Managing Single or Multi-Users Channel Allocation for the Priority Cognitive Access

M. Almasri[1], A. Mansour[1], C. Moy[2], A. Assoum[3], D. Le Jeune[1], C. Osswald[1],
[1] LABSTICC, UMR 6285 CNRS, ENSTA Bretagne, 2 rue F. Verny, 29806 Brest Cedex 9, France.
[2] Univ Rennes, CNRS, IETR - UMR 6164, F-35000, Rennes, France.
[3] Lebanese University, Faculty of sciences, Tripoli, Lebanon.

***Abstract*—This manuscript investigates the problem of the Multi-Armed Bandit (MAB) in the context of the Opportunistic Spectrum Access (OSA) case with priority management (e.g. military applications). The main aim of a Secondary User (SU) in OSA is to increase his transmission throughput by seeking the best channel with the highest vacancy probability. In this manuscript, we propose a novel MAB algorithm called ϵ-UCB in order to enhance the spectrum learning of a SU and decrease the regret, i.e. the loss of reward due to the selection of worst channels. We analytically prove, and corroborate with simulations, that the regret of the proposed algorithm has a logarithmic behavior. So, after a finite number of time slots, the SU can estimate the vacancy probability of channels in order to target the best one for transmitting. Hereinafter, we extend ϵ-UCB to consider multiple priority users, where a SU can selfishly estimate and access the channels according to his prior rank. The simulation results show the superiority of the proposed algorithm for a single or multi-user cases compared to the well-known MAB algorithms.**

***Index Terms*—Cognitive Networks, Multi-Armed Bandit, Priority Access, Logarithmic Regret**

I. Introduction

The rapid growth of wireless services is becoming a cause of major stress in limited spectrum. Indeed, fixed bandwidths result in low utilization of spectrum resources as per the spectrum assignment policy. Besides, new wireless applications (such as: Machine to Machine, Internet of Things, Vehicle to Vehicle, etc.) require more spectrum to provide efficient services. In order to tackle this problem, the Cognitive Radio (CR), a new communication paradigm, has been proposed to enhance the use of the spectrum [1].

A. Cognitive Radio

One of the CR schemes, called Opportunistic Spectrum Access (OSA), can improve the spectral efficiency by sharing the available spectrum between a Primary User (PU), i.e. a licensed user that has a priority access to his frequency band, and an opportunistic user called a secondary user (SU). IEEE 802.22 represents the first standard based on CR and is being used in the United States [2]. This standard allows the opportunistic use of the white spectrum allocated to the TV service. In this context, the Base Station (BS) can sense the spectrum and identify the white space of TV channels to provide broadband services for different users. In our work, we focus on decentralized secondary networks, where the users can identify selfishly the availabilities of channels without any need for a central unit. A SU can sense and access an unoccupied channel, but he must leave the targeted channel when a PU reuses his channel. The main target of a SU is to increase his transmission time by finding the best channel, i.e. the most vacant one. Due to the hardware constraint and the detection cost, i.e. delay and energy, it is impractical for the SU to scan all channels at each time slot. Therefore, under the partially observable (one channel/slot), the SU should select a channel in a time interval and decide if the detected channel is free to transmit his data or not. To help the SU making a decision, we formulate the access of channels as a Multi-Armed Bandit (MAB) problem.

B. Multi-Armed Bandit

Due to its generic nature, the MAB problem has a fundamental importance in the stochastic decision theory as it can be applied in many situations, such as: wireless channel access, jamming communication, object tracking, ads selection on web pages, etc. In MAB framework, an agent is facing K slot machines. At each time slot, the agent can choose one machine and obtain a reward according to a fixed distribution. The main goal of an agent is to maximize his expected reward by selecting the machine with the highest expected reward (exploitation) and to try from time to time another machine to gather more information about its expected reward (exploration). The most popular MAB algorithms to solve the MAB problem are: Thompson Sampling (TS) [3], Upper Confidence Bound (UCB) [4], ϵ-greedy [5], etc.

