Distributed Stochastic Online Learning Policies for Opportunistic Spectrum Access

Yi Gai, *Member, IEEE,* and Bhaskar Krishnamachari, *Member, IEEE, ACM*

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 distributed online learning policies that incur 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 consider two related problem formulations 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 DLP, 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 DLF, 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.

Index Terms—Online Learning, Dynamic Spectrum Access, Decentralized Multi-armed Bandit.

I. INTRODUCTION

Developing dynamic spectrum access mechanisms to enable more efficient spectrum utilization is one of the most challenging issues in cognitive radio systems [2]. 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 as an i.i.d. stochastic process 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.

The Multi-Armed Bandit problem (MAB, see [3]–[7]) is a fundamental mathematical framework for learning the unknown variables. In its simplest form of classic non-Bayesian version studied by Lai and Robbins [3], 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.* [4] 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 [8], [9], and Anandkumar *et al.* [10], [11] 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.* [10] 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 [10], [11] a randomized distributed policy that is logarithmic with respect to time and scales as $O(M^2 N)$ with respect to the number of arms and users. Liu and Zhao [8], [9] 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 [10], 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 [9] that the lower-bound of regret for distributed policies also scales as $\Omega(M(N-M))$, 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 [6]. $SL(K)$ provides a general solution for selecting an arm with the K-th largest expected rewards for classic MAB problems with N arms.

We focus in this work on the decentralized learning with i.i.d stochastic model, which has been studied essentially the same in the work [8]–[11]. To tackle other challenges in dynamic spectrum access, there has been other researchers [12]–[17] working within an MAB framework have considered other problem formulation for dynamic spectrum access: [12] considers the model of restless Markov process, where a Restless MAB is formulated; Liu and Zhao [13] also considered the restless Markovian formulation and established the indexability; the work by Lelarge *et al.* [14] set up an adversarial linear bandit formulation; a cooperative scheme was proposed in [15] to address the problem of distributed spectrum selection in presence of imperfect sensing from the secondary users; Tehrani *et al.* [16], [17] focus on unslotted primary systems. We believe that our contribution in this work will be a useful stepping stone for further theoretical research that integrates advances made along the various dimensions of this complicated problem.

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 $X_i(n)$, $1 \leq i \leq$ 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 $\Theta = \{\theta_i, 1 \leq i \leq N\}$.

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 $\pi = \{\pi_j(n), 1 \leq j \leq M\}$. 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:

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

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:

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

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. Then, if we denote $V_{i,j}^\pi(n) = \sum_{t=1}^{n} \mathbb{I}_{i,j}(t)$, we have:

$$E^\pi[\sum_{t=1}^{n} S_{\pi(t)}(t)] = \sum_{i=1}^{N} \sum_{j=1}^{M} \theta_i E[V_{i,j}^\pi(n)] \qquad (3)$$

The analysis in this paper works the same for both models. 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 ($1 \leq K \leq N$) 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.

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

1: // INITIALIZATION
2: **for** $t = 1$ to N **do**
3:     Let $i = t$ and play arm i;
4:     $\hat{\theta}_i(t) = X_i(t)$;
5:     $n_i(t) = 1$;
6: **end for**
7: // MAIN LOOP
8: **while** 1 **do**
9:     $t = t + 1$;
10:     Let the set $\mathcal{O}_K$ contains the K arms with the K largest values in (4)

$$\hat{\theta}_i(t-1) + \sqrt{\frac{2 \ln t}{n_i(t-1)}}; \qquad (4)$$

11:     Play arm k in $\mathcal{O}_K$ such that

$$k = \arg\min_{i \in \mathcal{O}_K} \hat{\theta}_i(t-1) - \sqrt{\frac{2 \ln t}{n_i(t-1)}}; \qquad (5)$$

12:     $\hat{\theta}_k(t) = \frac{\hat{\theta}_k(t-1)n_k(t-1) + X_k(t)}{n_k(t-1)+1}$;
13:     $n_k(t) = n_k(t-1) + 1$;
14: **end while**

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: While UCB1 can only pick the largest arm, SL(K) policy generalizes UCB1 in [6] and presents a general way to pick an arm with the K-th largest expected rewards for a classic multi-armed bandit problem with N arms (without the requirement of distinct expected rewards for different arms). Our analysis are build upon [6] and significantly extends those in [6].

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:

$$\frac{8 \ln n}{\Delta_{K,i}} + 1 + \frac{2\pi^2}{3}. \qquad (6)$$

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

Proof:

Denote $T_i(n)$ as the number of times that we pick arm $i \notin \mathcal{A}_K$ at time n. We now show the upper bound of $\mathbb{E}[T_i(n)]$.

