Opportunistic Spectrum Access with Multiple Users: Learning under Competition

Animashree Anandkumar*, Nithin Michael†, and Ao Tang†

*EECS Dept., MIT, Cambridge, MA 02139, USA. Email: animakum@mit.edu
†ECE Dept., Cornell University, Ithaca, NY 14853, USA. Email: {nm373@,atang@ece.}cornell.edu

***Abstract*—The problem of cooperative allocation among multiple secondary users to maximize cognitive system throughput is considered. The channel availability statistics are initially unknown to the secondary users and are learnt via sensing samples. Two distributed learning and allocation schemes which maximize the cognitive system throughput or equivalently minimize the total regret in distributed learning and allocation are proposed. The first scheme assumes minimal prior information in terms of pre-allocated ranks for secondary users while the second scheme is fully distributed and assumes no such prior information. The two schemes have sum regret which is provably logarithmic in the number of sensing time slots. A lower bound is derived for any learning scheme which is asymptotically logarithmic in the number of slots. Hence, our schemes achieve asymptotic order optimality in terms of regret in distributed learning and allocation.**

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

I. Introduction

Cognitive radio is an area of extensive research in communications, signal processing and networking [1]. Typically, there are two kinds of transmitting nodes in a cognitive network viz., the primary users who have priority in accessing the spectrum and the secondary users who only opportunistically access the spectrum when the primary user is idle. The secondary users are cognitive and can sense the spectrum before transmission. They take advantage of the empty spaces in the spectrum and use them for transmissions, thus improving spectral efficiency. However, due to resource and hardware constraints, they can sense only a part of the spectrum at any given time. It is then crucial for the secondary users to make optimal decisions about which parts of the spectrum to sense at different times.

We consider a slotted system where each secondary user can only sense and access one orthogonal channel in each slot (see Fig.1). Here, the optimal channel selection strategy for a secondary user is based on the availability statistics of the orthogonal channels, i.e., the probability that the primary user is not transmitting in a particular channel. In practical scenarios, the channels' availability statistics are initially unknown to the secondary users and need to be estimated via sensing samples. This gives rise to a tradeoff between *exploration:* sensing new channels in the hope of obtaining better availability and *exploitation:* ensuring successful transmission in the current time slot. Additionally, when there are multiple secondary users, there is competition among the users to access the channel with best availability. Hence, the system throughput is reduced due to collisions among the secondary users under decentralized channel selection.

The above tradeoffs in distributed learning and allocation among multiple secondary users have not been sufficiently examined in the literature before. Our goal is to analyze these tradeoffs and propose schemes achieving provable optimal system throughput through only implicit cooperation among the secondary users. In particular, we answer the following questions. Is it possible to ensure that the secondary users correctly estimate the ranks of channels (with respect to their mean availabilities) through sensing samples? Is it possible for the secondary users to allocate themselves to orthogonal channels without any explicit information exchange or coordination? If indeed so, is it possible to converge to this ideal state while maximizing the total system throughput (i.e., total number of successful transmissions of the secondary users assuming equal rate for all users) or equivalently, minimizing the regret in distributed learning and allocation?

Our contributions are three fold. First, we propose two distributed learning and allocation schemes for arbitrary numbers of secondary users and channels. The first scheme assumes minimal prior information in terms of pre-allocated ranks for secondary users while the second scheme is fully distributed and assumes no such prior information. Second, we derive upper bounds on the sum regret under the proposed schemes. Third, we derive an asymptotic lower bound on the sum regret experienced by any distributed learning and allocation scheme satisfying a certain uniformly-good criterion. By comparing the lower and upper bounds, we conclude that our two proposed schemes are asymptotically efficient in terms of the order of number of slots.

It should be noted that the parallels between cognitive medium access and the multi-armed bandit problem has been explored in various works such as [2], [3]. However, these works either do not consider competing secondary users or assume known channel parameters. The classical results in [4] and [5] proposed schemes for multi-armed bandits with asymptotic logarithmic regret based on upper confidence bounds on the unknown channel availabilities. Since then, simpler schemes have been proposed in [6], [7] which compute a statistic for each arm (channel), henceforth referred to as g-statistics. These schemes are directly applicable if there is only one secondary user. The works in [8], [9] consider

The first author is 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.

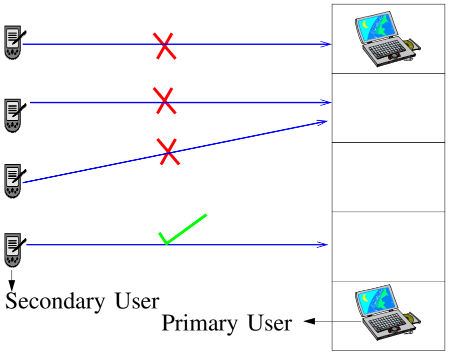


Fig. 1. Cognitive radio network with $U = 4$ secondary users and $C = 5$ channels. A secondary user is not allowed to transmit if the accessed channel is occupied by a primary user. If more than one secondary user transmits in the same free channel, then all the transmissions are unsuccessful.

centralized allocation schemes in contrast to distributed allocation here, [10] considers allocation through information exchange, [11] considers allocation under Q-learning for two users and channels where users can sense both the channels simultaneously, [12] considers learning through multiplicative updates in congestion games. In parallel with us, Liu and Zhao [13] developed a distributed learning and allocation policy for multiple secondary users when there is prior agreement among the users based on time division access and also first derived a lower bound on regret for any uniformly good time division policy. Our work differs in that we do not assume time division access and we consider two scenarios, one where there is prior agreement on the ranks of the secondary users and the second, where there is no prior agreement among the users. In [14], we consider extensions to the scenario where sensing is imperfect and has errors, and where the number of secondary users is unknown and needs to be additionally estimated.

