This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE ICC 2010 proceedings

Upper confidence bound based decision making strategies and dynamic spectrum access

Wassim Jouini
SUPELEC/IETR
wassim.jouini@supelec.fr

Damien Ernst
University of Liège
dernst@ulg.ac.be

Christophe Moy
SUPELEC/IETR
christophe.moy@supelec.fr

Jacques Palicot
SUPELEC/IETR
jacques.palicot@supelec.fr

***Abstract*—In this paper, we consider the problem of exploiting spectrum resources for a secondary user (SU) of a wireless communication network. We suggest that Upper Confidence Bound (UCB) algorithms could be useful to design decision making strategies for SUs to exploit intelligently the spectrum resources based on their past observations. The algorithms use an index that provides an optimistic estimation of the availability of the resources to the SU. The suggestion is supported by some experimental results carried out on a specific dynamic spectrum access (DSA) framework.**

***Index Terms*—Cognitive Radio, Dynamic Spectrum Access, Upper Confidence Bound Algorithm.**

I. Introduction

A. Dynamic spectrum access

During the last century, most of the meaningful frequency bands were licensed to the emerging wireless applications. Because of the static model of frequency allocation, the growing number of spectrum demanding services led to a spectrum scarcity. However, recently, series of measurements on the spectrum utilization [1] showed that the different frequency bands were underutilized (sometimes even unoccupied) and thus that the scarcity of the spectrum resource is virtual and only due to the static allocation of the different bands to specific wireless services. Moreover, the underutilization of the spectrum resource varies on different scales in time and space offering many opportunities to an unlicensed user or network to access the spectrum.

Dynamic Spectrum Access (DSA, also known as Opportunistic Spectrum Access: OSA) was introduced as a possible solution that could alleviate the spectrum scarcity issue. In general, DSA related issues consider a pool of users referred to as primary users (PUs). PUs access spectrum resources dedicated to the services provided (or available) to them. Consequently they have an unconstrained access to these resources. The primary users communicate in a primary network (PN) which is characterized by its environment, i.e., its geographical position as well as the resources provided during a certain amount of time.

The concept of DSA allows new users to access their surrounding PU's licensed bands even though they do not belong to the primary network. These users are referred to as secondary users (SUs). The main goal of a SU is to find in his surrounding environment new communication opportunities compared to the usual and current spectrum allocation scheme.

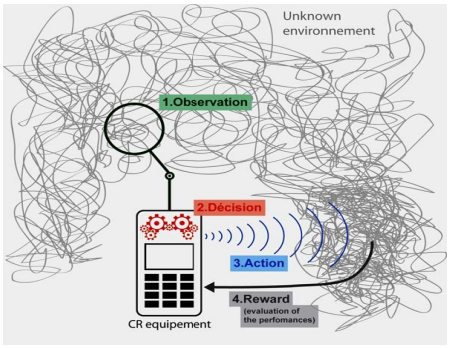

Fig. 1. Cognitive Radio context.

Usually an opportunity, in DSA related issues, is defined as: *a band of frequencies that are not being used by the primary users of that band at a particular time in a particular geographic area* [2]. However, a SU usually has no *a priori* information on the available opportunities surrounding him. To that issue, the Federal Communications Commission (USA) suggested the concept of Cognitive Radio, introduced by J. Mitola [3] in 1999, as a possible solution.

B. Decision making engine of a cognitive radio equipment

A Cognitive Radio (CR) device is a communication system aware of its environment as well as of its operational abilities and capable of using them intelligently. Thus it is a device that has the ability to collect information through it sensors and that can use the past observations on its surrounding environment to improve its behavior consequently. A simplified cognitive radio behavior in DSA is illustrated in Figure 1: the CR equipment observes its surrounding environment looking for opportunities. As illustrated by the magnifying glass, usually, a CR cannot see (or sense) the entire environment altogether. The results of these observations are taken into account by the decision making engine that decides on the next action to take (e.g. which part of the environment to sense? transmit or not transmit?). In some cases a numerical signal (reward or acknowledgment) is computed and help the CR equipment to evaluate its performance at that specific time.

