Decentralized Online Learning Algorithms for Opportunistic Spectrum Access

Yi Gai and Bhaskar Krishnamachari
Ming Hsieh Department of Electrical Engineering, University of Southern California, CA 90089, USA
Email: {ygai, bkrishna}@usc.edu

***Abstract*—The fundamental problem of multiple secondary users contending for opportunistic spectrum access over multiple channels in cognitive radio networks has been formulated recently as a decentralized multi-armed bandit (D-MAB) problem. In a D-MAB problem there are M users and N arms (channels) that each offer i.i.d. stochastic rewards with unknown means so long as they are accessed without collision. The goal is to design a decentralized online learning policy that incurs minimal regret, defined as the difference between the total expected rewards accumulated by a model-aware genie, and that obtained by all users applying the policy. We make two contributions in this paper. First, we consider the setting where the users have a prioritized ranking, such that it is desired for the K-th-ranked user to learn to access the arm offering the K-th highest mean reward. For this problem, we present the first distributed policy that yields regret that is uniformly logarithmic over time without requiring any prior assumption about the mean rewards. Second, we consider the case when a fair access policy is required, i.e., it is desired for all users to experience the same mean reward. For this problem, we present a distributed policy that yields order-optimal regret scaling with respect to the number of users and arms, better than previously proposed policies in the literature. Both of our distributed policies make use of an innovative modification of the well known UCB1 policy for the classic multi-armed bandit problem that allows a single user to learn how to play the arm that yields the K-th largest mean reward.**

I. Introduction

Developing dynamic spectrum access mechanisms to enable more efficient spectrum utilization is one of the most challenging issues in cognitive radio systems [1]. In this paper, we focus on a problem of opportunistic spectrum access in cognitive radio networks, where at every time slot, each of the M decentralized secondary users searches for idle channels which are not occupied by primary users temporarily among $N \geq M$ channels. We assume that the throughput of these N channels evolves i.i.d. over time with any arbitrary, bounded-support distribution, which is unknown to the users. These distributed players can only learn from their local observations and collide (with reward penalty) when choosing the same arm. The desired objective is to develop a sequential policy running at each user to make a selection among multiple choices, where there is no information exchange, such that the sum-throughput of all distributed users is maximized, assuming an interference model whereby at most one secondary user can derive benefit from any channel.

Multi-Armed Bandit problem (MAB, see [2]–[6]) is a fundamental mathematical framework for learning the unknown variables. In its simplest form of classic non-Bayesian version studied by Lai and Robbins [2], there are N arms, each providing stochastic rewards that are independent and identically distributed over time, with unknown means. A policy is desired to pick one arm at each time sequentially, to maximize the reward. Anantharam *et al.* [3] extend this work to the case when M simultaneous plays are allowed, with centralized scheduling of the players.

A fundamental tradeoff between exploration and exploitation is captured by MAB problems: on the one hand, various arms should be explored often enough in order to learn their parameters, and on the other hand, the prior observations should be exploited to gain the best possible immediate rewards. A key metric in evaluating a given policy for this problem is *regret*, which is defined as the difference between the expected reward gained by a *prior* that always makes the optimal choice and that obtain by the given policy. The regret achieved by a policy can be evaluated in terms of its growth over time. Many of the prior works on multi-armed bandits show logarithmic scaling of the regret over time.

While most of the prior work on MAB focused on the centralized policies, motivated by the problem of opportunistic access in cognitive radio networks, Liu and Zhao [7], [8], and Anandkumar *et al.* [9], [10] have both developed policies for the problem of M distributed players operating N independent arms. There are two problem formulations of interest when considering distributed MAB: a) the *prioritized access problem*, where it is desired to prioritize a ranked set of users so that the K-th ranked user learns to access the arm with the K-th highest reward, and b) the *fair access problem*, where the goal is to ensure that each user receives the same reward in expectation. For the prioritized access problem, Anandkumar *et al.* [9] present a distributed policy that yields regret that is logarithmic in time, but requires prior knowledge of the arm reward means. For the fair access problem, they propose in [9], [10] a randomized distributed policy that is logarithmic with respect to time and scales as $O(M^2N)$ with respect to the number of arms and users. Liu and Zhao [7], [8] also treat the fair access problem and present the TDFS policy which yields asymptotically logarithmic regret with respect to time and scales as $O(M(\max\{M^2, (N-M)M\}))$ with respect to the number of arms and users.

In this paper we make significant new contributions to both problem formulations. For the prioritized access problem, we present a distributed learning policy DLP that results in a regret that is uniformly logarithmic in time and, unlike the prior work in [9], does not require any prior knowledge about the arm reward means. For the fair access problem, we present another distributed learning policy DLF, which yields regret that is also uniformly logarithmic in time and that scales as $O(M(N-M))$ with respect to the number of users M and the number of arms N. As it has been shown in [8] that the lower-bound of regret for distributed policies also scales as $O(M(N-M))$, this is not only a better scaling than the previous state of the art, it is, in fact, order-optimal.