The paper is organized as follows. In Section II, we model the cognitive network with multiple secondary users learning the unknown channel availability statistics through sensing and allocating themselves to orthogonal channels. We propose two schemes for distributed learning and allocation. The first one in Section III assumes the presence of initial common information among the secondary users in the form of pre-allocated ranks. On the other hand, the second one in Section IV is fully distributed and assumes no such prior information. Analysis shows that our schemes are asymptotically optimal. Simulations in Section V then provide further insights into the behavior of regret with varying number of users and channels. We conclude in Section VI and discuss some possible directions to extend the current work.

Remark: We note a shortcoming of our approach, viz., that the i.i.d. model for primary transmissions is idealistic and a Markovian model may be more appropriate in practice [15]. However, the i.i.d. model is a good approximation if the time slots for transmissions are sufficiently long and/or the primary traffic is highly bursty. Note that by i.i.d. primary transmission model, we do not mean the presence of a single primary user but rather, this model captures the overall statistical behavior of all the primary users in the system. Our analysis in this paper provides important engineering insights towards dealing with learning, competition, and cooperation in practical cognitive systems.

II. System Model & Formulation

Let $U \geq 1$ be the number of secondary users[1] and $C > U$ be the number of orthogonal channels available for slotted transmissions with a fixed slot width. In each channel i and slot j, the primary user transmits i.i.d. with probability $1-\mu_i > 0$,

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

and $W_i(j) \overset{i.i.d.}{\sim} B(\mu_i)$.

We refer to the U highest entries in $\boldsymbol{\mu}$ as the U-best channels and the rest as the U-worst channels. Let $\sigma(T;\boldsymbol{\mu})$ denote the index of the T^{th} highest entry in $\boldsymbol{\mu}$. For ease of notation, we abbreviate $T^*{:=}\sigma(T;\boldsymbol{\mu})$, $D(1,2){:=}D(B(\mu_1);B(\mu_2))$ the Kullback-Leibler distance between the Bernoulli distributions $B(\mu_1)$ and $B(\mu_2)$ [16] and $\Delta(1,2){:=}\mu_1-\mu_2$.

The availability statistics[2] $\boldsymbol{\mu}{:=}[\mu_1,\ldots,\mu_C]$ and $\boldsymbol{\mu} \in (0,1)^C$ are initially unknown to all the secondary users and are learnt *independently* over time through perfect sensing samples without any information exchange among the users. To obtain the sensing samples, at the beginning of each slot k, each secondary user $j \in U$ selects exactly one channel $i \in C$ for (perfect) sensing, and gets the value of $W_i(k)$, which is the indicator variable if the channel is free. Each secondary user j records all the sensing samples of each channel i in a vector $\mathbf{X}^k_{i,j}{:=}[X_{i,j}(1),\ldots,X_{i,j}(T_{i,j}(k))]^T$ where $T_{i,j}(k)$ denotes the number of slots where channel i is accessed in k slots (not necessarily being the sole occupant of that channel). Let $\mathbf{X}^k_j{:=}\{X_{1,j}(k),\ldots,X_{C,j}(k)\}$ denote the collection of sensed samples for user j in k slots in all the C channels. Denote the sample mean availability of channel i as sensed by user j as

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

Our policies for channel selection will be based on the sample mean availabilities.

Recall the constraint that the secondary users cannot transmit if the channel is occupied by the primary user. Upon finding an available channel, we assume that a secondary user always transmits, and that if two or more than secondary users transmit then none of the transmissions are successful. Note that if instead, secondary users employ back-off protocol such as CSMA-CA, collisions can be avoided, and our policies and their performance guarantees are applicable in this scenario as well. At the end of each slot k, each user j receives an acknowledgement $Z_j(k)$ on whether the transmission in the k^{th} slot was successfully received. Hence, in general, the policy

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

[2] For simplicity, we limit to cases where all the μ_i are distinct.

employed by user j in the $(k+1)$-th slot, given by $\rho(\mathbf{X}_j^k, \mathbf{Z}_j^k)$ can be based on all the previous sensing and feedback results.

In this paper, we limit to scenarios where all the secondary users employ the same policy ρ but undertake distributed learning and allocation without any information exchange. We are interested in designing policies ρ which maximize the expected number of successful transmissions of all 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 users and policy ρ, 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 of channel i.

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 reward 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^{th}-highest entry in $\boldsymbol{\mu}$. It is clear that $S^*(n;\boldsymbol{\mu},U) > S(n;\boldsymbol{\mu},U,\rho)$ for any policy ρ and finite n. We are interested in minimizing the *regret* in learning and allocation,

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

The regret represents loss in secondary transmissions due to learning of the unknown availability statistics as well as collisions due to distributed allocation.

III. Learning Under Pre-allocation

We first consider the case where the secondary users have information in the form of an allocation order, prior to learning and transmission. But there is no other information exchange among the secondary users. In the next section, we drop this assumption of initial information and consider fully distributed schemes for learning and allocation. The assumption of pre-allocated ranks simplifies design since there is no need for the users to cooperatively arrive at an allocation order during the process of learning. Instead, any policy under pre-allocation only needs to ensure that the users learn the channel availabilities accurately and occupy channels according to their pre-assigned ranks. Moreover, pre-allocated ranks can incorporate heterogeneous secondary users with different priority rankings.

Pre-allocated ranks can be either provided by a central authority (base station) or arrived in a distributed manner through information exchange and consensus. For instance, each user generates a uniform random variable in $[0,1]$ and exchanges it with all other users. The ranks are then based on the order of the variables.

We assume that each user j, among the U users, is assigned a unique allocation rank $w_j\in\{1,\ldots,U\}$ and that all the users are aware of their ranks and agree to implement them. Hence, if each user j had perfect knowledge of the channel availabilities $\boldsymbol{\mu}$, then he/she would always access the assigned channel w_j^*, thereby resulting in no regret. However, in the absence of such knowledge, user make errors in estimating the correct order of the channels leading to positive regret.