The design of such CR equipments to tackle OSA issues has been, recently, the center of a lot of attention (e.g. [3] [4] [5]). We refer to as Cognitive Agent (CA) the decision making engine of the CR equipment that can be seen as the

978-1-4244-6404-3/10/$26.00 ©2010 IEEE

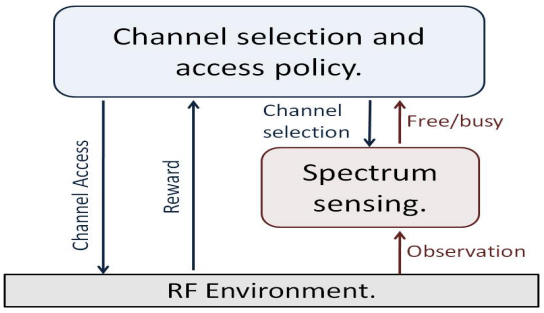

Fig. 2. Cognitive radio resource selection and access.

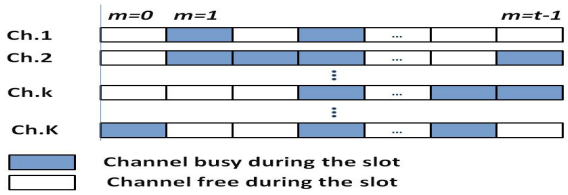

Fig. 3. Occupancy of the different channels considered by the SU.

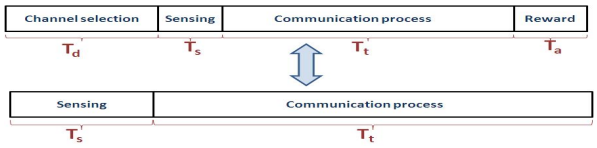

Fig. 4. Slot representation for a radio equipment controlled by a CA. It is assumed here that $T_d + T_a$ are small with respect to T_s and T_t.

brain of the CR device. At the level of the CA, the challenges are twofold: on the one hand, the SU must not compromise the efficiency of the primary network. Thus, a proper sensing of the environment must be done to avoid interfering with PUs. On the other hand, the SU has to find an allocation policy to select, and if possible, access the available resources. A simple representation of the different interactions between the environment and the cognitive agent is described in Figure 2.

In this paper, we assume that the CA can only take actions, (e.g. select and access a channel if possible) at discrete time instants $t = 0, 1, 2, ...$. At every instant t, the CA observes its radio frequency environment and can collect different kind of information (e.g., available frequency bands, noise level, position, throughput, etc.). All the information collected by the CA up to instant t is supposed to be gathered in a vector i_t. We assume that the CA has to select at every instant t an action a_t in a discrete set $\mathcal{A}$. Without loss of generality, the behavior of the CA can be seen as a policy (decision strategy) π that maps the information vector i_t into the action $a_t \in \mathcal{A}$, that is:

$$a_t = \pi(i_t) \tag{1}$$

The purpose of this paper is to study the performance of a particular policy on an academic DSA problem. The academic DSA problem is described in Section II. The policy which is based on the computation of upper confidence bound indexes is described in Section III. Section IV reports the simulation results and, finally, Section V concludes.

II. Dynamic spectrum access: network model

We consider a single secondary user (SU) operating in a primary network composed of K channels referenced by the integers $\{1, 2, \ldots, K\}$. The CR equipment of the SU can only sense (then access if possible) one channel at a time. As illustrated in Figure 3, we address the particular case where the time is divided into slots $t = 0, 1, 2, \ldots$, and that PUs are synchronous.

The temporal occupancy pattern of every channel k is supposed to follow an unknown Bernoulli distribution θ_k. Moreover, the distributions $\theta_1, \theta_2, \ldots, \theta_K$ are assumed to be stationary. When the SU senses a channel k at the slot number t, the cognitive agent computes a binary signal X_t that provides information on the availability of the sensed slot at that particular instant t. X_t is an independent realization of the distribution θ_k, at the slot t.

