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. We propose a basic policy structure that ensures a Time Division Fair Sharing (TDFS) of the channels. Based on this basic TDFS structure, decentralized policies can be constructed to achieve this optimal order while ensuring fairness among users. Furthermore, we show that the proposed TDFS 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 thus can be extended to i.i.d. processes. We assume that the state—(idle) or (busy)—of each channel evolves as an i.i.d. Bernoulli process across time slots with mean \( \theta \), and the parameter set \( \Theta = (\theta_1, \ldots, \theta_N) \) 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 \( M \) 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-player 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 regret 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. We propose a basic policy structure that ensures a Time Division Fair Sharing (TDFS) of the \( M \) best arms. Based on this basic TDFS structure, decentralized policies can be constructed to achieve this optimal order while ensuring fairness among users. While we present the TDFS policy using Lai and Robbins’s single-player policy as the building block, the basic structure of TDFS is not tied with a specific single-player policy. It can be used with any single-player policy to achieve an efficient and fair sharing of the \( M \) best arms among \( M \) users, and users can use different single-player policies to build their local policies. More specifically, if the single-player policies achieve the optimal logarithmic order in the centralized setting, then the corresponding TDFS policy achieves the optimal logarithmic order in the decentralized setting.

We also establish a lower bound on the achievable growth rate of the system regret for a general class of decentralized policies, to which the proposed TDFS policy belongs. This lower bound is tighter (i.e., larger) than the trivial bound provided by the centralized MAB considered in [2], which indicates, as one would expect, that decentralized MAB is likely to incur a larger leading constant than its centralized counterpart. All results hold for general reward models beyond Bernoulli and for collision models with or without carrier sensing.

Related Work Under the Bernoulli reward model in the context of cognitive radio, MAB with multiple users was considered in [3] and [4]. In [3], a heuristic policy based on histogram estimation of the unknown parameters was proposed. This policy provides a linear order of the system regret rate, thus cannot achieve the maximum average reward. In [4], Anandkumar et al. have independently established...
two order-optimal distributed policies by extending the order-optimal index-type single-user policies proposed in [5]. In the first policy, each of the M users has a pre-assigned rank of the M best channels to target at. In the second policy, the users eventually settle down in orthogonal channels without any pre-agreement [4]. Eliminating pre-agreement is an advantage over the proposed TDFS policy, where each player has a pre-assigned offset in the time division sharing of the M best arms. The policy proposed here, however, applies to more general reward models (for example, Gaussian and Poisson reward distributions that have infinite support). Furthermore, the policies proposed in [4] are specific to the single-user policies proposed in [5], whereas the TDFS policy can be used with any order-optimal single-player policy to achieve order optimality in the distributed setting. Another difference between the policies proposed in [4] and the TDFS policy is on user fairness. The two policies in [4] orthogonalize users into different channels that offer different throughput, whereas the TDFS policy ensures that each player achieves the same time-average reward at the same rate.

Recently, a variation of centralized MAB in the context of cognitive radio has been considered in [6] where a channel offers independent Bernoulli rewards with different (unknown) means for different users. This more general model captures contention among users experiencing different communication environments. A centralized policy that assumes full information exchange and cooperation among users is proposed which achieves the logarithmic order of the regret growth rate. We point out that the logarithmic order of the TDFS policy is preserved in the decentralized setting when we allow users to experience different reward distributions on the same arm provided that users have the common set of the M best channels and each of the M best channels has the same mean across users.

**Notation** For two sets A and B, let A\(\cap\)B denote the set consisting of all elements in A that do not belong to B. For two positive integers k and l, define \(k \otimes l := (k - 1) \mod l \) + 1, which is an integer taking values from 1, 2, · · · , l.

### 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 one arm (1 \(\leq \) M \(\leq\) 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 \(\theta_i\). The function \(f(\cdot; \cdot)\) is known and the parameter set \(\Theta = (\theta_1, \cdots, \theta_N)\) is unknown to the player. For the ease 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 \(\pi\) is measured by the system regret \(R_T(\Theta)\) over a time horizon of length T as defined below. Let \(\sigma\) be a permutation of \(\{1, \cdots, N\}\) such that \(\theta_{\sigma(1)} \geq \theta_{\sigma(2)} \geq \cdots \geq \theta_{\sigma(N)}\), we have

\[
R_T(\pi(\Theta)) = T \sum_{j=1}^{M} \theta_{\sigma(j)} - E_{\pi}[\sum_{t=1}^{T} Y_{\pi(t)}(t)],
\]

where \(Y_{\pi(t)}(t)\) is the random reward obtained in slot t under policy \(\pi\), and \(E_{\pi}[\cdot]\) denotes the expectation under the policy \(\pi\). In other words, \(R_T(\Theta)\) is the expected total reward loss up to time T under policy \(\pi\) compared to the perfect scenario that \(\Theta\) 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 \(R_T(\pi)\) grows with T under any parameter set \(\Theta\) by choosing the optimal policy \(\pi^*\).