The regret for each user j under a policy based on pre-allocated ranks can arise due to two possibilities in a time slot, viz., user j selects channel other than the one assigned or other users visit the w_j^*-th channel assigned to user j (and hence, potentially collide with user j). Thus the regret for user j satisfies the upper bound

$$R_j(n;\boldsymbol{\mu},\rho)\leq\mu(w_j^*)[\sum_{i\neq w_j^*}T_{i,j}(n)+\sum_{k=\{1,\ldots,U\}\backslash j}T_{w_j^*,k}(n)], \tag{3}$$

where $T_{ij}(n)$ is the number of slots that channel i is selected by user j in n slots. In the above upper bound, we effectively assign a zero reward to users not transmitting in channels according to their pre-allocated ranks, since in the worst case it results in collisions. Hence, the sum regret of all the users in (2) satisfies

$$R(n;\boldsymbol{\mu},\rho)\leq\sum_{j=1}^{U}\mu(w_j^*)[\sum_{i\neq w_j^*}T_{i,j}(n)+\sum_{k=\{1,\ldots,U\}\backslash j}T_{w_j^*,k}(n)], \tag{4}$$

We now describe a scheme based on greedy learning which has logarithmic regret under pre-allocated ranks.

A. ρ^{PRE}: Greedy Distributed Learning Under Pre-allocation

We now propose a distributed learning policy, referred to as ρ^{PRE}, for the users to settle down in channels according to their pre-allocated ranks. The ρ^{PRE} scheme is given in Fig.1, and is a generalization of the greedy scheme for single secondary user in [7]. The ρ^{PRE} scheme is based on the intuition that there needs to be a lot of experimentation or selection of different channels in the beginning and eventually the user settles in the channel according to the pre-allocated rank.

In every time slot n, with probability ϵ_n, a channel is selected for sensing uniformly at random, and with probability $1-\epsilon_n$, the channel with the w_j^{th} highest sample mean is selected, where w_j is the pre-allocated rank for user j. If $\epsilon_n\equiv\epsilon>0$ is chosen, then in every time slot, there is a finite probability for user j to not select his pre-allocated channel, leading to a linear growth of regret. Hence, we need the randomization probability ϵ_n to decay with n. At the same time, we cannot have ϵ_n decaying too quickly with n, since in this case, the user may settle down in a wrong channel. In fact, it can be shown that if ϵ_t decays faster than $\frac{1}{n}$, then there is linear growth of regret. Hence, $\epsilon_n{:=}\min[\frac{\beta}{n},1]$ is chosen with an appropriate choice of β to ensure logarithmic regret.

We now show logarithmic growth of both per-user and sum regrets in (3) and (4) under ρ^{PRE} policy. Without loss of generality, $w_j=j$ for $j=1,\ldots,U$. Note that for user j targeting the j-best channel, the expected time spent in any

Algorithm 1 Policy $\rho^{\text{PRE}}(\overline{\mathbf{X}}_j(n), w_j, C, \epsilon_n)$ for each user j under C channels, sample mean channel occupancies $\overline{\mathbf{X}}_j(n)$ and the pre-allocated rank w_j.

```
1) Input: $\overline{\mathbf{X}}_j(n){:=}\{\overline{X}_{i,j}(T_{i,j}(n)), i = 1,\ldots,C\}$ : Sample mean availabilities for $j^{\text{th}}$ user after $n$ slots,
   $\epsilon_n{:=}\min[\frac{\beta}{n}, 1]$: probability of random selection,
   $\sigma(T;\overline{\mathbf{X}}_j(n))$: index of $T^{\text{th}}$ highest entry in $\overline{\mathbf{X}}_j(n)$.
   $1 \leq w_j \leq U$: pre-allocated unique rank for user $j$
2) For each n=1,2,...
   Select channel $\sigma(w_j;\overline{\mathbf{X}}_j(n))$ for sensing with probability $1-\epsilon_n$, select a channel uniformly at random with probability $\epsilon_n$, update sample mean $\overline{\mathbf{X}}_j(n)$
```

channel i which is not the assigned channel is

$$\mathbb{E}[T_{i,j}(n)] \leq \sum_{t=1}^{n-1} \frac{\epsilon_{t+1}}{C} + \sum_{t=1}^{n-1}(1-\epsilon_{t+1})\mathbb{P}[\mathcal{E}_{i,j}(n)], \;\; i \neq j^*, \tag{5}$$

where $\mathcal{E}_{i,j}(n)$ is the error event

$$\mathcal{E}_{i,j}(n){:=}\overline{X}_{j^*,j}(T_{j^*,j}(t)) \underset{i\in j\text{-worst}}{\overset{i\in j\text{-best}}{\lessgtr}} \overline{X}_{i,j}(T_{i,j}(t)). \tag{6}$$

Recall that $\epsilon_n{:=}\min[\frac{\beta}{n}, 1]$ is chosen. Intuitively, there is a tradeoff in choosing β (i.e., rate of decay of ϵ_t). For a small β, the regret due to randomly selecting a bad channel is small (the second term in (5)) while the regret due to selecting a wrong channel as the best channel (and eventually settling down in that channel) is large (first term in (5)).

We now show that the expected times spent in channels other than the assigned channel, given by (5) are logarithmic under ρ^{PRE} scheme, as long as β is chosen appropriately. This implies logarithmic growth of regret, from (4).

Define the minimum separation between $U{+}1$-best channels

$$\Delta_{\min}{:=} \min_{i,j\in(U+1)\text{-best}} |\Delta_{i,j}|.$$

For $\beta > \max[20, \frac{4}{\Delta^2_{\min}}]$, define a constant,

$$\delta{:=}\frac{\beta}{C}(\gamma+1) + \beta e^{-\frac{\beta\gamma}{10}}(\zeta(\frac{\beta}{10}-1) + (\gamma+1)\zeta(\frac{\beta}{10})) + \frac{4e^{-\frac{\beta\gamma\Delta^2_{j^*,i}}{4}}}{\Delta^2_{j^*,i}}\zeta(\frac{\beta\Delta^2_{j^*,i}}{4}), \tag{7}$$

where $\zeta(\cdot)$ is the Riemann zeta function and γ is the Euler-Mascheroni constant [17].

