Upper Confidence Bound Algorithm for Opportunistic Spectrum Access with Sensing Errors

Wassim Jouini, Christophe Moy, and Jacques Palicot
SUPELEC, SCEE/IETR
Avenue de la Boulaie, CS 47601, 35576 Cesson Sévigné Cedex, France
Email: {wassim.jouini, christophe.moy, jacques.palicot}@supelec.fr

***Abstract*—In this paper we consider the problem of exploiting spectrum resources within the Opportunistic Spectrum Access context. We mainly focus on the case where one secondary user (SU) probes a pool of possibly available channels dedicated to a primary network. The SU is assumed to have imperfect sensing abilities. We, first, model the problem as a Multi-Armed Bandit problem with sensing errors. Then, we suggest to analyze the performances of the well known Upper Confidence Bound algorithm UCB_1 within this framework, and show that we still can obtain an order optimal channel selection behavior. Finally we compare these results to those obtained in the case of perfect sensing. Simulation results are provided to support the suggested approach.**

***Index Terms*—Cognitive Radio, Opportunistic Spectrum Access, Upper Confidence Bound Algorithm, Imperfect Sensing, Sensing Errors.**

I. Introduction

A. Opportunistic Spectrum Access and Cognitive Radio

The concept of Opportunistic Spectrum Access (OSA) has been suggested as a promising approach to exploit frequency band resources efficiently, taking advantage of the various available opportunities. As a matter of fact, during the last century, most of the meaningful spectrum resources were licensed to emerging wireless applications, where the static frequency allocation policy combined with a growing number of spectrum demanding services led to a spectrum scarcity. However, several measurements conducted in the United-States [1], first, and then in numerous other countries, showed a chronic underutilization of the frequency band resources, revealing substantial communication opportunities.

The general concept of OSA defines two types of users: primary users (PUs) and secondary users (SUs). PUs access spectrum resources dedicated to the services provided to them, while SUs refer to a pool of users willing to exploit the spectrum resources unoccupied by PUs at a particular time in a particular geographical area. Since SUs need to access the spectrum while ensuring minimum interference with PUs and without *a priori* knowledge on the behavior of PUs, cognitive abilities (sensing its environment[1], processing the gathered information, and finally adapting its behavior depending on the environment constraints and users' expectations) are required to enable the coexistence of SUs and PUs. To fulfill these requirements, Cognitive Radio (CR) has been suggested as a promising technology to enable the OSA concept [1] [2].

However several challenges arise to achieve an efficient spectrum use relying on CRs. On the one hand, an accurate and reliable detection of PUs activity, and on the other hand, a smart behavior enabling SUs' to adapt their channel selection and access policies to PUs' band occupation pattern. Proposing such algorithms to answer these challenges has been, in the last years, the center of a lot of attention [3] [4].

B. Multi-Armed Bandit models for Opportunistic Spectrum Access

Recently, the Cognitive Radio community gave a particular attention to the Multi-Armed Bandit (MAB) paradigm. In a nutshell, based on the analogy with the one-armed bandit (also known as slot machine), it models a gambler sequentially pulling one of the several levers (multi-armed bandit) on the gambling machine. Every time a lever is pulled, it provides the gambler with a random income usually referred to as reward. Although we assume that the gambler has no *a priori* information on the rewards' stochastic distributions, he aims at maximizing his cumulated income through iterative pulls. In the OSA framework, the SU is modeled as the gambler while the frequency bands represent the levers. The gambler faces at each trial a trade-off between pulling the lever with the highest estimated payoff (known as *exploitation* phase) and pulling another lever to acquire information about its expected payoff (known as *exploration* phase). We usually refer to this trade-off as the *exploration-exploitation* dilemma.

Thus, several algorithms were borrowed from the machine leaning community [5] [6] [7] and suggested as possible solutions to learn selecting and accessing the most available channels [8] [9] [10]. These algorithms, however assume perfect sensing. Namely, they assume that the SU can acquire an errorless knowledge on the state of the probed channel {idle, busy}. Under these assumptions, secondary users can maximize their cumulated income (channel access and/or throughput), in expectation, while completely avoiding harmful packet collisions with primary users.