II. Problem Formulation

Let K be the number of independent identically distributed (i.i.d.) channels. Assuming that the channels are sorted by their mean availability, i.e. $\mu_1 > \mu_2 > ... > \mu_K$, where μ_1 is the best channel with the highest availability probability and $\Gamma = \{\mu_i\}$ stands for the availability vector. As μ_i are unknown for the user, we define the regret as the sum of the reward loss due to the selection of a sub-optimal channel at each slot. In a single user case, the regret $R(n, \beta)$, up to the total number of slots n under a policy, β, can be defined as follows:

$$R(n, \beta) = n\mu_1 - \sum_{t=1}^{n} \mu_i^{\beta(t)}(t) \qquad (1)$$

where $n\mu_1$ means that the best channel has always been selected; $\beta(t)$ denotes the channel selected under the policy β

at time t; $\mu_i^{\beta(t)}$ is the mean reward obtained for the i^{th} channel selected at the time slot t and $\beta(t) = i$. Let $T_i(n)$ denote the number of time slots that the channel i was sensed by the SU up to the time n. As the user can only sense one channel at each time, then: $\sum_{i=1}^{K} T_i(n) = n$. The main target of a SU is to estimate channels' availabilities as soon as possible in order to target the highest available one. To reach this goal, several MAB algorithms have been applied in the context of CR.

TS represents one of the earlist MAB algorithm that is proposed in [3]. In TS, each channel has assigned an index $B_i(t, T_i(t))$ and at each time slot the agent selects the channel with the highest index $B_i(t, T_i(t))$:

$$B_i(t, T_i(t)) = \frac{W_i(t, T_i(t)) + a}{W_i(t, T_i(t)) + Z_i(t, T_i(t)) + a + b} \tag{2}$$

where $W_i(t, T_i(t))$ and $Z_i(t, T_i(t))$ represent respectively the success and failure access; a and b are constant numbers. Besides, its performance that can exceed the state-of-the-art of MAB algorithms [6], [7], TS is largly ignored in the literature by the Machine learning community. The rejection of TS may refer to the fact that this algorithm is proposed without any analytical proof. Recently, TS more and more attracts the attention and several works investigated the analytical proof of its convergence to the best choice [8], [9].

UCB, another important MAB algorithms, that represents the widely mentionned MAB algorithms in the literature. Several versions of UCB have been suggested in the literature in order to obtain good performance compared to the classical one: UCB1, UCB2, UCB-tuned, Bayes-UCB, KL-UCB [10]–[13]. In [10], the authors proposed a simple version of UCB, called UCB1, that represents the widely used UCB version. The importance of UCB1 can be justified by the fact that this version achieves a trade-off between the optimality and the complexity. Similarly to TS, in UCB1 each channel has assigned an index $B_i(t, T_i(t))$:

$$B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t)) \tag{3}$$

where $X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{j=1}^{t} r_i(j)$ represents the exploitation factor[1] and $A_i(t, T_i(t)) = \sqrt{\frac{2\ln(t)}{T_i(t)}}$ stands for the exploration factor.

Beside UCB1, ϵ-greedy represents another simple and important algorithm, firstly proposed in [5]. According to a recent version of ϵ-greedy proposed in [10], the user selects a random channel if χ (a random variable $\in$ [0,1]) $< \epsilon_t$ = min $\{1, \frac{H}{t}\}$ and H is a constant number, else SU selects the best channel with the highest availability probability. Recent works related to this topic are based on either UCB1 or ϵ-greedy [14]–[19]. In the next section, we propose a novel learning algorithm called ϵ-UCB for a single user, then we extend the proposed algorithm to deal with the case of multi-users for the priority access.

III. Distributed Learning and Access Algorithms

A. ϵ-UCB for a single user

Algorithm 1: ϵ-UCB for a single user

```
Initialization
for t = 1 to K do
    SU senses each channel once,
    SU updates $B_i(t, T_i(t))$, $X_i(T_i(t))$, $A_i(t, T_i(t))$,
for t = K+1 to n do
    if $\chi < \epsilon_t$, then
        $\beta(t) = \arg\max_i \; B_i\big(t-1, T_i(t-1)\big)$,
    else
        $\beta(t) = \arg\max_i \; X_i\big(T_i(t-1)\big)$,
    $i = \beta(t)$,
    $T_i(t)++$,
    SU updates $B_i(t, T_i(t))$, $X_i(T_i(t))$, $A_i(t, T_i(t))$
```

