Opportunistic Spectrum Access with Channel Switching Cost for Cognitive Radio Networks

Lin Chen*, Stefano Iellamo†, Marceau Coupechoux†

*Lab. de Recherche en Informatique (LRI)
University of Paris-Sud XI, Orsay, France
chen@lri.fr

†Department of Computer and Network Science
Telecom ParisTech - CNRS LTCI, Paris, France
{iellamo, coupecho}@enst.fr

***Abstract*—We study the spectrum access problem in cognitive networks consisting of multiple frequency channels, each characterized by a channel availability probability determined by the activity of the licensed primary users on the channel. The key challenge for the unlicensed secondary users to opportunistically access the unused spectrum of the primary users is to learn the channel availabilities and coordinate with others in order to choose the best channels for transmissions without collision in a distributed way. Moreover, due to the drastic cost of changing frequencies in current wireless devices in terms of delay, packet loss and protocol overhead, an efficient channel access policy should avoid frequently channel switching, unless necessarily. We address the spectrum access problem with channel switching cost by developing a block-based distributed channel access policy. Through mathematical analysis, we show that the proposed policy achieves logarithmic regret in spite of the channel switching cost. Extensive simulation studies show the performance gain of the proposed channel access policy.**

I. Introduction

Cognitive radio [1] has emerged in recent years as a promising paradigm to enable more efficient and spectrum utilization. Spectrum access models have been classified by [2] and include exclusive use (or operator sharing), commons and shared use of primary licensed spectrum. In the last model, unlicensed secondary users (SUs) are allowed to access the spectrum of licensed primary users (PUs) in an opportunistic way. In this case, a well-designed spectrum access policy is crucial to achieve efficient spectrum usage.

In this paper, we focus on the generic model of cognitive networks consisting of several frequency channels, each characterized by a channel availability probability determined by the activity of PUs on the channel. In such model, a challenging problem for SUs to opportunistically access the unused spectrum of PUs is to learn the channel availabilities and coordinate with other SUs in order to choose, in a distributed way, the best channels for transmissions without collision.

The model (with single SU) is closely related to the Multi-Armed Bandit (MAB) problem [3], a classical reinforcement learning problem where a SU should strike a balance between exploring the environment to find profitable channels and exploiting the best one as often as possible. Gittins developed an index policy in [4] that consists of selecting the arm with the highest index termed as Gittins index. This policy is shown to be optimal in the most general case. Lai and Robbins [5] and then Agrawal [6] studied the MAB problem by proposing policies based on the *upper confidence bounds* with logarithmic regret. Agrawal [7] proposed a block and frame based policy that achieves logarithmic regret for the MAB problem with switching cost, a variant of the original MAB problem. A detailed survey on the single player MAB problem with switching cost can be found in [8].

Despite of the similarity to the MAB problem, the spectrum access problem in cognitive radio networks has several specificities that make it especially challenging to tackle. One major specialty lies in the fact of multiple SUs that can cause collisions if they simultaneously access the same channel. Some recent work has investigated this issue, among which Anandkumar *et al.* proposed two algorithms with logarithmic regret, where the number of SUs is known [9] and unknown and estimated by each SU [10], Liu and Zhao developed a time-division fare share (TDFS) algorithm with convergence and logarithmic regret [11].

In our work, we investigate the channel access problem by taking into account the channel switching cost due to the change from one frequency band to another. Such channel switching cost is non-negligible in terms of delay (a radio reconfiguration may be needed), packet loss and protocol overhead (since SU transmitter and SU receiver have to coordinate). In such context, it is crucial to design channel access policies reluctant to switch channels unless necessary. The challenges of designing such channel access policies for cognitive radio networks are three-fold. Firstly, the uncertainty on the channel availability imposes a fundamental tradeoff between exploration, by probing new channels in order to learn it, and exploitation, by accessing the channel with the highest estimated availability probability based on current available information so as to achieve the best short-term reward. The second challenge stems from the competition among multiple SUs to access the best channel. Hence, the SUs should strike a balance between accessing the best channel and avoiding excessive collisions with others. Thirdly, the channel switching cost adds a new challenge to the design of efficient channel access policies. An efficient channel access policy should avoid frequent channel switching, unless necessary. To the best of our knowledge, if the first two challenges are attracting much attention in research community today, taking into account the switching cost in channel access policy design

*This work is supported by the project TEROPP (technologies for TERminal in OPPortunistic radio applications) funded by the French National Research Agency (ANR).

for cognitive radio networks has not yet been systematically addressed in the existing literature.