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., $i \notin \mathcal{A}_K$:

$$\begin{aligned}
T_i(n) &= 1 + \sum_{t=N+1}^{n} \mathbb{1}\{I_i(t)\} \\
&\leq l + \sum_{t=N+1}^{n} \mathbb{1}\{I_i(t), T_i(t-1) \geq l\} \\
&\leq l + \sum_{t=N+1}^{n} (\mathbb{1}\{I_i(t), \theta_i < \theta_K, T_i(t-1) \geq l\} \\
&\qquad + \mathbb{1}\{I_i(t), \theta_i > \theta_K, T_i(t-1) \geq l\})
\end{aligned} \qquad (7)$$

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:

$$\hat{\bar{\theta}}_{j(t),T_{j(t)}(t-1)} + C_{t-1,T_{j(t)}(t-1)} \leq \hat{\bar{\theta}}_{i,T_i(t-1)+C_{t-1,T_i(t-1)}}. \qquad (8)$$

Then, we have

$$\sum_{t=N+1}^{n} \mathbb{1}\{I_i(t), \theta_i < \theta_K, T_i(t-1) \geq l\}$$

$$\leq \sum_{t=N+1}^{n} \mathbb{1}\{\hat{\bar{\theta}}_{j(t),T_{j(t)}(t-1)} + C_{t-1,T_{j(t)}(t-1)}$$

$$\leq \hat{\bar{\theta}}_{i,T_i(t-1)} + C_{t-1,T_i(t-1)}, T_i(t-1) \geq l\}$$

$$\leq \sum_{t=N+1}^{n} \mathbb{1}\{\min_{0<n_{j(t)}<t} \hat{\bar{\theta}}_{j(t),n_{j(t)}} + C_{t-1,n_{j(t)}}$$

$$\leq \max_{l \leq n_i < t} \hat{\bar{\theta}}_{i,n_i} + C_{t-1,n_i}\}$$

$$\leq \sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=l}^{t-1} \mathbb{1}\{\hat{\bar{\theta}}_{j(t),n_{j(t)}} + C_{t,n_{j(t)}} \leq \hat{\bar{\theta}}_{i,n_i} + C_{t,n_i}\}$$

$$(9)$$

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

$$\hat{\bar{\theta}}_{j(t),n_{j(t)}} \leq \theta_{j(t)} - C_{t,n_{j(t)}}, \tag{10}$$

$$\hat{\bar{\theta}}_{i,n_i} \geq \theta_i + C_{t,n_i}, \tag{11}$$

$$\theta_{j(t)} < \theta_i + 2C_{t,n_i}. \tag{12}$$

Applying the Chernoff-Hoeffding bound [18], we could find the upper bound of (10) and (11) as,

$$Pr\{\hat{\bar{\theta}}_{j(t),n_{j(t)}} \leq \theta_{j(t)} - C_{t,n_{j(t)}}\} \leq e^{-4\ln t} = t^{-4}, \tag{13}$$

$$Pr\{\hat{\bar{\theta}}_{i,n_i} \geq \theta_i + C_{t,n_i}\} \leq e^{-4\ln t} = t^{-4} \tag{14}$$

For $l \geq \left\lceil \frac{8\ln n}{\Delta_{K,i}^2} \right\rceil$,

$$\theta_{j(t)} - \theta_i - 2C_{t,n_i}$$
$$\geq \theta_K - \theta_i - 2\sqrt{\frac{2\Delta_{K,i}^2 \ln t}{8\ln n}} \tag{15}$$
$$\geq \theta_K - \theta_i - \Delta_{K,i} = 0,$$

so (12) is false when $l \geq \left\lceil \frac{8\ln n}{\Delta_{K,i}^2} \right\rceil$.

Note that for the case $\theta_i > \theta_K$, when arm i is picked at time t, there are two possibilities: (1) $\mathcal{O}_K = \mathcal{O}_K^*$; (2) $\mathcal{O}_K \neq \mathcal{O}_K^*$. If $\mathcal{O}_K = \mathcal{O}_K^*$, the following inequality holds:

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

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:

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

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

$$\hat{\bar{\theta}}_{i,T_i(t-1)} - C_{t-1,T_i(t-1)} \leq \hat{\bar{\theta}}_{h(t),T_{h(t)}(t-1)} - C_{t-1,T_{h(t)}(t-1)}. \tag{16}$$

