Distributed Reinforcement Learning Based MAC Protocols for Autonomous Cognitive Secondary Users

Mario Bkassiny and Sudharman K. Jayaweera
Dept. of Electrical and Computer Engineering
University of New Mexico
Albuquerque, NM, USA
Email:{bkassiny, jayaweera}@ece.unm.edu

Keith A. Avery
Space Vehicle Directorate
Air Force Research Laboratory (AFRL)
Kirtland, AFB, Albuquerque, NM, USA

***Abstract*—We consider a decentralized cognitive radio network in which autonomous secondary users seek spectrum opportunities in licensed spectrum bands. We assume that the primary users' channel occupancy follows a Markovian evolution, and formulate the spectrum sensing problem as a Decentralized Partially Observable Markov Decision Process (DEC-POMDP). We develop a distributed Reinforcement Learning (RL) algorithm that allows each autonomous cognitive radio to distributively learn its own spectrum sensing policy. The resulting decentralized sensing policy enables secondary users to non-cooperatively reach an equilibrium that leads to high utilization of idle channels while minimizing the collisions among secondary cognitive radios. Moreover, we propose a decentralized channel access policy that permits controlling, with high accuracy, the collision probability with primary users. Our numerical results validate the robustness of this collision probability control as the sensing noise changes. They also show the efficiency of the proposed learning algorithm in improving the spectrum utilization.**

***Index Terms*—Cognitive radio, reinforcement learning.**

I. Introduction

Opportunistic Spectrum Access (OSA) [1] has been envisioned as a promising technique to exploit the spectrum vacancies, which permits unlicensed secondary users to access the primary channels opportunistically when the primary users who own the spectrum rights are not transmitting. Cognitive Radio (CR) devices have been founded as a platform to realize such OSA techniques. In general, CR's are assumed to be able to sense and adapt to their Radio Frequency (RF) environment.

In this paper, we consider a decentralized CR network in which each secondary user tries to obtain, independently, the best estimate of the status of the primary channels based on its own local information. In particular, when the primary channel states follow a Markovian evolution, a cognitive user can utilize its history of observations and actions in order to derive a better sensing/accessing policy. This problem can then be formulated as a Decentralized Partially Observable Markov Decision Process (DEC-POMDP) and has been discussed in several recent studies. For example, in [2], the authors suggested a Medium Access Control (MAC) protocol for decentralized ad-hoc CR networks by modeling the system as a POMDP that is equivalent to a Markov Decision Process (MDP) with an infinite number of states. The corresponding optimal sensing policy that maximizes the total discounted return was shown to be *computationally prohibitive*. Thus, an optimal myopic policy was derived such that it maximizes the instantaneous rewards. The myopic policy that was formulated in [2] is optimal for a single-user setup, and is suboptimal when applied to a multiuser setting because it would lead to collisions between secondary users when more than one user try to access the same channel. On the other hand, in [3] the authors proposed three different sensing policies for multiuser OSA: The first policy is based on a cooperative protocol in which secondary users exchange their beliefs about the channel states at each time slot. The second policy applies learning techniques to obtain an estimate of the other users' beliefs, and the third policy is based on a single-user approach in which the cognitive users act non-cooperatively. We note that [3] assumes perfect sensing of the primary channels, which we do not assume throughout this paper.

In [4], a suboptimal sensing/access policy was derived for *cooperative* cognitive networks since it is not easy to solve the Bellman equation that corresponds to the formulated POMDP model. However, the assumed model did not ensure full utilization of spectrum resources because only one primary channel was accessed at each time instant collectively by all secondary users. This leads to low network throughput since all the secondary users are assumed to sense the same primary channel at a time. The main advantage of this model, however, was that it achieves better sensing performance. The trade-off between the sensing accuracy and the secondary throughput has been discussed recently in [5].

