# Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret

Animashree Anandkumar, *Member, IEEE*, Nithin Michael, *Student Member, IEEE*,
Ao Kevin Tang, *Member, IEEE*, and Ananthram Swami, *Fellow, IEEE*

***Abstract*—The problem of distributed learning and channel access is considered in a cognitive network with multiple secondary users. The availability statistics of the channels are initially unknown to the secondary users and are estimated using sensing decisions. There is no explicit information exchange or prior agreement among the secondary users and sensing and access decisions are undertaken by them in a completely distributed manner. We propose policies for distributed learning and access which achieve order-optimal cognitive system throughput (number of successful secondary transmissions) under self play, i.e., when implemented at all the secondary users. Equivalently, our policies minimize the sum *regret* in distributed learning and access, which is the loss in secondary throughput due to learning and distributed access. For the scenario when the number of secondary users is known to the policy, we prove that the total regret is logarithmic in the number of transmission slots. This policy achieves order-optimal regret based on a logarithmic lower bound for regret under any uniformly-good learning and access policy. We then consider the case when the number of secondary users is fixed but unknown, and is estimated at each user through feedback. We propose a policy whose sum regret grows only slightly faster than logarithmic in the number of transmission slots.**

***Index Terms*—Cognitive medium access control, multi-armed bandits, distributed algorithms, logarithmic regret.**

## I. Introduction

THERE has been extensive research on cognitive radio network in the past decade to resolve many challenges not encountered previously in traditional communication networks (see e.g., [2]). One of the main challenges is to achieve coexistence of heterogeneous users accessing the same part of the spectrum. In a hierarchical cognitive network, there are two classes of transmitting users, viz., the primary users who have priority in accessing the spectrum and the secondary users who opportunistically transmit when the primary user is idle. The secondary users are *cognitive* and can sense the spectrum to detect the presence of a primary transmission. However, due to resource and hardware constraints, they can sense only a part of the spectrum at any given time.

We consider a slotted cognitive system where each secondary user can sense and access only one orthogonal channel in each transmission slot (see Fig. 1). Under sensing constraints, it is thus beneficial for the secondary users to select channels with higher mean availability, i.e., channels which are less likely to be occupied by the primary users. However, in practice, the channel availability statistics are unknown to the secondary users at the start of the transmissions.

Since the secondary users are required to sense the medium before transmission, can these sensing decisions be used to *learn* the channel availability statistics? If so, using these estimated channel availabilities, can we design channel access rules which maximize the transmission throughput? Designing provably efficient algorithms to accomplish the above goals forms the focus of our paper. Such algorithms need to be efficient, both in terms of learning and channel access.

For any learning algorithm, there are two important performance criteria: *consistency* and *regret bounds* [3]. A learning algorithm is said to be consistent if the learnt estimates converge to the true values as the number of samples goes to infinity. The regret[1] of a learning algorithm is a measure of the speed of convergence, and is thus meaningful only for consistent algorithms. In our context, the users need to learn the channel availability statistics consistently in a distributed manner. The regret in this case is defined as the loss in secondary throughput due to learning when compared to the ideal scenario where the channel statistics are known perfectly (see (2)). It is thus desirable to design distributed learning algorithms with small regret.

Additionally, we consider a distributed framework where there is no information exchange or prior agreement among the secondary users. This introduces additional challenges: it results in loss of throughput due to collisions among the secondary users, and there is now competition among the secondary users since they all tend to access channels with higher availabilities. It is imperative that the channel access policies overcome the above challenges. Hence, a distributed learning and access policy experiences regret both due to learning the unknown channel availabilities as well as due to collisions under distributed access.

Manuscript received 1 December 2009; revised 4 June 2010. During the stint of this work, the first author was supported by MURI through AFOSR Grant FA9550-06-1-0324. The second and the third authors are supported in part through NSF grant CCF-0835706. Parts of this paper were presented at [1].

A. Anandkumar is with the Center for Pervasive Communications and Computing at the Department of Electrical Engineering and Computer Science, University of California, Irvine, CA, 92697, USA (e-mail: a.anandkumar@uci.edu).

N. Michael and A. K. Tang are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA (e-mail: nm373@,atang@ece.cornell.edu).

A. Swami is with the Army Research Laboratory, Adelphi, MD 20783, USA (e-mail: a.swami@ieee.org).

Digital Object Identifier 10.1109/JSAC.2011.110406.

[1] Note that the sum regret is a finer measure of performance of a distributed algorithm than the time-averaged total throughput of the users since any sub-linear regret (with respect to time) implies optimal average throughput. On the other hand, regret is equivalent to total throughput and hence, optimal regret is equivalent to achieving optimal total throughput.

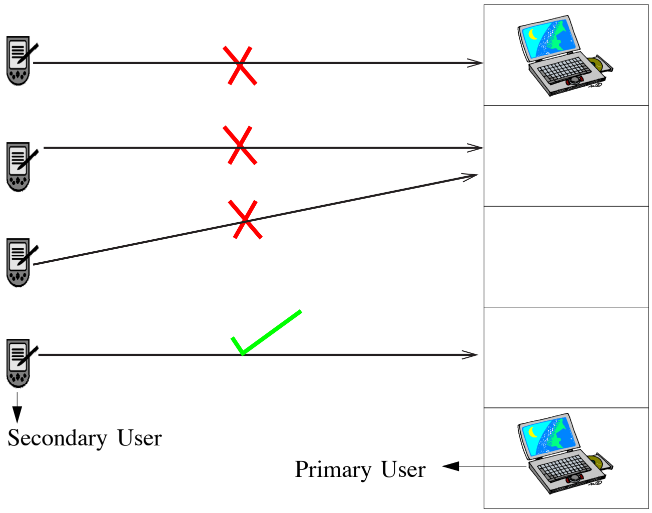


Fig. 1. Cognitive radio network with $U = 4$ secondary users and $C = 5$ channels. A secondary user is not allowed to transmit if the accessed channel is occupied by a primary user. If more than one secondary user transmits in the same free channel, then all the transmissions are unsuccessful.

### A. Our Contributions

The main contributions of this paper are two fold. First, we propose two distributed learning and access policies under multiple secondary users in a cognitive network. Second, we provide performance guarantees for these policies in terms of regret. Overall, we prove that one of our proposed algorithms achieves order-optimal regret and the other achieves nearly order-optimal regret, where the order is with respect to the number of transmission slots.

The first policy we propose assumes that the total number of secondary users in the system is known while our second policy removes this requirement. We provide bounds on the total regret experienced by the secondary users under self play, i.e., when implemented at all the secondary users. For the first policy, we prove that the regret is logarithmic, i.e., $O(\log n)$ where $n$ in the number of transmission slots. For the second policy, the regret grows slightly faster than logarithmic, i.e., $O(f(n) \log n)$, where we can choose any function $f(n)$ satisfying $f(n) \to \infty$, as $n \to \infty$. Hence, we provide performance guarantees for the proposed distributed learning and access policies.

A lower bound on regret under any *uniformly-good* distributed learning policy has been derived in [4], which is also logarithmic in the number of transmission slots (See (6) for definition of uniformly-good policies). Thus, our first policy (which requires knowledge of the number of secondary users) achieves order-optimal regret. The effects of the number of secondary users and the number of channels on regret are also explicitly characterized and verified via simulations.

The *exploration-exploitation* tradeoff for learning, combined with the *cooperation-competition* tradeoffs among multiple users for distributed medium access have not been sufficiently examined in the literature (see Section I-B for a discussion). Our analysis in this paper provides important engineering insights towards dealing with learning, competition, and cooperation in practical cognitive systems.

*Remark on Model Assumptions:* We note some of the limitations of our approach. We prove order-optimal regret for our proposed algorithms under an i.i.d. primary transmission model[2], where the probability of a channel being occupied by any primary user in any slot is assumed to be i.i.d. This is indeed is idealistic and in practice, a Markovian model may be more appropriate [5], [6]. The i.i.d. model leads to simple learning schemes for channel availabilities but is not crucial towards deriving regret bounds for our proposed schemes. Extensions of the classical multi-armed bandit problem to a Markovian model are considered in [7]. In principle, our results on achieving order-optimal regret for distributed learning and access can be similarly extended to a Markovian channel model but this entails more complex estimators and rules for evaluating the exploration-exploitation tradeoffs of different channels and is a topic of interest for future investigation.

### B. Related Work

Several results on the multi-armed bandit problem will be used and generalized to study our problem. Detailed discussion on multi-armed bandits can be found in [8]–[11]. Cognitive medium access is a topic of extensive research; see [12] for an overview. The connection between cognitive medium access and the multi-armed bandit problem is explored in [13], where a restless bandit formulation is employed. Under this formulation, indexability is established, Whittle's index for channel selection is obtained in closed-form, and the equivalence between the myopic policy and the Whittle's index is established. However, this work assumes known channel availability statistics and does not consider competing secondary users. The work in [14] considers allocation of two users to two channels under a Markovian channel model using a partially observable Markov decision process (POMDP) framework. The use of collision feedback information for learning, and spatial heterogeneity in spectrum opportunities were investigated. However, the difference from our work is that [14] assumes that the availability statistics (transition probabilities) of the channels are known to the secondary users while we consider learning of unknown channel statistics. The works in [15], [16] consider centralized access schemes in contrast to distributed access here, the work in [17] considers access through information exchange and studies the optimal choice of the amount of information to be exchanged given the cost of negotiation. The work in [18] considers access under $Q$-learning for two users and two channels where users can sense both the channels simultaneously. The work in [19] discusses a game-theoretic approach to cognitive medium access. In [20], learning in congestion games through multiplicative updates is considered and convergence to weakly-stable equilibria (which reduces to the pure Nash equilibrium for almost all games) is proven. However, the work assumes fixed costs (or equivalently rewards) in contrast to random rewards here, and that the players can fully observe the actions of other players.

[2] By i.i.d. primary transmission model, we do not mean the presence of a single primary user, but rather, this model is used to capture the overall statistical behavior of all the primary users in the system.

Recently, the work in [21] considers combinatorial bandits, where a more general model of different (unknown) channel availabilities is assumed for different secondary users, and a matching algorithm is proposed for jointly allocating users to channels. The algorithm is guaranteed to have logarithmic regret with respect to the number of transmission slots and polynomial storage requirements. A decentralized implementation of the proposed algorithm is proposed but it still requires information exchange and coordination among the users. In contrast, we propose algorithms which removes this requirement albeit in a more restrictive setting of identical channel availabilities for all users.

Recently, Liu and Zhao [4] proposed a family of distributed learning and access policies known as time-division fair share (TDFS), and proved logarithmic regret for these policies. They established a lower bound on the growth rate of system regret for a general class of uniformly-good decentralized polices. The TDFS policies in [4] can incorporate any order-optimal single-player policy while our work here is based on the single-user policy proposed in [11]. Another difference is that in [4], the users orthogonalize via settling at different offsets in their time-sharing schedule, while in our work here, users orthogonalize into different channels. Moreover, the TDFS policies ensure that each player achieves the same time-average reward while our policies here achieve probabilistic fairness, in the sense that the policies do not discriminate between different users. In [22], the TDFS policies are extended to incorporate imperfect sensing.

In our prior work [1], we first formulated the problem of decentralized learning and access for multiple secondary users. We considered two scenarios: one where there is initial common information among the secondary users in the form of pre-allocated ranks, and the other where no such information is available. In this paper, we analyze the distributed policy in detail and prove that it has logarithmic regret. In addition, we also consider the case when the number of secondary users is unknown, and provide bounds on regret in this scenario.

*Organization:* Section II deals with the system model, Section III deals with the special case of single secondary user and of multiple users with centralized access which can be directly solved using the classical results on multi-armed bandits. In Section IV, we propose distributed learning and access policy with provably logarithmic regret when the number of secondary users is known. Section V considers the scenario when the number of secondary users is unknown. Section VI provides a lower bound for distributed learning. Section VII includes some simulation results for the proposed schemes and Section VIII concludes the paper. Most of the proofs are found in the Appendix.