Then similarly, we can have:

$$\sum_{t=N+1}^{n} \mathbb{1}\{I_i(t), \theta_i > \theta_K, T_i(t-1) \geq l\}$$

$$\leq \sum_{t=1}^{\infty} \sum_{n_i=l}^{t-1} \sum_{n_{h(t)}=1}^{t-1} \mathbb{1}\{\hat{\bar{\theta}}_{i,n_i} - C_{t,n_i} \leq \hat{\bar{\theta}}_{h(t),n_{h(t)}} - C_{t,n_{h(t)}}\}$$

$$(17)$$

Note that $\hat{\bar{\theta}}_{i,n_i} - C_{t,n_i} \leq \hat{\bar{\theta}}_{h(t),n_{h(t)}} - C_{t,n_{h(t)}}$ implies one of the following must be true:

$$\hat{\bar{\theta}}_{i,n_i} \leq \theta_i - C_{t,n_i}, \tag{18}$$

$$\hat{\bar{\theta}}_{h(t),n_{h(t)}} \geq \theta_{h(t)} + C_{t,n_{h(t)}}, \tag{19}$$

$$\theta_i < \theta_{h(t)} + 2C_{t,n_i}. \tag{20}$$

We again apply the Chernoff-Hoeffding bound and get $Pr\{\hat{\bar{\theta}}_{i,n_i} \leq \theta_i - C_{t,n_i}\} \leq t^{-4}$, $Pr\{\hat{\bar{\theta}}_{h(t),n_{h(t)}} \geq \theta_{h(t)} + C_{t,n_{h(t)}}\} \leq t^{-4}$.

Also note that for $l \geq \left\lceil \frac{8\ln n}{\Delta_{K,i}^2} \right\rceil$,

$$\theta_i - \theta_{h(t)} - 2C_{t,n_i}$$
$$\geq \theta_i - \theta_K - \Delta_{K,i} \geq 0, \tag{21}$$

so (20) is false.

Hence, we have

$$\mathbb{E}[T_i(n)] \leq \left\lceil \frac{8\ln n}{\Delta_{K,i}^2} \right\rceil + \sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=\left\lceil (8\ln n)/\Delta_{K,i}^2 \right\rceil}^{t-1}$$

$$(Pr\{\hat{\bar{\theta}}_{j(t),n_{j(t)}} \leq \theta_{j(t)} - C_{t,n_{j(t)}}\} + Pr\{\hat{\bar{\theta}}_{i,n_i} \geq \theta_i + C_{t,n_i}\})$$

$$+ \sum_{t=1}^{\infty} \sum_{n_i=\left\lceil (8\ln n)/\Delta_{K,i}^2 \right\rceil}^{t-1} \sum_{n_{h(t)}=1}^{t-1}$$

$$(Pr\{\hat{\bar{\theta}}_{i,n_i} \leq \theta_i - C_{t,n_i}\} + Pr\{\hat{\bar{\theta}}_{h(t),n_{h(t)}} \geq \theta_{h(t)} + C_{t,n_{h(t)}}\})$$

$$\leq \frac{8\ln n}{\Delta_{K,i}^2} + 1 + 2\sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=1}^{t-1} 2t^{-4}$$

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

$$(22)$$

∎

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:

$$\mathfrak{R}_1^\pi(\Theta;n) = \sum_{t=1}^{n} |\theta_K - E^\pi[S_{\pi(t)}(t)]| \qquad (23)$$

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:

$$\mathfrak{R}_2^\pi(\Theta;n) = |n\theta_K - E^\pi[\sum_{t=1}^{n} S_{\pi(t)}(t)]| \qquad (24)$$

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

$$\sum_{i:i\notin\mathcal{A}_k} (\frac{8\ln n}{\Delta_{K,i}}) + (1 + \frac{2\pi^2}{3}) \sum_{i:i\notin\mathcal{A}_k} \Delta_{K,i}. \qquad (25)$$

Proof: Under the SL(K) policy, we have:

$$\begin{aligned}
\mathfrak{R}_2^\pi(\Theta;n) &\leq \mathfrak{R}_1^\pi(\Theta;n) \\
&= \sum_{t=1}^{n} |\theta_K - E^\pi[S_{\pi(t)}(t)]| \\
&= \sum_{i:i\notin\mathcal{A}_k} \Delta_{K,i} \mathbb{E}[T_i(n)] \\
&\leq \sum_{i:i\notin\mathcal{A}_k} (\frac{8\ln n}{\Delta_{K,i}}) + (1 + \frac{2\pi^2}{3}) \sum_{i:i\notin\mathcal{A}_k} \Delta_{K,i}.
\end{aligned} \qquad (26)$$

Here we note that $\forall n$, $\mathfrak{R}_2^\pi(\Theta;n) \leq \mathfrak{R}_1^\pi(\Theta;n)$ 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 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.

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 $(\hat{\theta}_i^m)_{1\times N}$ and $(n_i^m)_{1\times N}$. 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, $(\hat{\theta}_i^m)_{1\times N}$ and $(n_i^m)_{1\times N}$, 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.

