A Survey on Machine-Learning Techniques in Cognitive Radios[1]

Mario Bkassiny, *Student Member, IEEE*, Yang Li, *Student Member, IEEE* and Sudharman K. Jayaweera, *Senior Member, IEEE*

Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM, USA

Email: {bkassiny, yangli, jayaweera}@ece.unm.edu

Abstract

In this survey paper, we characterize the learning problem in cognitive radios and state the importance of artificial intelligence in achieving real cognitive systems. We review various learning approaches that have been proposed for cognitive radios classifying them under supervised and unsupervised learning paradigms. Unsupervised learning is presented as an autonomous learning procedure that is suitable for unknown RF environments, whereas supervised learning methods can be used to exploit prior information available to cognitive radios during the learning process. We describe some challenging learning problems that arise in cognitive radio networks, in particular in non-Markovian environments, and present their possible solution methods. Finally, we present some generic cognitive radio problems and show suitable machine learning approaches for learning in these contexts.

Index Terms

Cogntive radio, machine learning, artificial intelligence, unsupervised learning, supervised learning.

This work was supported in part by the National Science foundation (NSF) under the grant CCF-0830545.

I. INTRODUCTION

Since its inception, the term cognitive radio has been used to refer to radio devices that are capable of learning and adapting to their environment [1], [2]. A key aspect of any cognitive radio is the ability for self-programming [3]. In [4], Haykin envisioned cognitive radios to be *brain-empowered* wireless devices that are specifically aimed at improving the utilization of the electromagnetic spectrum. According to Haykin, a cognitive radio is assumed to use the methodology of *understanding-by-building* and is aimed to achieve two primary objectives, which are permanent reliable communications and efficient utilization of the spectrum resources [4]. With this interpretation of cognitive radios, a new era of cognitive radios began, focusing on dynamic spectrum sharing (DSS) techniques to improve the spectrum utilization [4]–[8]. This led to research on various aspects of communications and signal processing required for DSA networks [4], [9]–[24]. These included underlay, overlay and interweave paradigms for spectrum co-existence by secondary cognitive radios in licensed spectrum bands [8].

To perform its cognitive tasks, a cognitive radio should be aware of its RF environment. It should sense its surrounding environment and identify all types of RF activities. Thus, spectrum sensing was identified as a major ingredient in cognitive radios [4]. Many sensing techniques have been proposed over the last decade [25], based on matched filter, energy detection, cyclostationary detection, wavelet detection and covariance detection [18], [26]–[31]. In addition, cooperative spectrum sensing was proposed as a means of improving the sensing accuracy by addressing the hidden terminal problems inherent in wireless networks in [21], [22], [25], [27], [32]–[34]. In recent years, cooperative cognitive radios have also been considered in literature as in [35]–[38]. Recent surveys on cognitive radios can be found in [26], [39]–[41].

In addition to being aware of its environment, and in order to be really *cognitive*, a cognitive radio should be equipped with the abilities of learning and reasoning [1], [2]. These capabilities can be achieved through a cognitive engine which was identified as the core of a cognitive radio [42]–[47], following the pioneering vision of [2]. A cognitive engine coordinates the actions of the cognitive radio by applying machine learning algorithms. However, only in recent years there is a growing interest in applying machine learning algorithms to cognitive radios [48], [49], and these algorithms can be categorized under either supervised or unsupervised learning.

The authors in [44], [50], [51] have considered supervised learning based on neural networks

and support vector machines for cognitive radio applications. Unsupervised learning, such as reinforcement learning (RL), has been considered in [52], [53] for DSS applications. The distributed Q-learning algorithm has been shown to be effective in a certain cognitive radio application in [54]. For example, in [55], cognitive radios used the Q-learning to improve detection and classification performance of primary signals. Other applications of RL to cognitive radios can be found, for example, in [56]–[59]. Recent work in [60] introduces novel approaches to improve the efficiency of RL by adopting a weight-driven exploration. On the other hand, an unsupervised Bayesian non-parametric learning procedure based on the Dirichlet process was proposed in [61]. A robust signal classification algorithm was also proposed in [62], based on unsupervised learning.

Although the RL algorithms (such as Q-learning) may provide a suitable framework for autonomous unsupervised learning, their performance in partially observable, non-Markovian and multi-agent systems[1] can be unsatisfactory [64]–[67]. Other types of learning mechanisms such as evolutionary learning [65], [68], learning by imitation, learning by instruction [69] and policy-gradient methods [66], [67] have been shown to outperform RL on certain problems under such conditions. For example, the policy-gradient approach has been shown to be more efficient in partially observable environments since it searches directly for optimal policies in the policy space, as we shall discuss throughout this paper [66], [67].

Similarly, learning in multi-agent environments has been considered in recent years, especially when designing learning policies for cognitive radio networks (CRN's). For example, [70] compared a cognitive network to a human society that exhibits both individual and group behaviors, and a strategic learning framework for cognitive networks was proposed in [71]. An evolutionary game framework was proposed in [72] to provide adaptive learning to cognitive users during their strategic interactions. By taking into consideration the distributed nature of CRN's and the interactions among the cognitive radios, optimal learning methods can be obtained based on cooperative schemes, which helps avoid the selfish behaviors of individual nodes in a CRN.

[1]A multi-agent system can be defined as a group of autonomous, interacting entities sharing a common environment, which they perceive with sensors and upon which they act with actuators [63].

A. Purpose of this paper

This paper discusses the role of learning in cognitive radios and emphasizes how crucial the autonomous learning ability in realizing a real cognitive radio device. We present a survey of the state-of-the-art achievements in applying machine learning techniques to cognitive radios. We will focus on the special challenges that are encountered in applying machine learning techniques to cognitive radios. In particular, we describe different types of learning paradigms that have been proposed in the literature as well as those that might be reasonably applied to cognitive radios in the future. The advantages and limitations of these techniques are discussed to identify perhaps the most suitable learning methods in a particular context or in learning a particular aspect.

B. Organization of the paper

The remainder of this survey paper is organized as follows: Section II defines the learning problem in cognitive radios and presents the different learning paradigms. Sections III and IV present the unsupervised and supervised learning techniques, respectively. In Section V, we describe the learning problem for centralized and decentralized cognitive radio systems. Section VI presents the learning challenges in non-Markovian environments and we conclude in Section VII.

II. Need of learning in cognitive radios

A. Definition of the learning problem

Learning is defined as *the modification of behavior through practice, training, or experience* [73]. According to [74], the learning ability is an indispensable component of an intelligent behavior. A practical definition for the term *learning* was given in [74] to be *the ability of creating knowledge from the information acquired about the environment and the internal states*. Based on this definition, learning is related to the ability of synthesizing the acquired knowledge in order to improve the future behavior of the learning agent. This makes knowledge a fundamental component of the learning process and relates to the term *cognition* which is defined as *the act or process of knowing* or *perception* [73]. In Fig. 1, we depict the relations among intelligence,

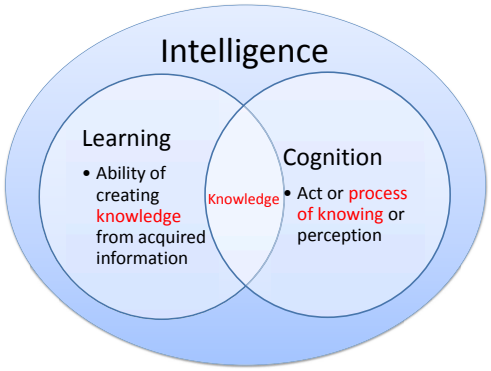

Fig. 1. Learning is a fundamental component of intelligence. It shares a common feature with cognition, which is knowledge.

learning and cognition, and illustrate the concept of knowledge as a common feature of both learning and cognition.

Thus, learning is indispensable to any cognitive system, and must be at the foundation of cognitive radios. By using its learning capability, an agent can classify, organize, synthesize and generalize information obtained from its sensors [74]. However, learning is not the unique feature of an intelligent device which should also be aware of its surrounding environment and must be capable of reasoning. Hence, the three main constituents of intelligence can be identified as: 1) perception, 2) learning and 3) reasoning [74].

We discuss, in the followings, how the above three constituents of intelligence can be realized through cognitive radios. First, *perception* can be achieved through the sensing measurements of the spectrum. This allows the cognitive radio to identify ongoing RF activities in its surrounding environment. After acquiring the sensing observations, the cognitive radio tries to *learn* from them in order to classify and organize the observations into suitable categories. This can be achieved through different types of learning algorithms that we discuss below in this survey. Finally, the *reasoning* ability allows the cognitive radio to use the knowledge acquired through learning to achieve its objectives. These characteristics were initially specified by Mitola in defining the so-called *cognition cycle* [1]. We illustrate in Fig. 2 an example of a simplified cognition cycle that was proposed in [75] for designing autonomous cognitive radios, referred to as *Radiobots*.