In this paper, we develop a channel access policy for cognitive radio networks with channel switching cost. Through mathematical analysis, we show that the proposed policy achieves logarithmic regret in spite of the channel switching cost. Extensive simulation studies show that the proposed policy outperforms the solutions in the literature.

The rest of the paper is structured as follows. Section II presents the system model and problem formulation. Section III describes the proposed block-based channel access policy and analyzes the system regret. In Section IV, extensive simulations are performed to evaluate the performance of the proposed policy. Section V concludes the paper.

II. System Model and Problem Formulation

We consider a cognitive radio network consisting of N independent channels $\mathcal{N} = \{1, \cdots, N\}$. There are M ($M \leq N$) secondary users (SUs) searching for idle channels temporarily unoccupied by the primary users (PUs) to transmit their own traffic in an opportunistic way. Both PUs and SUs in the network are operated in a synchronous time-slotted fashion. We assume that at each time slot, channel i is free with probability μ_i ($0 \leq \mu_i \leq 1$), i.e., in each channel i and time slot k, PUs transmit with an i.i.d. probability $1 - \mu_i$.[1] Without loss of generality, we assume throughout the paper that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_N$.

Having no initial knowledge on the channel statistics $\mu \triangleq \{\mu_i, i \in \mathcal{N}\}$, the SUs should learn independently in a distributed way over time through channel sensing samples without any information exchange. More specifically, at the beginning of each time slot k, each SU j chooses one channel $\phi_j(k)$ to sense and transmits its packet if the channel is unoccupied. Collisions occur when multiple SUs access the same channel.

The reward for a SU is 1 if the transmission is successful and 0 in case of collision. Moreover, we take into account the cost of channel switching, denoted as c, which corresponds to the normalized cost in terms of delay, packet loss and protocol overhead for SUs. The total reward of SU j after n slots, denoted as $U_j(n)$, can thus be calculated as

$$U_j(n) = \sum_{i=1}^{N} \mu_i \mathbb{E}[V_{i,j}(n)] - SW_j(n),$$

where $V_{i,j}(n)$ denotes the number of time slots during the n slots that SU j is the sole SU on channel i, $SW_j(n)$ is the channel switching cost of SU j during the n slots, shown as follows

$$SW_j(n) = c \sum_{i=1}^{N} \mathbb{E}[S_{i,j}(n)],$$

where $S_{i,j}(n)$ is the number of times SU j switches from another channel to channel i during the n slots, i.e.,

$$S_{i,j}(n) = \sum_{k=2}^{n} 1_{\{\phi_j(k-1) \neq i, \phi_j(k) = i\}}.$$

[1]Throughout the paper, we use i to refer to the channel index, k and n to refer to the time-slot index, j the index of the SUs.

where $\phi_j(k)$ denotes the channel chosen by SU j during slot k. The total reward for all SUs during the n slots, denoted as $U(n)$, are thus given as follows

$$U(n) = \sum_{j=1}^{M} U_j(n).$$

In the ideal case where μ is known a priori and a central scheduler orthogonally allocates the SUs to M channels with the highest values of μ_i (i.e., channel 1 to channel M), the expected global reward for all SUs after n slots, denoted as $U^*(n)$, is given by

$$U^*(n) = n \sum_{j=1}^{M} \mu_j.$$

Obviously $U^*(n)$ is the upper bound of $U(n)$ under any channel access policy ρ, i.e., $U^*(n) \geq U(n), \forall \rho$.

For a given channel access policy ρ, define the regret R_ρ as the expected reward loss with respect to the ideal case. More specifically, the regret represents the reward loss after n slots due to the lack of knowledge of the channel statistics, the competition among SUs and the channel switch. In our work, we seek to design asymptotically efficient channel access policies with sub-linear regret (more precisely, logarithmic regret, i.e., $R_\rho(n) \sim O(\log n)$ with $n \to \infty$). With such a policy, the time-averaged regret tends to zero.

III. Block-based Channel Access Policy

As argued in previous sections, the channel switching cost add a new element in the regret. Hence, in order to design asymptotically efficient channel access policies with logarithmic regret, we need to limit the frequency of channel switching at SUs. In this line of design, we develop the block-based channel access policy (BCA). The proposed channel access policy is inspired by the block allocation scheme in [7] on the single-player MAB problem with switching cost and adapted in our multiple-SU context. The main idea can be summarized as follows: we group time slots in blocks; at the beginning of each block, the SUs choose which channel to sense and stick to that channel for the whole block if no collision is experienced during the whole block; otherwise in case of collision, indicating more than one SU on the same channel, a channel randomization is performed such that each SU experiencing the collision switches randomly to another channel. The block structure is carefully constructed such that the total cost of channel switching and the loss due to collisions are both controlled to $O(\log n)$, resulting a global $O(\log n)$ regret.