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:     $\hat{\theta}_k^m(t) = X_k(t)$;
5:     $n_k^m(t) = 1$;
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:     $\hat{\theta}_k^m(t) = \frac{\hat{\theta}_k^m(t-1)n_k^m(t-1)+X_k(t)}{n_k^m(t-1)+1}$;
12:     $n_k^m(t) = n_k^m(t - 1) + 1$;
13: **end while**

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

$$\begin{aligned}
&\sum_{m=1}^{M} \sum_{i\neq o_m^*} (\frac{8\ln n}{\Delta_{o_m^*,i}^2} + 1 + \frac{2\pi^2}{3})\theta_{o_m^*} \\
&+ \sum_{m=1}^{M} \sum_{h\neq m} (\frac{8\ln n}{\Delta_{o_h^*,o_m^*}^2} + 1 + \frac{2\pi^2}{3})\theta_{o_m^*}.
\end{aligned} \qquad (27)$$

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 $\theta_{o_m^*}$. So for both cases, the regret of user m is upper bounded by:

$$\mathfrak{R}^\pi(\Theta,m;n) \leq \sum_{i\neq o_m^*} \mathbb{E}[T_{i,m}(n)]\theta_{o_m^*} + \sum_{h\neq m} \mathbb{E}[T_{o_m^*,h}(n)]\theta_{o_m^*} \qquad (28)$$

From Theorem 1, $T_{i,m}(n)$ and $T_{o_m^*,h}(n)$ are bounded by

$$\mathbb{E}[T_{i,m}(n)] \leq \frac{8\ln n}{\Delta_{o_m^*,i}^2} + 1 + \frac{2\pi^2}{3}, \qquad (29)$$

$$\mathbb{E}[T_{o_m^*,h}(n)] \leq \frac{8\ln n}{\Delta_{o_h^*,o_m^*}^2} + 1 + \frac{2\pi^2}{3}. \qquad (30)$$

So,

$$\begin{aligned}
\mathfrak{R}^\pi(\Theta,m;n) \leq &\sum_{i\neq o_m^*} (\frac{8\ln n}{\Delta_{o_m^*,i}^2} + 1 + \frac{2\pi^2}{3})\theta_{o_m^*} \\
&+ \sum_{h\neq m} (\frac{8\ln n}{\Delta_{o_h^*,o_m^*}^2} + 1 + \frac{2\pi^2}{3})\theta_{o_m^*}
\end{aligned} \qquad (31)$$

The upper bound for regret is:

$$\mathfrak{R}^\pi(\Theta; n) = \sum_{m=1}^{M} \mathfrak{R}^\pi(\Theta, m; n)$$

$$\leq \sum_{m=1}^{M} \sum_{i \neq o_m^*} \left(\frac{8 \ln n}{\Delta_{o_m^*, i}^2} + 1 + \frac{2\pi^2}{3} \right) \theta_{o_m^*} \qquad (32)$$

$$+ \sum_{m=1}^{M} \sum_{h \neq m} \left(\frac{8 \ln n}{\Delta_{o_h^*, o_m^*}^2} + 1 + \frac{2\pi^2}{3} \right) \theta_{o_m^*}$$

For a concise expression on the upper bound of the regret, if we define $\Delta_{\min} = \min\limits_{1 \leq i \leq N, 1 \leq j \leq M} \Delta_{i,j}$, and $\theta_{\max} = \max\limits_{1 \leq i \leq N} \theta_i$, we could get a more concise (but looser) upper bound as:

$$\mathfrak{R}^\pi(\Theta; n) \leq M(N + M - 2)\left(\frac{8 \ln n}{\Delta_{\min}^2} + 1 + \frac{2\pi^2}{3} \right) \theta_{\max}. \quad (33)$$

$\blacksquare$

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

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 $(\hat{\theta}_{j,i}^m)_{M \times N}$ and $(n_{j,i}^m)_{M \times N}$, where the j-th row stores only the observation values for the j-th prioritization vectors. This naive algorithm is shown in Algorithm 3.
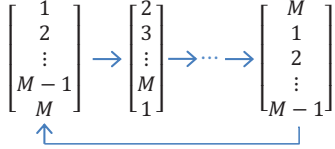
$$\begin{bmatrix} 1 \\ 2 \\ \vdots \\ M-1 \\ M \end{bmatrix} \rightarrow \begin{bmatrix} 2 \\ 3 \\ \vdots \\ M \\ 1 \end{bmatrix} \rightarrow \cdots \rightarrow \begin{bmatrix} M \\ 1 \\ 2 \\ \vdots \\ M-1 \end{bmatrix}$$

Fig. 1. Illustration of rotating the prioritization vector.

Algorithm 3 A Naive Algorithm for Distributed Learning Algorithm with Fairness (DLF-Naive) Running at User m

---

1: At time t, run Algorithm 2 with prioritization $K = ((m + t) \mod M) + 1$, then update the K-th row of $(\hat{\theta}_{j,i}^m)_{M \times N}$ and $(n_{j,i}^m)_{M \times N}$ accordingly.

