Multi-Armed Bandits with Dependent Arms for Cooperative Spectrum Sharing

Mario López-Martínez*, Juan J. Alcaraz*, Leonardo Badia† and Michele Zorzi†
*Technical University of Cartagena. Dept. of Information and Communications Technologies.
Email: {mario.lopez, juan.alcaraz}@upct.es
†University of Padova. Dept. of Information Engineering.
Email: {badia, zorzi}@dei.unipd.it

***Abstract*—Cooperative Spectrum Sharing (CSS) is an appealing approach for primary users (PUs) to share spectrum with secondary users (SUs) because it increases the transmission range or rate of the PUs. Most previous works are focused on developing complex algorithms which may not be fast enough for real-time variations such as in channel availability. Instead, we develop a learning mechanism for a PU to enable CSS in a strongly incomplete information scenario with low computational overhead. We model the learning mechanism of the PU to discover which SU to interact with and what offer to make to it with a combination of a Multi-Armed Bandit (MAB) and a Markov Decision Process (MDP). By means of Monte-Carlo simulations we show that, despite its low computational overhead, our proposed mechanism converges to the optimal solution and significantly outperforms the ϵ-greedy heuristic. This algorithm can be extended to include more sophisticated features while maintaining its desirable properties such as the fast speed of convergence.**

I. Introduction

The demand for wireless communications has kept growing in the recent years, to the point of making the traditional fixed licensing process obsolete [1], [2]. *Cooperative Spectrum Sharing* (CSS) [3] has been proposed as a more efficient management mechanism of spectrum access. The basic premise of CSS is that secondary users (SUs) without license may act as transmission relays for a licensed primary transmitter (PT) in exchange for transmission opportunities in the spectral resources of the PT. CSS fosters the creation of transmission opportunities in the PT spectrum: by increasing the transmission rate of the PT, it reduces its spectrum usage. A CSS system poses the following *key challenges*: 1) the PT has to undergo a negotiating process with the nearby SUs, having no previous information about them, in general; 2) the SUs may belong to self-interested networks different from the PT's, and thus, the PT should not expect that the SUs will collaborate in maximizing the PT's profit; 3) spectrum opportunities may happen on a short timescale (of the order of seconds or less), thus, for CSS to be effective, this negotiation must be carried out in real-time.

A. Related work

Previous works in CSS do not address these issues simultaneously. As we discussed in [4], they are focused on requirements 1 and 2, but the time required to reach elaborated solutions (requirement 3) is not adequately studied. Multiple factors in spectrum trading (*e.g.,* supply, demand, channel gains) vary rapidly with time, and trying to reach a complex allocation solution for a particular spectrum opportunity, as in [5], could take so long so as to render the solution impractical.

Works such as [3], [6] assume the PT holds previous information about the SUs. However, it is unlikely that SUs belonging to a different operator would reveal private information (*e.g.,* battery level) to the PT, taking into account that they may improve their utility functions by hiding it, at the cost of worsening the performance of the whole system. In [7] the authors do not make this assumption and develop a stochastic optimization based on contract theory. Nevertheless, in contrast to our proposal, they do not implement any learning process from successive interactions with SUs.

Although it is possible to design mechanisms that provide incentives to selfish SUs to collaborate with the PT (cooperation in terms of game theory [8]), they require the exchange of several messages between these individuals, and therefore a loss in transmission efficiency. These selfish entities, however, do have incentives to make strategies and/or collude against the PT. In our proposal the PT employs one-to-one bargaining instead of broadcast offers, as [9], but we consider more than one SU. The benefits of one-to-one transactions are: the reduction of the strategic power of the SUs, as they cannot overhear public offers or other information about their competitors; their robustness against collusions of SUs [10]; and the reduction of the communication overhead on the control channels in comparison to widespread mechanisms such as auctions [5] (*e.g.,* multiple rounds of bidding messages).

B. Contribution

We focus on meeting all the aforementioned requirements from the perspective of a PT, requesting help from selfish SUs to communicate with an intended primary receiver (PR). In each transmission period, the PT has to choose an SU to act as a relay for its transmission and an offer in terms of transmission time for the SU. The PT obtains a payoff related to the achieved transmission rate and offer. *We propose a novel algorithm for the PT to gradually learn the optimal SU and offer combination based on the rewards it observes*. Because in a realistic setting the situation around the PT can change quickly (*e.g.,* SUs arriving and leaving the system), our algorithm aims to achieve a balance between the time

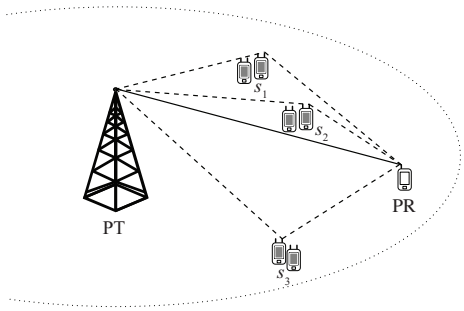

Fig. 1. Cooperative Spectrum Sharing scenario.