A policy is called uniformly good if for any parameter set \(\Theta\), we have \(R_T(\pi(\Theta)) = o(T^a)\) 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 \(\Theta\).

We present in the theorem below the result established in [1,2] on the logarithmic order as well as the leading constant of the minimum regret growth rate of the single-player MAB.

**Theorem 1.2.** For any uniformly good policy \(\pi\),

\[
\liminf_{T \to \infty} \frac{R_T(\pi(\Theta))}{\log T} \geq \sum_{j=1}^{M} \theta_{\pi(M)} - \theta_j
\]

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 \(\theta_i\) and \(\theta_j\), respectively.

Assuming single play (\(M = 1\)), Lai and Robbins [1] have constructed an optimal policy that achieves the lower bound on the regret growth rate given in (2).

**Lai and Robbins’ Policy for Single-User MAB with Single-Play [1]**

In the first N slots, play each arm once. Fix \(\delta > 1/N\). For all \(t > N\), let \(\theta_a(t)\) denote the sample mean of the reward obtained from arm \(a\) and \(\tau_a(t)\) the time arm \(a\) 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. Define the round-robin candidate \(r_t = \min \{a : \theta_a(t) \geq \theta_{l_t}(t)\}\). For all \(t > N\), let \(\theta_a(t) > \theta_{l_t}(t)\) and \(I(\theta_a(t), \theta_{l_t}(t)) > \log(t - 1)/\tau_{l_t}(t)\); otherwise the player plays the round-robin candidate \(r_t\).

### III. PROBLEM FORMULATION

Consider the spectrum consisting of \(N\) independent but nonidentical channels (arms). Let \(S(t) = \{S_1(t), \cdots, S_N(t)\} \in \{0, 1\}^N\) \((t \geq 1)\) denote the system state, where \(S_i(t)\) is the state of channel \(i\) in slot \(t\). Assume that \(\{S_i(t)\}_{t=1}^{\infty}\) is an i.i.d. Bernoulli process with unknown mean \(\theta_i\). 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 \(a_i(t) \in \{1, \cdots, N\}\) 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 \(\pi_i\) for user \(i\) as a sequence of functions \(\pi_i = \{\pi_i(t)\}_{t=1}^{\infty}\), where \(\pi_i(t)\) maps user i’s past observations and decisions to \(a_i(t)\) in slot \(t\). The decentralized policy \(\pi\) is thus given by the concatenation of the local policy for each user: \(\pi = \{\pi_1, \cdots, \pi_M\}\). 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 below.

**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 primary user does not experience any other secondary users have claimed the channel. Under this model, we have

\[
Y(t) = \sum_{j=1}^{N} I_j(t) S_j(t),
\]

where \(I_j(t)\) 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 valid
successful if and only if no other users have chosen the same channel. Under this model, we have
\[ Y(t) = \sum_{j=1}^{M} Y_j(t) S_j(t), \]
where \( Y_j(t) \) 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 \( \pi \).
\[ R_T^F(\pi) = T \sum_{j=1}^{M} \theta_{\sigma(j)} - E_{\pi^*}[\sum_{t=1}^{T} Y(t)], \]
The objective is to minimize the rate at which \( R_T(\pi) \) grows with time \( T \) under any parameter set \( \Theta \) by choosing the optimal decentralized policy \( \pi^* \). Similarly, we call a decentralized policy is uniformly good if for any parameter set \( \Theta \), we have \( R_T(\pi) = o(T^p) \) 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 \( \pi^* \), we have
\[ L(\Theta) \leq \liminf_{T \to \infty} \frac{R_T^F(\pi^*)}{\log T} \leq \limsup_{T \to \infty} \frac{R_T^F(\pi^*)}{\log T} \leq U(\Theta) \] (3)
for some constants \( L(\Theta) \) and \( U(\Theta) \) that depend on \( \Theta \).

**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 [7].

V. AN ORDER-OPTIMAL DECENTRALIZED POLICY

In this section, we construct a decentralized policy \( \pi_T^F \) that achieves the optimal logarithmic order of the system regret growth rate under the fairness constraint.

