} s}\right)\right]\\
&\le \sum_{s\in\mathcal{S}_i}\Pr\left[O_i^s(l) - 2^{k_l-1}\pi_s^i \ge \sqrt{L\log t_n}(\sqrt{2}-1)\sqrt{2^{k_l}}\left(\frac{1}{\sum_{\mathcal{S}_i} s}\right)\right]\\
&= |\mathcal{S}_i|\left(1 + \frac{(\sqrt{2}-1)\epsilon_i\sqrt{L\log t_n}}{10\sum_{s\in S_i} s}\frac{1}{\sqrt{2^{k_l}}}\right) \times N_{\mathbf{q}^i} t_n^{-((3-2\sqrt{2})L\epsilon_i)/(20(\sum_{s\in S_i} s)^2))} \qquad (36)
\end{aligned}$$

Thus, the regret caused by engaging bad arms in the nth exploitation epoch is bounded by

$$4^{n-1}2\sum_{j=2}^{N}(\mu_{\sigma(1)}-\mu_{\sigma(j)})\left[\left(\frac{1}{\log 2}+\frac{\sqrt{2}\epsilon_j\sqrt{L}}{10\sum_{s\in S_j}s}\right)|\mathcal{S}_j| + \left(\frac{1}{\log 2}+\frac{\sqrt{2}\epsilon_1\sqrt{L}}{10\sum_{s\in S_1}s}\right)|\mathcal{S}_1|\right]\frac{1}{\pi_{\min}}t_n^{-1}. \tag{43}$$

By (28) and $t_n \geq \frac{2}{3}4^{n-1}$, the bound in (43) becomes

$$3\lceil\log_4(\frac{3}{2}(t-N)+1)\rceil\frac{1}{\pi_{\min}}\left[\sum_{j=2}^{N}(\mu_{\sigma(1)}-\mu_{\sigma(j)}) \sum_{k=1,j}\left(\frac{1}{\log 2}+\frac{\sqrt{2}\epsilon_k\sqrt{L}}{10\sum_{s\in S_k}s}\right)|\mathcal{S}_k|\right]. \tag{44}$$

Combining (25), (29), (44), and (31), 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 [39] requires that all the observations used in calculating the sample means [$\bar{s}_i$ and $\bar{s}_j$ in (33)] 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 (33) is ensured by the carefully chosen (growing) lengths of the exploration and exploitation epochs.

Appendix B
Proof of Theorem 2

In this proof, we consider the zero-reward collision model. The extension to the reward-sharing collision model is straightforward. We first rewrite the definition of regret as

$$\begin{aligned} r_\Phi(t) &= t\sum_{i=1}^{M}\mu_{\sigma(i)} - \mathbb{E}_\Phi R(t) && (45)\\ &= \sum_{i=1}^{N}[\mu_i\mathbb{E}[T_i(t)] - \mathbb{E}[\sum_{n=1}^{T_i(t)} s_i(t_i(n))]] \\ &\quad + \left[t\sum_{i=1}^{M}\mu_{\sigma(i)} - \sum_{i=1}^{N}\mu_i\mathbb{E}[T_i(t)]\right] && (46)\end{aligned}$$

where $T_i(t)$ is the number of slots by t when arm i is played by one and only one player. Using Lemma 1 the first term in (46) can be bounded by (47) under the endogenous restless model (it is zero under the exogenous model):

$$(M\lceil\log_4(\frac{3t}{2M}+1)\rceil + N(\lfloor\log_4(3D\log t+1)\rfloor+1))MA_{\max} \tag{47}$$

which has a logarithmic order.

We are going to show that the second term in (45) has a logarithmic order. The upper bound we will find for the second term holds for both endogenous and exogenous restless model. It will be verified by bounding regret in both exploitation and exploration epochs by logarithmic order.