A key subroutine of both decentralized learning policies running at each user involves selecting an arm with the desired rank order of the mean reward. For this, we present a new policy that we refer to as SL(K), which is a non-trivial generalization of UCB1 in [5]. SL(K) provides a general solution for selecting an arm with the K-th largest expected rewards for classic MAB problems with N arms.

This paper is organized as follows. We present in section II the problem formulation. In section III, we first present our SL(K) policy, which is a general policy to play an arm with K-th largest expected reward for classic multi-armed bandits, and then present our decentralized DLP policy in section IV and DLF policy in section V based on SL(K) policy. Both policies are polynomial-storage polynomial-time-per-step learning policies. We show that the regrets of all policies we proposed are logarithmic in time and polynomial in the number of users and channels, and we compare the upper bound of the regrets of different policies. In section VI, we compare the decentralized learning policies with simulation results. Finally, section VII concludes the paper.

II. Problem Formulation

We consider a cognitive system with N channels (arms) and M decentralized secondary users (players). The throughput of N channels are defined by random processes $X_i(n)$, $1 \le i \le N$. Time is slotted and denoted by the index n. We assume that $X_i(n)$ evolves as an i.i.d. random process over time, with the only restriction that its distribution have a finite support. Without loss of generality, we normalize $X_i(n) \in [0,1]$. We do not require that $X_i(n)$ be independent across i. This random process is assumed to have a mean $\theta_i = E[X_i]$, that is unknown to the users and distinct from each other. We denote the set of all these means as $\Theta = \{\theta_i, 1 \le i \le N\}$.

At each decision period n (also referred to interchangeably as time slot), each of the M decentralized users selects an arm only based on its own observation histories under a decentralized policy. When a particular arm i is selected by user j, the value of $X_i(n)$ is only observed by user j, and if there is no other user playing the same arm, a reward of $X_i(n)$ is obtained. Else, if there are multiple users playing the same arm, then we assume that, due to collision, at most one of the conflicting users j' gets reward $X_i(n)$, while the other users get zero reward. This interference assumption covers practical models in networking research, such as the perfect collision model (in which none of the conflicting users derive any benefit) and CSMA with perfect sensing (in which exactly one of the conflicting user derives benefit from the channel). We denote the first model as $\mathbf{M}_1$ and the second model as $\mathbf{M}_2$.

We denote the decentralized policy for user j at time n as $\pi_j(n)$, and the set of policies for all users as $\pi = \{\pi_j(n), 1 \le j \le M\}$. We are interested in designing decentralized policies, under which there is no information exchange among users, and analyze them with respect to *regret*, which is defined as the gap between the expected reward that could be obtained by a genie-aided perfect selection and that obtained by the policy. We denote $\mathcal{O}^*_M$ as a set of M arms with M largest expected rewards. The regret can be expressed as:

$$\mathfrak{R}^\pi(\Theta;n) = n\sum_{i\in\mathcal{O}^*_M}\theta_i - E^\pi[\sum_{t=1}^{n} S_{\pi(t)}(t)] \tag{1}$$

where $S_{\pi(t)}(t)$ is the sum of the actual reward obtained by all users at time t under policy $\pi(t)$, which could be expressed as:

$$S_{\pi(t)}(t) = \sum_{i=1}^{N}\sum_{j=1}^{M} X_i(t)\mathbb{I}_{i,j}(t), \tag{2}$$

where for $\mathbf{M}_1$, $\mathbb{I}_{i,j}(t)$ is defined to be 1 if user j is the only user to play arm i, and 0 otherwise; for $\mathbf{M}_2$, $\mathbb{I}_{i,j}(t)$ is defined to be 1 if user j is the one with the smallest index among all users playing arm i at time t, and 0 otherwise.

Besides getting low total regret, there could be other system objectives for a given D-MAB. We consider two in this paper. In the prioritized access problem, we assume that each user has information of a distinct allocation order. Without loss of generality, we assume that the users are ranked in such a way that the m-th user seeks to access the arm with the m-th highest mean reward. In the fair access problem, users are treated equally to receive the same expected reward.

III. Selective Learning of the K-th Largest Expected Reward

We first propose a general policy to play an arm with the K-th largest expected reward ($1 \le K \le N$) for classic multi-armed bandit problem with N arms and one user, since the key idea of our proposed decentralized policies running at each user in section IV and V is that user m will run a learning policy targeting an arm with m-th largest expected reward.

Our proposed policy of learning an arm with K-th largest expected reward is shown in Algorithm 1.