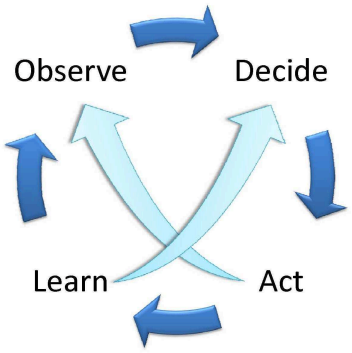

Fig. 2. The cognition cycle of an autonomous cognitive radio, referred to as the Radiobot [75].

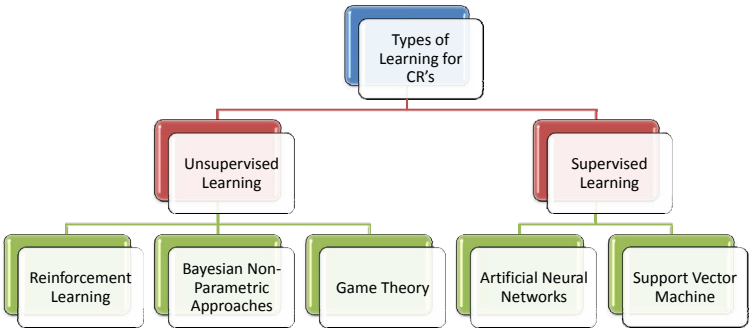

Fig. 3. Supervised and unsupervised learning approaches for cognitive radios.

B. Unique characteristics of cognitive radio learning problems

Although the term *cognitive radio* has been interpreted differently in different research communities [75], perhaps the most widely accepted definition is as radio that can sense and adapt to its environment [48]. The term *cognitive* implies *awareness*, *perception*, *reasoning* and *judgement*. However, as we have pointed out earlier, in order to make cognitive radios truly intelligent, the learning ability must also be present [74]. Learning implies that the current actions should be based on past and current observations of the environment [76]. This should not be confused with reasoning which consists of observing only the current state of the environment and making the decisions ignoring the past information [48]. Thus, the history plays a major role in the learning process of cognitive radios and forms a fundamental factor in optimizing the cognitive radio objectives.

Several learning problems are specific to cognitive radio applications due to the nature of the cognitive radios and the operating RF environments. First of all, due to the noisy observations and sensing errors, cognitive radios usually obtain partial observations of their state variables. The learning problem is thus equivalent to a learning process in partially observable environments and must be addressed accordingly.

Another problem that should be considered in cognitive radio learning problems is the multi-agent learning process. This situation arises, in particular, in CRN's in which multiple agents try to learn and optimize their behaviors simultaneously. Furthermore, the desired learning policy may be based on either cooperative or non-cooperative schemes and each cognitive radio might have either full or partial knowledge of the actions of the other cognitive users in the network. In the case of partial observability, a cognitive radio might apply special learning algorithms to estimate the actions of the other nodes in the network before selecting its appropriate actions, as in [64].

Finally, autonomous learning methods are desired in order to enable cognitive radios to learn in unknown RF environment. This is because, in contrast with licensed wireless users, a cognitive radio is supposed to operate in any available spectrum band, at any time and in any location. Thus, a cognitive radio may not have any prior knowledge of the operating RF environment such as the noise or interference levels, noise distribution or user traffics. Instead, it should be able to apply autonomous learning algorithms that reveal the underlying nature of the environment and its components. This makes the unsupervised learning a perfect candidate for the learning problem in cognitive radio applications, as we shall point out throughout this survey paper.

To sum up, we have identified three main characteristics that need to be considered when designing efficient learning algorithms for cognitive radios:

1) Learning in partially observable environments.
2) Multi-agent learning in distributed CRN's.
3) Autonomous learning in unknown RF environments.

A cognitive radio design that embeds the above capabilities will be able to operate efficiently and optimally in any RF environment.

C. Types of learning in cognitive radios

In this survey paper, we classify the learning algorithms for cognitive radios under two main categories: Supervised and unsupervised learning, as shown in Fig. 3. Unsupervised learning is particularly applicable for cognitive radios operating in alien environments. In this case, autonomous unsupervised learning algorithms permit exploring the environment characteristics and self-adapting actions accordingly without having any prior knowledge. However, if the cognitive radio has prior information about the environment, it might exploit this knowledge by using supervised learning techniques. For example, if certain signal waveform characteristics are known to the cognitive radio prior to its operation, training algorithms would help cognitive radios to better detect those signals. We present, in the following major learning algorithms under each of these categories, and describe some of their applications in cognitive radios.

In [69], the two categories of supervised and unsupervised learning are defined as learning by *instruction* and learning by *reinforcement*, respectively. A third learning regime is defined as the learning by *imitation* in which an agent learns by observing the actions of similar agents [69]. In [69], it was shown that the performance of a learning agent (learner) is influenced by its learning regime and its operating environment. Thus, for a cognitive radio to learn efficiently, it must adopt the best learning regime, whether it is learning by *imitation*, by *reinforcement* or by *instruction* [69]. Of course, some learning regimes may not be applicable under certain circumstances. For example, in the absence of an instructor, the cognitive radio may not be able to learn by instruction and may have to resort to learning by reinforcement. An effective cognitive radio architecture is the one that can switch between different learning regimes depending on its requirements, the available information and the environment characteristics.

III. Unsupervised Learning

A. Reinforcement learning (RL)

Reinforcement learning is a technique that permits an agent to modify its behavior by interacting with its environment. This type of learning can be used by agents to learn autonomously without supervision. In this case, the only source of knowledge is the feedback an agent receives from its environment after executing an action. Two main features characterize the reinforcement learning: *trial-and-error* and *delayed reward*. By *trial-and-error* it is assumed that an agent does

not have any prior knowledge about the environment, and it executes some actions blindly in order to *explore* the environment. The *delayed reward* is the feedback signal that an agent receives from the environment after executing each action. These rewards can be positive or negative quantities, telling *how good or bad* an action is. The agent's objective is to maximize these rewards by *exploiting* the system.

Reinforcement learning is distinguished from *supervised learning* by not having a supervisor to tell whether an action is correct or wrong. Therefore, the learning agent only relies on its interactions with the environment and tries to learn on its own. This makes the reinforcement learning a basic algorithm for autonomous learning.

A key concept in reinforcement learning is that the agent should observe the reward for each action in each situation. By repetition, the agent attempts to learn to favor the actions that lead to positive rewards, and avoids the actions that lead to negative rewards. Moreover, a learning agent can use the reinforcement learning to choose the actions that permit avoiding certain bad situations. After several repetitions, the agent acquires an optimal policy and adapts its actions and behavior to the environment.

The theory of reinforcement learning has evolved along three main threads. The first thread is the learning by *trial and error* which has its roots in the psychology of animals. This approach goes back to 1898 and has led to the revival of the reinforcement learning in the early 1980's [77]. For example, in his analysis of animal behavior, Thorndike observed that animals tend to reselect actions that are followed by good outcomes, and they try to avoid the actions that lead to bad outcomes [78].

The second thread originates from the problem of optimal control and its dynamic programming-based solution. One approach to this problem was developed in the mid 1950's by Bellman and others by extending the theory of Hamilton and Jacobi. The dynamic programming (DP) is found to be the most efficient solution to the optimal control problem. However it suffers from what Bellman called "*the curse of dimensionality*" because the complexity of DP increases exponentially with the number of state variables. Also, it requires complete knowledge of the system.

The third thread that led to the reinforcement learning is the *temporal difference* concept which was first applied to learning problems by Samuel [79]. This idea consists of updating

an evaluation function about the environment in order to improve the total reward. The three threads that constitute the reinforcement learning were joined together in 1989 by Watkins when he developed the Q-learning algorithm [80], [81].

It should be noted that many studies used the term *reinforcement learning* also to refer to *supervised learning*, and this distinction should be made clear since reinforcement learning is defined when an agent tries to learn from its *own* experience by evaluating the feedback signals that it receives after each action [82]. These feedback signals (reinforcement values) do not tell if an action is correct or wrong. They only reveal how good or bad the action is. On the other hand, supervised learning applies to the cases when a clear answer is available to the agent on whether its action was correct or wrong. Usually, supervised learning consists of training the agent for a certain duration by assigning the actions and revealing the correct answers.

