This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2011 proceedings.

Learning with the Bandit: A Cooperative Spectrum Selection Scheme for Cognitive Radio Networks

Marco Di Felice*, Kaushik Roy Chowdhury†, Luciano Bononi*
* Department of Computer Science, University of Bologna, Italy
Email: {difelice,bononi}@cs.unibo.it
† Department of Electrical and Computer Engineering, Northeastern University, Boston, USA
Email:krc@ece.neu.edu

***Abstract*—Distributed spectrum allocation in Cognitive Radio (CR) systems requires each Secondary User (SU) to learn the optimal spectrum policy which maximizes the network performance while minimizing the impact to the Primary Users (PUs). To this aim, each SU must rely on local sensing information which however can be biased by interference and fading effects on the received signal. Thus, if each SU works in isolation, the convergence to the system-wide optimal policy can not be guaranteed. In this paper, we formulate the spectrum allocation problem as a cooperative learning task in which each SU can learn the spectrum availability of each channel and share such knowledge with the other SUs. We propose a correlation model through which different SUs can leverage the experience of other nodes, and we integrate it into a distributed channel allocation scheme. At the same time, we investigate mechanisms to bound the cooperation overhead based on the performance of the distributed learning process. Simulation results confirm the ability of the cooperative learning scheme in providing higher sensing accuracy and convergence time when compared with non-cooperative spectrum allocation schemes for CR networks.**

I. Introduction

Recent advances in Software Defined Radio (SDR) platforms have opened the door to a new generation of wireless devices where most of the functionalities are implemented by software and thus reconfigurable on-the-fly. Among such functionalities, spectrum reconfigurability has attracted a considerable interest from academic and industrial research due to the possibility to solve the spectrum scarcity problem through Opportunistic Spectrum Access (OSA) mechanisms. Cognitive Radio (CR) networks constitute the enabling technology to achieve OSA systems [1]. The architecture of a CR network consists of two main components: a primary system, composed of Primary Users (PU) which are the licensed owner of the spectrum, and a secondary system, composed by Secondary Users (SUs) which can opportunistically access the licensed spectrum but must not cause any interference to the PUs.

Due to this constraint, spectrum allocation for SUs is a non-trivial problem involving at the same time *exploration*, *uncertainty* and *competition* issues. First, each SU must identify a PU-free channel where to transmit at each time, which means it must *explore* the licensed spectrum band in order to acquire knowledge of the availability of each channel. Sensing techniques are used at the physical layer to detect the presence of PUs on each channel [2]. However, the information provided by the sensing process can be biased by physical propagation effects (e.g. fading, shadowing), thus adding some ”*uncertainty*” to the selection process. Second, each SU must *compete* with other SUs in order to share the available spectrum resources.

The highly dynamic nature of spectrum conditions and the absence of a central controller of the secondary system make the spectrum allocation problem in CR networks quite difficult to be solved using traditional optimization techniques [1]. For this reason, there is an increasing interest in distributed Machine Learning (ML) techniques [3] which allow each SU to learn the optimal spectrum allocation policy that maximizes a system utility function, i.e. the cumulative throughput. Some recent works investigate the parallelism between the multi-armed bandit problem in ML theory and the spectrum allocation problem in CR networks [4] [5] [6] [9]. In some cases, the authors provide some interesting bounds on the performance of their learning-based selection algorithms [5] [6]. However, if each SU works in isolation, the system might converge to suboptimal solutions due the fact that each agent has a local and possibly biased vision of the environment.

In this paper, we look at the spectrum allocation problem from a different perspective. We formulate the spectrum allocation problem as a *collaborative* learning process in which the SUs can share information in order to enlarge the training set available to the learning algorithm. We consider a realistic formulation of the CR environment by taking into account the issues of: PU interference, SU interference and sensing accuracy. We integrate the cooperative learning scheme with correlation models which allows a SU to take benefit of the learning experience of another SU, by considering the locality of each vision of the environment. By introducing the cooperation factor into the learning process, we attempt to answer these questions: Can the cooperation process improve the accuracy of learning in presence of incorrect sensing? Can the cooperation process reduce the amount of exploration required from each SU? At the same time, we know that the cooperation process might introduce a significant cost in terms of network overhead required for its implementation. For this purpose, another fundamental question we address in this paper is: How can we bound the cooperation overhead without affecting the performance of the learning scheme? At the best of our knowledge, this is one of the few papers investigating all together the issues of cooperation, learning and competition