The upper bound on $T_O(t)$ in (30) still holds and consequently the regret caused by engaging bad arms in the exploration epochs by time t is upper bounded by

$$\frac{1}{3}[4(3D\log t+1)-1]\left(N\sum_{i=1}^{M}\mu_{\sigma(i)} - M\sum_{i=1}^{N}\mu_{\sigma(i)}\right). \tag{48}$$

The second reason for regret in the second term of (45) is not playing the expected arms in the exploitation epochs. If in the mth subepoch player k plays the $(m+k)\oslash M$ best arm, then every time the best M arms are played and there is no conflict. But arm $a^*_{(m+k)\oslash M}$ may not be the $(m+k)\oslash M$ 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 (43) 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 nth 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 nth exploitation epoch can be obtained as

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

By time t, we have

$$n_I(t) \leq \lceil\log_4(\frac{3t}{2M}+1)\rceil. \tag{49}$$

From the upper bound on the number of the exploitation epochs given in (28), 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 $r_{\Phi,I}(t)$):

$$\begin{aligned} &r_{\Phi,I}(t) \\ &\leq \sum_{m=1}^{M}\frac{\mu_m}{\pi_{\min}}\sum_{j=1}^{M}\sum_{i=1,i\neq j}^{N}\sum_{k=i,j}\left(\frac{1}{\log 2}+\frac{\sqrt{2}\epsilon_k\sqrt{L}}{10\sum_{s\in S_k}s}\right)|\mathcal{S}_k| \\ &\quad \times 3M\lceil\log_4(\frac{3t}{2M}+1)\rceil. && (50)\end{aligned}$$

Combining (45), (47), (48), and (50), we arrive at the upper bounds of regret given in (12).

APPENDIX C
PROOF OF THEOREM 3

At each time, regret incurs if one of the following three events happens: (i) at least one player is exploring, (ii) at least one player incorrectly identifies the set of M best arms in the exploitation sequence, and (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.

Clearly, the number of singular slots caused by (i) is logarithmic. We show that the number of singular slots caused by (ii) is logarithmic below. Define the event $F(k,t)$ as: Player k fails to identify the M best arms at time t. We have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} 1\{t \text{ is singular casued by (ii)}\}\right]\\
&= \sum_{t=1}^{T} \mathbb{E}\left[1\{t \text{ is singular casued by (ii)}\}\right]\\
&= \sum_{t=1}^{T} \Pr\left[1\{t \text{ is singular casued by (ii)}\}\right]\\
&\leq \sum_{t=1}^{T} \Pr\left[\cup_{k=1}^{M}\{F(k,t)\}\right]\\
&\leq \sum_{t=1}^{T}\sum_{k=1}^{M} \Pr\left[\{F(k,t)\}\right]\\
&= \sum_{k=1}^{M}\sum_{t=1}^{T} \Pr\left[\{F(k,t)\}\right]\\
&= \sum_{k=1}^{M} \mathbb{E}\Big[\sum_{t=1}^{T} 1\{F(k,t)\}\Big].
\end{aligned}$$

In the proofs for Theorems 1 and 3, it has been shown that the number of slots that each player fails to identify the M best arms is logarithmic. Thus, the number of slots caused by (ii) is also logarithmic.

To prove the theorem, it remains to show that the expected number of collisions in all nonsingular 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 nonsingular slots is bounded by the logarithmic order with time. We thus proved the theorem.

APPENDIX D
PROOF OF THEOREM 4

By any time t, for any l the expected time spent on a bad arm i is bounded by

$$\begin{aligned}
&2l + \sum_{t',T_i(t')>l,t'\leq t}\ \sum_{T_i(t')=1}^{t'}\\
&\sum_{T_\sigma(1)(t')=1}^{t'} 2T_i(t')\mathbf{1}\left[\bar{s}_i(t') \geq \bar{s}_{\sigma(1)}(t')\right]\\
&\leq 2l + \sum_{t',T_i(t')>l,t'\leq t} 2t' \sum_{T_i(t')=1}^{t'}\\
&\sum_{T_\sigma(1)(t')=1}^{t'} \mathbf{1}\left[\bar{s}_i(t') \geq \bar{s}_{\sigma(1)}(t')\right].
\end{aligned}$$