The basic structure of the proposed policy \( \pi_T^F \) is a time division structure at each user for selecting the \( M \) best channels. Users have different phases (offsets) in their time division selection schedule to avoid excessive collisions. Consider, for example, the case of \( M = 2 \). The time sequence is divided into two disjoint subsequences, where the first subsequence consists of all odd slots and the second one consists of all even slots. User 1 targets at the best channel during the first subsequence and the second best channel during the second subsequence, and user 2 does the opposite. Without loss of generality, consider user 1. In the first subsequence, user 1 applies Lai and Robbins’s single-user policy to efficiently learn and select the best channel. In the second subsequence, the second best channel is learned and identified by applying Lai and Robbins’s policy to the remaining \( N - 1 \) channels after removing the channel considered as the best in the first subsequence. Specifically, as illustrated in Fig. 1, the second subsequence is further divided into \( N \) disjoint mini-sequences, where the \( i \)-th mini-sequence consists of all slots that follow a slot in which channel \( i \) was sensed (i.e., channel \( i \) is considered as the best in the preceding slot that belongs to the first subsequence). In the \( i \)-th mini-sequence, user 1 applies Lai and Robbins’s policy to channels \( \{1, \ldots, i-1, i+1, \ldots, N\} \). In other words, the local policy for each user consists of \( N+1 \) parallel Lai and Robbins’s procedures: one is applied in the subsequence for targeting at the best channel and the rest \( N \) are applied in the \( N \) disjoint mini-sequences for targeting at the second best channel. These parallel procedures, however, are coupled through the common observation history since at each time, regardless of which subsequence or mini-sequence it belongs to, all the past observations are used in the decision making. A detailed implementation of the proposed policy for a general \( M \) is given in Fig. 2.

![Fig. 1. The structure of user 1’s local policy under \( \pi_T^F \) (\( M = 2, N = 3 \), and \( \alpha(1) \) denotes the channel considered as the best by user 1 in the first subsequence.).](image)

We point out that \( \pi_T^F \) 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 predetermined (for example, based on the user’s ID). The policy can thus be implemented in a distributed fashion.

While the basic building block of the proposed TDFS policy \( \pi_T^F \) is Lai and Robbins’s single-user policy that achieves the optimal logarithmic order in the single-user setting, it is highly nontrivial to establish the logarithmic regret growth rate of \( \pi_T^F \) in the decentralized setting with multiple players. Compared to the single-user-counterpart given in [1], [2], the difficulties in establishing the logarithmic regret growth rate of \( \pi_T^F \) 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 channel. Second, since players are learning the channel statistics based on their individual local observations, they do not always agree on the rank of the channels. Thus, collisions are bound to happen (for example, when an channel is considered as the best by one player but the second best by another). Such issues need to be carefully dealt with in establishing the logarithmic order of the regret growth rate.

**Theorem 2:** Under the decentralized policy \( \pi_T^F \), we have
\[ \limsup_{T \to \infty} \frac{R_T^F(\pi_T^F)}{\log T} \leq C(\Theta), \] (4)
where under collision model 1,
\[ C(\Theta) = M(\sum_{k=1}^{N} x_k \theta_{\sigma(k)} - \sum_{j=1}^{\sigma(M)} \theta_{\sigma(M)} (1/I(\theta_j, \theta_{\sigma(M)}))), \]
\[ x_k = \sum_{i=1}^{N} \theta_{\sigma(i)} 1/I(\theta_j, \theta_{\sigma(i)}), \]
and under collision model 2,
\[ C(\Theta) = M(\sum_{i=1}^{N} \sum_{k=1}^{M} x_k \theta_{\sigma(i)} - \sum_{j=1}^{\sigma(M)} \theta_{\sigma(M)} (1/I(\theta_j, \theta_{\sigma(M)})) - \sum_{i=1}^{M} 1/I(\theta_j, \theta_{\sigma(i)})), 0). \]

**Proof:** See [7].

From Theorem 1 and Theorem 2, the decentralized policy is order-optimal. Furthermore, the decentralized policy ensures fairness among players as given in Theorem 3 below.

**Theorem 3:** Define the local regret for user \( i \) as
\[ R_T^F(\pi_T^F) = \frac{1}{M} T \sum_{j=1}^{M} \theta_{\sigma(j)} - E_{\pi_T^F}[\sum_{i=1}^{N} Y_i(t)], \]
where \( Y_i(t) \) is the immediate reward obtained by user \( i \) in slot \( t \), we have
The Decentralized TDFS Policy $\pi^*_F$

Without loss of generality, consider player $i$.

- Notations and Inputs: let $\tilde{\theta}_n(t)$ denote the sample mean of the state of channel $n$ and $T_{\pi,n}$ the times channel $n$ is sensed up to (but excluding) slot $t$. The time is divided into $M$ subsequences. Let $m_k(t)$ denote the number of slots in the $k$th subsequence up to (and including) $t$.

At time $t$, user $i$ does the following.

