

Distributed Learning in Cognitive Radio Networks: Multi-Armed Bandit with Distributed Multiple Players

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

Abstract—We consider a cognitive radio network with distributed multiple secondary users, where each user independently searches for spectrum opportunities in multiple channels without exchanging information with others. The occupancy of each channel is modeled as an i.i.d. Bernoulli process with unknown mean. Users choosing the same channel collide, and none or only one receives reward depending on the collision model. This problem can be formulated as a decentralized multi-armed bandit problem. We measure the performance of a decentralized policy by the system regret, defined as the total reward loss with respect to the optimal performance under the perfect scenario where all channel parameters are known to all users and collisions among secondary users are eliminated through perfect scheduling. We show that the minimum system regret grows with time at the same logarithmic order as in the centralized counterpart, where users exchange observations and make decisions jointly. A decentralized policy is constructed to achieve this optimal order. Furthermore, we show that the proposed policy belongs to a general class of decentralized policies, for which a uniform performance benchmark is established. All results hold for general stochastic processes beyond Bernoulli and for collision models with or without carrier sensing.

Index Terms—cognitive radios, opportunistic spectrum access, decentralized multi-armed bandit, order-optimal policy.

I. INTRODUCTION

We consider a cognitive radio network consisting of N independent channels, where M ($M < N$) distributed secondary users independently search for idle channels temporarily unoccupied by the primary network. We assume that the state—1 (idle) or 0 (busy)—of each channel evolves as an i.i.d. Bernoulli process across time slots with mean θ_i and the parameter set $\Theta = (\theta_1, \dots, \theta_N)$ is unknown to all users. At the beginning of each slot, each secondary user chooses one channel to sense and subsequently transmits if the sensed channel is idle. Users do not exchange information on their decisions and observations. Collisions may occur when multiple users choose the same channel, and none or only one receives reward depending on the collision model. The objective is to design a decentralized channel selection policy for optimal network throughput.

The above problem motivates an extension to the classic multi-armed bandit (MAB) problem that considers only a single player. A commonly used performance measure under a non-Bayesian formulation is the so-called regret: the performance loss with respect to the genie-aided case where the distribution of each arm is perfectly known. The classic MAB with a single user under the assumption of single play (only one arm can be chosen at each time) was solved by Lai and Robbins in 1985 [1], where they showed that the minimum regret grows with time at a logarithmic order and constructed a policy that achieves the optimal regret growth rate. Anantharam *et al.* extended Lai and Robbins's results to MAB with multiple plays: exactly M ($M < N$) arms can be played simultaneously at each time [2]. They showed that allowing simultaneous multiple plays

changes only the constant but not the logarithmic order of the regret growth rate. They also extended Lai and Robbins's policy to achieve the optimal regret growth rate under multiple plays.

The single-user MAB with multiple plays considered in [2] is equivalent to a *centralized* MAB with multiple users. If all M users can exchange their observations and make decisions jointly, they act collectively as a single user who has the freedom of choosing M arms simultaneously. As a direct consequence, the optimal regret growth rate established in [2] provides a lower bound on the optimal performance of a *decentralized* MAB where users cannot exchange observations and must make decisions independently based on local observations.

Main Results The main questions we aim to answer in this paper are the optimal order of the regret growth rate in a decentralized MAB and how to construct decentralized policies to achieve the optimal order. We show that in the decentralized setting where users can only learn from their individual observations and collisions are bound to happen, the system can achieve the same logarithmic order in the total regret growth rate as in the centralized case. A decentralized policy is constructed to achieve this optimal order. Furthermore, we show that the order-optimal decentralized policy belongs to a general class of decentralized policies. For this class of decentralized policies, we establish a performance benchmark by constructing a lower bound on the achievable growth rate of the system regret. This lower bound is tighter than the trivial bound provided by the centralized system considered in [2]. All results hold for general stochastic processes beyond Bernoulli and for collision models with or without carrier sensing.