We can see that the storage of Algorithm 3 grows linear in MN, instead of N. And 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 4.

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

Line 11 in Algorithm 4 means user m play the arm with the K-th largest expected reward with Algorithm 1, where the

---

Algorithm 4 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:     $\hat{\theta}_k^m(t) = X_k(t)$;
5:     $n_k^m(t) = 1$;
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:     $\hat{\theta}_k^m(t) = \frac{\hat{\theta}_k^m(t-1) n_k^m(t-1) + X_k(t)}{n_k^m(t-1) + 1}$;
13:     $n_k^m(t) = n_k^m(t - 1) + 1$;
14: **end while**

---

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 specified in Algorithm 3 is at most

$$\sum_{o_m^* \in \mathcal{O}_m^*} \sum_{m=1}^{M} \sum_{i \neq o_m^*} \left(\frac{8 \ln \lceil n/M \rceil}{\Delta_{o_m^*, i}^2} + 1 + \frac{2\pi^2}{3} \right) \theta_{o_m^*}$$

$$+ \sum_{o_m^* \in \mathcal{O}_m^*} \sum_{m=1}^{M} \sum_{h \neq m} \left(\frac{8 \ln \lceil n/M \rceil}{\Delta_{o_h^*, o_m^*}^2} + 1 + \frac{2\pi^2}{3} \right) \theta_{o_m^*}. \tag{34}$$

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. $\blacksquare$

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

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

$$M \sum_{i=1}^{N} \left(\frac{8 \ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3} \right) \theta_{\max}$$

$$+ M(M - 1) \sum_{i \in \mathcal{O}_M^*} \left(\frac{8 \ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3} \right) \theta_i, \tag{35}$$

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

Proof:

Denote $K_m^*(t)$ as the index of the arm with the K-th (got by line 10 at time t in Algorithm 4 running at user m) largest expected reward. Denote $Q_i^m(n)$ as the number of times that user m pick arm $i \neq K_m^*(t)$ for $1 \leq t \leq n$.

The proof here is similar to the first part of proof of Theorem 1: we notice that for any arbitrary positive integer l and any time t, $Q_i^m(t) \geq l$ implies $n_i(t) \geq l$. So (7) to (21) in the proof of Theorem 1 still hold by replacing $T_i(n)$ with $Q_i^m(n)$ and replacing K with $K_m^*(t)$. Note that since all the channels are with different rewards, there is only one element in the set $\mathcal{A}_{K_m^*(t)}$.

To find the upper bound of $\mathbb{E}[Q_i^m(n)]$, we should let l to be $l \geq \left\lceil \frac{8 \ln n}{\Delta_{\min,i}^2} \right\rceil$ such that (12) and (20) are false for all t. So we have,

$$
\mathbb{E}[Q_i^m(n)] \leq \left\lceil \frac{8 \ln n}{\Delta_{\min,i}^2} \right\rceil + \sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=\left\lceil (8\ln n)/\Delta_{K_m^*(t),i}^2 \right\rceil}^{t-1}
$$

$$
(Pr\{\hat{\bar{\theta}}_{j(t),n_{j(t)}} \leq \theta_{j(t)} - C_{t,n_{j(t)}}\} + Pr\{\hat{\bar{\theta}}_{i,n_i} \geq \theta_i + C_{t,n_i}\})
$$

$$
+ \sum_{t=1}^{\infty} \sum_{n_i=\left\lceil (8\ln n)/\Delta_{K_m^*(t),i}^2 \right\rceil}^{t-1} \sum_{n_{h(t)}=1}^{t-1}
$$

$$
(Pr\{\hat{\bar{\theta}}_{i,n_i} \leq \theta_i - C_{t,n_i}\} + Pr\{\hat{\bar{\theta}}_{h(t),n_{h(t)}} \geq \theta_{h(t)} + C_{t,n_{h(t)}}\})
$$

$$
\leq \frac{8 \ln n}{\Delta_{\min,i}^2} + 1 + 2 \sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=1}^{t-1} 2t^{-4}
$$

$$
\leq \frac{8 \ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3}.
$$
(36)

Hence for user m, we could calculate the upper bound of regret considering the two possibilities as in the proof of Theorem 2 as:

$$
\mathfrak{R}^{\pi}(\Theta, m; n) \leq \sum_{i=1}^{N} Q_i^m(n)\theta_{\max} + \sum_{h \neq m} \sum_{i \in \mathcal{O}_M^*} Q_h^m(n)\theta_i \quad (37)
$$

So the upper bound for regret for m users is:

$$
\mathfrak{R}^{\pi}(\Theta; n) = \sum_{m=1}^{M} \mathfrak{R}^{\pi}(\Theta, m; n)
$$

$$
\leq M \sum_{i=1}^{N} \left(\frac{8 \ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3}\right)\theta_{\max}
$$
(38)

$$
+ M(M-1) \sum_{i \in \mathcal{O}_M^*} \left(\frac{8 \ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3}\right)\theta_i
$$

