Distributed Competitive Decision Making Using Multi-Armed Bandit Algorithms

Mahmoud Almasri, Ali Mansour, Christophe Moy, Ammar Assoum, Denis Le Jeune, Christophe Osswald

Noname manuscript No.
(will be inserted by the editor)

Distributed Competitive Decision Making Using Multi-Armed Bandit Algorithms

Mahmoud Almasri[1] · **Ali Mansour**[1] ·
Christophe Moy[2] · **Ammar Assoum**[3] ·
Denis Lejeune[1] · **Christophe Osswald**[1]

Received: date / Accepted: date

Abstract This paper tackles the problem of the Opportunistic Spectrum Access (OSA) in the Cognitive Radio (CR). The main challenge of a Secondary User (SU) in OSA is to learn the availability of existing channels in order to select and access the one with the highest vacancy probability. To reach this goal, we propose a novel Multi-Armed Bandit (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 by the selection of worst channels. We corroborate with simulations that the regret of the proposed algorithm has a logarithmic behavior. The last statement means that within a finite number of time slots, the SU can estimate the vacancy probability of targeted channels in order to select 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 algorithms for a single or multi-user cases compared to the existing MAB algorithms.

Keywords Opportunistic Spectrum Access · Cognitive Networks · Multi-Armed Bandit · Single or Multi-Users · Priority Access

1 Introduction

Opportunistic Spectrum Access (OSA) in the Cognitive Radio (CR) has been seen as a suitable solution to improve the spectrum efficiency. Several studies, initiated by the Federal Communications Commission (FCC) in the United States (US), have recently shown that the frequency bands are not well used. On the one hand, the demand of mobile and wireless networks, have experienced unprecedented advancement since 1990's, which makes the frequency

[1] LABSTICC, UMR 6285 CNRS, ENSTA Bretagne, 2 rue F. Verny, 29806 Brest, France.
[2] Univ Rennes, CNRS, IETR - UMR 6164, F-35000, Rennes, France.
[3] Faculté des sciences, Université Libanaise, Tripoli, Lebanon.

bands more and more crowded. On the other hand, many simulations conducted in several regions in the US showed that 60 % of the frequency bands are not well used [1].
In OSA, two categories of users are considered: a Primary User (PU) who has an exclusive license to use its frequency band at any time, and a Secondary User (SU), unlicensed user, who can access the frequency band in an opportunistic manner. At each time slot, the SU tries to search and access the vacant frequency let unused by the PU without causing any harumful intereference to the latter. Due to the hardware limitation, the high cost of energy detection and the delay, a SU is not able to sense all available frequency bands at the same time. Then, under a constraint detection (one channel/slot), a SU should decide which channel to select and transmit its data once the targeted channel is free. Otherwise, it should wait for another slot to choose another channel.

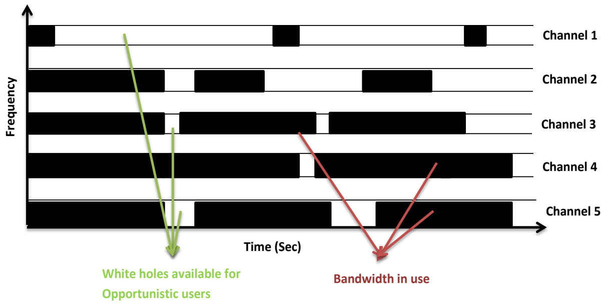

Fig. 1: Vacancy of licensed channels

In our work, we seek to enhance the licensed spectrum efficiency by helping the SU to find the best channel with the highest availability probability (i.e. the first channel in Fig.1). Indeed, this channel, on the one hand, can increase the transmission time and rate of the SU. On the other hand, accessing the best channel in the long-term may decrease the interference with the PU because this channel is not often used by this latter. In order to learn the vacancy probability of channels and thus reach the best one, we formulate our problem as the Multi-Armed Bandit (MAB) problem. After this formulation, we propose a novel MAB algorithm called e-Upper Confidence Bound (e-UCB) in order to help a SU making a decision. Hereinafter, we extend the proposed MAB algorithm to consider the multiple user case in which a policy called ALL-Powerful Learning (APL) is proposed. With a priority access, APL allows the SUs to learn collectively the vacancy probability of channels without any cooperation or prior knowledge about the vacancy probability of channels.

2 Multi-Armed Bandit Problem

Due to its generic nature, the MAB problem attracts more and more the attention in several fields including wireless channel access, jamming communication or object tracking. The classical MAB problem can be expressed as follows: An agent pulls one arm at each time slot and receives a fixed reward trying to maximize its long-term payoff. The main goal of the agent is to find the best arm with the highest expected reward. At each time, the agent has the choice to pull the current best arm (exploitation) or to try another arm in order to gain more (exploration). This problem is referred to the exploration-explotoitation dilemma in the MAB problem. Generally, a given policy can be considered as optimal when it balances between the exploration and the exploitation phases.

2.1 MAB Algorithms for a Single Agent

To solve the exploration-exploitation dilemma, different learning algorithms have been suggested in the literature, such as: e-greedy [2], Upper Confidence Bound (UCB) [3], Thompson Sampling (TS) [4], EXP3 [5], etc. Recently, the OSA is formulated as a MAB problem in order to help a SU makes a decision[1]. Subsequently, the MAB algorithms can be considered as a suitable solution for OSA. Indeed, the MAB algorithms are widely used to learn the vacancy probabilities of licensed channels according to the behavior of primary users. One of the simplest algorithm is referred to e-greedy firstly proposed in [2]. A recent version of e-greedy is proposed in [6] in order to achieve a better performance compared to several previous versions (see algorithm 1). Like several MAB algorithms, e-greedy contains two phases completly separated: exploration and exploitation. During the exploration phase, the user chooses a random channel in order to learn the vacancy probability of channels, while in the exploitation phase, the user usually selects the channel with the highest expected reward $X_i(T_i(t))$. The authors of [6] have also investigated the analytical convergence of the e-greedy and it have shown that the regret (i.e. the loss of reward by selecting worse channels) achieves a logarithmic asymptotic behavior. Upper Confidence Bound (UCB) represents the widely mentionned MAB algorithm in the literature and was first introduced in [3]. Several versions of UCB have been proposed to achieve a better performance compared to the classical one, such as: UCB1, UCB2, UCB-tuned, Bayes-UCB, Kullback-Leibler UCB (KL-UCB) [6–9]. In [6], the

[1] A SU in the context of OSA can be considered as an agent in the classic MAB problem, and the frequency channels become equivalent to different arms

authors proposed a simple and widely used version of UCB, called UCB1.