We believe that the solution to these issues is to make the so-called CR's indeed cognitive, i.e. to achieve smart performance, the CR's should have the ability to learn from their observed environment and the past actions. Indeed, it can be argued that learning from experience must be at the heart of any cognitive system. Recently, this view is gaining importance within the CR research community as is evident by the application of learning techniques to CR's. For example, the multi-agent Reinforcement Learning (RL) algorithm, known as

978-1-4577-0454-3/11/$26.00 ©2011 IEEE

Q-Learning, was applied in [6] to achieve interference control in decentralized Wireless Regional Area Networks (WRAN). In [7], the authors developed a Q-learning algorithm for an auction-based dynamic spectrum access protocol, which is different from the DEC-POMDP structure of our proposed model. To the best of our knowledge, none of the CR studies that assume an underlying POMDP structure has used the Q-learning algorithm to solve the OSA problem [2]–[4]. The literature on learning techniques to achieve CR goals is still at an infancy, although there is a rich literature on machine learning in computer science and classical statistical learning that provides a great starting point [8].

In this paper, we formulate the channel sensing in decentralized cognitive networks as a DEC-POMDP problem. Unlike [2], our approach considers a multi-user setting and we propose a channel sensing policy that takes into account the collisions among secondary users. Our proposed *sensing* policy is based on the distributed RL. Note that, we use the RL to derive the sensing policy rather than to obtain interference control as in [6]. This algorithm achieves two main goals: Deriving a sensing policy based on the history of actions and observations, and minimizing the collisions between secondary users while competing for channel access opportunities. On the other hand, we propose a channel *access* mechanism that limits the collisions between primary and secondary users when secondary users have noisy observations about the primary channels. Our channel access scheme ensures high accuracy and robustness in controlling the collision probability with primary channels, thus guaranteeing the Quality of Service (QoS) requirements of primary users.

The remainder of this paper is organized as follows: Section II defines the system model. In sections III and IV, we derive both the accessing and sensing policies for cognitive users. We show the simulation results in section V. Section VI concludes the paper.

II. System Model

We consider a wireless network having a set of primary channels $\mathcal{C} = \{1, ..., L\}$. The channels' occupancy states are assumed to be independent and following a Markovian evolution. A set of distributed users form a secondary network that is assumed to rely on cognitive techniques to access these primary channels when they are idle. The set of secondary users in the system is denoted by $\mathcal{K}_s = \{1, ..., K_s\}$. The secondary network forms a multiple access channel in which each secondary user independently searches for a spectrum opportunity in order to communicate with a secondary base station, as depicted in Fig. 1. Every secondary user $j \in \mathcal{K}_s$ is assumed to be able to sense only one primary channel at a time, and we assume that secondary users do not cooperate. This is a reasonable assumption in decentralized networks in which there is no control channels for ensuring collaboration among secondary users.

We identify the overall system made of primary channels and the K_s-secondary users as a DEC-POMDP [9] by defining the state of the system as $\mathbf{s}(k) = (s_1(k), ..., s_L(k)) \in \mathcal{S}$

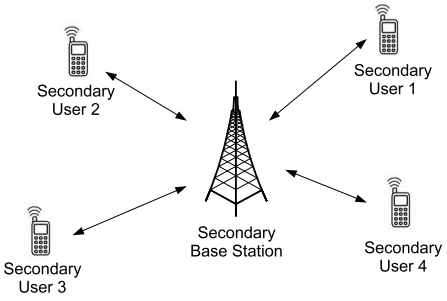

Fig. 1. Cognitive Radio Network (CRN) with distributed secondary nodes

where $s_i(k) \in \{0, 1\}$ represents the state of channel $i \in \mathcal{C}$ as being idle (0) or busy (1) in time slot k, and $\mathcal{S}$ is the set of all possible states $\mathbf{s}(k)$. We define $\mathbf{a} \triangleq (a_1, \cdots, a_{K_s})$ as the joint action of all secondary users (agents) and $P(s, \mathbf{a}, s')$ to be the probability of transition from state s to s' when taking the joint action $\mathbf{a}$. The transitions of every channel's state are independent of the other states and these transitions are assumed to follow a Markovian evolution as mentioned above. The state transition matrix $\mathbb{P}$ of the state vector $\mathbf{s}(k)$ is therefore $\mathbb{P} = \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_L$, where $\mathbb{P}_i$ is the state transition matrix of channel i, and $\otimes$ denotes the Kronecker product.

Note that, the transition probabilities $P(s, \mathbf{a}, s')$ (for $(s, s') \in \mathcal{S}^2$) are independent of the secondary user actions since they are determined by the evolution of the primary channels states, i.e. $P(s, \mathbf{a}, s') = P(s, s')$, where $P(s, s')$ is obtained from the state transition matrix $\mathbb{P}$. Similarly, for an individual channel $i \in \mathcal{C}$, the transition probabilities $P_i(l, l')$ (for $(l, l') \in \{0, 1\}^2$) are obtained from $\mathbb{P}_i$.