Theorem 1 (Logarithmic Regret Under ρ^{PRE}): Under ρ^{PRE} policy in Fig.1, the number of slots user j accesses a channel $i \neq j^*$ other than pre-allocated channel j^* satisfies,

$$\mathbb{E}T_{i,j}(n) \leq \frac{\beta}{C}\log n + \delta, \quad \forall i = 1,\ldots,C, i \neq j^*, \tag{8}$$

for $\epsilon_m{:=}\min[\frac{\beta}{m}, 1]$ and $\beta > \max[20, \frac{4}{\Delta^2_{\min}}]$. Hence, from (3) and (4), the regret for each user j and the sum regret are $O(\log n)$ under ρ^{PRE} policy.

On the other hand, if $\Delta^2_{\min} < \frac{1}{5}$ and an arbitrary $20 < \beta < \frac{4}{\Delta^2_{\min}}$ is chosen, the sum regret in (4) grows as

$$R(n;\boldsymbol{\mu},\rho^{\text{PRE}}) = O(n^a), \quad a{:=}1-\frac{\Delta^2_{\min}\beta}{4}. \tag{9}$$

Proof: See Appendix A. □

It is thus possible for users with pre-allocated ranks to attain their respective channels while experiencing only logarithmic regret under ρ^{PRE} policy. This is achieved by ensuring that there is enough experimentation in the beginning through random selection with probability ϵ_n.

Note that ρ^{PRE} policy requires knowledge of $\Delta_{\min}$ to choose β such that logarithmic regret is achieved, otherwise the regret growth is according to (9). Intuitively, this requirement cannot be removed under pre-allocation, since if two channels have the same mean availability, the users assigned to these two channels cannot distinguish the two channels and hence, have a finite probability of collisions in every transmission slot. In the subsequent section, we remove this requirement by using feedback (presence of collision) to adaptively allocate the users into orthogonal channels.

IV. Distributed Learning and Allocation

We now drop the assumption of prior agreement among the secondary users on the allocation order for accessing channels and design a fully distributed policy for learning and allocation without any information exchange. We show that logarithmic growth of regret is possible by adaptively changing the channel selection based on collisions experienced in the previous time slots.

To this end, we use a different method for estimating the rank of a channel, than the greedy scheme used in previous section where randomization with parameter ϵ_n is employed. We instead compute the so-called g-statistic [6], [7] for each channel, and the estimated ranks are based on the order of g-statistics of different channels.

Specifically, we use the sample-mean based g-statistic proposed in [7, Thm. 1], given by

$$g_j(i;n){:=}\overline{X}_{i,j}(T_{i,j}(n)) + \sqrt{\frac{2\log n}{T_{i,j}(n)}}. \tag{10}$$

For a single secondary user ($U=1$), selecting channel with the highest g-statistic in each time slot guarantees logarithmic regret. However, this policy results in a large number of collisions under multiple users since all the users end up targeting the same channel.

A. Bounds on Regret

We first provide a simple upper bound on the regret in (2) for distributed learning and allocation. This is later used to prove that our policy has logarithmic regret.

Proposition 1 (Upper Bound on Regret): The sum regret in (2) satisfies

$$UR(n) \leq \sum_{k=1}^{U}\mu(k^*)\left[\sum_{j=1}^{U}\sum_{i\in U\text{-worst}}\mathbb{E}[T_{i,j}(n)+M(n)]\right], \tag{11}$$

Algorithm 2 Policy $\rho^{\text{RAND}}(U, C, \mathbf{g}_j(n))$ for each user j under U users, C channels and statistic $\mathbf{g}_j(n)$.

```
1) Input: $(\mathbf{X}_j^n, \mathbf{Z}_j^n)$ : Channel occupancies, transmission
   selections and acknowledgement for $j^{\text{th}}$ user after $n$ slots,
   $\mathbf{g}_j(n)$: statistic based on $(\mathbf{X}_j^{n-1}, \mathbf{Z}_j^{n-1})$,
   $\sigma(T; \mathbf{g}_j(n))$: index of $T^{\text{th}}$ highest entry in $\mathbf{g}_j(n)$.
2) Init: Sense in each channel once, $n \leftarrow C$, $T \leftarrow 1$
3) Loop: $n \leftarrow n + 1$
4) Update $\mathbf{g}_j(n)$ based on $(\mathbf{X}_j^{n-1}, \mathbf{Z}_j^{n-1})$
   if $Z_j^{n-1} == 0$ then Draw a new $T \sim$ Unif$(U)$
   end if
   Select channel $i \leftarrow \sigma(T; \mathbf{g}_j(n))$ for sensing
5) $Z_j^n \leftarrow 1$ if successful transmission or unavailable
   channel, 0 o.w.
```

where $T_{i,j}(n)$ is the number of slots where user j selects channel i for sensing and $M(n)$ is the number of collisions faced by the users in the U-best channels in n slots.

Proof: Since a user can transmit at most once in each slot,

$$\sum_{j=1}^{U}\sum_{i=1}^{C} V_{i,j}(n) + Q(n) \leq nU$$

where $Q(n)$ is the total number of interferences faced by all the users during n slots. Hence, the regret in (2) satisfies

$$UR(n) \leq \sum_{k=1}^{U}\left[\sum_{j=1}^{U}\sum_{i=1}^{C}\Delta(U^*, i)\mathbb{E}[V_{i,j}(n)] + \mu(k^*)\mathbb{E}[Q(n)]\right].$$

Substitute the following in the above expression to obtain (11).

$$\sum_{j=1}^{U}\sum_{i \in U\text{-worst}}(T_{i,j}(n) - V_{i,j}(n)) = Q(n) - M(n).$$

□

Note that we can upper bound the sum regret by bounding $\mathbb{E}[T_{i,j}(n)]$ and $\mathbb{E}[M(n)]$ from (11). The first term $\mathbb{E}[T_{i,j}(n)]$ can be handled by manipulating the classical results of multi-armed bandits [6], [7]. On the other hand, quantifying the second term $\mathbb{E}[M(n)]$ is novel and requires new techniques.