The applications of reinforcement learning extend to a wide range of domains, such as robotics, distributed control, telecommunications, economics, data mining and active gesture recognition [82]–[84]. Recently, reinforcement learning was applied to the telecommunication field and especially to cognitive radio. RL is found to be effective in cognitive radio context because it presents an autonomous technique to make an agent to learn and adapt to its environment, which is a key feature of a cognitive radio. In particular, a cognitive radio can interact with its RF environment and can try to learn by observing the consequences of its actions. This method is useful if the cognitive radio does not have knowledge about certain parameters of its environment, and thus, tries to learn an optimal policy that leads to the best performance in a given RF environment.

A reinforcement learning-based cognition cycle for cognitive radios was defined in [53], as illustrated in Fig. 4. It shows the interactions between the cognitive radio and its RF environment. Based on this process, the learning agent receives an observation o_t of the state variable s_t at time instant t. The observation is accompanied with a delayed reward r_t representing the reward resulting from taking action a_{t-1} in state s_{t-1}. The learning agent uses the observation o_t and the reward r_t to compute the action a_t that should be taken at time t. Again, the action a_t results in a state transition from s_t to s_{t+1} and a delayed reward r_{t+1}. It should be noted that here the learning agent is not passive and does not only observe the outcomes from the environment, but can also affect the state of the system via its actions such that it might be able to drive the

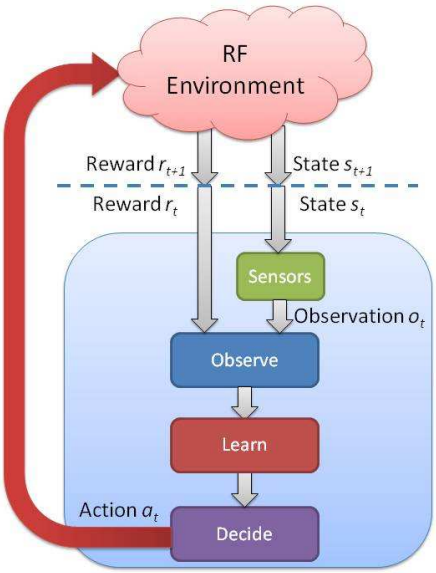

Fig. 4. Reinforcement learning cycle.

environment to a desired state that brings the highest reward to the agent.

In order to apply the above described RL procedure to cognitive radios, the learning problem can be formulated in several ways. As a specific example, we consider the model in [85] which assumes a primary and a secondary (cognitive) user that coexist in the same frequency band. The primary user (PU) is assumed to use a combination of time-division and frequency-division multiple access (TDMA, FDMA) schemes, which might result in spectral or temporal holes. Spectrum holes are the unused spectrum opportunities. They consist of frequency bands and/or time slots that are not used by any radio transmission at a particular time and at a particular location [8], [10]. These spectrum holes characterize the under-utilization of the frequency spectrum and form perfect candidates for secondary use in opportunistic spectrum access [24], [86], [87]. In the model proposed in [85], the SU is assumed to adopt an OFDM scheme such that each subcarrier can be switched on and off individually, depending on the PU allocation. It is assumed that the primary channel activity follows a Markov chain and the SU's try to access those channel resources whenever they are idle. Instead of using the dynamic programming approach to solve the dynamic spectrum access problem based on the Markov decision process (MDP) framework [88], the authors in [85] use the RL algorithm to obtain the optimal solution

of their MDP formulation. Similarly to the dynamic programming approach, the RL algorithm leads to optimal solution to the MDP problem, yet at a lower complexity [82]. Moreover, the RL algorithm does not require complete knowledge about the system model and can be applied as an online learning algorithm, as described in [85].

The authors in [85] propose two problem formulations for the dynamic spectrum access problem: In the first formulation, a simplistic model is assumed which considers that the switching cost between frequency bands is negligible. The resulting model is similar to the n-armed bandit problem and is solved by using the softmax exploration approach [82]. The softmax approach generates stochastic policies in which an action is selected with a probability proportional to the value of that action. In the second formulation, the authors assumed a certain switching cost among channels and introduced a state $s \in \{1, \cdots, N_{fb}\}$ which denotes the current sub-band of the SU, where N_{fb} is the total number of available frequency bands. The problem is thus modeled as an MDP characterized by the following parameters:

- A finite set $\mathcal{S}$ of states for the agent (i.e. SU).
- A finite set $\mathcal{A}$ of actions that are available to the agent. In particular, in each state $s \in \mathcal{S}$, a subset $\mathcal{A}_s \subseteq \mathcal{A}$ might be available.
- A state transition probability $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ defines the transition probability $p(s'|s, a)$ from state $s \in \mathcal{S}$ to $s' \in \mathcal{S}$, after performing the action $a \in \mathcal{A}$.
- A reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ defining a reward $r(s, a)$ that the agent receives when performing action $a \in \mathcal{A}$, while in state $s \in \mathcal{S}$.

The agent observes the current state s and chooses an action a for the next stage. This is done according to the stochastic policy $\pi : \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$, where $\pi(a, s)$ defines the probability of taking action a when the agent is in state s. An optimum policy maximizes the total expected rewards (i.e. the return function), which is usually discounted by a discount factor $\gamma \in [0, 1)$ in case of an infinite time horizon. Thus, the objective is to find the optimal policy π that maximizes the return function $R(t)$:

$$R(t) = \mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k}(s_{t+k}, a_{t+k})\right\} , \tag{1}$$

where r_t, s_t and a_t are, respectively, the reward, state and actions at time $t \in \mathbb{Z}$.

In [85], the state $s \in \{1, \cdots, N_{fb}\}$ denotes the current frequency band that the SU is using for transmitting. According to the assumed model, the set of available actions in state s is $\mathcal{A}_s = \{a_1, a_{2\tilde{s}}, a_{3\tilde{s}}\}$, where $\tilde{s} = \mathcal{S} \backslash s$ and

- a_1: perform a cycle of detection and transmission in the current frequency band s.
- $a_{2\tilde{s}}$: perform a detection phase in frequency band $\tilde{s}$ (out-of-band detection).
- $a_{3\tilde{s}}$: switch the SU system to frequency band $\tilde{s}$.

According to the proposed model in [85], a state transition occurs only if the action $a_{3\tilde{s}}$ is selected. In addition, the reward function $r(a, s)$ is defined as follows:

$$r(a,s) = \begin{cases} u_1(s) & \text{for} \quad a = a_1 \\ u_2 & \text{for} \quad a = a_{2\tilde{s}} \\ u_3 & \text{for} \quad a = a_{3\tilde{s}} \end{cases}, \tag{2}$$

where $u_1(s)$ is the number of radio resource goods (e.g. bits transmitted) that have been transmitted in the current step, while staying in the current frequency band. u_2 is the reward/cost for performing a detection in a different frequency band. u_3 is the cost of switching to another frequency band, which can represent a negative reward (i.e. a cost) associated with any transmission delay that is incurred due to switching (e.g. control data exchange overhead). Note that, in this setup, both u_2 and u_3 are independent of the current state s.

Several solutions were proposed for the MDP problem by following, for example, the *value-iteration* or the *linear programming* algorithms of [88]. The value-iteration algorithm is an iterative algorithm that is based on the Bellman's principle of optimality [88], [89]. This algorithm estimates the value function V^t at a given stage t in function of the value function V^{t-1} at the previous stage $t-1$, as follows:

$$V^t(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p\left(s'|s,a\right) V^{t-1}(s') \right\} . \tag{3}$$

Puterman showed that the value-iteration algorithm guarantees that the estimated value function is ϵ-optimal over an infinite horizon [88], [89].