Algorithm 1: *e*-greedy Algorithm

```
Input: $C, H, n,$
$C$: number of channels,
$H$: exploration constant,
$n$: total number of slots,
Parameters: $T_i(t),$
$T_i(t)$: number of times the channel is sensed up to time $t$,
$\chi$: a uniform random variable in [0,1],
Output: $X_i(T_i(t)),$
$X_i(T_i(t))$: the expected reward depends on $T_i(t),$
foreach $t = 1$ to $n$ do
    if $\chi < min\{1, \frac{H}{t}\}$ then
        SU makes a random action $a_t,$
    else
        $a_t = \max_i X_i(T_i(t)),$
        $T_i(t) + +,$
        $X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{\tau=1}^{t} S_i(\tau),$
        % $S_i(\tau)$ is the observed state from channel $i$ at $\tau,$
        % $S_i(\tau)$ = 1 if the $i^{th}$ channel is vacant and 0 otherwise,
```

The importance of UCB1 can be justified by the fact, that this version achieves a trade-off between the optimality and the complexity. In UCB1, each channel has assigned an index $B_i(t, T_i(t)))$ and at each time slot the user selects the channel with the highest index (see algorithm 2). The index $B_i(t, T_i(t))$ consists basically of two important factors: $X_i(T_i(t))$ and $A_i(t, T_i(t)))$ that represent respectively the exploitation (or the expected reward) and the exploration phases. The assigned index of each channel may be defined as follows:

$$B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t)) \tag{1}$$

where the exploitation and the exploration factors can be expressed as:

$$X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{j=1}^{t} r_i(j) \tag{2}$$

$$A_i(t, T_i(t)) = \sqrt{\frac{\alpha \ln(t)}{T_i(t)}} \tag{3}$$

Algorithm 2: UCB1 Algorithm

```
Input: $\alpha, C, n$,
$\alpha$: exploration-exploitation factor,
$C$: number of channels,
$n$: total number of slots,
Parameters: $T_i(t), X_i(T_i(t)), A_i(t, T_i(t))$,
$T_i(t)$: number of times the $i^{th}$ channel is sensed up to $t$,
$X_i(T_i(t))$: the exploitation contribution of $i^{th}$ channel,
$A_i(t, T_i(t))$: the exploration contribution of $i^{th}$ channel,
Output: $B_i(t, T_i(t))$,
$B_i(t, T_i(t))$: the index assigned for $i^{th}$ channel,
foreach $t = 1$ to $C$ do
    SU senses each channel once,
    SU updates its index $B_i(t, T_i(t))$,
foreach $t = C + 1$ to $n$ do
    $a_t = \arg\max_i B_i(t-1, T(t-1))$,
    $T_i(t) + +$,
    $X_i(T_i(t)) = \frac{1}{T_i(t)} \sum_{\tau=1}^{t} S_i(\tau)$,
    % $S_i(\tau)$ is the observed state from channel $i$ at $\tau$,
    % $S_i(\tau)$ = 1 if the channel $i$ is vacant and 0 otherwise,
    $A_i(t, T_i(t)) = \sqrt{\frac{\alpha \ln(t)}{T_i(t)}}$,
    $B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t))$,
```

After a certain number of time slots, the exploitation factor of the i^{th} channel $X_i(T_i(t))$ will be very close to its availability, μ_i. On the other hand, the importance of the exploration factor is to learn the vacancy probability of channels by reinforcing the algorithm to examinate the states of all available channels.

At the initialization period (i.e. from 1 up to the number of channels C), the user selects each channel once to have a prior information about the availability of channels. After this period, the user selects the channel that has the highest index $B_i(t, T_i(t))$:

$$a_t = \arg\max_i B_i(t, T_i(t)) \tag{4}$$

In [6], the authors suggest the upper bound of the sum of regret (i.e. the loss of reward by selecting the worst channels) and they also show that the regret achieves a logarithmic asymptotic behavior. Subsequently, after a finite number of time slots, the user distinguishes the optimal channel and can then always select this channel.

Thompson Sampling algorithm (TS), another important MAB algothim, represents one of the earliest MAB algorithms [4]. As in UCB1, 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 $B_i(t, T_i(t))$:

$$B_i(t, T_i(t)) = \frac{W_i(t, T_i(t)) + a}{W_i(t, T_i(t)) + Z_i(t, T_i(t)) + a + b} \tag{5}$$

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 compared to other MAB algorithms [7,10,11], TS was till recently largely ignored in the literature by the Machine learning community. The rejection of TS is due to the fact that this algorithm is proposed without any analytical proof. Recently, TS attracts more and more the attention and several works investigated the proof of its convergence [12,13,25].

Algorithm 3: Thompson Sampling Algorithm

```
Input: $C, n$,
$C$: number of channels,
$n$: total number of slots,
Parameters: $S_i(t)$, $T_i(t), W_i(t, T_i(t)), Z_i(t, T_i(t))$,
$S_i(t)$: state of the selected channel, equal one if the channel is free and
  0 otherwise,
$T_i(t)$: number of times the $i^{th}$ channel is sensed by SU,
$W_i(t, T_i(t))$: success access of the $i^{th}$ channel,
$Z_i(t, T_i(t))$: failure access of the $i^{th}$ channel,
Output: $B_i(t, T_i(t))$,
$B_i(t, T_i(t))$: index assigned for the $i^{th}$ channel,
foreach $t = 1$ to $n$ do
    $a_t = \arg\max_i B_i(t, T_i(t))$,
    Observe the State $S_i(t)$,
    $W_i(t, T_i(t)) = \sum_{t=0}^{n} S_i(t) 1_{a_t = i}$,
    % $1_{a_t=i}$: equals 1 if the user selects the $i^{th}$ channel and 0 otherwise,
    $Z_i(t, T_i(t)) = T_i(t) - W_i(t, T_i(t))$,
    $B_i(t, T_i(t)) = \frac{W_i(t,T_i(t))+a}{W_i(t,T_i(t))+Z_i(t,T_i(t))+a+b}$
```

After a finite number of time slots, the vacancy probability of the i^{th} channel μ_i becomes very close to $B_i(t, T_i(t))$. By choosing the channel with the highest index, the user usually selects the optimal channel.

2.2 Multiple Agents under Random or Priority Access

In the previous section, we investigated the well-known MAB algorithms, initially proposed for a single user, to help a SU making a decision. However, in OSA, several SUs may exist in the network and a decision making, then an appropriate action taken by a SU may affect the actions and the decisions of all other users. In order to help a SU to make a good decision without harming the transmission of others, several algorithms have been proposed. These algorithms can be classified into three main categories:

1. Centralized algorithms: in which the decision is made at the network-level.
2. Semi-distributed algorithms that represent the medley among the centralized and distributed algorithms in which the decision is often made at the network-level.
3. Distributed algorithms: In this type of algorithms, a SU makes its own action in a competitve manner without interacting with others. Although the competitve access can increase the number of collisions among users but also the complexity of the system will decrease compared to the cooperative access. Indeed, in the latter case, the decision made by a SU not only depends on its own observations, but also on the decision of other users. So that, the users would interact with each other in order to make a good decision.