The purpose of this paper is to introduce a more realistic scenario. Thus we consider the OSA problem as a MAB problem with sensing errors. First, the network model is

[1]The term *environment* is used in a broad sense refering to any source of information that could improve the CR's behavior (QoS, throughput, etc).

CROWNCOM 2011, June 01-03, Osaka, Japan

DOI 10.4108/icst.crowncom.2011.245851

detailed in Section II. Then we introduce the well known UCB_1 algorithm in Section III. In order to understand the behavior of this algorithm within this framework, Section III-B provides a theoretical analysis of its performances. Section IV reports simulation results that illustrate the analysis conducted in Section III and, finally, Section V concludes.

II. Network model

A. Framework

We consider the case of one secondary user willing to opportunistically exploit the available spectrum in its vicinity. The spectrum of interest is licensed to a primary network providing K independent but non-identical channels. We denote by $k \in \{1, \cdots, K\}$ the k^{th} most available channel. Every channel k can appear, when observed, in one of these two possible states {idle, busy}. In the rest of the paper, we associate the numerical value 0 to a busy channel and 1 to an idle channel. The temporal occupancy pattern of every channel k is thus supposed to follow an unknown Bernoulli distribution θ_k. Moreover, the distributions $\Theta = \{\theta_1, \theta_2, \cdots, \theta_K\}$ are assumed to be stationary.

In this paper we tackle the particular case where PUs are assumed to be synchronous and the time $t = 0, 1, 2 \cdots$, is divided into slots. We denote by $\mathbf{S}_t$ the channels' state at the slot number t: $\mathbf{S}_t = \{S_{1,t}, \cdots, S_{K,t}\} \in \{0,1\}^K$. For all $t \in \mathbb{N}$, the numerical value $S_{k,t}$ is assumed to be an independent random realization of the stationary distributions $\theta_k \in \Theta$. Moreover, the realizations $\{S_{k,t}\}_{t\in\mathbb{N}}$ drawn from a given distribution θ_k are assumed to be independent and identically distributed. The expected availability of a channel is characterized by its probability of being idle. Thus, we define the availability μ_k of a channel k, for all t as:

$$\mu_k \triangleq \mathbb{E}[\theta_k] = \mathbb{P}(\text{channel } k \text{ is free}) = \mathbb{P}(S_{k,t} = 1) \quad (1)$$

where $\mu_1 > \mu_2 \geq \cdots \geq \mu_k \geq \cdots \geq \mu_K$ without loss of generality.

Let us refer to the decision making engine of the CR equipment as Cognitive Agent (CA). The CA can be seen as the brain of the CR device. At every slot number t, the SU has to choose a channel to sense. To do so, the CA relies on the outcome of past trials. We denote by i_t the gathered information until the slot t. We assume that the SU can only sense one channel per slot. Thus selecting a channel can be seen as an action $a_t \in \mathcal{A}$ where the set of possible actions $\mathcal{A} = \{1, 2, \ldots, K\}$ refers to the set of channels available.

Thus, we can model the CA as a policy π that maps for all $t \in \mathbb{N}$, the information i_t to an action a_t:

$$a_t = \pi(i_t) \quad (2)$$

The outcome of the sensing process is denoted by the binary random variable $X_t \in \{0, 1\}$. In the case of perfect sensing, $X_t = S_{a_t,t}$, where a_t refers to the channel selected at the slot number t. However since we assumed that sensing errors can occur, the value of X_t depends on the receiver operating characteristic (ROC). The ROC defines the accuracy and the reliability of a sensor through the measure of two types of errors: on the one hand, detecting a PU on the channel when it is free usually referred to as *false alarm*. On the other hand, assuming the channel free when a PU is occupying it usually referred to as *miss detection*. Let us denote by ϵ and δ, respectively the probability of false alarm, and the probability of miss detection characterizing the CR equipment:

$$\begin{cases} \epsilon = \mathbb{P}_{fa} = \mathbb{P}\left(X_t = 0 | S_{a_t,t} = 1\right) \\ \delta = \mathbb{P}_{md} = \mathbb{P}\left(X_t = 1 | S_{a_t,t} = 0\right) \end{cases} \quad (3)$$

Finally, the outcome of the sensing process can be seen as the output of a random policy $\pi_s(\epsilon, \delta, S_{a_t,t})$ such that:

$$X_t = \pi_s(\epsilon, \delta, S_{a_t,t}) \quad (4)$$

The design of such policies [3] is however out of the scope of this paper.