On the other hand, the MDP can be solved by following the linear programming approach of

[88] as follows:

$$\min \sum_{s\in\mathcal{S}} V(s)$$
$$\text{s.t. } 0 \geq r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)V(s') - V(s); \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

The above solutions lead to optimal and near-optimal solutions to the MDP, but require knowledge of the transition probabilities of the MDP. The RL algorithm, on the other hand, finds the optimal solution to the MDP, yet without knowledge of the transition probabilities [82]. This makes the RL algorithm a desired approach for problems with partial knowledge of the MDP model, as in [85]. The RL algorithm in [85] is based on the temporal-difference (TD) learning approach which updates the value of each state $V(s)$, after each interaction, as follows:

$$V(s_t) \leftarrow V(s_t) + \beta \left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \right] \ , \tag{4}$$

where β is a positive step-size parameter, called the *learning rate*. Hence, after observing the reward r_{t+1} at time $t+1$, and knowing the old state s_t and the new state s_{t+1}, the agent updates $V(s_t)$ according to the rule described above. The obtained value function is thus used to update the policy π as follows:

$$\pi_t(s,a) = P\{a_t = a | s_t = s\} = \frac{e^{p(s,a)}}{\sum_b e^{p(s,b)}} \ , \tag{5}$$

where $p(s,a)$ are updated differently, depending on the type of action. Action a_1 is updated using a common update rule:

$$p(s,a_1) \leftarrow p(s,a_1) + \beta_1 \delta_t \ , \tag{6}$$

where β_1 is a positive step-size and $\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$. The above update rule reflects the amount of transmitted data when the system is in state s. The update rule of $p(s,a_{2\tilde{s}})$ is defined such that it favors the exploration of less reliable states. The update rule is defined as follows [85]:

$$p(s,a_{2\tilde{s}}) = \left(1 - \zeta(s)\right) V(s) \ , \tag{7}$$

where $\zeta(s) \in [0,1]$ is a reliability value. Finally, $p(s, a_{3\tilde{s}})$ is updated as:

$$p(s, a_{3\tilde{s}}) = \zeta(s)\,(V(\tilde{s}) - \frac{N_{fb}}{2}) + \frac{N_{fb}}{2}\ , \tag{8}$$

where N_{fb} is the number of frequency bands. Thus, this rule favors the switching to frequency bands having large number of resources and high reliability values $\zeta(s)$.

The TD algorithm is a combination of Monte Carlo and Dynamic Programming methods [82]. Like Monte Carlo, it can learn directly from experience, without a complete model of the system. Like Dynamic Programming, TD updates estimates based on other learned estimates without waiting for the final outcome [82]. In particular, a simple Monte Carlo algorithm for estimating the value of a state s_t can be defined as:

$$V(s_t) \leftarrow V(s_t) + \beta\,[R_t - V(s_t)]\ , \tag{9}$$

where β is a learning parameter, $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ is the return function at time t and γ is a discount factor. Obviously, the Monte Carlo method has to wait for the end of the episode (i.e. end of the time horizon) to update $V(s_t)$. On the other hand, the TD method updates $V(s_t)$ after the next time step as follows:

$$V(s_t) \leftarrow V(s_t) + \beta\,[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \tag{10}$$

The TD method has an advantage over the dynamic programming method since it does not require a model of the environment. Also, the TD method is more suitable for online learning, compared to the Monte Carlo method.

Moreover, it has been shown [82] that the value function in (10) converges in the mean to V^{π} for any fixed policy π if β is sufficiently small, and it converges with probability 1 if β satisfies the stochastic approximation conditions below:

$$\sum_{k=1}^{\infty} \beta_k(a) = \infty \text{ and } \sum_{k=1}^{\infty} \beta_k^2(a) < \infty, \tag{11}$$

where $\beta_k(a)$ is the step-size parameter used after executing action a for the k-th time.

Another reinforcement learning algorithm that has been applied to cognitive radios was based

on the Q-learning [54], [55], [90], [91]. This algorithm estimates the Q-values, $Q(s,a)$ of the joint state-action pairs (s,a). This function represents the return function of action a when the system is in state s and is defined as:

$$Q(s,a) = \mathbb{E}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = s, a_t = a\right\} . \tag{12}$$

The Q-learning algorithm is one of the most important TD methods that was developed by Watkins in 1989 [92]. The *one-step* Q-learning is defined as follows:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1},a) - Q(s_t,a_t)\right]. \tag{13}$$

The update function (13) directly approximates the optimal Q^* value. However, it is required that all state-action pairs need to be continuously updated in order to guarantee *correct convergence*. This can be achieved by applying an ε-greedy policy that ensures that all state-action pairs are updated with a non-zero probability, thus leading to an optimal policy [82].

In [54], the authors applied the Q-learning to derive the interference control in a cognitive network. The problem setup is illustrated in Fig. 5 in which multiple IEEE 802.22 WRAN cells are deployed around a Digital TV (DTV) cell such that the aggregated interference caused by the secondary networks to the DTV network is below a certain threshold. In this scenario, the cognitive radio (agents) constitutes a distributed network and each radio tries to determine how much power it can transmit so that the aggregated interference on the primary receivers does not exceed a certain threshold level.

In this system, the secondary base stations form the learning agents that are responsible for identifying the current environment state, selecting the action based on the Q-learning methodology and executing it. The state of the i-th WRAN network at time t consists of three components and is defined as [54]:

$$s_t^i = \{I_t^i, d_t^i, p_t^i\} , \tag{14}$$

where I_t^i is a binary indicator specifying whether the secondary network generates interference to the primary network above or below the specified threshold, d_t^i denotes an estimate of the distance between the secondary user and the interference contour, and p_t^i denotes the current

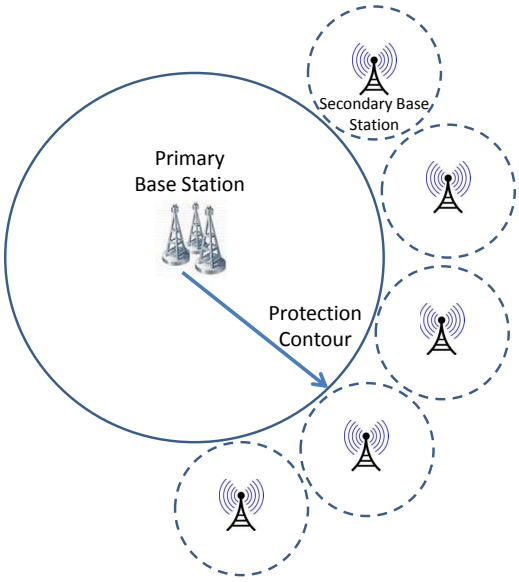

Fig. 5. System model of [54] which is formed of a Digital TV (DTV) cell and multiple WRAN cells.

power at which the secondary user i is transmitting. In the case of full state observability, the secondary user has complete knowledge of the state environment. However, in the partially observable environment, the agent i has a partial information of the actual state and uses a belief vector to represent the probability distribution of the state values. In this case, the randomness in s_t^i is only related to the parameter I_t^i which is characterized by two elements $\mathcal{B} = \{b(1), b(2)\}$, i.e. the values of the probability mass function of I_t^i.

The set of possible actions is the set P of power levels that the secondary base station can assign to the i-th user. The cost c_t^i denotes the immediate reward incurred due to the assignment of action a in state s and is defined as:

$$c = \left(SINR_t^i - SINR_{Th}\right)^2 \ , \tag{15}$$

where $SINR_t^i$ is the instantaneous SINR in the control point of WRAN cell i.

By applying the Q-learning algorithm, the results in [54] showed that it can control the interference to the primary receivers, even in the case of partial state observability.

In addition to the above system models in [54], [85] describing two different applications of RL to cognitive radios, there has been many other research works that applied RL to cognitive

radios. The popularity of RL is due to its simplicity, efficiency and perhaps, more importantly, the ability to learn autonomously, which makes it a perfect candidate for learning methods in unknown RF environments. For example, the authors in [86] used the multi-armed bandit problem as a reinforcement learning method to enhance the performance of SU's in dynamic environments, while providing a semi-dynamic parameter tuning scheme to achieve an online update of the multi-armed bandit parameters. The choice of the multi-armed bandit model is to balance simultaneously between 1) exploring the external environment and 2) exploiting the past acquired knowledge to decide which channel to access in the opportunistic spectrum access setup [86]. The authors in [55] proposed an RL framework based on Q-learning to identify the presence of primary signals and to access the primary channels whenever they are found to be idle. In particular, the proposed Q-learning algorithm in [55] identifies previously known primary signals and learns to detect the signals which otherwise could not be detected, and helps for efficient utilization of spectrum. The authors in [93] used the RL for routing in multi-hop cognitive radio networks. The proposed learning technique was based on the Q-learning and it permits learning the good routes efficiently.

The authors in [94] implemented a cognition cycle (CC) based on the RL for a cognitive secondary transmitter and a cognitive secondary receiver. The objective was to maximize the data throughput between the cognitive transmitter and receiver and minimize the transmission delay while avoiding the primary traffic. The authors in [94] analyzed the performance of the proposed method and justified that RL is a promising tool to implement the CC. The authors in [94] also investigated the effects of changes on RL parameters on network performance.