In our work, we focus on the competitve access in which a novel distributed algorithm for the piority access is proposed. By taking into account the priority access, the authors of [14] proposed the Selective Learning of the k^{th} largest expected reward (SLK) policy, based on UCB1. SLK has been shown as an efficient policy for the priority access. However, the number of users must be fixed and known by each user. Similarly to SLK, the authors of [15] proposed the $kth - MAB$ for the priority access, where each user has a prior knowledge about its prior rank. In $kth - MAB$, the time is slotted and each slot is divided into multi sub-slots depending on the users priority ranks. For instance, the slot of SU_U is divided into U sub-slots in order to find the U^{th} best channel and transmit its own data via this channel. Therefore, for a large number of users, the transmission time is insufficient for the high ranking users, which consists a major limitation of this algorithm.

For the random access, several learning policies can be found in the literature where the SU selects randomly its channel. For instance, the authors of [16] proposed the Musical chair policy where each user selects a random channel up to a fixed time T_0 in order to estimate the channels availabilities and the number of users, U, in the network. After T_0, each user should select a random channel in the set $\{1, ..., U\}$.

3 Problem Formulation

3.1 Single User Case

First, let us consider a SU is trying to access C available channels ordered by their vacancy probabilities, i.e. $\mu_1 > \mu_2 > \ldots > \mu_C$, and let $\Gamma = (\mu_i)$ being the vacancy probability vector initially considered as an unknown to the SU. A main target of SU is to provide an information about the vector Γ in order to access the best channel μ_1. Accessing this channel increases the trasnmission time and rate of the SU. At each time slot, the SU can select one channel and transmits its data if this channel is free; otherwise, it should wait to the next slot to select another channel. Each channel can be seen in two binary states $S_i(t)$: $S_i(t) = 1$ if the selected channel is free, and 0 otherwise. Without any loss of generality, the obtained reward $r_i(t)$ from the i^{th} channel selected at any time t may equal to its binary state, i.e. $r_i(t) = S_i(t)$. Let us define the memory for storing and accessing the i^{th} channel up to time t by $T_i(t)$. We should notice that $r_i(t)$ and $T_i(t)$ have important roles to estimate the expected reward obtained from the i^{th} channel up to t.

A SU should select a given policy that can maximize its long-term expected reward. In order to evaluate the performance of a given policy β, let us define the regret corresponding to the reward loss by the selection of the worse channels as follows:

$$R(n, \beta) = n\mu_1 - E\left[\sum_{t=1}^{n} \mu_i^{\beta}(t)\right] \tag{6}$$

where n denotes the total number of slots and $\mu_i^{\beta}(t)$ represents the vacancy probability of the selected channel at time t using a policy β, and $E[.]$ stands for the mathematical expectation.

3.2 Multi-User Case

In this section, we consider U SUs trying to access licensed bands containing C channels in an opportunistic manner. Since a SU can access only one channel at each time slot, then it should make a suitable decision and choose an appropriate channel (e.g. the channel with the highest vacancy probability). When several SUs existing in the spectrum, the main challenge for them is to learn collectively the vacancy probability of channels as mush as possible in order to access the best ones. Moreover, the number of collisions among users should be under a certain limit. Then, let us define the regret for multiple users that depends not only on the access of worse channels but also on the collision among users:

$$R(n, U, \beta) = n \sum_{k=1}^{U} \mu_k - \sum_{t=1}^{n} E\left[S^{\beta}(t)\right] \tag{7}$$

where μ_k represents the vacancy probability of the k^{th} optimal channel; $S^{\beta}(t)$ represents the global reward obtained by all users at time t using a given policy β. $S^{\beta}(t)$ can be defined as follows:

$$S^{\beta}(t) = \sum_{j=1}^{U} \sum_{i=1}^{C} S_i(t) I_{i,j}(t) \tag{8}$$

where $S_i(t)$ stands for the state of the i^{th} channel at time t [2] (i.e. $S_i(t) = 1$ if the channel selected by a SU is free at time t, otherwise $S_i(t) = 0$); $I_{i,j}(t)$ indicates that no-collision occured in the i^{th} channel by the user j at time t (i.e. $I_{i,j}(t)$ = 1

[2] The variable $S_i(t)$ may also represent the reward of the i^{th} channel at slot t.

if the j^{th} user is the sole occupant of channel i and 0 otherwise). Finally, we introduce in the equation below a simple definition of the regret that can be affected by the channel occupancy and the collision among users up to time n:

$$R(n, U, \beta) = n \sum_{k=1}^{U} \mu_k - \sum_{j=1}^{U} \sum_{i=1}^{C} P_{i,j}(n) \mu_i \tag{9}$$

where $P_{i,j}(n) = \sum_{t=1}^{n} E\left[I_{i,j}(t)\right]$ represents the expectation of times when the user j is the only occupant of the channel i up to n, and the mean of reward can be given by:

$$\mu_i = \frac{1}{n} \sum_{t=1}^{n} S_i(t)$$

4 Distributed Learning *e*-UCB

In this section, we propose a novel learning algorithm called *e*-UCB, based on UCB and *e*-greedy, to tackle the OSA problem and help a SU to find an opportunity in the frequency band. Here, it is worth mentioning that the well-known MAB algorithms that address the OSA problem are based or insipred either by UCB or *e*-greedy [14, 17–21]. Hereinafter, we extend *e*-UCB to consider the multiple SUs case in which a novel comptetive policy for the priority access is proposed. This policy does not require any cooperation or prior knowledge about the vacancy probabilities of channels. The proposed policy may achieve better performance compared to recent works that consider the priority access.

4.1 Learning MAB algorithm for a Single User

In order to learn the vacancy probabilities of channels and then access the best one, all suggested MAB algorithms have two phases that can be overlapped: exploration and exploitation. For their simplicity and optimality, UCB1 and *e*-greedy are widely suggested in the literature to solve the OSA problem. Besides the optimality achieved by UCB1 and *e*-greedy compared to other MAB algorithms, the two algorithms suffer from some drawbacks. On the one hand, in UCB1 the exploration phase, $A_i(t, T_i(t))$, has the same impact over time which may cause an additional increase of the regret. Nevertheless, the importance of the exploration factor should be only during the learning phase when gathering information about the vacancy probability of channels. Thus, the big challenge is to restrict the exploration factor $A_i(t, T_i(t))$ after the learning period while giving an additionel weight to the exploitation factor $X_i(T_i(t))$. On the other hand, due to the random selection in *e*-greedy during the exploration phase, the user may access many bad channels. Subsequently, the regret may increase significantly. Due to the random access in the multi-user case, a large number of collisions may occur among users and many transmission periods will be lost. Our *e*-UCB can solve the main drawbacks of UCB1 and *e*-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 vacancy probability $X_i(T_i(t))$ (see algorithm 4).