Let us define μ_k as follows: $\forall k$,

$$\mu_k \triangleq E[\theta_k] = P(\text{channel } k \text{ is free})$$

Without loss of generality, we assume that $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_{K-1} < \mu_K$. Moreover, we assume in this paper that the outcome of the sensing process is error free. However the distribution probabilities $\theta_1, \theta_2, \ldots, \theta_K$ are assumed to be unknown to the CA.

At every instant t, and for every channel k the state of the channel observed by the SU can be either free or busy. If the channel is free, the CR equipment can transmit a certain number of bits B_t. Otherwise, the CR equipment waits until the next slot and selects a new channel to sense. A slot is divided into 4 periods (cf. Figure 4). During the first period, the CA chooses the next channel to access. During the second period the CA senses the selected channel before communicating if it is possible (channel free during the slot). At the end of every slot t, the CA computes a numerical signal referred to as reward r_t that depends on the occupancy state of the selected channel and evaluates the CA's performance (e.g., throughput in this paper) during the communication process. The added information at the end of every slot is used to improve the decision making behavior of the CA which is characterized by the policy π. As mentioned earlier, this policy takes an information vector i_t as input and outputs the action to be selected at time t. The action is here the channel to select, $\mathcal{A} = \{1, 2, \ldots, K\}$, and the information vector is $i_t = [a_0, r_0, a_1, r_1, \ldots, a_{t-1}, r_{t-1}]$.

The throughput achieved by the CR equipment at the slot number t can be defined as:

$$r_t \triangleq B_t . X_t \tag{2}$$

which is the reward considered in this particular framework. For the sake of simplicity we assume here that if the channel is free the CR can always transmit $B_t = B$ bits. Thus, the cumulated throughput after t slots can be written:

$$W_t^\pi = \sum_{m=0}^{t-1} r_m = B \sum_{m=0}^{t-1} X_m$$

where the suffix π is used to emphasize that the CR equipment uses the policy π to select the channels.

The purpose of the CA is to maximize the expected cumulated throughput of the CR equipment:

$$E[W_t^\pi] = B \sum_{m=0}^{t-1} E[X_m] \tag{3}$$

Let R_t^π denote the regret of the CA at the slot number t, using a policy π. The regret R_t^π is defined as:

$$R_t^\pi = B.\mu_K.t - W_t^\pi \tag{4}$$

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 channel with the highest expected availability, i.e., μ_K. Unfortunately, the CA usually lacks that information and has to learn it. For that purpose, the CA has to explore the channels in order to have better estimations of their temporal occupancy pattern. While exploring it should also exploit the already collected information to minimize the regret during the learning process. This leads to an exploration-exploitation tradeoff. 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. The expected cumulated regret can be written as follows:

$$E[R_t^\pi] = B.\sum_{k=1}^{K} \Delta_k.E[T_k(t)] = B.E[\tilde{R}_t^\pi] \tag{5}$$

where $\tilde{R}_t^\pi = \frac{R_t^\pi}{B}$, $\Delta_k = \mu_K - \mu_k$ and $T_k(t)$ refers to the number of times the channel k has been selected from instant 0 to instant $t-1$.

We propose in the next section policies π that upper bound the expected cumulated regret of the CR equipment by a logarithmic function of the slot number.

III. Upper Confidence Bound index

A. UCB index

Building a cognitive agent to tackle the DSA issue requires to find a policy π for this agent that offers a good solution to the exploration-exploitation tradeoff behind the notion of regret's minimization. The general approach suggested in this section aims at selecting actions based on indexes that provide upper confidence bounds (UCB) on the rewards associated to the channels the secondary user can potentially exploit. Policies based on the computation of UCB indexes were initially introduced in the machine learning community to solve the so-called multi-armed bandit problem (see [6] and [7]).