In the following, we give a detailed description on the proposed channel access policy BCA, followed by a quantitative analysis on the resulting regret.

A. Description of the Block-based Channel Access Policy

In the proposed approach, time is divided into *frames* numbered $0, 1, 2, \cdots$. Each frame f is further subdivided into *blocks* numbered $0, 1, 2, \cdots$. All the blocks in a frame are of equal length. Let N_f denote the last time slot of frame f, b_f denote the block length in time slots of each block in

frame f. We choose the block lengths b_f and the frame length $N_f - N_{f-1}$ as follows:

$$\begin{cases} b_f = f \\ N_f - N_{f-1} = \left\lfloor \dfrac{2^{f^2} - 2^{(f-1)^2}}{f} \right\rfloor \end{cases},$$

where $\lfloor x \rfloor$ denotes the largest integer not more than x.

At the SU side, each SU j maintains two vectors $\mathbf{T_j}(n) \triangleq \{T_{i,j}(n), i \in \mathcal{N}\}$ and $\mathbf{X_j}(n) \triangleq \{X_{i,j}(n), i \in \mathcal{N}\}$, where $T_{i,j}(n)$ denotes the number of slots that SU j is on channel i during the past n slots, $X_{i,j}(n)$ denotes the number of slots that channel i is sensed unoccupied by PUs in the past n slots (note that SU j is not necessarily the sole occupant of that channel). With these two vectors, the mean availability of channel i $\overline{X}_{i,j}(n)$ sensed by SU j can be estimated as

$$\overline{X}_{i,j}(n) = \frac{X_{i,j}(n)}{T_{i,j}(n)}.$$

Each SU j then uses the sample-mean based g-statistic proposed in [12] to rank the availability of channels. The g-statistic is computed at each SU j as follows:

$$g_{i,j}(n) \triangleq \overline{X}_{i,j}(n) + \sqrt{\frac{2\log n}{T_{i,j}(n)}}.$$

Algorithm 1 Block-based distributed channel access policy

1: **Initialization:** $k \leftarrow M+1$, $I \leftarrow 1$
2: Sense each channel once
3: **loop**
4: Update channel statistics $\mathbf{T_j}(k)$, $\mathbf{X_j}(k)$, $\overline{\mathbf{X}}_\mathbf{j}(k)$, $\mathbf{g_j}(k)$
5: **if** k is the first slot of a block **then**
6: Sense the Ith best channel in terms of g statistic
7: **end if**
8: **if** collision **then**
9: Draw a new integer I randomly from $[1, M]$ and switch to the Ith best channel next slot
10: **end if**
11: $k \leftarrow k+1$:
12: **end loop**

The proposed block-based channel access policy BCA is detailed in Algorithm 1, which is executed at each SU j. Each SU j starts by sensing each channel once (line 2) to get the initial channel statistics $\overline{\mathbf{X}}_\mathbf{j}(\mathbf{0}) \triangleq \{\overline{X}_{i,j}(0), i \in \mathcal{N}\}$ and $\mathbf{g_j}(\mathbf{0}) \triangleq \{g_{i,j}(0), i \in \mathcal{N}\}$, which are then updated each slot (line 4). The SU j senses the Ith best channel (i.e., the channel with the Ith highest value of $g_{i,j}(k)$) at the beginning of each block and stays in that channel if no collision is experienced (line $5-6$). In case of collision, the channel randomization is performed such that the SU switches to the Ith best channel next slot (lines $8-10$).

B. Regret Analysis on the Block-based Channel Access Policy

In this subsection, we provide a quantitative analysis on the system regret of BCA. To this end, we first derive an upper bound of the regret and then show that the upper bound is logarithmic in time.

Theorem 1. *On the regret of the proposed block-based channel access policy, denoted as R_{BCA}, it holds that*

$$R_{BCA}(n) \leq \mu_1 \left[\sum_{i=M+1}^{N} \sum_{j=1}^{M} \mathbb{E}[T_{i,j}(n)] \right] + M\mathbb{E}[Y(n)] + \mathbb{E}[SW(n)],$$

where $Y(n)$ is the number of collisions on channels 1 to M during n slots, $SW(n)$ is the channel switching cost during n slots.