B. ρ^{RAND} Policy for Distributed Learning & Allocation

We present the ρ^{RAND} policy in Fig.2. The scheme is based on the fact that the users need to randomize their channel selections to ensure that there is a finite probability of having an orthogonal allocation. Note that the users only need to randomize over the top-U entries of the g-statistic to ensure that the regret due to selection of a U-worst channel is likely avoided. However, if the users randomize in every slot, it results in linear growth of regret with the number of time slots since there is a finite probability of collisions in every slot and the users do not settle down in an orthogonal configuration asymptotically.

There is hence a need for careful design of randomization to ensure that the regret in decentralized learning and allocation in (2) is only logarithmic in the number of time slots. This is incorporated in ρ^{RAND} policy through adaptive randomization based on feedback, where each user randomizes only if there is a collision in the previous slot; otherwise, the previously generated random rank for the user is retained. It is easy to see that ρ^{RAND} policy ensures that the users are allocated orthogonally to the top U channels as the number of transmission slots goes to infinity. It is however not clear if we can guarantee that the regret in achieving this orthogonal configuration is logarithmic in the number of slots and we prove it below.

Recall that the sum regret satisfies the upper bound in (11) involving two terms, viz., the slots spent in the U-worst channels and the number of collisions in the U-best channels. The first term decouples among the different users and can be analyzed solely through the marginal distribution of the g-statistic of each individual user. On the other hand, the second term requires the joint distribution of the g-statistics of multiple users, which are correlated variables, and is intractable to analyze directly.

We first give a logarithmic upper bound on the number of slots spent by each user in any U-worst channel. Hence, the first term in the bound on regret in (11) is also logarithmic.

Lemma 1 (Time Spent in U-worst Channels): Under the ρ^{RAND} scheme in Fig.2, the total time spent by any user $j = 1, \ldots, U$, in any $i \in U$-worst channel is given by

$$\mathbb{E}[T_{i,j}(n)] \leq \mathbb{E}[L_{i,j}(n)] \leq \sum_{k=1}^{U}\Big[\frac{8\log n}{\Delta(i,k^*)^2} + 1 + \frac{\pi^2}{3}\Big], \quad (12)$$

where $L_{i,j}(n)$ is the number of times when the channel i appears in the top-U entries of the g-statistic, given by

$$L_{i,j}(n) := \bigcup_{a=1}^{U} I[g_j(a^*; k) \leq g_j(i; k)]. \quad (13)$$

Proof: On lines of [7, Thm. 1], we have the result. □

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 are only concerned about reaching an orthogonal configuration through randomization and there is no issue of learning.

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

$$\Pi(U) := \mathbb{E}[M(n); \rho^{\text{RAND}}(U, C, \boldsymbol{\mu})] \leq U\left[\binom{2U-1}{U} - 1\right]. (14)$$

Proof: The expected number of collisions can be bounded by the mean time to absorption in a finite state Markov chain. See Appendix B. □

So there are a finite number of expected collisions $\Pi(U)$ for the random allocation scheme 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, $\Pi(U)$ represents additional regret due to the absence of pre-allocated ranks.

We use the result of Lemma 2 for analyzing the number of collisions under distributed learning of $\boldsymbol{\mu}$ by showing that if the users are able to learn the correct order of the different channels with only logarithmic regret then only an additional finite expected number of collisions occur before reaching an orthogonal configuration. Below we prove that the times spent in the U-best channels with a wrong order of the channels is only logarithmic in the number of slots.

Lemma 3 (Wrong Order of g-statistics): Under the ρ^{RAND} scheme in Fig.2, the total time spent by any user j in the U-best channels, given the event that the user has the U-best channels as the top-U entries of its g-statistic but in an order different from the true order of the channel availabilities $\boldsymbol{\mu}$, is

$$\mathbb{E}[T'_j(n)] \leq (U-1)! \sum_{a=1}^{U} \sum_{b=a+1}^{U} \left[\frac{8\log n}{\Delta(a^*,b^*)^2} +1+ \frac{\pi^2}{3}\right].(15)$$

Proof: See Appendix C. □

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

Theorem 2 (Logarithmic Number of Collisions Under ρ^{RAND}): The expected number of collisions in the U-best channels under $\rho^{\text{RAND}}(U, C, \mathbf{g})$ scheme satisfies

$$\mathbb{E}[M(n)] \leq (\Pi(U)+U)\sum_{j=1}^{U}(\mathbb{E}[T'_j(n)]+\sum_{i\in U\text{-worst}}\mathbb{E}[L_{i,j}(n)]), \quad (16)$$

where $L_{i,j}(n)$ is given by (13). Hence, from (12), (15) and (14), $M(n) = O(\log n)$.

Proof: Define good state as all users having correct top U order of the channels,

$$\mathcal{G}(n):=I[\bigcap_{j=1}^{U} \text{Top } U \text{ entries of } \mathbf{g}_j(n) \text{ are in correct order}].$$

Hence, the bad state $\mathcal{G}^c(n)$ is the complementary event, which is the union of the events that a U-worst channel occurs in the top U entries of the $\mathbf{g}_j(n)$ statistic or that the top U entries of $\mathbf{g}_j(n)$ has the U-best channels but in the wrong order. Hence, the number of slots with bad events is at most

$$\sum_{k=1}^{n} I(\mathcal{G}^c(k)) \leq \sum_{j=1}^{U}[\sum_{i\in U_{\text{worst}}} L_{i,j}(k) + T'_j(k)].$$

In each slot, either a good or a bad event occurs. Let γ be the total number of collisions in U-best channels between two bad events, i.e., under a run of good states. In this case, all users have the correct top U order. Hence, by Lemma 2,

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

where $\Pi(U)$ is given by (14). Hence, each transition from bad to a good event results in at most $\Pi(U)$ number of expected collisions in the U-best channels. Under bad event in each slot, there are at most U collisions among U users. Hence, (16) holds. □

Hence, the expected number of collisions before the users settle in orthogonal channels is logarithmic. Combining this result with the result of Lemma 1, we immediately have one of the main results of this paper that the sum regret under distributed learning and allocation is logarithmic.