Parameters: *K, exploration coefficient α*
Input: $i_t = [a_0, r_0, a_1, r_1, \ldots, a_{t-1}, r_{t-1}]$
Output: a_t
Algorithm:
If: $t \leq K$ *return* $a_t = t+1$
Else:

- $T_k(t) \leftarrow \sum_{m=0}^{t-1} \mathbf{1}_{\{a_m=k\}}$, $\forall k$
- $A_{k,t,T_k(t)} \leftarrow \sqrt{\frac{\alpha.\ln(t)}{T_k(t)}}$, $\forall k$
- $B_{k,t,T_k(t)} \leftarrow \overline{X}_{k,T_k(t)} + A_{k,t,T_k(t)}$, $\forall k$
- return $a_t = \arg\max_k (B_{k,t,T_k(t)})$

Fig. 5. A tabular version of a policy $\pi(i_t)$ using a UCB_1 algorithm for computing actions a_t.

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:

$$\overline{X}_{k,T_k(t)} = \frac{\sum_{m=0}^{t-1} r_m.\mathbf{1}_{\{a_m=k\}}}{t} \tag{6}$$

[1]

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 optimistic estimation of the expected reward obtained when the CA selects the resource k at a time t after being tested $T_k(t)$.

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

$$B_{k,t,T_k(t)} = \overline{X}_{k,T_k(t)} + A_{k,t,T_k(t)} \tag{7}$$

where $A_{k,t,T_k(t)}$ is an upper confidence bias added to the sample mean.

An upper confidence bound (UCB) based cognitive agent uses a policy π to compute from i_t these indexes from which it selects a resource a_t as follows:

$$a_t = \pi(i_t) = \arg\max_k (B_{k,t,T_k(t)}) \tag{8}$$

1) UCB_1 [8] [9]: When using the following upper confidence bias:

$$A_{k,t,T_k(t)} = \sqrt{\frac{\alpha.\ln(t)}{T_k(t)}} \tag{9}$$

with $\alpha > 1$, we obtain an upper confidence bound index referred to as UCB_1 in the literature. A fully detailed version of the policy using UCB_1 indexes is given in Figure 5.

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

2) UCB_V [8]: The UCB_1 index uses only first order statistic information (empirical mean). It was suggested in [9] that adding the second order statistic information (empirical variance) to the UCB indexes could lead to better performances. The UCB_V index exploits the empirical variance of the estimated rewards. More specifically it uses the following upper confidence bias:

$$A_{k,t,T_k(t)} = \sqrt{\frac{2\xi.V_k(t).\ln(t)}{T_k(t)}} + \frac{3.c.\xi.\ln(t)}{T_k(t)} \qquad (10)$$

with $c \geq 1$ and $3.\xi.c > 1$ and where $V_k(t)$ refers to the empirical variance of the channel k.

In Section IV we will compare the performances of UCB_1 and UCB_V policies on the dynamic spectrum access problem introduced in Section II.

B. Performance evaluation

When using a policy π, an interesting way to analyze its behavior is to consider the notion of consistency. This notion gives information on the growth rate of the regret. A policy π is said to be β-*consistent*, $0 < \beta \leq 1$, if it satisfies:

$$\lim_{t\to\infty} \frac{E[R_t^\pi]}{t^\beta} = 0 \qquad (11)$$

We expect a good policy to be at least *1-consistent*. As a matter of fact, this property ensures that asymptotically the mean expected reward is optimal, i.e.:

$$\lim_{t\to\infty} \frac{\sum_{m=0}^{t-1} r_m}{t} = B.\mu_K \qquad (12)$$

Theorem 1: (cf. [8] for proofs) For all $K \geq 2$, if policy $UCB_1(\alpha > 1)$ is run on K channels having arbitrary reward distributions $\theta_1, ..., \theta_K$ with support in [0,1], then:

$$E[\tilde{R}_t^{\pi=UCB_1}] \leq \sum_{k:\Delta_k>0} \frac{4.\alpha}{\Delta_k}.\ln(t) \qquad (13)$$

Notice that a similar theorem could be written if the reward distributions had a bounded support rather than a support in [0,1].