In order to estimate the availability of channels, all MAB algorithms such as UCB1 or ϵ-greedy have two phases that can be overlapped: exploration and exploitation. The exploration phase has an important role during the learning period but it becomes less important with an increasing number of slots. However, UCB1 gives the same importance to this factor at each time slot up to n by using the exploration factor $A_i(t, T_i(t))$. On the other hand, in the case of ϵ-greedy, the exploration phase can be achieved if $\chi < \epsilon_t$ where the user selects a random channel. Subsequently, the user may select many worst channels in the learning phase, i.e. $\chi < \epsilon_t$. Moreover, by selecting random channels in the multi-user case, several transmission periods can be lost due to the large number of collisions among users. Our ϵ-UCB can solve the limitation of UCB1 and ϵ-greedy, where the user accesses the channel that has the highest index $B_i(t, T_i(t))$ if $\chi < \epsilon_t$, otherwise the user selects the best channel with the highest availability probability $X_i(T_i(t))$ (see algorithm 1). In the next step, we show that the regret of our ϵ-UCB has a logarithmic asymptotic behavior. Thus, after a finite number of iterations, the user may converge to the best channel, μ_1. The regret in eq. (1) for a single user can be expressed as follows:

$$R(n, \beta) = \sum_{i=1}^{K} E\big[T_i(n)\big]\Delta_i \tag{4}$$

where $E[.]$ is the expectation and $\Delta_i = (\mu_1 - \mu_i)$. Normally, the user senses the channel i once during the initialization stage and every time $\beta(t) = i$; therefore, $T_i(n)$ can be expressed as follows:

$$T_i(n) = 1 + \sum_{t=K+1}^{n} \mathbb{1}_{\{\beta(t)=i\}} \tag{5}$$

where the logic operator $\mathbb{1}_{\{\beta(t)=i\}}$ equals 1 if $\beta(t) = i$ and 0 otherwise. Suppose that SU senses at least l times each channel up to n; Then, according to (5), $T_i(n)$ should be bounded as follows:

$$T_i(n) \leq l + \sum_{t=K+1}^{n} \mathbb{1}_{\{\beta(t)=i; T_i(t-1) \geq l\}} \tag{6}$$

[1] $X_i(T_i(t))$ can also be the expected reward; Where $r_i(j) = 1$ if the channel i is free at an instant j, $1 \leq j \leq t$, and 0 otherwise.

In our algorithm, the user may access the i^{th} worst channel during the exploration or exploitation phases. Subsequently, we define respectively M_i and N_i as the event that the user accesses the i^{th} worst channel in the exploration and the exploitation phases, and let $\mathbb{C}$ be the event that $T_i(t-1) \geq l$. In this case, $T_i(n)$ can be expressed as follows:

$$T_i(n) \;\leq\; l + \sum_{t=K+1}^{n} \mathbb{1}_{\{M_i(t);\mathbb{C}\}} + \sum_{t=K+1}^{n} \mathbb{1}_{\{N_i(t);\mathbb{C}\}} \quad (7)$$

The summation argument in the above equation follows the Bernoulli's distribution $\big($i.e. $E\{X\} = p\{X = 1\}$ where X is a random variable in $\{0, 1\}\big)$. In this case, the expectation of $T_i(n)$ should satisfy the following constraint:

$$E[T_i(n)] \leq l + \sum_{t=K+1}^{n} \underbrace{p\{M_i(t);\mathbb{C}\}}_{\mathbb{A}} + \sum_{t=K+1}^{n} \underbrace{p\{N_i(t);\mathbb{C}\}}_{\mathbb{B}} \quad (8)$$

Concerning the probability $\mathbb{A}$ and based on our algorithm, the user selects the i^{th} channel during the exploration phase if at $(t-1)$, $B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1))$. Subsequently, $\mathbb{A}$ can be expressed as follows:

$$\mathbb{A} = p\big\{\chi < \epsilon_t; B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1)); \mathbb{C}\big\}$$

The event $\chi < \epsilon_t$ is independent of the selection procedure. Then, we obtain:

$$\mathbb{A} = \epsilon_t \times p\big\{B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1)); \mathbb{C}\big\}$$