Proof: Let $V_i(n) \triangleq \sum_{j=1}^{M} V_{i,j}(n)$ denote the number of slots where there is exactly one SU on channel i. The global utility can be written as

$$U(n) = \sum_{i=1}^{N} \mu_i V_i(n) - SW(n).$$

Let $Y_i(n)$ denote the number of collisions on channel i during n slots, noticing that a collision involves at most M SUs, it holds that

$$V_i(n) + MY_i(n) \geq \sum_{j=1}^{M} T_{i,j}(n) \quad \forall i \in \mathcal{N}.$$

It then leads to

$$\begin{aligned} U(n) + SW(n) &= \sum_{i=1}^{N} \mu_i V_i(n) \geq \sum_{i=1}^{M} \mu_i V_i(n) \\ &\geq \sum_{j=1}^{M} \sum_{i=1}^{M} \mu_i T_{i,j}(n) - MY(n) = \sum_{i=1}^{M} \mu_i \sum_{j=1}^{M} T_{i,j}(n) - MY(n), \end{aligned}$$

where $Y(n) = \sum_{i=1}^{M} Y_i(n)$.

On the other hand, the optimal utility is

$$U^*(n) = n \sum_{i=1}^{M} \mu_i.$$

It follows that

$$\begin{aligned} R_{BCA}(n) &= U^*(n) - \mathbb{E}[U(n)] \\ &= \sum_{i=1}^{M} \mu_i \left[n - \sum_{j=1}^{M} \mathbb{E}[T_{i,j}(n)] \right] + M\mathbb{E}[Y(n)] + \mathbb{E}[SW(n)] \\ &\leq \mu_1 \left[Mn - \sum_{i=1}^{M} \sum_{j=1}^{M} \mathbb{E}[T_{i,j}(n)] \right] + M\mathbb{E}[Y(n)] + \mathbb{E}[SW(n)] \\ &= \mu_1 \left[\sum_{i=M+1}^{N} \sum_{j=1}^{M} \mathbb{E}[T_{i,j}(n)] \right] + M\mathbb{E}[Y(n)] + \mathbb{E}[SW(n)]. \end{aligned}$$

which concludes the proof. ∎

The upper bound of the regret derived in Theorem 1 is composed of three terms. The first term $\mu_1 \sum_{i=M+1}^{N} \sum_{j=1}^{M} \mathbb{E}[T_{i,j}(n)]$ is the sum of the number of slots that each SU chooses channel $M+1$ to N, multiplied by a constant μ_1. The second term is the utility loss due to collisions. The third term corresponds to the channel switching cost. In the subsequent analysis we establish the logarithmic bound for the regret, i.e., $R_{BCA}(n) \sim O(\log n)$. To this end, we show that the three terms in the regret are all logarithmic in n.

Lemma 1. *It holds that*

$$\sum_{i=M+1}^{N}\sum_{j=1}^{M}\mathbb{E}[T_{i,j}(n)] \sim O(\log n).$$

Proof: The proof follows Theorem 4.1 in [7]. ■

Lemma 2. *On the expected number of collisions, it holds that*

$$\mathbb{E}[Y(n)] \sim O(\log n).$$

Proof: The proof follows Theorem 3 in [10]. ■

Lemma 3. *On the channel switching cost, it holds that*

$$\mathbb{E}[SW(n)] \sim O(\log n).$$

Proof: A SU switches from the current channel to a new channel in the following two cases:

- at the beginning of a block;
- upon a collision.

Let $SW_1(n)$ and $SW_2(n)$ denote the channel switching cost for the first and the second cases, respectively, it holds that

$$SW(n) = SW_1(n) + SW_2(n).$$

It follows from Theorem 4.1 in [7] that

$$\mathbb{E}[SW_1(n)] \sim o(\log n).$$

On the other hand, we have

$$\mathbb{E}[SW_2(n)] \leq cM\mathbb{E}[Y(n)] \sim O(\log n)$$

in that a collision involves at most M SUs.

It then follows that

$$\mathbb{E}[SW(n)] \sim O(\log n),$$

which concludes the proof. ■

Combining the results of the above lemmas, we can establish the logarithmic bound on the total regret of BCA, as stated in the following theorem.

Theorem 2. *The block-based channel access policy BCA has logarithmic regret, i.e.,* $R_{BCA}(n) \sim O(\log n)$.