An equivalent theorem also exists for the index UCB_V:

Theorem 2: (cf. [8] for proofs) For all $K \geq 2$, if policy $UCB_V(\xi \geq 1, c = 1)$ is run on K channels having arbitrary reward distributions $\theta_1, ..., \theta_K$ with support in [0,1], then $\exists C_\xi > 0$ s.t.

$$E[\tilde{R}_t^{\pi=UCB_V}] \leq C_\xi \sum_{k:\Delta_k>0} (\frac{\sigma_k^2}{\Delta_k} + 2).\ln(t) \qquad (14)$$

Actually a similar result would still hold if $c \geq 1$ but satisfies nonetheless $3.\xi.c > 1$.

These results are of a particular interest for many reasons:

- They bound the expected regret of the UCB policies by a logarithmic functions for all t. This guarantees that the suggested policies are β consistent for all $0 < \beta \leq 1$. Thus these policies converge quickly to the optimal channel K.
- Moreover, the indexes these policies rely on to select actions can be computed incrementally [10]. Thus, their complexity, in terms of memory usage and computational needs, are low.
- Last but not least, it has been proven in [6] that when having no *a priori* information on the temporal occupancy pattern of the different channels $\theta_1, \theta_2, \ldots, \theta_K$, a logarithmic upper bound is the best we can expect.

IV. Simulations

In our simulations, we consider that the CA agent can choose between 10 channels. The parameters of the Bernouilli distributions which characterize the temporal occupancy of these channels are: $[\mu_1, \mu_2, \ldots, \mu_{10}] = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9]$. We consider that the number of bits a SU can transmit on a free channel is $B = 1$bit.

Every numerical result reported hereafter is the average of the values obtained over 100 experiments.

In this section, the parameter α of the UCB_1 algorithm is chosen equal to 1.2. The parameters ξ and c of the UCB_V algorithm are equal to 1 and 0.4, respectively. With such values for c and ξ, the condition $3.\xi.c > 1$ is satisfied and the bound on the expected cumulated regret given by Equation (14) still holds. The simulation results depend on the paramaters values, however we chose these values to be close to the critical ones ($\alpha = 1$, $\xi = 1$ and $c = 1/3$) without being too conservative.

Figure 6-top shows the evolution of the average cumulated regret for the different UCB policies. For both policies, the cumulated regret first increases rather rapidly with the slot number and then more and more slowly. This shows that UCB policies are able to process the past information in an appropriate way such that most available resources are favored with time. This is further illustrated by the 3 graphics on the bottom of Figure 6. These graphics show the average throughput achieved by the UCB policies. As we observe, the throughput increases with time. Actually, one has the theoretical guarantee that it will converge to 0.9, which is the largest probability of availability of a channel. Figure 7 shows the percentage p of times a UCB policy selects the optimal channel until the slot number t ($p = 100.\frac{\sum_{m=0}^{t-1} \mathbf{1}_{\{a_m=K\}}}{t}$). As one can observe, this percentage tends to get closer and closer to 100 as the slot number increases.

In our simulations results, we have always found out that UCB_1 seems to outperform UCB_V at the beginning of the learning process and that, afterwards, UCB_V outperforms UCB_1. This may be explained by the fact at the beginning of the learning UCB_V spends more time collecting information on the different channels than UCB_1 since it also depends on the variances of the different channels and not only on their empirical mean. During this phase, it mainly has a pure exploration strategy while UCB_1 starts already exploiting the information that has been gathered. However, once it starts having good estimates of these variances, it can address the