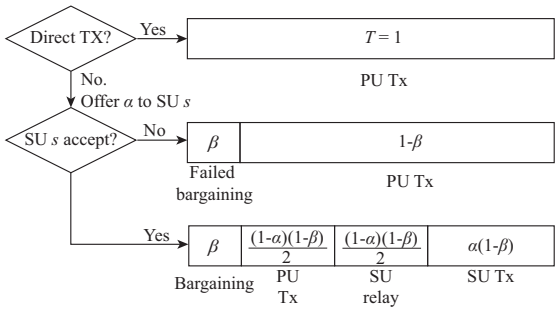

Fig. 2. Frame structures.

devoted by the PT to exploration of options and the time spent exploiting the best alternative known so far, to maximize the PT's payoff over time. In Section II, *we formulate the problem by means of a multi-armed bandit problem (MAB) [11] with dependent arms*. Multi-armed bandits have been previously used in spectrum sharing, but mainly for sensing or resource allocation [12]. *We solve the MAB problem in Section III, combining stochastic MAB index policies and a Markov Decision Process (MDP),* inspired by [13]. However, our dependency model is different enough to require substantial changes in the algorithm, such as the integration with an MDP. We evaluate the performance, scalability and robustness of our approach with Monte Carlo simulations as shown in Section IV. Our solution is not only directly implementable and without strong assumptions, but also extensible to more complex scenarios, as indicated in Section V.

II. System model

The protocol considers a PU transmitter (PT) and receiver (PR) pair and a set of SU cognitive pairs in the coverage area of the PT denoted by $\mathcal{S} \equiv \{s_1, s_2, \ldots, s_S\}$, as in Fig. 1. The system is under the "exclusive-use" coexistence model by which the PUs are the only entities with the right to transmit in a certain band. When the PU pair's channel conditions are not suitable for direct transmission, the PT would be willing to use the SUs as relays.

In exchange for its services, the PT makes an offer to the SU, consisting of a certain amount of time for SU data transmission over the PU channels. The SUs transmit with fixed power, the same for relaying and their own transmissions. The SUs are assumed to have their own, but limited, spectral resources. Therefore, although it is not crucial for their communication purposes, the SUs may benefit from the additional spectrum resources obtained from the PU. As a consequence, it is the PT who contacts the SU and initiates the bargaining.

Time is divided into fixed transmission periods or *frames*, which we will consider of duration $T = 1$ units of time and numbered as $n = 1, 2, \ldots$. The offer that the PT makes, denoted by $\alpha \in [0, 1]$, is the fraction of the transmission period during which the SU is allowed to transmit its own data.[1] For tractability, we discretize that interval in equal increments and $\mathcal{A} = \{\alpha_1, \ldots, \alpha_A\}$ denotes the set of possible offers.

[1]In practice we will assume $\alpha \in [z, 0.9]$, with $z > 0$, as neither the SU nor the PT may "work for free."

Each transmission period is composed of a decision phase, an optional bargaining phase, and a communication phase as in [3] and [9]. See also Fig. 2. During the *decision phase*, the PT has to choose which SU to bargain with and what offer to make to it: $(s, \alpha) \in \mathcal{S} \times \mathcal{A}$, aiming to maximize its utility function, and based on previous interactions. We consider the PT has full bargaining power and makes a take-it-or-leave-it offer to an SU. The *bargaining phase* models the time spent in sending the offer and waiting for an answer. Finally, during the *communication phases* the PT directly transmits for the rest of the frame to the PR if the offer is rejected by the SU. If the offer is accepted by the SU: first, the PT transmits its information to its receiver and to the selected SU; secondly, the selected SU re-transmits the information to the PR; and finally, SU's own transmission takes place.

A. Payoff Functions

PT's payoff. For the cooperative transmission, the system employs the decode and forward relay communication scheme from [14]. For a given SU $s \in \mathcal{S}$, the links initially involved are: PT-PR, PT-s, and s-PR. Let $\Gamma_{\text{P,P}}^{(n)}$, $\Gamma_{\text{P},s}^{(n)}$, and $\Gamma_s^{(n)}$ be their respective SNR values averaged over the duration of transmission frame n. The achievable data rate satisfies $R_{\text{P}}^{(n)} = K \min\{\log(1+\Gamma_{\text{P},s}^{(n)}), \log(1+\Gamma_{\text{P,P}}^{(n)}+\Gamma_s^{(n)})\}$, where K is a constant. We assume that the SUs always decode the PT data correctly in the first phase, and thus we focus on the s-PR link and we have $R_{\text{P}}^{(n)} = K \log(1 + \Gamma_{\text{P,P}}^{(n)} + \Gamma_s^{(n)})$. Since the PT-PR and s-PR links are considered to be in bad and good propagation conditions respectively, *i.e.,* $\Gamma_{\text{P,P}}^{(n)} \ll \Gamma_s^{(n)}$, and $1 \ll \Gamma_s^{(n)}$ $\forall n$, the achievable data rate can be approximated as $R_{\text{P}}^{(n)} \approx K \log(\Gamma_s^{(n)})$.