Since Section III mostly deals with a recap of the classical results on multi-armed bandits, those familiar with the literature (e.g., [8]–[11]) may skip this section without loss of continuity.

## II. System Model & Formulation

*Notation:* For any two functions $f(n), g(n)$, $f(n) = O(g(n))$ if there exists a constant $c$ such that $f(n) \leq cg(n)$ for all $n \geq n_0$ for a fixed $n_0 \in \mathbb{N}$. Similarly, $f(n) = \Omega(g(n))$ if there exists a constant $c'$ such that $f(n) \geq c'g(n)$ for all $n \geq n_0$ for a fixed $n_0 \in \mathbb{N}$, and $f(n) = \Theta(g(n))$ if $f(n) = \Omega(g(n))$ and $f(n) = O(g(n))$. Also, $f(n) = o(g(n))$ when $f(n)/g(n) \to 0$ and $f(n) = \omega(g(n))$ when $f(n)/g(n) \to \infty$ as $n \to \infty$.

For a vector $\boldsymbol{\mu}$, let $|\boldsymbol{\mu}|$ denote its cardinality, and let $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_{|\boldsymbol{\mu}|}]^T$. We refer to the $U$ highest entries in a vector $\boldsymbol{\mu}$ as the $U$-best channels and the rest as the $U$-worst channels, where $1 \leq U \leq |\boldsymbol{\mu}|$. Let $\sigma(a; \boldsymbol{\mu})$ denote the index of the $a^{\text{th}}$ highest entry in $\boldsymbol{\mu}$. Alternatively, we abbreviate $a^*{:=}\sigma(a; \boldsymbol{\mu})$ for ease of notation. Let $B(\mu)$ denote the Bernoulli distribution with mean $\mu$. With abuse of notation, let $D(\mu_1, \mu_2){:=}D(B(\mu_1); B(\mu_2))$ be the Kullback-Leibler distance between the Bernoulli distributions $B(\mu_1)$ and $B(\mu_2)$ [23] and let $\Delta(1,2){:=}\mu_1 - \mu_2$.

### A. Sensing & Channel Models

Let $U \geq 1$ be the number of secondary users[3] and $C \geq U$ be the number[4] of orthogonal channels available for slotted transmissions with a fixed slot width. In each channel $i$ and slot $k$, the primary user transmits i.i.d. with probability $1 - \mu_i > 0$. In other words, let $W_i(k)$ denote the indicator variable if the channel is free

$$W_i(k) = \begin{cases} 0, & \text{channel } i \text{ occupied in slot } k \\ 1, & \text{o.w,} \end{cases}$$

and we assume that $W_i(k) \overset{i.i.d.}{\sim} B(\mu_i)$.

The mean availability vector $\boldsymbol{\mu}$ consists of mean availabilities $\mu_i$ of all channels, i.e., is $\boldsymbol{\mu}{:=}[\mu_1, \ldots, \mu_C]$, where all $\mu_i \in (0,1)$ and are distinct. $\boldsymbol{\mu}$ is initially unknown to all the secondary users and is learnt *independently* over time using the past sensing decisions without any information exchange among the users. We assume that channel sensing is perfect at all the users.

Let $T_{i,j}(k)$ denote the number of slots where channel $i$ is sensed in $k$ slots by user $j$ (not necessarily being the sole user to sense that channel). Note that $\sum_{i=1}^{C} T_{i,j}(k) = k$ for all users $j$, since each user senses exactly one channel in every time slot. The sensing variables are obtained as follows: at the beginning of the $k^{\text{th}}$ slot, each secondary user $j \in U$ selects exactly one channel $i \in C$ for sensing, and hence, obtains the value of $W_i(k)$, indicating if the channel is free. User $j$ then records all the sensing decisions of each channel $i$ in a vector $\mathbf{X}_{i,j}^k{:=}[X_{i,j}(1), \ldots, X_{i,j}(T_{i,j}(k))]^T$. Hence, $\cup_{i=1}^{C} \mathbf{X}_{i,j}^k$ is the collection of sensed decisions for user $j$ in $k$ slots for all the $C$ channels.

We assume the collision model under which if two or more users transmit in the same channel then none of the transmissions go through. At the end of each slot $k$, each user $j$ receives acknowledgement $Z_j(k)$ on whether its transmission in the $k^{\text{th}}$ slot was received. Hence, in general, any policy employed by user $j$ in the $(k+1)^{\text{th}}$ slot, denoted by $\rho(\cup_{i=1}^{C} \mathbf{X}_{i,j}^k, \mathbf{Z}_j^k)$, is based on all the previous sensing and feedback results.

[3] A user refers to a secondary user unless otherwise mentioned.

[4] When $U \geq C$, learning availability statistics is less crucial, since all channels need to be accessed to avoid collisions. In this case, design of medium access is more crucial.

**Algorithm 1** Single User Policy $\rho^1(\mathbf{g}(n))$ in [10].

```
Input: $\{\bar{X}_i(n)\}_{i=1,\ldots,C}$ : Sample-mean availabilities after $n$ rounds, $g(i;n)$: statistic based on $\bar{X}_{i,j}(n)$,
$\sigma(a;\mathbf{g}(n))$: index of $a^{\text{th}}$ highest entry in $\mathbf{g}(n)$.
Init: Sense in each channel once, $n \leftarrow C$, $Curr\_Sel \leftarrow C$.
Loop: $n \leftarrow n+1$
$Curr\_Sel \leftarrow$ channel corresponding to highest entry in $\mathbf{g}(n)$ for sensing. If free, transmit.
```

### B. Regret of a Policy

Under the above model, we are interested in designing policies $\rho$ which maximize the expected number of successful transmissions of the secondary users subject to the non-interference constraint for the primary users. Let $S(n;\boldsymbol{\mu},U,\rho)$ be the expected total number of successful transmissions after $n$ slots under $U$ number of secondary users and policy $\rho$.

In the ideal scenario where the availability statistics $\boldsymbol{\mu}$ are known a priori and a central agent orthogonally allocates the secondary users to the $U$-best channels, the expected number of successful transmissions after $n$ slots is given by

$$S^*(n;\boldsymbol{\mu},U):=n\sum_{j=1}^{U}\mu(j^*), \tag{1}$$

where $j^*$ is the $j^{\text{th}}$-highest entry in $\boldsymbol{\mu}$.

It is clear that $S^*(n;\boldsymbol{\mu},U) > S(n;\boldsymbol{\mu},U,\rho)$ for any policy $\rho$ and any finite $n \in \mathbb{N}$. We are interested in minimizing the *regret* in learning and access, given by

$$R(n;\boldsymbol{\mu},U,\rho):=S^*(n;\boldsymbol{\mu},U) - S(n;\boldsymbol{\mu},U,\rho) > 0. \tag{2}$$

We are interested in minimizing regret under any given $\boldsymbol{\mu} \in (0,1)^C$ with distinct elements.

By incorporating the collision channel model assumption with no avoidance mechanisms[5], the expected throughput under the policy $\rho$ is given by

$$S(n;\boldsymbol{\mu},U,\rho) = \sum_{i=1}^{C}\sum_{j=1}^{U}\mu(i)\mathbb{E}[V_{i,j}(n)],$$

where $V_{i,j}(n)$ is the number of times in $n$ slots where user $j$ is the sole user to sense channel $i$. Hence, the regret in (2) simplifies as

$$R(n;\rho) = \sum_{k=1}^{U} n\mu(k^*) - \sum_{i=1}^{C}\sum_{j=1}^{U}\mu(i)\mathbb{E}[V_{i,j}(n)]. \tag{3}$$

## III. Special Cases From Known Results

We recap the bounds for the regret under the special cases of a single secondary user ($U=1$) and multiple users with centralized learning and access by appealing to the classical results on the multi-armed bandit process [8]–[10].

### A. Single Secondary User ($U=1$)

When there is only one secondary user, the problem of finding policies with minimum regret reduces to that of a multi-armed bandit process. Lai and Robbins [8] first analyzed schemes for multi-armed bandits with asymptotic logarithmic regret based on the upper confidence bounds on the unknown channel availabilities. Since then, simpler schemes have been proposed in [10], [11] which compute a statistic or an index for each arm (channel), henceforth referred to as the *g-statistic*, based only on its sample mean and the number of slots where the particular arm is sensed. The arm with the highest index is selected in each slot in these works. We summarize the policy in Algorithm 1 and denote it $\rho^1(\mathbf{g}(n))$, where $\mathbf{g}(n)$ is the vector of scores assigned to the channels after $n$ transmission slots.

The sample-mean based policy in [11, Thm. 1] proposes an index for each channel $i$ and user $j$ at time $n$ is given by

$$g_j^{\text{MEAN}}(i;n):=\bar{X}_{i,j}(T_{i,j}(n)) + \sqrt{\frac{2\log n}{T_{i,j}(n)}}, \tag{4}$$

where $T_{i,j}(n)$ is the number of slots where user $j$ selects channel $i$ for sensing,

$$\bar{X}_{i,j}(T_{i,j}(n)):= \sum_{k=1}^{T_{i,j}(n)} \frac{X_{i,j}(k)}{T_{i,j}(n)}$$

is the sample-mean availability of channel $i$, as sensed by user $j$.

The statistic in (4) captures the *exploration-exploitation* tradeoff between sensing the channel with the best predicted availability to maximize immediate throughput and sensing different channels to obtain improved estimates of their availabilities. The sample-mean term in (4) corresponds to exploitation while the other term involving $T_{i,j}(n)$ corresponds to exploration since it penalizes channels which are not sensed often. The normalization of the exploration term with $\log n$ in (4) implies that the term is significant when $T_{i,j}(n)$ is much smaller than $\log n$. On the other hand, if all the channels are sensed $\Theta(\log n)$ number of times, the exploration terms become unimportant in the $g$-statistics of the channels and the exploitation term dominates, thereby, favoring sensing of the channel with the highest sample mean.

The regret based on the above statistic in (4) is logarithmic for any finite number of slots $n$ but does not have the optimal scaling constant. The sample-mean based statistic in [10, Example 5.7] leads to the optimal scaling constant for regret and is given by

$$g_j^{\text{OPT}}(i;n):=\bar{X}_{i,j}(T_{i,j}(n)) + \min\left[\sqrt{\frac{\log n}{2T_{i,j}(n)}}, 1\right]. \tag{5}$$

In this paper, we design policies based on the $g^{\text{MEAN}}$ statistic since it is simpler to analyze than the $g^{\text{OPT}}$ statistic. This is because $g^{\text{MEAN}}$ is a continuous function in $T_{i,j}(n)$ while $g^{\text{OPT}}$ has a threshold on $T_{i,j}(n)$.

We now recap the results which show logarithmic regret in learning the best channel. In this context, we define *uniformly good* policies $\rho$ [8] as those with regret

$$R(n;\boldsymbol{\mu},U,\rho) = o(n^\alpha), \quad \forall \alpha > 0, \boldsymbol{\mu} \in (0,1)^C. \tag{6}$$

[5] If the users employ CSMA-CA to avoid collisions then the regret is reduced. The bounds derived in this paper are applicable for this case as well.