A distinct feature of the proposed decentralized policy is that it ensures *fairness* among users, *i.e.*, all users achieve the same average reward at the same rate. Note that without knowing the reward rate that each channel can offer, ensuring fairness requires that each user identify the entire set of the M best arms and share each of these M arms evenly with other users. As a consequence, each user needs to learn which of the C_M^N possibilities is the correct choice. Another approach to sharing the M best arms is that each user aims at identifying and exploiting a single arm with a specific rank (for example, the i th user targets solely at the i th best arm). In this case, each user only needs to distinguish one arm (with a specific rank) from the rest. The uncertainty facing each user, consequently, the amount of learning required, is thus reduced from C_M^N to N . Unfortunately, fairness among users is lost. We further point out that under the proposed policy, each user identifies not only the set of the M best arms, but also the entire rank of these M arms.

Related Work The problem of searching for time-varying spectrum opportunities in multiple channels has been considered in [3]–[5] from a single user's perspective. A Markovian model on the channel occupancy is adopted in [3]–[5], which is more general than the i.i.d. Bernoulli process considered in this paper. On the other hand, even with known transition probabilities and with a single user,

⁰This work was supported by the Army Research Laboratory under Grant DAAD19-01-C-0062, by the Army Research Office under Grant W911NF-08-1-0467 and by the National Science Foundation under Grant CCF-0830685.

the Markovian channel model results in a Partially Observable Markovian Decision Process (POMDP) or a restless multi-armed bandit problem [3]–[5]. Both are significantly more complex classes of sequential decision problems as compared to the classic MAB.

The assumption on known channel transition probabilities can be relaxed under certain conditions while maintaining the optimal performance defined by a known channel model, as shown in [4], [5], which establish the universal or semi-universal structures of the optimal policies. Extensions to multiuser case under the Markovian channel model have been considered in [6], [7], where the problem is formulated as a decentralized POMDP and heuristic policies are proposed. Unfortunately, finding the optimal policy for a decentralized POMDP is, in general, intractable [8].

In this paper, we adopt a simpler channel occupancy model—i.i.d. Bernoulli processes with unknown means. Without time correlation in the channel occupancy model, addressing multiple distributed users and unknown channel parameters becomes tractable. The same problem has been considered in [9], where a heuristic randomized policy is proposed. This policy, however, only achieves the linear order of the system regret and thus cannot achieve the maximum time-average reward.

In a broader context, the problem considered in this paper gives rise to a decentralized formulation of the classic MAB with multiple players, which has not been clearly formulated or optimally solved in the literature. The significance of this work lies in establishing the same logarithmic order of the system regret as in the centralized counterpart and constructing a decentralized policy to achieve this optimal order.

Notation For two positive integers K and L , define $K \oslash L$ as the integer in $\{1, \dots, L\}$ satisfying $K \oslash L = K - DL$ for some integer $D \geq 0$.

II. CLASSIC RESULTS ON SINGLE-USER MAB

In this section, we give a brief review of the main results established in [1], [2] for the classic MAB with a single player.

Consider an N -arm bandit with a single player. At each time t , the player can choose exactly M ($1 \leq M < N$) arms to play. Playing arm i yields a random reward Y_i drawn from a univariate density function $f(y; \theta_i)$ parameterized by θ_i . The function $f(\cdot; \cdot)$ is known and the parameter set $\Theta = (\theta_1, \dots, \theta_N)$ is unknown to the player. For the simplicity of the presentation, we assume that $y \in \{0, 1\}$ and $f(y; \theta_i) = \theta_i^y (1 - \theta_i)^{1-y}$, i.e., the reward process on each arm is an i.i.d. bernoulli process as in the adopted channel model for cognitive radio networks.

A policy $\pi = \{\pi(t)\}_{t=1}^{\infty}$ is a series of functions, where $\pi(t)$ maps the previous observations of rewards to the current action that specifies the set of M arms to play in slot t . The system performance under policy π is measured by the system regret $R_T^\pi(\Theta)$ over a time horizon of length T as defined below. Let σ be a permutation of $\{1, \dots, N\}$ such that $\theta_{\sigma(1)} \geq \theta_{\sigma(2)} \geq \dots \geq \theta_{\sigma(N)}$, we have

$$R_T^\pi(\Theta) \triangleq T \sum_{j=1}^M \theta_{\sigma(j)} - \mathbb{E}_\pi[\sum_{t=1}^T Y_{\pi(t)}(t)], \quad (1)$$

where $Y_{\pi(t)}(t)$ is the random reward obtained in slot t under $\pi(t)$, and $\mathbb{E}_\pi[\cdot]$ denotes the expectation under the policy π . In other words, $R_T^\pi(\Theta)$ is the expected total reward loss up to time T under policy π compared to the perfect scenario that Θ is known to the player (leading to a total reward of $T \sum_{j=1}^M \theta_{\sigma(j)}$, the first term of (1)). The objective is to minimize the rate at which $R_T(\Theta)$ grows with T under any parameter set Θ by choosing the optimal policy π^* .