978-1-4244-9268-8/11/$26.00 ©2011 IEEE

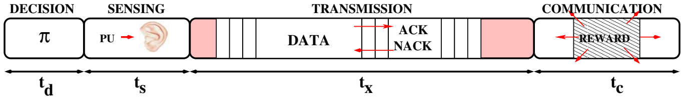


Fig. 1. The slot structure.

in the CR spectrum allocation problem.

The paper is organized as follows. In Section II we review existing proposals to solve the spectrum allocation problem in CR networks based on ML techniques. Section III discusses the system model and the problem formulation. The collaborative spectrum learning scheme is described in Section IV. Simulation results are provided in Section V. Conclusions and Future works follow in Section VI.

II. Related Works

Machine Learning (ML) techniques constitute a promising although not well investigated approach to implement adaptive decision making strategies in CR networks. In [3] we have discussed the application of reinforcement learning algorithms for protocol design in CR systems, and we have shown how the spectrum allocation problem can be formulated through a distributed learning problem [9]. Similarly to the approach followed in this paper, there are some recent works [4] [5] [6] which investigate the parallelism between the multi-armed bandit problem and the spectrum allocation problem in CR networks. Some fundamental results on the multi-armed bandit problems are provided in [7] [8]. In [4], the authors propose to use Upper Confidence Bound (UCB) policies to allow a cognitive node to decide the spectrum to use at each slot, for a network scenario composed by a single CR node. The authors of [5] [6] extend the analysis by also considering the impact of interference from SUs. To this aim, they propose distributed channel allocation strategy for SUs which are demonstrated to minimize the cumulative regret over time. However, both [5] [6] consider a special case of the allocation problem when the number of SUs is lower than the available channels, and thus interference-free allocation can be guaranteed. Compared to the previous works, we provide these novel contributions: (*i*) we model the case of incorrect sensing caused by spatial-correlated shadowing, (*ii*) we consider a realistic network model with an arbitrary number of PUs and SUs and (*iii*) we propose collaborative learning algorithms to improve the accuracy and convergence of the spectrum selection process, and we empirically evaluate their performance.

III. Network Model and Architecture

We consider a network model composed of N Primary Users (PU), which transmits on N different channels of the licensed band, and K secondary users (SU). Each SU is equipped with a spectrum-agile transceiver, through which it can sense and access the licensed band in an opportunistic way. Without loss of generality, we assume the time is divided into discrete slots t=0,1,2,.... At each slot, spectrum j, with $0 \leq j < N$ can be occupied by PU_j with probability θ_j. However, the values of $\theta_{0 \leq j < N}$ are unknown to the SUs, which must learn them in order to minimize the interference caused to the PUs. To this aim, we assume that each SU, e.g. SU_i divides the time slot in four periods, as shown in Figure 1. In the following, we describe the operations performed by SU_i in each period.

1) *Decision Period.* The SU_i decides the channel to use for the next transmission, i.e. channel j and communicates such information to the SU receiver on the Common Control Channel (CCC). Then, both SU sender and receiver tune their radio agile interface to channel j. The policy function: π^i decides the channel to be used by SU_i at each slot.

2) *Sensing Period.* The SU_i senses channel j for an interval t_s to detect the presence of PU_j on the channel. As a result, the SU_i must choose between the two hypothesis: H_0 (i.e. channel j is currently free) and H_1 (i.e. channel j is currently occupied by a PU). Conversely to most of the cited papers, we do not assume perfect sensing and we model the accuracy of sensing through the probability of correct detection (i.e. P_d) and the probability of false positive (i.e. P_f) [2]. Based on such values, the SU_i will decide for H_1 with probability $P^{i,j}(H_1)$ equal to:

$$P^{i,j}(H_1) = \theta_k \cdot P_d^{i,j} + (1 - \theta_k) \cdot P_f^{i,j} \qquad (1)$$