The time allocated to the s-PR transmission within a frame can be considered long compared to fast fading variations. That is, the effect of multipath is assumed to be negligible in terms of the average SNR, $\Gamma_s^{(n)}$. Therefore, $\Gamma_s^{(n)}$ is mainly determined by pathloss attenuation and shadowing, remaining constant during an s-PR transmission. Successive SU relay phases between s and the PR are sufficiently distant in time for the channel to decorrelate. Thus, the SNR samples $\Gamma_s^{(n)}$ are modeled as i.i.d. random variables, following a log-normal distribution which typically characterizes shadowing [15], *i.e.,* $\log(\Gamma_s^{(n)}) = \gamma_s^{(n)} \sim N(\mu_{\gamma_s}, \sigma_{\gamma_s})$. The PR can observe the value of $\gamma_s^{(n)}$ and feed it back to the PT at the end of each

transmission frame. Then, the reward for the PT in frame n is the transmission rate it gets, multiplied by the time the transmission lasts:

$$W^{(n)}(s,\alpha) = \begin{cases} (1-\beta)(1-\alpha)\log(\Gamma_s^{(n)}) & \text{if } \alpha \text{ is accepted} \\ (1-\beta)\log(1+\Gamma_{\text{P,P}}^{(n)}) & \text{otherwise} \end{cases} \quad (1)$$

For the sake of clarity, from now on we set $\beta = 0$ and $\log(1+\Gamma_{\text{P,P}}^{(n)}) = 0$ for all n. Nevertheless, our model is also applicable with different values of these parameters.

SU's payoff. The payoff obtained by an SU acting as relay is the difference between its net transmitted data during a frame when using the PT channel, αR_s^{P}, and the net data it would transmit when using its own spectral resources, R_s. An SU accepts any offer that provides positive payoff, which results in a threshold behavior. SU s accepts an offer $\alpha_a \in \mathcal{A}$ whenever $\alpha_a \geq \alpha_s^*$, where α_s^* is the minimum offer SU s is willing to accept. The *type of an SU*, τ_s, is the index of the smallest offer that this SU accepts, i.e. $\tau_s = \arg\min_a\{\alpha_a : \alpha_a \geq \alpha_s^*, \alpha_a \in \mathcal{A}\}$. If the link between the SU pair is stable (*e.g.,* a close ad hoc connection) and the offered PT bandwidth is constant (only the time offered changes), the thresholds, and therefore the type of each SU, remain constant over multiple transmission frames.

B. Multi-armed bandit formulation

Mathematically, we model the sequential decisions of the PT as a multi-armed bandit (MAB) problem. The PT selects SU-offer pairs (s,α) from the action set $\mathcal{U} = \mathcal{S} \times \mathcal{A}$. In the MAB model each $u = (s,\alpha)$ is an arm, and $\mathcal{U}$ is the set of arms. At round n, the arm pulled is $u^{(n)} = (s^{(n)}, \alpha^{(n)})$, and the reward received by the PT is $W^{(n)} = W^{(n)}(u^{(n)})$. The history of the system up to time n, $h^{(n)}$, is defined as the sequence of decisions and observed rewards: $h^{(n)} = W^{(0)}, u^{(1)}, W^{(1)}, \ldots u^{(n)}, W^{(n)}$ (where $W^{(0)}$ corresponds to initial samples $\gamma_s^{(0)}$). A *policy* π is a function that, at each stage n, prescribes a decision $u^{(n+1)}$ based on $h^{(n)}$. Therefore π induces a history $W^{(0)}, u_\pi^{(1)}, W_\pi^{(1)}, u_\pi^{(2)}, W_\pi^{(2)}, \ldots$. The usual performance metric in learning problems is the *regret*. At decision stage n, we define the regret of a policy π as $r^{(n)} = \max_u \mathbb{E}\left[\sum_{k=1}^n \left(W^{(k)}(u) - W_\pi^{(k)}\right)\right]$. The regret quantifies the performance loss of a policy with respect to a policy that knows the average values μ_{γ_s} and the type of each SU. However, in a realistic setting, the PT has no initial information and faces the challenge of learning both the SU types and μ_{γ_s}, while trying to maximize the reward. This is known as the *exploration - exploitation* tradeoff, which implies balancing immediate gains (pulling the arms that seem to be better in expectation) with gaining information to make better decisions in the next rounds (pulling arms that seem to be worse initially but could potentially be the best).

Learning about the arms of the MAB should not be treated independently. The reward obtained from an SU s when offered a particular α will give the PT information about the rewards it can obtain with all the other offers or associated arms. The following example illustrates why.