Depending on the sensing outcome $X_t \in \{0, 1\}$, the CA can choose to access the channel or not. We denote by $\pi_a(X_t) \in \{0, 1\}$ the access decision, where 0 refers to *access denied* and 1 refers to *access granted*. The access policy π_a chosen in this paper can be described as: "*access the channel if sensed available*", i.e. $\pi_a(X_t) = \mathbf{1}_{\{X_t=1\}}$[2]. Notice that we assume the ROC to be designed such that the probability of miss detection δ is smaller or equal to a given interference level allowed by the primary network, although $\{\epsilon, \delta\}$ are not necessarily known. Moreover, we assume that if interference occurs, it is detected and the transmission of the secondary user fails. When channel access is granted, the CA receives a numerical acknowledgment. This feedback informs the CA of the state of the transmission {succeeded, failed}. Finally, we assume that for every transmission attempt, a packet D_t is sent. At the end of every slot t, the CA can use the different information available to compute a numerical value, usually referred to as reward r_t in the MAB literature. This reward informs the CA of its current performance. The form of the reward as well as the evaluation of the selection policy π are described and discussed in the next subsection.

Finally, the sequential steps described hereabove formalize the OSA framework we are dealing with as a MAB problem with sensing errors. A schematic representation of a CA

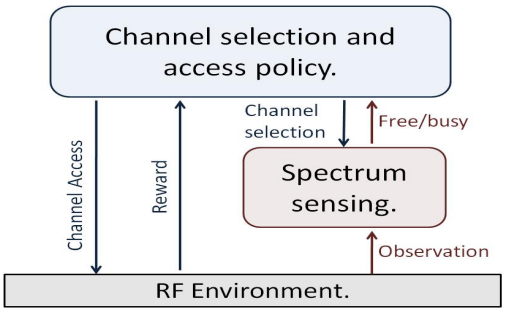

Fig. 1. Representation of a CA observing and accessing an RF environment.

[2]Indicator function: $\mathbf{1}_{\{logical\_expression\}}$={1 if logical_expression=true ; 0 if logical_expression=false}.

observing and accessing an RF environment is illustrated in Figure 1.

B. Performance evaluation

Thus, at the end of every slot t, the CA can compute a numerical value that evaluates its performance. In the case of OSA, we focus on the transmitted throughput. Relying on the previously introduced notations, the throughput achieved by the SU at the slot number t can be defined as:

$$r_t \triangleq D_t S_{a_t,t} \pi_a(X_t) \tag{5}$$

which is the reward considered in this particular framework. For the sake of simplicity we assume a normalized transmitted packet for all channels and all t, $D_t = 1$ bit. We can notice that the choices made on the access policy π_a and D_t, simplify the expression of the reward such that:

$$r_t = S_{a_t,t} X_t \tag{6}$$

where r_t equals 1 only if the channel is free and the CA senses it free. Consequently, the expected reward achievable using a channel $a_t \in \mathcal{A}$ can be easily computed:

$$\mathbb{E}[r_t] = \mathbb{P}(X_t = 1 | S_{a_t,t} = 1)\mathbb{P}(S_{a_t,t} = 1) = (1-\epsilon)\mu_{a_t} \tag{7}$$

Thus, we refer to the channel $\mu_1 = \max_k \mu_k$, that maximizes the reward, as optimal whereas the other channels are said to be suboptimal. We usually evaluate the performance of a policy by its expected cumulated throughput after t slots defined as:

$$W_t^\pi = \mathbb{E}\left[\sum_{m=0}^{t-1} r_m\right] \tag{8}$$

A good policy π is assumed to maximize the quantity W_t^π.