The exact value of $P_d^{i,j}$ and $P_f^{i,j}$ depends on the locations of SU_i and PU_j and on the channel propagation characteristics in between. In this paper, we model the impact of the shadowing fading on the received signal through the well-known log-normal path loss model, also considering the spatial correlation among sensing observations at different locations. To this aim, we use the correlation function proposed in [10]:

$$Corr(S_i, S_k) = e^{-\frac{d}{d_{corr}} \cdot ln2} \qquad (2)$$

where d is the distance between S_i and S_k and d_{corr} is the uncorrelation distance (e.g. 20m in an urban environment).

3) *Transmission Period.* If channel j is found free, than the SU_i accesses the channel by using the CSMA/CA scheme and transmits the data to the SU receiver. In case of successful transmission, it receives an ACK message from the receiver. Otherwise, in case of PU interference, it receives a NACK message. We assume that a SU is able to distinguish between interference caused by PUs or by other SUs. This is reasonable if cyclo-stationary or filter-matching techniques are implemented at the physical layer [2]. At the end of the transmission period, the SU_i computes a numerical reward $x_t^{i,j}$ which reflects the outcome of the communication process for node SU_i and channel j.

4) *Communication Period.* After its transmission period, SU_i broadcasts the reward $x_t^{i,j}$ on the CCC, and updates its current policy π^i based on the local reward and on the rewards received from other SUs.

For our purpose, we are interested in allocating the K users over the N channels in a distributed way in order to: maximize the expected number of successful transmissions in the network while minimizing the interference to the PUs. This is equivalent in determining the policies which minimize the cumulated expected regret over time. Let $E[X_t^{\pi}]$ the expected total reward after t slots produced by the current policy π.

Analogously, let $E[X_t^{\pi^*}]$ the expected total reward produced by the optimal policy allocation π^*. Then, the regret R_t is defined as the loss in performance due to selection of suboptimal actions in t slots:

$$R_t = E[X_t^{\pi^*}] - E[X_t^{\pi}] \tag{3}$$

IV. Machine Learning based Spectrum Selection

We formulate the problem of spectrum selection as a special instance of the distributed multi-armed bandit problem [7] [8], where we have a set of K learning agents (i.e. the SUs in the network). At each step, each agent must choose one of the N available channels, e.g. channel j. After its transmission period, each agent SU_i receives a binary reward $x_t^{i,j}$ which is defined as follows:

$$x_t^{i,j} = \begin{cases} 0 & \text{if PU-interfered slot} \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

A slot is considered *PU-interfered* for SU_i if: (*i*) it finds channel j busy during the sensing period OR (*ii*) it receives a NACK message from the SU receiver. Otherwise, SU_i receives a positive reward. Note that in this formulation we do not take into account the interference from other SUs in the reward process. Instead, we propose to tackle separately the PU- and SU- interference problem through a 2-step approach:

Step 1) First, we make each SU_i learn the values of θ_j, for each channel $0 \leq j < N$. To this aim, we allow nodes to cooperate and share the rewards received at each slot in order to: (*i*) improve the accuracy of learning by reducing the impact of imperfect sensing on local observation and (*ii*) speed up the convergence of the learning process by reducing the amount of exploration required from each node.

Step 2) Then, we design distributed randomized policies π^i based on the spectrum availability of each channel j (i.e. θ_j). The randomization process allows to reduce the impact of self-interference caused by the secondary system.