Example. Consider the following set of offers $\mathcal{A} = \{0.3, 0.5, 0.7, 0.9\}$, and assume that the PT's initial beliefs about the probability that an SU, $s \in \mathcal{S}$, is of a particular type are equiprobable, *i.e.,* $P(\alpha_a < \alpha_s^* \leq \alpha_{a+1}) = 0.25$ for $a = 1 \ldots A-1; P(\alpha_s^* \leq \alpha_1) = 0.25$. Thus, the beliefs about the probabilities of acceptance of each offer by s are $P(\alpha_s^* \leq \alpha_1) = 0.25$, $P(\alpha_s^* \leq \alpha_2) = 0.5$, and so forth. With these probabilities we can build the following initial *belief vector*: $[0.25, 0.50, 0.75, 1]$. If the PT makes the offer $\alpha = 0.5$ to SU s and s accepts it, then the PT will learn that s is neither type 3 nor type 4. Therefore, the belief vector is updated to $[0.5, 1, 1, 1]$. In addition, the PT observes a sample of γ_s and updates its sample mean $\overline{\gamma}_s^{(1)}$, affecting the belief about the reward of all the arms $u = (s, \alpha_a)$ of that SU s.

In the following section we explain how, inspired by [13], we handle correlation by grouping the arms of an SU, exploiting mutual information.

III. MAB - MDP ALGORITHM

We could think of a reduced MAB with S arms, one for each SU, where each arm integrates the information the PT has observed about each SU up to stage n, that is, the belief about the acceptance of offers and the sample average SNR $\overline{\gamma}_s^{(n)}$. We must then answer two questions: *how do we represent an SU in the reduced MAB?* Once an SU has been chosen by means of the reduced MAB, *what offer should the PT make to the selected SU?*

Offer selection policy. We are interested in finding a policy π_s^{MDP} that, given an SU $s \in \mathcal{S}$, maps PT's knowledge about the type of s, to the next offer α to s. Given α, the reward ψ that the MDP observes is $(1-\alpha)$ if the offer is accepted, and 0 otherwise. Therefore, a given policy π_s^{MDP} induces a sequence of rewards $\psi^{(1)}, \psi^{(2)}, \ldots, \psi^{(n)}$. The uncertainty about the time horizon n is captured by a discount factor $\delta < 1$, characterizing the expected lifetime of the system, e.g., the probability that the s-PR pair remains active in each frame. The search problem of discovering the optimal offer to s consists of finding the policy maximizing $\mathbb{E}\left[\sum_{n=1}^\infty \delta^n \psi^{(n)}\right]$. This problem is formulated as an MDP as follows.

The set of *states* of the MDP for SU s, denoted by $\mathcal{X}_s$, are all the possible knowledge states about the type of s. Assuming uniform probability for the type of an SU, the state of the MDP is completely defined by a two dimensional vector x, whose elements, h and l, contain the index of the highest offer rejected by s, and the index of the lowest offer that s accepted, respectively. The PT knows that the SU accepts every offer α_a with index $a \geq l$, and rejects those with index $a \leq h$, $\alpha_a \in \mathcal{A}$. The initial state is $(0, A)$, since the PT does not know anything about rejections initially ($h = 0$) and α_A is known to be surely accepted ($l = A$).

Each *transition probability* $P_s(x, x', \alpha_a)$ from a state x to a state x', with $x, x' \in \mathcal{X}_s$, given an offer α_a, is determined by the PT's beliefs about the acceptance probability of the offer

α_a at current knowledge state, x. This probability is defined as:

$$P(\alpha_s^* \leq \alpha_a | x) = \begin{cases} 0 & \text{for } 0 < a \leq h \\ \frac{a-h}{l-h} & \text{for } h < a < l \\ 1 & \text{for } l \leq a \leq A \end{cases} \quad (2)$$

Therefore, the transition probabilities between every pair of states are given by:

$$P_s(x, x', \alpha_a) = \begin{cases} P(\alpha_s^* \leq \alpha_a | x) & \text{for } a = l' \\ 1 - P(\alpha_s^* \leq \alpha_a | x) & \text{for } a = h' \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Finally, the reward associated to a transition of the MDP is:

$$\psi(x', \alpha_a) = \begin{cases} 1 - \alpha_a & \text{for } a = l' \\ 0 & \text{for } a = h' \end{cases} \quad (4)$$

We can now formulate the Bellman equation that allows us to obtain the value function V_s for each state:

$$V_s(x) = \max_{\alpha_a \in \mathcal{A}} \sum_{x' \in \mathcal{X}_s} P_s(x, x', \alpha_a)(\psi(x', \alpha_a) + \delta V_s(x')) \quad (5)$$

which can be readily solved offline by standard algorithms such as policy iteration [16].

