QoS driven Channel Selection Algorithm for Cognitive Radio Network: Multi-User Multi-armed Bandit Approach

Navikkumar Modi, Philippe Mary, Christophe Moy

QoS driven Channel Selection Algorithm for Cognitive Radio Network: Multi-User Multi-armed Bandit Approach

Navikkumar Modi, Philippe Mary, *Member, IEEE,* and Christophe Moy, *Member, IEEE*

***Abstract*—In this paper, we deal with the problem of opportunistic spectrum access (OSA) in infrastructure-less cognitive networks. Each secondary user (SU) Tx is allowed to select one frequency channel at each transmission trial. We assume that there is no information exchange between SUs, and they have no knowledge of channel quality, availability and other SUs actions, hence, each SU selfishly tries to select the best band to transmit. This particular problem is designed as a multi-user restless Markov multi-armed bandit (MAB) problem, in which multiple SUs collect a priori unknown reward by selecting a channel. The main contribution of the paper is to propose an online learning policy for distributed SUs, that takes into account not only the availability criterion of a band but also a quality metric linked to the interference power from the neighboring cells experienced on the sensed band. We also prove that the policy, named distributed restless QoS-UCB (RQoS-UCB), achieves at most logarithmic order regret, for a single-user in a first time and then for multi-user in a second time. Moreover, studies on the achievable throughput, average bit error rate obtained with the proposed policy are conducted and compared to well-known reinforcement learning algorithms.**

***Index Terms*—Cognitive Radio, Upper Confidence Bound, Opportunistic Spectrum Access, Machine Learning.**

I. Introduction

A. Bandit Theory and Opportunistic Spectrum Access

Opportunistic spectrum access (OSA) has emerged as an effective alternative to alleviate the spectrum scarcity issue and to improve the spectrum efficiency. In OSA, secondary users (SUs) also referred as unlicensed users identifies vacant spectrum through sensing and transmit opportunistically into the selected band without interfering with primary users (PUs). One of the main challenges of the cognitive networks is to achieve coexistence of heterogeneous SUs trying to access the same part of the spectrum. Due to resource and hardware constraints, SUs may sense only a part of the spectrum at any given time and hence they need to learn about statistics of the PUs' channels without having any *a priori* informations or very few. OSA scenario is hence generally tackled with reinforcement learning (RL) approaches whose the principle is to reward good trials and penalize bad ones.

This work has received a French government support granted to the CominLabs excellence laboratory and managed by the National Research Agency in the "Investing for the Future" program under reference No. ANR-10-LABX-07-01. The authors would also like to thank the Region Bretagne, France, for its support of this work.

N. Modi and C. Moy are with CentraleSupélec, IETR/SCEE, UMR CNRS 6164 France (e-mail: navikkumar.modi@centralesupelec.fr).

P. Mary is with INSA de Rennes, IETR, UMR CNRS 6164, France.

It has been shown in several works that multi-armed bandit (MAB) approach is particularly well adapted to tackle the OSA problem. Indeed, each SU senses, at each time step, one of the K frequency bands available in the PUs spectrum, according to a given policy, and searches for learning the best channel given a certain criterium, e.g. availability. The performance is analyzed using two criteria: convergence and regret [1]. The former is the guarantee that the policy, at the end, finishes to play *almost* always the best channel, and the latter measures the speed at which the convergence is achieved and is defined as the difference between the expected reward obtained using an infeasible ideal policy and the selected policy.

Additionally in this paper, we consider a distributed framework where there is no information exchange or prior agreement among the different SUs, thus it introduces additional challenges such as: loss in the collected reward due to collisions among the different SUs trying to access same channel, and also competition among different SUs since they all sense and access a channel with higher reward in the long run. It is necessary for the channel access policies to overcome the above challenges.

B. Related Work

Detailed discussions on MAB problems can be found in [2], [3], [4], [5]. In [6], [7], the authors have provided a lower bound on the regret of single user MAB policies and developed simple mean based index policies for independent and identical distributed (iid) rewards to achieve this bound. In [8], the authors have proposed a simple mean based index policy for single SU, named upper confidence bound 1 (UCB1) for iid reward which achieves logarithmic order regret uniformly over time and not only asymptotically.

The reward distribution model is fundamental in the machine learning analysis and the iid assumption is not the only one used. For instance, Markovian reward is also particularly well adapted to OSA scenario in which the probability for a channel to be free at the current time is correlated to its state at the previous time instant. Logarithmic lower bound on regret has also been shown for Markov MAB in [9]. In Markov MAB framework, two use cases can be considered, i.e. rested or restless MAB. In the former case, the state of arms (or bands in OSA scenario) that are not played do not change and only the arm played continues to evolve, while in the later case, the state of non-played arms may change [5], [10]. In [4], [11], the authors have proved that the single

user regenerative cycle algorithm (RCA) are order optimal for rested and restless MAB, respectively. In general, there have been relatively few works on MAB problem with Markovian reward, however, the authors in [11], [10], have proposed few policies for restless Markov MAB problem when a single user accesses the spectrum.

The authors in [11] and [12] have considered centralized spectrum access schemes in contrast to distributed spectrum access here. In a centralized spectrum access, multiple channels are selected by a single centralized user, at each iteration, and receives the reward which is a linear combination of the collected rewards from the selected channels. However, it requires extensive information flow among users, and this type of learning cannot generally be used in cognitive networks where multiple users act selfishly and their collected rewards are affected by the actions of other users operating in a same spectrum.

Moreover, the MAB framework have been also designed to address the distributed channel selection problem, as discussed in [13], [14]. In [14], the authors have proposed a set of policies for multiple-user iid and rested MAB problems, whereas in this paper we propose a policy for restless MAB problems. [14] has assumed that each SU declares its actions to others e.g. the selected channel, which can be a strong constraint. Liu and Zhao, in [13], have proposed a distributed learning and spectrum access policy, time-division fair share (TDFS), for iid rewards and have proved that the policy has a logarithmic order regret. In [13] and [15], the SUs orthogonalize their transmission with different offsets in their time-sharing schedule, while we consider that SUs orthogonalize into different channels. Moreover, the TDFS policies consider that each SU collects almost the same time-average reward while policy proposed in this paper achieves probabilistic fairness.

In [16], the authors have formulated the OSA problem of decentralized learning and spectrum access for multiple SUs, however they have considered MAB framework with iid reward distribution only. In this paper, we consider a similar probabilistic channel access framework as in [16], however contrary to the previous works, we consider restless Markov MAB, which has been proven to be an NP-hard problem, and we prove that the presented policy achieves logarithmic order regret for restless Markovian rewards. Moreover in the above presented works, upon availability, each channel provides the same reward which is a restrictive hypothesis since it ignores different channel qualities. Moreover, the authors in [17] have modeled the OSA problem with partially observable Markov decision process (POMDP) framework which considers channel quality along with availability statistics to decide about the channel to sense. However, their solution has comparatively an higher complexity and also there is no theoretical guarantee on the convergence property of the proposed policies. On the contrary, the OSA problem is modeled under the MAB framework which turns to be very easy and less complex to implement.

Most of the previous works, [10], [11], [14], on restless Markov MAB models a considered fixed positive reward (i.e. expectation of the reward distribution of each channel). In this paper, we consider not only a reward coming from the availability of the channel, which evolves according to a Markov chain, but also a positive random reward iid that we decide to be related to the quality of a channel when the channel is accessed by the user. It is worth mentioning that the distribution of this reward is unknown from the user. Also, an important difference compared to previous approaches, is that the quality reward changes time to time; it is not fixed. Moreover in wireless communication field, the separation of both functionalities, i.e. availability and quality, is necessary for applications where certain level of QoS is required. Hence, by making the engine able to learn on the availability and on the quality, we can choose to emphasize on one or the other criteria.

C. Contributions

Our contributions are to propose single and multi-player policies able to learn on two different criteria, i.e. availability and quality, and to prove the logarithmic order of the regret for these policies. The contributions of this paper are summarized in the following:

- A new version of the quality of service UCB (QoS-UCB), proposed in [18], for the restless Markov MAB framework is proposed and named RQoS-UCB that takes into account not only the availability but also the quality for rating a channel. This new metric allows not only to opportunistically use the spectrum holes but also maximizing the data rate achieved by the unlicensed users.
- A single-user followed by the multi-user version of the algorithm are proposed. The regret bound analysis is provided for the both cases and is proved to be logarithmic over time.
- No information exchange or prior agreement between different users is assumed thereby minimizing the communication overhead. In that context we show that the number of collisions is bounded.

D. Paper Structure

Section II introduces the CR model and hypothesis on the restless Markov MAB problem. In Section III, the new distributed RQoS-UCB algorithm taking into account channel availability and quality is presented for the single-user restless Markov MAB problem and the regret law is derived and Section IV extends to the multi-user case. Numerical results are presented in Section V, which validates the efficiency of the proposed distributed RQoS-UCB policy compared to state-of-the-art algorithms in literature. Finally, Section VI concludes the paper.

II. Model and Problem Formulation

A. System Model

Let $U \geq 1$ be the number of SUs which opportunistically access the spectrum, and $K \geq U$ be the number of PUs' channels available[1]. We consider that each SU can sense only

[1]When $U \geq K$ all channels need to be accessed to avoid collisions and hence learning makes less sense.

one channel in each time slot. Moreover SUs operate in a completely uncorrelated manner w.r.t. primary users, hence the actions of individual SU does not affect the PUs policy. A policy is defined as a one-to-one mapping $\mathcal{A}$ such as at each time n, a frequency band $i \in \mathbb{K}$, $\mathbb{K} = \{1, \cdots, K\}$, is selected:

$$\begin{array}{rccc} \mathcal{A}: & \mathbb{N} & \longrightarrow & \mathbb{K} \\ & n & \longmapsto & i \end{array}$$

Each band is modeled as an aperiodic, irreducible and discrete time Markov chain with finite state space $S^i = \{q_0, q_1\}$, where q_0 and q_1 are the states occupied and free respectively, and the transition probability matrix of the $i-$th band is $P^i = \left\{p^i_{kl}, k, l \in S^i\right\}$. Markov chains are independent from each other and $\boldsymbol{\pi}^i$ is the stationary distribution of the $i-$th Markov chain with $\pi^i_q(n) = \pi^i_q \ \forall n$. $S^{i,j}(n)$ denotes the state observed from band i by SU $j, \forall j \in \mathcal{U}$ at time n. The reward achieved in state $q \in S^i$ from band i by SU j at time n is $r^{i,j}_q(n) \in \mathbb{R}$. Without loss of generality, one can consider that $r^{i,j}_q(n) = S^{i,j}(n)$. The mean reward $\mu^S_{i,j}$ associated with state $S^{i,j}$ of the $i-$th band and $j-$th user under stationary distribution $\boldsymbol{\pi}^i$ is given by: $\mu^S_{i,j} = \sum_{q \in S^{i,j}} r^{i,j}_q \pi^i_q$.

Furthermore, the band quality is rated according to the interference temperature recorded on it. We assume that the band quality in a given state is stationary in the wide sense, meaning that its statistical properties, i.e. first and second moment, are not evolving over time, but the instantaneous value $R^{i,j}_q(n)$ may vary. $G^{i,j}_q(T^{i,j}(n)) = \frac{1}{T^{i,j}(n)} \sum_{k=1}^{T^{i,j}(n)} R^{i,j}_q(k)$ denotes the empirical mean of quality observations $R^{i,j}_q$ collected from band i by SU j in state q and $T^{i,j}(n)$ denotes the total number of times band i has been sensed up to time n by SU j. The global mean reward, i.e. taking into account the quality as well as the state of each band i, is defined as:

$$\mu^R_{i,j} = \sum_{q \in S^{i,j}} G^{i,j}_q r^{i,j}_q \pi^i_q. \tag{1}$$

Without loss of generality, let us consider that $\mu^R_1 > \mu^R_i > \mu^R_K$, $\forall\ i \in \{2, \cdots, K-1\}$. It is important to note that the optimal bands are the ones having the highest global mean reward, i.e. $\{\mu^R_j\}_{\forall j \in U}$. The global mean reward can be seen as the expectation of the reward function of the user j in band i, i.e. $G^{i,j}_q r^{i,j}_q$. This function can be seen as a weighting of the channel availability reward, i.e. $r^{i,j}_q$, by a random variable reflecting its quality, i.e. $G^{i,j}_q$. Other combinations of quality and availability might be envisaged but fall out of the scope of this paper and are left for further works.

Let ALOHA-like protocol be considered under which if two or more SUs transmit in the same channel then none of the transmissions are successful, and no collision avoidance mechanisms are considered[2]. $C^j_o(i,n)$ is the collision indicator function at the $n-$th slot at channel i for SU j. At the end of each slot n, each SU j receives the reward $r^{i,j}_q(n)$ and $R^{i,j}_q(n)$. Under this model, we are interested in designing a policy $\mathcal{A}$, maximizing the expected number of successful transmissions with good quality in the long run. Let $\Phi^{\mathcal{A}}(n)$ be the regret and defined as the reward loss after n slots for U SUs and policy $\mathcal{A}$. In the ideal scenario, we assume that the channel mean reward statistics μ^R_i are known a priori by a central agent and it selects U optimal channels for U SUs.

We are interested in minimizing the *regret* $\Phi^{\mathcal{A}}(n)$ associated with the learning and access scheme, defined as:

$$\Phi^{\mathcal{A}}(n) = \sum_{j=1}^{U} n\mu^R_j - \sum_{j=1}^{U} \mathbb{E}\left[\sum_{t=1}^{n} G^{\mathcal{A}(t,j)}_{q_{\mathcal{A}(t,j)}}(t) r^{\mathcal{A}(t,j)}_{q_{\mathcal{A}(t,j)}}(t)\right] \tag{2}$$

where the expectation $\mathbb{E}$ is taken over the states and quality. Let $q_{\mathcal{A}(t,j)}$ being the state observed by SU j under the policy $\mathcal{A}$ at time t. Frequency bands whose mean reward is strictly less than $\{\mu^R_j\}_{\forall j \in U}$ are referred as suboptimal frequency bands.

B. Bandit Theoretical Model of Wireless Network

A centralized primary network is considered where a radio access point serves K frequency bands. In the same cell, Tx-Rx SU pairs are considered. The CR system seeks to interweave the SU's signals with the PUs transmissions in the set of frequency bands. Due to the frequency reuse factor and partial utilization of the frequency band i by PUs in the neighboring cells, the interference level is not the same for all bands which leads to a varying quality according to the band considered. This noise process is considered as stationary in wide sense.