A channel selection scheme was proposed in [90] for multi-user and multi-channel cognitive radio systems. In this paper, the SU's avoid the negotiation overhead by applying a multi-agent RL (MARL) algorithm. As opposed to single-agent RL (or SARL), MARL refers to the RL algorithms implemented on multiple agents in a multi-agent system introduced at the beginning of Section I. A comprehensive survey of MARL is provided in [63] with detailed discussion on the benefits and challenges of MARL. As discussed in [63], including the curse of dimensionality and the exploration-exploitation tradeoff, several common challenges in MARL are: 1) the difficulty of specifying a learning goal, 2) the nonstationarity of the learning problem, and 3) the need for coordination. The proof of convergence of the proposed algorithm in [90] was

also provided via similarity between the Q-learning and Robinson-Monro algorithm[2] [96]. In [59], a machine-learning technique was proposed to ensure effective opportunistic spectrum access (OSA) in cognitive radio networks. The model in [59] uses RL to learn by interacting with the environment. Recognizing the importance of the efficiency of a RL process for cognitive radios and the balancing between exploration and exploitation in RL, two novel exploration schemes were proposed in [60]. A first pre-partitioning exploration scheme that randomly partitions the action space to ensure faster exploration was presented, followed by a second weight-driven exploration scheme in which the action selection is influenced by the knowledge gained during exploration. In order to provide a measure of how efficient the learning process is, the authors in [60] defined the learning efficiency as

$$\text{Learning efficiency} = \frac{\text{Useful learning cost}}{\text{Total learning cost}}, \tag{16}$$

where the total learning cost is the time consumed by a learning agent to finish a task, and the useful learning cost is the time consumed to exploit the obtained optimal strategy. Simulation results were provided in [60] to show that the learning efficiencies of both the pre-partitioning and the weight-driven exploration schemes are significantly improved compared to the traditional uniform random exploration scheme.

A distributed multi-agent multi-band RL based sensing policy was proposed in [57] for ad-hoc cognitive networks. The proposed sensing policy employs secondary user (SU) local collaborations. The goal is to maximize the amount of available spectrum found for secondary use given a desired diversity order, i.e. a desired number of SUs sensing simultaneously each frequency band. The RL algorithm formulated is employed by each SU to update the local action values. The action value is approximated by a linear function in order to reduce the dimensionality of the spectrum sensing state-action space in a multiagent scenario, allowing computationally efficient learning also in networks with high numbers of secondary users and different frequency bands. The authors in [91] proposed a medium access control (MAC) protocols for autonomous cognitive radios. The protocol is based on the Q-learning and allows learning an efficient sensing policy in a multi-agent decentralized partially observable Markov decision process (DEC-

[2]Robinson-Monro algorithm is a stochastic approximation [95] method that functions by placing conditions on iterative step sizes and whose convergence is guaranteed under mild conditions [96].

POMDP) [97] environment. The DEC-POMDP framework is a model to represent multiple agents making decisions under uncertainty. It is an extension of the partially observable Markov decision process (POMDP) [98], [99] framework and a specific case of a partially observable stochastic game (POSG) [100]. The optimal solution of the POMDP was derived in [98] by considering the POMDP as an Markov decision process (MDP) [88] with an infinite state space. This solution was obtained by following the dynamic programming approach. However, it suffers from high computational complexity due to the infinite dimension of the state space, which makes it computationally intractable [101]. Hence, approximate solutions with low complexity are usually suggested for POMDP problems in order to avoid the high complexity of the optimal solution [54], [101]. In particular, several RL algorithms were shown to provide efficient near-optimal solutions to the POMDP's, yet with low complexity [54], [102], [103].

In [104], RL was employed for learning problems in a dynamic spectrum leasing (DSL) framework. The algorithms allows to reach an equilibrium for the proposed auction game with both centralized and distributed cognitive networks architectures. The authors in [105] proposed a stochastic game framework for anti-jamming defense in cognitive radios. In particular, the minimax Q-learning [106] was used to learn the optimal secondary policy so as to maximize the spectrum-efficient throughput. The minimax Q-learning is essentially identical to the standard Q-learning algorithm with a minimax replacing the max in (13) [106]. The essence of minimax is to behave so as to maximize your reward in the worst case: For sometimes, the performance of an agent depends critically on the actions of the opponent. In the game theory literature, the resolution to this problem is to eliminate the choice and evaluate each policy with respect to the opponent that makes it look the worst. This performance measure prefers conservative strategies that can force any opponent to a draw to more daring ones that accrue a great deal of reward against some opponents and lose a great deal to others [106]. Using the minimax Q-learning, the authors in [105] made the secondary users gradually learn the optimal policy, which maximizes the expected sum of discounted payoffs defined as the spectrum-efficient throughput. Simulation results showed that the optimal policy obtained from the minimax Q-learning can achieve much better performance in terms of spectrum-efficient throughput, compared to the myopic learning policy which only maximizes the payoff at each stage without considering the dynamics of the environment and the cognitive capability of attackers.

B. Non-parametric Learning: The Dirichlet Process Mixture Model (DPMM)

A major challenge an autonomous cognitive radio can face is the lack of knowledge about the surrounding RF environment, in particular, when operating in the presence of unknown primary signals. Even in such situations, a cognitive radio is assumed to be able to adapt to its environment while satisfying certain requirements. For example, in DSA, a cognitive radio cannot exceed a certain collision probability with primary users, under any circumstance. For this reason, a cognitive radio should be equipped with the ability to autonomously explore its surrounding environment and to make decisions about the primary activity based on the observed data. In particular, a cognitive radio must be able to extract knowledge concerning the statistics of the primary signals based on measurements. This makes unsupervised learning an appealing approach for cognitive radios in this context. The RL has been shown to ensure efficient learning for cognitive radios in Markovian environments. In this section, however, we will focus on non-parametric learning techniques [107] that do not rely on the Markovian property of the environment, yet ensure efficient learning and adaptation. In particular, we will explore a Dirichlet process prior based [108]–[111] technique as a framework for non-parametric learning and point out its potentials and limitations. The Dirichlet process prior based techniques are considered as unsupervised learning methods since they make few assumptions about the distribution from which the data is drawn [112], [113], as can been seen from this sub-section.

First, a Dirichlet process $DP(\alpha_0, G_0)$ is defined to be the distribution of a random probability measure G that is defined over a measurable space $(\Theta, \mathcal{B})$, such that, for any finite measurable partition $(A_1, \cdots, A_r)$ of Θ, the random vector $(G(A_1), \cdots, G(A_r))$ is distributed as a finite dimensional Dirichlet distribution with parameters $(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_r))$, where $\alpha_0 > 0$ [112]. We denote:

$$(G(A_1), \cdots, G(A_r)) \sim Dir(\alpha_0 G_0(A_1), \cdots, \alpha_0 G_0(A_r)) \ , \tag{17}$$

where $G \sim DP(\alpha_0, G_0)$, denotes that the probability measure G is drawn from the Dirichlet process $DP(\alpha_0, G_0)$. In other words, G is a *random probability measure* whose distribution is given by the Dirichlet process $DP(\alpha_0, G_0)$ [112].

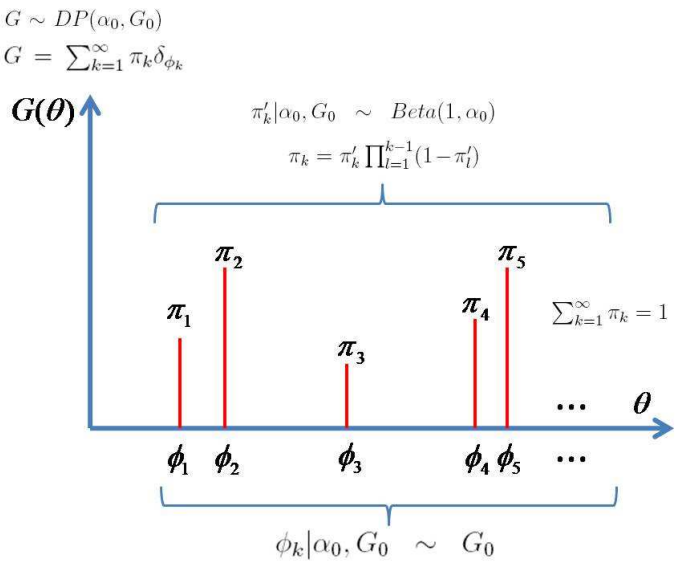

Fig. 6. One realization of the Dirichlet process.

1) Construction of the Dirichlet process: Teh [112] describes several ways of constructing the Dirichlet process. A first method is a direct approach that constructs the random probability distribution G based on the *stick-breaking* method. The *stick-breaking* construction of G can be summarized as follows **[112]**:

1) Generate independent i.i.d. sequences $\{\pi'_k\}_{k=1}^{\infty}$ and $\{\phi_k\}_{k=1}^{\infty}$ such that

$$\begin{cases} \pi'_k|\alpha_0, G_0 & \sim & Beta(1, \alpha_0) \\ \phi_k|\alpha_0, G_0 & \sim & G_0 \end{cases}, \tag{18}$$

where Beta(a, b) is the beta distribution whose probability density function (pdf) is given by $f(x, a, b) = \frac{x^{a-1}(1-x)^{b-1}}{\int_0^1 u^{a-1}(1-u)^{b-1}du}$.