An alternative representation of the expected performance of a policy π until the slot number t is described through the notion of *regret* R_t^π (or expected regret). The regret is defined as the gap between the maximum achievable performance in expectation, if the most available channel were chosen, and the expected cumulated throughput achieved by the policy π:

$$R_t^\pi = \sum_{m=0}^{t-1} \max_{a_t \in \mathbb{A}} \mathbb{E}[r_t] - W_t^\pi \tag{9}$$

Hence, we define the regret of a channel selection policy π when sensing errors can occur as:

$$R_t^\pi = \sum_{m=0}^{t-1} (1-\epsilon)\mu_1 - W_t^\pi \tag{10}$$

The general idea behind the notion of *regret* can be explained as follows: if the CA knew *a priori* the values of $\{\mu_k\}_{k\in\mathcal{A}}$, the best choice would be to always select the optimal channel μ_1. Unfortunately, since usually the CA lacks that information, it has to learn it. For that purpose, the CA explores the different channels to acquire better estimations of their expected availability. While exploring it should also exploit the already collected information to minimize the regret during the learning process. This leads to an exploration-exploitation trade-off. Thus, the *regret* represents the loss due to suboptimal channel selections during the learning process.

Maximizing the expected throughput is equivalent to minimizing the cumulated expected regret. In the rest of the paper, we will use the following equivalent formula of the regret:

$$R_t^\pi = (1-\epsilon)\sum_{k=1}^{K} \Delta_k . \mathbb{E}\left[T_k(t)\right] \tag{11}$$

where $\Delta_k = \mu_1 - \mu_k$ and $T_k(t)$ refers to the number of times the channel k has been selected from instant 0 to instant $t-1$.

Finally we introduce a loss function $\mathcal{L}^\pi(t)$ that evaluates the loss of performance due to sensing errors compared to the perfect sensing framework.

$$\mathcal{L}^\pi(t) = t \max_{a_t \in \mathbb{A}} \mu_{a_t} - W_t^\pi \tag{12}$$

The next section reminds, first, the form of the UCB_1 algorithm. Then we prove that even if the characteristics of the receiver, $\{\epsilon, \delta\}$, are unknown, this algorithm suffers a number of suboptimal channel selections upper bounded by a logarithmic function of the slot number t. Finally we conclude by an evaluation of both the regret R^π and the loss function $\mathcal{L}^\pi$.

III. Upper Confidence Bound index for OSA with sensing errors

A. UCB_1 index

In a previous work [9] the authors suggested and discussed the use of Upper Confidence Bound (UCB) algorithms to build an efficient cognitive agent in order to tackle the OSA issue with perfect sensing. As a matter of fact, UCB based policies are known to offer a good solution to the exploration-exploitation trade-off. The general approach suggested by the UCB algorithms aims at selecting actions based on indexes that provide an optimistic evaluation on the rewards associated to the channels the secondary user can potentially exploit.

A usual approach to evaluate the average reward provided by a resource k is to consider a confidence bound for its sample mean. Let $\overline{X}_{k,T_k(t)}$ be the sample mean of the resource $k \in \mathcal{A}$ after being selected $T_k(t)$ times at the step t:

$$\overline{X}_{k,T_k(t)} = \frac{\sum_{m=0}^{t-1} r_m . \mathbf{1}_{\{a_m = k\}}}{T_k(t)} \tag{13}$$

For every $k \in \mathcal{A}$ and at every step $t = 0, 1, 2, \ldots$, an upper bound confidence index (UCB index), $B_{k,t,T_k(t)}$, is a numerical value computed from i_t. For all k, $B_{k,t,T_k(t)}$ gives an over estimation of the expected reward obtained when the resource k is selected at a time t after being sensed $T_k(t)$.

The UCB indexes we use in this paper have the following general expression:

$$B_{k,t,T_k(t)} = \overline{X}_{k,T_k(t)} + A_{k,t,T_k(t)} \tag{14}$$

where $A_{k,t,T_k(t)}$ is an upper confidence bias added to the sample mean. In this paper we consider the UCB_1 index which has the upper confidence bias $A_{k,t,T_k(t)}$ form:

$$A_{k,t,T_k(t)} = \sqrt{\frac{\alpha.\ln(t)}{T_k(t)}} \tag{15}$$

An UCB policy π selects the next channel a_t based on the past information i_t such that:

$$a_t = \pi(i_t) = \arg\max_k (B_{k,t,T_k(t)}) \tag{16}$$

A detailed version of the implementation of the algorithm UCB_1 was described in a previous work [9].