Algorithm 1 shows the main steps of the cooperative learning process. We assume that each SU_i records the rewards received for each channel j in a vector: $X^{i,j}$. After a transmission slot on channel j, SU_i receives a reward $x_t^{i,j}$, adds such value to $X^{i,j}$ and assigns to it a numerical weight $0 \leq w_t^{i,j} \leq 1$, which quantifies the relevance of the received reward in the learning process of SU_i. The weight assignment scheme is described in Section IV-A. During the communication period, SU_i can decide to share its own experience with the other SUs. This is done by broadcasting a REWARD message on the CCC, including the values of j and $x_t^{i,j}$. To reduce the network overhead, each SU_i broadcasts its REWARD message with probability p_{REW}^i, which is computed by each SU_i through the scheme described in Section IV-B. When receiving a *REWARD* message from node SU_k with the reward value $x_t^{k,m}$ for channel m, SU_i computes the weight $x_t^{i,m}$ associated to the reward, and updates its $X^{i,m}$ and $W^{i,m}$ vectors accordingly. Through the $X^{i,j}$ and $W^{i,j}$ vectors, SU_i can estimate the Upper Confidence Bounds (UCB) on the spectrum availability of each channel j by using the scheme in Section IV-C. Then, it can update the spectrum selection policy π^i to choose the channel to be used for the next slot. The channel selection scheme should take into account the spectrum availability of each channel (provided by the UCB value) and the interference from other SUs. To this aim, in Section IV-D we propose different formulations of π^i, based on deterministic or probabilistic channel selection approaches.

Algorithm 1 Cooperative Learning Algorithm for SU_i

```
Decision Period:
Decide channel $j$ based on $\pi^i$ through Equation 15 or 16

Sensing Period:
Sense channel $j$ and decides between $H_0$ and $H_1$

Transmission Period:
if decision = $H_0$ then
  Transmit DATA and waits for ACK/NACK
  if NACK received then
    Set reward $x_t^{i,j} = 0$
  else
    Set reward $x_t^{i,j} = 1$
  end if
else
  Set reward $x_t^{i,j} = 0$
end if
Compute $w_t^{i,j}$ through Equations 5-7
Update $X^{i,j} = X^{i,j} \bigcup x_t^{i,j}$ and $X_L^{i,j} = X_L^{i,j} \bigcup x_t^{i,j}$
Update $W^{i,j} = W^{i,j} \bigcup w_t^{i,j}$

Communication Period:
Broadcast REWARD=$\{j, E(i,j)\}$ on the CCC
for each received REWARD=$\{m, E(k,m)\}$ from $SU_k$ do
  Compute $w_t^{i,m}$ through Equations 5-8
  Update $X^{i,m} = X^{i,m} \bigcup x_t^{i,m}$
  Update $W^{i,m} = W^{i,m} \bigcup w_t^{i,m}$
  Update UCB index $B^{i,m}$ through Equations 10-13
end for
```

A. Weight Computation

Each time SU_i receives a reward $x_t^{k,m}$ from agent SU_k for channel m, it adds it to $X^{i,m}$, and computes the weight $w_t^{i,m}$ as follows:

$$w_t^{i,m} = E(k,m) \cdot C(i,k) \tag{5}$$

Here, the first factor i.e. $0 \leq E(k,m) \leq 1$ is a function which measures the actual knowledge of spectrum conditions of channel m from node SU_k. For this reason, we call it the *expertise* value and it is included into the REWARD message broadcasted by SU_k. The second factor, i.e. $0 \leq C(i,k) \leq 1$ is a function which defines the degree of similarity between the vision of SU_i and SU_k. Basically, it quantifies how much expertise of SU_k can be transferred to SU_i, considering the spatial correlation between the two nodes. In case of local rewards, i.e. rewards received after the transmission period of node SU_i, we set $C(i,i) = 1$, i.e. the weight $w_t^{i,j}$ is a function of the expertise function of node SU_i only.

In our approach, we bound the expertise value e.g. $E(k,m)$ to the variance of the received experienced by SU_k after each transmission on channel m, based on the intuitive idea that high variances in the received rewards can be an indicator of a faded channel. To this aim, we consider $X_L^{k,m}$ as the subset of $X^{k,m}$ corresponding to local rewards (i.e. not received from other nodes) and we compute the confidence interval of $X_L^{k,m}$:

$$CI^{k,m} = t_{n-1,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{S^2}{q}} \tag{6}$$

where S^2 is the sample variance of $X_L^{k,m}$, q is the number of samples in $X_L^{k,m}$ and $t_{n-1,1-\frac{\alpha}{2}}$ is the value of the t-student for a given value of accuracy (95% in our experiments). The higher is the variance in the received rewards, the lower should be the expertise of a node. Thus:

$$E(k,m) = min\{1 - CI^{k,m}, 1\} \tag{7}$$

Analogously, $0 \leq C(i,k) \leq 1$ is a function which measures the spatial *correlation* between SU_i and SU_k. Since close SUs might experience correlated shadowing effects due to the presence of obstacles between the SU and the PU transmitter, we assign higher weights to samples received from far nodes, based on the function of correlation defined by Equation 2:

$$C(i,k) = 1 - \frac{Corr(SU_i, SU_k)}{max(Corr(SU_i, SU_h))} \forall h \in \{0, ..., K-1\} \tag{8}$$