In order to find an upper bound of $\mathbb{A}$, we previously proved [20] that l should be higher than $\frac{4\alpha \ln n}{\Delta_i^2}$ where α is the exploitation-exploration factor. Using similar steps to our previous work in [20], we may obtain:

$$p\big\{B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1)); \mathbb{C}\big\} \leq 2t^{-2\alpha+2}$$

Then, we get:

$$\mathbb{A} \leq 2H \times t^{-2\alpha+1}$$

According to the theorem of Cauchy [21], a serie of the form $\sum_{t=1}^{n} t^{-2\alpha+1}$ can converge if $\alpha > 1$. Let $\alpha = 2$ (in order to achieve a balance between the exploration-exploitation phases), then we get:

$$\sum_{t=K+1}^{n} \mathbb{A} \leq 2H \times \sum_{t=1}^{n} t^{-3} \leq \frac{\pi^2 H}{3}$$

To find an upper bound of $\mathbb{B}$ in eq. (8), the user selects the i^{th} channel during the exploitation phase if at $t-1$, $X_i(T_i(t-1)) \geq X_1(T_1(t-1))$. Then, we obtain:

$$\begin{aligned}\mathbb{B} =& p\big\{\chi \geq \epsilon_t; X_i(T_i(t-1)) \geq X_1(T_1(t-1)); \mathbb{C}\big\}\\ =& \big(1-\epsilon_t\big) \times p\big\{X_i(T_i(t-1)) \geq X_1(T_1(t-1)); \mathbb{C}\big\} \quad (9)\end{aligned}$$

The second term in the above equation can be expressed as follows:

$$p\big\{X_i(T_i(t-1)) \;\geq\; X_1(T_1(t-1)); \mathbb{C}\big\} \;\leq\; Y + Z \quad (10)$$

where $Y = p\big\{X_i(T_i(t-1)) \geq a; \mathbb{C}\big\}$, $Z = p\big\{X_1(T_1(t-1)) \leq a; \mathbb{C}\big\}$, and a is a constant number that can be chosen as follows: $a = \frac{\mu_1+\mu_i}{2} = \mu_1 - \frac{\Delta_i}{2} = \mu_i + \frac{\Delta_i}{2}$. We start with the first term in eq. (10):

$$\begin{aligned}Y =& \sum_{y=l}^{n} p\big\{X_i(T_i(t-1)) \geq \mu_i + \frac{\Delta_i}{2}; T_i(t-1) = y\big\}\\ \leq& \sum_{y=l}^{n} p\big\{X_i(y) \geq \mu_i + \frac{\Delta_i}{2}\big\}\\ \leq& \sum_{y=l}^{n} \exp^{-\frac{2\Delta_i^2 y^2}{4y}} \;\text{(Using the Chernoff-Hoeffding in [22])}\\ \leq& n \exp^{\frac{-l\Delta_i^2}{2}} \leq n \exp^{-4\ln n} = \frac{1}{n^3} \;\big(\text{Using } l = \frac{8 \ln n}{\Delta_i^2}\big)\end{aligned} \quad (11)$$

Using the same steps for Z, we obtain $Z = \frac{1}{n^3}$. Finally, $E[T_i(n)]$ can be upper bounded by:

$$E[T_i(n)] \leq \frac{8 \ln n}{\Delta_i^2} + \frac{\pi^2 H}{3} + \frac{2}{n^3} \quad (12)$$

B. ϵ-UCB for the Multi-Priority Access

Based on our policy, All-Powerful Learning (APL), proposed in [23], we extend ϵ-UCB to consider multiple SUs (see algorithm 2). According to APL, each user has a fixed rank, $k \in \{1, ..., U\}$, and his target remains to access the k^{th} best channel. In addition, we consider the competitive priority access where users selfishly estimate the availability probability of channels. In a classical priority access, the target of SU_1 (i.e. the highest priority user) is to access the best channel, μ_1, at each time slot. As the second priority user