1. If $t$ belongs to the $i$th subsequence (i.e., $t \gg M = i$), user $i$ targets at the best channel by carrying out the following procedure.
   - If $m_i(t) \leq N$, sense channel $m_i(t)$. Otherwise, the user chooses between a leader $l_i$ and a round-robin candidate $r_i = m_i(t) \subseteq N$, where the leader $l_i$ is the channel with the largest sample mean among all channels that have been sensed for at least $(m_i(t) - 1)\delta$ times. The user senses the leader $l_i$ if $\tilde{\theta}_{l_i}(t) > \tilde{\theta}_{r_i}(t)$ and $I(\tilde{\theta}_{l_i}(t), \tilde{\theta}_{l_i}(t)) > \log(t - 1)/\tau_{r_i}$; otherwise the user senses the round-robin candidate $r_i$.

2. If $t$ belongs to the $k$th ($k \neq i$) subsequence (i.e., $t \gg M = k$), the user targets at the $j$th best channel where $j = (k - i + M + 1) \subseteq M$ by carrying out the following procedure.
   - If $m_k(t) \leq N$, play channel $m_k(t)$. Otherwise, the users target at the $j$th best channel. Let $\mathcal{A}_t$ denote the set of $j - 1$ channels played in the previous $j - 1$ slots. $t$ thus belongs to the mini-sequence associated with the subset $\mathcal{N}_t = \{1, \ldots, N\} \setminus \mathcal{A}_t$ of channels. The user chooses between a leader and a round-robin candidate defined within $\mathcal{N}_t$. Specifically, let $m_{\mathcal{N}_t}(t)$ denote the number of slots in this mini-sequence up to (and including) $t$. Among all channels that have been played for at least $(m_{\mathcal{N}_t}(t) - 1)\delta$ times, let $l_i$ denote the leader with the largest sample mean. If $r_i = m_{\mathcal{N}_t}(t) \subseteq (N - j + 1)$ be the round-robin candidate where, for simplicity, we have assumed that channels in $\mathcal{N}_t$ are indexed by $1, \ldots, N - j + 1$. The user senses the leader $l_i$ if $\tilde{\theta}_{l_i}(t) > \tilde{\theta}_{r_i}(t)$ and $I(\tilde{\theta}_{l_i}(t), \tilde{\theta}_{l_i}(t)) > \log(t - 1)/\tau_{r_i}$; otherwise the user senses the round-robin candidate $r_i$.

Fig. 2. The decentralized TDFS policy $\pi^*_F$.

$$\limsup_{T \to \infty} \frac{R^T_F(\Theta)}{\log T} = \frac{1}{M} \limsup_{T \to \infty} \frac{R^T_F(\Theta)}{\log T}. \tag{5}$$

Theorem 3 follows directly from the symmetry among players under $\pi^*_F$. It shows that each player achieves the same time average reward $\frac{1}{T} \sum_{j=1}^{M} \mu(\theta_{\sigma(1)})$ at the same rate.

Note that the proposed policy can be used with any single-user policy, which can also be different for different users. More importantly, the order optimality of the TDFS policy is preserved as long as each user’s single-user policy achieves the optimal logarithmic order in the single-user setting. A more detailed discussion can be found in [7].

VI. A LOWER BOUND FOR A CLASS OF DECENTRALIZED POLICIES

In this section, we establish a lower bound on the growth rate of the system regret for a general class of decentralized policies, to which the proposed policy $\pi^*_F$ 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 $\pi = [\pi_1, \ldots, \pi_M]$ that satisfy the following property: under any local policy $\pi_i$, 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 $\Theta$ 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 $a_{ij}T - o(T^2)$ for all $b > 0$.

A policy in the TDS class essentially allows a user to efficiently select each of the $M$ best channels according to a fixed time portion that does not depend on the parameter set $\Theta$. It is easy to see that the TDFS $\pi^*_F$ belongs to the TDS class with $a_{ij} = 1/M$ for all $1 \leq i, j \leq M$.

**Theorem 4:** For any uniformly good decentralized policy $\pi$ in the TDS class, we have

$$\liminf_{T \to \infty} \frac{R^T_F(\Theta)}{\log T} \geq \sum_{j=1}^{M} a_{ij} \log \frac{\theta_{\sigma(M)} - \theta_{\sigma(1)}}{T(\theta_j, \theta_{\sigma(1)})}. \tag{6}$$

**Proof:** See [7].

VII. CONCLUSION AND DISCUSSION

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. A lower bound on the leading constant of the logarithmic order is established for policies in the TDS class, to which the proposed policy belongs. While we have focused on a Bernoulli reward model, all results hold under a general reward model as detailed in [7]. With its general reward model, the proposed TDFS policy finds a wide area of potential applications including multi-channel communication systems, multi-agent systems, web search and advertising, and social networks [7].

**References**