SU representation in the reduced MAB. In a classic stochastic MAB, rewards when pulling an arm are drawn from a probability distribution associated to that arm. Proposed policies in the literature compute an index for each arm (dependent on that arm only), $I_s^{(n)}$, and dictate to pull the arm with the highest index on each decision round. A commonly used family of index policies are the *Upper Confidence Bound (UCB)* policies proposed by [11]. The index of these policies consists of the sample average reward obtained from arm s up to round n, plus an additional term, the UCB, related to the uncertainty of that estimation.

In our case, the reward W when pulling arm s depends not only on s but also on α: $W^{(n)}(s, \alpha) = (1-\alpha)\log(\Gamma_s^{(n)})$. When choosing arm s and offering α, the reward is drawn from the Gaussian distribution $\log(\Gamma_s^{(n)}) \sim N(\mu_{\gamma_s}, \sigma_{\gamma_s})$, multiplied by a constant $(1-\alpha)$. Let us denote by $x_s^{(n)} = (h_s^{(n)}, l_s^{(n)}) \in \mathcal{X}_s$ the state of SU s at round n. We characterize the SU in the reduced MAB by its *presumed best achievable offer* $\alpha^{\min}(x_s^{(n)})$, denoting the minimum $\alpha \in \mathcal{A}$ in state $x_s^{(n)}$ with positive belief of being accepted and also being achievable by the offer selection policy π_s^{MDP}.[2]

By applying this SU characterization to the UCB index for Gaussian distributions shown in [11], we obtain:

$$I_s^{(n)} = \widehat{W}_s^{(n)} + 4\hat{\sigma}_{W_s}(x_s^{(n)}) \sqrt{\frac{\ln(n_{\{W>0\}} + 2)}{n_s + 1}} \quad (6)$$

where $\widehat{W}_s^{(n)}$ is the estimated average reward of arm s up to round n, $\hat{\sigma}_{W_s}(x_s^{(n)})$ is the estimated standard deviation of the rewards of arm s, n_s is the number of times arm s has been pulled up to round n, and $n_{\{W>0\}}$ is the number of rounds in which the PT obtained a positive reward from any SU. Since s is characterized by its presumed best achievable offer in state $x_s^{(n)}$, we have that $\widehat{W}_s^{(n)} = (1 - \alpha^{\min}(x_s^{(n)}))\overline{\gamma}_s^{(n)}$ and $\hat{\sigma}_{W_s}(x_s^{(n)}) = (1 - \alpha^{\min}(x_s^{(n)}))\sigma_{\gamma_s}$, with $\overline{\gamma}_s^{(n)}$ and σ_{γ_s} denoting the sample mean and known standard deviation of $\gamma_s^{(n)}$, respectively. In practice, removing the factor 4 from the UCB yields a significantly lower regret, while still achieving convergence. However, despite this empirical evidence, there are no theoretical bounds on the regret for this version of the UCB.

[2]For a policy π_s^{MDP} that ends up exploring all the offers of an SU, $\alpha^{\min}(x_s^{(n)})$ is simply the lowest α not rejected for the current state x. For more conservative policies that may dictate not to explore all offers, $\alpha^{\min}(x_s^{(n)})$ is the lowest α that policy is willing to explore given the state $x_s^{(n)}$. Note that, in general, $\alpha^{\min}(x_s^{(n)}) \neq \alpha_a$, with $\alpha_a = \pi_s^{\text{MDP}}(x_s^{(n)})$.

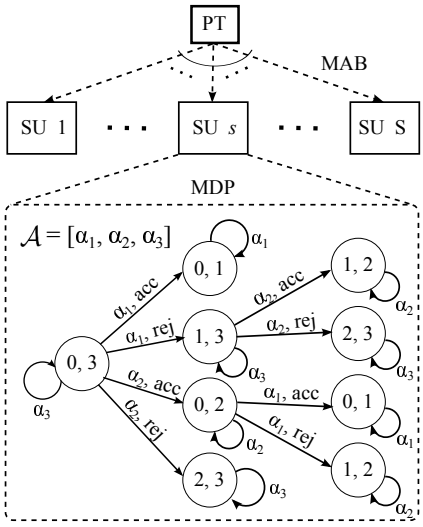

Fig. 3. MAB-MDP algorithm for S SUs and $A = 3$ possible offers . Each SU is chosen according to a MAB and the offer according to an MDP. "acc" represents the event that the offer is accepted and "rej" that it is rejected.

We proceed to give a detailed **description of the MAB - MDP algorithm**. Its diagram can be found in Fig. 3. Initially, for each SU $s \in \mathcal{S}$:

1) The PT computes the exploration-exploitation policy π_s^{MDP} for the offers $\alpha \in \mathcal{A}$.
2) Given that policy, for the initial state $x_s^{(0)} = (0, A)$, the PT computes the presumed best achievable offer $\alpha^{\min}(x_s^{(0)})$ of each SU. The PT chooses that offer as representative of the SU.
3) With $\alpha^{\min}(x_s^{(0)})$ and the initial sample of the SNR $\gamma_s^{(0)}$, the PT builds the UCB index $I_s^{(0)}$ (6).