A slotted frame structure for CR is considered. During the frame duration, i.e. T_{total}, SUs sense the channel during T_{sens}, learn during T_{lear}, and transmit (or not depending on the result of the channel sensing) during T_{trans}. The aim of the learning policy is to decide which band should be explored in the next time slot and can be implemented in parallel with transmission requiring a very small time compared to the sensing and transmission duration [19].

At the current time slot, SUs sense the spectrum on the $i-$th frequency band and utilize it for communication only when there is no PU. The spectrum sensing part is error-prone but the imperfect sensing has no effect on the optimal solution achieved by algorithms [20]. The discrete received signal at SUs Rx can be written according to both hypothesis: i.e. $\mathcal{H}_0$ band i is used by a PU and $\mathcal{H}_1$ band i is vacant:

$$\mathcal{H}_0 : y^i_{q_0}[m] = p^i[m] + u^i[m], \tag{3}$$

$$\mathcal{H}_1 : y^i_{q_1}[m] = u^i[m] \tag{4}$$

where, $p^i[m]$ and $u^i[m]$ are the signal and noise component respectively for the $i-$th band. The PU signal $p^i[m]$ is a zero mean independent and identically distributed (i.i.d.) random process with variance $\mathbb{E}[|p^i[m]|^2] = \sigma^2_{p,i}$. The noise components $u^i[m]$ are assumed to be zero mean and complex Gaussian distributed with variance $\mathbb{E}[|u^i[m]|^2] = \sigma^2_{u,i}$ and independent from the primary users' signal p^i. We remind that $u^i[m]$ counts for other cells interference and background noise. Theoretically in OSA context, SUs should transmit only when no PU occupies band i. However, they may miss the detection of a primary user and transmit anyway.

[2]The effect of employing CSMA-CA is not taken into an account here although the use of CSMA-CA increases spectrum usage and consecutively decreases the regret, thus, the bound we derive in this paper is applicable.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

1) Energy Detector (ED) as state and quality information metric: ED senses band i and measures a power level. If the measured power level is above a certain threshold υ, ED decides the band is occupied and if the power level is below υ the band is decided to be available. Moreover, in our work, the measured spectrum level is recorded and used as a quality information metric for learning policy. Indeed, instead of having only a binary variable at the output of ED, we get a soft-metric representing the measured power level which will be used later to rate the band quality. Let F_s being the sampling frequency at the receiver, then $N_s = T_{\text{sens}} F_s$ is the number of samples acquired during the sensing phase. User j measures the power in the $i-$th band as $\mathcal{P}_q^{i,j} = 1/N_s \sum_{m=1}^{N_s} \|y_q^{i,j}[m]\|^2$. The false alarm probability $P_f(\upsilon, N_s)$ under the threshold υ and number of samples N_s is given as $P_f(\upsilon, N_s) = Q\left(\left(\upsilon/\sigma_{u,i}^2 - 1\right)\sqrt{N_s}\right)$ [21], $Q(\cdot)$ being the Gaussian Q-function. The band with the highest quality has the lowest interference plus noise power level. Hence, the quality metric should be inversely proportional to the ED output, i.e. $R_q^{i,j} = 1/\left(\mathcal{P}_q^{i,j} + c_e\right)$ where c_e is a constant added to the received power in order to avoid taking the inverse of very small numbers. The proposed algorithms in this paper is able to take into account quality metric related to soft output of any spectrum sensing detector [22]. There have been some attempts in [23], [24], to consider the energy detector soft output as a reward for general reinforcement learning algorithms, but they lack from significant theoretical guarantee and a relation with achievable throughput.

2) Achievable Throughput Analysis: The SU's average achievable throughput $\Xi^{i,j}$ in the $i-$th frequency band is given by the sum of the achievable throughput under $\mathcal{H}_0$ and the achievable throughput under $\mathcal{H}_1$, which can be written as:

$$\Xi^{i,j}(\upsilon, N_s) = \Xi_{q_1}^{i,j}(\upsilon, N_s) + \Xi_{q_0}^{i,j}(\upsilon, N_s), \tag{5}$$

where $\Xi_{q_1}^{i,j}$ and $\Xi_{q_0}^{i,j}$ are defined as:

$$\Xi_{q_1}^{i,j}(\upsilon, T_{sens}) = \frac{T_{\text{total}} - T_{\text{sens}}}{T_{\text{total}}} \pi_{q_1}^i C\left(\Gamma_1^{i,j}\right)\left(1 - P_f(\upsilon, N_s)\right)$$

$$\Xi_{q_0}^{i,j}(\upsilon, T_{sens}) = \frac{T_{\text{total}} - T_{\text{sens}}}{T_{\text{total}}} \pi_{q_0}^i C\left(\Gamma_0^{i,j}\right)\left(1 - P_d(\upsilon, N_s)\right)$$

where $C(\Gamma)$ denotes the achievable rate with SINR Γ, i.e. $C(\Gamma) = \log_2(1+\Gamma)$. Hence, $C\left(\Gamma_1^{i,j}\right)$ is the achievable rate in the $i-$th band without any PU and $C\left(\Gamma_0^{i,j}\right)$ is the achievable rate in the $i-$th band when a PU has not been detected. The achievable throughput under hypothesis $\mathcal{H}_1$ is hence $C\left(\Gamma_1^{i,j}\right)$ multiplied by the probability band i is available, i.e. $\pi_{q_1}^i$, and the probability not to generate a false alarm, i.e. $1 - P_f(\upsilon, N_s)$. On the other hand, the achievable throughput under $\mathcal{H}_0$ is $C\left(\Gamma_0^{i,j}\right)$ weighted by the probability band i is occupied, i.e. $\pi_{q_0}^i$ and the miss detection probability i.e. $P_{md}(\upsilon, N_s) = 1 - P_d(\upsilon, N_s)$ where $P_d(\upsilon, N_s)$ is the probability of correct detection. Both rates are weighted by the effective transmission time ratio, i.e. $\frac{T_{\text{total}} - T_{\text{sens}}}{T_{\text{total}}}$.

III. Single user restless Markovian MAB problem

Let us consider in a first place that our secondary network only contains 1 transceiver pair. We construct an algorithm called RQoS-UCB, as shown in Algorithm 1, and prove that it achieves logarithmic order regret uniformly over time same as original QoS-UCB policy for rested problem [18]. On a way to solve the restless MAB problem is to apply regenerative cycle algorithm (RCA), [25], [11], or restless UCB (RUCB), [10], policies and ignore the separate weight for channel quality and availability criteria. In this paper, we assume that channel set consists in different transmission quality, and also a CR may have certain preferences, i.e., quality or availability or both, for channel selection. Unlike the rested problem dealt with in [18], the Markov chains of arms evolve even if they are not observed. Table I summarizes the notation we use for RQoS-UCB algorithm, where the dependence on index j vanishes for 1 SU.

RQoS-UCB operates in a block structure as shown in Fig. 1. For each arm, a state ζ^i is chosen and defined as a *regenerative* state. Each block is further divided into three sub-blocks (SBs), i.e. SB1, SB2 and SB3. SB1 consists in all time slots from the start of the block to right before the first visit to ζ^i, SB2 contains all time slots from the first visit to ζ^i up to but excluding the second visit to ζ^i where state and quality of the band are recorded, and finally SB3 consists in a single time slot with the second visit to state ζ^i. At the end of SB3, the policy index is computed and is compared with the index of other arms and the highest one gives the next arm to play, e.g. arm K for the second block in Fig. 1. Note that the sub-block division is relevant for regret analysis purpose. Indeed, all SB2 blocks are virtually assembled to construct a regenerative cycle of the Markov chains. The newly constructed sample path has exactly the same statistics as the original transition probability matrix P^i which translates restless problem into a tractable problem. However, it is important to emphasize that the CR does not run only during SB2 block but also in SB1 and SB3 blocks in which channels are sensed and transmissions are performed, if bands are found free.

Initially, all the channels are observed at least once and ζ^i is fixed as a first state observed for each arm, i.e. steps 1 to 3 in Algorithm 1. After the initialization, at the beginning of a new block b, RQoS-UCB selects the channel which maximizes the policy index $B^i(n_2, T_2^i(n_2))\ \forall\ i \in \mathbb{K}$, step 5, according to three terms:

$$\begin{aligned} B^i(n_2, T_2^i(n_2)) &= \bar{S}^i(T_2^i(n_2)) - Q^i(n_2, T_2^i(n_2)) \\ &+ A^i(n_2, T_2^i(n_2)),\ \ \forall i \end{aligned} \tag{6}$$

where $\bar{S}^i(T_2^i(n_2))$ being the empirical mean of the states of the $i-$th band (occupied or free) at time n_2, defined as:

$$\bar{S}^i(T_2^i(n_2)) = \frac{S^i(1) + S^i(2) + \cdots + S^i(T_2^i(n_2))}{T_2^i(n_2)}, \forall i. \tag{7}$$

The second term, i.e. $Q^i(n_2, T_2^i(n_2))$, is computed same as of rested policy:

$$Q^i(n_2, T_2^i(n_2)) = \frac{\beta M^i(n_2, T_2^i(n_2)) \log(n_2)}{T_2^i(n_2)}, \forall i, \tag{8}$$

where,

$$M^i(n_2, T_2^i(n_2)) = G_{\max}^{q_1}(n_2) - G_{q_1}^i(T_2^i(n_2)),\ \ \forall i,$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

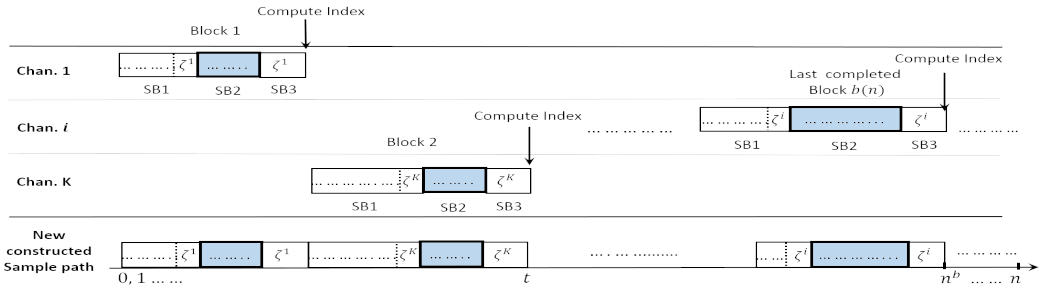


Fig. 1. Example of block (i.e. SB1, SB2 and SB3 sub-block) operation of RQoS-UCB policy. At the end of block 1, RQoS-UCB policy computes index based on the observations collected in SB2 block, finds a channel having highest the index among the set of channels $\mathcal{K}$, and moves to channel (for example K) with the highest index for block 2.

Algorithm 1 Single SU RQoS-UCB policy

Input: $b = 1$, $n = 0$, $n_2 = 0$, $T_2^i = 0$, α, β, $\mathcal{A}(0)$, $R_{q_1}^i(0)$ $\forall$ $i \in \mathbb{K}$.

Output: $\mathcal{A}(n+1)$

for n_2 = b **to** K **do**
  Initialize policy by sensing each channel for at least one block (i.e. SB1, SB2 and SB3)
end for
while (1) **do**
  $B^i(n_2, T_2^i(n_2)) = \bar{S}^i(T_2^i(n_2)) - Q^i(n_2, T_2^i(n_2)) + A^i(n_2, T_2^i(n_2))$, $\forall i$
  $\mathcal{A}(n) = \arg\max_i B^i(n_2, T^i(n_2))$
  Sense $i = \mathcal{A}(n)$ and Observe $S^i(n_2)$
  while $S^i(n_2) \neq \zeta^i$ **do**
    $n = n+1$, $\mathcal{A}(n) = i$ // Start SB1 sub-block
    Sense channel i and Observe $S^i(n_2)$
  end while
  $n = n+1$, $n_2 = n_2+1$, $T_2^i(n_2) = T_2^i(n_2)+1$, $\mathcal{A}(n) = i$; // End of SB1, start SB2
  Observe current state $S^i(n_2)$ and update $R_{q_1}^i(n_2)$
  Update $\bar{S}^i(T_2^i(n_2))$, $Q^i(n_2, T_2^i(n_2))$ and $A^i(n_2, T_2^i(n_2))$ as of (7), (8) and (9), respectively
  while $S^i(n_2) \neq \zeta^i$ **do**
    $n = n+1$, $n_2 = n_2+1$, $T_2^i(n_2) = T_2^i(n_2)+1$, $\mathcal{A}(n) = i$; // Start SB2 sub-block
    Observe current state $S^i(n_2)$ and update $R_{q_1}^i(n_2)$
    Update $\bar{S}^i(T_2^i(n_2))$, $Q^i(n_2, T_2^i(n_2))$ and $A^i(n_2, T_2^i(n_2))$ as of (7), (8) and (9), respectively
  end while
  $b = b+1$, $n = n+1$ // Start of SB3 sub-block
end while

where $G_{q_1}^i(T_2^i(n_2))$ and $G_{\max}^{q_1}(n_2)$ denote the empirical mean of quality observations $R_{q_1}^i$ and the maximum expected quality within the set of frequency bands respectively, defined as in Table I. Finally, the bias term $A^i(n_2, T_2^i(n_2))$, is defined as

$$A^i(n_2, T_2^i(n_2)) = \sqrt{\frac{\alpha \log(n_2)}{T_2^i(n_2)}}, \quad \forall i. \tag{9}$$

Two coefficients come into play in (8) and (9), i.e. β and α respectively, and are introduced to balance the trade-off between exploration and exploitation. Parameter α in (9) forces the exploration of the other bands to check their availability while the new parameter β forces the algorithm to give some weight to the quality in the index computation.

We first give an upper bound on the total expected number of plays of suboptimal arms in Theorem 1 and then the regret of RQoS-UCB policy in Theorem 2.

Condition 1. *All arms are finite-state, irreducible, aperiodic restless Markov chains whose transition probability matrices have irreducible multiplicative symmetrization, and the state of non-played arms evolve. Let,* $G_q^i \geq \frac{1}{\hat{\pi}_{\max} + \pi_q^i}$ *and* $\forall \beta \geq 84 S_{\max}^2 r_{\max}^2 G_{\max}^2 \hat{\pi}_{\max}^2 / \left(\gamma_{\min} \Delta\mu_i^R M_{\min}\right)$.