*Theorem 1 (Logarithmic Regret for $U = 1$ [10], [11]):* For any uniformly good policy $\rho$ satisfying (6), the expected time spent in any suboptimal channel $i \neq 1^*$ satisfies

$$\lim_{n\to\infty} \mathbb{P}\left[T_{i,1}(n) \geq \frac{(1-\epsilon)\log n}{D(\mu_i, \mu_{1^*})}; \boldsymbol{\mu}\right] = 1, \tag{7}$$

where $1^*$ is the channel with the best availability. Hence, the regret satisfies

$$\liminf_{n\to\infty} \frac{R(n;\boldsymbol{\mu},1,\rho)}{\log n} \geq \sum_{i\in 1\text{-worst}} \frac{\Delta(1^*,i)}{D(\mu_i,\mu_{1^*})}. \tag{8}$$

The regret under the $g^{\text{OPT}}$ statistic in (5) achieves the above bound.

$$\lim_{n\to\infty} \frac{R(n;\boldsymbol{\mu},1,\rho^1(\mathbf{g}_j^{\text{OPT}}))}{\log n} = \sum_{i\in 1\text{-worst}} \frac{\Delta(1^*,i)}{D(\mu_i,\mu_{1^*})}. \tag{9}$$

The regret under $g^{\text{MEAN}}$ statistic in (34) satisfies

$$R(n;\boldsymbol{\mu},1,\rho^1(\mathbf{g}_j^{\text{MEAN}})) \leq \sum_{i\neq 1^*} \Delta(1^*,i)\left[\frac{8\log n}{\Delta(j^*,i)^2} + 1 + \frac{\pi^2}{3}\right].$$

### *B. Centralized Learning & Access for Multiple Users*

We now consider multiple secondary users under centralized access policies where there is joint learning and access by a central agent on behalf of all the $U$ users. Here, to minimize the sum regret, the centralized policy allocates the $U$ users to orthogonal channels to avoid collisions. Let $\rho^{\text{CENT}}(\mathcal{X}^k)$, with $\mathcal{X}^k := \cup_{j=1}^U \cup_{i=1}^C \mathbf{X}_{i,j}^k$, denote a centralized policy based on the sensing variables of all the users. The policy under centralized learning is a simple generalization of the single-user policy and is given in Algorithm 2. We now recap the results of [9].

*Theorem 2 (Regret Under Centralized Policy $\rho^{\text{CENT}}$ [9]):* For any uniformly good centralized policy $\rho^{\text{CENT}}$ satisfying (6), the expected times spent in a $U$-worst channel $i$ satisfies

$$\lim_{n\to\infty} \mathbb{P}\left[\sum_{j=1}^{U} T_{i,j}(n) \geq \frac{(1-\epsilon)\log n}{D(\mu_i,\mu_{U^*})}; \boldsymbol{\mu}\right] = 1, \tag{10}$$

where $U^*$ is the channel with the $U^{\text{th}}$ best availability. Hence, the regret satisfies

$$\liminf_{n\to\infty} \frac{R(n;\boldsymbol{\mu},1,\rho^{\text{CENT}})}{\log n} \geq \sum_{i\in U\text{-worst}} \frac{\Delta(U^*,i)}{D(\mu_i,\mu_{U^*})}. \tag{11}$$

The scheme in Algorithm 2 based on $g^{\text{OPT}}$ achieves the above bound.

$$\lim_{n\to\infty} \frac{R(n;\boldsymbol{\mu},1,\rho^{\text{CENT}}(\mathbf{g}^{\text{OPT}}))}{\log n} = \sum_{i\in U\text{-worst}} \frac{\Delta(U^*,i)}{D(\mu_i,\mu_{U^*})}. \tag{12}$$

The scheme in Algorithm 2 based on the $g^{\text{MEAN}}$ satisfies for any $n > 0$,

$$R(n;\boldsymbol{\mu},U,\rho^{\text{CENT}}(\mathbf{g}^{\text{MEAN}})) \leq \sum_{m=1}^{U} \sum_{i\in U\text{-worst}} \sum_{k=1}^{U} \frac{\Delta(m^*,i)}{U}\left[\frac{8\log n}{\Delta(m^*,i)^2} + 1 + \frac{\pi^2}{3}\right]. \tag{13}$$

*Proof:* See Appendix A. □

**Algorithm 2** Centralized Learning Policy $\rho^{\text{CENT}}$ in [9].

```
Input: $\mathcal{X}^n := \cup_{j=1}^U \cup_{i=1}^C \mathbf{X}_{i,j}^n$ : Channel availability after $n$ slots, $\mathbf{g}(n)$: statistic based on $\mathcal{X}^n$,
$\sigma(a;\mathbf{g}(n))$: index of $a^{\text{th}}$ highest entry in $\mathbf{g}(n)$.
Init: Sense in each channel once, $n \leftarrow C$
Loop: $n \leftarrow n+1$
$Curr\_Sel$ ← channels with $U$-best entries in $\mathbf{g}(n)$. If free, transmit.
```

Notice that for both the single user case $U = 1$ (Theorem 1) and the centralized multi-user case (Theorem 2), the number of time slots spent in the U-worst channels is $\Theta(\log n)$ and hence, the regret is also $\Theta(\log n)$.

## IV. Main Results

Armed with the classical results on multi-armed bandits, we now design distributed learning and allocation policies.

### *A. Preliminaries: Bounds on Regret*

We first provide simple bounds on the regret in (3) for any distributed learning and access policy $\rho$.

*Proposition 1 (Lower and Upper Bounds on Regret):* The regret under any distributed policy $\rho$ satisfies

$$R(n;\rho) \geq \sum_{j=1}^{U} \sum_{i\in U\text{-worst}} \Delta(U^*,i)\mathbb{E}[T_{i,j}(n)], \tag{14}$$

$$R(n;\rho) \leq \mu(1^*)\left[\sum_{j=1}^{U} \sum_{i\in U\text{-worst}} \mathbb{E}[T_{i,j}(n)] + \mathbb{E}[M(n)]\right], \tag{15}$$

where $T_{i,j}(n)$ is the number of slots where user $j$ selects channel $i$ for sensing, $M(n)$ is the number of collisions faced by the users in the $U$-best channels in $n$ slots, $\Delta(i,j) = \mu(i) - \mu(j)$ and $\mu(1^*)$ is the highest mean availability.

*Proof:* See Appendix B. □

In the subsequent sections, we propose distributed learning and access policies and provide regret guarantees for the policies using the upper bound in (15). The lower bound in (14) can be used to derive lower bound on regret for any uniformly-good policy.

The first term in (15) represents the lost transmission opportunities due to selection of $U$-worst channels (with lower mean availabilities), while the second term represents the performance loss due to collisions among the users in the $U$-best channels. The first term in (15) decouples among the different users and can be analyzed solely through the marginal distributions of the $g$-statistics at the users. This in turn, can be analyzed by manipulating the classical results on multi-armed bandits [10], [11]. On the other hand, the second term in (15), involving collisions in the $U$-best channels, requires the joint distribution of the $g$-statistics at different users which are correlated variables. This is intractable to analyze directly and we develop techniques to bound this term.

**Algorithm 3** Policy $\rho^{\text{RAND}}(U, C, \mathbf{g}_j(n))$ for each user $j$ under $U$ users, $C$ channels and statistic $\mathbf{g}_j(n)$.

```
Input: $\{\bar{X}_{i,j}(n)\}_{i=1,\ldots,C}$ : Sample-mean availabilities at user $j$ after $n$ rounds, $g_j(i;n)$: statistic based on $\bar{X}_{i,j}(n)$, $\sigma(a;\mathbf{g}_j(n))$: index of $a^{\text{th}}$ highest entry in $\mathbf{g}_j(n)$.
$\zeta_j(i;n)$: indicator of collision at $n^{\text{th}}$ slot at channel $i$
Init: Sense in each channel once, $n \leftarrow C$, $Curr\_Rank \leftarrow 1$, $\zeta_j(i;m) \leftarrow 0$
Loop: $n \leftarrow n+1$
if $\zeta_j(Curr\_Sel; n-1) = 1$ then
    Draw a new $Curr\_Rank \sim$ Unif$(U)$
end if
Select channel for sensing. If free, transmit.
$Curr\_Sel \leftarrow \sigma(Curr\_Rank; \mathbf{g}_j(n))$.
If collision $\zeta_j(Curr\_Sel; m) \leftarrow 1$, Else 0.
```

### B. $\rho^{\text{RAND}}$ : Distributed Learning and Access

We present the $\rho^{\text{RAND}}$ policy in Algorithm 3. Before describing this policy, we make some simple observations. If each user implemented the single-user policy in Algorithm 1, then it would result in collisions, since all the users target the best channel. When there are multiple users and there is no direct communication among them, the users need to randomize channel access in order to avoid collisions. At the same time, accessing the $U$-worst channels needs to be avoided since they contribute to regret. Hence, users can avoid collisions by randomizing access over the $U$-best channels, based on their estimates of the channel ranks. However, if the users randomize in every slot, there is a finite probability of collisions in every slot and this results in a linear growth of regret with the number of time slots. Hence, the users need to converge to a collision-free configuration to ensure that the regret is logarithmic.

In Algorithm 3, there is adaptive randomization based on feedback regarding the previous transmission. Each user randomizes *only* if there is a collision in the previous slot; otherwise, the previously generated random rank for the user is retained. The estimation for the channel ranks is through the $g$-statistic, along the lines of the single-user case.

### C. Regret Bounds under $\rho^{\text{RAND}}$

It is easy to see that the $\rho^{\text{RAND}}$ policy ensures that the users are allocated orthogonally to the $U$-best channels as the number of transmission slots goes to infinity. The regret bounds on $\rho^{\text{RAND}}$ are however not immediately clear and we provide guarantees below.

We first provide a logarithmic upper bound[6] on the number of slots spent by each user in any $U$-worst channel. Hence, the first term in the bound on regret in (15) is also logarithmic.

*Lemma 1 (Time Spent in U-worst Channels):* Under the $\rho^{\text{RAND}}$ scheme in Algorithm 3, the total time spent by any user $j = 1, \ldots, U$, in any $i \in U$-worst channel is given by

$$\mathbb{E}[T_{i,j}(n)] \le \sum_{k=1}^{U} \Big[\frac{8 \log n}{\Delta(i, k^*)^2} + 1 + \frac{\pi^2}{3}\Big]. \tag{16}$$

*Proof:* The proof is similar along the lines of Theorem 2, given in Appendix A. The only difference is that here, we need to consider the probability of transmission of each user in each $U$-worst channel while in Appendix A, the probability of transmission by all users is considered. □

We now focus on analyzing the number of collisions $M(n)$ in the $U$-best channels. We first give a result on the expected number of collisions in the ideal scenario where each user has perfect knowledge of the channel availability statistics $\boldsymbol{\mu}$. In this case, the users attempt to reach an orthogonal (collision-free) configuration by uniformly randomizing over the $U$-best channels.

The stochastic process in this case is a finite-state Markov chain. A state in this Markov chain corresponds to a configuration of $U$ number of (identical) users in $U$ number of channels. The number of states in the Markov chain is the number of *compositions* of $U$, given by $\binom{2U-1}{U}$ [24, Thm. 5.1]. The orthogonal configuration corresponds to the absorbing state. There are two other classes of states in the Markov chain. One class consists of states in which each channel either has more than one user or no user at all. The second class consists of the remaining states where some channels may have just one user. For the first class, the transition probability to any state of the Markov chain (including self transition and absorption probabilities) is uniform. For a state in the second class, where certain channels have exactly one user, there are only transitions to states which have identical configuration of the single users and the transition probabilities are uniform. Let $\Upsilon(U, U)$ denote the maximum time to absorption in the above Markov chain starting from any initial distribution. We have the following result

*Lemma 2 (No. of Collisions Under Perfect Knowledge):* The expected number of collisions under $\rho^{\text{RAND}}$ scheme in Algorithm 3, assuming that each user has perfect knowledge of the mean channel availabilities $\boldsymbol{\mu}$, is given by

$$\begin{aligned}\mathbb{E}[M(n); \rho^{\text{RAND}}(U, C, \boldsymbol{\mu})] &\le U\mathbb{E}[\Upsilon(U, U)] \\ &\le U\left[\binom{2U-1}{U} - 1\right]. \end{aligned} \tag{17}$$

*Proof:* See Appendix C. □

The above result states that there is at most a finite number of expected collisions, bounded by $U\mathbb{E}[\Upsilon(U, U)]$ under perfect knowledge of $\boldsymbol{\mu}$. In contrast, recall from the previous section, that there are no collisions under perfect knowledge of $\boldsymbol{\mu}$ in the presence of pre-allocated ranks. Hence, $U\mathbb{E}[\Upsilon(U, U)]$ represents a bound on the additional regret due to the lack of direct communication among the users to negotiate their ranks.