Algorithm 4: *e*-UCB Algorithm

```
Input: $C, H, \alpha, n,$
$C$: number of channels,
$H$ and $\alpha$: represent the exploration-exploitation factors,
$n$: total number of slots,
Parameters: $\chi$, $\epsilon_t$, $T_i(t), X_i(T_i(t)), A_i(t, T_i(t))$
$\chi$: random variable in [0,1] generated at time $t$,
$\epsilon_t$: decreasing number with respect to time $\in [0, 1]$,
$T_i(t)$: number of times the $i^{th}$ channel is sensed up to $t$,
$X_i(T_i(t))$: exploitation contribution of $i^{th}$ channel,
$A_i(t, T_i(t))$: exploration contribution of $i^{th}$ channel,
Output: $B_i(t, T_i(t))$,
$B_i(t, T_i(t))$: index assigned for $i^{th}$ channel,
Initialization
foreach $t = 1$ to $C$ do
    $SU$ senses each channel once,
    $SU$ updates $B_i(t, T_i(t))$, $X_i(T_i(t))$, $A_i(t, T_i(t))$,
foreach $t = C + 1$ to $n$ do
    if $\chi < \epsilon_t$, then
        $a_t = \arg\max_i \; B_i\big(t - 1, T_i(t - 1)\big)$,
    else
        $a_t = \arg\max_i \; X_i\big(T_i(t - 1)\big)$,
    $T_i(t) + +$,
    $SU$ updates $B_i(t, T_i(t))$, $X_i(T_i(t))$, $A_i(t, T_i(t))$
```

Hereinafter, we show that the upper bound of regret of *e*-UCB achieves a logarithmic asymptotic behavior. Then, after a finite number of time slots, the user can learn the vacancy probabilities of channels and may select always the best one, μ_1.

Due to the hardware constraints, we suppose that the user can sense only one channel at each time slot, then: $\sum_{i=1}^{C} T_i(n) = n$. The regret for a single user in eq. (6) can be expressed as follows:

$$\begin{aligned} R(n,\beta) &= n\mu_1 - \sum_{i=1}^{C} E[T_i(n)]\mu_i \\ &= \sum_{i=1}^{C} E\big[T_i(n)\big]\mu_1 - \sum_{i=1}^{C} E\big[T_i(n)\big]\mu_i \\ &= \sum_{i=1}^{C} E\big[T_i(n)\big]\Delta_i \qquad (10) \end{aligned}$$

where $E[.]$ is the expectation and $\Delta_i = \mu_1 - \mu_i$.

According to e-UCB, the user selects the i^{th} channel once during the initialization phase and every time this channel has the highest index (see eq. (4)); therefore, $T_i(n)$ can be written as follows:

$$T_i(n) = 1 + \sum_{t=C+1}^{n} \mathbb{1}_{\{a_t=i\}} \qquad (11)$$

where $\mathbb{1}_{\{a_t=i\}}$ equals 1 if $a_t = i$ and 0 otherwise. Up to time n, the user may select each channel at least l times; then, according to eq. (11), $T_i(n)$ can be bounded as follows:

$$T_i(n) \leq l + \sum_{t=C+1}^{n} \mathbb{1}_{\{a_t=i;T_i(t-1)\geq l\}} \qquad (12)$$

In e-UCB, the user may select the i^{th} non-optimal channel either in the exploration or exploitation phases. Therefore, let M_i and N_i be the events that the user selects the i^{th} channel during the exploration and exploitation respectively, and let $\mathbb{D}$ be the event that $T_i(t-1) \geq l$. In this case, $T_i(n)$ can be expressed as follows:

$$T_i(n) \leq l + \sum_{t=C+1}^{n} \mathbb{1}_{\{M_i(t);\mathbb{D}\}} + \sum_{t=C+1}^{n} \mathbb{1}_{\{N_i(t);\mathbb{D}\}} \qquad (13)$$

In the above equation, the second and third terms follow the Bernoulli 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 $T_i(n)$ can be written as:

$$E[T_i(n)] \leq l + \sum_{t=C+1}^{n} \underbrace{p\{M_i(t);\mathbb{D}\}}_{\mathbb{A}} + \sum_{t=C+1}^{n} \underbrace{p\{N_i(t);\mathbb{D}\}}_{\mathbb{B}} \qquad (14)$$

According to e-greedy, the user selects the i^{th} channel during the exploration phase if at $(t-1)$, $B_i(t-1, T_i(t-1)) > B_1(t-1, T_1(t-1))$. Subsequently, $\mathbb{A}$ can be expressed as follows:

$$\mathbb{A} = p\big\{\chi < \epsilon_t; B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1)); \mathbb{D}\big\} \qquad (15)$$

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

$$\mathbb{A} = p\{\chi < \epsilon_t\} \times p\big\{B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1)); \mathbb{D}\big\} \tag{16}$$

As χ is uniformly distributed in $[0, 1]$, and $0 \leq \epsilon_t \leq 1$, then $p\{\chi < \epsilon_t\} = \epsilon_t$. So, we get:

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

Proof in appendix A

where H and α are constant numbers. According to Cauchy theorem [22], a series 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 and the exploitation phases), then we obtain:

$$\sum_{t=C+1}^{n} \mathbb{A} \leq 2H \times \sum_{t=1}^{n} t^{-3} \leq \frac{\pi^2 H}{3} \tag{17}$$

Let us seek an upper bound of $\mathbb{B}$ in eq. (14) that referred to the the probability to access the i^{th} channel during the exploitation phase. Indeed, during the exploitation phase, the user may select the i^{th} channel at time slot t, whereas $X_i(T_i(t-1)) > X_1(T_1(t-1))$ at $t-1$. Then, we get the following inequality:

$$\mathbb{B} = p\big\{\chi \geq \epsilon_t; X_i(T_i(t-1)) \geq X_1(T_1(t-1)); \mathbb{D}\big\} \tag{18}$$

As the two events $\chi \geq \epsilon_t$ and $X_i(T_i(t-1)) \geq X_1(T_1(t-1))$ are independent, we obtain:

$$\mathbb{B} = \big(1 - \epsilon_t\big) \times p\big\{X_i(T_i(t-1)) \geq X_1(T_1(t-1)); \mathbb{D}\big\} \tag{19}$$