Then, in each round n:

4) The PT selects the SU z with the highest UCB index $I_z^{(n)}$ and the offer $\alpha_a \in \mathcal{A}$ is indicated by the MDP policy π_z^{MDP}.
5) If the SU *rejects* the offer, the PT updates its knowledge state (MDP) $x_z^{(n+1)}$ and checks if $\alpha^{\min}(x_z^{(n+1)}) \neq \alpha^{\min}(x_z^{(n)})$. If that is the case, the PT computes $I_z^{(n+1)}$ according to (6). Otherwise, $I_z^{(n+1)} = I_z^{(n)}$.
6) If the SU *accepts*, the PT updates $x_z^{(n+1)}$, and sets $n_z = n_z + 1$. The indices I_s, for all $s \in \mathcal{S}$, are also updated

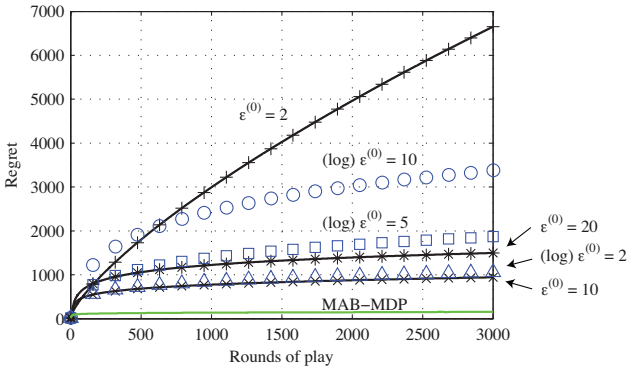

Fig. 4. Regret over time for the MAB-MDP and the ϵ-greedy algorithms. The results are averaged over 2000 independent experiments.

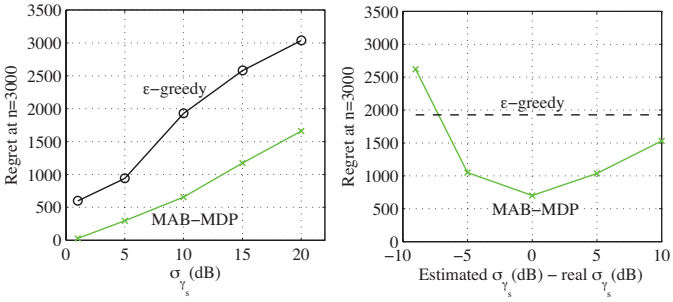

Fig. 5. Mean SNRs and SU types are fixed over experiments. $\boldsymbol{\mu_\gamma} = \{35, 32.5, 32.5, 30, 30, 27, 27\}$ and SU types $= \{1, 1, 1, 2, 2, 3, 3\}$. On the left, regret at $n = 3000$ for the MAB-MDP and the ϵ-greedy algorithms for different values of σ_{γ_s}. On the right, regret at $n = 3000$ for different estimations of σ_{γ_s}, with the real value of $\sigma_{\gamma_s} = 10$ dB.

because $n_{\{W>0\}} = n_{\{W>0\}} + 1$.[3]

Note that it cannot be assured that this algorithm performs a fully optimal learning. An optimal learning algorithm would imply solving a continuous-space MDP comprising all the information gathered by the PT, which is intractable in practice. Our MAB-MDP algorithm decomposes the learning problem into two simpler sub-problems (the MAB and the MDP). Each sub-problem can be solved very efficiently because it uses only partial information about the global system. Despite its simplicity, MAB-MDP shows a remarkably low regret in numerical evaluation.

IV. Numerical results

Benchmark strategy. Given the main features of our proposed mechanism (real-time operation and exploration-exploitation tradeoff), the best candidates for comparison are the families of reinforcement learning heuristics [17] known as ϵ-greedy policies. Specifically, we will compare our algorithm to the ϵ-descending strategy. This strategy chooses the alternative $u = (s, \alpha)$ with best expected reward with probability $1 - \epsilon^{(n)}$, and performs random exploration with probability $\epsilon^{(n)}$, which, in each round n, takes the value $\epsilon^{(n)} = \min\left(\epsilon^{(0)}/n, 1\right)$, or $\epsilon^{(n)} = \min\left(\epsilon^{(0)} \ln(n)/n, 1\right)$ for a log-descending variant, where $\epsilon^{(0)}$ is a tunable parameter. The alternative u is chosen according to: $u = (s, \alpha) = \arg\max_{s,\alpha}(1-\alpha)\overline{\gamma}_s^{(n)} P(\alpha_s^* \leq \alpha)$. As noted in [11], [17], the ϵ-descending greedy strategy performs as well as (and most of the time, even better than) many other complex policies. Its main drawback is that the $\epsilon^{(0)}$ parameter has to be carefully chosen and there is little that can be said theoretically about its optimal value except for distributions with support on $[0, 1]$ (see [11]). We show the performance of ϵ-greedy for several values of $\epsilon^{(0)}$ in Fig. 4, for both variants of the strategy. We empirically set $\epsilon^{(0)} = 10$ for all other figures. We set the discount factor δ of the MAB-MDP to 0.98.