Theorem 1. *Assume all arms follow condition 1. We can upper bound the total expected number of block spent in suboptimal arms as:*

$$\mathbb{E}[F^i(b(n))|b(n) = b] \leq \frac{4\alpha \log n}{(\Delta\mu_i^R)^2} + \frac{|S^1| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}$$

and the time spent in suboptimal arm:

$$\sum_{i \in K} (\mu_1^R - \mu_i^R) \mathbb{E}[T^i(n)] \leq Z_1 \log n + Z_2$$

where,

$$Z_1 = \sum_{i=2}^{K} \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + 1 \right) \frac{4\alpha}{\Delta\mu_i^R}$$

$$Z_2 = \sum_{i=2}^{K} \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + 1 \right) \Delta\mu_i^R \left[\frac{|S^1| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right]$$

Proof: The proof is given in Appendix A. ■

Theorem 2. *Assume all arms follow condition 1, the regret of RQoS-UCB can be bounded by*

$$\Phi^R(n) \leq Z_3 \log n + Z_4, \tag{10}$$

$$Z_3 = Z_1 + Z_5 \quad \text{and} \quad Z_4 = Z_2 + Z_6 + Z_7$$

$$Z_5 = \sum_{i=2}^{K} \frac{4\alpha}{\left(\Delta\mu_i^R\right)^2} \left[\mu_i^R \left(1 + \Omega_{\max}^i\right) + \mu_1^R \Omega_{\max}^1 \right]$$

$$Z_6 = \sum_{i=2}^{K} \left[\frac{|S^1| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right] \left[\mu_i^R \left(1 + \Omega_{\max}^i\right) + \mu_1^R \Omega_{\max}^1 \right]$$

$$Z_7 = \mu_1^R \left(\frac{1}{\Pi_{\min}} + \max_{i \in \{1,\ldots,K\}} \Omega_{\max}^i + 1 \right)$$

K and U:	Number of arms (channels) and users respectively
P^i and S^i:	transition matrix and state space of the Markov chain of channel i
$\{1,\cdots,U\}$ and $\{U+1,\cdots,K\}$:	denotes set of optimal and suboptimal channels
$S^{i,j}(n)$:	observed state of channel i at time n for SU j
$r_q^{i,j}(n)$:	reward achieved in state $q \in S^i$ from a band i by SU j at time n
$R_q^{i,j}(n)$:	Instantaneous observed quality of band i by SU j at time n
$T^i(n) = \sum_{j=1}^{U} T^{i,j}(n)$:	total number of times band i has been sensed up to time n
$G_q^{i,j}(T^{i,j}(n))$:	the empirical mean of quality observations $R_q^{i,j}$ by SU j
$G_{\max}^{q,j}(n)$:	maximum expected quality within the set by SU j
$M^{i,j}(n, T^{i,j}(n))$:	$G_{\max}^{q_1,j}(n) - G_{q_1}^{i,j}(T^{i,j}(n))$
$\mathcal{A}(n,j)$	channel index which has to be sensed in the next time slot by SU j
$B^{i,j}(n, T^{i,j}(n))$:	policy index for the set of bands $\mathcal{K}$ at SU j
α and β:	exploration coefficients with respect to availability and quality, respectively
$\zeta^{i,j}$:	state that determines regenerative cycles for band i by SU j
n_2:	total number of time slots spend in SB2 block up to block b.
$T_2^{i,j}(n_2)$:	total number of time band i is sensed by SU j during SB2 block up to n_2 time.
W:	Frame size for multi-user distributed RQoS-UCB policy
$C_o^j(i,n)$:	indicator of collision at $n-$th slot at channel i for SU j
$Rank(j)$:	$Rank(j)-$th highest entry in $B^{i,j}(n, T^i(n)), \forall i \in K$ for SU j
π_q^i:	stationary distribution for state q of the Markov chain associated with i
μ_i^R:	global mean reward
$\Delta\mu_{i,j}^R$:	$\mu_j^R - \mu_i^R$
$G_{\max}^j$:	$\max_{q\in S^i} G_{\max}^{q,j}$
$\pi_{\min}^i$ and $\pi_{\min}$:	$\min_{q\in S^i} \pi_q^i$ and $\min_{i\in\mathbb{K}} \pi_{\min}^i$ respectively
$\hat{\pi}_q^i$, $\tilde{\pi}_{\max}$ and $r_{\max}$:	$\max\left\{\pi_q^i, 1-\pi_q^i\right\}$, $\max_{q\in S^i, i\in\mathbb{K}} \hat{\pi}_q^i$ and $\max_{q\in S^i, i\in\mathbb{K}} r_q^i$ respectively
$S_{\max}$:	$\max_{i\in\mathbb{K}} \lvert S^i\rvert$, where $\left\lvert S^i\right\rvert$ stands for the cardinality of the state space of arm i
$M_{\min}^j$ and $M_{\max}^j$:	$\min_{i\in\mathbb{K}} M^{i,j}\left(n, T^{i,j}(n)\right)$ and $\max_{i\in\mathbb{K}} M^{i,j}\left(n, T^{i,j}(n)\right)$ respectively
γ^i, $\gamma_{\min}$ and $\gamma_{\max}$:	eigenvalue gap of the $i-$th channel, $\min_{i\in\mathbb{K}} \gamma^i$ and $\max_{i\in\mathbb{K}} \gamma^i$, respectively
$\Omega_{k,l}^i$:	mean hitting time of state l starting from an initial state k for the ith arm.
$\Omega_{\max}^i$ and $\Omega_{\max}$:	$\max_{k,l\in S^i, k\neq l} \Omega_{k,l}^i$ and $\max_{i\in\{1,\cdots,K\}} \Omega_{\max}^i$ respectively
$b(n)$ and $f(n)$:	total number of completed blocks and frame up to time n
n^b	time at the end of the last completed block $b(n)$
$f_w(n)$ and $T_w(n)$:	number of frame and time where any one of the U optimal channel's estimated ranks is wrong
$F^{i,j}(b(n))$:	total number of block in which arm i is played by SU j up to block $b(n)$
$\mathbb{E}(C_o(n))$:	expected number of collisions in U optimal channels
Υ:	refers to the time required to reach the absorbing state from any initial state
$S_1^{i,j}(k)$:	vector of observed states from SB1 of $k-$th block in which band i is sensed by SU j
$S_2^{i,j}(k)$:	vector of observed states from SB2 of $k-$th block in which band i is sensed by SU j
$S^{i,j}(k)$:	vector of observed states from $k-$th block $S^i(j) = \left[S_1^i(j), S_2^i(j), \zeta^i\right]$
$\bar{X}(j)$:	j-th combined block in which the optimal band is sensed, i.e. $\bar{X}(j) = \left[\bar{X}_1(j), \bar{X}_2(j), \zeta^i\right]$
$\bar{b}^{i,j}$:	total number of joined blocks up to current block b for optimal band i for SU j

TABLE I
NOTATION FOR ALGORITHMS 1 AND 2, AND REGRET ANALYSIS

Proof: Proof of Theorem 2 is given in Appendix B. ■

IV. DISTRIBUTED MULTI-USER LEARNING AND ACCESS POLICY

In this part, we extend the previous approach to the multi-users case and present the multi-user version of RQoS-UCB, sometimes referred as *distributed* RQoS-UCB policy. If each SU applies naively the single user RQoS-UCB presented in Algorithm 1, then the number of collisions will likely increase since all SUs go for the best channel. Hence, we introduce the so-called channel access *rank* [16]. Let assume that each SU j keeps a decreasing ordering set $B^{i,j} \forall i \in \mathbb{K}$ indexes. For instance, a user having a *rank* equal to 3, goes to the channel having the third entry in the ordering set of its index. Moreover, suboptimal channels are expected to be played as less as possible, in order not to increase the regret of the system.

Fig. 2 illustrates the functioning of the distributed RQoS-UCB policy for 2 SUs. Let us consider a slotted system with a frame size W, where each SU can be synchronized for their index calculation. Each SU computes its own $B^{i,j}$ index at the end of W if possible, i.e. the sub block SB3 has been encountered for user j in channel i when the frame ends. If not, the computation is delayed to the next frame and user j continues to play the same channel. If two or more users go to the same channel, they collide and then they draw a random number from the channels' set $\mathcal{K}$ as their new rank. Letting a player randomizes among its U optimal ranked arms can help alleviate this problem, and focuses to eventually orthogonalize the U players in their choice of arms.

This random rank is the same idea that has been used in Anandkumar *et al.* [16]. The difference in this paper is that, in Anandkumar *et al.* [16] the randomization is performed at each time step a collision occurs under an iid reward model, whereas in our case the randomization is performed at the end of a completed frame of length W and is therefore less frequent as block lengths are random. The reason for this is because with the Markovian reward model, index updates can

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

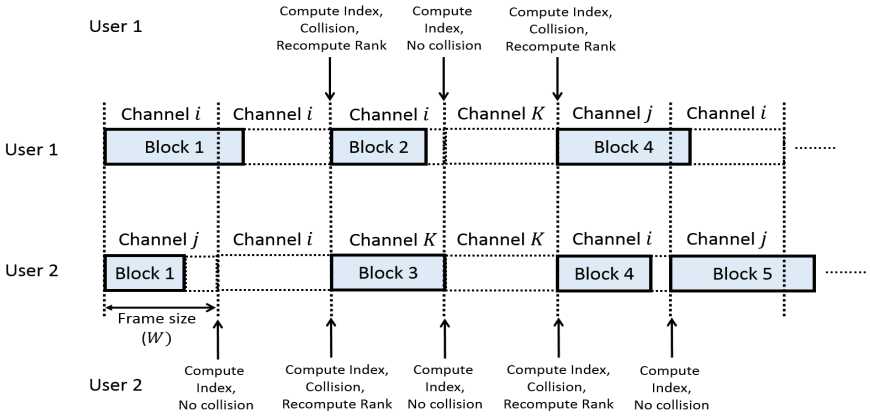


Fig. 2. Running cycle for 2 different SUs using the distributed RQoS-UCB policy. Actions of user 1 and 2 are listed on top and bottom, respectively

only be performed after finishing a regenerative cycle, and switching a channel before a completion of regenerative block will waste the state observations made within that incomplete block. Algorithm 2 summarizes the steps followed by each user. Note that policy operation in each block (containing SB1, SB2 and SB3) follows the same steps as the single user policy detailed in Algorithm 1, but not reported here for the sake of simplicity.

Algorithm 2 Multi-User Distributed RQoS-UCB policy

Input: U: Number of users, K: Number of channels,
$B^{i,j}(n, T^i(n))$: single-user policy index for each SU $j \in U$ and channel $i \in \mathcal{K}$,
$C_o^j(i,n)$: collision indicator in channel i for SU j,
$Rank(j)$: the $Rank(j)-$th highest entry in $B^{i,j}(n, T^i(n)), \forall i \in K$ for SU j

Output: $\mathcal{A}(n,j)$

if SB3 sub-block observed by SU j in last frame **then**
  Calculate RQoS-UCB policy index $B^{i,j}(n, T^i(n))$ as in Algorithm 1
  if $C_o^j(\mathcal{A}(n-1,j), n-1) = 1$ **then**
    Draw a new $Rank(j)$ randomly from the set $\{1, \cdots, U\}$ for SU j
  else
    Maintain the same $Rank(j)$ for SU j
  end if
  $\mathcal{A}(n,j)$: channel having $Rank(j)-$th highest entry in $B^{i,j}(n, T^i(n))$
  Sense $\mathcal{A}(n,j)$ channel
  if collision **then**
    $C_o^j(\mathcal{A}(n,j), n) \leftarrow 1$,
  else
    $C_o^j(\mathcal{A}(n,j), n) \leftarrow 0$
  end if
else
  $\mathcal{A}(n,j) = \mathcal{A}(n-1,j)$
  Follow on SB1 and SB2 block same as Algorithm 1
end if

We now present a logarithmic upper bound on the number of time one of the suboptimal channel $i \in \{U+1, \cdots, K\}$ is sensed by one of the U SUs employing the same distributed RQoS-UCB policy[3].

Theorem 3 (Time spent in suboptimal channels under multi--user distributed RQoS-UCB policy)**.** *Assume all arms follow condition 1. Under the distributed RQoS-UCB scheme, total time spent by any SU $j \in \{1, \cdots, U\}$ in any suboptimal channel $i \in \{U+1, \cdots, K\}$ is given by:*

$$\begin{aligned}\mathbb{E}[T^{i,j}(n)] &\leq \sum_{j=1}^{U} \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + W \right) \mathbb{E}\left[F^{i,j}(f(n))\right] \\ &\leq \sum_{j=1}^{U} \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + W \right) \left[\frac{4\alpha \log n}{(\Delta\mu_{i,j}^R)^2} \right. \\ &\quad \left. + \left[\frac{|S^j| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right] \right] \end{aligned} \tag{11}$$

Proof: Proof of Theorem 3 is given in Appendix C. ■

Let now focus on the analysis of the number of collisions $C_o(n)$ in the U optimal channels up to time n. First, we state a bound on the expected number of collisions in the ideal scenario where each SU has perfect knowledge of the mean reward μ_i^R. In this case, SUs try to reach an orthogonal configuration by uniformly randomizing over the U optimal channels. We use the following Lemma from [16] and [26] to bound the number of collisions arising due to the distributed scenario:

Lemma 1 (Number of collisions under perfect knowledge of μ_i^R, [16], [26])**.** *The expected number of collisions under the random allocation access scheme in Algorithm 2, assuming that each SU has perfect knowledge of the mean reward μ_i^R, is given by*

$$\mathbb{E}[C_o(n)|\mu_i^R] \quad \leq \quad U\mathbb{E}[\Upsilon] \leq U\left[\binom{2U-1}{U} - 1 \right].$$

[3]Note that the upper bound on $\mathbb{E}[T^{i,j}(n)]$ is still valid even if other SUs use a different policy than distributed RQoS-UCB. However on the contrary, we need to ensure that every SU must implement the same random access mechanism in order to analyze the expected number of collisions

The above Lemma 1 states that there is a finite number of collisions, bounded by $U\mathbb{E}[\Upsilon]$ under the perfect knowledge of μ_i^R. However as stated before, there are no collisions in a case where all SUs have the perfect knowledge of μ_i^R in the presence of pre-allocated ranks. Thus, $U\mathbb{E}[\Upsilon]$ gives a bound on the additional number of collisions due to absence of pre-allocated ranks or the lack of direct communication among the SUs to negotiate their rank. To analyze the number of collisions under multi-user distributed RQoS-UCB learning of the unknown μ_i^R, we show that all SUs are able to learn the correct order of the different channels with only logarithmic regret, and then we show that only an additional finite number of collisions occurs before reaching collision-free configuration.