To be more concise, we could also write the above upper bound as:

$$
\mathfrak{R}^{\pi}(\Theta; n) \leq M(N + M(M-1))\left(\frac{8 \ln n}{\Delta_{\min}} + 1 + \frac{2\pi^2}{3}\right)\theta_{\max}.
$$
(39)

∎

Theorem 5: When time n is large enough such that

$$
\frac{n}{\ln n} \geq \frac{8(N+M)}{\Delta_{\min}^2} + \left(1 + \frac{2\pi^2}{3}\right)N + M, \quad (40)
$$

the expected regret under the DLF policy specified in Algorithm 4 is at most

$$
M \sum_{i \notin \mathcal{O}_M^*} \left(\frac{8 \ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3}\right)\theta_{\max} + M^2\left(1 + \frac{2\pi^2}{3}\right)\theta_{\max}
$$

$$
+ M(M-1)\left(1 + \frac{2\pi^2}{3}\right) \sum_{i \in \mathcal{O}_M^*} \theta_i
$$
(41)

Proof: The inequality (36) implies that the total number of times that the desired arms are picked by user m at time n

is lower bounded by $n - \sum_{i=1}^{N}\left(\frac{8\ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3}\right)$. Since all the arms with M largest expected rewards are picked in turn by the algorithm, $\forall i \in \mathcal{O}_M^*$, we have

$$
n_i(n) \geq \frac{1}{M}\left(n - \sum_{i=1}^{N}\left(\frac{8\ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3}\right)\right). \quad (42)
$$

where $n_i(n)$ refers to the number of times that arm i has been observed up to time n at user m. (For the purpose of simplicity, we omit m in the notation of n_i.)

Note that when n is big enough such that $\frac{n}{\ln n} \geq \frac{8(N+M)}{\Delta_{\min}^2} + \left(1 + \frac{2\pi^2}{3}\right)N + M$, we have,

$$
n_i(n) \geq \frac{1}{M}\left(n - \sum_{i=1}^{N}\left(\frac{8\ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3}\right)\right) \geq \left\lceil\frac{8\ln n}{\Delta_{\min}^2}\right\rceil.
$$
(43)

When (43) holds, both (12) and (20) are false. Then $\forall i \in \mathcal{O}_M^*$, when n is large enough to satisfy (43),

$$
\mathbb{E}[Q_i^m(n)] = \sum_{t=N+1}^{n} \mathbb{1}\{I_i(t)\}
$$

$$
= \sum_{t=N+1}^{n} \left(\mathbb{1}\{I_i(t), \theta_i < \theta_K\} + \mathbb{1}\{I_i(t), \theta_i > \theta_K\}\right)
$$

$$
\leq \sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=\left\lceil (8\ln n)/\Delta_{\min}^2 \right\rceil}^{t-1}
$$

$$
(Pr\{\hat{\bar{\theta}}_{j(t),n_{j(t)}} \leq \theta_{j(t)} - C_{t,n_{j(t)}}\} + Pr\{\hat{\bar{\theta}}_{i,n_i} \geq \theta_i + C_{t,n_i}\})
$$

$$
+ \sum_{t=1}^{\infty} \sum_{n_i=\left\lceil (8\ln n)/\Delta_{\min}^2 \right\rceil}^{t-1} \sum_{n_{h(t)}=1}^{t-1}
$$

$$
(Pr\{\hat{\bar{\theta}}_{i,n_i} \leq \theta_i - C_{t,n_i}\} + Pr\{\hat{\bar{\theta}}_{h(t),n_{h(t)}} \geq \theta_{h(t)} + C_{t,n_{h(t)}}\})
$$

$$
\leq 1 + 2 \sum_{t=1}^{\infty} \sum_{n_{j(t)}=1}^{t-1} \sum_{n_i=1}^{t-1} 2t^{-4} \leq 1 + \frac{2\pi^2}{3}.
$$
(44)

So when (43) is satisfied, a tighter bound for the regret in (35) is:

$$
\mathfrak{R}^{\pi}(\Theta; n) \leq M \sum_{i \notin \mathcal{O}_M^*} \left(\frac{8\ln n}{\Delta_{\min,i}^2} + 1 + \frac{2\pi^2}{3}\right)\theta_{\max}
$$

$$
+ M^2\left(1 + \frac{2\pi^2}{3}\right)\theta_{\max} + M(M-1)\left(1 + \frac{2\pi^2}{3}\right) \sum_{i \in \mathcal{O}_M^*} \theta_i.
$$
(45)