We use the result of Lemma 2 for analyzing the number of collisions under distributed learning of the unknown availabilities $\boldsymbol{\mu}$ as follows: if we show that the users are able to learn the correct order of the different channels with only logarithmic regret then only an additional finite expected

[6] Note that the bound on $\mathbb{E}[T_{i,j}(n)]$ in (16) holds for user $j$ even if the other users are using a policy other than $\rho^{\text{RAND}}$. But on the other hand, to analyze the number of collisions $\mathbb{E}[M(n)]$ in (19), we need every user to implement $\rho^{\text{RAND}}$.

number of collisions occur before reaching an orthogonal configuration.

Define $T'(n;\rho^{\text{RAND}})$ as the number of slots where any one of the top $U$-estimated ranks of the channels at some user is wrong under the $\rho^{\text{RAND}}$ policy. Below we prove that its expected value is logarithmic in the number of transmissions.

*Lemma 3 (Wrong Order of g-statistics):* Under the $\rho^{\text{RAND}}$ scheme in Algorithm 3,

$$\mathbb{E}[T'(n;\rho^{\text{RAND}})] \leq U \sum_{a=1}^{U} \sum_{b=a+1}^{C} \left[ \frac{8 \log n}{\Delta(a^*, b^*)^2} + 1 + \frac{\pi^2}{3} \right]. \quad (18)$$

*Proof:* See Appendix D. □

We now provide an upper bound on the number of collisions $M(n)$ in the $U$-best channels by incorporating the above result on $\mathbb{E}[T'(n)]$, the result on the average number of slots $\mathbb{E}[T_{i,j}]$ spent in the $i \in U$-worst channels in Lemma 1 and the average number of collisions $U\mathbb{E}[\Upsilon(U,U)]$ under perfect knowledge of $\boldsymbol{\mu}$ in Lemma 2.

*Theorem 3 (Logarithmic Number of Collisions Under $\rho^{\text{RAND}}$):* The expected number of collisions in the $U$-best channels under $\rho^{\text{RAND}}(U, C, \mathbf{g}^{\text{MEAN}})$ scheme satisfies

$$\mathbb{E}[M(n)] \leq U(\mathbb{E}[\Upsilon(U,U)] + 1)\, \mathbb{E}[T'_j(n)]. \quad (19)$$

Hence, from (16), (18) and (17), $M(n) = O(\log n)$.

*Proof:* See Appendix E. □

Hence, there are only a logarithmic number of expected collisions before the users settle in the orthogonal channels. Combining this result with Lemma 1 that the number of slots spent in the $U$-worst channels is also logarithmic, we immediately have one of the main results of this paper that the sum regret under distributed learning and access is logarithmic.

*Theorem 4 (Logarithmic Regret Under $\rho^{\text{RAND}}$):* The policy $\rho^{\text{RAND}}(U, C, \mathbf{g}^{\text{MEAN}})$ in Algorithm 3 has $\Theta(\log n)$ regret.

*Proof:* Substituting (19) and (16) in (15). □

Hence, we prove that distributed learning and channel access among multiple secondary users is possible with logarithmic regret without any explicit communication among the users. This implies that the number of lost opportunities for successful transmissions at all secondary users is only logarithmic in the number of transmissions, which is negligible when there are a large number of transmissions.

We have so far focused on designing schemes that maximize system or social throughput. We now briefly discuss the fairness for an individual user under $\rho^{\text{RAND}}$. Since $\rho^{\text{RAND}}$ does not distinguish any of the users, in the sense that each user has equal probability of "settling" down in one of the $U$-best channels while experiencing only logarithmic regret in doing so. Simulations in Section VII (in Fig. 4) demonstrate this phenomenon.

We also note that the bound on regret under $\rho^{\text{RAND}}$ grows rapidly with the number of users $U$, due to the bound in (17). This is due to uniform channel access by the users without any coordination or information exchange. It is of interest to explore better distributed access schemes, which when combined with learning algorithms yield low regret in the number of users.

**Algorithm 4** Policy $\rho^{\text{EST}}(n, C, \mathbf{g}_j(m), \xi)$ for each user $j$ under $n$ transmission slots (horizon length), $C$ channels, statistic $\mathbf{g}_j(m)$ and threshold function $\xi$.

```
1) Input: $\{\bar{X}_{i,j}(n)\}_{i=1,\ldots,C}$ : Sample-mean availabilities
   at user $j$, $g_j(i;n)$: statistic based on $\bar{X}_{i,j}(n)$,
   $\sigma(a;\mathbf{g}_j(n))$: index of $a^{\text{th}}$ highest entry in $\mathbf{g}_j(n)$.
   $\zeta_j(i;n)$: indicator of collision at $n^{\text{th}}$ slot at channel $i$
   $\widehat{U}$: current estimate of the number of users.
   $n$: horizon (total number of slots for transmission)
   $\xi(n;k)$: threshold functions for $k = 1,\ldots,C$.
2) Init: Sense each channel once, $m \leftarrow C$, $Curr\_Sel \leftarrow$
   $C$, $Curr\_Rank \leftarrow 1$, $\widehat{U} \leftarrow 1$, $\zeta_j(i;m) \leftarrow 0$ for all
   $i = 1,\ldots,C$
3) Loop 4 to 6: $m \leftarrow m+1$, stop when $m = n$ or $\widehat{U} = C$.
4) If $\zeta_j(Curr\_Sel; m-1) = 1$ then
     Draw a new $Curr\_Rank \sim \text{Unif}(\widehat{U})$. end if
   Select channel for sensing. If free, transmit.
   $Curr\_Sel \leftarrow \sigma(Curr\_Rank; \mathbf{g}_j(m))$
5) $\zeta_j(Curr\_Sel; m) \leftarrow 1$ if collision, 0 o.w.
6) If $\sum_{a=1}^{m} \sum_{k=1}^{\widehat{U}} \zeta_j(\sigma(k;\mathbf{g}_j(m)); a) > \xi(n;\widehat{U})$ then
     $\widehat{U} \leftarrow \widehat{U} + 1$, $\zeta_j(i;a) \leftarrow 0$, $i = 1,\ldots C$, $a = 1,\ldots,m$.
   end if
```

## V. Distributed Learning and Access under Unknown Number of Users

We have so far assumed that the number of secondary users is known, and is required for the implementation of the $\rho^{\text{RAND}}$ policy. In practice, this entails initial announcement from each of the secondary users to indicate their presence in the cognitive network. However, in a truly distributed setting without any information exchange among the users, such an announcement may not be possible.

In this section, we consider the scenario, where the number of users $U$ is unknown (but fixed throughout the duration of transmissions and $U \leq C$, the number of channels). In this case, the policy needs to estimate the number of secondary users in the system, in addition to learning the channel availability statistics and designing channel access rules based on collision feedback. Note that if the policy assumed the worst-case scenario that $U = C$, then the regret grows linearly since there is a positive probability that the $U$-worst channels are selected for sensing in any time slot.

### A. Description of $\rho^{\text{EST}}$ Policy

We now propose a policy $\rho^{\text{EST}}$ in Algorithm 4. This policy incorporates two broad aspects in each transmission slot, viz., execution of the $\rho^{\text{RAND}}$ policy in Algorithm 3, based on the current estimate of the number of users $\widehat{U}$, and updating of the estimate $\widehat{U}$ based on the number of collisions experienced by the user.

The updating is based on the idea that if there is under-estimation of $U$ at all the users ($\widehat{U}_j < U$ at all the users $j$), collisions necessarily build up and the collision count serves as a criterion for incrementing $\widehat{U}$. This is because after a long learning period, the users learn the true ranks of the channels, and target the same set of channels. However, when there is under-estimation, the number of users exceeds the number of channels targeted by the users. Hence, collisions among the users accumulate, and can be used as a test for incrementing $\widehat{U}$.

Denote the collision count used by the $\rho^{\text{EST}}$ policy as

$$\Phi_{k,j}(m) := \sum_{a=1}^{m}\sum_{b=1}^{k}\zeta_j(\sigma(b;\mathbf{g}_j(m));a). \tag{20}$$

which is the total number of collisions experienced by user $j$ so far (till the $m^{\text{th}}$ transmission slot) in the top $\widehat{U}_j$-channels, where the ranks of the channels are estimated using the $g$-statistics. The collision count is tested against a threshold $\xi(n;\widehat{U}_j)$, which is a function of the horizon length[7] and current estimate $\widehat{U}_j$. When the threshold is exceeded, $\widehat{U}_j$ is incremented, and the collision samples collected so far are discarded (by setting them to zero) (line 6 in Algorithm 4). The choice of the threshold for incrementing $\widehat{U}_j$ is critical; if it is too small it can result in over-estimation of the number of users. On the other hand, if it is too large, it can result in slow learning and large regret. Threshold selection is studied in detail in the subsequent section.

### B. Regret Bounds under $\rho^{\text{EST}}$

We analyze regret bounds under the $\rho^{\text{EST}}$ policy, where the regret is defined in (3). Let the maximum threshold function for the number of consecutive collisions under the $\rho^{\text{EST}}$ policy be denoted by

$$\xi^*(n;U) := \max_{k=1,\ldots,U}\xi(n;k). \tag{21}$$

We prove that the $\rho^{\text{EST}}$ policy has $O(\xi^*(n;U))$ regret when $\xi^*(n;U)=\omega(\log n)$, and where $n$ is the number of transmission slots.

The proof for the regret bound under the $\rho^{\text{EST}}$ policy consists of two main parts: we prove bounds on regret conditioned on the event that none of the users over-estimate $U$. Second, we show that the probability of over-estimation at any of the users goes to zero asymptotically in the horizon length. Taken together, we obtain the regret bound for the $\rho^{\text{EST}}$ policy.

Note that in order to have small regret, it is crucial that none of the users over-estimate $U$. This is because when there is over-estimation, there is a finite probability of selecting the $U$-worst channels even upon learning the true ranks of the channels. Note that regret is incurred whenever a U-worst channel is selected since under perfect knowledge this channel would not be selected. Hence, under over-estimation, the regret grows linearly in the number of transmissions.

In a nutshell, under the $\rho^{\text{EST}}$ policy, the decision to increment the estimate $\widehat{U}$ reduces to a hypothesis-testing problem with hypotheses $\mathcal{H}_0$: number of users is less than or equal to the current estimate and $\mathcal{H}_1$: number of users is greater than the current estimate. In order to have a sub-linear regret, the false-alarm probability (deciding $\mathcal{H}_1$ under $\mathcal{H}_0$) needs to decay asymptotically. This is ensured by selecting appropriate thresholds $\xi(n)$ to test against the collision counts obtained through feedback.

*Conditional Regret:* We now give the result for the first part. Define the "good event" $\mathcal{C}(n;U)$ that none of the users over-estimates $U$ under $\rho^{\text{EST}}$ as

$$\mathcal{C}(n;U) := \{\bigcap_{j=1}^{U}\widehat{U}_j^{\text{EST}}(n)\le U\}. \tag{22}$$

The regret conditioned on $\mathcal{C}(n;U)$, denoted by $R(n;\boldsymbol{\mu},U,\rho^{\text{EST}})|\mathcal{C}(n;U)$, is given by

$$n\sum_{k=1}^{U}\mu(k^*)-\sum_{i=1}^{C}\sum_{j=1}^{U}\mu(i)\mathbb{E}[V_{i,j}(n)|\mathcal{C}(n;U)],$$

where $V_{i,j}(n)$ is the number of times that user $j$ is the sole user of channel $i$. Similarly, we have conditional expectations of $\mathbb{E}[T_{i,j}(n)|\mathcal{C}(n;U)]$ and of the number of collisions in $U$-best channels, given by $\mathbb{E}[M(n)|\mathcal{C}(n;U)]$. We now show that the regret conditioned on $\mathcal{C}(n;U)$ is $O(\max(\xi^*(n;U),\log n))$.