2) Define $\pi_k = \pi'_k \prod_{l=1}^{k-1}(1-\pi'_l)$. We can write $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots) \sim GEM(\alpha_0)$, where GEM stands for Griffiths, Engen and McCloskey [112]. The $GEM(\alpha)$ process generates the vector $\boldsymbol{\pi}$ as described above, given a parameter α in (18).

3) Define $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$, where δ_{ϕ} is a probability measure concentrated at ϕ (and $\sum_{k=1}^{\infty} \pi_k = 1$).

In the above construction G is a random probability measure distributed according to $DP(\alpha_0, G_0)$. The randomness in G stems from the random nature of both the weights π_k and the weights positions ϕ_k. A sample distribution G of a Dirichlet process is illustrated in Fig. 6, using the steps described above in the *stick-breaking* method. Since G has an infinite discrete support (i.e. $\{\phi_k\}_{k=1}^{\infty}$), this makes it a suitable candidate for non-parametric Bayesian classification problems in which the number of clusters is unknown *a priori* (i.e. allowing for infinite number of clusters), with the infinite discrete support (i.e. $\{\phi_k\}_{k=1}^{\infty}$ being the set of clusters. However, due to the infinite sum in G, it may not be practical to construct G directly by using this approach in many applications. An alternative approach to construct G is by using either the Polya urn model [111] or the Chinese Restaurant Process (CRP) [114]. The CRP is a discrete-time stochastic process. A typical example of this process can be described by a Chinese restaurant with infinitely many tables and each table (cluster) having infinite capacity. Each customer (feature point) that arrives to the restaurant (RF spectrum) will choose a table with a probability proportional to the number of customers on that table. It may also choose a new table with a certain fixed probability.

A second approach does not define G explicitly. Instead, it characterizes the distribution of the drawings θ of G. Note that G is discrete with probability 1. The Polya urn model [111] does not construct G directly, but it characterizes the draws from G. Let $\theta_1, \theta_2, \cdots$ be i.i.d. random variables distributed according to G. These random variables are independent, given G. However, if G is integrated out, $\theta_1, \theta_2, \cdots$ are no more conditionally independent and they can be characterized as:

$$\theta_i|\theta_1, \cdots, \theta_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{K} \frac{m_k}{i-1+\alpha_0}\delta_{\phi_k} + \frac{\alpha_0}{i-1+\alpha_0}G_0 \ , \tag{19}$$

where $\{\phi_k\}_{k=1}^{K}$ are the K distinct values of θ_i's and m_k is the number of values θ_i that are equal to ϕ_k. Note that this conditional distribution is not necessarily discrete since G_0 might be a continuous distribution (in contrast with G which is discrete with probability 1). The θ_i's that are drawn from G exhibit a clustering behavior since a certain value of θ_i is most likely to reoccur with a nonnegative probability (due to the point mass functions in the conditional distribution). Moreover, the number of distinct θ_i values is infinite, in general, since there is a nonnegative probability that the new θ_i value is distinct from the previous $\theta_1, \cdots, \theta_{i-1}$. This

conforms with the definition of G as a probability mass function (pmf) over an infinite discrete set. Since θ_i's are distributed according to G, given G, we denote:

$$\theta_i|G \sim G \ . \tag{20}$$

2) Dirichlet Process Mixture Model (DPMM): The Dirichlet process makes a perfect candidate for non-parametric classification problems through the Dirichlet process mixture model (DPMM). The DPMM imposes a non-parametric prior on the parameters of the mixture model [112]. The DPMM can be modeled as follows:

$$\begin{cases} G & \sim & DP(\alpha_0, G_0) \\ \theta_i|G & \sim & G \\ y_i|\theta_i & \sim & f(\theta_i) \end{cases} , \tag{21}$$

where θ_i's denote the mixture components and the y_i is drawn according to this mixture model with a density function f given a certain mixture component θ_i.

3) Data clustering based on the DPMM and the Gibbs sampling: Consider a sequence of observations $\{y_i\}_{i=1}^N$ and assume that these observations are drawn from a mixture model. If the number of mixture components is unknown, it is reasonable to assume a non-parametric model, such as the DPMM. Thus, the mixture components θ_i are drawn from $G \sim DP(\alpha_0, G_0)$, where G can be expressed as $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$, ϕ_k's are the unique values of θ_i, and π_k are their corresponding probabilities. Denote $\mathbf{y} = (y_1, \cdots, y_N)$.

The problem is to estimate the mixture component $\hat{\theta}_i$ for each observation y_i, for all $i \in \{1, \cdots, N\}$. This can be achieved by applying the Gibbs sampling [115] method proposed in [116] which has been applied for several unsupervised clustering problems, such as speaker clustering problem in [117]. The Gibbs sampling is a technique for generating random variables from a (marginal) distribution indirectly, without having to calculate the density. As a result, by using te Gibbs sampling, we are able to avoid difficult calculations, replacing them instead with a sequence of easier calculations. Although the roots of the Gibbs sampling can be traced back to at least Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) [115], the Gibbs sampling became popular after the paper of Geman and Geman (1984) [118], who studied image-processing models. More recently, Gelfand and Smith (1990) [119] generated new interest in the

Gibbs sampler by revealing its potential in a wide variety of conventional statistical problems. A good tutorial on the Gibbs sampling can be found in [120].

In the Gibbs sampling method proposed in [116], the estimates $\hat{\theta}_i$ will be sampled from the conditional distribution of θ_i, given all the other feature points and the observation vector $\mathbf{y}$. This distribution was obtained in [116] to be

$$\theta_i|\{\theta_j\}_{j\neq i},\mathbf{y} = \begin{cases} \theta_j & \text{with Pr. } \frac{f_{\theta_j}(y_i)}{A(y_i)+\sum_{l=1,l\neq i}^{N} f_{\theta_l}(y_i)} \\ \sim h(\theta|y_i) & \text{with Pr. } \frac{A(y_i)}{A(y_i)+\sum_{l=1,l\neq i}^{N} f_{\theta_l}(y_i)} \end{cases}, \tag{22}$$

where $h(\theta_i|y_i) = \frac{\alpha_0}{A(y_i)} f_{\theta_i}(y_i) G_0(\theta_i)$ and $A(y) = \alpha_0 \int f_\theta(y) G_0(\theta) d\theta$.

In order to illustrate this clustering method, consider a simple example summarizing the process. We assume a set of mixture components $\theta \in \mathbb{R}$. Also, we assume $G_0(\theta)$ to be uniform over the range $[\theta_{min}, \theta_{max}]$. Note that this is a worst-case scenario assumption whenever there is no prior knowledge of the distribution of θ, except its range. Let $f_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\theta)^2}{2\sigma^2}}$.

Hence, $A(y) = \frac{\alpha_0}{\theta_{max}-\theta_{min}} \left[Q\left(\frac{\theta_{min}-y}{\sigma}\right) - Q\left(\frac{\theta_{max}-y}{\sigma}\right)\right]$ and

$$h(\theta_i|y_i) = \begin{cases} B \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta_j)^2}{2\sigma^2}} & \text{if} \quad \theta_{min} \leq \theta_i \leq \theta_{max} \\ 0 & \text{otherwise} \end{cases}, \tag{23}$$

where $B = \frac{1}{Q\left(\frac{\theta_{min}-y_i}{\sigma}\right) - Q\left(\frac{\theta_{max}-y_i}{\sigma}\right)}$. Initially, we set $\theta_i = y_i$ for all $i \in \{1, \cdots, N\}$. The algorithm is described in Algorithm 1.