B. Reward Broadcasting Scheme

At each cooperation period, SU_i broadcasts a `REWARD` message with this format: $< x_t^{i,j}, E(i,j) >$ where $x_t^{i,j}$ is the last reward received on channel j and $E(i,j)$ is the expertise of SU_i on channel j. It is easy to see that the network overhead in terms of messages sent after t slots is equal to $t \cdot K$. A natural way to reduce such overhead without affecting the performance of the collaborative learning scheme is to provide a metric to quantify the relevance of information contained in each `REWARD` message. If a message is not considered relevant for the scope of the distributed learning process, then SU_i can avoid to transmit it. In our case, the relevance of each `REWARD` message (i.e. p_{REW}^i) can be expressed by comparing the expertise of S_i on channel j, i.e. $E(i,j)$ with the average expertise of its neighbors, i.e. $\overline{E(j)}$ as follows:

$$p_{REW}^i = min\{1, \frac{E(i,j)}{\overline{E(j)}}\} \tag{9}$$

SU_i computes the relevance of each `REWARD` and decides to broadcast it with probability p_{REW}^i during the communication period. It is easy to see that if a SU has lower expertise on a given channel than its neighbors than it will be discouraged from broadcasting its reward, thus reducing the overhead involved by the cooperation process. We discuss the performance of such approach in Section V-C.

C. Upper Confidence Bound (UCB) indexes

In classical multi-armed bandit problem, the Upper Confidence Bound (UCB) indexes are used to represent the average reward associated to each arm [4] [8]. Here, we use the UCB indexes to represent the spatial availability of each channel j for each SU_i (i.e $B^{i,j}$). Using the general definition in [4] [8], $B^{i,j}$ values are defined as follows:

$$B^{i,j} = \overline{T^{i,j}} + A^{i,j} \tag{10}$$

where $\overline{T^{i,j}}$ is the mean of the $x_t^{i,j}$ values and $A^{i,j}$ is an upper bias added to the mean. As a main difference with the formulation of the UCB policies in [4], here we take into account the weight $w_t^{i,j}$ associated to each $x_t^{i,j}$. Let $n = |X^{i,j}|$. Thus, we define $\overline{W^{i,j}}$ as follows:

$$\overline{W^{i,j}} = \sum_{t=0}^{n} w_t^{i,j} \tag{11}$$

Analogously, we denote with $\overline{T^{i,j}}$ the average mean availability of channel j for SU_i:

$$\overline{T^{i,j}} = \frac{\sum_{t=0}^{n} x_t^{i,j} \cdot w_t^{i,j}}{\overline{W^{i,j}}} \tag{12}$$

The upper confidence bias is defined as:

$$A^{i,j} = \sqrt{\frac{\alpha \cdot ln(t)}{\overline{W^{i,j}}}} \tag{13}$$

D. Channel Selection

In this section we describe how the channel selection policy is decided for each node SU_i (i.e. π^i). We consider two types of policies: (*i*) a *deterministic* policy, through which the SU_i attempt to minimize the impact on PUs (but without maximizing their own performance) and (*ii*) a *randomized* policy, through which the SU_i attempt to maximize the probability of successful transmission while considering the spectrum availability of each channel.

