Cooperative Q-Learning for Multiple Secondary Users in Dynamic Spectrum Access

Pavithra Venkatraman and Bechir Hamdaoui

School of EECS, Oregon State University
{venkatrp,hamdaoui}@eecs.oregonstate.edu

***Abstract*— In this paper, we present and evaluate learning schemes that allow multiple secondary users to locate and use spectrum opportunities effectively, thus improving efficiency of dynamic spectrum access (DSA) systems. Using simulations, we show that the proposed schemes achieve good performances in terms of throughput and fairness, and does so by interacting with and learning from the environment only, without requiring prediction models of the environment's dynamics and behaviors.**

Keywords: Machine learning; dynamic spectrum access.

I. Introduction

FCC has been observing a huge demand for radio spectrum due to the rapid growth in wireless technology. Unfortunately, the spectrum supply has not catered to this growing demand. The shortage in spectrum supply has primarily been due to the inefficient, inflexible, static nature of the existing spectrum allocation methods, and not due to the scarcity of available spectrum [1]. This fact is well supported by measurement-based studies [2, 3], which show that the average occupancy of spectrum over all frequencies is a paltry 5.2% and that the occupancy of some bands in the 30-300 MHz range is less than 1%. This measurement data confirms the availability of many spectrum opportunities along time, frequency, and space that wireless devices and networks can potentially utilize. Therefore, it is imperative to develop mechanisms that enable effective exploitation of these spectrum opportunities.

In order to meet the growing demand for spectrum resources, FCC has resorted to more flexible spectrum allotment policies and usage rights. Here, spectrum will be managed and controlled dynamically by network entities and end-user devices themselves with little to no involvement of any centralized regulatory bodies. In this regard, FCC has promoted the *dynamic spectrum access* (DSA), which increases the spectrum *efficiency* by giving the right to the unlicensed users to exploit unused licensed spectrum, but in a way that limits interference to licensed users.

Due to its potentials, DSA has resulted in numerous works ranging from protocol design and optimization [4–8] to market-oriented access and management strategies [9–12]. More recently, some research efforts have been given to the development of learning based approaches [13–15]. Generally, these reported techniques require models that predict the environment's dynamics and characteristics. However, the DSA environment has very unique characteristics that make it too difficult to derive models that can predict its behaviors accurately enough. Therefore, it is imperative to develop techniques that can achieve good performance, but without needing models that predict the environment's behaviors.

In this paper, we present two multi-agent schemes, non-cooperative and cooperative Q-learning, that improve spectrum efficiency of DSA systems through reinforcement learning. We evaluate the performance of these two proposed schemes and compare them with the random access scheme. Simulation results show that partial and fully cooperative schemes perform better than the non-cooperative and the random schemes in terms of achieved throughput and balanced traffic loads. Depending on the communication overhead due to the extra traffic in exchanging information between the cooperating users, different levels of partial cooperation can be used. Overall, the proposed learning technique does not require prior knowledge of the environment's characteristics and dynamics, yet achieves high throughput performance by learning from interaction with the environment and intelligently locating and exploiting spectrum opportunities.

The paper is organized as follows. In Section II, we present a background on DSA. In Section III, we formulate the RL framework, and present the two multi-agent RL schemes. Section IV evaluates the proposed approach, and finally, Section V concludes the paper.

II. Dynamic Spectrum Access

The spectrum has traditionally been divided by FCC into frequency bands, and assigned to licensed or primary users (PUs) who have exclusive and flexible rights to use them. PUs are also protected against interference when using their assigned bands. Due to recent findings, showing that large portions of the licensed bands are lightly used or unused at all, and in order to address the spectrum scarcity problem, FCC opens up for DSA.

The basic idea behind DSA is to allow unlicensed users, also referred to as *secondary users* (SUs), to exploit unused licensed spectrum on an instant-by-instant basis, but in a manner that limits interference to PUs so as to maintain compatibility with legacy systems. In DSA, an agent is a group of two or more SUs who want to communicate together. In order to communicate with each other, all SUs in the same group must be tuned to the same spectrum band. Throughout, agents will also be referred to as secondary

user groups (SUGs); the terms agent and SUG will then be used interchangeably to mean the same thing.