A policy is called *uniformly good* if for any parameter set Θ , we have $R_T^\pi(\Theta) = o(T^b)$ for any $b > 0$. Note that a uniformly good policy implies sub-linear growth of the system regret and achieves

the maximum time-average reward $\sum_{j=1}^M \theta_{\sigma(j)}$ which is the same as in the case with perfect knowledge of Θ .

Theorem [1,2]: For any uniformly good policy π ,

$$\liminf_{T \rightarrow \infty} \frac{R_T^\pi(\Theta)}{\log T} \geq \sum_{j: \theta_j < \theta_{\sigma(M)}} \frac{\theta_{\sigma(M)} - \theta_j}{I(\theta_j, \theta_{\sigma(M)}),} \quad (2)$$

where $I(\theta_i, \theta_j) = \theta_i \log(\theta_i/\theta_j) + (1 - \theta_i) \log((1 - \theta_i)/(1 - \theta_j))$ is the K-L distance between the reward distributions parameterized by θ_i and θ_j , respectively. Lai and Robbins [1] in 1985 have constructed a policy to solve the classic MAB with single play ($M = 1$), which is presented as follows.

Lai and Robbins's Policy for Single-User MAB with Single-Play [1] In the first N slots, play each arm once. Fix δ ($0 < \delta < 1/N$). For all $t > N$, let $\tilde{\theta}_n(t)$ denote the sample mean of the reward obtained from arm n and $\tau_{n,t}$ the times arm n is played up to (but excluding) slot t . Among all arms that have been played at least $(t - 1)\delta$ times, select the leader l_t that has the largest sample mean. Let $j = t \oslash N$. The player plays arm j if $\tilde{\theta}_j(t) \geq \tilde{\theta}_{l_t}(t)$ or $I(\tilde{\theta}_j(t), \tilde{\theta}_{l_t}(t)) \leq \log t / \tau_{j,t}$; otherwise the player plays arm l_t .

Lai and Robbins [1] have shown that their policy is asymptotically optimal (i.e., it achieves the lower bound on the regret growth rate given in (2)). Anantharam *et al.* [2] in 1987 have extended Lai and Robbins's results to multiple plays ($M > 1$).

III. PROBLEM FORMULATION

Consider the spectrum consisting of N independent but nonidentical channels (arms). Let $\mathbf{S}(t) = [S_1(t), \dots, S_N(t)] \in \{0, 1\}^N$ ($t \geq 1$) denote the system state, where $S_i(t)$ is the state of channel i in slot t . Assume that $\{S_i(t)\}_{t \geq 1}$ is an i.i.d. Bernoulli process with unknown mean θ_i . We assume that the M largest means are distinct.

In slot t , a user (say user i ($1 \leq i \leq M$)) chooses a sensing action $a_i(t) \in \{1, \dots, N\}$ that specifies the channel to sense based on its sensing and observation history. Sensing is assumed to be accurate. If the channel is sensed to be busy, the user refrains from transmission. If the channel is idle, the user transmits with or without carrier sensing as detailed below. Note that the user can also learn the system from previous identified collisions to others. Its local observation history thus consists of both the previous observed channel states and the collision history.

We define a local policy π_i for user i as a sequence of functions $\pi_i = \{\pi_i(t)\}_{t \geq 1}$, where $\pi_i(t)$ maps user i 's past observations and decisions to $a_i(t)$ in slot t . The decentralized policy π is thus given by the concatenation of the local policy for each user:

$$\pi = [\pi_1, \dots, \pi_M].$$

Define immediate reward $Y(t)$ as the total number of successful transmissions by all users in slot t , which depends on the system collision model given below¹.