Let define $T_w(n)$ and $f_w(n)$ as the number of time and frames where any one of the U optimal channel's estimated ranks is wrong under the distributed RQoS-UCB policy.

Lemma 2 (Wrong order of distributed RQoS-UCB index statistics without perfect knowledge of μ_i^R)**.** *Under the distributed RQoS-UCB scheme in Algorithm 2, the total expected number of frames and time slots for which the estimated $B^{i,j}(n, T^{i,j}(n))$ indices of distributed RQoS-UCB policy is not in a same order as the mean reward μ_i^R, is:*

$$\mathbb{E}[T_w(n)] \leq U \sum_{a=1}^{U} \sum_{b=1}^{K} \left(\frac{1}{\pi_{\min}^b} + \Omega_{\max}^b + W \right) \left[\frac{4\alpha \log n}{(\Delta\mu_{a,b}^R)^2} + \frac{|S^a| + |S^b|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2} \right] \tag{12}$$

Proof: Proof of Lemma 2 is achieved by following same argument and steps as of [16]. ∎

With the help of Theorem 3 and Theorem 3 from [16], the number of time slots spent in the suboptimal channels is also logarithmic, we proceed to prove one of the main results of this paper that the sum regret under multi-user distributed RQoS-UCB policy is logarithmic for restless Markovian reward.

Theorem 4 (Regret analysis of multi-user distributed RQoS-UCB policy)**.** *Assume all arms follow condition 1. Then, the regret of distributed RQoS-UCB policy can be bounded by $O(\log n)$*

$$\Phi^M(n) \quad \leq \quad X_3 \log n + X_4 \tag{13}$$

where X_3 and X_4 are as stated in Appendix D.

Proof: Proof of Theorem 4 is given in Appendix D. ∎

We can note that the presented upper bound does not depend on the regenerative state ζ. Thus, with minimal information about the bands, an SU can still have a logarithmic regret by selecting appropriate exploration coefficients.

V. Numerical Analysis

The performance of distributed RQoS-UCB policy is investigated on a set of 10 bands, and with different number of SUs $U \in \{1, \cdots, 10\}$. Simulation is performed over 10^3 runs, each with a duration about 1000 seconds. QPSK signaling are assumed to be used for PU signals same as [21]. The threshold υ is set to have a high detection probability, i.e. $P_d(\upsilon, T_s) = 0.95$ for each channel i. Note that the learning phase represents a very low computational complexity and it may be neglected compared to sensing [19]. Learning can be done in parallel with the transmission, and so uses no time that could prevent from transmitting frames, and thus be considered as having no impact on bandwidth usage. The exploration coefficients of RQoS-UCB policy are $\alpha = 0.25$, $\beta = 0.32$, like in [18], and will be used throughout the numerical analysis unless otherwise mentioned.

Table II summarizes the Markov chain parameters modeling the primary network, such as the transition probabilities P^i, selected arbitrarily, for each channel on the first and second rows, the vacancy probability $\pi_{q_1}^i$ calculated from P^i, the empirical average of the band quality $G_{q_1}^i$ calculated as explained in Section II and estimated at Rx and feedback to Tx, on the fourth row and the global mean reward, μ_i^R calculated with (1) taking into account availability and quality on the fifth row.

A. Single user case

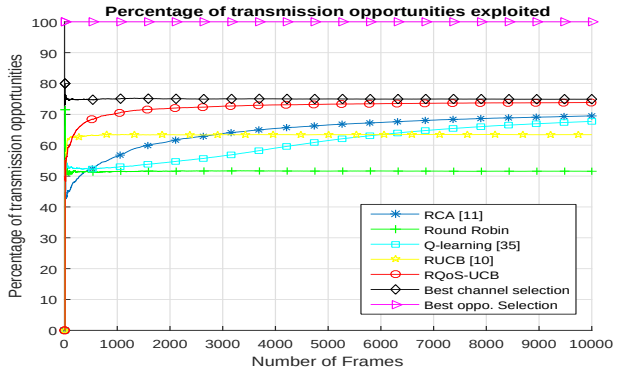
In this Section, only one cognitive transceiver pair is considered trying to exploit the frequency bands of a primary network. Theorems 1 and 2 state that the new metric introduced to rate the quality in the learning phase does not prevent from achieving a logarithmic order regret in the restless case, as it has been also observed but not proved in the rested case [18]. Due to the application to OSA context, regret analysis is not sufficient to characterize the performance of a learning policy and its ability to provide high reliable data rate is of great interest for telecommunication purposes. Hence, RQoS-UCB policy is compared to several other learning policies found in literature, such as RCA [11], RUCB [10] and Q-learning [27] with the optimal exploration parameters suggested by the authors. The classical RCA and RUCB policies are modified such as the unique reward takes into account the combination of availability and quality, i.e. $s_q^{i,j} = R_q^{i,j} r_q^{i,j}$. Here, $r_q^{i,j}$ is the fixed reward selected for channel i and user j in state q, whereas $R_q^{i,j}$ is the sample drawn from the fixed iid reward distribution modeling the quality of channel i in state q for user j which is not considered in previous works. The algorithms are also compared with two other policies used as baseline for comparison, i.e. best channel selection and best opportunistic selection policies. The best channel selection policy always selects the optimal channel, i.e. channel with the highest mean reward, and if it is occupied does not transmit, whereas the best opportunistic selection policy is a "god driven" policy which knows *a priori* all holes in the spectrum as well as the quality associated to. Then at a given time, it exploits the best available channel among all channels to transmit.

Fig. 3(a) presents the percentage of transmission opportunities defined as the ratio between the number of times a policy selects an available channel and the total number of trials. Best opportunistic transmission policy is an upper bound since it possesses a prior information about the spectrum occupancy. The proposed RQoS-UCB but also RCA [11], RUCB [10] and Q-learning policies are able to find an optimal channel in the long run, and match the performance of best channel selection policy which always selects a channel having the

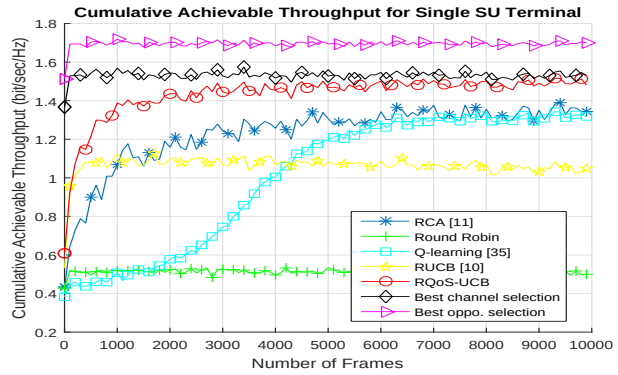
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

channel (i)	1	2	3	4	5	6	7	8	9	10
$p^i_{q_0q_1}$	0.3	0.65	0.75	0.6	0.8	0.4	0.2	0.65	0.45	0.35
$p^i_{q_1q_0}$	0.7	0.35	0.25	0.4	0.2	0.6	0.8	0.35	0.55	0.65
$\pi^i_{q_1}$	0.3	0.65	0.75	0.6	0.8	0.4	0.2	0.65	0.45	0.35
$G^i_{q_1}$	0.67	0.64	0.79	0.77	0.70	0.80	0.67	0.63	0.64	0.79
μ^R_i	0.31	0.48	0.60	0.47	0.58	0.33	0.26	0.46	0.39	0.29

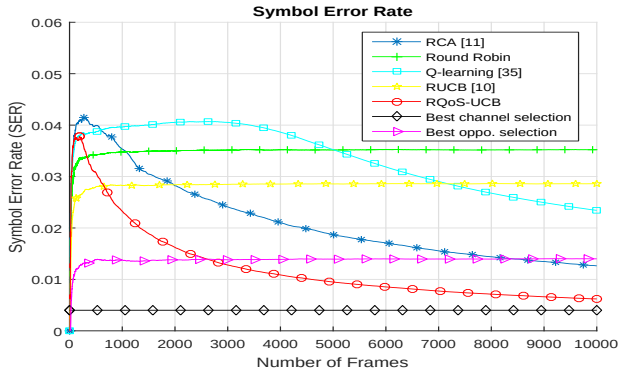
TABLE II
STATE TRANSITION PROBABILITIES, MEAN AVAILABILITY, EMPIRICAL MEAN QUALITY AND GLOBAL MEAN REWARD.



(a) Opportunities w.r.t. the number of frames



(b) Achievable throughput w.r.t. the number of frames



(c) Symbol error rate (SER) w.r.t. the number of frames

Fig. 3. Percentage of transmission opportunities exploited, achievable throughput and SER for single SUs w.r.t. the number of frames.

highest mean reward μ^R_i. However, the proposed RQoS-UCB policy greatly outperforms other approaches in term of convergence and achieves the higher number of opportunities. This result highlights the benefits of using separate optimization of both availability and quality as we propose with QoS-UCB compared to use a single reward in the UCB policy.

The achievable throughput is investigated in Fig. 3(b) and is computed with (5). The best opportunistic selection policy logically upper bounds the performance due to the higher number of transmission opportunities it exploits as it can be seen in Fig. 3(a). RQoS-UCB, RCA and Q-learning converge toward the best channel selection because of their ability to learn the band with the best weighted combination of quality and availability but with different convergence speed. RQoS-UCB achieves 87% of the best channel selection policy rate in 1000 frames while more than 4000 frames is needed to achieve less than 87% of this value for RCA and Q-learning policies. Moreover, Fig. 3(b) also depicts that RQoS-UCB outperforms RUCB, in [10], in convergence speed which is relatively far from the best channel and the best opportunistic transmission policies. This behavior is probably due to the exploitation-exploration epoch structure which has an exponentially growing length. Hence, if the channel selected is not the optimal one, the policy in [10] has difficulties to change in a real communication scenario. On the other hand, the exploration and exploitation are done in the *regenerative cycle* of a near constant length in our policy, which makes it more suitable to try other channels and hence converge faster to the optimal one. These results also demonstrate the efficiency of controlling the learning phase with two rewards instead of one when channels are characterized by not only their availability but also by their quality. As expected, a simple round-robin technique cannot compete in this scenario.

In Fig. 3(c), the average symbol error rate (SER) is investigated. The SER of RQoS-UCB, RCA and Q-learning converge toward the SER obtained with the best channel selection policy, i.e. $5 \cdot 10^{-3}$ which is the SER of QPSK signaling under 9 dB of SNR, i.e. SNR of band 3. Again, the slower convergence characteristic of Q-learning and RCA, w.r.t. RQoS-UCB, can be emphasized in this figure. This figure also reveals the limitation of the RUCB policy, which likely does not select the band with the best SNR as often as the competing policies, i.e. RQoS-UCB, RCA and Q-learning. We can even notice that the SER of the best opportunistic transmission policy is higher than the RQoS-UCB policy, because it selects a sub-optimal bands to continue transmission when optimal band is occupied. Indeed, the best opportunistic policy is an ideal scheme which exploits at each time the best (in

Learning Algorithms	Running time complexity	Space complexity	Selection Criteria	Theoretical guarantee	Convergence speed
Q-learning [27]	$\mathcal{O}(N(6K+3))$	$\mathcal{O}(N(3K+3))$	availability and quality	Rested and Restless	Medium
RCA [11]	$\mathcal{O}(N(3K+3))$	$\mathcal{O}((4K+5))$	availability only	Rested and Restless	Fast
RQoS-UCB	$\mathcal{O}(N(8K+6))$	$\mathcal{O}((4K+7))$	availability and quality	Rested and Restless	Fast

TABLE III
ALGORITHMS COMPLEXITY

term of quality) available channel, but not necessarily the best one globally. Hence, rather than stopping transmission because the best channel in term of quality is occupied, it goes to a suboptimal (in quality) channel but available. Hence, during the transmission the secondary link experiences a degraded SINR which increases the SER compared to the case where it would have used the optimal channel, however it allows to transmit anyway. In (5), the achievable throughput depends not only on SINR but also on the transmission opportunities exploited. In other words, transmitting on a link with a better SINR leads to an increase of throughput but attenuated by the $\log$ function. On the other hand, if the link is of bad quality, this results in a decrease of throughput but marginally due to the $\log$ function which can be compensated by using more often this link (if it is more available than another one with a better SINR). This what explains that the best opportunistic policy may have a larger SER than the best channel selection policy but a larger throughput than the latter, Fig. 3(b).

Complexity Analysis

To conclude this first part, a comparative study, in terms of complexity, optimality and convergence speed, between the three learning algorithms, i.e. RQoS-UCB, Q-learning and RCA, has been summarized in Table III. The running time complexity is the number of operation performed and space complexity is related to the storage space (memory) needed to run [28]. For the running time complexity, RCA, RQoS-UCB and Q-learning policies behave in $\mathcal{O}(NK)$ for large N and K, where N and K are the number of time slots and channels. Time complexity of these algorithms are comparable however, RQoS-UCB performs better than the others as seen in numerical analysis. Time complexity of reinforcement learning is negligible and it is approximately 1% of the sensing time complexity of energy detector that is also required for OSA as stated in [19]. On the other hand, it is clear that space complexity (expected memory requirement) is the drawback of Q-learning that needs to store all past observations contrary to RCA and RQoS-UCB policies whose complexity is about $O(K)$.