We could also write a concise (but looser) upper bound as:

$$
\mathfrak{R}^{\pi}(\Theta; n) \leq M(N - M)\left(\frac{8\ln n}{\Delta_{\min}} + 1 + \frac{2\pi^2}{3}\right)\theta_{\max}
$$

$$
+ M^3\left(1 + \frac{2\pi^2}{3}\right)\theta_{\max}.
$$
(46)

∎

Comparing Theorem 3 with Theorem 4 and Theorem 5, if we define $C = \frac{8(N+M)}{\Delta_{\min}^2} + \left(1 + \frac{2\pi^2}{3}\right)N + M$, we can see that the regret of the naive policy DLF-Naive grows as
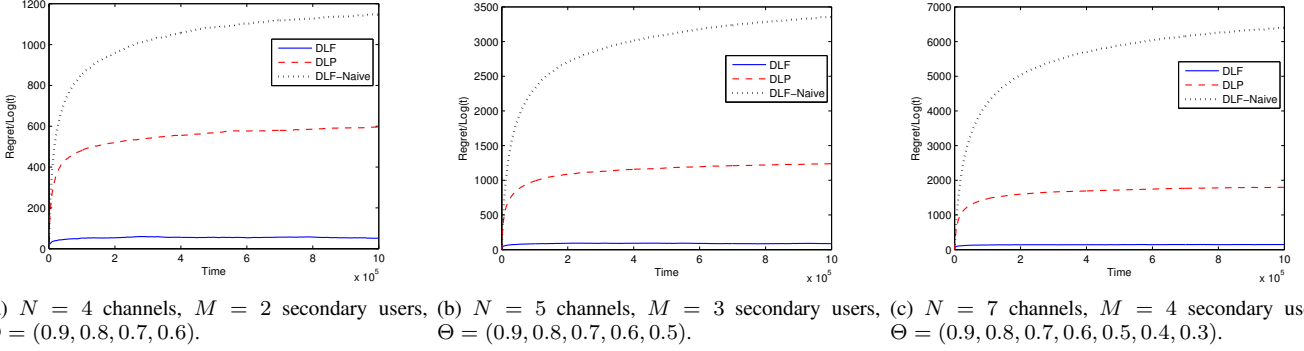
(a) $N = 4$ channels, $M = 2$ secondary users, $\Theta = (0.9, 0.8, 0.7, 0.6)$.

(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 $\frac{\Re(n)}{\ln n}$ vs. n time slots.

(n_i^m)

	i=1	i=2	i=3	i=4	i=5
m=1	995835	3030	659	659	184
m=2	2914	994062	2138	623	263
m=3	711	2570	993495	2640	584

(a) DLP policy.

(n_i^m)

	i=1	i=2	i=3	i=4	i=5
m=1	333328	333322	331068	1761	521
m=2	333328	333269	330982	1872	549
m=3	333326	333316	330764	2098	496

(b) DLF policy.

$(n_{j,i}^m), m = 1$

	i=1	i=2	i=3	i=4	i=5
j=1	329604	2749	474	330	176
j=2	2184	327301	3031	540	277
j=3	659	2282	327684	2109	600

$(n_{j,i}^m), m = 2$

	i=1	i=2	i=3	i=4	i=5
j=1	330062	2154	589	374	155
j=2	2952	327695	1759	703	224
j=3	743	1978	328347	1738	527

$(n_{j,i}^m), m = 3$

	i=1	i=2	i=3	i=4	i=5
j=1	329891	2303	732	261	146
j=2	2197	327574	2344	825	394
j=3	515	2519	327619	1932	748

(c) DLF-Naive policy.

Fig. 3. Number of times that channel i has been chosen by user m up to time $n = 10^6$, with $N = 5$ channels, $M = 3$ secondary users and $\Theta = (0.9, 0.8, 0.7, 0.6, 0.5)$.

$O(M^2(N+M)\ln n)$, while the regret of the DLF policy grows as $O(M(N + M^2)\ln n)$ when $\frac{n}{\ln n} < C$, $O(M(N - M)\ln n)$ when $\frac{n}{\ln n} \geq C$. So when n is large, the regret of DLF grows as $O(M(N - M)\ln n)$.

We also note that the following theorem has been shown in [11] on the lower bound of regret under any distributed policy.

Theorem 6 (Proposition 1 from [11]): The regret of any distributed policy π is lower-bounded by

$$\Re^\pi(\Theta; n) \geq \sum_{m=1}^{M} \sum_{i \notin \mathcal{O}_M^*} \Delta_{\min,i} \mathbb{E}[Q_i^m]. \tag{47}$$

Lai and Robbins [3] showed that for any uniformly good policy, the lower bound of Q_i^m for a single user i grows as $\Omega(\ln n)$. So DLF is a decentralized algorithm with finite-time order-optimal regret bound for fair access.