Prior to using a licensed band, SUs must first sense the band to assess whether it is vacant, and if it is, then they can switch to and use it for so long as no PUs are present. Upon the detection of the return of any PUs to their band, SUs must immediately vacate the band. DSA has great potentials for improving spectrum efficiency, but in order to enable it, SUs must be capable of *sensing*, the ability to observe and locate spectrum opportunities; *identifying*, the ability to analyze and characterize these opportunities; and *switching*, the ability to configure and tune to the best available opportunities.

III. Reinforcement Learning for DSA

Reinforcement learning (RL) is the concept of learning from past and present experience to decide what to do best in the future. That is, the learner, also referred to as *agent*, learns from experience by interacting with the environment, and uses its acquired knowledge to select the *action* that maximizes a cumulative *reward* signal. RL is well suited for systems whose behaviors are, by nature, too complex to predict, but the reward, or reinforcement, resulting from taking an action can easily be assessed or observed. For example, in DSA, albeit it may be difficult to predict which spectrum band will be available in the near future, the reward resulting from the use of a spectrum band can easily be determined. The reward can, for example, be assessed through the amount of obtained throughput, the experienced interference, the packet success rate, etc. Thus, RL techniques are a natural choice for DSA where it is difficult to precisely specify an explicit model of the environment, but it is easy to provide a reward function.

RL is typically formalized in the context of Markov Decision Processes (MDPs). An MDP represents a dynamic system, and is specified by giving a finite set of states $\mathcal{S}$, representing the possible states of the system, a set of control actions $\mathcal{A}$, a transition function δ, and a reward function r. For this work, we formulate DSA as a finite MDP, defined by its state set $\mathcal{S}$ consisting of one state s only ($\mathcal{S} = \{s\}$), the action set $\mathcal{A}$ and the reward function r described as follows.

Action set. At each time step, the agent chooses an action from the action set $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$, where m is the number of bands. The number of actions is equal to the number of spectrum bands in the system. Taking action a_i while using spectrum band b_j makes an SUG enter and use spectrum band b_i.

Reward function. The reward perceived by the agent when taking action a_i and entering spectrum band b_i is a function of the quality level the SUG receives when using the band. We assume that each band b_i has its own bandwidth capacity V_i, and when more than one SUG use a spectrum band, the bandwidth is equally divided among all the SUGs using the band. For example, if there is a total number of 3 SUGs, A, B, and C, each taking action i, j, and k respectively, then the reward of SUG A, denoted by ra_{ijk}, can be calculated as

$$ra_{ijk} = \begin{cases} V_i/3 & \text{when } i = j = k \\ V_i/2 & \text{when } i = j \neq k \text{ or } i = k \neq j \\ V_i & \text{when } i \neq j \neq k \end{cases}$$

Non-cooperative Q-learning. The goal of the agent is to learn a policy, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, for choosing the next action a_i that produces the greatest possible expected cumulative reward. A function, $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, is defined so that its value for each state-action (s, a_i) pair corresponds to the maximum discounted cumulative reward that can be achieved when starting from state s and taking action a_i. Q can be constructed recursively [16] as follows.

$$\begin{aligned} Q(s,a_i)(t+1) = \quad & Q(s,a_i)(t) + \alpha \\ & \times (E[r(s,a_i)] - Q(s,a_i)(t)) \end{aligned}$$

where $0 < \alpha < 1$ is the learning rate. When using the non-cooperative Q-learning scheme, each SUG calculates its Q table independently from other SUGs.

Action selection. The action selection mechanism plays a very important role in Q-learning. During the learning process, this selection mechanism is what enables the agent to choose its actions. We consider the ϵ-greedy exploration as the action selection mechanism, where the action corresponding to the highest Q value in that time step is chosen with a probability of $(1 - \epsilon) + \epsilon/m$, and any other action from the action set $\mathcal{A}$ is chosen with a probability of ϵ/m. The ϵ-greedy mechanism balances between exploration and exploitation.

Probability vector. Based on the ϵ-greedy exploration, we define the probability vector over the action set as follows. $X = (x_1, x_2, \ldots, x_m)$, where x_i is the probability of taking action i

$$x_i = \begin{cases} (1-\epsilon) + \epsilon/m & \text{if } Q_i \text{ is the highest value} \\ \epsilon/m & \text{otherwise} \end{cases}$$

where again m is the number of actions.