Algorithm 2: ϵ-UCB for multiple users

```
Parameters: $k$, $\xi_k(t)$,
$k$: indicates the $k^{th}$ user,
$\xi_k(t)$: indicates a collision for the $k^{th}$ user at time $t$,
Initialization
for $t$ = 1 to $K$ do
    $SU_k$ senses each channel once,
    $SU_k$ updates $B_{i,k}(t)$, $X_{i,k}(t)$, $A_{i,k}(t)$,
    $SU_k$ generates a rank in the set $\{1, ..., k\}$,
for $t$ = $K$+1 to $n$ do
    if $\chi_k < \epsilon_t$ then
        $SU_k$ senses a channel in his index $B_{i,k}(t)$
          according to his rank,
        if $\xi_k(t) = 1$, then
            $SU_k$ regenerates his rank in the set
              $\{1, ..., k\}$,
        else
            $SU_k$ keeps his previous rank,
    else
        $SU_k$ senses the channel that has the $k^{th}$
          expected of reward,
    $SU_k$ updates $B_{i,k}(t)$, $X_{i,k}(t)$, $A_{i,k}(t)$,
```

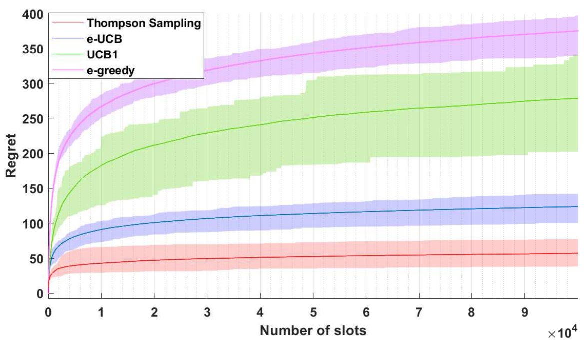

Fig. 1. ϵ-UCB compared to TS, UCB1 and ϵ-greedy

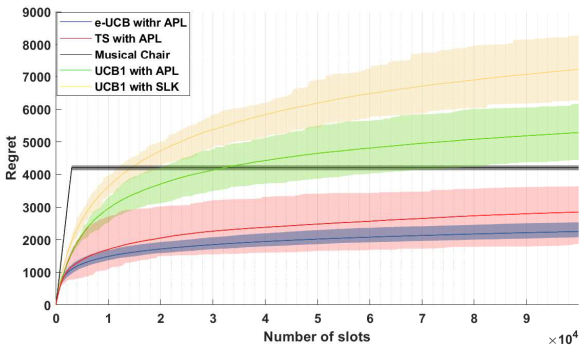

Fig. 2. ϵ-UCB, TS, UCB1 with APL compared to SLK and Musical Chair

SU_2 should avoid the best channel and try to access the second best channel. To reach his goal, SU_2 should find the two best channels at the same time in order to compute their availabilities, i.e. μ_1 and μ_2, and then access the second best channel when possible. However, this approach is a costly and impractical method to settle down each user to his dedicated channel. Our algorithm ϵ-UCB based on APL can solve this problem by making each user generate a rank around his prior rank if χ_k (a random variable generated by the k^{th} user) $< \epsilon_t$ to have information about the channels availability. In this case, SU_k can scan the k best channels and his target is the k^{th} best one. However, if the generated rank of SU_k is different than k then he accesses a channel of the set $\mu_1, \mu_2, ..., \mu_{k-1}$ and he may collide with a higher priority user, i.e. $SU_1, SU_2, ..., SU_{k-1}$. After each collision, the user can regenerate his rank to access his assigned channel; Otherwise, he retains his rank. On the other hand, when $\chi_k > \epsilon_t$, the k^{th} user may have a good estimation about the availability probability and he should access the channels according to his rank, i.e. the k^{th} best channel.

IV. Simulation and Results

In our simulation, we consider two main scenarios: In the first one, a SU tries to learn the vacancy of channels in order to access the best one. We investigate the performance of ϵ-UCB to the well-known MAB algorithms: TS, UCB1 and ϵ-greedy. In the second scenario, we consider 4 SUs trying to learn collectively the vacancy of channels with a low number of collisions. Based on our APL policy [23], we extend the well-known MAB algorithms TS, UCB1 and ϵ-greedy to consider the case of multiple users in order to compare their performance with our proposed ϵ-UCB.