We use two 1 by N vectors to store the information after we play an arm at each time slot. One is $(\hat{\theta}_i)_{1\times N}$ in which $\hat{\theta}_i$ is the average (sample mean) of all the observed values of X_i up to the current time slot (obtained through potentially different sets of arms over time). The other one is $(n_i)_{1\times N}$ in which n_i is the number of times that X_i has been observed up to the current time slot.

Note that while we indicate the time index in Algorithm 1 for notational clarity, it is not necessary to store the matrices from previous time steps while running the algorithm. So SL(K) policy requires storage linear in N.

Remark: SL(K) policy generalizes UCB1 in [5] and presents a general way to pick an arm with the K-th largest

Algorithm 1 Selective learning of the K-th largest expected rewards (SL(K))

```
// INITIALIZATION
for $t = 1$ to $N$ do
    Let $i = t$ and play arm $i$;
    $\hat{\theta}_i(t) = X_i(t)$;
    $n_i(t) = 1$;
end for
// MAIN LOOP
while 1 do
    $t = t + 1$;
    Let the set $\mathcal{O}_K$ contains the $K$ arms with the $K$ largest values in (3)
        $$\hat{\theta}_i(t-1) + \sqrt{\frac{2\ln t}{n_i(t-1)}};  (3)$$
    Play arm $k$ in $\mathcal{O}_K$ such that
        $$k = \arg\min_{i\in\mathcal{O}_K} \hat{\theta}_i(t-1) - \sqrt{\frac{2\ln t}{n_i(t-1)}};  (4)$$
    $\hat{\theta}_k(t) = \frac{\hat{\theta}_k(t-1)n_k(t-1)+X_k(t)}{n_k(t-1)+1}$;
    $n_k(t) = n_k(t-1) + 1$;
end while
```

expected rewards for a classic multi-armed bandit problem with N arms (without the requirement of distinct expected rewards for different arms).

Now we present the analysis of the upper bound of regret, and show that it is linear in N and logarithmic in time. We denote $\mathcal{A}_K$ as the set of arms with K-th largest expected reward. Note that Algorithm 1 is a general algorithm for picking an arm with the K-th largest expected reward for the classic multi-armed bandit problems, where we allow multiple arms with K-th largest expected reward, and all these arms retreated as optimal arms. The following theorem holds for Algorithm 1.

Theorem 1: Under the policy specified in Algorithm 1, the expected number of times that we pick any arm $i \notin \mathcal{A}_K$ after n time slots is at most:

$$\frac{8\ln n}{\Delta_{K,i}} + 1 + \frac{2\pi^2}{3}. \tag{5}$$

where $\Delta_{K,i} = |\theta_K - \theta_i|$, θ_K is the K-th largest expected reward.

Proof: Below is a sketch of the proof. A detailed proof can be found in [12].

Denote $T_i(n)$ as the number of times that we pick arm $i \notin \mathcal{A}_K$ at time n. Denote C_{t,n_i} as $\sqrt{\frac{(L+1)\ln t}{n_i}}$. Denote $\hat{\bar{\theta}}_{i,n_i}$ as the average (sample mean) of all the observed values of X_i when it is observed n_i time. $\mathcal{O}^*_K$ is denoted as the set of K arms with K largest expected rewards.

Denote by $I_i(n)$ the indicator function which is equal to 1 if $T_i(n)$ is added by one at time n. Let l be an arbitrary positive integer. Then, for any arm i which is not a desired arm, i.e., $i \notin \mathcal{A}_K$:

$$\begin{aligned} T_i(n) &= 1 + \sum_{t=N+1}^{n} \mathbb{1}\{I_i(t)\} \\ &\le l + \sum_{t=N+1}^{n} (\mathbb{1}\{I_i(t), \theta_i < \theta_K, T_i(t-1) \ge l\} \\ &\qquad + \mathbb{1}\{I_i(t), \theta_i > \theta_K, T_i(t-1) \ge l\}) \end{aligned} \tag{6}$$

where $\mathbb{1}(x)$ is the indicator function defined to be 1 when the predicate x is true, and 0 when it is false.