Cooperative Q-learning. Our multi-agent cooperative scheme is based on the multi-agent Q-learning approach derived in [17]. To illustrate, suppose that SUG A with probability vector X is going to cooperate with two other SUGs, B and C, with probability vectors Y and Z, respectively. The Q table entry for SUG A choosing action i can be calculated as [17]:

$$\begin{aligned} & Q(s,a_i)(t+1) = Q(s,a_i)(t) + x_i(t)\alpha \\ & \times [(\textstyle\sum_{j=1}^{j=m} y_j(t) \sum_{k=1}^{k=m} (ra_{ijk})(z_k(t))) - Q(s,a_i)(t)] \end{aligned}$$

Similarly, each SUG can compute its Q table values based on the probability vectors of the other SUGs.

IV. Evaluation

In this section, we evaluate the performance of the proposed schemes. We show the importance of cooperation in multi-agent DSA systems by comparing the per SUG average received throughput of the cooperative scheme with that of a non-cooperative scheme. Specifically, we study the effect that cooperation has on network load balancing by allowing SUGs to make better action decision, leading to more effective exploitation of bandwidth opportunities. This also ensures fairness among SUGs by making sure that all SUGs receive (approximately) equal throughput shares.

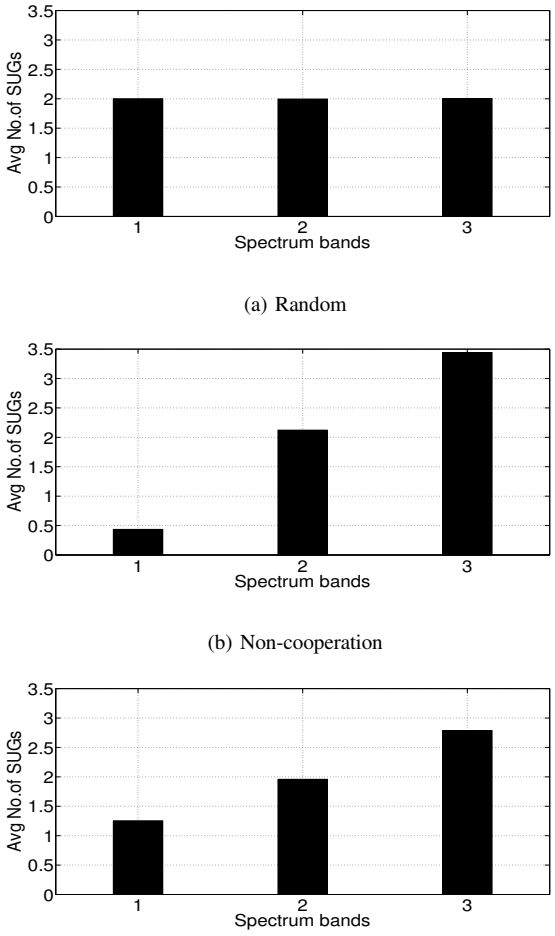
A. Simulated Access Schemes

We consider that the spectrum is divided into m non-overlapping spectrum bands with n SUGs. We mimic the presence of PUs by considering different spectrum bands with different bandwidth capacities. Let V_j denote the bandwidth capacity of band j. A spectrum band with a higher bandwidth capacity is meant to have a lower PU activity, and vice versa. We consider a time-slotted system, and assume that SUGs interact with the environment in accordance with these time slots. That is, SUGs can only enter or leave a band at the beginning and at the end of these time steps. We now summarize the three access schemes that are evaluated in this section.

Random Access Scheme. At the end of each time slot/step, an SUG using the random access scheme selects a spectrum band among the m available bands randomly, and uses it during the next time slot. If more than one SUG happen to select the same spectrum band, they share the bandwidth of the band equally.

Non-cooperative Access Scheme. In the non-cooperative access scheme, each SUG uses the non-cooperative Q-learning policy discussed in Section III to create and update its own Q table. Each SUG enters the environment and takes actions based on its own Q table without cooperating with any of the other SUGs. When two or more SUGs choose the same band during the same time step, they share its bandwidth equally. Although the SUGs are typically unaware of the other agent's actions and act independently, the effect of the other SUG's actions are reflected in the reward that the SUGs receive from the spectrum band.

Cooperative Access Scheme. In the cooperative access scheme, each SUG maintains its own Q table using the cooperative Q-learning, discussed in Section III. Here, an agent's Q table is formulated by taking into account the probabilities associated with the actions of the other SUGs with which it cooperates. In this case, at each time step, the SUG is provided with the probability vector of every other SUG with which it cooperates. The tradeoff here is between the communication overhead caused by extra traffic needed for exchanging the probability vectors among the cooperating SUGs and the performance gains due to improved action selections because of cooperation.