Deterministic Policy. In this case, we do not consider the impact of interference from other SUs operating in the same channel. We analyze first this (ideal) scenario to evaluate the performance of the learning process in terms of cumulative regret over time. In fact, it is easy to see that under the assumption of SU-free interference the best policy for each SU consists in selecting the channel O that provides the highest availability, i.e. $O = argmax_j \theta_j$. Thus:

$$E[X^{\pi^*}] = K \cdot t \cdot \theta_O \tag{14}$$

In order to converge to the best policy, each SU_i must select the channel providing the highest UCB index, i.e.:

$$\pi^i = argmax_j B^{i,j} \;\; 0 \leq j < N \tag{15}$$

Randomized Policy. In this case, we take into account the interferences caused by SUs operating in the same channel. We highlight that the regret function can not be formulated for the general case, because we need a close form of $E[X^{\pi^*}]$. Here, we propose an heuristic solution which randomizes the

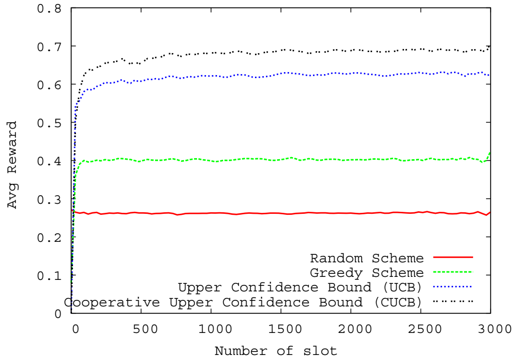


Fig. 2. The average reward over time is shown for K=10.

channel selection based on the spectrum availability of each channel j, which is reflected by the value of $B_t^{i,j}$. More specifically, each SU_i defines a probabilistic distribution π^i over the set of available channels N, where the probability to select channel j is given by:

$$\pi^i(j) = \frac{e^{B^{i,j}}}{\sum_{m=0}^{N} e^{B^{i,m}}} \tag{16}$$

V. Performance Evaluation

In this section, we evaluate the performance of the cooperative learning scheme described in Section IV. We consider a CR network composed by 10 PUs (i.e. N=10) and a variable number of SUs, randomly distributed in an area of 1000x1000 m^2. We consider the following PU probability of transmission on each slot: $\theta = \{1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1\}$. Moreover, we model the presence of obstacles in between the PUs and the SUs and the effect of correlated shadowing as in [10]. As a result, each SU_i experiences different values of $P_d^{i,j}$ and $P_f^{i,j}$, on each channel j. Each result shown in this analysis is computed as the average of 100 runs. In the following, we evaluate the performance of the cooperative learning scheme under three different analysis:

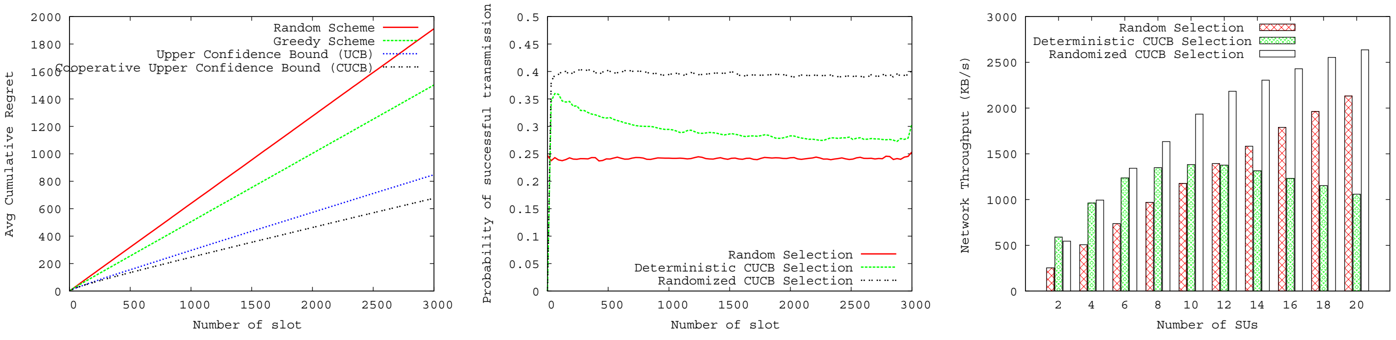
- *Regret Analysis*. We analyze the ability of the cooperative learning scheme in discovering the best policy for each SU. To this aim, we analyze the regret and the convergence time of the cooperative learning scheme when the deterministic policy (Equation 15) is used and the interference from SUs is temporary neglected.
- *Throughput Analysis*. We evaluate the performance of the randomized selection policy (Equation 16) when both the interference from SUs and from PUs are modeled.
- *Overhead Analysis*. We evaluate the overhead introduced by our cooperation scheme, and the tradeoff between cooperation overhead and accuracy of learning through the optimization scheme described in Section IV-B.