*Lemma 4: (Conditional Regret):* When all the $U$ secondary users implement the $\rho^{\text{EST}}$ policy, we have for all $i\in U$-worst channel and each user $j=1,\ldots,U$,

$$\mathbb{E}[T_{i,j}(n)|\mathcal{C}(n)]\le\sum_{k=1}^{U}\left[\frac{8\log n}{\Delta(i,k^*)^2}+1+\frac{\pi^2}{3}\right]. \tag{23}$$

The conditional expectation on number of collisions $M(n)$ in the $U$-best channel satisfies

$$\mathbb{E}[M(n)|\mathcal{C}(n;U)]\le U\sum_{k=1}^{U}\xi(n;k)\le U^2\xi^*(n;U). \tag{24}$$

From (15), we have $R(n)|\mathcal{C}(n;U)$ is $O(\max(\xi^*(n;U),\log n))$ for any $n\in\mathbb{N}$.

*Proof:* See Appendix F. □

*Probability of Over-estimation:* We now prove that none of the users over-estimates[8] $U$ under the $\rho^{\text{EST}}$ policy, i.e., the probability of the event $\mathcal{C}(n;U)$ in (22) approaches one as $n\to\infty$, when the thresholds $\xi(n;\widehat{U})$ for testing against the collision count are chosen appropriately (see line 6 in Algorithm 4). Trivially, we can set $\xi(n;1)=1$ since a single collision is enough to indicate that there is more than one user. For any other $k>1$, we choose function $\xi$ satisfying

$$\xi(n;k)=\omega(\log n),\quad\forall k>1. \tag{25}$$

We prove that the above condition ensures that over-estimation does not occur.

Recall that $T'(n;\rho^{\text{EST}})$ is the number of slots where any one of the top $U$-estimated ranks of the channels at some user is wrong under the $\rho^{\text{EST}}$ policy. We show that $\mathbb{E}[T'(n)]$ is $O(\log n)$.

*Lemma 5 (Time spent with wrong estimates):* The expected number of slots where any of the top $U$-estimated

[7] In this section, we assume that the users are aware of the horizon length $n$ for transmission. Note that this is not a limitation and can be extended to case of unknown horizon length as follows: implement the algorithm by fixing horizon lengths to $n_0, 2n_0, 4n_0\ldots$ for a fixed $n_0\in\mathbb{N}$ and discarding estimates from previous stages.

[8] Note that the $\rho^{\text{EST}}$ policy automatically ensures that all the users do not under-estimate $U$, since it increments $\widehat{U}$ based on collision estimate. This implies that the probability of the event that all the users under-estimate $U$ goes to zero asymptotically.

ranks of the channels at any user is wrong under the $\rho^{\text{EST}}$ policy satisfies

$$\mathbb{E}[T'(n)] \leq U \sum_{a=1}^{U} \sum_{b=a+1}^{C} \left[ \frac{8 \log n}{\Delta(a^*, b^*)^2} + 1 + \frac{\pi^2}{3} \right]. \quad (26)$$

*Proof:* The proof is on the lines of Lemma 3 □

Recall the definition of $\Upsilon(U, U)$ in the previous section, as the maximum time to absorption starting from any initial distribution of the finite-state Markov chain, where the states correspond to different user configurations and the absorbing state corresponds to the collision-free configuration. We now generalize the definition to $\Upsilon(U, k)$, as the time to absorption in a new Markov chain, where the state space is the set of configurations of $U$ users in $k$ channels, and the transition probabilities are defined on similar lines. Note that $\Upsilon(U, k)$ is almost-surely finite when $k \geq U$ and $\infty$ otherwise (since there is no absorbing state in the latter case).

We now bound the maximum value of the collision count $\Phi_{k,j}(m)$ under the $\rho^{\text{EST}}$ policy in (20) using $T'(m)$, the total time spent with wrong channel estimates, and $\Upsilon(U, k)$, the time to absorption in the Markov chain. Let $\overset{st}{\leq}$ denote the stochastic order for two random variables [25].

*Proposition 2:* The maximum collision count in (20) over all users under the $\rho^{\text{EST}}$ policy satisfies

$$\max_{j=1,\ldots,U} \Phi_{k,j}(m) \overset{st}{\leq} (T'(m) + 1)\Upsilon(U, k), \quad \forall m \in \mathbb{N}. \quad (27)$$

*Proof:* The proof is on the lines of Theorem 3. See Appendix G. □

We now prove that the probability of over-estimation goes to zero asymptotically.

*Lemma 6 (No Over-estimation Under $\rho^{\text{EST}}$):* For threshold function $\xi$ satisfying (25), the event $\mathcal{C}(n; U)$ in (22) satisfies

$$\lim_{n \to \infty} \mathbb{P}[\mathcal{C}(n; U)] = 1, \quad (28)$$

and hence, none of the users over-estimates $U$ under the $\rho^{\text{EST}}$ policy.

*Proof:* See Appendix H. □

We now give the main result of this section that $\rho^{\text{EST}}$ has slightly more than logarithmic regret asymptotically and this depends on the threshold function $\xi^*(n; U)$ in (21).

*Theorem 5 (Asymptotic Regret Under $\rho^{\text{EST}}$):* With threshold function $\xi$ satisfying conditions in (25), the policy $\rho^{\text{EST}}(n, C, \mathbf{g}_j(m), \xi)$ in Algorithm 4 satisfies

$$\limsup_{n \to \infty} \frac{R(n; \boldsymbol{\mu}, U, \rho^{\text{EST}})}{\xi^*(n; U)} < \infty. \quad (29)$$

*Proof:* From Lemma 4 and Lemma 6. □

Hence, the regret under the proposed $\rho^{\text{EST}}$ policy is $O(\xi^*(n; U))$ under fully decentralized setting without the knowledge of number of users when $\xi^*(n; U) = \omega(\log n)$. Hence, $O(f(n) \log n)$ regret is achievable for all functions $f(n) \to \infty$ as $n \to \infty$. The question of whether logarithmic regret is possible under unknown number of users is of interest.

Note the difference between the $\rho^{\text{EST}}$ policy in Algorithm 4 under unknown number of users with the $\rho^{\text{RAND}}$ policy with known number of users in Algorithm 3. The regret under $\rho^{\text{EST}}$ is $O(f(n) \log n)$ for any function $f(n) = \omega(1)$, while it is $O(\log n)$ under the $\rho^{\text{RAND}}$ policy. Hence, we are able to quantify the degradation of performance when the number of users is unknown.

## VI. Lower Bound & Effect of Number of Users

### A. Lower Bound For Distributed Learning & access

We have so far designed distributed learning and access policies with provable bounds on regret. We now discuss the relative performance of these policies, compared to the optimal learning and access policies. This is accomplished by noting a lower bound on regret for any *uniformly-good* policy, first derived in [4] for a general class of uniformly-good time-division policies. We restate the result below.

*Theorem 6 (Lower Bound [4]):* For any uniformly good distributed learning and access policy $\rho$, the sum regret in (2) satisfies

$$\liminf_{n \to \infty} \frac{R(n; \boldsymbol{\mu}, U, \rho)}{\log n} \geq \sum_{i \in U\text{-worst}} \sum_{j=1}^{U} \frac{\Delta(U^*, i)}{D(\mu_i, \mu_{j^*})}. \quad (30)$$

The lower bound derived in [9] for centralized learning and access holds for distributed learning and access considered here. But a better lower bound is obtained above by considering the distributed nature of learning. The lower bound for distributed policies is worse than the bound for the centralized policies in (11). This is because each user independently learns the channel availabilities $\boldsymbol{\mu}$ in a distributed policy, whereas sensing decisions from all the users are used for learning in a centralized policy.

Our distributed learning and access policy $\rho^{\text{RAND}}$ matches the lower bound on regret in (15) in the order $(\log n)$ but the scaling factors are different. It is not clear if the regret lower bound in (30) can be achieved by any policy under no explicit information exchange and is a topic for future investigation.

### B. Behavior with Number of Users

We have so far analyzed the sum regret under our policies under a fixed number of users $U$. We now analyze the behavior of regret growth as $U$ increases while keeping the number of channels $C > U$ fixed.

*Theorem 7 (Varying Number of Users):* When the number of channels $C$ is fixed and the number of users $U < C$ is varied, the sum regret under centralized learning and access $\rho^{\text{CENT}}$ in (12) decreases as $U$ increases while the upper bounds on the sum regret under $\rho^{\text{RAND}}$ in (15) monotonically increases with $U$.

*Proof:* The proof involves analysis of (12) and (15). To prove that the sum regret under centralized learning and access in (12) decreases with the number of users $U$, it suffices to show that for $i \in U$-worst channel,

$$\frac{\Delta(U^*, i)}{D(\mu_i, \mu_{U^*})}$$

decreases as $U$ increases. Note that $\mu(U^*)$ and $D(\mu_i, \mu_{U^*})$ decrease as $U$ increases. Hence, it suffices to show that

$$\frac{\mu(U^*)}{D(\mu_i, \mu_{U^*})}$$

decreases with $U$. This is true since its derivative with respect to $U$ is negative.

For the upper bound on regret under $\rho^{\text{RAND}}$ in (15), when $U$ is increased, the number of $U$-worst channels decreases and hence, the first term in (15) decreases. However, the second term consisting of collisions $M(n)$ increases to a far greater extent. □

Note that the above results are for the upper bound on regret under the $\rho^{\text{RAND}}$ policy and not the regret itself. Simulations in Section VII reveal that the actual regret also increases with $U$. Under the centralized scheme $\rho^{\text{CENT}}$, as $U$ increases, the number of $U$-worst channels decreases. Hence, the regret decreases, since there are less number of possibilities of making bad decisions. However, for distributed schemes although this effect exists, it is far outweighed by the increase in regret due to the increase in collisions among the $U$ users.

In contrast, the distributed lower bound in (30) displays anomalous behavior with $U$ since it fails to account for collisions among the users. Here, as $U$ increases there are two competing effects: a decrease in regret due to decrease in the number of $U$-worst channels and an increase in regret due to increase in the number of users visiting these $U$-worst channels.

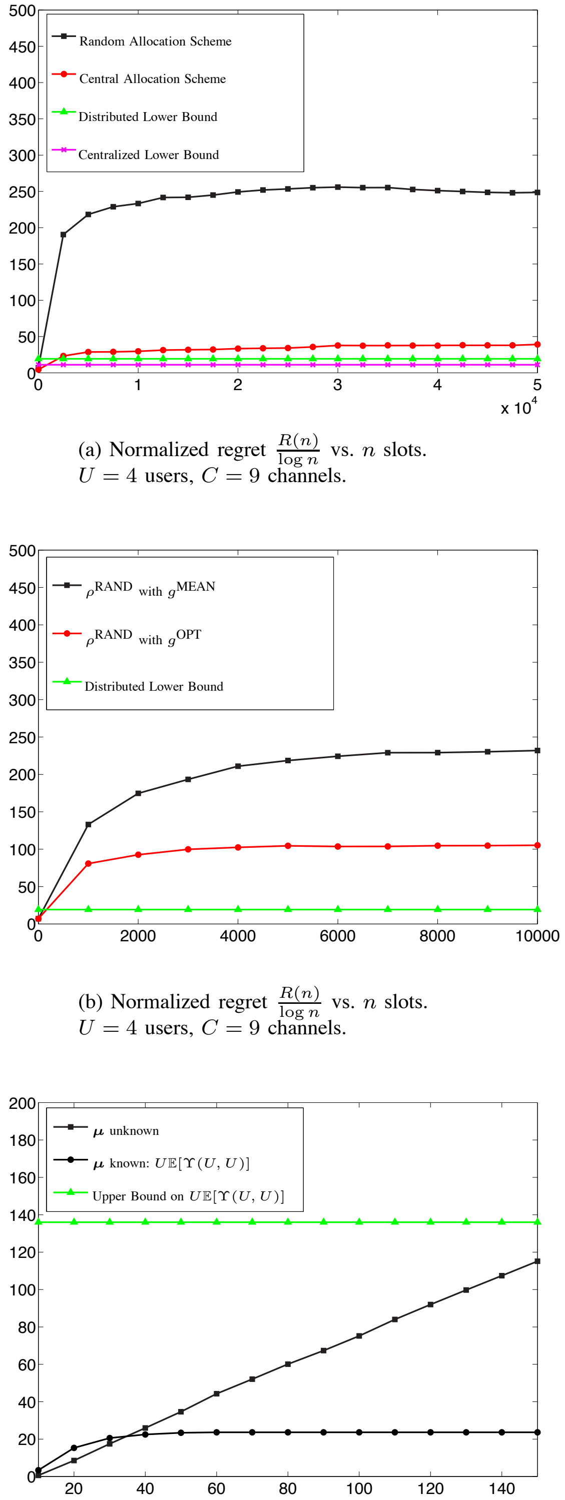
## VII. Numerical Results

We present simulations for the algorithms developed in the paper, varying the number of users and channels to verify the performance of the algorithms detailed earlier. We consider $C$= 9 channels (or a subset of them when the number of channels is varying) with probabilities of availability characterized by Bernoulli distributions with evenly spaced parameters ranging from 0.1 to 0.9.