The action of secondary user $j \in \mathcal{K}_s$ at time k is denoted by $a_j(k) \in \mathcal{C}$ which represents the index of the primary channel that user $j \in \mathcal{K}_s$ should sense during time slot k. We define $Y_i(k, j)$ to be the observation of secondary $j \in \mathcal{K}_s$ on channel $i \in \mathcal{C}$ in time slot k which is assumed to be the output of a Binary Symmetric Channel (BSC) where $\Pr\{Y_i(k, j) \neq s_i(k)\} = \nu_i$ is the crossover probability. As a result, $Y_i(k, j)$ is a discrete random variable with distinct probability mass functions (pmf) f_0 and f_1 when $s_i(k) = 0$ and $s_i(k) = 1$, respectively.

Let $\mathbf{Y}_i^k(j)$ denotes the vector of observations up to time slot k obtained by secondary $j \in \mathcal{K}_s$ on channel $i \in \mathcal{C}$. Let $\mathbf{K}_i^k(j)$ denote the time slot indices up to slot k when channel i was sensed by secondary user j. Also, let $\mathbf{Y}^k(j) = \{\mathbf{Y}_i^k(j) : i \in \mathcal{C}\}$ be the collection of observations up to slot k on all primary channels obtained by the j-th secondary user.

III. Channel Access Mechanism

The sensing and access operations of the secondary users are scheduled as is shown in Fig. 2, where we consider that a secondary user senses a primary channel during the sensing period τ. Primary users are assumed to always start their

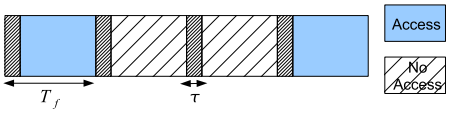

Fig. 2. Channel Access Policies

transmission at the beginning of a frame of duration T_f so that a primary channel will remain free during the secondary access duration if it was free during the corresponding sensing period.

A cognitive device that has sensed a channel can access that channel during the remaining frame duration of $T_f - \tau$. In order to avoid collisions among secondary users, we assume that each secondary user generates a random *backoff time* before transmitting [2]. If more than one secondary users decide to access the same channel, the channel access will be granted to the secondary user that has the smallest *backoff time*.

After sensing channel $i = a_j(k)$, secondary user $j \in \mathcal{K}_s$ decides whether to access channel i based on its observation sequence $\mathbf{y}_i^k(j) \triangleq \{y_i(k', j) : k' \in \mathbf{K}_i^k\}$ where $y_i(k', j)$ is a realization of $Y_i(k', j)$. In order to achieve a probability of collision below a certain bound, we may apply a Neyman-Pearson (NP) type detector [10]. An optimal access decision for the j-th secondary user would choose one of the two possible hypothesis $H_1 = \{s_i(k) = 0\}$ or $H_0 = \{s_i(k) = 1\}$ in time slot k based on the whole observation sequence $\mathbf{y}_i^k(j)$. However, implementing such an optimal detector becomes too complicated due to the need for computing the distribution of the likelihood ratio of $\mathbf{Y}_i^k(j)$ which is a random sequence whose length increases linearly with time. Hence, we simplify the detection rule by assuming that the decision to access a channel in time slot k is based only on the current observation.

Let α be the false alarm probability such that $\alpha \leq 0.5$. The optimal NP detector then is as randomized access decision rule $\tilde{\delta}_i(k, j)$ for secondary j to access channel i at time k. This access decision can be viewed as a binomial random variable denoted by $\delta_i(k, j)$ whose parameter $\tilde{\delta}_i(k, j)$ is given by:

$$\tilde{\delta}_i(k,j) = \begin{cases} \frac{\alpha}{\nu_i} \mathcal{I}_{\{y_i(k,j)=0\}} I_{i,j}^{(k)} & \text{if} \quad \alpha < \nu_i \\ \left(\mathcal{I}_{\{y_i(k,j)=0\}} + \frac{\alpha - \nu_i}{1-\nu_i} \mathcal{I}_{\{y_i(k,j)=1\}} \right) I_{i,j}^{(k)} & \text{if} \quad \alpha \geq \nu_i \end{cases}$$

where $I_{i,j}^{(k)} = \mathcal{I}_{\{a_j(k)=i\}}$, and $\mathcal{I}_B = 1$ if condition B is satisfied, and 0 otherwise. Therefore, secondary user j decides to access a sensed channel i in time slot k only if $\delta_i(k, j) = 1$, which happens with probability $\tilde{\delta}_i(k, j)$.