A. Regret Analysis

In Figure 2 we show the average reward over time (expressed in slot numbers) received by each SU after the transmission period. We consider a configuration with 10 SUs. We compare four different spectrum selection schemes: (*i*) a Random scheme, which chooses a channel randomly among the N available at each slot (*ii*) a ϵ-Greedy scheme, which chooses the channel providing the highest mean reward with probability 1-ϵ and performs random exploration with probability ϵ (0.2 in our experiments), (*iii*) an Upper Confidence Bound (UCB) scheme which implements the classical UCB indexes proposed in [4], and (*iv*) a Cooperative Upper Confidence Bound (CUCB) scheme which is our proposal described in Section IV. Both the UCB and CUCB schemes use the deterministic spectrum policy given by Equation 15. As expected, the Random policy produces the worst performance due to the fact that channel selection does not take into account the spectrum availability. The UCB scheme enhances both the Random and the Greedy schemes, but it still suffers of suboptimal selection due to the fact that individual sensing might be biased by sensing errors. Instead, the CUCB scheme allows to mitigate the impact of sensing errors through the weight system described in Section IV-A. As a result, the CUCB scheme guarantees the highest probability to select PU-free channels, and thus minimizes the impact on PUs transmissions. In Figure 3(a) we show the cumulative regret over time for the four evaluated schemes. In our configuration, the optimal policy π^* consists in selecting channel 9 with θ_9=0.9. Figure 3(a) confirms the ability of the CUCB scheme in reducing the impact of suboptimal action selection from the SUs.

B. Throughput Analysis

In Figure 3(b) we show the average probability of successful transmission per slot over time (on the x-axis), for the case K=10. A transmission from SU_i on channel j is considered successful if: (*i*) it does not experience interference from PU_j or from other SU operating on the same channel AND (*ii*) SU_i receives an ACK message. We assume that the SUs implement a CSMA/CA protocol at MAC Layer, and that the transmission period is slotted. Thus, multiple SUs might share the same channel during the transmission period. We consider three spectrum selection schemes: (*i*) a pure Random selection scheme, (*ii*) our CUCB scheme with deterministic policy (Equation 15) and (*ii*) our CUCB scheme with randomized policy (Equation 16). As before, the pure Random selection scheme provides the lowest performance due to the interference from PUs. The CUCB scheme with deterministic policy guarantees much better PUs detection but suffers of interference from other SUs, due to the fact the all the SU_i will select channel 9 for their operations. The CUCB scheme with randomized policy takes into account both PUs and SUs interference, and thus provides the highest probability of successful transmission. Figure 3(c) shows the network throughput of the three protocols for different values of K (on the x-axis). From Figure 3(c) we can see that the CUCB scheme with deterministic policy enhances the pure Random scheme for values of $K < 10$, while it is the opposite for $K > 10$ due to the MAC collisions caused by SU interference. Instead, the CUCB scheme with randomized policy combines



(a)

(b)

(c)

Fig. 3. The cumulative regret is shown in Figure 3(a). The network throughput over time and as a function of K is shown in Figures 3(b) and 3(c).

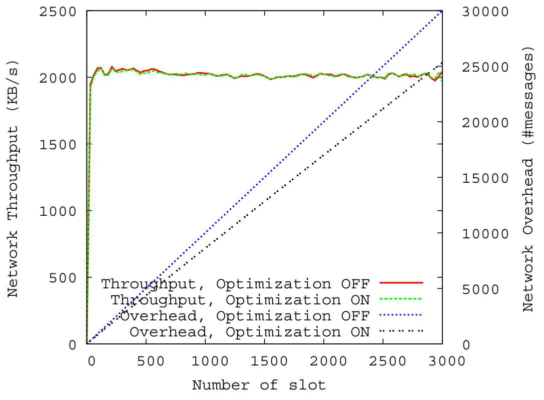


Fig. 4. The network overhead (x-$y2$) is shown for $K = 10$.

the advantages of the previous two approaches, and thus it provides the highest throughput for all the values of K.