(a) Random

(b) Non-cooperation

(c) Full cooperation

Fig. 1. SUG distribution: $m = 3$, $n = 6$, $V_j = [5 \;\; 10 \;\; 15]$.

B. Cooperation Vs. Non-cooperation

First, we consider a DSA system with $m = 3$ spectrum bands and $n = 6$ SUGs. Bandwidth capacities are set to $V_j = [5 \quad 10 \quad 15]$. In this scenario, an ideal balanced spectrum load is reached when each of the SUGs gets a reward of 5 units, which implies that the 1^{st} band has 1 SUG, the 2^{nd} has 2 SUGs, and the 3^{rd} band has 3 SUGs. We simulate the three different access schemes for this scenario, and plot the average number of SUGs (averaged over 10000 episodes) in each of the three spectrum bands (i.e., the distribution of SUGs) in Fig. 1.

The figure shows the average number of SUGs that end up choosing each of the three spectrum bands for each of the three studied schemes. It can be observed that the fully cooperative access scheme leads to the ideal balanced system load. As explained earlier, this is because in the fully cooperative method, each SUG accounts for all the possible actions that could be taken by its counterparts when making a decision. On the other hand, when SUGs do not cooperate, they may not select the best available band, as they have no clue what other SUGs will select, leading

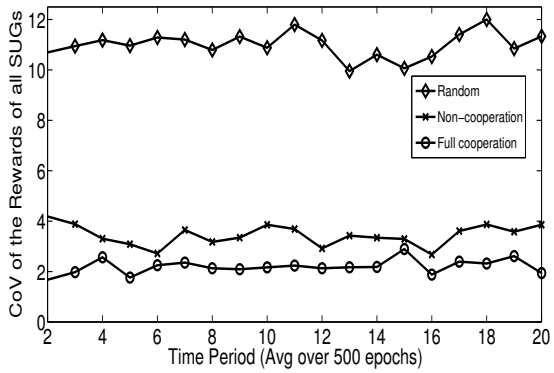


Fig. 2. Coefficient of variation of the rewards of all the SUGs at each time period: $m = 3$, $n = 6$, $V_j = [5\ \ 10\ \ 15]$.

to a lesser balanced load distribution when compared with that of the cooperative scheme. Clearly and as expected, the Random access scheme results in an equally distributed SUGs among all bands, leading to the worst load balance when compared with the other two schemes[1].

Fairness is another important metric that we also evaluate in this work. To do this, we plot in Fig. 2 the coefficient of variation (CoV) of the received rewards of all the SUGs as a function of time period (each time period corresponds to 500 epochs). Observe that the fully cooperative access scheme has the lowest CoV among the three schemes. The lower the CoV is, the closer the SUGs' received rewards are to one another, indicating a fairer access scheme. It can also be seen that the CoV of the non-cooperative access scheme is approximately twice that of the fully cooperative access scheme, and the CoV of the random access scheme is substantially higher than the other two. Therefore, cooperation improves performances not only in terms of network load balancing, but also in terms of ensuring fairness among all SUGs.

C. Impact of Degree of Cooperation

Recall that cooperation increases the performance because it allows the SUGs to make a better decision when selecting their next actions. This is because the SUGs take into account what other SUGs will select when making their action decisions. However, acquiring such information would necessitate the exchange of messages among cooperative SUGs, which clearly incurs extra overhead. Therefore, the challenge is to strike a good balance between the desire for a higher level of cooperation that enables a better action selection and the need for a lower level of cooperation so as to keep the cooperation overhead to a minimum. Cooperation overhead comes from the extra traffic needed to exchange the probability vectors and also from the computing delay/time resulting from solving the complex equations involved in updating the Q table entries of the cooperative SUGs.

[1]We want to mention that these above results do not account for the communication overhead caused by message exchange needed to share the probability vectors among cooperative SUGs.

We now study the impact of degree of cooperation on the achievable performances of a DSA system with $m = 3$ spectrum bands and $n = 12$ SUGs. The bandwidth capacities of the spectrum bands are set to $V_j = [10\ \ 20\ \ 30]$. In this scenario, an ideal balanced load is reached when each of the SUGs earns a reward of 5 units, corresponding to when the 1^{st} band houses 2 SUGs, the 2^{nd} band houses 4 SUGs, and the 3^{rd} band houses 6 SUGs. For this simulation scenario, we evaluate and compare the performances of the cooperative access scheme by considering three degrees of cooperation: 2 (i.e, each SUG cooperates with 2 other SUGs), 4 (i.e, each SUG cooperates with 4 other SUGs), and 6 (i.e, each SUG cooperates with 6 other SUGs).