B. UCB_1: channel selection with sensing errors

The following theorem shows that although the CA suffers imperfect sensing, it still can converge quickly to the most available channel.

Theorem 1 (Logarithmic suboptimal channel selection): Let us consider a receiver with sensing characteristics $\{\epsilon, \delta\}$, and an "*access the channel if sensed available*" policy. We consider the instantaneous normalized throughput as the CA's reward.

Then for all $K \geq 2$, if the receiver runs the $UCB_1(\alpha > 1)$ policy on K channels having Bernoulli occupation pattern distributions $\theta_1, \cdots, \theta_K$ with support in $[0, 1]$, the expected number of selections $\mathbb{E}[T_k(t)]$ for all suboptimal channels $k \in \{2, \cdots, K\}$ after t slots is upper bounded by a logarithmic function such that:

$$\mathbb{E}[T_k(t)] \leq \frac{4\alpha \ln(t)}{((1-\epsilon)\Delta_k)^2} \tag{17}$$

Proof: Due to space limitations and in order to make this paper as self content as possible we provide an intuitive proof:

Let us consider Bernoulli occupation pattern distributions $\Theta = \{\theta_1, \cdots, \theta_K\}$ with support in $[0, 1]$. As noticed previously, CR equipment' sensors can be seen as functions $\pi_s(\epsilon, \delta, \cdot)$ with parameters $\{\epsilon, \delta\}$ that map a random realisation $S_{k,t}$ drawn form the distribution θ_k, at the slot number $t \in \mathbb{N}$, into a binary value $X_t \in \{0, 1\}$ such that:

$$X_t = \pi_s(\epsilon, \delta, S_{k,t}) \tag{18}$$

Let us define the set of reward distributions $\tilde{\Theta} = \{\tilde{\theta}_1, \cdots, \tilde{\theta}_K\}$ such that: $\forall t \in \mathbb{N}$, the reward $r_t = S_{k,t}X_t$ computed when the channel k is selected follows the distribution $\tilde{\theta}_k$. Then the distributions $\tilde{\Theta} = \{\tilde{\theta}_1, \cdots, \tilde{\theta}_K\}$ are bounded distributions with support in $[0, 1]$.

Moreover let us define:

$$\forall k \in \{1, 2, \cdots, K\}, \tilde{\mu}_k \triangleq \mathbb{E}[\tilde{\theta}_k] \tag{19}$$

Under the assumptions of this theorem, we can write for all $k \in \{1, 2, \cdots, K\}$:

$$\begin{cases} \tilde{\mu}_k = (1-\epsilon)\mu_k \\ \tilde{\Delta}_k = (1-\epsilon)\Delta_k \end{cases} \tag{20}$$

Consequently we can apply the following theorem (Cf. [7] for proof):

For all $K \geq 2$, if policy $UCB_1(\alpha > 1)$ is run on K channels having arbitrary reward distributions $\theta_1, \cdots, \theta_K$ with support in [0,1], then:

$$\mathbb{E}[T_k(t)] \leq \frac{4\alpha}{\Delta_k^2} \ln(t) \tag{21}$$

Finally, by substituting: $\mu_k \rightleftharpoons (1-\epsilon)\mu_k$ and $\Delta_k \rightleftharpoons (1-\epsilon)\Delta_k$ we obtain the stated result:

$$\mathbb{E}[T_k(t)] \leq \frac{4\alpha \ln(t)}{((1-\epsilon)\Delta_k)^2} \tag{22}$$

■

The consequences of Theorem 1 are twofold: on the one hand, as in the case of perfect sensing, UCB_1 based policies used in the case of OSA with sensing errors spend exponentially more time probing the optimal channel than suboptimal channels[3]. On the other hand, we notice that the exploration phase, characterized by the time spent on suboptimal channels increases with a scale $\frac{1}{(1-\epsilon)^2}$ compared to the perfect sensing framework. Thus, as expected the accuracy of the sensor is crucial in order the maximize SUs' profit.