Collision model 1 (with carrier sensing): When the channel is sensed to be idle, the user generates a random backoff time and transmits when the time expires and no other secondary users have claimed the channel. Under this model, we have

$$Y(t) = \sum_{j=1}^N \mathbb{I}_j(t) S_j(t),$$

where $\mathbb{I}_j(t)$ is the indicator function that equals to 1 if channel j is sensed by at least one user, and 0 otherwise.

Collision model 2 (without carrier sensing): When the channel is sensed to be idle, the user will transmit. The transmission will be successful if and only if no other users have chosen the same channel. Under this model, we have

$$Y(t) = \sum_{j=1}^N \mathbb{I}'_j(t) S_j(t),$$

¹The results apply to more general collision models, including the case that all users involved in a collision receive reward.

where $\mathbb{I}_j'(t)$ is the indicator function that equals to 1 if channel j is sensed by only one user, and 0 otherwise.

Similar to the classic MAB, we use the following system regret as the performance measurement of a decentralized policy π .

$$R_T^\pi(\Theta) = T \sum_{j=1}^M \theta_{\sigma(j)} - \mathbb{E}_\pi[\sum_{t=1}^T Y(t)].$$

The objective is to minimize the rate at which $R_T(\Theta)$ grows with time T under any parameter set Θ by choosing the optimal decentralized policy π^* . Similarly, we call a decentralized policy is *uniformly good* if for any parameter set Θ , we have $R_T(\Theta) = o(T^b)$ for any $b > 0$. To address the optimal order of the regret, it is sufficient to focus on uniformly good decentralized policies provided that such policies exist.

IV. THE OPTIMAL ORDER OF THE SYSTEM REGRET

In this section, we show that the optimal order of the rate at which the system regret grows with T in the decentralized MAB is logarithmic.

Theorem 1: Under both collision models, the optimal order of the rate at which the system regret grows with T in the decentralized MAB is logarithmic, *i.e.*, for an optimal decentralized policy π^* , we have

$$L(\Theta) \leq \liminf_{T \rightarrow \infty} \frac{R_T^{\pi^*}(\Theta)}{\log T} \leq \limsup_{T \rightarrow \infty} \frac{R_T^{\pi^*}(\Theta)}{\log T} \leq U(\Theta) \quad (3)$$

for some constants $L(\Theta)$ and $U(\Theta)$ that depend on Θ .

Proof: Note that the growth rate of the system regret in the centralized MAB given in (2) provides a lower bound for the decentralized MAB. For the upper bound, we construct a decentralized policy (see Sec. V) that achieves the logarithmic order of the regret growth rate. Details can be found in [10]. ■

V. AN ORDER-OPTIMAL DECENTRALIZED POLICY

In this section, we construct a decentralized policy $\tilde{\pi}$ that achieves the optimal logarithmic order of the system regret growth rate.

The basic structure of the proposed policy $\tilde{\pi}$ is a round-robin structure at each user for selecting those M best channels. Users have different phases (offsets) in their round-robin schedule to avoid excessive collisions. Consider, for example, the case of $M = 2$. User 1 targets at the best channel in odd slots and the second best channel in even slots, and user 2 does the opposite. For each user, selecting the best channel can be done efficiently using Lai and Robbins's single-user policy. The question here is how to select the second best channel efficiently. One intuitive approach is to apply Lai and Robbins's policy to the remaining $N - 1$ channels after removing the channel considered as the best channel. Specifically, consider user 1 who targets at the second best channel in even slots. All the even slots are divided into N interleaving subsequences, where the i th subsequence consists of all the even slots that follow an odd slot in which channel i is considered as the best channel. In the i th subsequence, user 1 applies Lai and Robbins's policy to channels $\{1, \dots, i - 1, i + 1, \dots, N\}$. In other words, the local policy for each user consists of multiple parallel Lai and Robbins's procedures applied in different subsequences of time slots to different subsets of channels. These parallel procedures, however, are coupled through the common observation history since in each slot (regardless of which subsequence it belongs to), all the past observations are used in decision making. A detailed implementation of the proposed policy for a general M is given in Fig. 1.