*Comparison of Different Schemes:* Fig. 2a compares the regret under the centralized and random allocation schemes in a scenario with $U = 4$ cognitive users vying for access to the $C = 9$ channels. The theoretical lower bound for the regret in the centralized case from Theorem 2 and the distributed case from Theorem 6 are also plotted. The upper bounds on the random allocation scheme from Theorem 4 is not plotted here, since the bounds are loose especially as the number of users $U$ increases. Finding tight upper bounds is a subject of future study.

As expected, centralized allocation has the least regret. Another important observation is the gap between the lower bounds on the regret and the actual regret in both the distributed and the centralized cases. In the centralized scenario, this is simply due to using the $g^{\text{MEAN}}$ statistic in (34) instead of the optimal $g^{\text{OPT}}$ statistic in (5). However, in the distributed case, there is an additional gap since we do not account for collisions among the users. Hence, the schemes under consideration are $O(\log n)$ and achieve order optimality although they are not optimal in the scaling constant.

*Performance with Varying $U$ and $C$:* Fig. 3a explores the impact of increasing the number of secondary users $U$ on the regret experienced by the different policies with the number of channels $C$ fixed. With increasing $U$, the regret decreases for the centralized schemes and increases for the distributed schemes, as predicted in Theorem 7. The monotonic increase of regret under random allocation $\rho^{\text{RAND}}$ is a result of the increase in the collisions as $U$ increases. On the other hand, the monotonic decreasing behavior in the centralized case is because as the number of users increases, the number of $U$-

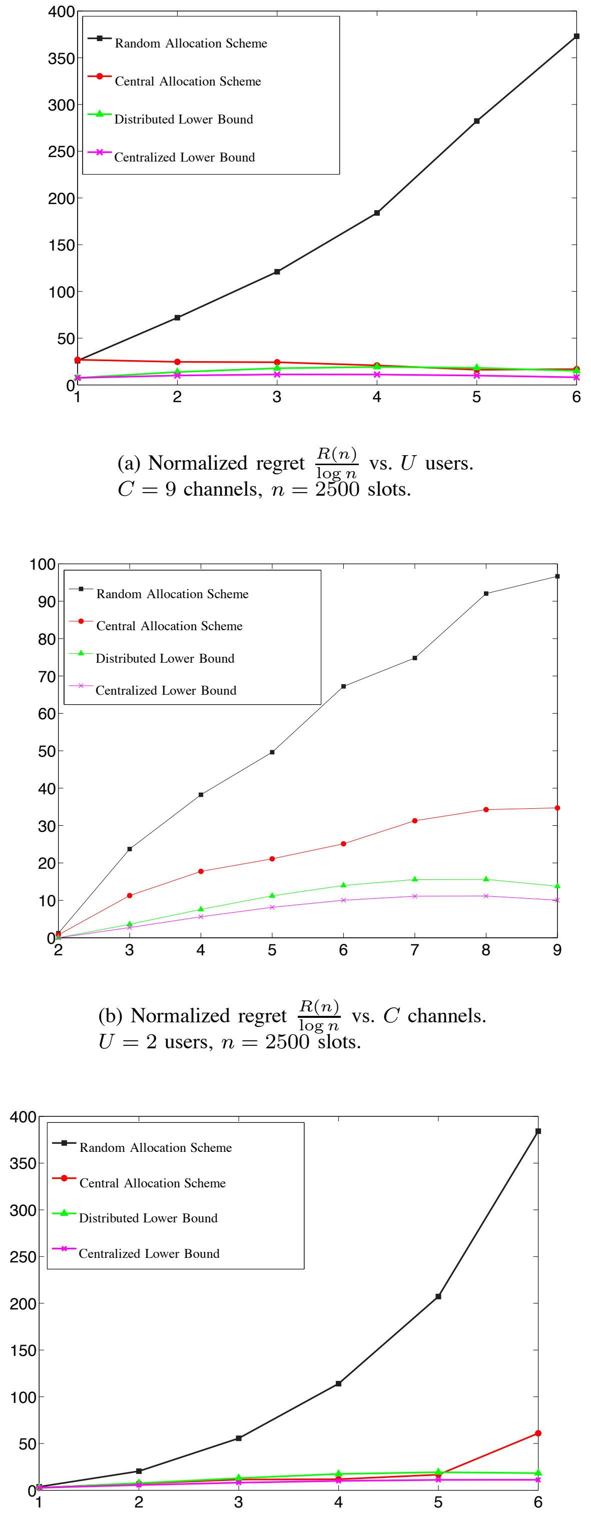


(a) Normalized regret $\frac{R(n)}{\log n}$ vs. $n$ slots. $U = 4$ users, $C = 9$ channels.

(b) Normalized regret $\frac{R(n)}{\log n}$ vs. $n$ slots. $U = 4$ users, $C = 9$ channels.

(c) No. of collisions $M(n)$ vs. $n$ slots. $U = 4$ users, $C = 9$ channels, $\rho^{\text{RAND}}$ policy.

Fig. 2. Simulation Results. Probability of Availability $\boldsymbol{\mu} = [0.1, 0.2, \ldots, 0.9]$.



(a) Normalized regret $\frac{R(n)}{\log n}$ vs. $U$ users. $C = 9$ channels, $n = 2500$ slots.

(b) Normalized regret $\frac{R(n)}{\log n}$ vs. $C$ channels. $U = 2$ users, $n = 2500$ slots.

(c) Normalized regret $\frac{R(n)}{\log n}$ vs. $U$ users. User-channel ratio $U/C = 0.5$, $n = 2500$ slots.

Fig. 3. Simulation Results. Probability of Availability $\boldsymbol{\mu} = [0.1, 0.2, \ldots, 0.9]$.

worst channels decreases resulting in lower regret. Also, the lower bound for the distributed case in (30) initially increases and then decreases with $U$. This is because as $U$ increases there are two competing effects: decrease in regret due to decrease in number of $U$-worst channels and increase in regret due to increase in number of users visiting these $U$-worst channels.

Fig. 3b evaluates the performance of the different algorithms as the number of channels $C$ is varied while fixing the number of users $U$. The probability of availability of each additional channel is set higher than those already present. Here, the regret monotonically increases with $C$ in all cases. When the number of channels increases along with the quality of the channels, the regret increases as a result of an increase in the number of $U$-worst channels as well as the increasing gap in quality between the $U$-best and $U$-worst channels.

Also, the situation where the ratio $U/C$ is fixed to be 0.5 and both the number of users and channels along with their quality increase is considered in Fig. 3c and we find that the regret grows in this case as well. Once again, this is in agreement with theory since the number of $U$-worst channels increases as $U$ and $C$ increase while keeping $U/C$ fixed.

*Collisions and Learning:* Fig. 2c verifies the logarithmic nature of the number of collisions under the random allocation scheme $\rho^{\text{RAND}}$. Additionally, we also plot the number of collisions under $\rho^{\text{RAND}}$ in the ideal scenario when the channel availability statistics $\boldsymbol{\mu}$ are known to see the effect of learning on the number of collisions. The low value of the number of collisions obtained under known channel parameters in the simulations is in agreement with theoretical predictions, analyzed as $U\mathbb{E}[\Upsilon(U, U)]$ in Lemma 2. As the number of slots $n$ increases, the gap between the number of collisions under the known and unknown parameters increases since the former converges to a finite constant while the latter grows as $O(\log n)$. The logarithmic behavior of the cumulative number of collisions can be inferred from Fig. 2a. However, the curve in Fig. 2c for the unknown parameter case appears linear in $n$ due to the small value of $n$.

*Difference between $g^{OPT}$ and $g^{MEAN}$:* Since the statistic $g^{\text{MEAN}}$ used in the schemes in this paper differs from the optimal statistic $g^{\text{OPT}}$ in (5), a simulation is done to compare the performance of the schemes under both the statistics. As expected, in Fig. 2b, the optimal scheme has better performance. However, the use of $g^{\text{MEAN}}$ enables us to provide finite-time bounds, as described earlier.

*Fairness:* One of the important features of $\rho^{\text{RAND}}$ is that it does not favor any one user over another. Each user has an equal chance of settling down in any one of the $U$-best channels. Fig. 4 evaluates the fairness characteristics of $\rho^{\text{RAND}}$. The simulation assumes $U = 4$ cognitive users vying for access to $C = 9$ channels. The graph depicts which user asymptotically gets the best channel over 1000 runs of the random allocation scheme. As can be seen, each user has approximately the same frequency of being allotted the best channel indicating that the random allocation scheme is indeed fair.

## VIII. Conclusion

In this paper, we proposed novel policies for distributed learning of channel availability statistics and channel access of multiple secondary users in a cognitive network. The first

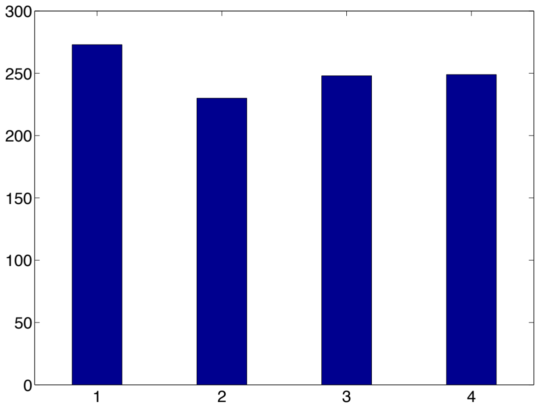


Fig. 4. Simulation Results. Probability of Availability $\boldsymbol{\mu} = [0.1, 0.2, \ldots, 0.9]$. No. of slots where user has best channel vs. user. $U = 4$, $C = 9$, $n = 2500$ slots, 1000 runs, $\rho^{\text{RAND}}$.

policy assumed that the number of secondary users in the network is known, while the second policy removed this requirement. We provide provable guarantees for our policies in terms of sum regret. By noting the lower bound on regret for any uniformly-good learning and access policy, we find that our first policy achieves order-optimal regret while our second policy is also nearly order optimal. Our analysis in this paper provides insights on incorporating learning and distributed medium access control in a practical cognitive network.

The results of this paper open up an interesting array of problems for future investigation. Simulations suggest that our lower and upper bounds are not tight in terms of the scaling constant and that better bounds are needed. Our assumptions of an i.i.d. model for primary user transmissions and perfect sensing at the secondary users need to be relaxed. Our policy allows for an unknown but fixed number of secondary users, and it is of interest to incorporate users dynamically entering and leaving the system. Moreover, our model ignores dynamic traffic at the secondary nodes and extension to a queueing-theoretic formulation is desirable. We consider the worst-case scenario that there is no information exchange among the secondary users. Extension to the case with limited information exchange is of interest. The proposed schemes are designed to be simple and tractable for mathematical analysis. It is desirable to obtain more robust schemes with additional design considerations for real-world deployment.