Note that for the case $\theta_i < \theta_K$, arm i is picked at time t means that there exists an arm $j(t) \in \mathcal{O}^*_K$, such that $j(t) \notin \mathcal{O}_K$. This means the following inequality holds:

$$\hat{\bar{\theta}}_{j(t),T_{j(t)}(t-1)} + C_{t-1,T_{j(t)}(t-1)} \le \hat{\bar{\theta}}_{i,T_i(t-1)+C_{t-1,T_i(t-1)}}. \tag{7}$$

Then, we have

$$\begin{aligned} &\sum_{t=N+1}^{n} \mathbb{1}\{I_i(t), \theta_i < \theta_K, T_i(t-1) \ge l\} \\ &\le \sum_{t=N+1}^{n} \mathbb{1}\{\hat{\bar{\theta}}_{j(t),T_{j(t)}(t-1)} + C_{t-1,T_{j(t)}(t-1)} \\ &\qquad \le \hat{\bar{\theta}}_{i,T_i(t-1)} + C_{t-1,T_i(t-1)}, T_i(t-1) \ge l\} \\ &\le \sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=l}^{t-1} \mathbb{1}\{\hat{\bar{\theta}}_{j(t),n_{j(t)}} + C_{t,n_{j(t)}} \le \hat{\bar{\theta}}_{i,n_i} + C_{t,n_i}\} \end{aligned} \tag{8}$$

$\hat{\bar{\theta}}_{j(t),n_{j(t)}} + C_{t,n_{j(t)}} \le \hat{\bar{\theta}}_{i,n_i} + C_{t,n_i}$ implies that at least one of the following must be true:

$$\hat{\bar{\theta}}_{j(t),n_{j(t)}} \le \theta_{j(t)} - C_{t,n_{j(t)}}, \tag{9}$$

$$\hat{\bar{\theta}}_{i,n_i} \ge \theta_i + C_{t,n_i}, \tag{10}$$

$$\theta_{j(t)} < \theta_i + 2C_{t,n_i}. \tag{11}$$

Applying the Chernoff-Hoeffding bound [11], we could find the upper bound of (9) and (10) as,

$$Pr\{\hat{\bar{\theta}}_{j(t),n_{j(t)}} \le \theta_{j(t)} - C_{t,n_{j(t)}}\} \le e^{-4\ln t} = t^{-4}, \tag{12}$$

$$Pr\{\hat{\bar{\theta}}_{i,n_i} \ge \theta_i + C_{t,n_i}\} \le e^{-4\ln t} = t^{-4} \tag{13}$$

For $l \ge \left\lceil \frac{8\ln n}{\Delta^2_{K,i}} \right\rceil$, $\theta_{j(t)} - \theta_i - 2C_{t,n_i} \ge \theta_K - \theta_i - 2\sqrt{\frac{2\Delta^2_{K,i}\ln t}{8\ln n}} \ge \theta_K - \theta_i - \Delta_{K,i} = 0$, so (11) is false when $l \ge \left\lceil \frac{8\ln n}{\Delta^2_{K,i}} \right\rceil$.

Note that for the case $\theta_i > \theta_K$, when arm i is picked at time t, there are two possibilities: either $\mathcal{O}_K = \mathcal{O}^*_K$, or $\mathcal{O}_K \neq \mathcal{O}^*_K$. If $\mathcal{O}_K = \mathcal{O}^*_K$, the following inequality holds:

$$\hat{\bar{\theta}}_{i,T_i(t-1)} - C_{t-1,T_i(t-1)} \le \hat{\bar{\theta}}_{K,T_K(t-1)} - C_{t-1,T_K(t-1)}.$$

If $\mathcal{O}_K \neq \mathcal{O}^*_K$, $\mathcal{O}_K$ has at least one arm $h(t) \notin \mathcal{O}^*_K$. Then, we have:

$$\hat{\bar{\theta}}_{i,T_i(t-1)} - C_{t-1,T_i(t-1)} \le \hat{\bar{\theta}}_{h(t),T_{h(t)}(t-1)} - C_{t-1,T_{h(t)}(t-1)}.$$

So to conclude both possibilities for the case $\theta_i > \theta_K$, if we denote $\mathcal{O}^*_{K-1} = \mathcal{O}^*_K - \mathcal{A}_K$, at each time t when arm i is picked, these exists an arm $h(t) \notin \mathcal{O}^*_{K-1}$, such that

$$\hat{\bar{\theta}}_{i,T_i(t-1)} - C_{t-1,T_i(t-1)} \leq \hat{\bar{\theta}}_{h(t),T_{h(t)}(t-1)} - C_{t-1,T_{h(t)}(t-1)}. \tag{14}$$

We could have the similar analysis for (14) as for $\hat{\bar{\theta}}_{j(t),n_{j(t)}} + C_{t,n_{j(t)}} \leq \hat{\bar{\theta}}_{i,n_i} + C_{t,n_i}$.