Theorem 3 (Logarithmic Regret Under ρ^{RAND}): The policy $\rho^{\text{RAND}}(U, C, \mathbf{g})$ in Fig.2 has $O(\log n)$ regret.

Proof: Substituting (16) and (12) in (11). □

Comparing the ρ^{RAND} policy with ρ^{PRE} policy, we note that both policies have logarithmic regret. However, ρ^{RAND} operates under less prior information, viz., without pre-allocated ranks and knowledge of minimum channel separation $\Delta_{\min}$, but requires more dynamic information, in the form of feedback at the end of each transmission slot. Also, while ρ^{RAND} uses the g-statistic, ρ^{PRE} uses the sample mean of the sensing results and ϵ_n-randomization. Simulations (see Section V) indicate that ρ^{PRE} performs worse than ρ^{RAND}. We believe that this happens because ρ^{RAND} is an adaptive algorithm that seeks to avoid collisions while ρ^{PRE} is primarily interested in allocating the users to the channels according to their pre-allocated ranks and does not incorporate feedback information.

C. Lower Bound For Distributed Learning & Allocation

The lower bound derived in [5] for centralized learning and allocation obviously holds for distributed learning and allocation considered here. But a better lower bound can be obtained by considering the distributed nature of learning.

We derive a lower bound for any uniformly good distributed policy ρ which enables users to ultimately settle down in orthogonal channels, defined as

$$\mathbb{E}_{\boldsymbol{\mu}}[n - T_{i,j}(n)] = o(n^{\alpha}), \quad \forall \alpha > 0, \boldsymbol{\mu} \in (0,1)^C. \quad (17)$$

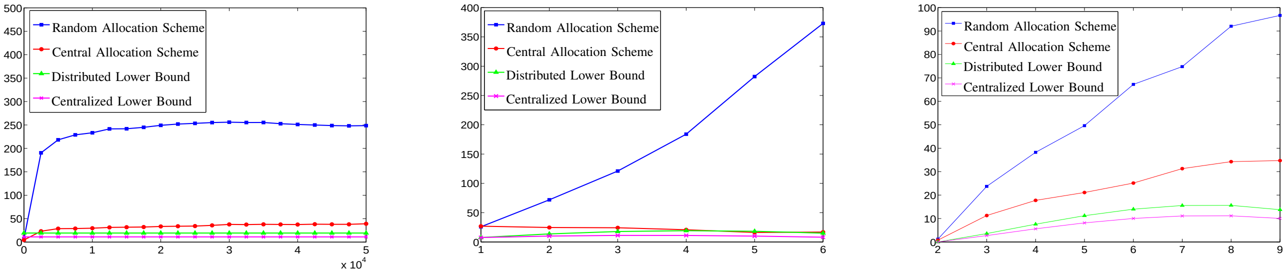
for some $i \in U$-best channel and some user j implying that user j accesses channel i most of the time. Note that the uniformly good requirement in (17) is stronger than the one in [5], where it suffices for the expected regret to be $o(n^{\alpha})$. It is easy to verify that ρ^{PRE} and ρ^{RAND} satisfy (17). The lower bound was first derived in [13] under a more general class of time division policies.

Theorem 4 (Lower Bound): For any uniformly good distributed learning and allocation policy ρ satisfying (17), the sum regret in (2) satisfies

$$\liminf_{n\to\infty} \frac{R(n;\boldsymbol{\mu},U,\rho)}{\log n} \geq \sum_{i\in U\text{-worst}} \sum_{j=1}^{U} \frac{\Delta(U^*,i)}{D(\mu_i,\mu_{j^*})}. \quad (18)$$

Proof: On the lines of [5], consider a channel i which is U-worst under a fixed parameter set $\boldsymbol{\mu}_0 = [\mu_1, \mu_2, \ldots, \mu_i, \ldots, \mu_C]$. Applying the change of measure argument, as in [5], consider $\boldsymbol{\mu}_1 = [\mu_1, \mu_2, \ldots, \lambda, \ldots, \mu_C]$, where μ_i is replaced with λ such that it becomes the k^{th} best channel under $\boldsymbol{\mu}_1$ for some $1 \leq k \leq U$. From the uniformly-good requirement in (17), we have

$$\mathbb{E}_{\boldsymbol{\mu}_1}[n - T_{i,j}(n)] = o(n^{\alpha}),$$



(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. U users. $C = 9$ channels, $n = 2500$ slots.

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

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

for some user j, and on lines of [5], we have for this particular user j,

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\mu}_0} \left[T_{i,j}(n) \geq \frac{(1-\epsilon)\log n}{D(\mu_j, \mu_{k^*})} \right] = 1.$$

Since the above expression holds for some combinations of user j and rank k, for each $1 \leq j \leq U$ and $1 \leq k \leq U$, we have

$$\lim_{n\to\infty} \mathbb{P}_{\boldsymbol{\mu}_0} \left[\sum_{j=1}^{U} T_{i,j}(n) \geq \sum_{k=1}^{U} \frac{(1-\epsilon)\log n}{D(\mu_i, \mu_{k^*})} \right] = 1.$$

We obtain (18) since

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

□

Hence, the lower bound for distributed policies under (17) is worse than the bound for the centralized policies in [5]. Intuitively, this is because each user independently learns the channel availabilities $\boldsymbol{\mu}$ in a distributed policy, whereas sensing samples from all the users is used for learning in a centralized policy.

Our distributed learning and allocation scheme matches the lower bound on regret in (11) in the order $(\log n)$ but the scaling factors are different. It is not clear if the lower bound on regret in (18) can be achieved by any policy under no explicit information exchange and is a topic of future investigation.