Corollary 1 (Regret and Loss function): Assuming that we verify the assumptions and conditions of Theorem 1, the regret and the loss function can be upper bounded as follows:

$$\begin{cases} R_t^\pi \leq \sum_{k=1}^K \frac{4\alpha \ln(t)}{((1-\epsilon)\Delta_k)} \\ \\ \mathcal{L}^\pi(t) \leq \epsilon t + \sum_{k=1}^K \frac{4\alpha \ln(t)}{((1-\epsilon)\Delta_k)} \end{cases} \tag{23}$$

Proof: First,we can notice that:

$$\mathcal{L}^\pi(t) = \epsilon t + R_t^\pi \tag{24}$$

The rest of the proof is an immediate application of the result of Equation 17 of Theorem 1, to Equation 11 and Equation 12. ■

The first result of the corollary shows that the regret, as defined in machine learning, is still upper bounded by a logarithmic function of the slot number t. However, as for $\mathbb{E}[T_k(t)]$, due to sensing errors, the regret increases by a scaling factor equal to $1/(1-\epsilon)$. The second result shows that compared to the perfect sensing framework, the SU suffers unavoidable linear expected loss due to sensing errors.

IV. Simulations

In this section we present and comment simulation curves focusing on the regret and on the optimal channel selection. The curves compare the behavior of the UCB_1 algorithm under various sensing characteristics.

We consider, in our simulations, one SU willing to exploit a pool of 10 channels. The parameters of the Bernoulli distributions are $[\mu_1, \mu_9, \ldots, \mu_{10}] = [0.9, 0.8, 0.8 : -0.1 : 0.1]$. These distribution characterize the temporal occupancy of these channels. To avoid causing interference to PU's, we assume that an adequate δ is guaranteed. Since, ϵ and δ are related to one another through their ROC, the values of ϵ are

[3]Notice that $\mathbb{E}[T_k(t)]$ only depends explicitly on ϵ because of the feedback. This latter avoids considering failed transmissions as rewards (Equation 7)!

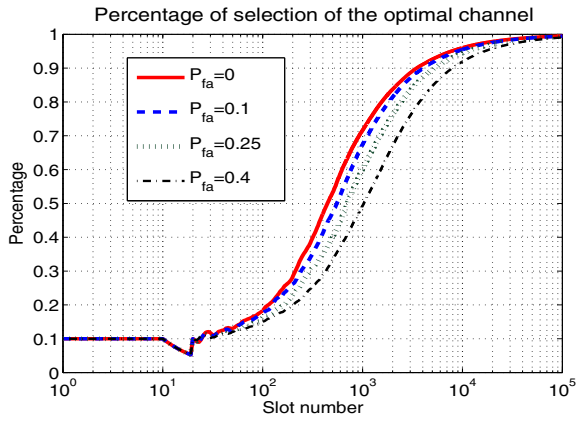

Fig. 2. Percentage of time the UCB_1-based CA selects the optimal channel under various sensing errors frameworks (over 10 available channels).

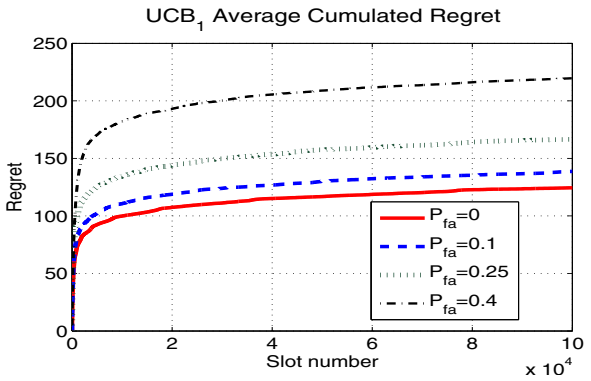

Fig. 3. UCB_1 algorithm and Opportunistic Spectrum Access problem with sensing errors: regret simulation results.

imposed depending on the channels' conditions.In order to evaluate the impact of these parameters on the CA's behavior, we chose to simulate the UCB_1 algorithm with four different sensors: $\epsilon = [0, 0.1, 0.25, 0.4]$. Moreover, in order to respect the conditions stated in Theorem 1, UCB_1 was run with the parameter $\alpha = 1.2$. Every numerical result reported hereafter is the average of the values obtained over 100 experiments.