If we choose l to be $4L_1\ln(t)/((\mu_{\sigma(1)} - \mu_i)^2)$, then (40) still holds with t_n replaced by a general t, L replaced by L_1. Similarly (42) also holds. So we have

$$\begin{aligned}
\mathbb{E}[T_i(t)] \leq& \sum_{t'=1}^{\infty}\frac{2t'^{-2}}{\pi_{\min}}\left[\left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_{\sigma(1)}\sqrt{L_1}}{10\sum_{s\in S_{\sigma(1)}} s}\right)|\mathcal{S}_{\sigma(j)}|\right]\\
&+\sum_{t'=1}^{\infty}\frac{2t'^{-2}}{\pi_{\min}}\left[\left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_{i}\sqrt{L_1}}{10\sum_{s\in S_{i}} s}\right)|\mathcal{S}_{i}|\right] + 2l\\
\leq& \frac{2\pi}{3}\sum_{k=\sigma(1),i}\left(\frac{1}{\log 2} + \frac{\sqrt{2}\epsilon_{\sigma(1)}\sqrt{L_1}}{10\sum_{s\in S_{k}} s}\right)|\mathcal{S}_{k}|\frac{1}{\pi_{\min}}\\
&+\frac{8L_1\ln(t)}{(\mu_{\sigma(1)} - \mu_i)^2}.
\end{aligned}$$

Combined with the fact that the number of epochs by time t is $N(\lceil\log_2(t/N)\rceil + 1)$, we have the regret bound in (18).

APPENDIX E
PROOF OF THEOREM 5

Recall in Theorems 1 and 2, L and D are fixed *a priori*. Now we choose $L(t) \to \infty$ as $t \to \infty$ and $\frac{D(t)}{L(t)} \to \infty$ 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 of the regret is on a lower order or on the same order of $f(t)\log t$. In this proof, we show details for single player and the corresponding details for decentralized multiple players can be done similarly.

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

$$N\log_2(t/N+1)A_{\max}. \quad (51)$$

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

$$\lim_{t\to\infty}\frac{N\log_2(t/N+1)\max_i A_i}{f(t)\log t} = 0. \quad (52)$$

Thus, the regret caused by arm switching is on a lower order than $f(t)\log t$. By similar reasoning, regret from arm switching for decentralized player is on a lower order than $f(t)\log t$.

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

$$\frac{1}{3}[4(3D(t)\log t+1)-1]\left(N\mu_{\sigma(1)}-\sum_{i=1}^{N}\mu_{\sigma(i)}\right). \tag{53}$$

Thus, the regret caused by playing bad arms in the exploration epochs is on the same order of $f(t)\log t$. The corresponding regret for decentralized multiple player is

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

which 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 $\frac{D(t)}{L(t)}\rightarrow\infty$ 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 (32) to (44)]

$$\frac{\pi^2}{2}\frac{|\mathcal{S}_i|+|\mathcal{S}_{\sigma(1)}|}{\pi_{\min}}(1+\frac{\epsilon_{\max}\sqrt{L(t_5)}}{10s_{\min}}). \tag{54}$$

An upper bound for the corresponding regret is

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

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 $f(t)\log t$. The corresponding regret for decentralized multiple player 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 $f(t)\log t$.

REFERENCES

[1] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[2] R. Agrawal, "Sample mean based index policies with O(log n) regret for the multi-armed bandit problem," *Adv. Appl. Probabil.*, vol. 27, pp. 1054–1078, 1995.

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

[4] K. Liu and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," presented at the Allerton Conf. Commun., Control, Comput., Sep. 2011.

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

[6] C. Tekin and M. Liu, "Online algorithms for the multi-armed bandit problem with Markovian rewards," presented at the Allerton Conf. Commun., Control, Comput., Sep. 2010.

[7] H. Liu, K. Liu, and Q. Zhao, Learning in a changing world: Non-Bayesian restless multi-armed bandit Oct. 2010 [Online]. Available: http://arxiv.org/abs/1011.4969

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

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

[10] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, Aug. 2012.

[11] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao, "The non-Bayesian restless multi-armed bandit: A case of near-logarithmic regret," presented at the Int. Conf. Acoust., Speech Signal Process., May 2011.

[12] C. Tekin and M. Liu, "Approximately optimal adaptive learning in opportunistic spectrum access," presented at the Int. Conf. Comput. Commun., Orlando, FL, Mar. 2012.

[13] C. Tekin and M. Liu, "Adaptive learning of uncontrolled restless bandits with logarithmic regret," presented at the Allerton Conf. Commun., Control, Comput., Sep. 2011.

[14] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multichannel opportunistic access: Structure, optimality, and performance," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5431–5440, Dec. 2008.

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

[16] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," *J. ACM*, vol. 58, no. 1, Dec. 2010.

[17] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, pp. 1563–1600, Apr. 2010.