Hence, we have

$$\begin{aligned}
\mathbb{E}[T_i(n)] \leq & \left\lceil \frac{8\ln n}{\Delta^2_{K,i}} \right\rceil + \sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=\lceil (8\ln n)/\Delta^2_{K,i} \rceil}^{t-1} \\
& (Pr\{\hat{\bar{\theta}}_{j(t),n_{j(t)}} \leq \theta_{j(t)} - C_{t,n_{j(t)}}\} + Pr\{\hat{\bar{\theta}}_{i,n_i} \geq \theta_i + C_{t,n_i}\}) \\
& + \sum_{t=1}^{\infty} \sum_{n_i=\lceil (8\ln n)/\Delta^2_{K,i} \rceil}^{t-1} \sum_{n_{h(t)}=1}^{t-1} \\
& (Pr\{\hat{\bar{\theta}}_{i,n_i} \leq \theta_i - C_{t,n_i}\} + Pr\{\hat{\bar{\theta}}_{h(t),n_{h(t)}} \geq \theta_{h(t)} + C_{t,n_{h(t)}}\}) \\
& \leq \frac{8\ln n}{\Delta^2_{K,i}} + 1 + 2\sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=1}^{t-1} 2t^{-4} \\
& \leq \frac{8\ln n}{\Delta^2_{K,i}} + 1 + \frac{2\pi^2}{3}.
\end{aligned} \tag{15}$$

■

The definition of *regret* for the above problem is different from the traditional multi-armed bandit problem with the goal of maximization or minimization, since our goal now is to pick the arm with the K-th largest expected reward and we wish we could minimize the number of times that we pick the wrong arm. Here we give two definitions of the regret to evaluate the SL(K) policy.

Definition 1: We define the *regret of type 1 at each time slot* as the absolute difference between the expected reward that could be obtained by a genie that can pick an arm with K-th largest expected reward, and that obtained by the given policy at each time slot. Then the *total regret of type 1 by time* n is defined as sum of the regret at each time slot, which is:

$$\mathfrak{R}^\pi_1(\Theta; n) = \sum_{t=1}^{n} |\theta_K - E^\pi[S_{\pi(t)}(t)]| \tag{16}$$

Definition 2: We define the *total regret of type 2 by time* n as the absolute difference between the expected reward that could be obtained by a genie that can pick an arm with K-th largest expected reward, and that obtained by the given policy after n plays, which is:

$$\mathfrak{R}^\pi_2(\Theta; n) = |n\theta_K - E^\pi[\sum_{t=1}^{n} S_{\pi(t)}(t)]| \tag{17}$$

Here we note that $\forall n$, $\mathfrak{R}^\pi_2(\Theta; n) \leq \mathfrak{R}^\pi_1(\Theta; n)$ because $|n\theta_K - E^\pi[\sum_{t=1}^{n} S_{\pi(t)}(t)]| = |n\theta_K - \sum_{t=1}^{n} E^\pi[S_{\pi(t)}(t)]| \leq \sum_{t=1}^{n} |\theta_K - E^\pi[S_{\pi(t)}(t)]|$.

Corollary 1: The expected regret under both definitions is at most

$$\sum_{i:i\notin\mathcal{A}_k} (\frac{8\ln n}{\Delta_{K,i}}) + (1 + \frac{2\pi^2}{3}) \sum_{i:i\notin\mathcal{A}_k} \Delta_{K,i}. \tag{18}$$

Proof: See [12]. ■

Corollary 1 shows the upper bound of the regret of SL(K) policy. It grows logarithmical in time and linearly in the number of arms.

IV. Distributed Learning with Prioritization

We now consider the distributed multi-armed bandit problem with prioritized access. Our proposed decentralized policy for N arms with M users is shown in Algorithm 2.

Algorithm 2 Distributed Learning Algorithm with Prioritization for N Arms with M Users Running at User m (DLP)

```
1:  // Initialization
2:  for $t = 1$ to $N$ do
3:      Play arm $k$ such that $k = ((m + t) \mod N) + 1$;
4:      $\hat{\theta}^m_k(t) = X_k(t)$;
5:      $n^m_k(t) = 1$;
6:  end for
7:  // Main loop
8:  while 1 do
9:      $t = t + 1$;
10:     Play an arm $k$ according to policy SL($m$) specified in Algorithm 1;
11:     $\hat{\theta}^m_k(t) = \frac{\hat{\theta}^m_k(t-1)n^m_k(t-1)+X_k(t)}{n^m_k(t-1)+1}$;
12:     $n^m_k(t) = n^m_k(t-1) + 1$;
13: end while
```

In the above algorithm, line 2 to 6 is the initialization part, for which user m will play each arm once to have the initial value in $(\hat{\theta}^m_i)_{1\times N}$ and $(n^m_i)_{1\times N}$. Line 3 ensures that there will be no collisions among users. Similar as in Algorithm 1, we indicate the time index for notational clarity. Only two 1 by N vectors, $(\hat{\theta}^m_i)_{1\times N}$ and $(n^m_i)_{1\times N}$, are used by user m to store the information after we play an arm at each time slot.