The probability in the above equation can be bounded as follows:

$$p\big\{X_i(T_i(t-1)) \geq X_1(T_1(t-1)); \mathbb{D}\big\} \leq Y + Z \tag{20}$$

where $Y = p\big\{X_i(T_i(t-1)) \geq a; \mathbb{D}\big\}$, $Z = p\big\{X_1(T_1(t-1)) \leq a; \mathbb{D}\big\}$, and a is a constant number that can be chosen as: $a = \frac{\mu_1+\mu_i}{2} = \mu_1 - \frac{\Delta_i}{2} = \mu_i + \frac{\Delta_i}{2}$.
Let us first consider, the first term of eq. (20):

$$\begin{aligned} Y &= \sum_{y=l}^{n} p\big\{X_i(T_i(t-1)) \geq \mu_i + \frac{\Delta_i}{2}; T_i(t-1) = y\big\} \\ &= \sum_{y=l}^{n} p\big\{X_i(y) \geq \mu_i + \frac{\Delta_i}{2}; T_i(t-1) = y\big\} \\ &\leq \sum_{y=l}^{n} p\big\{X_i(y) \geq \mu_i + \frac{\Delta_i}{2}\big\} \end{aligned} \tag{21}$$

Using the Chernoff-Hoeffding theorem in [23][3], we can upper bound the above equation as follows:

$$Y \leq \sum_{y=l}^{n} \exp^{-\frac{2\Delta_i^2 y^2}{4y}} \leq n \exp^{\frac{-l\Delta_i^2}{2}} \tag{22}$$

According to the proof provided in appendix A, we consider $l = \lceil \frac{4\alpha \ln(n)}{\Delta_i^2} \rceil$ where $\alpha = 2$. So, we get:

$$Y \leq n \exp^{-4 \ln n} = \frac{1}{n^3} \tag{23}$$

The upper bound of Z can be expressed as:

$$Z \leq \frac{1}{n^3} \tag{24}$$

Proof in appendix B

Based on eq (14), (17), (19), (23) and (24), $E[T_i(n)]$ can be upper bounded by:

$$E[T_i(n)] \leq \frac{8 \ln n}{\Delta_i^2} + \frac{\pi^2 H}{3} + \frac{2}{n^3} \tag{25}$$

From the above inequation, we conclude that the user plays each arm no more than $\frac{8 \ln n}{\Delta_i^2}$ plus a constant number. Finally, based on eq (10) and (25), the regret of e-UCB, $R(n, e\text{-UCB})$, can be upper bounded by the following equation:

$$R(n, e\text{-UCB}) \leq 8 \ln n \sum_{i=2}^{C} \frac{1}{\Delta_i} + \left(\frac{\pi^2 H}{3} + \frac{2}{n^3}\right) \sum_{i=1}^{C} \Delta_i \tag{26}$$

4.2 ϵ-UCB for the Multi-Priority Access

In our work, we are interested in the priority access where the SUs should access the channels based on their priority ranks. Our goal is to ensure that the U users are accessing separately the U-best channels without any prior information on best channels.

[3] According to [23], Chernoff-Hoeffding theorem is defined as follows: Let $X_1, ..., X_n$ be random variables in $\{0, 1\}$, and $E[X_t] = \mu$, and let $S_n = \sum_{i=1}^{n} X_i$. Then $\forall\ a \geq 0$, $p\{S_n \geq n\mu + a\} \leq \exp^{\frac{-2a^2}{n}}$ and $p\{S_n \leq n\mu - a\} \leq \exp^{\frac{-2a^2}{n}}$.

Based on our policy, All-Powerful Learning (APL), proposed in [26], we extend ϵ-UCB to consider multiple SUs (see algorithm 5).

Algorithm 5: ϵ-UCB for multiple users

```
Input: $k, C, H, \alpha, n,$
$k$: indicates the $k^{th}$ user,
$C$: is the number of channels,
$H$ and $\alpha$: represent the exploration and the exploitation factors,
$n$: total number of slots,
Parameters: $\chi_k$, $\xi_k(t)$, $\epsilon_t$, $X_{i,k}(t)$, $A_{i,k}(t)$
$\chi_k$: a random variable in [0,1] generated by the $k$-th user,
$\xi_k(t)$: indicates a collision for the $k^{th}$ user at time $t$,
$\epsilon_t$: a decreasing number with respect to time $\in [0,1]$,
$X_{i,k}(t)$: the exploitation contribution of $i^{th}$ channel for the $k$-th user,
$A_{i,k}(t)$: the exploration contribution of $i^{th}$ channel for the $k$-th user,
Output: $B_{i,k}(t)$,
$B_{i,k}(t)$: the index assigned of $i^{th}$ channel for the $k$-th user,
Initialization
foreach $t$ = 1 to C 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 $\{1, ..., k\}$,
foreach $t$ = C+1 to n do
    if $\chi_k < \epsilon_t$ then
        $SU_k$ senses a channel in its index $B_{i,k}(t)$ according to its rank,
        if $\xi_k(t) = 1$, then
            $SU_k$ regenerates its rank in the set $\{1, ..., k\}$,
        else
            $SU_k$ keeps its previous rank,
    else
        $SU_k$ senses the channel that has the $k^{th}$ expected of reward,
    $SU_k$ updates $B_{i,k}(t)$, $X_{i,k}(t)$, $A_{i,k}(t)$,
```

According to APL, each user has a fixed rank, $k \in \{1, ..., U\}$, and its 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, each channel has assigned an index and the highest priority user SU_1 should sense and access the channel with the highest index, i.e. μ_1, at each time slot. Indeed, the best channel, after a finite number of time slots, will be corresponding to the highest index. As for the second priority user, SU_2, he should avoid the best channel and try to access the second best channel, μ_2. To reach its goal, SU_2 should sense the first and second best channels at each time slot in order to estimate their vacancy probabilities and then access the second best channel when available. For

the U^{th} SU, it should estimate the vacancy probability of all the first U best channels at each time slot to access the U^{th} best one. Therefore, the complexity of the hardware is increased, and we conclude that a classical priority access represents a costly and impractical method to settle down each user to its dedicated channel.
Our algorithm ϵ-UCB, based on APL, can overcome this limitation by making each user generate a rank around its prior rank if χ_k (a random variable generated by the k^{th} user) $< \epsilon_t$ in order to have information about the channels availability. In this case, SU_k can scan the k best channels and its 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 its rank to access its assigned channel; Otherwise, he retains its rank. On the other hand, after the learning period when $\chi_k > \epsilon_t$, the k^{th} user may have a good estimation about the availability probability and should access the channels according to its rank, i.e. the k^{th} best channel.

5 Simulation and Results

In our simulation, we consider three main scenarios: In the first one, a SU tries to learn the vacancy of channels in order to access the best one, i.e. that has the highest availability probability. We evaluate the performance of e-UCB with respect to well-known versions of MAB algorithm: TS, UCB1 and e-greedy. In a second scenario, we consider 4 SUs trying to learn collectively the vacancy of channels with a low number of collisions. It has been shown that, based on our policy APL, the users reach their dedicated channel faster than several existing algorithms while decreasing the total regret. Let us initially consider a SU accessing channels with the following availability probabilities:

$$\Gamma = [0.9\ \ 0.8\ \ 0.7\ \ 0.6\ \ 0.5\ \ 0.4\ \ 0.3\ \ 0.2\ \ 0.1]$$

and trying to reach the best channel, i.e. $\mu_1 = 0.9$. Fig. 2 represents the regret of e-UCB compared to TS, UCB1 and e-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, e-UCB, UCB1 and e-greedy has a logarithmic asymptotic behavior in function of the number of slots. Moreover, for 1000 simulations, e-UCB produces a lower regret compared to UCB1 and e-greedy while TS achieves the best performance.