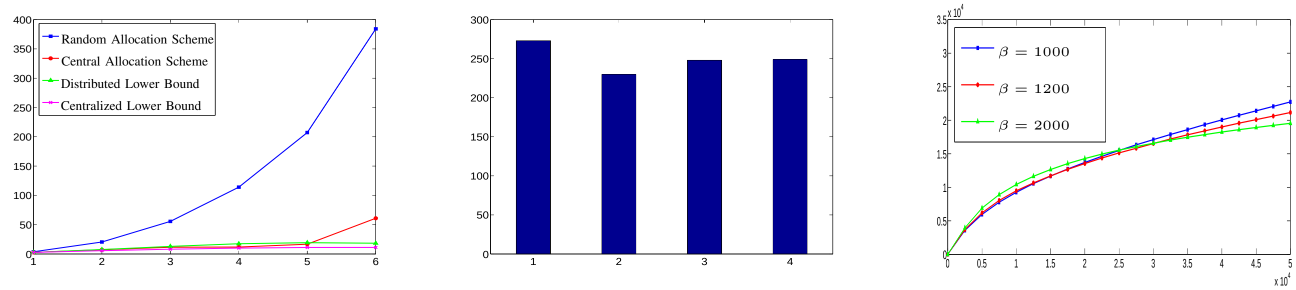
V. Numerical Results

We present simulations that vary the schemes and the number of users and channels to verify the performance of the algorithms detailed earlier. We consider C= 9 channels and U=4 cognitive users (or a subset of them when we vary their numbers) with probabilities of availability characterized by Bernoulli distributions with evenly spaced parameters ranging from 0.1 to 0.9. We discuss four topics.

Comparison of Different Schemes: Fig.2a compares the regret under the random allocation and the central allocation scheme from [5] (implemented using g-statistic in (10) for simplicity). Theoretical lower bounds for the regret in both the centralized case from [5, Thm 3.1] and the distributed case in Theorem 4 are also plotted. The upper bound on the regret for ρ^{RAND} is not plotted as it is very loose. As expected, centralized allocation has the least regret. The gap between the lower bound on the regret and the actual regret in the centralized scenario, is simply due to using the g-statistic instead of the optimal statistic described in [5]. However, in the distributed case, there is additional gap since we do not account for collisions among the users in deriving the lower bound. 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.2b explores the impact of increasing the number of secondary users U on the regret experienced by ρ^{RAND} and central allocation while fixing the number of channels C. The monotonic increase of regret under ρ^{RAND} is a result of the increase in collisions as U increases while the monotonic decreasing behavior in the centralized case is due to the decrease in the number of U-worst channels resulting in lower regret. These observations can also be rigorously proven to hold for any U. Also, the lower bound for the distributed case in (18) initially increases and then decreases with U 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.2c 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. The situation where the ratio $\frac{U}{C}$ is fixed to be 0.5 and both the



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

(b) No. of runs with top rank after n slots vs. user. $U = 4$, $C = 9$, $n = 2500$ slots, 1000 runs, ρ^{RAND}.

(c) Regret $R(n)$ vs. n slots (varying β). $U = 4$ users, $C = 9$ channels, ρ^{PRE}.

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

number of users and channels along with their quality increase is considered in Fig.3a. As the number of users increases the regret increases as the number of channels C and their quality are both increasing. Once again, this is in agreement with theory as the number of U-worst channels increases as U and C increase while keeping $\frac{U}{C}$ fixed.

Fairness: One of the important features of ρ^{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 top U channels. Fig.3b depicts the frequency with which each user ultimately gets the top channel over 1000 runs of ρ^{RAND}. As can be seen, each user has approximately the same frequency of being allotted the top channel indicating that the random allocation scheme is indeed fair.

ρ^{PRE} with varying β: Fig. 3c confirms that the regret under ρ^{PRE} is logarithmic. The regret increases with larger β at small number of time slots n, while the behavior is reversed for large n. Intuitively, this is because a large β causes the user to explore (i.e., randomize over) the channels more initially, but eventually yields a better estimate of the order of the channels with respect to their mean availabilities.

VI. Conclusion

In this paper, we design two schemes for distributed learning of channel availability statistics and cooperative allocation for secondary users. We prove that the schemes have logarithmic regret in the number of slots when compared to the ideal scenario with known availability statistics and centralized allocation. We also prove that all uniformly-good schemes suffer at least logarithmic asymptotic regret implying that our schemes have order-optimal regret.

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. We need to incorporate more realistic primary-user behavior by relaxing the i.i.d. assumption. Our schemes assume knowledge of the number of users and cooperation among them which may not hold always. We also plan to investigate the effect of limited information exchange among the secondary users.

Acknowledgement

The authors thank Prof. L. Tong and Prof. R. Kleinberg at Cornell, Prof. B. Krishnamachari at USC and Dr. I. Menache at MIT for comments, and K. Liu and Prof. Q. Zhao at UC Davis for extensive discussions, for pointing out an error in the lower bound and in simulations in the preprint and for providing their simulation code.

Appendix

A. *Proof of Theorem 1*

We drop the subscript corresponding to user j in $T_{i,j}$ and give the derivation only for $j = 1$. Similar analysis holds for other j. Define $2x_0(n) := \sum_{t=1}^{n} \epsilon_t$.

$$\mathbb{E}[T_i(n)] = \sum_{t=1}^{n} \frac{\epsilon_t}{C} + (1 - \frac{\epsilon_t}{C})\mathbb{P}[\overline{X}_{1^*}(T_{1^*}(t-1)) \leq \overline{X}_i(T_i(t-1))]$$

From union bound, $\mathbb{P}[\overline{X}_{1^*}(T_{1^*}(k)) \leq \overline{X}_i(T_i(k))] \leq$

$$\mathbb{P}[\overline{X}_{1^*}(T_{1^*}(k)) \leq \mu_{1^*} + \frac{\Delta_{1^*,i}}{2}] + \mathbb{P}[\overline{X}_i(T_i(k)) \geq \mu_i + \frac{\Delta_{1^*,i}}{2}].$$