We point out that $\tilde{\pi}$ is distributed in the sense that users do not exchange information and make decisions solely based on local observations. The offset in each user's round-robin schedule can be

The Decentralized Policy $\tilde{\pi}$

Consider user i . Choose a δ such that $0 < \delta < 1/N$. Consider a slot t . Let $k_t = t \oslash M$ and $m_t(j)$ the total number of events that $k_t = j$ ($1 \leq j \leq M$) up to time t . Let $\hat{\theta}_n(t)$ denote the sample mean of the state of channel n and $\tau_{n,t}$ the times channel n is played up to (but excluding) slot t .

1. If $k = i$, go to Step 2; otherwise go to Step 3.
2. If $m_t(i) \leq N$, sense channel $m_t(i)$; otherwise, the user does the following. Among all channels that have been sensed for at least $\delta(m_t(i) - 1)$ slots, first select a leader l_t that has the largest sample mean. Let $n = m_t(i) \oslash N$. The user senses channel n if and only if $\hat{\theta}_n(t) \geq \hat{\theta}_{l_t}(t)$ or $I(\hat{\theta}_n(t), \hat{\theta}_{l_t}(t)) \leq \log(t - 1)/\tau_{n,t}$; otherwise the user chooses channel l_t .
3. If $m_t(k_t) \leq N$, sense channel $m_t(k_t)$; otherwise, the user does the following. Let $j = (k_t - i + M + 1) \oslash M$. The user removes the $j - 1$ arms played in the previous $j - 1$ slots from the arm set $\mathcal{N} = \{1, \dots, N\}$. Assume a channel set $a \subset \mathcal{N}$ has been removed. Let $m^a(k_t)$ denote the number of events that channel set a has been removed up to time t . We relabel the remaining channels in $\mathcal{N} \setminus a$ with indices $\{1, 2, \dots, N - j + 1\}$ according to the order of their original indices. Among all the remaining channels that have been sensed for at least $\delta(m^a(k_t) - 1)$ slots, first select a leader l_t that has the largest sample mean. Let $n = m^a(k_t) \oslash (N - j + 1)$. The user senses channel n if and only if $\hat{\theta}_n(t) \geq \hat{\theta}_{l_t}(t)$ or $I(\hat{\theta}_n(t), \hat{\theta}_{l_t}(t)) \leq \log(t - 1)/\tau_{n,t}$; otherwise the user chooses channel l_t . Reset the indices of the remaining channels as the original ones.

Fig. 1. The Structure of the Decentralized Policy $\tilde{\pi}$

predetermined (for example, based on the user's ID). The policy can thus be implemented in a distributed fashion.

The theorem below establishes the order optimality of the proposed policy $\tilde{\pi}$. The difficulties in establishing the logarithmic order of $\tilde{\pi}$ as compared to the centralized counterpart given in [1], [2] are two folds. First, since the rank of any channel can never be perfectly identified, the mistakes in identifying the i th ($1 \leq i < M$) best channel will propagate into the learning process for identifying the $(i + 1)$ th, up to the M th best channels. Second, since users are learning the channel statistics based on their individual local observations, they do not always agree on the rank of the channels. Collisions are bound to happen (for example, when a channel is considered as the best by one user but the second best by the other). Such issues need to be carefully dealt with in establishing the logarithmic order of the regret growth rate.

Theorem 2: Under the decentralized policy $\tilde{\pi}$, we have

$$\limsup_{T \rightarrow \infty} \frac{R_T^{\tilde{\pi}}(\Theta)}{\log T} \leq C(\Theta), \quad (4)$$

where under collision model 1,

$$\begin{aligned} C(\Theta) &= M(\sum_{k=1}^M x_k \theta_{\sigma(k)} - \sum_{j: \theta_j < \theta_{\sigma(M)}} \theta_j / I(\theta_j, \theta_{\sigma(M)})), \\ x_k &= \sum_{i=1}^k \sum_{j: \theta_j < \theta_{\sigma(i)}} 1 / I(\theta_j, \theta_{\sigma(i)}), \end{aligned}$$

and under collision model 2,

$$\begin{aligned} C(\Theta) &= M(\sum_{j=1}^M \sum_{k=1}^M x_k \theta_{\sigma(j)} - \sum_{j: \theta_j < \theta_{\sigma(M)}} \theta_j \max\{1 / I(\theta_j, \theta_{\sigma(M)}) \\ &\quad - \sum_{i=1}^{M-1} 1 / I(\theta_j, \theta_{\sigma(i)}), 0\}). \end{aligned}$$