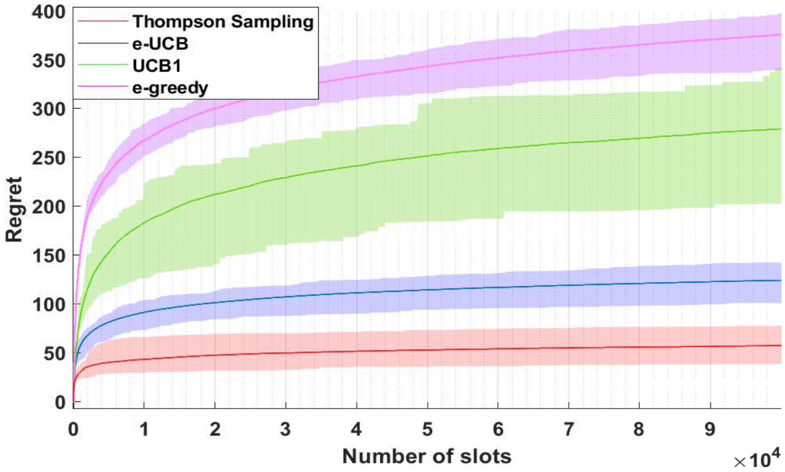

Fig. 2: *e*-UCB compared to TS, UCB1 and *e*-greedy

In our simulation, we also use the percentage accessing the best channel by the user, P_{Best}, given by:

$$P_{Best} = 100 \times \sum_{t=1}^{n} \frac{\mathbb{1}_{(a_t=\mu_1)}}{t}$$

where $\mathbb{1}_{(a=b)} = \begin{cases} 1 \text{ if } a = b \\ 0 \text{ otherwise} \end{cases}$
In Fig. 3, P_{Best} shows three parts:

1. The first part, from slot 1 to C, represents the initialization part where the SU should access each channel once in order to get some information about the vacancy of channels.
2. The second part, from slot C+1 to around 2000, represents the adaptation phase.
3. In the last part, the user converges asymptotically towards the optimal channel μ_1.

In the adaptation part, *e*-greedy achieves the worst results compared to other MAB algorithms. While in the convergence part, *e*-greedy can reach the best channel faster than UCB1. However, *e*-greedy spends more time to gather more information about the vacancy of channels in the exploration phase, then it exploits efficiently the obtained information to reach the best channel. While UCB1 seems to have a balance between its exploration-exploitation phases at any given time up to the total number of slots, n. The same figure shows that TS achieves the best performance followed by our proposed *e*-UCB, *e*-greedy, and UCB1.

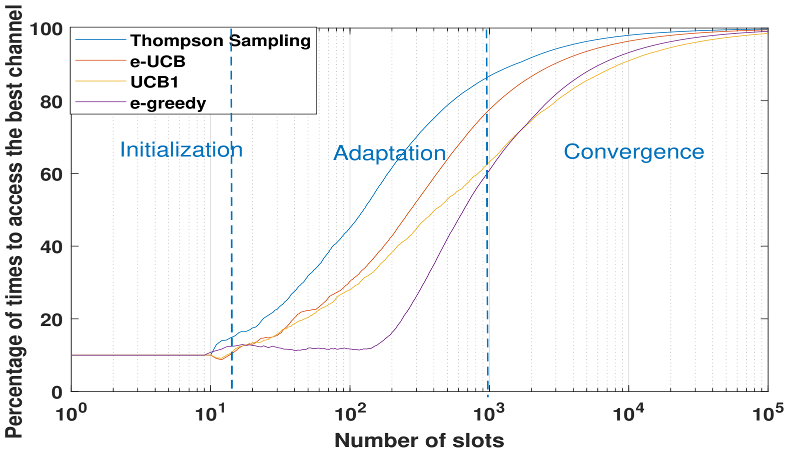

Fig. 3: The percentage of times where the user selects its optimal channel using the 4 MAB algorithms

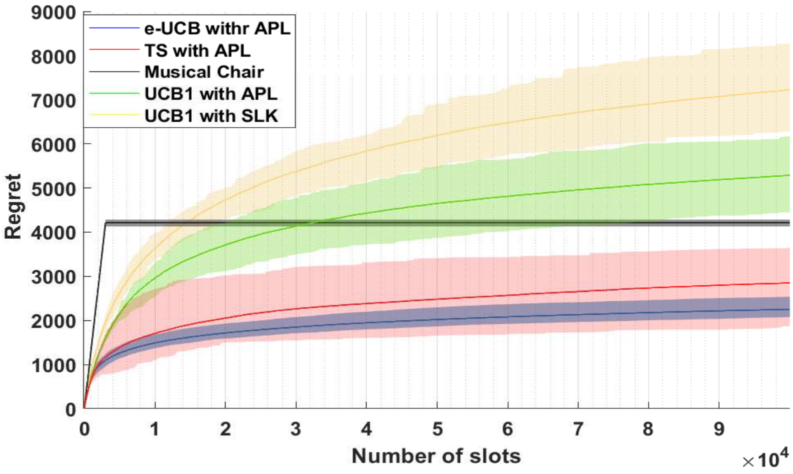

Fig. 4: *e*-UCB, TS, UCB1 with APL compared to SLK and Musical Chair

According to many recent works, the performance of TS seems to exceed that 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. 4. In fact, in the multi-user case, the performance for a given MAB algorithm not only depends on the access of worse channels but also on the number of collisions among users. The access of worse channels and the number of collisions are mainly related to the exploration impact. Similarly, the effect

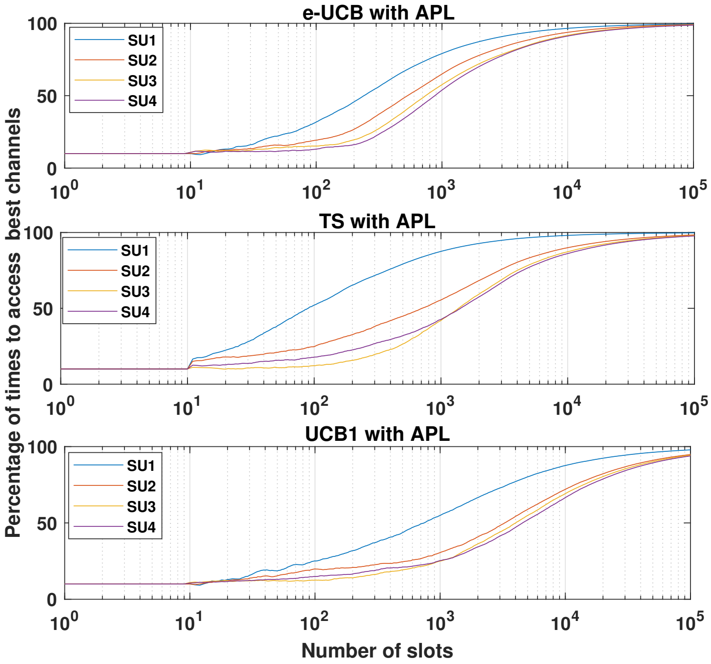

Fig. 5: The percentage of times where each SU_k selects its optimal channel using e-UCB, TS, UCB1 based on the APL policy

of the exploration factor should be restricted, as does our proposed e-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 e-UCB. That explains why, for multiple users using APL, e-UCB attains a lower regret and gives a good result compared to TS.