It can be observed that the collision probability on a particular channel can go beyond the desired threshold because the accessing rule in a decentralized network follows an OR-rule. For that reason, we will design a channel access mechanism that guarantees a certain collision probability with the primary channels.

We define $E_{j,i}(k)$ to be the event that secondary user $j \in \mathcal{K}_s$ decides to access channel $i \in \mathcal{C}$ at time k, given that secondary user j has sensed channel i at time k. Also, we let $E_i(k)$ to be the event that channel $i \in \mathcal{C}$ is busy at time k. When several secondary users sense and try to access the same primary channel $i \in \mathcal{C}$, we define the resulting collision probability as $P_c(i) = \Pr\left\{\bigcup_{j \in \mathcal{Z}_i(k)} E_{j,i}(k) | E_i(k)\right\}$, where $\mathcal{Z}_i(k)$ is the set of secondary users that sense channel i in time slot k.

Note that the events $\{E_{j,i}(k)|E_i(k) : j \in \mathcal{K}_s\}$ are independent because each secondary user makes its access decision independently of the other users, after having sensed the channel i. As a result, the collision probability on channel i can be expressed as $P_c(i) = 1 - (1-\alpha)^{Z_i(k)}$, where $Z_i(k) = |\mathcal{Z}_i(k)|$ and $\alpha = \Pr\{E_{j,i}(k)|E_i(k)\}$ is the false alarm probability of each secondary detector that results from claiming $H_1 = \{s_i(k) = 0\}$ (or equivalently $\{\delta_i(k, j) = 1\}$) when $H_0 = \{s_i(k) = 1\}$ is true. Therefore, in order to ensure an overall collision probability $P_c(i) = \alpha_0$ in channel i, each secondary user $j \in \mathcal{Z}_i(k)$ should set its false alarm probability to $\alpha = 1 - (1-\alpha_0)^{1/Z_i(k)}$.

Since each secondary user does not know the total number of users $Z_i(k)$ that are sensing primary channel $i \in \mathcal{C}$ at a particular time k, it uses the expected value of $Z_i(k)$ to compute its false alarm probability such that $\alpha = 1 - (1-\alpha_0)^{1/\mathbb{E}\{Z_i(k)\}}$. We will compute this expected value in the followings and show, through simulations, that the proposed access technique can guarantee an upper bound on the collision between primary and secondary users.

IV. Sensing policies of Distributed Secondary Users

We define the belief vector of channel $i \in \mathcal{C}$ as $\mathbf{p}(k, j, i) = [p_0(k, j, i), p_1(k, j, i)]$ where $p_l(k, j, i) = \Pr\{s_i(k) = l|\mathbf{Y}_i^{k-1}(j)\}$ which represents the probability of $s_i(k)$ being in state $l \in \{0, 1\}$ in time slot k, given the past observations $\mathbf{Y}_i^{k-1}(j)$. Let $\mathbf{b}_j(k) = \left[b_j(1, k), \cdots, b_j(2^L, k)\right]$ be the belief vector of the primary system according to secondary user j, where

$$b_j\left(u\left(\mathbf{s}\left(k\right)\right), k\right) = \prod_{i=1}^{L} p_{s_i(k)}(k, j, i), \tag{1}$$

given that $u(\mathbf{s}) \in \mathcal{U} = \left\{1, \cdots, 2^L\right\}$ is the index of state $\mathbf{s}(k) = (s_1(k), \cdots, s_L(k))$. The belief vector $\mathbf{b}_j(k)$ is a sufficient statistic for an optimal OSA protocol in a single-user setup [2]. However, in our case, we consider a distributed multi-user scenario and $\mathbf{b}_j(k)$ is no longer a sufficient statistic for optimal decisions. But since we are interested in applying RL techniques to solve the DEC-POMDP problem, we may still use belief vector $\mathbf{b}_j(k)$ to obtain a reasonably good suboptimal solution in a distributed multi-user setting, as shown in [6]. This would simplify the problem, yet leading to near-optimal solutions.