Define the local regret for user i as

$$R_{T,i}^{\tilde{\pi}}(\Theta) \triangleq \frac{1}{M} T \sum_{j=1}^M \theta_{\sigma(j)} - \mathbb{E}_{\tilde{\pi}}[\sum_{t=1}^T Y_i(t)],$$

where $Y_i(t)$ is the immediate reward obtained by user i in slot t , we have

$$\limsup_{T \rightarrow \infty} \frac{R_{T,i}^{\tilde{\pi}}(\Theta)}{\log T} = \frac{1}{M} \limsup_{T \rightarrow \infty} \frac{R_T^{\tilde{\pi}}(\Theta)}{\log T}. \quad (5)$$

Proof: See [10]. ■

From Theorem 1 and Theorem 2, the decentralized policy is order-optimal. Furthermore, (5) shows that $\tilde{\pi}$ ensures fairness among users: each user achieves the same time-average reward $\frac{1}{M} T \sum_{j=1}^M \theta_{\sigma(j)}$ at the same rate.

VI. A LOWER BOUND FOR A CLASS OF DECENTRALIZED POLICES

In this section, we establish a uniform lower bound on the growth rate of the system regret for a general class of decentralized policies, to which the proposed policy $\tilde{\pi}$ belongs. This lower bound provides a tighter performance benchmark compared to the one defined by the centralized MAB. The definition of this class of decentralized policies is given below.

Definition 1: Time Division Selection Policies The class of time division selection (TDS) policies consists of all decentralized policies $\pi = [\pi_1, \dots, \pi_M]$ that satisfy the following property: under any policy π_i , there exists $0 \leq a_{ij} \leq 1$ ($1 \leq i, j \leq M$, $\sum_{j=1}^M a_{ij} = 1 \forall i$) independent of the parameter set Θ such that the expected number of slots that user i chooses the j th ($1 \leq j \leq M$) best channel to sense up to time T is equal to $a_{ij}T - o(T^b)$ for all $b > 0$.

A policy in the TDS class essentially allows a user to efficiently select each of the M best arms according to a fixed time portion that does not depend on the parameter set Θ . It can be shown that the order-optimal decentralized policy $\tilde{\pi}$ given in Fig. 1 belongs to the TDS class with $a_{ij} = 1/M$ for all $1 \leq i, j \leq M$.

In the following, we establish a lower bound on the rate at which the system regret grows with T for all uniformly good decentralized policies in the TDS class.

Theorem 3: For any uniformly good decentralized policy π in the TDS class, we have

$$\liminf_{T \rightarrow \infty} \frac{R_T^{\pi}(\Theta)}{\log T} \geq \sum_{j:\theta_j < \theta_{\sigma(M)}} \sum_{i=1}^M \frac{\theta_{\sigma(M)} - \theta_j}{I(\theta_j, \theta_{\sigma(i)})}. \quad (6)$$

Proof: See [10]. ■

VII. EXTENSIONS TO GENERAL STOCHASTIC PROCESSES

In this section, we generalize all results to arbitrary i.i.d. stochastic processes.

For general stochastic i.i.d. processes, the state $S_n(t) \in \mathcal{R}$ of arm n is drawn from a univariate density function $f(s; \theta_n)$ with unknown parameter $\theta_n \in \Theta$. Let $\mu(\theta)$ denote the mean of a random variable S under the density function $f(s; \theta)$.

While we have focused on Bernoulli processes in Sec. II, the classic results on centralized MAB given in [1,2] hold for general stochastic processes that satisfy the following conditions.

- C1. $\mu(\theta)$ exists for any $\theta \in \Theta$.
- C2. $0 < I(\theta, \theta') < \infty$ whenever $\mu(\theta) < \mu(\theta')$.
- C3. $\forall \epsilon > 0$ and $\mu(\theta) < \mu(\theta')$, $\exists \delta > 0$ such that $|I(\theta, \theta') - I(\theta, \theta'')| < \epsilon$ whenever $\mu(\theta'') \leq \mu(\theta') \leq \mu(\theta) + \delta$.
- C4. $\forall \theta$ and $\delta > 0$, $\exists \theta' > 0$ such that $\mu(\theta) < \mu(\theta') < \mu(\theta) + \delta$.
- C5. Let X_1, X_2, \dots be i.i.d. random variables with the common density function $f(x, \theta)$. Assume there exist Borel functions $h_i(X_1, \dots, X_i)$ for $i \geq 1$ and $g_{t,i}(X_1, \dots, X_i)$ for $t \geq 1, 1 \leq i < t$ such that for any $\theta \in \Theta$,