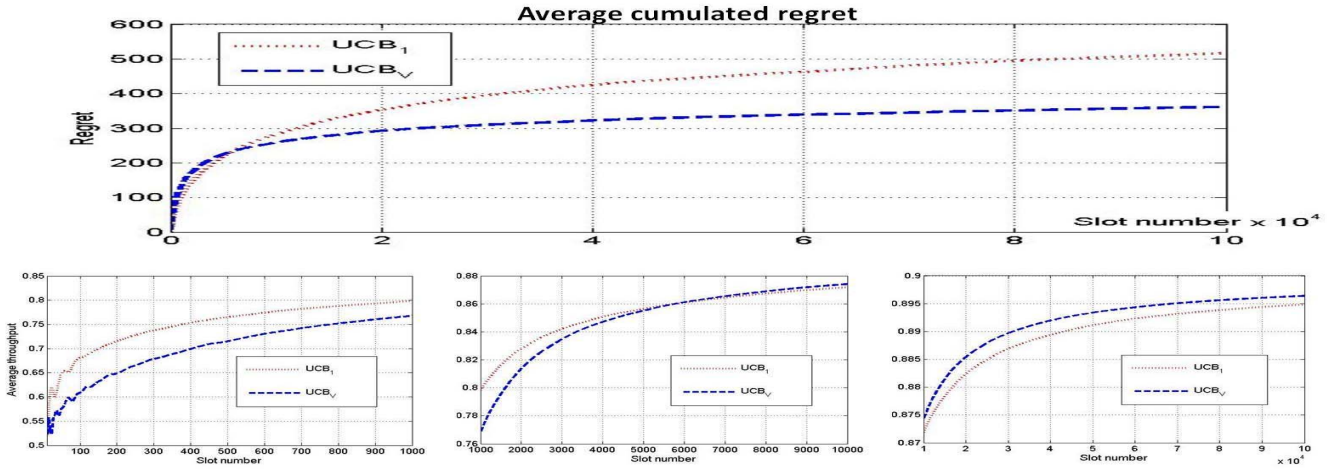

Fig. 6. UCB based policies and dynamic spectrum access problem: simulation results. Figure on top plots the average cumulated reward as a function of the number of slots for the different UCB based policies. The figures on the bottom represent the evolution of the normalized average throughput achieved by these policies.

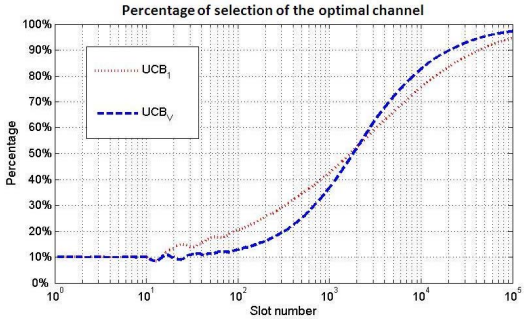

Fig. 7. Percentage of time a UCB-based policy selects the optimal channel.

exploration-exploitation tradeoff in a more efficient way than UCB_1.

V. CONCLUSION

We presented in this paper a new approach to tackle the resource selection and access problem in dynamic spectrum access in the case of one secondary user in a primary network. This approach exploits some upper confidence based algorithms introduced in the machine learning community for solving the multi-armed bandit problems. Although this research is still in its infancy, we believe that this approach can lead to efficient CAs to address DSA problems. However many questions still need to be answered especially when the temporal occupancy pattern of the channels do not follow Bernoulli distributions or when many SUs use these UCB based policies to access the same primary network.

ACKNOWLEDGMENT

Damien Ernst is a Research Associate of the Belgian FRS-FNRS of which he acknowledges the financial support.

REFERENCES

[1] Federal Communications Commission. Spectrum policy task force report. November 2002.
[2] P. Kolodzy and al. Next generation communications: Kickoff meeting. *In Proc. DARPA*, October 2001.
[3] J. Mitola and G.Q. Maguire. Cognitive radio: making software radios more personal. *Personal Communications, IEEE*, 6:13–18, August 1999.
[4] S. Haykin. Cognitive radio: brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 23, no. 2:201–220, Feb 2005.
[5] 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.
[6] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.
[7] 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.
[8] 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.
[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of multi-armed bandit problems. *Machine learning*, 47(2/3):235–256, 2002.
[10] W. Jouini, D. Ernst, C. Moy, and J. Palicot. Multi-armed bandit based policies for cognitive radio's decision making issues. *In Proceedings of the 3rd international conference on Signals, Circuits and Systems (SCS)*, November 2009.