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

[19] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE JSAC Adv. Cognit. Radio Network. Commun.*, vol. 29, no. 4, pp. 731–745, Mar. 2011.

[20] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," presented at the IEEE Global Commun. Conf., Houston, TX, Dec. 2011.

[21] C. Tekin and M. Liu, "Performance and convergence of multiuser online learning," presented at the Int. Conf. Game Theory Netw., Apr. 2011.

[22] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized multi-armed bandits," presented at the Inf. Theory Appl., Feb. 2012.

[23] N. Cesa-Bianchi and P. Fischer, "Finite-time regret bounds for the multiarmed bandit problem," in *Proc. Int. Conf. Mach. Learn.*, 1998, pp. 100–108.

[24] R. Bellman, "A problem in the sequential design of experiments," *Sankhia*, vol. 16, pp. 221–229, 1956.

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

[26] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25, pp. 287–298, 1988.

[27] 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.

[28] R. R. Weber and G. Weiss, "Addendum to 'On an index policy for restless bandits'," *Adv. Appl. Probab.*, vol. 23, no. 2, pp. 429–430, Jun. 1991.

[29] K. D. Glazebrook and H. M. Mitchell, "An index policy for a stochastic scheduling model with improving/deteriorating jobs," *Naval Res. Logist.*, vol. 49, pp. 706–721, Mar. 2002.

[30] P. S. Ansell, K. D. Glazebrook, J. E. Nino-Mora, and M. O'Keeffe, "Whittle's index policy for a multi-class queueing system with convex holding costs," *Math. Meth. Operat. Res.*, vol. 57, pp. 21–39, 2003.

[31] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride, "Some indexable families of restless bandit problems," *Adv. Appl. Probab.*, vol. 38, pp. 643–672, 2006.

[32] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5547–5567, Nov. 2010.

[33] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

[34] M. Agarwal, V. S. Borkar, and A. Karandikar, "Structural properties of optimal transmission policies over a randomly varying channel," *IEEE Trans. Wireless Commun.*, vol. 53, no. 6, pp. 1476–1491, Jul. 2008.

[35] S. Ali, V. Krishnamurthy, and V. Leung, "Optimal and approximate mobility assisted opportunistic scheduling in cellular data networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 6, pp. 633–648, Jun. 2007.

[36] L. Johnston and V. Krishnamurthy, "Opportunistic file transfer over a fading channel—A POMDP search theory formulation with optimal threshold policies," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, pp. 394–405, Feb. 2006.

[37] M. Sorensen, "Learning by investing: Evidence from venture capital," presented at the Amer. Finance Assoc. Annu. Meet., Feb. 2008.

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

[39] D. Gillman, "A Chernoff bound for random walks on expander graphs," in *Proc. 34th IEEE Symp. Found. Comput. Sci.*, 1998, pp. 680–691.

Haoyang Liu received the B.S. degree in Electrical Engineering from Shanghai Jiao Tong University, Shanghai, China in 2009. He is currently a Ph.D. candidate in Graduate Group of Applied Mathematics at University of California, Davis. His research interests include stochastic optimization, time series and high dimensional statistics.

Keqin Liu (S'07–M'11) received the B.S. degree in automation from Southeast University, Nanjing, China, in 2005 and the M.S. and Ph.D. degrees in electrical engineering from the University of California at Davis, Davis, in 2008 and 2010, respectively. He received the 2012 Zuhair A. Munir Award for Best Doctoral Dissertation from UC-Davis.

Qing Zhao (S'97–M'02–SM'08–F'13) received the Ph.D. degree in Electrical Engineering in 2001 from Cornell University, Ithaca, NY. In August 2004, she joined the Department of Electrical and Computer Engineering at University of California, Davis, where she is currently a Professor. Her research interests are in the general area of stochastic optimization, decision theory, and algorithmic theory in dynamic systems and communication and social networks.

She received the 2010 IEEE Signal Processing Magazine Best Paper Award and the 2000 Young Author Best Paper Award from the IEEE Signal Processing Society. She holds the title of UC Davis Chancellor's Fellow and received the 2008 Outstanding Junior Faculty Award from the UC Davis College of Engineering. She was a plenary speaker at the 11th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2010. She is also a co-author of two papers that received student paper awards at ICASSP 2006 and the IEEE Asilomar Conference 2006.