C. Overhead Analysis

In Figure 4 we show the network throughput over time (on the x-y axis) for two configurations of the CUCB scheme: (*i*) a configuration which implements the optimization scheme for overhead reduction described in Section IV-B (Optimization ON) and (*ii*) a configuration in which each SU_i broadcasts its REWARD message during the communication process (Optimization OFF). In the same Figure, we show the communication overhead computed as cumulative number of REWARD messages sent in the network (on the x-$y2$ axis). We consider a network scenario with K=10. We place the SUs so that 3 of them experience high shadowing on the received PU signal while the remaining 7 SUs experience error-free sensing. Since the high faded SUs will experience high variance in the received rewards when compared with their neighbors, they will suppress the transmissions of the REWARD messages when the optimization scheme is used. This is confirmed by Figure 4, which shows the overhead reduction provided by the optimization scheme. Moreover, Figure 4 shows that such reduction does not come at expense of the performance of the learning scheme, since the curves of the network throughput for the two configurations are almost overlapped.

VI. Conclusion and Future Works

In this paper we have addressed the problem of distributed spectrum selection in CR networks in presence of imperfect sensing from the SUs. We have formulated the problem as a cooperative learning process, through which each SU can learn the spectrum occupancy of each channel and share such knowledge with the other SUs. Then, we have integrated the spectrum information in a distributed spectrum selection policy which takes into account the interference from PUs and from other SUs. Future works will include: analytical estimation of the convergence time of the cooperative learning scheme and its implementation on a SDR platform.

VII. Acknowledgment

The authors wish to thank prof. Cristophe Moy and Dr. Wassim Jouini for their support and useful comments.

References

[1] I. F. Akyildiz, W. Y. Lee, M. C. Vuran, and S. Mohanty. NeXt Generation/Dynamic Spectrum Access/Cognitive Radio Wireless Networks: A Survey. *Computer Networks Journal*, 50(1), pp. 2127-2159, 2006.

[2] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan. Cooperative Spectrum Sensing in Cognitive Radio Networks: A Survey. *Physical Communication Journal*, 4(1), pp. 40-62, 2011.

[3] M. Di Felice, K. Chowdhury, C. Wu, W. Meleis and L. Bononi. Learning-based Spectrum Selection in Cognitive Radio Ad Hoc Networks. *Proc. of IEEE WWIC'10*, Lulea, Sweden, pp. 133-145, 2010.

[4] W. Jouini, D. Ernst, C. Moy and J. Palicot. Upper Confidence Bound Based Decision Making Strategies and Dynamic Spectrum Access. *Proc. of IEEE ICC'10*, Cape Town, South Africa, pp. 1-5, 2010.

[5] K. Liu, Q. Zhao and B. Krishnamachari. Distributed Learning Under Imperfect Sensing in Cognitive Radio Networks. *Proc. of IEEE ACSSC'10*, Pacific Grove, USA, 2010.

[6] A. Anandukumar, N. Michael and A. Tang. Opportunistic Spectrum Access with Multiple Users: Learning Under Competition. *Proc. of ACM INFOCOM'10*, San Diego, USA, pp. 1-9, 2010.

[7] T. Lai and H. Robbins. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1), pp. 4-22, 1985.

[8] P. Auer, N. Cesa-Bianchi and P. Fischer. Finite Time Analysis of Multi-Armed Bandit Problems. *Machine Learning*, 47(2), pp. 235-256, 2002.

[9] K. L. A. Yau, P. Komisarczuk and P. D. Teal. A Context-aware and Intelligent Dynamic Channel Selection Scheme for Cognitive Radio Networks. *Proc. of CROWNCOM'09*, Hannover, 2009.

[10] M. Gudmundson. Correlation Model for Shadow Fading in Mobile Radio Systems. IEEE Electronics Letters, 27(30), pp. 2145-2146, 1991.