C. Discussion

The intrinsic idea behind the proposed block-based channel access policy is to limit the frequency of channel switching. To this end, each SU switches to another channel once every block if no collision is detected, i.e., every f slots at frame f, by deciding whether to switch in the first slot of each block. A randomization is performed at each SU if more than one SU accesses the same channel which leads to a collision.

The proposed policy is especially suited in the scenario where the channel switching cost is extremely significant in that it contains a conservative switching mechanism in which a SU stays at least for b_f slots in case of absence of collision in a channel to get more sensing samples.

The BCA policy can be synchronous or asynchronous. In the first case, all SUs have the same vision of the frame and block structure of time and thus take decisions simultaneously. In the asynchronous case, each SU has its own vision of the frame and block structure. The asynchronous version is more practical to implement in many applications in that no system synchronization are required among SUs. Note that our proof on the regret bound holds in both versions. The simulations presented in the next section show that the asynchronous version slightly outperforms the synchronous one.

IV. Simulation analysis

In this section we conduct extensive simulations to evaluate the performance of the proposed block-based channel access policy.

A. Simulation Setting

We simulate a cognitive radio network of $N = 9$ channels, whose availabilities are characterized by stationary Bernoulli distributions with evenly spaced parameters μ_i ranging from 0.1 to 0.9. A channel switching cost c (c is set to 1 if not specified) occurs when a SU switches from a channel to another in two adjacent time slots. Two versions of the proposed channel access policy BCA are investigated: the synchronous version (BCA-SYN) where the SUs are synchronous in terms of block and the asynchronous version (BCA-ASYN) where the SUs do not follow the synchronous block structure. We take the solution proposed in [10], termed as ρ^{RAND}, as the reference scheme to evaluate our proposed policy. In the numerical results presented in this section, each plot represents the average of 50 independent realizations.

B. Total Regret

As analyzed in previous sections, the learning process of a channel access policy produces a regret which depends on three factors: (a) time spent in the $(N-M)$ worst channels; (b) the number of collisions; and (c) channel switching cost. (a) and (b) represent the classical definition of the regret without switching cost (e.g. in [10]). The *total regret* considered in this paper also includes the third factor (c). Its logarithmic property is shown in Fig. 1 (BCA solid line) and in Fig. 4 for different switching costs and for both BCA-SYN and BCA-ASYN. We now analyze the different parts of the total regret.

C. Collisions and Time Spent in Worst Channels

In Fig. 1 the total regret (solid line) is decomposed into two components (in the figure, we only plot the regret of BCA-ASYN for the sake of clearness in that the curves of BCA-SYN are only slightly above those of BCA-ASYN). The dotted line in Fig. 1 plots the part of regret caused by factors (a) and (b). This part of regret is further decomposed in Fig. 2 into two parts: the regret caused by collisions and that by the time spent in worst channels.

We observe the fact that in ρ^{RAND} the SUs spend less time on the $(N-M)$ worst channels while in BCA the SUs experience less collisions. Globally, we can see that the effect of collisions has a major impact on the total regret compared with the time spent on the $(N-M)$ worst channels. This can be explained by the fact that a SU accessing a bad channel still obtains some reward, while a SU experiencing a collision gets nothing at all.

This phenomena is more clearly demonstrated in Fig. 3 where the depicted curves on the total regret slots show a

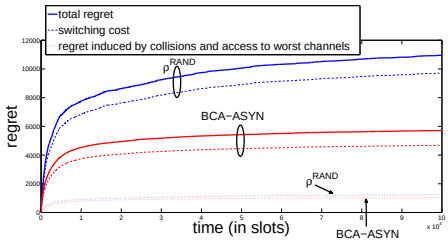


Fig. 1. Total regret, regret induced by collisions and time spent in worst channels, and switching cost

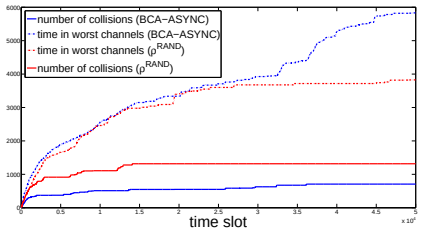


Fig. 2. Regret induced by collisions and time spent in worst channels

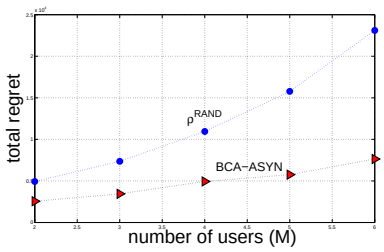
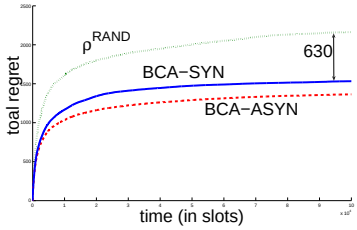
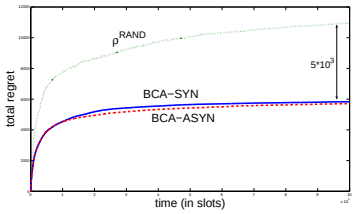