Fig. 4 shows the performance of our proposal over time compared to the ϵ-greedy, for $S = 7$ SUs and $A = 7$ possible α offers. The standard deviation of the SNR of each arm s is $\sigma_{\gamma_s} = 5$ dB. Our proposal, MAB-MDP, makes a more effective use of the information than ϵ-greedy, and therefore one can expect a significantly better performance in terms of regret, as shown in the figure.

1) Performance versus standard deviation: On the left of Fig. 5 the performance of the algorithms is shown versus different values of σ_{γ_s}. The higher the deviations, the harder the problem, as the samples of each distribution can overlap. The MAB-MDP standard deviation term is adjusted accordingly. The ϵ-greedy policy experiences a more notable performance degradation under high variance. This is because it relies just on estimates of the mean reward. The higher the variance of the samples, the more the rounds needed for the sample mean to get close to the true average value. Nevertheless, due to its random exploration nature, ϵ-greedy manages to eventually find the optimal arm. The proposed UCB-based algorithm shows certain robustness under high variances, since it takes the standard deviation into account when making its decisions.

2) Performance under mis-estimation of standard deviation: On the right of Fig. 5 the effect of misestimating the standard deviation σ_{γ_s} in the MAB-MDP algorithm is shown. ϵ-greedy performs a frequentist inference and therefore does not make use of the variance. Underestimating or overestimating σ_{γ_s} implies being less or more optimistic, respectively, about the values of the true average rewards of the SUs, based on the sample means observed. Thus, an overestimation implies more exploration of each SU and an increased probability of finding the optimal arm. An underestimation implies more exploitation of SUs and less exploration. As we can see in the figure, overestimation is safer than underestimation in terms of regret. Overestimation leads to a slower convergence towards the optimal arm but underestimation could lead to convergence to a suboptimal solution and thus, linear regret with the number of rounds: the difference between the rewards of the optimal and that suboptimal arm on each round accumulates over time. Note that, under severe underestimation, MAB-MDP can even do worse than the ϵ-greedy algorithm. Underestimation may be justified in a scenario where the SUs are expected to stay a short time in the coverage area of the PT. Then, the PT

[3]In a classic stochastic MAB, every arm pull increases $n_{\{W>0\}}$ because every arm pull gives a reward from a probability distribution. This is not our case due to the possibility that an offer is rejected.

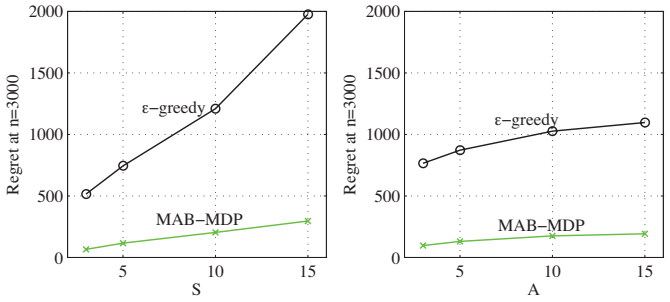

Fig. 6. Regret at $n = 3000$ for different number of SUs, S, and offers, A.

may be better off sacrificing the search of an optimal arm for exploiting suboptimal arms with "good enough" rewards.

3) Performance against number of SUs and discretization levels of the offers: Fig. 6 illustrate how regret grows with respect to the optimal policy for the different algorithms when increasing the number of SUs S in the system or the granularity of the offers A the PT can make. Note that a higher regret represents worse performance with respect to the optimal policy. With a higher number of SUs or higher number of possible offers, the optimal omniscient policy obtains better rewards. In short: finding the optimal arm is harder but that optimal arm provides a higher reward than the optimal arm of a scenario with fewer SUs or fewer possible offers. That said, both figures show that the proposed MAB-MDP algorithm scales well when increasing the size of the problem, as the regret grows dramatically less than for ϵ-greedy. This is because of the more effective use of the available information made by the UCB-based algorithms. Moreover, the computational cost or memory requirements do not experience a significant growth, even for the worst case, which is increasing the number of offers. The state space of the MDP of an SU, $\mathcal{X}_s$, grows with the number of offers, A, as $|\mathcal{X}_s| = A(A+1)/2$, which for a fine-grained range of 20 possible offers becomes 210 states and 20 actions on each state.