Figure 3 shows the evolution of the average regret achieved by the UCB_1 policy under various sensing characteristics. As expected (Cf. Corollory 1), we observe that the regret first increases rather rapidly with the slot number and then more and more slowly. We remind that the smaller the regret is, the better is the algorithm behaving. This shows that the UCB policy is able to process the past information in an appropriate way even if there are sensing errors such that most available resources are favored with time. Actually, one has the theoretical guarantee that it will converge to $(1-\epsilon)\mu_1$, which is the largest probability of availability of the optimal channel within the herein modeled imperfect sensing framework. We however notice that the sensing errors increase the cumulated regret. The smallest regret is achieved as expected in the case of perfect sensing ($\epsilon = \mathbb{P}_{fa} = 0$). Moreover, we can notive that the ratio of the regret in the case of perfect sensing and in the case of sensing errors characterized by $\epsilon \neq 0$ is approximately equal to $1/(1-\epsilon)$ which supports the theoretical results.

The optimal channel selection percentage p achieved by the UCB_1 algorithm until the slot number t is illustrated in Figure 2, where $p = 100.\frac{\sum_{m=0}^{t-1} \mathbf{1}_{\{a_m=1\}}}{t}$. As one can observe the percentage of optimal channel selection increases progressively and tends to get closer and closer to 100% as the slot number increases.

As for the regret analysis, we observe that the performance of the UCB_1 algorithm decreases when the $\mathbb{P}_{fa}$ increase. Thus, the UCB_1 with perfect sensing performs best. The increasing rate of the other curves is slower depending on their sensing capabilities. As proven in the theoretical analysis provided hereabove, all UCB_1 algorithms converge to the best channel, however the less accurate is their sensing outcome, the slower becomes their convergence rate.

V. CONCLUSION

We tackled in this paper the OSA problem with sensing errors within a MAB framework. We argued that the UCB_1 algorithm used as channel selection policy can still offer a good trade-off to the exploration-exploitation dilemma faced by the SU. Thus, we showed that the time spent on suboptimal channels is upper-bounded by a logarithmic function of the slot number, ensuring a quick convergence to the optimal channel. Although these preliminary results are promising, many questions still need to be answered especially when several SUs compete to access the same resources.

ACKNOWLEDGMENT

This work was also supported by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (contract n. 216715).

REFERENCES

[1] Federal Communications Commission. Spectrum policy task force report. November 2002.

[2] J. Mitola and G.Q. Maguire. Cognitive radio: making software radios more personal. *Personal Communications, IEEE*, 6:13–18, August 1999.

[3] T. Yucek and H. Arslan. A survey of spectrum sensing algorithms for cognitive radio applications. *In IEEE Communications Surveys and Tutorials*, 11, no.1, 2009.

[4] Q. Zhao and B. M. Sadler. A survey of dynamic spectrum access: signal processing, networking, and regulatory policy. In *in IEEE Signal Processing Magazine*, pages 79–89, 2007.

[5] R. Agrawal. Sample mean based index policies with O(log(n)) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078, 1995.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of multi-armed bandit problems. *Machine learning*, 47(2/3):235–256, 2002.

[7] J.-Y. Audibert, R. Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. In *Proceedings of the 18th international conference on Algorithmic Learning Theory*, 2007.

[8] L. Lai, H.E. Gamal, H.J. Jiang, and V. Poor. Cognitive medium access: Exploration, exploitation and competition. *[Online]. Available: http://arxiv.org/abs/0710.1385*.

[9] W. Jouini, D. Ernst, C. Moy, and J. Palicot. Upper confidence bound based decision making strategies and dynamic spectrum access. *Proceedings of the 2010 IEEE International Conference on Communications (ICC)*, May 2010.

[10] L. Keqin and Q. Zhao. Distributed learning in cognitive radio networks: Multi-armed bandit with distributed multiple players. 2010.