At each time slot, each secondary user updates its belief vector about the states of the channels in the next slot. Suppose secondary user j senses channel $i = a_j(k)$ in time slot k and observes $Y_i(k, j)$. Then it updates its belief about the state of

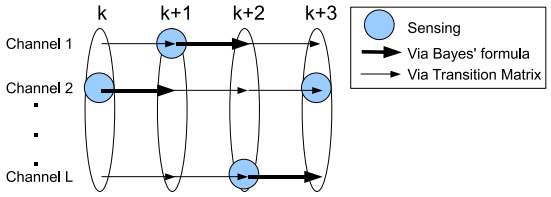

Fig. 3. Sensing and Updating the Beliefs

channel i in time $k+1$ using Bayes' formula as follows:

$$p_m(k+1,j,i) = \frac{\sum_{l=0}^{1} P_i(l,m) f_l(Y_i(k,j)) p_l(k,j,i)}{\sum_{l=0}^{1} f_l(Y_i(k,j)) p_l(k,j,i)}, \quad (2)$$

where $m \in \{0,1\}$. For the *unsensed* primary channels $i' \neq a_j(k)$, the j-th secondary user's belief vector is simply updated based on the assumed Markovian evolution: $\mathbf{p}(k+1,j,i') = \mathbf{p}(k,j,i')\mathbb{P}_{i'}, \forall i' \neq a_j(k)$.

Figure 3 shows the update procedure in which thick arrows represent the updates using Bayes' formula, whereas thin arrows represent the updating of beliefs based only on the assumed Markovian nature of the channels.

A. The Reward and Value functions

We define the total discounted return of user $j \in \mathcal{K}_s$ in time slot k as $R_j(k) = \sum_{n=0}^{\infty} \gamma^n r_j(k+n)$, where $r_j(k)$ is the reward of secondary user j in time slot k and $\gamma \in (0,1)$ is a discounting factor. In a fully observable MDP, an agent $j \in \mathcal{K}_s$ may define the value of a state $\mathbf{s}$ in slot k and under a policy π_j as [8]:

$$V_j^{\pi_j}(\mathbf{s},k) = \mathbb{E}\left\{R_j(k) | \mathbf{s}(k) = \mathbf{s}\right\}. \quad (3)$$

Similarly, the function $Q_j(s,a)$ is defined as the expected return starting from state s, taking the action a, and then following a policy π_j thereafter as:

$$Q_j^{\pi_j}(\mathbf{s},a,k) = \mathbb{E}\left\{R_j(k) | \mathbf{s}(k) = \mathbf{s}, a_j(k) = a\right\}. \quad (4)$$

In the case of a POMDP, however, the actual state of the system is the belief vector $\mathbf{b}_j(k)$. Hence, the resulting process is an infinite state MDP which makes the solutions of (3) and (4) computationally expensive. In particular, our assumed model of a DEC-POMDP is a non-cooperative multi-agent system whose solution is shown to be NEXP-hard [9]. Hence, we will solve this problem by finding the Q values of the DEC-POMDP model by using the underlying MDP model [11], as explained in the next section.

B. Reinforcement Learning for DEC-POMDP

In the following, we extend the Q-learning algorithm that is defined for centralized fully observable environments in [8] by extending it to the partially observable channel sensing problem. This can be made by assigning a $Q(s,a)$ table for each secondary user j, where $s \in \mathcal{S}$ is the channels' states vector with $u(s) \in \mathcal{U} = \left\{1, \cdots, 2^L\right\}$ being the index of state s and $a \in \mathcal{C}$ is the index of the sensed channel. However, we do not use the belief vector $\mathbf{b}_j(k)$ as the actual state. Instead, we solve for the values of $Q(s,a)$ in the underlying MDP model by using $\mathbf{b}_j(k)$ as a weighting vector, as described in [11]. Although this is not the optimal solution of the DEC-POMDP problem, [11] shows that this approach leads to a near-optimal solution with a very low computational complexity if the algorithm adopts an ε-greedy policy [8].