V. Conclusion and Future Work

We have proposed a spectrum trading mechanism in cooperative secondary spectrum access using multi-armed bandits, from the perspective of a primary transmitter (PT), modeling channels under shadowing effects. We have focused on a scenario where the PT has no knowledge of the performance of the SUs acting as relays, nor of the offers they are willing to accept. We have built an algorithm that learns payoff-maximizing actions for the PT with little communication or computation overhead. Our numerical results show that, despite its simplicity, MAB-MDP significantly outperforms the classical exploration-exploitation ϵ-greedy algorithm. MAB-MDP has also been shown to be robust to inaccuracies in the little information it needs and to scale well when the size of the problem increases, *i.e.,* for more SUs and available offers. This work can be the starting point to develop more complex scenarios. Considering the explosion of MAB variants in the recent literature, as the next steps it would be possible and interesting to study: 1) how to exploit the spatial fading correlation across different SUs, 2) extension of the algorithms to a multiple PT and/or multiple PR case, 3) inclusion of more dimensions to learning, such as learning the staying time of SUs in the PT coverage area or the distribution of the SU types, and 4) inclusion of SU and PU strategic behavior.

Acknowledgment

This work was supported by project grant MINECO/FEDER COINS TEC2013-47016-C2-2-R and by the project "A Novel Approach to Wireless Networking based on Cognitive Science and Distributed Intelligence," funded by Fondazione CaRiPaRo under the framework Progetto di Eccellenza 2012. Mario López Martínez also acknowledges the personal grant BES-2011-051051.

References

[1] T. M. Valletti, "Spectrum trading," *Telecommunications Policy*, vol. 25, no. 10-11, pp. 655–670, Oct. 2001.

[2] E. A. Jorswieck, L. Badia, T. Fahldieck, E. Karipidis, and J. Luo, "Spectrum sharing improves the network efficiency for cellular operators," *IEEE Commun. Mag.*, vol. 52, no. 3, pp. 129–136, Dec. 2013.

[3] O. Simeone, I. Stanojev, S. Savazzi, U. Spagnolini, and R. Pickholtz, "Spectrum leasing to cooperating secondary ad hoc networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 203–213, Jan. 2008.

[4] M. López-Martínez, J. J. Alcaraz, J. Vales-Alonso, and J. Garcia-Haro, "Automated spectrum trading mechanisms: Understanding the big picture," *Wireless Networks*, vol. 21, no. 2, pp. 685–708, Jan. 2015.

[5] X. Feng, G. Sun, X. Gan, F. Yang, and X. Tian, "Cooperative spectrum sharing in cognitive radio networks: A distributed matching approach," *IEEE Trans. Commun.*, vol. 62, no. 8, pp. 2651–2644, Aug. 2014.

[6] T. Nadkar, V. Thumar, G. Shenoy, A. Mehta, U. B. Desai, and S. N. Merchant, "A cross-layer framework for symbiotic relaying in cognitive radio networks," in *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, May 2011, pp. 498–509.

[7] L. Duan, L. Gao, and J. Huang, "Cooperative spectrum sharing: A contract-based approach," *IEEE Trans. Mobile Comput.*, vol. 13, no. 1, pp. 174–187, Jan. 2014.

[8] G. Zhang, K. Yang, J. Song, and Y. Li, "Fair and efficient spectrum splitting for unlicensed secondary users in cooperative cognitive radio networks," *Wireless Personal Communications*, vol. 71, no. 1, pp. 299–316, Aug. 2012.

[9] Y. Yan, J. Huang, and J. Wang, "Dynamic bargaining for relay-based cooperative spectrum sharing," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 8, pp. 1480–1493, Aug. 2013.

[10] J. Alcaraz and M. van der Schaar, "Coalitional games with intervention: Application to spectrum leasing in cognitive radio," *IEEE Trans. Wireless Commun.*, vol. 13, no. 11, pp. 6166–6179, Nov. 2014.

[11] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, May 2002.

[12] P. Si, H. Ji, F. R. Yu, and V. C. M. Leung, "Optimal cooperative internetwork spectrum sharing for cognitive radio systems with spectrum pooling," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1760–1768, May 2010.

[13] S. Pandey, D. Chakrabarti, and D. Agarwal, "Multi-armed bandit problems with dependent arms," in *ICML Proceedings of the 24th International Conference on Machine learning*, 2007, pp. 721–728.

[14] J. N. Laneman and G. W. Wornell, "An efficient protocol for realizing distributed spatial diversity in wireless ad-hoc networks," in *Proc. ARL FedLab Symp. on Adv. Telecomm. and Inform. Distrib. Prog. (ATIRP)*, Washington, DC, 2001.

[15] A. Goldsmith, S. Jafar Ali, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," *Proc. IEEE*, vol. 97, no. 5, pp. 894 – 914, May 2009.

[16] D. Bertsekas, *Dynamic Programming and Optimal Control*. Nashua, NH: Athena Scientific, 2000.

[17] J. Vermorel and M. Mohri, "Multi-armed bandit algorithms and empirical evaluation," in *Machine Learning: ECML*, ser. Lecture Notes in Computer Science. Springer, 2005, vol. 3720, pp. 437–448.