$$\Pr_{\theta}\{g_{t,i}(X_1, \dots, X_i) \geq r \text{ for all } i < t\} = 1 - o(t^{-1}) \text{ for every } r < \mu(\theta);$$

$$\begin{aligned} \lim_{\epsilon \downarrow 0} (\limsup_{t \rightarrow \infty} \sum_{i=1}^{t-1} \Pr_{\theta}\{g_{t,i}(X_1, \dots, X_i) \geq \mu(\theta') - \epsilon\} / \log t) \\ \leq 1/I(\theta, \theta') \text{ whenever } \mu(\theta') > \mu(\theta); \\ g_{t,i} \text{ is nondecreasing in } t \text{ for every fixed } i = 1, 2, \dots; \\ g_{t,i} \geq h_i \forall t; \\ \Pr\{\max_{\theta} \max_{\delta t \leq i \leq t} |h_i - \mu(\theta)| > \epsilon\} = o(t^{-1}) \forall \epsilon > 0, 0 < \delta < 1. \end{aligned}$$

Note that conditions C1–C5 are satisfied for Gaussian, Bernoulli, Poisson, and double exponential distributions, where h_i and $g_{t,i}$ have been established in [1].

For the decentralized MAB, we show that all results established in Sec. IV–VI can be extended to general stochastic processes. In addition to conditions C1–C5 required by the centralized MAB, we require that the M best arms have distinct nonnegative means. For any i.i.d. stochastic processes satisfying these conditions, Theorem 1 holds. Furthermore, both the decentralized policy $\tilde{\pi}$ and the lower bound given in Theorem 3 can be obtained in general forms. Details can be found in [10].

VIII. CONCLUSION

In this paper, we have considered a decentralized multi-armed bandit problem with distributed multiple players. This problem is motivated by distributed spectrum sharing in cognitive radio networks. We have shown that the optimal system regret in the decentralized MAB grows at the same logarithmic order as that in the centralized MAB considered in the classic work by Lai and Robbins [1] and Anantharam, *et al.* [2]. A decentralized policy that achieves the optimal order has been constructed.

REFERENCES

- [1] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4C22, 1985.
- [2] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part I: IID rewards," *IEEE Tran. on Auto. Control*, vol. 32, no. 11, pp. 968C976, 1987.
- [3] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework," in *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589-600, April, 2007.
- [4] Y. Chen, Q. Zhao, and A. Swami, "Joint Design and Separation Principle for Opportunistic Spectrum Access in the Presence of Sensing Errors," in *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2053-2071, May, 2008.
- [5] Q. Zhao, B. Krishnamachari, and K. Liu, "On Myopic Sensing for Multi-Channel Opportunistic Access: Structure, Optimality, and Performance," in *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5431-5440, Dec., 2008.
- [6] K. Liu, Q. Zhao, and Y. Chen "Distributed Sensing and Access in Cognitive Radio Networks," in *Proc. of 10th International Symposium on Spread Spectrum Techniques and Applications (ISSSTA)*, August, 2008.
- [7] H. Liu, B. Krishnamachari, and Q. Zhao, "Cooperation and Learning in Multiuser Opportunistic Spectrum Access," in *Proc. of IEEE International Conference on Communications (ICC) Workshops*, May, 2008.
- [8] D. Bernstein, "Complexity Analysis and Optimal Algorithms for Decentralized Decision Making," *PhD thesis*, University of Massachusetts Amherst, September 2005.
- [9] L. Lai, H. Jiang and H. Vincent Poor, "Medium Access in Cognitive Radio Networks: A Competitive Multi-armed Bandit Framework," in *Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2008.
- [10] K. Liu and Q. Zhao, "Decentralized Multi-Armed Bandit with Multiple Distributed Players," Technical Report (TR-09-03), Sep. 2009, <http://www.ece.ucdavis.edu/~qzhao/TR-09-03.pdf>.