We denote o^*_m as the index of arm with the m-th largest expected reward. Note that $\{o^*_m\}_{1\leq m\leq M} = \mathcal{O}^*_M$. Denote $\Delta_{i,j} = |\theta_i - \theta_j|$ for arm i, j. We now state the main theorem of this section.

***Theorem** 2:* The expected regret under the DLP policy specified in Algorithm 2 is at most

$$\begin{aligned}
& \sum_{m=1}^{M} \sum_{i\neq o^*_m} (\frac{8\ln n}{\Delta^2_{o^*_m,i}} + 1 + \frac{2\pi^2}{3})\theta_{o^*_m} \\
& + \sum_{m=1}^{M} \sum_{h\neq m} (\frac{8\ln n}{\Delta^2_{o^*_h,o^*_m}} + 1 + \frac{2\pi^2}{3})\theta_{o^*_m}.
\end{aligned} \tag{19}$$

Proof: Denote $T_{i,m}(n)$ the number of times that user m pick arm i at time n.

For each user m, the regret under DLP policy can arise due to two possibilities: (1) user m plays an arm $i \neq o^*_m$; (2) other user $h \neq m$ plays arm o_m^*. In both cases, collisions may happen, resulting a loss which is at most $\theta_{o_m^*}$. Considering these two possibilities, the regret of user m is upper bounded by: $\mathfrak{R}^\pi(\Theta, m; n) \leq \sum_{i \neq o_m^*} \mathbb{E}[T_{i,m}(n)]\theta_{o_m^*} + \sum_{h \neq m} \mathbb{E}[T_{o_m^*,h}(n)]\theta_{o_m^*}$.

From Theorem 1, $T_{i,m}(n)$ and $T_{o_m^*,h}(n)$ are bounded by $\mathbb{E}[T_{i,m}(n)] \leq \frac{8 \ln n}{\Delta^2_{o_m^*,i}} + 1 + \frac{2\pi^2}{3}$, $\mathbb{E}[T_{o_m^*,h}(n)] \leq \frac{8 \ln n}{\Delta^2_{o_h^*,o_m^*}} + 1 + \frac{2\pi^2}{3}$. So, $\mathfrak{R}^\pi(\Theta, m; n) \leq \sum_{i \neq o_m^*} (\frac{8 \ln n}{\Delta^2_{o_m^*,i}} + 1 + \frac{2\pi^2}{3})\theta_{o_m^*} + \sum_{h \neq m} (\frac{8 \ln n}{\Delta^2_{o_h^*,o_m^*}} + 1 + \frac{2\pi^2}{3})\theta_{o_m^*}$. We note that $\mathfrak{R}^\pi(\Theta; n) = \sum_{m=1}^{M} \mathfrak{R}^\pi(\Theta, m; n)$, so Theorem 2 stands. ■

Theorem 2 shows that the regret of our DLP algorithm is uniformly upper-bounded for all time n by a function that grows as $O(M(N + M - 2) \ln n)$.

V. Distributed Learning with Fairness

For the purpose of fairness consideration, secondary users should be treated equally, and there should be no prioritization for the users. In this scenario, a naive algorithm is to apply Algorithm 2 directly by rotating the prioritization as shown in Figure 1. Each user maintains two 1 by N vectors $(\hat{\theta}^m_{i,j})_{M \times N}$ and $(n^m_{i,j})_{M \times N}$, where the i-th row stores only the observation values for the i-th prioritization vectors. We denote this naive algorithm as DLF-Naive. Since the storage of DLF-Naive grows linear in MN, instead of N, it does not utilize the observations under different allocation order, which will result a worse regret as shown in the analysis of this section. To utilize all the observations, we propose our distributed learning algorithm with fairness (DLF) in Algorithm 3.

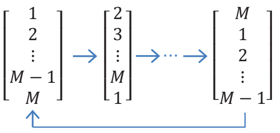


Fig. 1. Illustration of rotating the prioritization vector.

Same as in Algorithm 2, only two 1 by N vectors, $(\hat{\theta}^m_i)_{1 \times N}$ and $(n^m_i)_{1 \times N}$, are used by user m to store the information after we play an arm at each time slot.

Line 11 in Algorithm 3 means user m play the arm with the K-th largest expected reward with Algorithm 1, where the value of K is calculated in line 10 to ensure the desired arm to pick for each user is different, and the users play arms from the estimated largest to the estimated smallest in turns to ensure the fairness.