### *Acknowledgement*

The authors thank the guest editors and anonymous reviewers for valuable comments that vastly improved this paper, and for pointing out an error in Proposition 1. The authors thank Keqin Liu and Prof. Qing Zhao for extensive discussions, feedback on the proofs in an earlier version of the manuscript, and for sharing their simulation code. The authors also thank Prof. Lang Tong and Prof. Robert Kleinberg at Cornell, Prof. Bhaskar Krishnamachari at USC and Prof. John Tsitsiklis and Dr. Ishai Menache at MIT for helpful comments.

## APPENDIX

### *A. Proof of Theorem 2*

The result in (13) involves extending the results of [11, Thm. 1]. Define $T_i(n) := \sum_{j=1}^{U} T_{i,j}(n)$ as the number of times a channel $i$ is sensed in $n$ rounds for all users. We will show that

$$\mathbb{E}[T_i(n)] \le \sum_{k \in U\text{-best}} \left[ \frac{8 \log n}{\Delta(k^*, i)^2} + 1 + \frac{\pi^2}{3} \right], \quad \forall i \in U\text{-worst}. \tag{31}$$

We have

$$\begin{aligned} &\mathbb{P}[\text{Tx. in } i \text{ in } n^{\text{th}} \text{ slot}] = \mathbb{P}[g(U^*; n) \le g(i; n)], \\ &= \mathbb{P}[\mathcal{A}(i; n) \cap (g(U^*; n) \le g(i; n))] \\ &\quad + \mathbb{P}[\mathcal{A}^c(i; n) \cap (g(U^*; n) \le g(i; n))], \end{aligned}$$

where

$$\mathcal{A}(i; n) := \bigcup_{k \in U\text{-best}} (g(k; n) \le g(i; n))$$

is the event that at least one of the $U$-best channels has $g$-statistic less than $i$. Hence, from union bound we have

$$\mathbb{P}[\mathcal{A}(i; n)] \le \sum_{k \in U\text{-best}} \mathbb{P}[g(k; n) \le g(i; n)].$$

We have for $C > U$,

$$\mathbb{P}[\mathcal{A}^c(i; n) \cap (g(U^*; n) \le g(i; n))] = 0,$$

Hence,

$$\mathbb{P}[\text{Tx. in } i \text{ in } n^{\text{th}} \text{ round}] \le \sum_{k \in \text{-best}} \mathbb{P}[g(k; n) \le g(i; n)].$$

On the lines of [11, Thm. 1], we have $\forall k, i \;:\; k \text{ is } U\text{-best}, i \text{ is } U\text{-worst}$

$$\sum_{l=1}^{n} I[g(k; l) \le g(i; l)] \le \frac{8 \log n}{\Delta(k^*, i)^2} + 1 + \frac{\pi^2}{3}.$$

Hence, we have (31). For the bound on regret, we can break $R$ in (2) into two terms

$$\begin{aligned} R(n; \boldsymbol{\mu}, U, \rho^{\text{CENT}}) &= \sum_{i \in U\text{-worst}} \Big[ \frac{1}{U} \sum_{l=1}^{U} \Delta(l^*, i) \Big] \mathbb{E}[T_i(n)] \\ &\quad + \sum_{i \in U\text{-best}} \Big[ \frac{1}{U} \sum_{l=1}^{U} \Delta(l^*, i) \Big] \mathbb{E}[T_i(n)]. \end{aligned}$$

For the second term, we have

$$\begin{aligned} &\sum_{i \in U\text{-best}} \Big[ \frac{1}{U} \sum_{l=1}^{U} \Delta(l^*, i) \Big] \mathbb{E}[T_i(n)] \\ &\le \mathbb{E}[T^*(n)] \sum_{i \in U\text{-best}} \Big[ \frac{1}{U} \sum_{l=1}^{U} \Delta(l^*, i) \Big] = 0, \end{aligned}$$

where $T^*(n) := \max_{i \in U\text{-best}} T_i(n)$. Hence, we have the bound. □

### *B. Proof of Proposition 1*

For convenience, let $T_i(n) := \sum_{j=1}^{U} T_{i,j}(n)$, $V_i(n) := \sum_{j=1}^{U} V_{i,j}(n)$. Note that $\sum_{i=1}^{C} T_i(n) = nU$, since each user selects one channel for sensing in each slot and there are $U$ users. From (3),

$$\begin{aligned} R(n) =& n\sum_{i=1}^{U}\mu(i^*) - \sum_{i=1}^{C}\mu(i)\mathbb{E}[V_i(n)], \\ \leq& \sum_{i\in U\text{-best}} \mu(i)(n - \mathbb{E}[V_i(n)]) \\ \leq& \mu(1^*)(nU - \sum_{i\in U\text{-best}} \mathbb{E}[V_i(n)]) \qquad (32) \\ =& \mu(1^*)(\mathbb{E}[M(n)] + \sum_{i\in U\text{-worst}} \mathbb{E}[T_i(n)]), \qquad (33) \end{aligned}$$

where Eqn.(32) uses the fact that $V_i(n) \leq n$ since total number of sole occupancies in $n$ slots of channel $i$ is at most $n$, and Eqn.(33) uses the fact that $M(n) = \sum_{i\in U\text{-best}}(T_i(n) - V_i(n))$.

For the lower bound, since each user selects one channel for sensing in each slot, $\sum_{i=1}^{C}\sum_{j=1}^{U} T_{i,j}(n) = nU$. Now $T_{i,j}(n) \geq V_{i,j}(n)$.

$$\begin{aligned} R(n;\boldsymbol{\mu},U,\rho) \geq& \frac{1}{U}\left[\sum_{k=1}^{U}\sum_{j=1}^{U}\sum_{i=1}^{C}\Delta(U^*,i)\mathbb{E}[T_{i,j}(n)]\right], \\ \geq& \sum_{j=1}^{U}\sum_{i\in U\text{-worst}}\Delta(U^*,i)\mathbb{E}[T_{i,j}(n)]. \end{aligned}$$

□

### *C. Proof of Lemma 2*

Although, we could directly compute the time to absorption of the Markov chain, we give a simple bound $\mathbb{E}[\Upsilon(U,U)]$ by considering an i.i.d process over the same state space. We term this process as a genie-aided modification of random allocation scheme, since this can be realized as follows: in each slot, a genie checks if any collision occurred, in which case, a new random variable is drawn from Unif$(U)$ by all users. This is in contrast to the original random allocation scheme where a new random variable is drawn only when the particular user experiences a collision. Note that for $U = 2$ users, the two scenarios coincide.

For the genie-aided scheme, the expected number of slots to hit orthogonality is just the mean of the geometric distribution

$$\sum_{k=1}^{\infty} k(1-p)^k p = \frac{1-p}{p} < \infty, \tag{34}$$

where $p$ is the probability of having an orthogonal configuration in a slot. This is in fact the reciprocal of the number of *compositions* of $U$ [24, Thm. 5.1], given by

$$p = \binom{2U-1}{U}^{-1}. \tag{35}$$

The above expression is nothing but the reciprocal of number of ways $U$ identical balls (users) can be placed in $U$ different bins (channels): there are $2U - 1$ possible positions to form $U$ partitions of the balls.

Now for the random allocation scheme without the genie, any user not experiencing collision does *not* draw a new variable from Unif$(U)$. Hence, the number of possible configurations in any slot is lower than under genie-aided scheme. Since there is only one configuration satisfying orthogonality[9], the probability of orthogonality increases in the absence of the genie and is at least (35). Hence, the number of slots to reach orthogonality without the genie is at most (34). Since in any slot, at most $U$ collisions occur, (17) holds. □

### *D. Proof of Lemma 3*

Let $c_{n,m}:=\sqrt{\frac{2\log n}{m}}$.

**Case 1:** Consider $U = C = 2$ first. Let $\mathcal{A}(t,l):=\{g_j^{\text{MEAN}}(1^*;t-1) \leq g_j^{\text{MEAN}}(2^*;t-1), T_j'(t-1) \geq l\}$. On lines of [11, Thm. 1],

$$\begin{aligned} T'(n) \leq& l + \sum_{t=2}^{n} I[\mathcal{A}(t,l)], \\ \leq& l + \sum_{t=1}^{\infty}\sum_{m+h=l}^{t} I\left(\bar{X}_{1^*,j}(h) + c_{t,h} \leq \bar{X}_{2^*,j}(m) + c_{t,m}\right). \end{aligned}$$

The above event is implied by

$$\bar{X}_{1^*,j}(h) + c_{t,h} \leq \bar{X}_{2^*,j}(h) + c_{t,h+m}$$

since $c_{t,m} > c_{t,h+m}$.

The above event implies at least one of the following events and hence, we can use the union bound.

$$\begin{aligned} \bar{X}_{1^*,j}(h) \leq& \mu_{1^*} - c_{t,h}, \\ \bar{X}_{2^*,j}(m) \geq& \mu_{2^*} + c_{t,h+m}, \\ \mu_{1^*} <& \mu_{2^*} + 2c_{t,h+m}. \end{aligned}$$

From the Chernoff-Hoeffding bound,

$$\begin{aligned} \mathbb{P}[\bar{X}_{1^*,j}(t) \leq \mu_{1^*} - c_{t,h}] \leq& t^{-4}, \\ \mathbb{P}[\bar{X}_{2^*,j} \geq \mu_{2^*} + c_{t,h+m}] \leq& t^{-4}, \end{aligned}$$

and the event that $\mu_{1^*} < \mu_{2^*} + 2c_{t,h+m}$ implies that

$$h + m < \left\lceil \frac{8\log t}{\Delta^2_{1^*,2^*}} \right\rceil.$$

Since

$$\sum_{t=1}^{\infty}\sum_{m=1}^{t}\sum_{h=1}^{t} 2t^{-4} = \frac{\pi^2}{3},$$

,

$$\mathbb{E}[T'(n;U=C=2)] \leq \frac{8\log n}{\Delta^2_{1^*,2^*}} + 1 + \frac{\pi^2}{3}.$$

**Case 2:** For $\min(U,C) > 2$, we have

$$T'(n) \leq U\sum_{a=1}^{U}\sum_{b=a+1}^{C}\sum_{m=1}^{n} I(g_j^{\text{MEAN}}(a^*;m) < g_j^{\text{MEAN}}(b^*;m)),$$

where $a^*$ and $b^*$ represent channels with $a^{\text{th}}$ and $b^{\text{th}}$ highest availabilities. On lines of the result for $U = C = 2$, we can show that

$$\sum_{m=1}^{n} \mathbb{E}I[g_j^{\text{MEAN}}(a^*;m) < g_j^{\text{MEAN}}(b^*;m)] \leq \frac{8\log n}{\Delta^2_{a^*,b^*}} + 1 + \frac{\pi^2}{3}.$$

Hence, (18) holds. □

[9] since all users are identical for this analysis.