B. Multiple users case

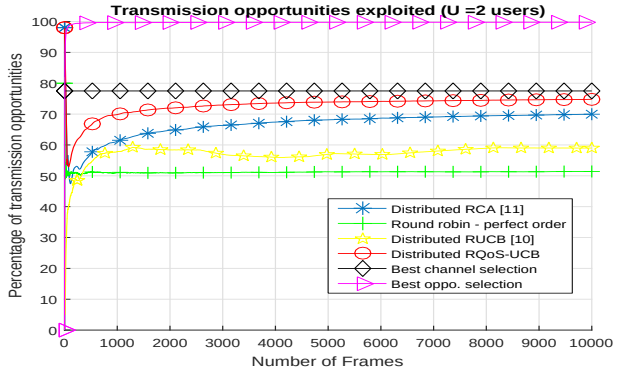
In this part, RQoS-UCB is implemented as presented in Algorithm 2 and compared to *distributed* RCA, distributed RUCB, and round robin-perfect order (R2PO) policy. Only the single user version of RCA can be found in literature, however multi-user version can be implemented easily following the same structure as distributed RQoS-UCB leading to the *distributed RCA* in the following. Moreover, the *round robin-perfect order* policy plays each channel consecutively without suffering from collisions, and *best channel selection* and *best opportunistic selection* are defined in a same manner as for the single-user policy.

Fig. 4 presents the average percentage of transmission opportunities exploited by the RQoS-UCB, RCA, RUCB, R2PO and the best channel selection policies w.r.t. the number of frames for 2 SUs, Fig. 4(a), and w.r.t. the number of users in Fig. 4(b). The best opportunistic selection policy logically upper bounds the performance. From Fig. 4(a), one can remark that distributed RQoS-UCB policy achieves better performance compared to RCA converging towards the best channel selection policy due to separation of two optimization criteria, i.e. availability and quality, and both significantly outperform R2PO which finds less transmission opportunities even if all SUs have a perfect knowledge about each others' actions. This is also confirmed by Fig. 4(b) where the percentage of transmission opportunities is investigated w.r.t. the number of SUs. The distributed RQoS-UCB outperforms RCA in general, and also percentage of opportunities decreases for both when the number of SUs in the network increases and achieves similar performance than R2PO when 9 secondary transceivers are considered. Note that R2PO is not an interesting solution for CR systems, because requiring a predefined agreement or information exchange among SUs. Hence, the distributed implementation of RQoS-UCB is able to find sufficiently high number of transmission opportunities without additional signaling overhead.

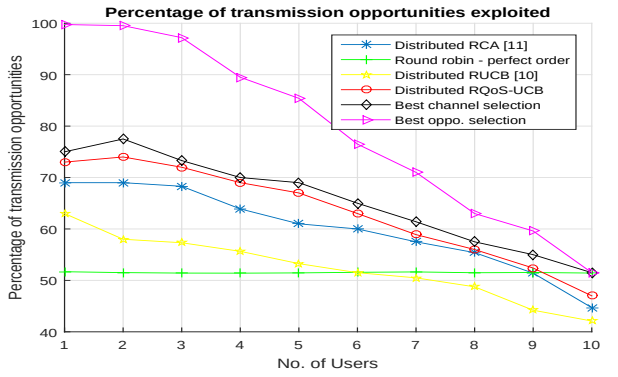
The average achievable throughput is investigated in Fig. 5 w.r.t. the number of frames for 2 users in the network, in Fig. 5(a), and w.r.t. the number of SUs in Fig. 5(b). Distributed RQoS-UCB converges rapidly towards the best channel selection policy while distributed RCA achieves similar performance after a larger learning time. Our policy, RQoS-UCB, outperforms RUCB for all values of the number of users and converges to the same performance than RCA when the number of users is larger than or equal to 6, Fig. 5(b). Note that these particular values are not absolute but depend on the scenario considered and the total number of primary users' channels. Hence, our proposal is not only beneficial for up to 6 SUs but can be greater if the number of channels is higher.

Average SER obtained with several policies is investigated in Fig. 6, w.r.t. the number of frames and 2 SUs in Fig. 6(a), and w.r.t. the number of SUs in 6(b). The SER of distributed RQoS-UCB converges towards the SER obtained with the best channel selection policy. Like in Figs. 3, 4 and 5, the distributed RUCB [10] and RCA [11] converge to an optimal band but with a slower convergence speed due to the single reward mixing the availability and quality, and hence the average SER achieved with RUCB and RCA is poorer than the one with RQoS-UCB. We remark in Fig. 6(b) that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

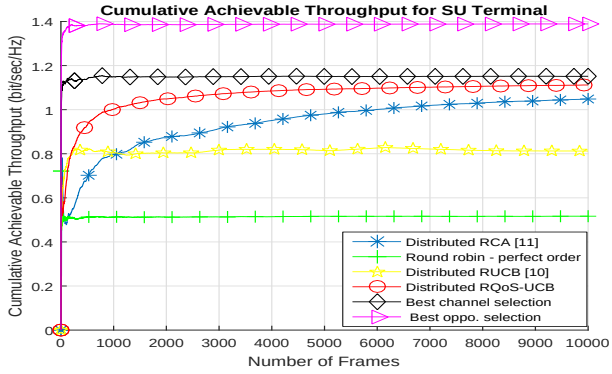


(a) Opportunities w.r.t. the number of frames

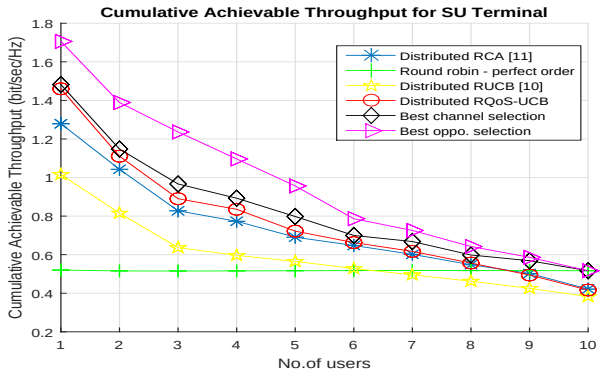


(b) Opportunities w.r.t. the number of SUs

Fig. 4. Opportunities transmission percentage for 2 SUs w.r.t. the number of frames a) and w.r.t. the number of SUs operating in the primary network b).



(a) Throughput w.r.t. the number of frames



(b) Throughput w.r.t. the number of SUs

Fig. 5. Average achievable throughput for 2 SUs w.r.t. the number of frames a) and w.r.t. the number of SUs operating in the primary network b).

RQoS-UCB matches the best channel selection policy which lower bounds the SER of all other strategies when the number of users is varying. The SER of RQoS-UCB increases as the number of users increases, since the number of channels with worse quality increases but is still lower than the SER of RCA and RUCB. At a point, i.e. more than 7 SUs in network, the difference between all learning approaches and round robin-perfect order becomes negligible as they finish to select an important proportion of the same channels. Furthermore, if K would be much larger, the number of users at which our policy would offer better performance than round-robin would also be larger.

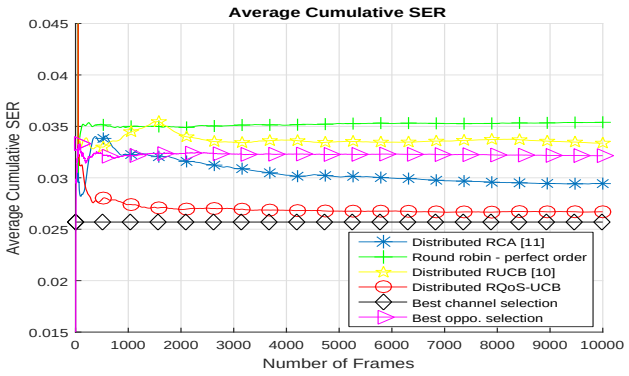
Fairness in Channel Access

One of the important features of the proposed restless RQoS-UCB policy is that it does not favor one specific user over another in order to access optimal arms. In the proposed distributed RQoS-UCB approach, each user has an equal chance to sense and transmit in any one of the $U-$optimal channels. Fig. 7(a) illustrates the percentage of opportunities exploited and Fig. 7(b) the optimal arm selection percentage for 4 SUs when $K = 10$ channels w.r.t. the number of frames. As it can be observed, each user exploits approximately the same amount of transmission opportunities and select the optimal arm more or less the same proportion of time. This demonstrate that the proposed distributed RQoS-UCB scheme is indeed fair in user allocation.

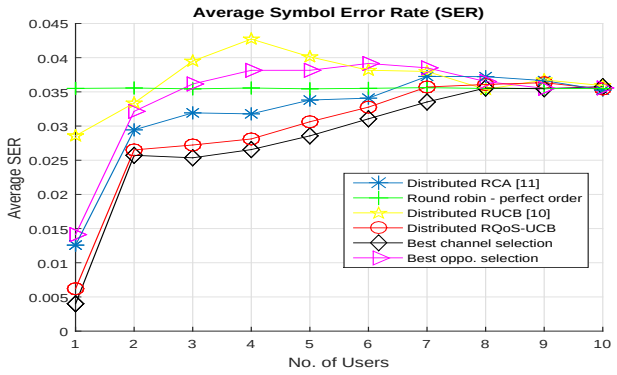
VI. Conclusion

This paper has dealt with OSA problem in multi-user infrastructure-less cognitive wireless network. A new machine learning algorithm, called distributed RQoS-UCB, has been proposed to address the multi-user OSA problem modeled as a restless Markov MAB formulation. The proposed policy takes into account a quality information metric instead of only availability for most of algorithms in literature. Moreover, we have proved that our proposed policies achieve a logarithmic order regret uniformly over time for the restless Markov MAB. In cognitive radio applications, the ability to learn on a different criteria than the traditional free or occupied status of a channel is of particular interest in order to improve the QoS of SUs transmissions. Using the soft-output of ED as a quality metric for the sensed band, we have shown that the proposed policies are able to achieve a larger throughput than the state-of-the-art algorithms which suffer from a larger convergence time compared to the proposed policies. The idea proposed in this paper can be used to learn on many other criteria such as energy efficiency or actual SINR on the secondary link and will be investigated in further works. Moreover, our model ignores dynamic traffic at the secondary nodes and extension to a queueing-theoretic formulation is left for future work.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

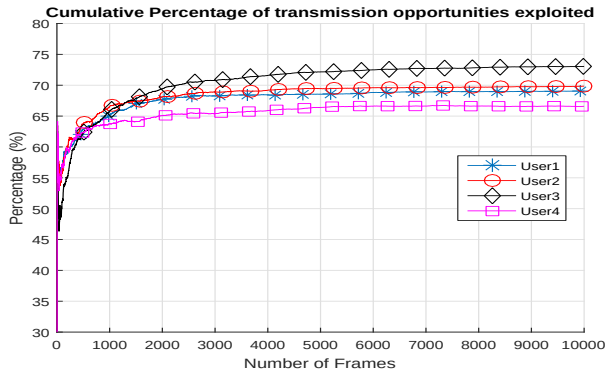


(a) SER w.r.t. the number of frames

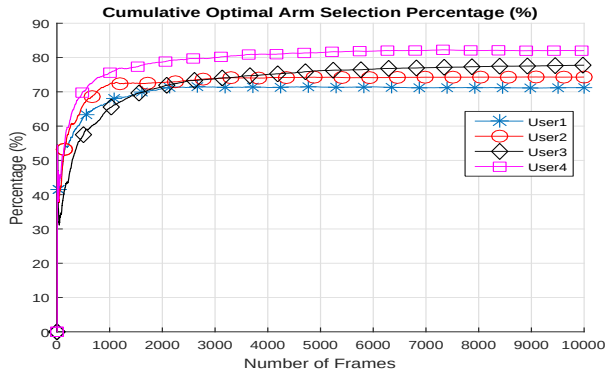


(b) SER w.r.t. the number of SUs

Fig. 6. Average SER for 2 SUs w.r.t. the number of frames a) and w.r.t. the number of SUs operating in the primary network b).



(a) Opportunities w.r.t. the number of frames



(b) Optimal arm selection percentage w.r.t. the number of frames

Fig. 7. Fairness analysis for 4 SUs implementing RQoS-UCB policy w.r.t. the number of frames.

APPENDIX A
PROOF OF THEOREM 1

In order to bound the regret, we need to bound the expected number of blocks, $\mathbb{E}[F^i(b)]$, for any suboptimal band $i > 1$. Let l being a positive integer and $\mathcal{A}(b)$ the action performed by policy $\mathcal{A}$ in block b. $n_2(b)$ represents total time spent in SB2 block up to block b. Following the steps as in [4], the number of blocks a band i has been visited up to block b can be expressed as

$$F^i(b) = 1 + \sum_{m=K+1}^{b} \mathbb{1}\left((\mathcal{A}(m) = i)\right) \tag{14}$$

$$F^i(b) = l + \sum_{m=K+1}^{b} \mathbb{1}\left(\left(\mathcal{A}(m) = i, F^i(m-1) \geq l\right)\right) \tag{15}$$

$$= l + \sum_{m=K+1}^{b} \mathbb{1}\Big(B^1\left(T_2^1(n_2(m)), n_2(m)\right) \leq B^i\left(T_2^1(n_2(m)), n_2(m)\right), F^i(m) \geq l \Big) \tag{16}$$

$$\leq l + \sum_{m=K+1}^{b} \mathbb{1}\left(\exists \omega^i : l \leq \omega^i \leq n_2(m), B^i(\omega^i, n_2(m)) > \mu_1^R\right) + \mathbb{1}\left(\left(\exists \omega^1 : 1 \leq \omega^1 \leq n_2(m), B^1(\omega^1, n_2(m)) \leq \mu_1^R\right)\right) \tag{17}$$

where (15) comes from the fact that each band has been sensed at least l blocks up to block b. (16) comes from the reason why suboptimal band i is chosen up to $n_2(m-1)$ time at the end of block $m-1$, i.e. the index of an optimal band at block $m-1$, i.e. $B^1\left(T_2^1(n_2(m-1)), n_2(m-1)\right)$, is below the index of the suboptimal band i. Moreover (16) is upper bounded by (17) because these two conditions are not exclusive. Taking the expectation on both sides and using union bound we get:

$$\begin{aligned}\mathbb{E}[F^i(b)] &\leq l + \sum_{m=K+1}^{b} \sum_{\omega^i=n_2(l)}^{n_2(m-1)} \mathbb{P}(B^i(\omega^i, n_2(m)) > \mu_1^R) \\ &+ \sum_{m=K+1}^{b} \sum_{\omega^1=1}^{n_2(m-1)} \mathbb{P}(B^1(\omega^1, n_2(m)) \leq \mu_1^R) \\ &\leq l + \sum_{t=1}^{n_2(b)} \sum_{\omega^i=l}^{t-1} \mathbb{P}(B^i(\omega^i, t) > \mu_1^R) \\ &+ \sum_{t=1}^{n_2(b)} \sum_{\omega^1=1}^{t-1} \mathbb{P}(B^1(\omega^1, t) \leq \mu_1^R)\end{aligned} \tag{18}$$