Fig. 3. Total regret per SU after 10^5 slots as a function of the number of users





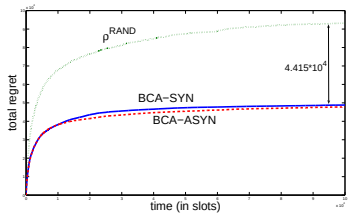


Fig. 4. Total regret for low (left), average (middle) and high (right) channel switching cost

monotonic increase w.r.t. the number of SUs after 10^5 time slots. As the number of SUs increases, the impact of collisions on the regret clearly outweighs that of time spent in the $(N-M)$ worst channels.

D. Channel Switching Cost

We now study the contribution of the channel switching cost in the total regret. We run three sets of simulations characterized by low, average and high channel switching costs with $c = 0.1, 1$ and 10, respectively (Fig. 4). We observe that BCA-ASYN performs slightly better than BCA-SYN in terms of regret. This can be explained by the fact that in BCA-SYN, the SUs starts new block and sense potential new channels in the synchronous way, leading to a more severe collision situation than BCA-ASYN where the blocks are asynchronous among SUs. The results also show that both BCA-SYN and BCA-ASYN outperform ρ^{RAND} with the regret gap increasing with the channel switching cost. This is due to the fact that our algorithm tries to limit the number of unnecessary switches and the switching cost becomes predominant when c increases.

E. Comparison with ρ^{RAND}

We now focus on a more systematic comparison between BCA-ASYN and ρ^{RAND}. We observe that our proposed policies outperforms ρ^{RAND} in the simulated scenarios in terms of system regret. The gap is stepped up as the channel switching cost becomes more predominant. Secondly, Fig. 3 shows the better system scalability of our scheme as compared to ρ^{RAND} the average individual regret in our proposed scheme increases only slightly as the system scales. Moreover, our scheme shows a comparable convergence speed w.r.t. ρ^{RAND} before the SUs are stabilized in their channels.

V. Conclusion

In this paper, we study the channel access problem in cognitive radio networks by taking into account the channel switching cost. We develop a channel access policy for cognitive radio networks with channel switching cost. Through mathematical analysis, we show that the proposed policy achieves logarithmic regret in spite of the channel switching cost. Extensive simulation studies show the performance gain of the proposed policy. An important direction of future work is to consider the more dynamic scenario with random arrival and departure of SUs and investigate efficient channel access policies in that case.

References

[1] S. Haykin. Cognitive radio: Brain-empowered wireless communications. *IEEE J. on Selected Areas in Communications*, 23(2):201–220, 2005.

[2] M. Buddhikot. Understanding dynamic spectrum access: models, taxonomy and challenges. In *Proc. IEEE DySPAN*, April 2007.

[3] A. Mahajan and D. Teneketzis. Multi-armed Bandit Problems. *Foundations and Applications of Sensor Management, Springer-Verlag*, 2007.

[4] J. C. Gittins. Multi-armed Bandit Allocation Indices. *Wiley-Interscience Series in Systems and Optimization, John Wiley & Sons*, 1989.

[5] T. L. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Probability*, 6(1), 1985.

[6] R. Agrawal. Sample Mean Based Index Policies with O(logn) Regret for the Multi-Armed Bandit Problem. *Advances in Applied Probability*, 27(4), Dec. 1995.

[7] R. Agrawal, D. Teneketzis, and V. Anantharam. Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching. *IEEE Trans. on Automatic Control*, 33(10):899–906, 1988.

[8] T. Jun. A Survey on the Bandit Problem with Switching Costs. *De Economist*, 152(4), Dec. 2004.

[9] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami. Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret. *IEEE J. on Selected Areas in Communications (to appear)*.

[10] A. Anandkumar, N. Michael, and A. Tang. Opportunistic spectrum access with multiple users: Learning under competition. In *Proc. IEEE Infocom*, San Diego, CA, 2010.

[11] K. Liu and Q. Zhao. Distributed Learning in Multi-Armed Bandit with Multiple Players. *Arxiv 0910.2065*, 2009.

[12] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.