### E. Proof of Theorem 3

Define the good event as all users having correct top $U$-order of the $g$-statistics, given by

$$\mathcal{G}(n) := \bigcap_{j=1}^{U} \{\text{Top-}U \text{ entries of } \mathbf{g}_j(n) \text{ are same as in } \boldsymbol{\mu}\}.$$

The number of slots under the bad event is

$$\sum_{m=1}^{n} I[\mathcal{G}^c(m)] = T'(n),$$

by definition of $T'(n)$. In each slot, either a good or a bad event occurs. Let $\gamma$ be the total number of collisions in $U$-best channels between two bad events, i.e., under a run of good events. In this case, all the users have the correct top $U$-ranks of channels and hence,

$$\mathbb{E}[\gamma|\mathcal{G}(n)] \le U\mathbb{E}[\Upsilon(U,U)] < \infty,$$

where $\mathbb{E}[\Upsilon(U,U)]$ is given by (17). Hence, each transition from the bad to the good state results in at most $U\mathbb{E}[\Upsilon(U,U)]$ expected number of collisions in the $U$-best channels. The expected number of collisions under the bad event is at most $U\mathbb{E}[T'(n)]$. Hence, (19) holds. □

### F. Proof of Lemma 4

Under $\mathcal{C}(n;U)$, a $U$-worst channel is sensed only if it is mistaken to be a $U$-best channel. Hence, on lines of Lemma 1,

$$\mathbb{E}[T_{i,j}(n)|\mathcal{C}(n;U)] = O(\log n), \quad \forall i \in U\text{-worst}, j = 1,\dots,U.$$

For the number of collisions $M(n)$ in the $U$-best channels, there can be at most $U\sum_{k=1}^{a}\xi(n;k)$ collisions in the $U$-best channels where $a := \max_{j=1,\dots,U}\widehat{U}_j$ is the maximum estimate of number of users. Conditioned on $\mathcal{C}(n;U,)$, $a \le U$, and hence, we have (24). □

### G. Proof of Proposition 2

Define the good event as all users having correct top $U$-order, given by

$$\mathcal{G}(n) := \bigcap_{j=1}^{U} \{\text{Top-}U \text{ entries of } \mathbf{g}_j(n) \text{ are same as in } \boldsymbol{\mu}\}.$$

The number of slots under the bad event is

$$\sum_{m=1}^{n} I[\mathcal{G}^c(m)] = T'(n),$$

by definition of $T'(n)$. In each slot, either a good or a bad event occurs. Let $\gamma$ be the total number of collisions in $k$-best channels between two bad events, i.e., under a run of good events. In this case, all the users have the correct top $U$-ranks of channels and hence,

$$\gamma|\mathcal{G}(n) \overset{st}{\le} U\Upsilon(U,k),$$

The number of collisions under the bad event is at most $T'(n)$. Hence, (27) holds. □

### H. Proof of Lemma 6

We are interested in

$$\begin{aligned}\mathbb{P}[\mathcal{C}^c(n);U] &= \mathbb{P}[\cup_{j=1}^{U}\widehat{U}_j^{\text{EST}}(n) > U],\\ &= \mathbb{P}[\bigcup_{m=1}^{n}\bigcup_{j=1}^{U}\{\Phi_{U,j}(m) > \xi(n;U)\}],\\ &= \mathbb{P}[\max_{j=1,\dots,U}\Phi_{U,j}(n) > \xi(n;U)],\end{aligned}$$

where $\Phi$ is given by (20). For $U=1$, we have $\mathbb{P}[\mathcal{C}^c(n);U]=0$ since no collisions occur.

Using (27) in Proposition 2,

$$\begin{aligned}&\mathbb{P}[\max_{j=1}^{k}\Phi_{k,j}(n) > \xi(n;k)]\\ &\le \mathbb{P}[k\Upsilon(U,k)(T'(n)+1) > \xi(n;k)]\\ &\le \mathbb{P}[k(T'(n)+1) > \frac{\xi(n;k)}{\alpha_n}] + \mathbb{P}[\Upsilon(U,k) > \alpha_n]\\ &\le \frac{k\alpha_n(\mathbb{E}[T'(n)]+1)}{\xi(n;k)} + \mathbb{P}[\Upsilon(U,k) > \alpha_n], \qquad (36)\end{aligned}$$

using Markov inequality. By choosing $\alpha_n = \omega(1)$, the second term in (36), viz., $\mathbb{P}[\Upsilon(U,k) > \alpha_n] \to 0$ as $n \to \infty$, for $k \ge U$. For the first term, from (26) in Lemma 5, $\mathbb{E}[T'(n)] = O(\log n)$. Hence, by choosing $\alpha_n = o(\xi^*(n;k)/\log n)$, the first term decays to zero. Since $\xi^*(n;U) = \omega(\log n)$, we can choose $\alpha_n$ satisfying both the conditions. By letting $k=U$ in (36), we have $\mathbb{P}[\mathcal{C}^c(n);U] \to 0$ as $n \to \infty$, and (28) holds. □

## REFERENCES

[1] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic Spectrum Access with Multiple Users: Learning under Competition," in *Proc. IEEE INFOCOM*, San Deigo, USA, March 2010.

[2] Q. Zhao and B. Sadler, "A Survey of Dynamic Spectrum Access," *IEEE Signal Proc. Mag.*, vol. 24, no. 3, pp. 79–89, 2007.

[3] L. Kaelbling, M. Littman, and A. Moore, "Reinforcement learning: A Survey," *J. Artificial Intelligence research*, vol. 4, no. 237-285, pp. 102–138, 1996.

[4] K. Liu and Q. Zhao, "Decentralized Multi-Armed Bandit with Multiple Distributed Players," *submitted to IEEE Trans. Signal Process.*, Oct. 2009.

[5] A. Konrad, B. Zhao, A. Joseph, and R. Ludwig, "A Markov-based channel model algorithm for wireless networks," *Wireless Networks*, vol. 9, no. 3, pp. 189–199, 2003.

[6] S. Geirhofer, L. Tong, and B. Sadler, "Cognitive Medium Access: Constraining Interference Based on Experimental Models," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, p. 95, 2008.

[7] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part II: Markovian Rewards," *IEEE Tran. Autom. Control*, vol. 32, no. 11, pp. 977–982, 1987.

[8] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[9] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part I: IID rewards," *IEEE Tran. Autom. Control*, vol. 32, no. 11, pp. 968–976, 1987.

[10] R. Agrawal, "Sample Mean Based Index Policies with $O(\log n)$ Regret for the Multi-Armed Bandit Problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, 2002.

[12] Q. Zhao, Y. Chen, and A. Swami, "Cognitive MAC Protocols for Dynamic Spectrum Access," in *Cognitive Wireless Communication Networks*. Springer, 2007, pp. 271–301.

[13] K. Liu and Q. Zhao, "A restless bandit formulation of opportunistic access: Indexablity and index policy," *Submitted IEEE Trans. Inf. Theory, available on Arxiv*, 2008.

[14] H. Liu, B. Krishnamachari, and Q. Zhao, "Cooperation and Learning in multiuser opportunistic spectrum access," in *IEEE Intl. Conf. on Comm. (ICC)*, Beijing, China, May 2008.

[15] F. Fu and M. van der Schaar, "Learning to compete for resources in wireless stochastic games," *IEEE Trans. Veh. Technol.*, vol. 58, no. 4, pp. 1904–1919, May 2009.

[16] H. Gang, Z. Qian, and X. Ming, "Contention-Aware Spectrum Sensing and Access Algorithm of Cognitive Network," in *Intl. Conf. on Cognitive Radio Oriented Wireless Networks and Comm.*, Singapore, May 2008.

[17] H. Liu, L. Huang, B. Krishnamachari, and Q. Zhao, "A Negotiation Game for Multichannel Access in Cognitive Radio Networks," in *Proc. Intl. Conf. on Wireless Internet*, Las Vegas, NV, Nov. 2008.

[18] H. Li, "Multi-agent Q-Learning of Channel Selection in Multi-user Cognitive Radio Systems: A Two by Two Case," in *IEEE Conf. on System, Man and Cybernetics*, Istanbul, Turkey, 2009.

[19] M. Maskery, V. Krishnamurthy, and Q. Zhao, "Game Theoretic Learning and Pricing for Dynamic Spectrum Access in Cognitive Radio," in *Cognitive Wireless Comm. Networks*. Springer, 2007.

[20] R. Kleinberg, G. Piliouras, and E. Tardos, "Multiplicative Updates Outperform Generic No-regret Learning in Congestion Games," in *Proc. ACM Symp. on theory of computing (STOC)*, Bethesda, MD, May-June 2009.

[21] Y. Gai, B. Krishnamachari, and R. Jain, "Learning Multiuser Channel Allocations in Cognitive Radio Networks: A Combinatorial Multi-Armed Bandit Formulation," in *IEEE Symp. on Dynamic Spectrum Access Networks (DySPAN)*, Singapore, April 2010.

[22] K. Liu, Q. Zhao, and B. Krishnamachari, "Distributed learning under imperfect sensing in cognitive radio networks," in *Submitted to Proc. IEEE Asilomar Conf. on Signals, Sys., and Comp.*, Monterey, CA, Oct. 2010.

[23] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

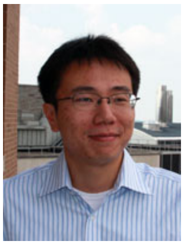
[24] M. Bona, *A Walk Through Combinatorics: An Introduction to Enumeration and Graph Theory*. World Scientific Pub. Co. Inc., 2006.

[25] M. Shaked and J. Shanthikumar, *Stochastic Orders*. Springer, 2007.

**Animashree Anandkumar** (S' 02, M' 09) received her B.Tech in Electrical Engineering from the Indian Institute of Technology Madras in 2004 and her Ph.D. in Electrical and Computer Engineering with a minor in Applied Math from Cornell University in 2009. She was a post-doctoral researcher at the Stochastic Systems Group at Massachusetts Institute of Technology between 2009 and 2010. She is currently an assistant professor at the Electrical Engineering and Computer Science Department at University of California, Irvine. She is also a member of the center for pervasive communications and computing at UCI.

She is the recipient of the 2008 IEEE Signal Processing Society (SPS) Young Author award for her paper co-authored with Lang Tong which appeared in the IEEE Transactions on Signal Processing. She is the recipient of the Fran Allen IBM Ph.D fellowship 2008-09, presented in conjunction with the IBM Ph.D. Fellowship Award. She received the Best Thesis Award 2009 by the ACM Sigmetrics society. Her research interests are in the area of statistical-signal processing, information theory and networking with a focus on distributed inference and learning of graphical models. She serves on the Technical Program Committee of IEEE INFOCOM 2011, ACM MOBIHOC 2011, IEEE ICC 2010, MILCOM 2010 and IEEE PIMRC 2010. She has served as a reviewer for IEEE Transactions on Signal Processing, IEEE Journal on Selected Areas in Communications, IEEE Transactions on Information Theory, IEEE Transactions on Wireless Communications and IEEE Signal Processing Letters.

**Nithin Michael** (S '06) received the B.S. degree summa cum laude from Drexel University, Philadelphia, PA, in 2008 in electrical engineering. He is currently pursuing the M.S./Ph.D. joint degree program in electrical engineering at Cornell University. Mr. Michael was the recipient of first Honors in ECE and the Highest Academic Achievement Award at Drexel University in 2008. At Cornell University, he has been awarded the Jacobs Fellowship in 2008 and 2010.



**Ao Tang** (S' 01 M' 07) received the B.E. in Electronics Engineering from Tsinghua University, Beijing, China, and the Ph.D. in Electrical Engineering with a minor in applied and computational mathematics from the California Institute of Technology, Pasadena, CA, in 1999 and 2006, respectively.

He is an Assistant Professor in the School of Electrical and Computer Engineering at Cornell University, where his current research is on control and optimization of networks including communication networks, power networks and on-chip networks. Dr. Tang was the recipient of the 2006 George B. Dantzig Best Dissertation Award, the 2007 Charles Wilts best dissertation Prize, and a 2009 IBM Faculty Award.

**Ananthram Swami** (F'08) received the B.Tech. degree from the Indian Institute of Technology (IIT), Bombay; the M.S. degree from Rice University, Houston, TX, and the Ph.D. degree from the University of Southern California (USC), Los Angeles, all in electrical engineering.

He has held positions with Unocal Corporation, USC, CS-3 and Malgudi Systems. He was a Statistical Consultant to the California Lottery, developed HOSAT, a MATLAB-based toolbox for non-Gaussian signal processing, and has held visiting faculty positions at INP, Toulouse, France. He is the ST for Network Science at the U.S. Army Research Laboratory. His work is in the broad areas of signal processing, wireless communications, sensor and mobile ad hoc networks. He is the co-editor of the book Wireless Sensor Networks: Signal Processing & Communications Perspectives (New York: Wiley, 2007).

Dr. Swami is a member of the IEEE Signal Processing Society's (SPS) Technical Committee on Sensor Array and Multichannel systems and serves on the Senior Editorial Board of the IEEE Journal on Selected Topics in Signal Processing. He was a tutorial speaker on Networking Cognitive Radios for Dynamic Spectrum Access at IEEE ICC 2010 and co-chair of IEEE SPAWC 2010.