Now $\mathbb{P}[\overline{X}_i(T_i(k)) \geq \mu_i + \frac{\Delta_{1^*,i}}{2}]$

$$\begin{aligned} &\leq \sum_{m=1}^{k} \mathbb{P}[T_i(k) = m] \exp[-\frac{\Delta_{1^*,i}^2 m}{2}] \\ &\leq \sum_{m=1}^{k} \mathbb{P}[T_i^R(k) \leq m] \exp[-\mathbb{P}[T_i^R(k) \leq m]] \\ &\leq \sum_{m=1}^{\lfloor x_0(k) \rfloor} \mathbb{P}[T_i^R(k) \leq m] + \frac{2}{\Delta_{1^*,i}^2} \exp[-\frac{\Delta_{1^*,i}^2 \lfloor x_0 \rfloor}{2}], \end{aligned}$$

since $\sum_{m=x+1}^{\infty} e^{-am} \leq \frac{e^{-ax}}{a}$ and $T_i^R(k)$ is the number of slots where channel i is chosen at random in k runs. From Bernstein's inequality,

$$\mathbb{P}[T_i^R(k) \leq x_0(k)] \leq \exp[-\frac{x_0(k)}{5}].$$

$$\begin{aligned} \mathbb{P}[\overline{X}_{1^*}(T_{1^*}(k)) \leq \overline{X}_i(T_i(k))] \leq& 2x_0(k) e^{-\frac{x_0(k)}{5}} \\ &+ \frac{4}{\Delta_{1^*,i}^2} \exp[-\frac{\Delta_{1^*,i}^2 \lfloor x_0(k) \rfloor}{2}], \end{aligned}$$

We have for harmonic series [17],

$$\beta[\log n+\gamma+\frac{1}{2(n+1)}]\le 2x_0(n)\le\beta[\log n+\gamma+\frac{1}{2n}].$$

$$\begin{aligned}\sum_{k=1}^{n}2x_0(k)e^{-\frac{x_0(k)}{5}}&\le\beta e^{-\frac{\beta\gamma}{10}}[\sum_{k=1}^{n}\frac{\log k}{k^{\frac{\beta}{10}}}+\sum_{k=1}^{n}\frac{\gamma+1}{k^{\frac{\beta}{10}}}],\\&\le\beta e^{-\frac{\beta\gamma}{10}}[\sum_{k=1}^{n}\frac{1}{k^{\frac{\beta}{10}-1}}+\sum_{k=1}^{n}\frac{\gamma+1}{k^{\frac{\beta}{10}}}]\end{aligned}$$

Hence, (5) holds. □

B. Proof of Lemma 2

Consider a "genie"-aided modification of random allocation scheme where 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. Note that in ρ^{RAND}, a new random variable is drawn only when the particular user experiences a collision. For the genie scheme, the mean hitting time for orthogonality is just the mean of the geometric distribution

$$\sum_{k=1}^{\infty}k(1-p)^k p=\frac{1-p}{p}<\infty, \tag{19}$$

where p is the probability of having an orthogonal configuration in a slot. This is given by the reciprocal of the number of *compositions* of U [18, Thm. 5.1], as

$$p=\binom{2U-1}{U}^{-1}. \tag{20}$$

For ρ^{RAND} scheme without the genie, any user not experiencing collision does not draw a new random variable from Unif(U). Hence, the number of possible configurations in any slot is lower than under genie-aided scheme, since the user retains his previous choice. Since there is only one configuration satisfying orthogonality and all users are identical for this analysis, the probability of orthogonality increases in the absence of the genie and is at least (20). Hence, the number of slots to reach orthogonality without the genie is at most (19). Since in any slot, at most U collisions occur, (14) holds. □

C. Proof of Lemma 3

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

$$\mathcal{A}(t,l){:=}\{g_j(1^*;t-1)\le g_j(2^*;t-1),T'_j(t-1)\ge l\}.$$

On lines of [7, Thm. 1],

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

Following the proof in [7, Thm. 1], we have

$$T'_j(n;U=2)\le\frac{8\log n}{\Delta^2_{1^*,2^*}}+1+\frac{\pi^2}{3}.$$

For $U>2$, we have to consider all $U!-1$ possible wrong orders of the top U entries of $\mathbf{g}(n)$. If we choose any two numbers $a>b$ from $1,\ldots,U$ and consider the number of times that the g-statistic has the wrong order

$$\sum_{n=1}^{\infty}I[g_j(a^*;n)<g_j(b^*;n)],$$

where a^* and b^* represent channels with a^{th} and b^{th} highest availabilities. On lines of the result for $U=2$, we have

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

The possible orderings of the g-statistic entries of the U-best channels where $g_j(a^*;n)<g_j(b^*;n)$ and no other g-statistic entry occurs in between is $(U-1)!$ implying (15). □

References

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

[2] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad hoc Networks: A POMDP Framework," *IEEE J. on Selected Areas in Comm.*, vol. 25, no. 3, pp. 589–600, 2007.

[3] K. Liu and Q. Zhao, "A restless bandit formulation of opportunistic access: Indexablity and index policy," in *Proc. of IEEE Conf. on Sensor, Mesh and Ad Hoc Comm. and Networks (SECON)*, 2008.

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

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

[6] 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.

[7] 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.

[8] F. Fu and M. van der Schaar, "Learning to compete for resources in wireless stochastic games," *Vehicular Tech., IEEE Tran. on*, vol. 58, no. 4, pp. 1904–1919, May 2009.

[9] 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.*, 2008, pp. 1–8.

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

[11] 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*, 2009.

[12] R. Kleinberg, G. Piliouras, and E. Tardos, "Multiplicative Updates Outperform Generic No-regret Learning in Congestion Games," in *Proc. of ACM Symp. on theory of computing (STOC)*, 2009, pp. 533–542.

[13] K. Liu and Q. Zhao, "Decentralized Multi-Armed Bandit with Multiple Distributed Players," *Arxiv 0910.2065v1*, 2009.

[14] A. Anandkumar, N. Michael, A. Tang, and A. Swami, "Distributed Learning and Allocation of Cognitive Users with Logarithmic Regret," *Under Submission*, Dec. 2009.

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

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

[17] D. Detemple and S. Wang, "Half Integer Approximations for the Partial Sums of the Harmonic Series," *Elsevier J. of Math. Analysis and App.*, vol. 160, no. 1, pp. 149–156, 1991.

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