Fig. 5 depicts the percentage 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 to access the 4 best channels (i.e. $\mu_1 = 0.9$, $\mu_2 = 0.8$, $\mu_3 = 0.7$ and $\mu_4 = 0.6$). If two or more users access the same channel, a collision occurs and all the collided users receive zero reward. The percentage of times that the user SU_k accesses successfully its dedicated channel up to time n using our policy APL is defined as follows:

$$P_k(n) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{(a_t = k)} \tag{27}$$

According to Fig. 5, the users may reach their dedicated channels quickly compared to TS or UCB1.

6 Conclusion

In this study, we propose a novel MAB algorithm called e-Upper Confidence Bound (e-UCB) in order to help a single SU to learn Opportunistic Spectrum Access (OSA) channels. We also evaluate the performance of e-UCB compared to the well-known Multi-Armed Bandit (MAB) algorithms, such as: Thompson Sampling (TS), UCB1 and e-greedy. In order to tackle the problem of OSA with several Secondary Users (SUs), we extend e-UCB based on our novel All-Powerful Learning (APL) policy for the priority access. This policy allows several SUs to learn collectively the available channels without any cooperation or prior knowledge about the vacancy probabilities of the channels. We should notice that the priority access is not widely considered in the literature, meanwhile SLK is one of rare policy with priority access and it is based on UCB1 algorithm. We also compare the extended e-UCB to recent existing policies such as: SLK and Musical Chair. In future works, we plan to undertake more comprehensive simulations based on e-UCB and APL; we will also investigate the analytical upper bound of e-UCB with APL.

Appendix A

In this Appendix, we investigate the upper bound of the term $\mathbb{A} = \epsilon_t \times Prob$ in e-UCB where $Prob$ can be expressed as follows:

$$Prob \leq P\big\{B_i(t-1, T_i(t-1)) \geq B_1(t-1, T_1(t-1)); T_i(t-1) \geq l\big\}$$

The index of the i-th channel $B_i(t, T_i(t))$ is the sum of the exploration factor, $X_i(T_i(t))$, and the exploitation factor, $A_i(t, T_i(t))$:

$$B_i(t, T_i(t)) = X_i(T_i(t)) + A_i(t, T_i(t)) \tag{28}$$

Then, we obtain:

$$\begin{aligned} Prob \leq P\Bigg\{ & X_1(T_1(t-1)) + A_1(t-1, T_1(t-1)) \leq \\ & X_i(T_i(t-1)) + A_i(t-1, T_i(t-1)) \text{ and } T_i(t-1) \geq l \Bigg\} \end{aligned} \tag{29}$$

By taking the minimum value of $X_1(T_1(t-1)) + A_1(t-1, T_1(t-1))$ and the maximum value of $X_i(T_i(t-1)) + A_i(t-1, T_i(t-1))$ at each time slot, we can upper bound $Prob$ by the following equation:

$$Prob \leq P\left\{ \min_{0<S_1<t} \Big[X_1(S_1) + A_1(t, S_1) \Big] \leq \max_{l \leq S_i < t} \Big[X_i(S_i) + A_i(t, S_i) \Big] \right\} \tag{30}$$

where $S_i \geq l$ to fulfill the condition $T_i(t-1) \geq l$. Then, we obtain:

$$Prob \leq \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} P\left\{ X_1(S_1) + A_1(t, S_1) < X_i(S_i) + A_i(t, S_i) \right\} \tag{31}$$

The above probability can be upper bounded by:

$$\begin{aligned} Prob \leq \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} P\Big\{ X_1(S_1) + A_1(t, S_1) \leq \mu_1 \Big\} + \\ P\Big\{ \mu_1 < \mu_i + 2A_i(t, S_i) \Big\} + \\ P\Big\{ X_i(S_i) + A_i(t, S_i) \geq \mu_i + 2A_i(t, S_i) \Big\} \end{aligned} \tag{32}$$

Using the ceiling operator $\lceil \rceil$, let $l = \lceil \frac{4\alpha \ln(n)}{\Delta_i^2} \rceil$, where $\Delta_i = \mu_1 - \mu_i$ and $S_i \geq l$, then the inequality $\mu_1 < \mu_i + 2A_i(t, S_i)$ in eq (32) becomes false, in fact:

$$\begin{aligned} \mu_1 - \mu_i - 2A_i(t, S_i) &= \mu_1 - \mu_i - 2\sqrt{\frac{\alpha \ln(t)}{S_i}} \\ &\geq \mu_1 - \mu_i - 2\sqrt{\frac{\alpha \ln(n)}{l}} \\ &\geq \mu_1 - \mu_i - \Delta_i = 0 \end{aligned}$$

Based on eq (32), we obtain:

$$Prob \leq \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} P\left\{ X_1(S_1) \leq \mu_1 - A_1(t, S_1) \right\} + P\left\{ X_i(S_i) \geq \mu_i + A_i(t, S_i) \right\} \tag{33}$$

Using Chernoff-Hoeffding bound[4] [23], we can prove that:

$$\begin{aligned} P\Big\{ X_1(S_1) \leq \mu_1 - A_1(t, S_1) \Big\} &\leq \exp^{\frac{-2}{S_1}\left[S_1 \sqrt{\frac{\alpha \ln(t)}{S_1}} \right]^2} \\ &= t^{-2\alpha} \end{aligned} \tag{34}$$

$$\begin{aligned} P\Big\{ X_i(S_i) \geq \mu_i + A_i(t, S_i) \Big\} &\leq \exp^{\frac{-2}{S_i}\left[S_i \sqrt{\frac{\alpha \ln(t)}{S_i}} \right]^2} \\ &= t^{-2\alpha} \end{aligned} \tag{35}$$

[4] According to [23], Chernoff-Hoeffding theorem is defined as follows: Let $X_1, ..., X_n$ be random variables in [0,1], and $E[X_t] = \mu$, and let $S_n = \sum_{i=1}^{n} X_i$. Then $\forall\ a \geq 0$, we have $P\{S_n \geq n\mu + a\} \leq \exp^{\frac{-2a^2}{n}}$ and $P\{S_n \leq n\mu - a\} \leq \exp^{\frac{-2a^2}{n}}$.