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. Figure 2(a) shows the simulations for $N = 4$ channels, $M = 2$ users, with $\Theta = (0.9, 0.8, 0.7, 0.6)$. In Figure 2(b), we have $N = 5$ channels, $M = 3$ users, and $\Theta = (0.9, 0.8, 0.7, 0.6, 0.5)$. In Figure 2(c), there are $N = 7$ channels, and $M = 4$ users, with $\Theta = (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3)$.

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. To show this fact, we present the number of times that a channel is accessed by all M users up to time $n = 10^6$ in Figure 3.

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

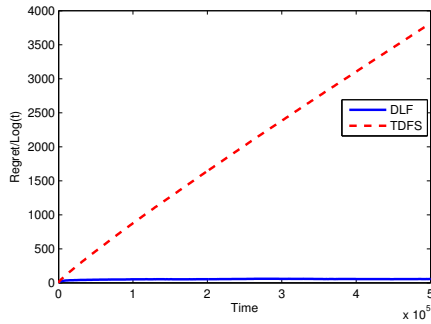the regret gap between different algorithms increases.

Fig. 4. Comparison of DLF and TDFS [8], [9].

In Figure 4, we compare the normalized regret $\frac{\Re(n)}{\ln n}$ of DLF logarithm and the TDFS algorithm proposed by Liu and Zhao [8], [9], in a system with $N = 4$ channels and $M = 2$ secondary users. $\Theta = (0.9, 0.8, 0.7, 0.6)$. The results are got by averaging 50 runs up to half million time slots. We can see that compared with TDFS, our proposed DLF algorithm not only has a better theoretical upper bound of regret, it also performs better for practical use. Also, TDFS only works for problems with single-parameterized distribution. We don't have this requirement for DFS. Besides, the storage of TDFS is not polynomial.

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 DLP, 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 DLF, 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 have further shown that the overall regret is lower for the fair access policy. We have considered the i.i.d. formulation throughout this work.

References

[1] Y. Gai and B. Krishnamachari, "Decentralized Online Learning Algorithms for Opportunistic Spectrum Access", *IEEE Global Communications Conference (GLOBECOM)*, December, 2011.

[2] 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.

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

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

[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 Transactions on Automatic Control*, vol. 32, pp. 977-982, 1987.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] A. Anandkumar, N. Michael, A. Tang, and A. Swami, "Distributed learning and allocation of cognitive users with logarithmic regret", *IEEE JSAC on Advances in Cognitive Radio Networking and Communications*, vol. 29, pp. 731-745, April, 2011.

[12] S.H.A. Ahmad and M. Liu, "Multi-channel Opportunistic Access: A Case of Restless Bandits with Multiple Plays", *the 47th Annual Allerton Conference on Communication, Control, and Computing*, September, 2009.

[13] K. Liu and Q. Zhao, "Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access", *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5547-5567, November, 2010.

[14] M. Lelarge, A. Proutiere, and M.S. Talebi, "Spectrum Bandit Optimization", *Information Theory Workshop (ITW)*, September, 2013.

[15] M. Di Felice, K. R. Chowdhury, and L. Bononi, "Learning with the Bandit: A Cooperative Spectrum Selection Scheme for Cognitive Radio Networks", *IEEE Global Communications Conference (GLOBECOM)*, December, 2011.

[16] P. Tehrani, Q. Zhao, and L. Tong, "Multi-Channel Opportunistic Spectrum Access in Unslotted Primary Systems with Unknown Models", *IEEE Workshops on CAMSAP*, December, 2011.

[17] P. Tehrani, K. Liu, and Q. Zhao, "Opportunistic Spectrum Access in Unslotted Primary Systems", *Journal of The Franklin Institute*, vol. 349, no. 3, pp. 985-1010, April, 2012.

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

Yi Gai received her B.E. and M.S. degrees in Electrical Engineering at Tsinghua University, China, in 2005 and 2007 respectively. She received her Ph.D. degree in the Department of Electrical Engineering, University of Southern California (USC), where she was a recipient of an Annenberg Graduate Fellowship, WiSE Merit Fellowship and Oakley Fellowship. She received the USC center for applied mathematical sciences (CAMS) prize in 2011. She is currently with Intel Labs as a research scientist. Her research interests are in the areas of machine learning, security, big data, mobile networks, and ad-hoc networks.

Bhaskar Krishnamachari received his B.E. in Electrical Engineering at The Cooper Union, New York, in 1998, and his M.S. and Ph.D. degrees from Cornell University in 1999 and 2002 respectively. He is currently an Associate Professor and a Ming Hsieh Faculty Fellow in the Department of Electrical Engineering at the University of Southern Californias Viterbi School of Engineering. His primary research interest is in the design and analysis of algorithms and protocols for next-generation wireless networks.