Theorem 3: The expected regret under the DLF-Naive policy is at most

$$\sum_{o_m^* \in \mathcal{O}_m^*} \sum_{m=1}^{M} \sum_{i \neq o_m^*} (\frac{8 \ln \lceil n/M \rceil}{\Delta^2_{o_m^*,i}} + 1 + \frac{2\pi^2}{3})\theta_{o_m^*} + \sum_{o_m^* \in \mathcal{O}_m^*} \sum_{m=1}^{M} \sum_{h \neq m} (\frac{8 \ln \lceil n/M \rceil}{\Delta^2_{o_h^*,o_m^*}} + 1 + \frac{2\pi^2}{3})\theta_{o_m^*}. \tag{20}$$

Algorithm 3 Distributed Learning Algorithm with Fairness for N Arms with M Users Running at User m (DLF)

```
1:  // Initialization
2:  for $t = 1$ to $N$ do
3:      Play arm $k$ such that $k = ((m + t) \mod N) + 1$;
4:      $\hat{\theta}^m_k(t) = X_k(t)$;
5:      $n^m_k(t) = 1$;
6:  end for
7:  // Main loop
8:  while 1 do
9:      $t = t + 1$;
10:     $K = ((m + t) \mod M) + 1$;
11:     Play an arm $k$ according to policy SL($K$) specified in Algorithm 1;
12:     $\hat{\theta}^m_k(t) = \frac{\hat{\theta}^m_k(t-1) n^m_k(t-1) + X_k(t)}{n^m_k(t-1) + 1}$;
13:     $n^m_k(t) = n^m_k(t - 1) + 1$;
14: end while
```

Proof: Theorem 3 is a direct conclusion from Theorem 2 by replacing n with $\lceil n/M \rceil$, and then take the sum over all M best arms which are played in the algorithm. ■

The above theorem shows that the regret of the DLF-Naive policy grows as $O(M^2(N + M - 2) \ln n)$.

Theorem 4: The expected regret under the DLF policy specified in Algorithm 3 is at most

$$M \sum_{i=1}^{N} (\frac{8 \ln n}{\Delta^2_{\min,i}} + 1 + \frac{2\pi^2}{3})\theta_{\max} + M(M - 1) \sum_{i \in \mathcal{O}_M^*} (\frac{8 \ln n}{\Delta^2_{\min,i}} + 1 + \frac{2\pi^2}{3})\theta_i, \tag{21}$$

where $\Delta_{\min,i} = \min_{1 \leq m \leq M} \Delta_{o_m^*,i}$.

Proof: See [12]. ■

Theorem 5: When time n is large enough such that

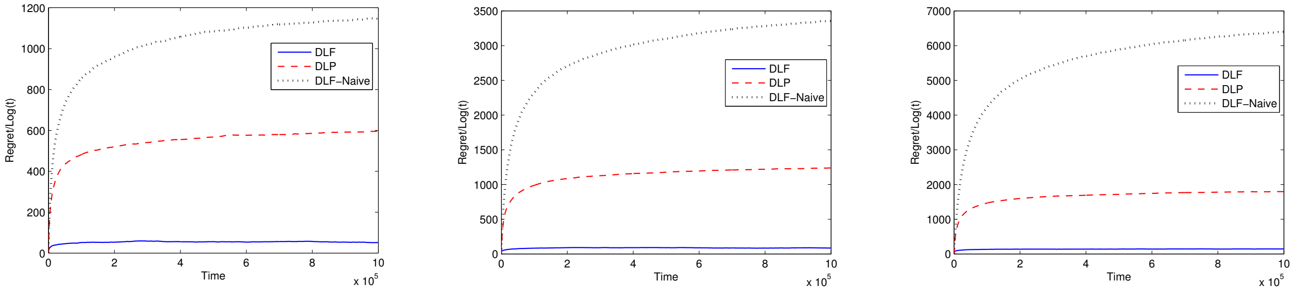
$$\frac{n}{\ln n} \geq \frac{8(N + M)}{\Delta^2_{\min}} + (1 + \frac{2\pi^2}{3})N + M, \tag{22}$$

the expected regret under the DLF policy specified in Algorithm 3 is at most

$$M \sum_{i \notin \mathcal{O}_M^*} (\frac{8 \ln n}{\Delta^2_{\min,i}} + 1 + \frac{2\pi^2}{3})\theta_{\max} + M^2(1 + \frac{2\pi^2}{3})\theta_{\max} + M(M - 1)(1 + \frac{2\pi^2}{3}) \sum_{i \in \mathcal{O}_M^*} \theta_i. \tag{23}$$

Proof: See [12]. ■

Comparing Theorem 3 with Theorem 4 and Theorem 5, if we define $C = \frac{8(N+M)}{\Delta^2_{\min}} + (1 + \frac{2\pi^2}{3})N + M$, we can see that the regret of the naive policy DLF-Naive grows as $O(M(NM + M^2 - 2M) \ln n)$, while the regret of the DLF policy grows as $O(M(N + M^2 - 1) \ln n)$ when $\frac{n}{\ln n} < C$, $O(M(N - M) \ln n)$ when $\frac{n}{\ln n} \geq C$.