The two inequations above and inequation (33) lead us to:

$$Prob \leq \sum_{S_1=1}^{t-1} \sum_{S_i=l}^{t-1} 2t^{-2\alpha} \leq 2t^{-2\alpha+2} \tag{36}$$

Finally, we obtain:

$$\mathbb{A} \leq \frac{H}{t} \times 2t^{-2\alpha+2} = 2H \times t^{-2\alpha+1} \tag{37}$$

Appendix B

This appendix stands for finding an upper bound of Z that contributes to finding an upper bound of e-UCB:

$$Z = p\{X_1(T_1(t-1)) \leq a; T_i(t-1) \geq l\ \} \tag{38}$$

where a is a constant number that can be chosen as follows: $a = \frac{\mu_1+\mu_i}{2} = \mu_1 - \frac{\Delta_i}{2} = \mu_i + \frac{\Delta_i}{2}$, and $\Delta_i = \mu_1 - \mu_i$. After the learning period where $T_i(t-1) \geq l$, we have: $T_1(t-1) >> T_i(t-1)$. Then Z can be upper bounded by:

$$\begin{aligned} Z \leq & p\{X_1(T_1(t-1)) \leq a; T_1(t-1) \geq l\ \} \qquad (39) \\ \leq & \sum_{z=l}^{n} p\{X_1(T_1(t-1)) \leq \mu_1 - \frac{\Delta_i}{2}; T_1(t-1) = z\} \\ \leq & \sum_{z=l}^{n} p\{X_1(z) \leq \mu_1 - \frac{\Delta_i}{2}\} \end{aligned} \tag{40}$$

Using the Chernoff-Hoeffding [23], we can upper bound the above equation as follows:

$$Z \leq \sum_{z=l}^{n} \exp^{-\frac{2\Delta_i^2 z^2}{4z}} \leq n \exp^{\frac{-l\Delta_i^2}{2}} \tag{41}$$

According to the proof provided in appendix A, we have $l = \lceil \frac{4\alpha \ln(n)}{\Delta_i^2} \rceil$ where $\alpha = 2$. So, we obtain:

$$Z \leq n \exp^{-4 \ln n} = \frac{1}{n^3} \tag{42}$$

Author's contributions

All the authors have contributed to the analytic and numerical results. All authors read and approved the final manuscript.

Availability of data and materials

The authors declare that all the data and materials in this manuscript are available from the authors.

Competing interests

The authors declare that they have no competing interests.

References

1. M. Marcus and C.J. Burtle and B. Franca and A. Lahjouji and N. McNeil, Federal Communications Commission (FCC): Spectrum Policy Task Force, ET Docket no. 02-135, (November 2002).
2. C. Watkins, Learning from delayed rewards, University of Cambridge (1989).
3. T. Lai and H. Robbins, Asymptotically efficient adaptive allocation rules, Advances in Applied Mathematics, 6, 4-22 (1985).
4. W.R. Thompson, On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, Biometrika, 25, 285-294 (1933).
5. P. Auer, N. Cesa-Bianchi, Y. Freund and R. Schapire, The nonstochastic multiarmed bandit problem, SIAM journal on computing, 32, 48-77 (2002).
6. P. Auer and N. Cesa-Bianchi and P. Fischer, Finite-time Analysis of the Multiarmed Bandit Problem, Machine Learning, 47, 235-256 (2002).
7. 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).
8. E. Kaufmann, O. Cappé and A. Garivier, On Bayesian upper confidence bounds for bandit problems, Artificial intelligence and statistics (La Palma, Canary Islands, April 2012).
9. O. Maillard, R. Munos and G. Stoltz, A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences, Annual conference On Learning Theory (Budapest, Hungary, July 2011).
10. S. Scott, A modern Bayesian look at the multi-armed bandit, Applied Stochastic Models in Business and Industry, 26, 639-658 (2010).
11. O. Chapelle and L. Li, An empirical evaluation of thompson sampling, Advances in neural information processing systems (Granada, Spain, December 2011).
12. E. Kaufmann, N. Korda and R. Munos, Thompson sampling: An asymptotically optimal finite-time analysis, International conference on Algorithmic Learning Theory (Lyon, France, October 2012).
13. S. Agrawal and N. Goyal, Further optimal regret bounds for thompson sampling, Artificial intelligence and statistics (Scottsdale, USA, April 2013).
14. Y. Gai and B. Krishnamachari, Decentralized Online Learning Algorithms for Opportunistic Spectrum Access, Global Communications Conference (Texas, USA, December 2011).
15. N. Torabi and K. Rostamzadeh and V. C. Leung, Rank-optimal channel selection strategy in cognitive networks, Global Communications Conference (California, USA, December 2012).
16. J. Rosenski and O. Shamir and L. Szlak, Multi-player bandits-a musical chairs approach, International Conference on Machine Learning (New York, USA, June 2016).
17. O. Avner and S. Mannor, Concurrent bandit and cognitive radio networks, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (Nancy, France, September 2014).
18. 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, International Symposium on Communications and Information Technologies (HoChiMinh, Vietnam, September 2019).
19. M. Almasri, A. Mansour, C. Moy, A. Assoum, C. Osswald and D. Lejeune, Opportunistic Spectrum Access in Cognitive Radio for Tactical Network, European Conference on Electrical Engineering and Computer Science (Bern, Switzerland, December 2018).
20. 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 and Networking, 3, 1-6 (2017).
21. C. Tekin and M. Liu, Online learning in opportunistic spectrum access: A restless bandit approach, International Conference on Computer Communications (Shanghai, China, April 2011).
22. A.Cauchy, Sur la convergence des séries, Oeuvres completes Sér 2, 2, 267-279 (1889).
23. W. Hoeffding, Probability inequalities for sums of bounded random variables, Journal of the American statistical association, 58, 13-30 (1963).
24. R. Kumar and S. J. Darak and A. Yadav and A. K. Sharma and R. K. Tripathi, Channel Selection for Secondary Users in Decentralized Network of Unknown Size, IEEE Communications Letters, 21, 2186-2189 (2017).

25. S. Agrawal and N. Goyal, Analysis of thompson sampling for the multi-armed bandit, Annual Conference on Learning Theory (Edinburgh, Scotland, June 2012).
26. M. Almasri and A. Mansour and C. Moy and A. Assoum and C. Osswald and D. Lejeune, All-Powerful Learning Algorithm for the Priority Access in Cognitive Network, European Signal Processing Conference (A Coruña, Spain, September 2019).