Since the secondary users cannot fully observe the state of the primary system in the POMDP environment, the sensing policy of each secondary user is based on the belief vector $\mathbf{b}_j(k) = \left[b_j(1,k), ..., b_j(2^L,k)\right]$. We describe the Q-learning

Algorithm 1 Q-learning Algorithm for agent $j \in \mathcal{K}_s$

for each $s \in \mathcal{S}$. $a \in \mathcal{C}$ **do**
  Initialize $Q(s,a) = 0$.
end for
Initialize the belief vector $\mathbf{b}$ arbitrarily.
for each time slot k **do**
  Generate a random number rnd between 0 and 1.
  if $rnd < \varepsilon$ **then**
    Select action a^* randomly.
  else
    Select action $a^* = \arg\max_a Q_{\mathbf{b}}(a)$.
  end if
  Execute action a^* (i.e. sense channel a^*).
  Receive the immediate reward $r_j(k)$.
  Update $p_0(a^*,k,j)$ using the observation $y(k)$:
  $p_0(a^*,k,j) \leftarrow \frac{f_0(y(k)) p_0(a^*,k,j)}{\sum_{l=0}^{1} f_l(y(k)) p_l(a^*,k,j)}$
  Update the current belief $\mathbf{b}$ according to $p_0(a^*,k,j)$.
  Evaluate the next belief vector $\mathbf{b}'$ based on (2).
  Update the table entries as follows:
  $Q(s,a^*) \leftarrow Q(s,a^*) + \Delta Q_{\mathbf{b}}(s,a^*), \forall s \in \mathcal{S}$.
  $\mathbf{b} \leftarrow \mathbf{b}'$.
end for

procedure for each user $j \in \mathcal{K}_s$ in Algorithm 1. Given a belief vector $\mathbf{b} = [b(1), \cdots, b(2^L)]$, we define the Q-value of the belief vector $\mathbf{b}$ as:

$$Q_{\mathbf{b}}(a) = \sum_{s \in \mathcal{S}} b(u(s)) Q(s,a), \quad (5)$$

and the update function as:

$$\Delta Q_{\mathbf{b}}(s,a) = \xi b(u(s)) \left[r_j(k) + \gamma \max_{a' \in \mathcal{C}} Q_{b'}(a') - Q(s,a) \right].$$

We define ξ to be the learning rate. The Q-value $Q(s,a)$ is updated after taking every action using:

$$Q(s,a) \leftarrow Q(s,a) + \Delta Q_{\mathbf{b}}(s,a). \quad (6)$$

This update is done for every state $s \in \mathcal{S}$.

V. Simulation Results

We assume that all primary channels $i \in \mathcal{C}$ have the same transition probabilities that are governed by the transition matrix:

$$\mathbb{P}_i = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}. \quad (7)$$

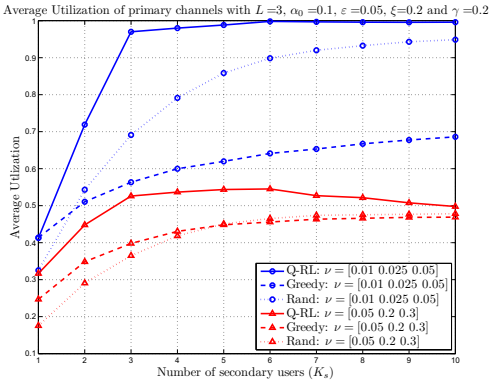

Fig. 4. Average Utilization of Primary channels for $\alpha_0 = 0.1$.

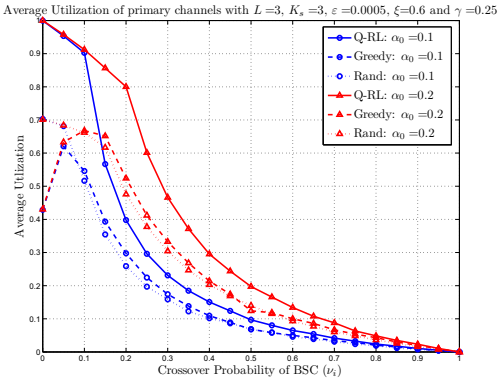

Fig. 5. Average Utilization of Primary channels for $K_s = 3$.