The summation over t starts from 1 instead of $K+1$ because it does not change the validity of the upper bound. Note that channel 1 is optimal in terms of mean reward, μ_1^R, i.e. both in vacancy and quality. G^1 is hence the empirical mean of the quality reward of this channel, it does not mean necessarily that $G^1 = G^{q_1}_{\max}$. Moreover, let's remind that $\Delta\mu_i^R = \mu_1^R - \mu_i^R$. let's choose $l = \left\lceil \frac{4\alpha \ln n}{(\Delta\mu_i^R)^2} \right\rceil$, take expectation on both sides and relaxing the outer sum in (18) from $n_2(b)$ to ∞, and proceed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

from (18):

$$\mathbb{E}[F^i(b)] \leq l + \sum_{t=1}^{\infty}\sum_{\omega^i=l}^{t-1} \mathbb{P}(B^i(\omega^i,t) > \mu_1^R) + \sum_{t=1}^{\infty}\sum_{\omega^1=1}^{t-1} \mathbb{P}(B^1(\omega^1,t) \leq \mu_1^R) \quad (19)$$

Let's start with the first part of (19), i.e. $\mathbb{P}(B^i(\omega^i,t) > \mu_1^R)$. By writing $\mu_1^R = \mu_i^R + \Delta\mu_i^R$ and replacing the $B^i(\omega^i,t)$ by its expression, we get

$$\mathbb{P}(B^i(\omega^i,t) > \mu_1^R) = \mathbb{P}\left(\bar{S}^i(\omega^i) - \frac{\beta M^i(\omega^i)\ln(t)}{\omega^i} + \sqrt{\frac{\alpha\ln(t)}{\omega^i}} > \mu_i^R + \Delta\mu_i^R\right) \quad (20)$$

For sake of notational simplicity, let's note $D^i(\omega^i,t) = \frac{\beta M^i(\omega^i)\ln(t)}{\omega^i}$. Moreover, using $l = \left\lceil \frac{4\alpha\ln n}{(\Delta\mu_i^R)^2} \right\rceil$ and $\omega^i \geq l$, the third term in (20) can be upper-bounded by:

$$\sqrt{\frac{\alpha\ln t}{\omega^i}} \leq \sqrt{\frac{\alpha\ln t}{l}} \leq \sqrt{\frac{\alpha\ln t(\Delta\mu_i^R)^2}{4\alpha\ln t}} = \frac{\Delta\mu_i^R}{2}$$

Substituting this last bound into (20) and because all terms are positive we get

$$\mathbb{P}(B^i(\omega^i,t) > \mu_1^R) \leq \mathbb{P}\left(\bar{S}^i(\omega^i) - \mu_i^R > \frac{\Delta\mu_i^R}{2} + D^i(\omega^i,t)\right) \quad (21)$$

Let, $O_q^i(t)$ being the number of times reward r_q^i associated with state q of arm i has been observed up to time t, hence $\bar{S}^i(\omega^i) = \frac{1}{\omega^i}\sum_{q\in S^i} r_q^i O_q^i(\omega^i)$. Following from (21):

$$\begin{aligned}\mathbb{P}\left(\bar{S}^i(\omega^i) - \mu_i^R \geq \frac{\Delta\mu_i^R}{2} + D^i(\omega^i,t)\right) &= \mathbb{P}\Bigg(\sum_{q\in S^i}\left(-r_q^i O_q^i(\omega^i) + \omega^i G_q^i r_q^i \pi_q^i\right) \\ &\leq -\omega^i\left(\frac{\Delta\mu_i^R}{2} + D^i(\omega^i,t)\right)\Bigg)\end{aligned} \quad (22)$$

Following same steps as of [4], (22) is upper bounded as:

$$\leq \sum_{q\in S^i} N_{\mathbf{h}^i} \exp\left(-\frac{\omega^i\left(\frac{\frac{\Delta\mu_i^R}{2}+D^i(\omega^i,t)}{r_q^i|S^i|G_q^i\hat{\pi}_q^i}\right)^2\gamma^i}{28}\right) \quad (23)$$

where $|S^i|$ is the arm i state space cardinality, $\hat{\pi}_q^i = \max\left\{\pi_q^i, 1-\pi_q^i\right\}$ and $\hat{\pi}_{\max} = \max_{i\in\mathbb{K}}\hat{\pi}_q^i$. Moreover, (23) follows from Theorem 3.3 from [29] by considering $n = \omega^i$, $f(X_t^i) = \frac{\mathbb{1}(S_t^i=q) - G_q^i\pi_q^i}{G_q^i\hat{\pi}_q^i}$. The Theorem 3.3 from [29] conditions are fulfilled if $G_q^i \geq \frac{1}{\hat{\pi}_{\max}+\pi_q^i}$. Consider an initial distribution $\mathbf{h}^i$ as defined in [4] and eigenvalue gap γ^i for the ith arm, then

$$N_{\mathbf{h}^i} = \left\|\left(\frac{h_q^i}{\pi_q^i}, q\in S^i\right)\right\|_2 \leq \sum_{q\in S^i}\left\|\frac{h_q^i}{\pi_q^i}\right\|_2 \leq \frac{1}{\pi_{\min}}, \quad (24)$$

In order to lighten the notation, let us redefine the following variables. $G_{\max} \equiv G_{\max}^{q_1}$ but the superscript is dropped and $r_{\max} = \max_{q\in S^i, i\in\mathbb{K}} r_q^i$. Moreover, let's define $M_{\min} = \min_{i\in\mathbb{K}} M^i\left(\omega^i\right)$, $\omega_{\max} = \max_{i\in\mathbb{K}}\omega^i$ and $\omega_{\min} = 1$. From (23),

$$\begin{aligned}&\leq \frac{|S^i|}{\pi_{\min}}\exp\left(-\frac{\omega^i\left(\frac{\frac{\Delta\mu_i^R}{2}+D^i(\omega^i,t)}{|S^i|r_q^iG_q^i\hat{\pi}_q^i}\right)^2\gamma^i}{28}\right) \\ &\leq \frac{|S^i|}{\pi_{\min}} t^{-\frac{\Delta\mu_i^R\beta M_{\min}\gamma_{\min}}{28S_{\max}^2 r_{\max}^2 G_{\max}^2\hat{\pi}_{\max}^2}}\end{aligned} \quad (25)$$

where (25) is achieved by noting that $\exp\left(-\frac{\left(\Delta\mu_i^R\right)^2\gamma_{\min}\omega_{\min}}{112S_{\max}^2 r_{\max}^2 G_{\max}^2\hat{\pi}_{\max}^2}\right) \geq 0$. Inserting (25) into first part of (19), we get

$$\sum_{t=1}^{\infty}\sum_{\omega^i=l}^{t-1}\mathbb{P}(B^i(\omega^i,t) \geq \mu_1^R) \leq \frac{|S^i|}{\pi_{\min}}\sum_{t=1}^{\infty} t^{-2} \quad (26)$$

where, $\beta \geq 84S_{\max}^2 r_{\max}^2 G_{\max}^2\hat{\pi}_{\max}^2/\left(\gamma_{\min}\Delta\mu_i^R M_{\min}\right)$ is considered to obtain (26).

Similarly, we prove the second part of (19):

$$\mathbb{P}(B^1(\omega^1,t) \leq \mu_1^R) = \mathbb{P}\left(\bar{S}^1(\omega^1) - \frac{\beta M^1(\omega^1)\ln t}{\omega^1} + \sqrt{\frac{\alpha\ln t}{\omega^1}} \leq \mu_1^R\right) \quad (27)$$

Let, $C(\omega^1,t) = \sqrt{\frac{\alpha\ln t}{\omega^1}}$ and $D^1(\omega^1,t) = \frac{\beta M^1(\omega^1)\ln(t)}{\omega^1}$ for notation simplification.

$$\mathbb{P}(B^1(\omega^1,t) \leq \mu_1^R) = \mathbb{P}\left(\bar{S}^1(\omega^1) - \mu_1^R \leq -(C(\omega^1,t) - D^1(\omega^1,t))\right) \quad (28)$$

Similarly as shown in (23), we obtain

$$\begin{aligned}&\mathbb{P}\left(\bar{S}^1(\omega^1) - \mu_1^R \leq -(C(\omega^1,t) - D^1(\omega^1,t))\right) \\ &= \mathbb{P}\Bigg(\sum_{q\in S^1}(r_q^1 O_q^1(\omega^1) - \omega^1 G_q^1 r_q^1\pi_q^1) \\ &\qquad \leq -\omega^1(C(\omega^1,t) - D^1(\omega^1,t))\Bigg)\end{aligned} \quad (29)$$

$$\leq \sum_{q_1\in S^1} N_{\mathbf{h}^1}\exp\left(-w^1\frac{\left(C(\omega^1,t) - D^1(\omega^1,t)\right)^2\gamma^1}{28\left(|S^1|G_q^1r_q^1\hat{\pi}_q^1\right)^2}\right) \quad (30)$$

where, (30) follows from Theorem 3.3 from [29] with $C(\omega^1,t) - D^1(\omega^1,t)$ can be proved to be positive from a certain time. Indeed, $C(\omega^1,t) - D^1(\omega^1,t) = \sqrt{\frac{\ln t}{w^1}}\left(\sqrt{\alpha} - \beta M^1\left(w^1\right)\sqrt{\frac{\ln t}{w^1}}\right)$ and $\exists A\in\mathbb{N}$, $\exists\epsilon > 0$ such that $\forall t > A$, $\sqrt{\frac{\ln t}{w^1}} < \epsilon$. This can be justified by the fact that $\ln t$ grows always slower than t and so w^1 which can be viewed as a fraction of t. Using inequalities in (24) and after

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

replacing $C\left(w^1,t\right)$ and $D^1\left(w^1,t\right)$ by their values and after some calculus we get

$$\mathbb{P}(B^1(\omega^1,t)\leq\mu_1^R)\leq\frac{\left|S^1\right|}{\pi_{\min}}t^{-\frac{\gamma_{\min}\left(\alpha-2\sqrt{\alpha}\beta M^1(\omega^1)\sqrt{\frac{\ln t}{\omega^1}}\right)}{28(S_{\max}G_{\max}r_{\max}\tilde{\pi}_{\max})^2}}$$
$$t^{-\frac{\gamma_{\min}\left(\beta M^1(\omega^1)\right)^2\ln t}{28(S_{\max}G_{\max}r_{\max}\tilde{\pi}_{\max})^2\omega^1}} \quad (31)$$

Moreover, $\exists A\in\mathbb{N},\exists\epsilon>0$, such that $\forall t>A$, $\frac{\ln t}{\omega^1}<\sqrt{\frac{\ln t}{\omega^1}}<\epsilon<1$. We get

$$\mathbb{P}(B^1(\omega^1,t)\leq\mu_1^R)\leq\frac{\left|S^1\right|}{\pi_{\min}}t^{-\frac{\gamma_{\min}\left(\alpha-2\sqrt{\alpha}\beta M^1(\omega^1)\right)}{28(S_{\max}G_{\max}r_{\max}\tilde{\pi}_{\max})^2}}$$
$$\leq\frac{\left|S^1\right|}{\pi_{\min}}t^{-\frac{\gamma_{\min}\left(\alpha-2\sqrt{\alpha}\beta M_{\max}\right)}{28(S_{\max}G_{\max}r_{\max}\tilde{\pi}_{\max})^2}} \quad (32)$$

where $M_{\max}=\max_{i\in\mathbb{K}}M^i\left(\omega^i\right)$ and where from (31) to (32) the second term in t is upper bounded by 1. By choosing α such that $\frac{\gamma_{\min}\left(\alpha-2\sqrt{\alpha}\beta M_{\max}\right)}{28(S_{\max}G_{\max}r_{\max}\tilde{\pi}_{\max})^2}\geq 3$ we obtain

$$\mathbb{P}(B^1(\omega^1,t)\leq\mu_1^R)\leq\frac{\left|S^1\right|}{\pi_{\min}}t^{-3} \quad (33)$$

Replacing (33) into second part of (19), we get

$$\sum_{t=1}^{\infty}\sum_{\omega^1=1}^{t-1}\mathbb{P}(B^1\left(\omega^1,t\right)\leq\mu_1^R)\leq\frac{|S^1|}{\pi_{\min}}\sum_{t=1}^{\infty}\sum_{\omega^1=1}^{t}t^{-3}$$
$$=\frac{|S^1|}{\pi_{\min}}\sum_{t=1}^{\infty}t^{-2} \quad (34)$$

Then the bound follows from combining (26) and (34):

$$\mathbb{E}[F^i(b(n))|b(n)=b]\leq\frac{4\alpha\ln n}{(\Delta\mu_i^R)^2}+\frac{|S^1|+|S^i|}{\pi_{\min}}\sum_{t=1}^{\infty}t^{-2} \quad (35)$$

The SB2 block begins with the state ζ^i and ends with a return to the same state. The total number of plays of sub-optimal arm i at the end of block $b(n)$ is estimated by considering the observations acquired in: i) the total number of plays of sub-optimal arm i during SB2 sub-block (upper bounded by $\frac{1}{\pi_{\min}^i}$), ii) the total number of plays in SB1 before entering in SB2 (upper bounded by $\Omega_{\max}^i$), and iii) one more play during SB3. Thus, we have

$$\mathbb{E}\left[T^i(n)\right]\leq\left(\frac{1}{\pi_{\min}^i}+\Omega_{\max}^i+1\right)\mathbb{E}\left[F^i(b(n))\right]$$

Thus,

$$\sum_{i\in K}(\mu_1^R-\mu_i^R)\mathbb{E}[T^i(n)]\leq Z_1\ln n+Z_2,$$

where,

$$Z_1=\sum_{i:\mu_i^R<\mu_1^R}\left(\frac{1}{\pi_{\min}^i}+\Omega_{\max}^i+1\right)\frac{4\alpha}{\Delta\mu_i^R}$$
$$Z_2=\sum_{i:\mu_i^R<\mu_1^R}\left(\frac{1}{\pi_{\min}^i}+\Omega_{\max}^i+1\right)\Delta\mu_i^R$$
$$\left[\frac{|S^1|+|S^i|}{\pi_{\min}}\sum_{t=1}^{\infty}t^{-2}\right]$$