(a) $N = 4$ channels, $M = 2$ secondary users, $\Theta = (0.9, 0.8, 0.7, 0.6)$.

(b) $N = 5$ channels, $M = 3$ secondary users, $\Theta = (0.9, 0.8, 0.7, 0.6, 0.5)$.

(c) $N = 7$ channels, $M = 4$ secondary users, $\Theta = (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3)$.

Fig. 2. Normalized regret $\frac{\mathfrak{R}(n)}{\ln n}$ vs. n time slots.

VI. Numerical Results

We present simulation results for the algorithms developed in this work, varying the number of users and channels to verify the performance of our proposed algorithms detailed earlier. In the simulations, we assume channels are in either idle state (with throughput 1) or busy state (with throughput 0). The state of each N channel evolves as an i.i.d. Bernoulli process across time slots, with the parameter set Θ unknown to the M users.

Figure 2 shows the simulation results averaged over 50 runs using the three algorithms, DLP, DLF-Naive, and DLF, and the regrets are compared. As expected, DLF has the least regret, since one of the key features of DLF is that it does not favor any one user over another. The chance for each user to use any one of the M best channels are the same. It utilizes its observations on all the M best channels, and thus makes less mistakes for exploring. DLF-Naive not only has the greatest regret, also uses more storage. DLP has greater regret than DLF since user m has to spend time on exploring the $M - 1$ channels in the M best channels expect channel $k \neq o^*_m$. Not only this results in a loss of reward, this also results in the collisions among users.

Figure 2 also explores the impact of increasing the number of channels N, and secondary users M on the regret experienced by the different policies with the minimum distance between arms $\Delta_{\min}$ fixed. It is clearly that as the number of channels and secondary users increases, the regret, as well as the regret gap between different algorithms increases.

VII. Conclusion

The problem of distributed multi-armed bandits is a fundamental extension of the classic online learning framework that finds application in the context of opportunistic spectrum access for cognitive radio networks. We have made two key algorithmic contributions to this problem. For the case of prioritized users, we presented the first distributed policy that yields logarithmic regret over time without prior assumptions about the mean arm rewards. For the case of fair access, we presented a policy that yields order-optimal regret scaling in terms of the numbers of users and arms, which is also an improvement over prior results.

Through simulations, we further show that the overall regret is lower for the fair access policy. In future work, we plan to undertake more comprehensive simulation based comparison of the proposed policy with previously proposed schemes, including over more realistic channel models. We are also interested in considering extensions of our distributed policies to multi-armed bandits with dependent arms, such as the combinatorial model considered in [6].

References

[1] T. Yucek and H. Arslan, “A Survey of Spectrum Sensing Algorithms for Cognitive Radio Applications”, *IEEE Communications Surveys & Tutorials*, vol. 11, pp. 116-130, 2009.

[2] T. Lai and H. Robbins, “Asymptotically Efficient Adaptive Allocation Rules”, *Advances in Applied Mathematics*, vol. 6, pp. 4-22, 1985.

[3] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays - Part I: I.I.D. Rewards”, *IEEE Transactions on Automatic Control*, vol. 32, pp. 968-976, 1987.

[4] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays - Part II: Markovian Rewards”, *IEEE Transactions on Automatic Control*, vol. 32, pp. 977-982, 1987.

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

[6] Y. Gai, B. Krishnamachari, and R. Jain, “Learning Multiuser Channel Allocations in Cognitive Radio Networks: A Combinatorial Multi-armed Bandit Formulation”, *IEEE International Dynamic Spectrum Access Networks (DySPAN) Symposium*, April, 2010.

[7] K. Liu and Q. Zhao, “Distributed Learning in Multi-Armed Bandit With Multiple Players”, *IEEE Transactions on Signal Processing*, vol. 58, pp. 5667 - 5681, November, 2010.

[8] K. Liu and Q. Zhao, “Distributed Learning in Cognitive Radio Networks: Multi-Armed Bandit with Distributed Multiple Players”, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March, 2010.

[9] A. Anandkumar, N. Michael, and A. K. Tang, “Opportunistic Spectrum Access with Multiple Users: Learning under Competition”, *International Conference on Computer Communications (INFOCOM)*, March, 2010.

[10] A. Anandkumar, N. Michael, A. Tang, and A. Swami, “Distributed Learning and Allocation of Cognitive Users with Logarithmic Regret”, to appear in *IEEE JSAC on Advances in Cognitive Radio Networking and Communications*.

[11] D. Pollard, *Convergence of Stochastic Processes*. Berlin: Springer, 1984.

[12] Y. Gai and B. Krishnamachari, “Decentralized Online Learning Algorithms for Opportunistic Spectrum Access”, Technical Report, March, 2011. Available at http://anrg.usc.edu/www/publications/papers/DMAB2011.pdf.