We define the average spectrum hole utilization as:

$$U = \frac{\sum_{j=1}^{K_s} \sum_{k=1}^{\infty} \mathcal{I}_{\{r_j(k)=1\}}}{\sum_{i=1}^{L} \sum_{k=1}^{\infty} \mathcal{I}_{\{s_i(k)=0\}}}. \tag{8}$$

The reinforcement values (rewards) are selected as follows:

1) $r_j(k) = 1$ if secondary j successfully accesses channel $a_j(k)$ at time k.
2) $r_j(k) = -0.5$ if secondary j back-off due to collision with another secondary user, and conditioned on the channel being idle.
3) $r_j(k) = 0$ if the sensed channel is busy.

In the random sensing scenario, the average number of secondary users that are sensing a given primary channel is $\mathbb{E}\{Z_i(k)\} = \frac{K_s}{L\left(1-(1-1/L)^{K_s}\right)}$, where $Z_i(k) \in \{1, \cdots, K_s\}$ is a zero-truncated binomial random variable with parameters K_s and $1/L$. Thus, in the random sensing scenario, we set the false alarm probability of each secondary user to $\alpha = 1 - (1-\alpha_0)^{1/\mathbb{E}\{Z_i(k)\}}$.

On the other hand, when applying the Q-learning algorithm, the secondary users will be evenly distributed over the channels. Therefore, $\mathbb{E}\{Z_i(k)\} = \frac{K_s}{L}$ if $K_s \geq L$, and $\mathbb{E}\{Z_i(k)\} = 1$ otherwise.

We note that $\mathbb{E}\{Z_i(k)\}$ is conditioned on the channel i being sensed (i.e. conditioned on $\{Z_i(k) \neq 0\}$).

In the following simulations, we model the sensing observations of channel $i \in \mathcal{C}$ as the output of a BSC with crossover probability ν_i, and we let $\nu = [\nu_1, \cdots, \nu_L]$. The use of a BSC permits to simplify the analysis, yet it is applicable to different channel environments since ν_i can depend on the channel fading model, the detector type, the signal and noise power, and the prior distributions of the information message. Interested readers are referred to [12]–[14] for the computation of ν_i under different channel conditions and with different detection methods.

We compare the performance of our proposed channel access/sensing mechanism to the greedy approach that was proposed in [2]. This greedy approach is equivalent to the *single-user approach* that is defined in [3] and which is applied as a non-cooperative myopic policy in multiuser OSA. In Fig.

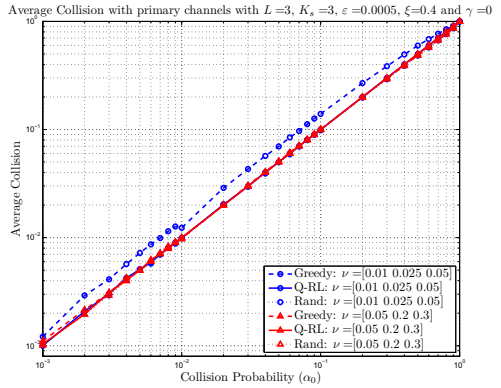

Fig. 6. Collision rates with Primary channels for $K_s = 3$.

4, we observe that RL permits to achieve high utilization of the spectrum opportunities in the primary channels. In particular, in the low-noise regime, the spectrum utilization approaches 100%. Moreover, the RL algorithm has a significant advantage over the greedy algorithm of [2] because the greedy algorithm makes most of the secondary users to sense the channel that is most likely to be idle, thus ignoring other possible spectrum opportunities and causing collisions among secondary users, as stated in [3]. This is expected because the greedy algorithm is an optimal *myopic* strategy for a single-user case and can only be a suboptimal strategy in a multiuser context. On the other hand, a simple random sensing policy that selects randomly a channel at each time instant can outperform the greedy algorithm of [2] as the number of secondary users K_s increases. That is because a random policy reduces the collisions among the secondary users, compared to the greedy policy of [2].

Next, we assume all primary channels to have the same crossover probability ν_i and we show in Fig. 5 the impact of the sensing noise on the performance of both the Q-learning and random sensing systems. We see that the performance drops at a higher rate when the crossover probability of the sensing BSC (ν_i) becomes greater than the false alarm probability α of *each* secondary user.