APPENDIX B
PROOF OF THEOREM 2

Assume that the regenerative states are denoted by $\zeta=[\zeta^1,\cdots,\zeta^K]$. The expectation w.r.t. the modified sample path is defined as $\mathbb{E}_\zeta$. Let n^b be the time at the end of the last completed block $b(n)$ for all SUs.

$$\Phi^R(n)=n\mu_1^R-\mathbb{E}\left[\sum_{t=1}^{n}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}(t)r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}(t)\right]$$
$$\Phi^R(n)=\mu_1^R\mathbb{E}_\zeta[n^b]-\mathbb{E}_\zeta\left[\sum_{t=1}^{n^b}r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right]$$
$$+\mu_1^R\mathbb{E}_\zeta[n-n^b]-\mathbb{E}_\zeta\left[\sum_{t=n^b+1}^{n}r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right]$$
$$=\left\{\mu_1^R\mathbb{E}_\zeta[n^b]-\sum_{i=1}^{K}\mu_i^R\mathbb{E}_\zeta\left[T^i(n)\right]\right\}$$
$$+\sum_{i=1}^{K}\mu_i^R\mathbb{E}_\zeta\left[T^i(n)\right]-\mathbb{E}_\zeta\left[\sum_{t=1}^{n^b}r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right]$$
$$+\mu_1^R\mathbb{E}_\zeta\left[n-n^b\right]-\mathbb{E}_\zeta\left[\sum_{t=n^b+1}^{n}r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right] \quad (36)$$

First difference in (36) is bounded logarithmically with the help of Theorem 1 as:

$$\mu_1^R\mathbb{E}_\zeta[n^b]-\sum_{i=1}^{K}\mu_i^R\mathbb{E}_\zeta\left[T^i(n)\right]$$
$$\leq\sum_{i=2}^{K}\left(\mu_1^R-\mu_i^R\right)\mathbb{E}_\zeta\left[T^i(n)\right]\leq Z_1\ln n+Z_2 \quad (37)$$

We have to bound only the two remaining differences in (36). We can bound second difference in (36) by following same steps as of Theorem 2 in [4], as:

$$\sum_{i=1}^{K}\mu_i^R\mathbb{E}_\zeta\left[T^i(n)\right]-\mathbb{E}_\zeta\left[\sum_{t=1}^{n^b}r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right]$$
$$\leq\sum_{i=2}^{K}\mu_i^R(\Omega_{\max}^i+1)\mathbb{E}_\zeta\left[F^i(b(n))\right]$$
$$+\mu_1^R\Omega_{\max}^1\sum_{i=2}^{K}\mathbb{E}_\zeta\left[F^i(b(n))\right] \quad (38)$$

by following same steps as of Theorem 2 in [11]. Finally, the last part in (36) is bounded as:

$$\mu_1^R\mathbb{E}_\zeta\left[n-n^b\right]-\mathbb{E}_\zeta\left[\sum_{t=n^b+1}^{n}r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t)}\right]$$
$$\leq\mu_1^R\left(\frac{1}{\pi_{\min}}+\max_{i\in\{1,\ldots,K\}}\Omega_{\max}^i+1\right). \quad (39)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

From (37), (38), (39) and Theorem 1, the upper bound on the regret of RQoS-UCB policy is:

$$\Phi^R(n) \leq Z_1 \ln n + Z_2 + \sum_{i=2}^{K} \mu_i^R (\Omega_{\max}^i + 1)\mathbb{E}_\zeta \left[F^i(b(n))\right] + \mu_1^R \Omega_{\max}^1 \sum_{i=2}^{K} \mathbb{E}_\zeta[F^i(b(n))] + \mu_1^R \left(\frac{1}{\pi_{\min}} + \max_{i\in\{1,\ldots,K\}} \Omega_{\max}^i + 1\right) \leq Z_3 \ln n + Z_4$$

where $Z_1, Z_2, Z_3, Z_4, Z_5, Z_6$ and Z_7 are as stated in Theorems 1 and 2 and obtained by identification with previous quantities and the proof is complete.

APPENDIX C
PROOF OF THEOREM 3

The total number of frames $f(n)$ up to time n for which the distributed RQoS-UCB policy suggested suboptimal arms i to sense is bounded in the same way as for the RQoS-UCB policy in a single-user restless Markov MAB setting. For convenience, let $T^i(n) := \sum_{j=1}^{U} T^{i,j}(n)$ and $\sum_{i=1}^{K} T^i(n) = nU$, since each SU selects at least one channel to sense in each slot and there are U SUs. Let, $\forall j, i : j \in \{1, \cdots, U\}$ and $i \in \{U+1, \cdots, K\}$ denote the set of optimal and suboptimal channel respectively. Following the same steps as in [14], [16], the total expected number of blocks $\mathbb{E}\left[F^{i,j}(f(n))\right]$ up to frame $f(n)$ for which SU j implementing distributed RQoS-UCB policy selected suboptimal channels i can be bounded as:

$$\mathbb{E}\left[F^{i,j}(f(n))\right] = \mathbb{P}\left[B^j\left(f^j(n), n\right) \leq B^i\left(f^i(n), n\right)\right] \leq \sum_{j=1}^{U} \mathbb{P}\left[B^j\left(f^j(n), n\right) \leq B^i\left(f^i(n), n\right)\right]$$

For one user, Theorem 1 gives the expected number of blocks used to sense a suboptimal band:

$$\mathbb{E}\left[F^{i,1}(b(n))\right] = \sum_{t=1}^{n} \mathbb{1}\left(B^1\left(f^1(t), t\right) \leq B^i\left(f^i(t), t\right)\right) \leq \frac{4\alpha \ln n}{(\Delta\mu_i^R)^2} + \left[\frac{|S^1| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}\right]$$

Thus, we have for U optimal channels for SU j:

$$\mathbb{E}\left[F^{i,j}(f(n))\right] \leq \sum_{j=1}^{U} \left[\frac{4\alpha \ln n}{(\Delta\mu_{i,j}^R)^2} + \left[\frac{|S^j| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}\right]\right]$$

The total number of senses of the suboptimal channel i at the end of the frame $f(n)$ is estimated by considering the observations acquired during i) the entire block duration (i.e. SB1, SB2 and SB3), and ii) after the end of SB3 and before finishing of current frame $f(n)$ within the fixed time slot W. Thus, we have

$$\mathbb{E}\left[T^{i,j}(n)\right] \leq \sum_{j=1}^{U} \left(\frac{1}{\pi_{\min}^i} + \Omega_{\max}^i + W\right) \left[\frac{4\alpha \ln n}{(\Delta\mu_{i,j}^R)^2} + \left[\frac{|S^j| + |S^i|}{\pi_{\min}} \sum_{t=1}^{\infty} t^{-2}\right]\right]$$

APPENDIX D
PROOF OF THEOREM 4

Let $\{1, \cdots, U\}$ and $\{U+1, \cdots, K\}$ denote the set of optimal and suboptimal channels, respectively, and n^b is the time at the end of the last completed block $b(n)$ as detailed in Fig. 1. Following the same spirit as in the proof of Theorem 2, we have

$$\begin{aligned}\Phi^M(n) &= \sum_{j=1}^{U} n\mu_j^R - \sum_{j=1}^{U} \mathbb{E}\left[\sum_{t=1}^{n} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}(t) r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}(t)\right] \\ &= \left\{\mathbb{E}_\zeta[n^b] \sum_{j=1}^{U} \mu_j^R - \sum_{j=1}^{U}\sum_{i=1}^{K} \mu_i^R \mathbb{E}_\zeta\left[V^{i,j}(n)\right]\right\} \\ &+ \left\{\sum_{j=1}^{U}\sum_{i=1}^{K} \mu_i^R \mathbb{E}_\zeta\left[V^{i,j}(n)\right] - \sum_{j=1}^{U} \mathbb{E}_\zeta\left[\sum_{t=1}^{n^b} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}\right]\right\} \\ &+ \left\{\mathbb{E}_\zeta[n - n^b] \sum_{j=1}^{U} \mu_j^R - \sum_{j=1}^{U} \mathbb{E}_\zeta\left[\sum_{t=n^b+1}^{n} r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j} G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}\right]\right\} \end{aligned} \tag{40}$$

where $V^{i,j}(n)$ is the total number of times where an SU j is the only one to sense and access the channel i up to time n. Working with the first part of (40) we have:

$$\begin{aligned}&\left\{\mathbb{E}_\zeta[n^b] \sum_{j=1}^{U} \mu_j^R - \sum_{j=1}^{U}\sum_{i=1}^{K} \mu_i^R \mathbb{E}_\zeta\left[V^{i,j}(n)\right]\right\} \\ &\leq \sum_{i=1}^{U} \mu_i^R \left(\mathbb{E}_\zeta[n^b] - \mathbb{E}_\zeta\left[V^i(n)\right]\right) \\ &\leq \mu_1^R \left(U\mathbb{E}_\zeta[n^b] - \sum_{i=1}^{U} \mathbb{E}_\zeta\left[V^i(n)\right]\right) \qquad (41) \\ &= \mu_1^R \left(U\mathbb{E}_\zeta[n^b] + \mathbb{E}_\zeta[C_o(n)] - \sum_{i=1}^{U} \mathbb{E}_\zeta\left[T^i(n)\right]\right) \\ &\leq \mu_1^R \left(\mathbb{E}_\zeta[C_o(n)] + \sum_{i=U+1}^{K} \mathbb{E}_\zeta\left[T^i(n^b)\right]\right) \qquad (42) \\ &\leq \mu_1^R \left(U(\mathbb{E}[\Upsilon(U,U)] + 1)\,\mathbb{E}[T_w(n)] + \sum_{i=U+1}^{K} \mathbb{E}_\zeta\left[T^i(n)\right]\right) \qquad (43)\end{aligned}$$

where in (41), we use the fact that $\mathbb{E}_\zeta\left[V^i(n)\right] = \sum_{j=1}^{U} \mathbb{E}_\zeta\left[V^{i,j}(n)\right]$ because of $V^i(n) < \mathbb{E}_\zeta[n^b]$, since the total number of time a unique SU occupies the channel i is at most $\mathbb{E}_\zeta[n^b]$. In (42), we use the fact that the total number of collisions in U optimal channels is defined as $C_o(n) = \sum_{i=1}^{U} \left(T^i(n) - V^i(n)\right)$. Moreover, in (42), we have $U\mathbb{E}_\zeta[n^b] = \left(\sum_{i=1}^{U} T^i(n^b) + \sum_{i=U+1}^{K} T^i(n^b)\right)$, and $\mathbb{E}_\zeta\left[T^i(n)\right] \geq \mathbb{E}_\zeta\left[T^i(n^b)\right]$. Finally (43) is achieved by applying Theorem 3 and Theorem 3 from [16].

Concerning the second part of (40), we can upper bound as shown in (44) at the top of next page, where the first and the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2017.2675901, IEEE Transactions on Cognitive Communications and Networking

$$\sum_{j=1}^{U}\sum_{i=1}^{K}\mu_i^R\mathbb{E}_\zeta\left[V^{i,j}(n)\right]-\sum_{j=1}^{U}\mathbb{E}_\zeta\left[\sum_{t=1}^{n^b}r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}\right]$$
$$\leq\left\{\sum_{j=1}^{U}\sum_{i=1}^{U}\mu_i^R\mathbb{E}_\zeta\left[V^{i,j}(n)\right]-\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{F^{i,j}(f(n))}\sum_{S(t)\in S^{i,j}(k)}\mathbb{1}_{(S(t)=q)}\right]\right\}+\left\{\sum_{j=1}^{U}\sum_{i=U+1}^{K}\mu_i^R\mathbb{E}_\zeta\left[V^{i,j}(n)\right]\right.$$
$$\left.-\sum_{j=1}^{U}\sum_{i=U+1}^{K}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{F^{i,j}(f(n))}\sum_{S(t)\in S_2^{i,j}(k)}\mathbb{1}_{(S(t)=q)}\right]\right\}+\left\{\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[C_o(n)\right]+\sum_{j=1}^{U}\mathbb{E}_\zeta\left[C_o(n)\right]\sum_{i=U+1}^{K}\sum_{q\in S^i}r_q^iG_q^i\right\} \tag{44}$$

second part of (44) are the rewards collected in optimal and suboptimal arms separately. The last part is the reward loss due to the collisions in optimal and suboptimal channels, where $\mathbb{E}_\zeta\left[C_o(n)\right]$ is the expected number of collisions in optimal channels.

Let us start with the first part of (44) and following the same reasoning than in [11], we have:

$$\mathbb{E}_\zeta[\bar{b}^{i,j}(n)]\leq\sum_{k=U+1}^{K}\mathbb{E}_\zeta[F^{k,j}(f(n))]. \tag{45}$$

where, $\{\bar{b}^{i,j}\}$ is the total number of the joined blocks, and is always less than or equal to the total number of discontinuities. Thus, each successive combined block $\bar{X}^{i,j}$ can be separated into two sub-blocks: i) $\bar{X}_1^{i,j}$ consisting in the states observed from the beginning of $\bar{X}^{i,j}$ (empty if the first state is $\zeta^{i,j}$) to the state right before observing $\zeta^{i,j}$, and ii) $\bar{X}_2^{i,j}$ consisting in the rest of $\bar{X}^{i,j}$.

Therefore, the first part of (44) can be upper bound as shown in (48) at the top of next page, where (46) comes from counting the rewards in two different sub-blocks SB1 and SB2. The inequality in (48) is obtained by observing that $\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}|\bar{X}_2^{i,j}(k)|\right]\leq\frac{1}{\pi_\zeta^i}\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]$ in SB2 and $\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}|\bar{X}_1^{i,j}(k)|\right]\leq\Omega_{\max}^i\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]$ in SB1. Since rewards are positive, the last part of (46) is larger than 0, and applying Lemma 2 from [11], [30] to the second part of (46).