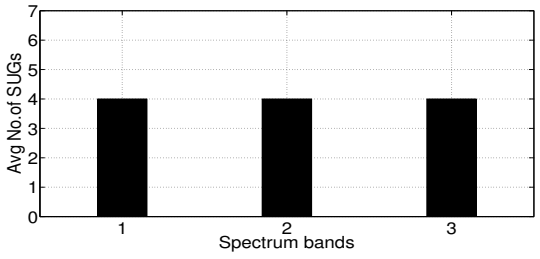
Fig. 3 shows the average number of SUGs that end up choosing each of the three spectrum bands for the random scheme, non-cooperative scheme, and cooperative access scheme with 2, 4 and 6 degree of cooperation. Note that as the degree of cooperation increases, the system load becomes more balanced. That is, the cooperative access scheme with degree of cooperation equal to 6 leads to a better balanced system load when compared with the other two degrees of cooperation.

We also study fairness achieved under each of the three cooperation degrees, and plot the CoV of the received rewards of the SUGs in Fig. 4. Observe that cooperation with a degree of 6 has the lowest CoV, followed by a degree of 4, and then followed by a degree of 2. This indicates that a higher degree of cooperation leads to a lower CoV, meaning that SUGs receive closer amounts of rewards, thus ensuring fairness among SUGs. Therefore, cooperation improves performances not only in terms of network load balancing, but also in terms of ensuring fairness among all SUGs. Note that each of the three degrees of cooperation has a lower CoV when compared with the non-cooperative and random access schemes.

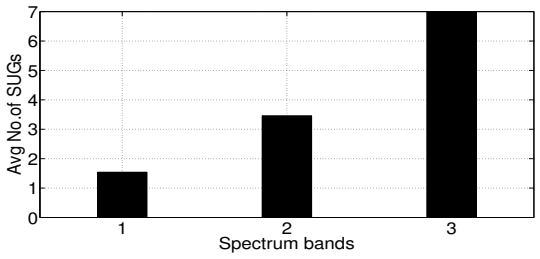
It is important to mention again that although higher degree of cooperation results in improved action selection decisions, it also incurs more communication overhead and execution times. Therefore, one must choose the degree of cooperation that balances between good selection decision and minimum overhead so as to lead to an increased overall system performance.

V. Conclusion

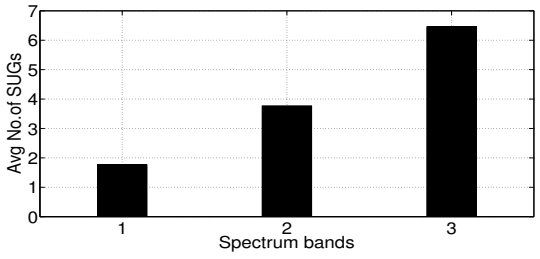
In this paper, we developed a reinforcement learning based framework for DSA system with multiple secondary users. We evaluated and compared two multi-agent Q-learning algorithms, namely the non-cooperative and the cooperative Q-learning schemes along with the random scheme. Simulation results showed that partial and fully cooperative access schemes perform better than the non-cooperative and the random access schemes in terms of achieving a higher throughput and a better balanced traffic loads. We also showed that cooperation improves performances not only in terms of network load balancing, but also in terms of ensuring fairness among all users.



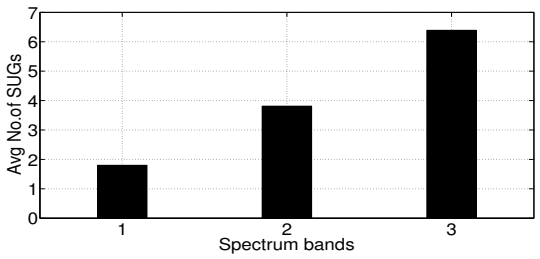
(a) Random



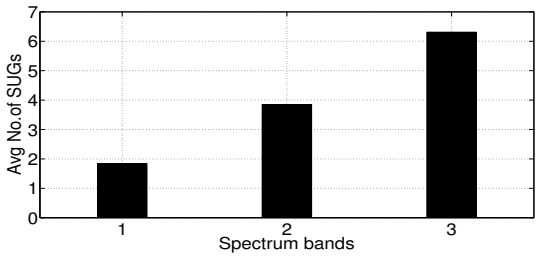
(b) Non-cooperation



(c) Cooperation with 2 SUGs



(d) Cooperation with 4 SUGs



(e) Cooperation with 6 SUGs

Fig. 3. SUG distribution: $m = 3$, $n = 12$, $V_j = [10 \;\; 20 \;\; 30]$.