In Fig. 6, we analyze the collision probability that results from our designed NP detectors. Here we are controlling the collision probability with the primary channels during the time slots in which a primary channel is being sensed. Figure 6 shows the accuracy of the proposed decentralized collision probability control in maintaining the collision rate equal to the prescribed threshold α_0, by using either of the RL or the random sensing protocols that are proposed in this paper. From Fig. 6 it can be seen that these algorithms are robust against channel impairments as captured by ν_i. The efficiency of these algorithms is due to the fact that they estimate the number of secondary users that are sensing each channel, and based on this information, the channel access rule is updated so that the collision rate with primary users is maintained within the required bound. We observe also that the greedy policy violates the prescribed collision probability with primary users when the observation noise ν_i is low. However, in this case, the excess in collision probability is not very large, compared with α_0, because most of the users sense the most likely idle channel, whereas a small number of users would sense a busy channel according to the greedy approach.

VI. Conclusion

In this paper, we derived channel sensing and accessing protocols for secondary users in decentralized cognitive networks. The sensing policy is completely decentralized and is obtained by using RL. The proposed policy ensures efficient utilization of the spectrum resources since it exploits the Markovian nature of the primary channel traffic and limits the collisions among competing secondary users. Also, we have designed a secondary detector that maximizes the detection probability of the idle channels while satisfying the collision probability constraint imposed by primary users. The designed policies are characterized by their robustness and accuracy, and help to enhance the cognitive capabilities of secondary users.

Acknowledgment

This research was supported in part by the Space Vehicles Directorate of the Air Force Research Laboratory (AFRL), Kirtland AFB, Albuquerque, NM, and the National Science Foundation (NSF) under the grant CCF-0830545.

References

[1] Q. Xiao, Q. Gao, L. Xiao, S. Zhou, and J. Wang, "An optimal opportunistic spectrum access approach," in *IEEE International Conference on Communications Workshops, 2009*, Dresden, Germany, 14-18 2009, pp. 1 –5.

[2] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 589 –600, Apr. 2007.

[3] H. Liu, B. Krishnamachari, and Q. Zhao, "Cooperation and learning in multiuser opportunistic spectrum access," in *IEEE International Conference on Communications Workshops, 2008*, Beijing, China, May. 2008, pp. 487 –492.

[4] J. Unnikrishnan and V. Veeravalli, "Algorithms for dynamic spectrum access with learning for cognitive radio," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 750–760, Feb. 2010.

[5] T. Zhang, Y. Wu, K. Lang, and D. H. K. Tsang, "Optimal scheduling of cooperative spectrum sensing in cognitive radio networks," *IEEE Systems Journal*, vol. 4, no. 4, pp. 535 –549, Dec. 2010.

[6] A. Galindo-Serrano and L. Giupponi, "Distributed Q-Learning for aggregated interference control in cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 4, pp. 1823 –1834, May 2010.

[7] Y. Teng, Y. Zhang, F. Niu, C. Dai, and M. Song, "Reinforcement learning based auction algorithm for dynamic spectrum access in cognitive radio networks," in *IEEE 72nd Vehicular Technology Conference Fall (VTC '10-Fall)*, Ottawa, ON, Sep. 2010, pp. 1 –5.

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[9] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov Decision Processes," *Mathematics of Operations Research*, vol. 27, no. 4, pp. 819–840, Nov. 2002.

[10] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer, 1998.

[11] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, "Learning policies for partially observable environments: Scaling up," *Readings in agents*, pp. 495–503, 1998.

[12] F. F. Digham, M.-S. Alouini, and M. Simon, "On the energy detection of unknown signals over fading channels," in *IEEE International Conference on Communications.(ICC '03)*, vol. 5, Anchorage, AK, May 2003, pp. 3575 – 3579 vol.5.

[13] M. Bkassiny, S. K. Jayaweera, Y. Li, and K. A. Avery, "Optimal and low-complexity algorithms for dynamic spectrum access in centralized cognitive radio networks with fading channels," in *IEEE 73rd Vehicular Technology Conference (VTC Spring 2011)*, Budapest, Hungary, May 2011, [Accepted].

[14] Y. Li, S. K. Jayaweera, M. Bkassiny, and K. A. Avery, "Optimal myopic sensing and dynamic spectrum access in centralized secondary cognitive radio networks with low-complexity implementations," in *IEEE 73rd Vehicular Technology Conference. (VTC Spring 2011)*, Budapest, Hungary, May 2011, [Accepted].