The second part of (44) can be upper bound as shown:

$$\leq\sum_{j=1}^{U}\sum_{i=U+1}^{K}\mu_i^R\mathbb{E}_\zeta\left[T^{i,j}(n)\right]-\sum_{j=1}^{U}\sum_{i=U+1}^{K}\frac{\mu_i^R}{\pi_\zeta^i}\mathbb{E}_\zeta\left[F^{i,j}(f(n))\right]$$
$$=\sum_{j=1}^{U}\sum_{i=U+1}^{K}\mu_i^R\left(\Omega_{\max}^i+W\right)\mathbb{E}_\zeta\left[F^{i,j}(f(n))\right] \tag{49}$$

where (49) comes from $V^{i,j}(n)\leq T^{i,j}(n)$, and applying Lemma 2 from [11], [30]. (49) is obtained with Theorem 3.

Now we bound the last part of (44) in (50):

$$\leq\mathbb{E}_\zeta\left[C_o(n)\right]\left[\sum_{i=1}^{U}\frac{\mu_i^R}{\pi_{\min}^i}+K\sum_{i=U+1}^{K}\frac{\mu_i^R}{\pi_{\min}^i}\right]$$
$$\leq\left[K^2-KU+U\right]\frac{\mu_1^R}{\pi_{\min}}U(\mathbb{E}[\Upsilon(U,U)]+1)\,\mathbb{E}[T_w(n)] \tag{50}$$

where (50) follows from Theorem 3 and Theorem 3 from [16].

Combining (50), (49) and (48) into (44), we immediately upper bound (44) as (51):

$$\sum_{j=1}^{U}\sum_{i=1}^{K}\mu_i^R\mathbb{E}_\zeta\left[V^{i,j}(n)\right]-\mathbb{E}_\zeta[\sum_{j=1}^{U}\sum_{t=1}^{n^b}r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}]$$
$$\leq\sum_{j=1}^{U}\sum_{i=1}^{U}\mu_i^R\Omega_{\max}^i\sum_{K=U+1}^{K}\mathbb{E}_\zeta[F^{k,j}(f(n))]$$
$$+\sum_{j=1}^{U}\sum_{i=U+1}^{K}\mu_i^R\left(\Omega_{\max}^i+W\right)\mathbb{E}_\zeta\left[F^{i,j}(f(n))\right]$$
$$+\left[K^2-KU+U\right]\frac{\mu_1^R}{\pi_{\min}}U(\mathbb{E}[\Upsilon(U,U)]+1)\,\mathbb{E}[T_w(n)] \tag{51}$$

The last part of (40) is bounded as:

$$\mathbb{E}_\zeta[n-n^b]\sum_{j=1}^{U}\mu_j^R-\mathbb{E}_\zeta[\sum_{j=1}^{U}\sum_{t=n^b+1}^{n}r_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}G_{q_{\mathcal{A}(t)}}^{\mathcal{A}(t),j}]$$
$$\leq\sum_{j=1}^{U}\mu_j^R\left(\frac{1}{\pi_\zeta}+\Omega_{\max}+W\right) \tag{52}$$

Combining (43), (51), (52) and using Lemma 2 and Theorem 3, the upper bound on the regret $\Phi^M(n)$ of the multi-user distributed RQoS-UCB policy is obtained as:

$$\Phi^M(n)\leq X_3\ln n+X_4$$
$$X_3=X_1+X_5\quad\text{and}\quad X_4=X_2+X_6+X_8$$
$$X_1=\left(\frac{\left[K^2-KU+U\right]}{\pi_{\min}}+1\right)\mu_1^RU^2\Bigg(\mathbb{E}[\Upsilon(U,U)]+1\Bigg)\sum_{a=1}^{U}\sum_{b=1}^{K}\frac{4\alpha}{\Delta\mu_{a,b}^R}X_9$$
$$X_2=\left(\frac{\left[K^2-KU+U\right]}{\pi_{\min}}+1\right)\mu_1^RU^2\Bigg(\mathbb{E}[\Upsilon(U,U)]+1\Bigg)\sum_{a=1}^{U}\sum_{b=1}^{K}\left[\frac{|S^a|+|S^b|}{\pi_{\min}}\sum_{t=1}^{\infty}t^{-2}\right]X_9$$
$$X_5=\sum_{i=U+1}^{K}\sum_{k=1}^{U}\frac{4\alpha}{\left(\Delta\mu_{i,k}^R\right)^2}X_7$$

$$\sum_{j=1}^{U}\sum_{i=1}^{U}\mu_i^R\mathbb{E}_\zeta\left[V^{i,j}(n)\right]-\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{F^{i,j}(f(n))}\sum_{S_t^{i,j}\in S^i(k)}\mathbb{1}_{(S_t^{i,j}=q)}\right]$$

$$\leq\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\pi_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}|\bar{X}_2^{i,j}(k)|\right]-\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}\sum_{S(t)\in\bar{S}_2^{i,j}(k)}\mathbb{1}_{(S(t)=q)}\right]$$

$$+\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\pi_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}|\bar{X}_1^{i,j}(k)|\right]-\sum_{j=1}^{U}\sum_{i=1}^{U}\sum_{q\in S^i}r_q^iG_q^i\mathbb{E}_\zeta\left[\sum_{k=1}^{\bar{b}^{i,j}(n)}\sum_{S(t)\in\bar{S}_1^{i,j}(k)}\mathbb{1}_{(S(t)=q)}\right] \tag{46}$$

$$\leq\sum_{j=1}^{U}\sum_{i=1}^{U}\frac{\mu_i^R}{\pi_\zeta^i}\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]-\sum_{j=1}^{U}\sum_{i=1}^{U}\frac{\mu_i^R}{\pi_\zeta^i}\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]+\sum_{j=1}^{U}\sum_{i=1}^{U}\mu_i^R\Omega_{\max}^i\mathbb{E}_\zeta\left[\bar{b}^{i,j}(n)\right]-0 \tag{47}$$

$$<\sum_{j=1}^{U}\sum_{i=1}^{U}\mu_i^R\Omega_{\max}^i\sum_{K=U+1}^{K}\mathbb{E}_\zeta[F^{k,j}(f(n))] \tag{48}$$

$$X_6=\sum_{i=U+1}^{K}\sum_{k=1}^{U}\left[\frac{|S^k|+|S^i|}{\pi_{\min}}\sum_{t=1}^{\infty}t^{-2}\right]X_7$$

$$X_7=\left[\mu_i^R\left(W+\Omega_{\max}^i\right)+\sum_{j=1}^{U}\mu_j^R\Omega_{\max}^j+\mu_1^RX_9\right]$$

$$X_8=U\mu_1^R\left(\frac{1}{\pi_\zeta}+\Omega_{\max}+W\right)$$

$$X_9=\left(\frac{1}{\pi_\zeta}+\Omega_{\max}+W\right)$$

REFERENCES

[1] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.

[2] W. Jouini, D. Ernst, C. Moy, and J. Palicot, "Upper confidence bound based decision making strategies and dynamic spectrum access," in *International Conference on Communications, ICC'10*, May 2010.

[3] W. Jouini, C. Moy, and J. Palicot, "Decision making for cognitive radio equipment: analysis of the first 10 years of exploration," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 26, Jan. 2012.

[4] C. Tekin and M. Liu, "Online algorithms for the multi-armed bandit problem with markovian rewards," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*. IEEE, 2010, pp. 1675–1682.

[5] J. Oksanen, V. Koivunen, and H. V. Poor, "A sensing policy based on confidence bounds and a restless multi-armed bandit model," *CoRR*, vol. abs/1211.4384, 2012.

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

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

[8] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, May 2002. [Online]. Available: http://dx.doi.org/10.1023/A:1013689704352

[9] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part ii: Markovian rewards," *Automatic Control, IEEE Transactions on*, vol. 32, no. 11, pp. 977–982, Nov 1987.

[10] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2013.

[11] C. Tekin and M. Liu, "Online learning in opportunistic spectrum access: A restless bandit approach," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 2462–2470.

[12] K. Liu and Q. Zhao, "Cooperative game in dynamic spectrum access with unknown model and imperfect sensing," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1596–1604, April 2012.

[13] K. Liu, Q. Zhao, and B. Krishnamachari, "Decentralized multi-armed bandit with imperfect observations," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, Sept 2010, pp. 1669–1674.

[14] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multi-player multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, April 2014.

[15] K. Liu, Q. Zhao, and B. Krishnamachari, "Distributed learning under imperfect sensing in cognitive radio networks," in *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers*, Nov 2010, pp. 671–675.

[16] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731–745, April 2011.

[17] Y. Wang, Y. Xu, L. Shen, C. Xu, and Y. Cheng, "Two-dimensional pomdp-based opportunistic spectrum access in time-varying environment with fading channels," *Journal of Communications and Networks*, vol. 16, no. 2, pp. 217–226, April 2014.

[18] N. Modi, P. Mary, and C. Moy, "QoS driven channel selection algorithm for opportunistic spectrum access," in *IEEE Globecom 2015 Workshop on Advances in Software Defined Radio Access Networks and Context-aware Cognitive Networks (IEEE SDRANCAN 2015)*, San Diego, USA, Dec. 2015.

[19] L. Melián-Gutiérrez, N. Modi, C. Moy, I. Pérez-lvarez, F. Bader, and S. Zazo, "Upper confidence bound learning approach for real HF measurements," in *IEEE ICC 2015 - Workshop on Advances in Software Defined and Context Aware Cognitive Networks 2015 (IEEE SCAN-2015) (ICC'15 - Workshops 10)*, London, United Kingdom, Jun. 2015, pp. 387–392.

[20] W. Jouini, C. Moy, J. Palicot *et al.*, "Upper confidence bound algorithm for opportunistic spectrum access with sensing errors," *6th International ICST Conference on Cognitive Radio Oriented Wireless Networks and Communications, Osaka, Japan*, p. 17, 2011.

[21] Y.-C. Liang, Y. Zeng, E. C. Peh, and A. T. Hoang, "Sensing-throughput tradeoff for cognitive radio networks," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 4, pp. 1326–1337, 2008.

[22] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Communications Surveys Tutorials*, vol. 11, no. 1, pp. 116–130, First 2009.

[23] M. Bkassiny, S. K. Jayaweera, Y. Li, and K. A. Avery, "Wideband spectrum sensing and non-parametric signal classification for autonomous self-learning cognitive radios," *IEEE Transactions on Wireless Communications*, vol. 11, no. 7, pp. 2596–2605, July 2012.

[24] K. Wanuga, N. Gulati, H. Saarnisaari, and K. R. Dandekar, "Online learning for spectrum sensing and reconfigurable antenna control," in *2014 9th International Conference on Cognitive Radio Oriented Wireless*

Networks and Communications (CROWNCOM), June 2014, pp. 508–513.

[25] Y. Gai, B. Krishnamachari, and M. Liu, "Online learning for combinatorial network optimization with restless markovian rewards," in *Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2012 9th Annual IEEE Communications Society Conference on*, June 2012, pp. 28–36.

[26] M. Bóna, *A Walk Through Combinatorics: An Introduction to Enumeration and Graph Theory*, 2nd ed. World Scientific Publishing Company, Oct. 2006.

[27] X. Chen, Z. Zhao, and H. Zhang, "Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks," *Mobile Computing, IEEE Transactions on*, vol. 12, no. 11, pp. 2155–2166, Nov 2013.

[28] S. Koenig and R. G. Simmons, "Complexity analysis of real-time reinforcement learning," in *Proceedings of the 11th National Conference on Artificial Intelligence. Washington, DC, USA, July 11-15, 1993.*, 1993, pp. 99–107.

[29] P. Lezaud, "Chernoff-type bound for finite markov chains," in *Annals of Applied Probability, Vol. 8 (1998), no. 3, pp. 849–867*, 1998.

[30] P. Bremaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues (Texts in Applied Mathematics)*, corrected ed. Springer, Feb 2008.

Navikkumar MODI received in 2010 his bachelor degree of Technology (B.Tech.) in Electrical, Electronics and Communications Engineering from U.V.Patel College of Engineering, India. He then received in 2013 the Master of Science (M.S.) in Communications and Signal Processing from TU Ilmenau. He is now a PhD student with CentraleSupelec (created in January 2015), formerly Supelec, since March 2014. He is studying machine learning algorithms for green cognitive radio systems in TEPN CominLabs Labex project. He was before Student Research Assistant at Digital Broadcasting Research Laboratory, TU Ilmenau.

Philippe MARY (S'06-M'09) received his M.Sc. in Signal Processing and Digital Communications from the University of Nice Sophia-Antipolis and the Dipl. Ing. degree in electrical engineering from the Polytechnic University School of Nice Sophia-Antipolis (France) both in 2004. He received his PhD in electrical engineering from the National Institute of Applied Sciences of Lyon in 2008. During his PhD, he was with France Telecom R&D and he worked on the analytical performance study for mobile communications considering shadowing and fading and multi-user detectors for wireless communications. From 2008 to 2009, he was post-doctoral researcher at ETIS Laboratory in Cergy-Pontoise.

In September 2009, Dr. Mary joined the Digital Communication Systems department of INSA Rennes and IETR laboratory as associate professor. In 2011, he served as TPC chair for the third international workshop on cross-layer design (IWCLD 2011) and he is serving as TPC member of various IEEE conferences, e.g. ICC, Globecom, PIMRC, WCNC, VTC. His research interests include analytical performance analysis and signal processing for digital communications, resource allocation and finite block length communication theory.

Christophe MOY (M'06) received the M.Sc. degree in engineering, and the Ph.D. degree in electronics from the INSA (National Institute of Applied Sciences), Rennes, France, in 1995 and 1999, respectively. He then worked for six years in Mitsubishi Electric ITE focused on software radio systems and concepts. Since 2005, he has been a Professor with Supelec, now CentraleSupelec since January 2015, in the IETR Laboratory of CNRS (UMR 6164): Institute of Electronics and Telecommunications of Rennes. His research interests include software radio, cognitive radio, and green radio. He addresses heterogeneous design techniques for SDR, high-level design methodologies for cognitive management, decentralized decision making and learning for cognitive radio equipment and systems, and dynamic spectrum access. He applies cognitive radio strategies for the Internet of Things, Smart Grids, HetNets, etc. He is working in the following collaborative projects: French ANR SoGreen and EPHYL, CominLabs labex TEPN. He was previously involved in several European collaborative projects: Networks of Excellence NEWCOM and NEWCOM++, EULER, and E2R-phase2, E2R. He participated also in the following French collaborative projects (ANR and Images et Reseaux cluster): WiNoCoD, Mopcom, Idromel, A3S and SoftRF. He is currently the Head of the Signal and Communication Department of IETR, and a Member of B-COM Institute of Research and Technology.