Let us initially consider that a SU accessing channels with the following availability probabilities:

$$\Gamma = [0.9 \;\; 0.79 \;\; 0.73 \;\; 0.65 \;\; 0.54 \;\; 0.41 \;\; 0.32 \;\; 0.22 \;\; 0.11]$$

and trying to reach the best channel, i.e. $\mu_1 = 0.9$. Fig. 1 represents the regret of ϵ-UCB compared to the ones of TS, UCB1 and ϵ-greedy over 1000 Monte Carlo runs. The simulation outcomes are presented with a shaded region enveloping the average regret. As we can see, the regret of the 4 MAB algorithms TS, ϵ-UCB, UCB1 and ϵ-greedy has a logarithmic asymptotic behavior with respect to the number of slots. Moreover, for 1000 simulations, ϵ-UCB produces a lower regret compared to UCB1 and ϵ-greedy while TS achieves the best performance. According to many recent works, TS seems to exceed the performance of the state-of-the-art MAB algorithms. Its performance is widely suggested for a single user and several studies found an upper bound for its optimal regret. Despite its optimal convergence to the best channel for a single user, TS may not achieve a better result for multiple users as shown in Fig. 2. In fact, in the multi-user case, the performance for a given MAB algorithm does not only depend on the access of worst channels but also on the number of collisions among users. The access of worst channels and the number of collisions are mainly related to the exploration impact. Similarly, the effect of the exploration factor should be restricted, as do our proposed ϵ-UCB, after the learning period, where the user collects sufficient information about the availability of channels. While in the case of TS, the exploration factor still has the same impact all the time, which basically produces a large number of collisions compared to ϵ-UCB. That explains why, for multiple users using APL, ϵ-UCB attains a lower regret and gives a good result compared to TS. The same figure also evaluates the performance of APL compared to SLK (Selective learning of the K-th largest expected rewards) [24] and Musical Chair [25] two existing policies in the literature to manage a seondary network. SLK, as APL, takes into consideration the priority access while Musical Chair is proposed for the random access. Moreover, SLK is based on the UCB1 algorithm, and can be used only under UCB1. While, APL can be used with any learning algorithm. Fig. 2 shows that APL achieves a better result compared to SLK and Musical Chair while achieving a lower regret

Fig. 3 depicts the percentage of times to access the best channels by each SU using our policy APL. After estimating the availabilities of the communication channels, and based on APL, the targets of users SU_1, SU_2, SU_3 and SU_4 are the 4 best channels: $\mu_1 = 0.9$, $\mu_2 = 0.8$, $\mu_3 = 0.7$ and $\mu_4 = 0.6$ respectively. The percentage of times that the user SU_k accesses successfully its dedicated channel up to n using our

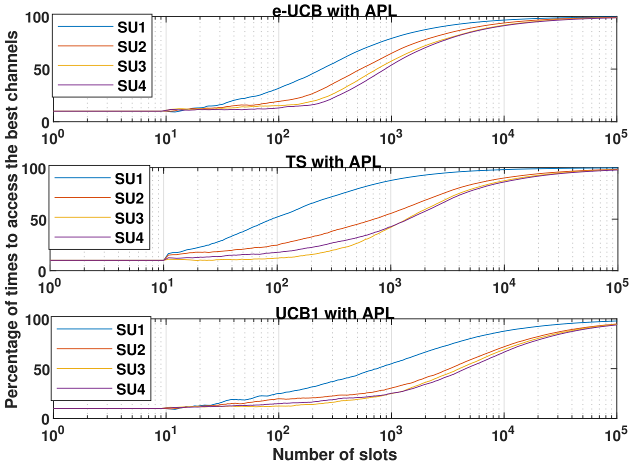

Fig. 3. The percentage of times where each SU_k selects its optimal channel using ϵ-UCB, TS, UCB1 based on the policy APL

algorithm APL is defined as follows:

$$P_k(n) = \frac{1}{n}\sum_{t=1}^{n} 1_{(\text{if } \beta^l_{APL}(t)=k)} \tag{13}$$

According to Fig. 3, the users may reach their dedicated channels quickly using ϵ-UCB compared to TS or UCB1.