Algorithm 1 Clustering algorithm.

```
Initialize θ̂_i = y_i, ∀i ∈ {1, ··· , N}.
while Convergence condition not satisfied do
  for i = shuffle {1, ··· , N} do
    Use Gibbs sampling to obtain θ̂_i from the distribution in (22).
  end for
end while
```

If the observation points $y_i \in \mathbb{R}^k$ (with $k > 1$), the distribution of $h(\theta_i|y_i)$ becomes too complicated to be used in the sampling process of θ_i's. In [116], if $G_0(\theta)$ is constant in a large area around y_i, $h(\theta|y_i)$ was shown to be approximated by the Gaussian distribution (assuming that the observation pdf $f_\theta(y_i)$ is Gaussian). In our case, assuming a large uniform prior distribution

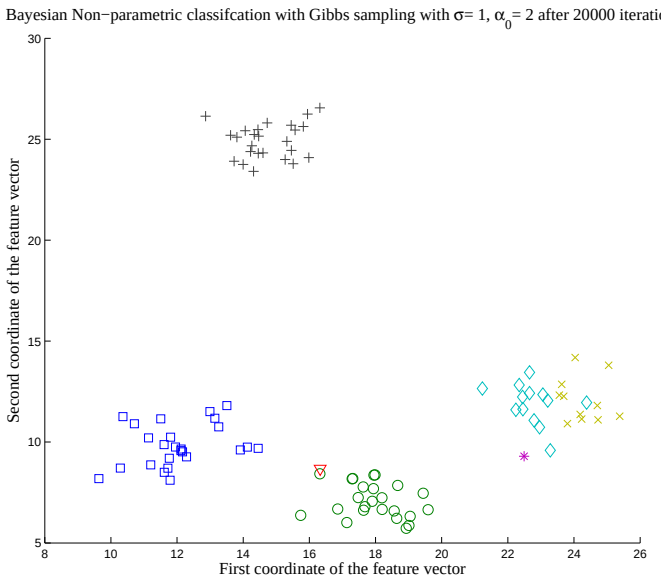

Fig. 7. The observation points y_i are classified into different clusters, denoted with different marker shapes. The original data points are generated from a Gaussian mixture model with 4 mixture components and with an identity covariance matrix.

on θ, we can approximate $h(\theta|y)$ by the Gaussian pdf. Thus, (23) becomes:

$$h(\theta_i|y_i) = \mathcal{N}(y_i, \Sigma) \ , \tag{24}$$

where Σ is the covariance matrix.

In order to illustrate this approach in a multidimensional scenario, we may generate a Gaussian mixture model having 4 mixture components. The mixture components have different means in $\mathbb{R}^2$ and they have an identity covariance matrix. We assume that the covariance matrix is known.

We plot in Fig. 7 the results of the clustering algorithm based on DPMM. Three of the clusters were almost perfectly identified, whereas the forth cluster was split into three parts. The main advantage of this technique is its ability of learning the number of clusters from the data itself, without any prior knowledge. As opposed to heuristic or supervised classification approaches that assume a fixed number of clusters (such as the K-mean approach), the DPMM-based clustering technique is completely unsupervised, yet, provides effective classification results. This makes it a perfect choice for autonomous cognitive radios that rely on unsupervised learning for decision-making.

4) Applications of DP to cognitive radios: The Dirichlet process has been used as a framework for non-parametric Bayesian learning in cognitive radios in [61], [121]. The approach was used for identifying and classifying wireless systems in [121], based on the CRP. The method consists of extracting two features from the observed signals (in particular, the center frequency and frequency spread) and to classify these feature points in a feature space by adopting an unsupervised clustering technique, based on the CRP. The objective is to identify both the number and types of primary systems that exist in a certain frequency band at a certain moment. One application of this could be when multiple wireless systems co-exist in the same frequency band and try to communicate without interfering with each other. Such scenarios could arise in ISM bands where wireless local area networks (WLAN IEEE 802.11) coexist with personal area networks (PAN), such as Zigbee (IEEE 802.15.4) and Bluetooth (IEEE 802.15.1). In that case, a PAN should sense the ISM band before selecting its communication channel so that it does not interfere with the WLAN or other PAN systems. A practical assumption, in that case, is that individual wireless users do not know the number of the other coexisting wireless users. Instead, these unknown variables should be learnt based on appropriate autonomous learning algorithms. Moreover, the designed learning algorithms should account for the dynamics of the RF environment. For example, the number of wireless users might change over time. These dynamics should be handled by an embedded flexibility offered by non-parametric learning approaches.

The advantages of the DP-based learning technique in [121] is that it does not rely on training data, making it suitable for identifying unknown signals by using unsupervised learning techniques. In this survey, we do not delve into details of choosing and computing appropriate feature points for the particular application considered in [121]. Instead, our focus is below on the implementation of the unsupervised learning and clustering technique.

After sensing a certain signal, the radio extracts a feature point that captures certain spectrum characteristics. Usually, the extracted feature points are noisy and might be affected by estimation errors, receiver noise, path loss, etc. Moreover, the statistical distribution of these observations might be unknown itself. It is assumed that feature points that are extracted from a particular system belong to the same cluster in the feature space. Depending on the feature definition, different systems might result in different clusters that are located at different places in the feature

space. For example, if the feature point represents the center frequency, two systems transmitting at different carrier frequencies will result in feature points that are distributed around different mean points.

The authors in [121] argue that the clusters of a certain system are random themselves and might be drawn from a certain distribution. That is, not to mention the randomness in the observed data, given a particular cluster. To illustrate this idea, assume two WiFi transmitters located at different distances from the receiver that both uses WLAN channel 1. Although the two transmitters belong to the same system (i.e. WiFi channel 1), their received powers might be different, resulting in variations of the features extracted from the signals of the same system. To capture this randomness, it can be assumed that the position and structure of the clusters formed (i.e. mean, variance, etc.) are themselves drawn from some distribution.

To be concrete, denote x as the derived feature point and assum that x is normally distributed (i.e. $x \sim \mathcal{N}(\mu_c, \Sigma)$) with mean μ_c and covariance matrix Σ_c. These two parameters characterize a certain cluster and are drawn from certain distribution. For example, it can be assumed that $\mu_c \sim \mathcal{N}(\mu_M, \Sigma_M)$ and $\Sigma_c \sim \mathcal{W}(V, n)$, where $\mathcal{W}$ denotes the Wishart distribution, which can be used to model the distribution of the covariance matrix of multivariate Gaussian variables.

In the method proposed in [121], a training process[3] is required to estimate the parameters μ_M and Σ_M. The estimation is performed by sensing a certain system (e.g. WiFi, or Zigbee) under different scenarios and estimating the centers of the clusters resulting from each experiment (i.e. estimating μ_c). The average of all μ_c's forms a maximum-likelihood (ML) estimate of the parameter μ_M of the corresponding wireless system. This step is equivalent to estimating the hyperparameters of a Dirichlet process [113]. Similar estimation method can also be performed to estimate Σ_M.

The knowledge of μ_M and Σ_M helps identify the corresponding wireless system of each cluster. That is, the maximum a posteriori (MAP) detection can be applied to a cluster center μ_c to estimate the wireless system that it belongs to. However, the classification of feature points into clusters can be done based on the CRP.

[3]Note that the training process used in [121] refers to the cluster formation process. The training used in [121] is done without data labeling nor human instructions, but done with the CRP [114] and the Gibbs sampling [116], thus still qualifies for the unsupervised learning schemes.

The classification of a feature point into a certain cluster is made based on the Gibbs sampling applied to the CRP. The algorithm fixes the cluster assignments of all other feature points. Given that assignment, it generates a cluster index for the current feature point. This sampling process is applied to all the feature points separately until certain convergence criterion is satisfied. Other examples of the CRP-based feature classification can be found in speaker clustering [117] and document clustering applications [122].

C. Game theory-based Learning

Game theory [123] presents a suitable platform for implementing rational behavior among cognitive radios in CRN's. There is a rich literature on game theoretic applications in cognitive radio, such as in [124]–[135]. A survey on game theoretic approaches for multiple access wireless systems can be found in [136].

Game theory [123] is a mathematical tool that implements the behavior of rational entities in an environment of conflict. This branch of mathematics has primarily been popular in economics, and was later applied to biology, political science, engineering and philosophy [136]. In wireless communications, game theory has been applied to data communication networking, in particular, to model and analyze routing and resource allocation in competitive environments. A game model consists of several rational entities that are denoted as the players. Each player has a set of available actions and a utility function. The utility function of an individual player depends on the actions taken by all the players, in general. Each player selects its strategy (i.e. action sequence) in order to maximize its utility function. A Nash equilibrium of a game is defined as the point at which the utility function of each player does not increase if the player deviates from that point, given that the other players' actions are fixed.

A key advantage of applying game theoretic solutions to cognitive radio protocols is in reducing the complexity of adaptation algorithms in large cognitive networks. While optimal centralized control is computationally prohibitive in most CRN's, due to communication overhead and algorithm complexity, game theory presents a platform to handle such situation, distributively [137]. Another reason for applying game theoretic approaches to cognitive radios is the assumed cognition in the cognitive radio behavior, which induces *rationality* among cognitive radios, similar to the players in a game.