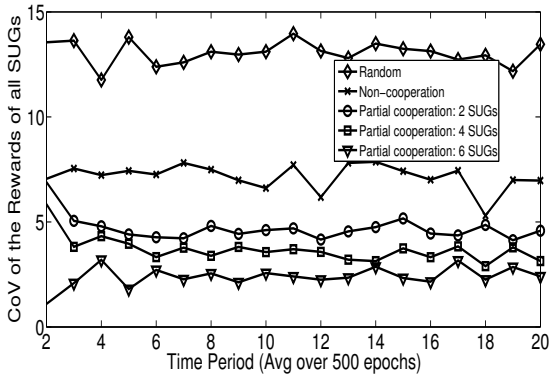


Fig. 4. Coefficient of variation of the rewards of all the SUGs at each time period: $m = 3$, $n = 12$, $V_j = [10 \;\; 20 \;\; 30]$.

REFERENCES

[1] M. Vilimpoc and M. McHenry, "Dupont circle spectrum ulitzation during peak hours," in *www.newamerica.net/files/archive/Doc_File_183_1.pdf*, 2006.

[2] M. McHenry, "Reports on spectrum occupancy measurements, shared spectrum company," in *www.sharedspectrum.com/?section=nsf_summary*.

[3] M. McHenry and D. McCloskey, "New york city spectrum occupancy measurements," *Shared Spectrum Conference*, September 2004.

[4] C.-T. Chou, S. Shankar, H. Kim, and K. G. Shin, "What and how much to gain by spectrum agility," *IEEE Journal on Selected Areas in Communications*, April 2007.

[5] S. Srinivasa and S. A. Jafar, "Cognitive radio networks: how much spectrum sharing is optimal?," in *Proceedings of IEEE GLOBECOM*, 2007.

[6] A. Ghasemi and E. S. Sousa, "Interference aggregation in spectrum-sensing cognitive wireless networks," *IEEE Journal of Selected Topics in Signal Processing*, Februray 2008.

[7] Z. Quan, S. Cui, and A. H. Sayed, "Optimal linear cooperation for spectrum sensing in cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing*, Februray 2008.

[8] H. Su and X. Zhang, "Cross-layer based oppotrunitic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE Journal on Selected Areas in Communications*, January 2008.

[9] Z. Ji and K. J. R. Liu, "Multi-stage pricing game for collusion-resistant dynamic spectrum allocation," *IEEE Journal on Selected Areas in Communications*, January 2008.

[10] S. Delaere and P. Ballon, "Flexible spectrum management and the need for controlling entities for reconfigurable wireless systems," in *Proceedings of IEEE DySPAN*, 2007.

[11] Y. T. Hou, Y. Shi, and H. D. Sherali, "Spectrum sharing for multi-hop networking with cognitive radios," *IEEE Journal on Selected Areas in Communications*, January 2008.

[12] S. Yarkan and H. Arslan, "Exploiting location awareness toward improved wireless system design in cognitive radio," *IEEE Communications Magazine*, January 2008.

[13] Z. Han, C. Pandana, and K. J. R. Liu, "Distributive opportunistic spectrum access for cognitive radio using correlated equilibrium and no-regret learning," in *Proceedings of IEEE WCNC*, 2007.

[14] H. Kim and K. G. Shin, "Efficient discovery of spectrum oppotunities with MAC-layer sensing in cognitive radio networks," *IEEE Transactions on Mobile Computing*, May 2008.

[15] U. Berthold, M. Van Der Schaar, and F. K. Jondral, "Detection of spectral resources in cognitive radios using reinforcement learning," in *Proceedings of IEEE DySPAN*, 2008, pp. 1–5.

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning*, The MIT Press, 1998.

[17] E.R. Gomes and R. Kowalczyk, "Dynamic analysis of multiagent Q-learning with ϵ-greedy exploration," in *Proceedings of the 26th International Conference on Machine Learning*, 2009, pp. 369–376.