V. Conclusion

In this study, we propose a novel Multi-Armed Bandit (MAB) algorithm called ϵ-Upper Confidence Bound (ϵ-UCB) to find the best channel in the context of cognitive radio. We have proven that the upper bound of regret of ϵ-UCB has a logarithmic behavior. In the context of several SUs, our main objective is to learn collectively the available spectrum and decrease the number of collisions among users. For this reason, we extend ϵ-UCB to consider the priority competitive access where the k^{th} user should access the k^{th} best channel. Our simulations compare the performance of ϵ-UCB algorithm to the most known MAB algorithms, such as: Thompson Sampling (TS), UCB1 and ϵ-greedy for a single or multiple users. It has shown that TS represents an optimal solution for a single SU but not necessary for multiple SUs. We should notice that the priority access is not widely considered in the literature, meanwhile SLK is one of the rare algorithms with priority access and it is based on UCB1 algorithm. As future work, we plan to undertake more comprehensive simulations based on ϵ-UCB for a single or multiple users. We will also investigate the analytical upper bound of ϵ-UCB for the multi-user case.

References

[1] J. Mitola and G.Maguire, "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13–18, 1999.

[2] C. Cordeiro, K. Challapali, D. Birru, and S. Shankar, "IEEE 802.22: the first worldwide wireless standard based on cognitive radios," in *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, Baltimore, USA, November 2005.

[3] W. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3, pp. 285–294, 1933.

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

[5] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, University of Cambridge, 1989.

[6] S. L. Scott, "A modern bayesian look at the multi-armed bandit," *Applied Stochastic Models in Business and Industry*, vol. 26, no. 6, pp. 639–658, 2010.

[7] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," *Advances in neural information processing systems*, Granada Spain, December 2011.

[8] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *conf. on Learning Theory*, Edinburgh, Scotland, June 2012.

[9] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *International conf. on Algorithmic Learning Theory*, Lyon, France, October 2012.

[10] 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.

[11] E. Kaufmann, O. Cappé, and A. Garivier, "On bayesian upper confidence bounds for bandit problems," in *Artificial intelligence and statistics*, La Palma, Canary Islands, April 2012.

[12] O.-A. Maillard, R. Munos, and G. Stoltz, "A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences," in *Annual conf. On Learning Theory*, Budapest, Hungary, July 2011.

[13] G. Burtini, J. Loeppky, and R. Lawrence, "A survey of online experiment design with the stochastic multi-armed bandit," *arXiv preprint arXiv:1510.00757*, 2015.

[14] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," in *Global Communications Conference (GLOBECOM)*, Texas, USA, December 2011.

[15] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune, "Distributed algorithm to learn osa channels availability and enhance the transmission rate of secondary users," in *International Symposium on Communications and Information Technologies (ISCIT)*, Ho Chi Minh, Vietnam, September 2019.

[16] M. Almasri, A. Mansour, C. Moy, A. Asoum, C. Osswald, and D. Lejeune, "Opportunistic spectrum access in cognitive radio for tactical network," in *European Conference on Electrical Engineering and Computer Science*, Bern, Switzerland, December 2018.

[17] O. Avner and S. Mannor, "Concurrent bandit and cognitive radio networks," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, Nancy, France, September 2014.

[18] N. Modi, P. Mary, and C. Moy, "Qos driven channel selection algorithm for cognitive radio network: Multi-user multi-armed bandit approach," *IEEE Transactions on Cognitive Communications & Networking*, vol. 3, no. 1, pp. 1–6, 2017.

[19] C. Tekin and L. Mingyan, "Online learning in opportunistic spectrum access: A restless bandit approach," in *International Conference on Computer Communications (INFOCOM)*, Shanghai, China, April 2011.

[20] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune, "Distributed algorithm under cooperative or competitive users with priority access in cognitive networks," *EURASIP journal on wireless communications and networking*, (Accepted).

[21] A. Cauchy, "Sur la convergence des séries," *Oeuvres completes Sér 2*, vol. 2, pp. 267–279, 1889.

[22] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.

[23] M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald, and D. Lejeune, "All-powerful learning algorithm for the priority access in cognitive network," in *European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, September 2019.

[24] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," in *Global Communications Conference (GLOBECOM)*, Texas, USA, December 2011.

[25] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits-a musical chairs approach," in *International Conference on Machine Learning (ICML)*, New York, USA, June 2016.