Several types of games have been adapted to model different situations in cognitive radio networks [137]. For example, supermodular games [138] (the games having an important and useful property: there exists at least one pure strategy Nash equilibrium) are used for distributed power control [139], [140] and rate adaptation [141]. Repeated games were applied for dynamic spectrum access (DSA) by multiple SU's that share the same spectrum hole [142]. In this context, repeated games are useful in building reputations and applying punishments in order to reinforce a certain desired outcome. The Stackelberg game model can be used as a model for implementing cognitive radio behavior in cooperative spectrum leasing where the primary users act as the game-leaders and secondary cognitive users as the followers [35].

Auctions are one of the most popular methods used for selling a variety of items, ranging from antiques to wireless spectrum. In auction games the players are the buyers who must select the appropriate bidding strategy in order to maximize their perceived utility (i.e., the value of the acquired items minus the payment to the seller). The auction games were applied to cooperative dynamic spectrum leasing (DSL) applications, as in [104], as well as to spectrum allocation problems, as in [143]. The basics of the auction games and the open challenges of auction games to the field of spectrum management are provided in [144].

Stochastic games [145] can be used to model the greedy selfish behavior of cognitive radios in a cognitive radio network, where cognitive radios try to learn their best response and improve their strategies over time [146]. In the context of cognitive radios, stochastic games are dynamic, competitive games with probabilistic actions played by SU's. The game is played in a sequence of stages. At the beginning of each stage, the game is in a certain state. The SU's choose their actions, and each SU receives a reward that depends on both its current state and its selected actions. The game then moves to the next stage having a new state with a certain probability, which depends on the previous state and the actions selected by the SU's. The process continues for a finite or infinite number of stages. The stochastic games are generalizations of repeated games that only have one single state.

D. Threshold Learning

A cognitive radio can be implemented on a mobile device that changes location over time and switches transmissions among several channels. This mobility and multi-band/multi-channels

operability causes a major problem for cognitive radios in adapting to their RF environments. A cognitive radio may encounter different noise or interference levels when switching between different bands or when moving from one place to another. Hence, the operating parameters (e.g. test thresholds, sampling rate, etc.) of cognitive radios need to be adapted with respect to each particular situation. Moreover, cognitive radios may be operating in unknown RF environments and may not have perfect knowledge of the characteristics of the other existing primary or secondary signals, which require special learning algorithms to allow the cognitive radio to explore and adapt to its surrounding environment. In this context, special types of learning can be applied to directly learn the optimal setup of certain design and operation parameters.

Threshold learning presents a technique that permits such dynamic adaptation of operating parameters to satisfy the performance requirements, while continuously learning from the past experience. By assessing the effect of previous parameter values on the system performance, the learning algorithm optimizes the parameters values in order to ensure a desired performance. For example, when considering energy detection, after measuring the energy levels at each frequency, a cognitive radio decides on the occupancy of a certain frequency band by comparing the measured energy levels to a certain threshold. The threshold levels are usually designed based on Neyman-Pearson tests in order to maximize the detection probability of primary signals, while satisfying a constraint on the false alarm. However, in such tests, the optimal threshold depends on the noise level. A bad estimation of the noise levels might cause sub-optimal behavior and violation of the operation constraints (for example, exceeding a tolerable collision probability with primary users). In this case, and in the absence of perfect knowledge about the noise levels, threshold-learning algorithms can be devised to learn the optimal threshold values. Given each choice of a threshold, the resulting false alarm rate determines how the test threshold should be regulated to achieve a desired false alarm probability. An example of threshold learning algorithms can be found in [147] where a threshold learning process was derived for optimizing spectrum sensing in cognitive radios. The resulting algorithm was shown to converge to the optimal threshold that satisfies a given false alarm probability.

IV. Supervised Learning

Unlike the unsupervised learning techniques discussed in the previous section that may be used in alien environments without having any prior knowledge, supervised learning techniques are generally used in certain familiar/known environments, with prior knowledge about the characteristics of the environment. In the following, we introduce some of the major supervised learning techniques that have been applied to the cognitive radio literature.

A. Artificial Neural Network (ANN)

The work on ANN has been motivated by the recognition that human brain computes in an entirely different way compared to the conventional digital computers [148]. A neural network is defined to be *a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use* [148]. An ANN resembles the brain in two respects [148]: 1) knowledge is acquired by the network from its environment through a learning process, and 2) interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge.

Some of the top beneficial properties and capabilities of ANN's includes: 1) nonlinearity fitness to underlying physical mechanisms; 2) adaptive to minor changes of surrounding environment; 3) in the context of pattern classification, the ANN provides information not only about which particular pattern to select, but also the confidence in the decision made. However, the disadvantages of ANN's is that 1) they require a large diversity of training for real-world operations, which can lead to excessive hardware necessities and efforts; 2) the training outcome of an ANN can sometimes be nondeterministic and depend crucially on the choice of initial parameters.

Various applications of ANN to cognitive radios can be found in recent literature [149]–[154]. The authors in [149] proposed the use of Multilayered Feedforward Neural Networks (MFNN) as a technique to synthesize performance evaluation functions in cognitive radios. The benefit of using MFNNs is that they provide a general-purpose black-box modeling of the performance as a function of the measurements collected by the cognitive radio; furthermore, this characterization can be obtained and updated by a cognitive radio at run-time, thus effectively achieving a certain level of learning capability. The authors in [149] also demonstrated the concept in several IEEE

802.11 based environments to show how these modeling capabilities can be used for optimizing the configuration of a cognitive radio.

In [150], the authors proposed an ANN-based cognitive engine that learns how environmental measurements and the status of the network affect its performance on different channels. In particular, an implementation of the proposed Cognitive Controller for dynamic channel selection in IEEE 802.11 wireless networks was presented. Performance evaluation carried out on an IEEE 802.11 wireless network deployment demonstrated that the Cognitive Controller is able to effectively learn how the network performance is affected by changes in the environment, and to perform dynamic channel selection thereby providing significant throughput enhancements.

In [151], an application of a Feedbackward ANN in conjunction with the cyclostationarity-based spectrum sensing was presented to perform spectrum sensing. The results showed that the proposed approach was appropriate to detect the signals under considerably low signal-to-noise ratio (SNR) environment. In [152], the authors designed a channel status predictor using a MFNN model. The authors argued that their proposed MFNN-based prediction is superior to the hidden Markov model (HMM) based approaches, by pointing out that the HMM based approaches require a huge memory space to store a large number of past observations with high computational complexity.

In [153], the authors proposed a methodology for spectrum prediction by modeling licensed user features as a multivariate chaotic time series, which is then given as input to an ANN, that predicts the evolution of RF time series to decide if the unlicensed user can exploit the spectrum band. Experimental results show a similar trend between predicted and observed values. This proposed spectrum evolution prediction method was done by exploiting the cyclostationary signal features to construct a RF multivariate time series that contain more information than the univariate time series [155], in contrast to most of the modeling methodologies which focus on the univariate time series prediction [156].

In [154], a feedforward ANN-based automatic modulation classification (AMC) algorithm was applied for signal sensing and detection of primary users in cognitive radio environments. An eight-dimension feature was used as inputs to the feedforward network, and 13 neurons at the output layer corresponding to the number of targets: 12 analog and digital modulation schemes and noise signal. The results showed the high recognition-success rate of the proposed classifier

in additive white Gaussian noise (AWGN) channels. However, the classification performance for AWGN channels with fading and other types of channels were not provided.

B. Support Vector Machine

The Support Vector Machine (SVM), developed by Vapnik and others [157], [158], is used for many machine learning tasks such as pattern recognition and object classifications. The SVM is characterized by the absence of local minima, the sparseness of the solution and the capacity control obtained by acting on the margin, or on other dimension independent quantities such as the number of support vectors [157], [158]. SVM based techniques have achieved superior performances in a wide variety of real world problems due to their generalization ability and robustness against noise and outliers [159].

The basic idea of SVM's is to map the input vectors into a high-dimensional feature space in which they become linearly separable. The mapping from the input vector space to the feature space is a non-linear mapping which can be done by using kernel functions. Depending on the application different types of kernel functions can be used. A common choice for classification problems is the Gaussian kernel which is a polynomial kernel of infinite degree. When performing classification, a hyperplane which allows for the largest generalization in this high-dimensional space is found. This is so-called a maximal margin classifier. As shown in Fig. 8, there could be many possible separating hyperplanes between the two classes of data, but only one of them allows for the a maximum margin. A margin is the distance from a separating hyperplane to the closest data points. These closest data points are named support vectors and the hyperplane allowing for the maximum margin is called an optimal separating hyperplane. The interested reader is referred to [160], [161] for insightful coverage of SVM's.

Many applications of SVM's to cognitive radio can be found in current literatures, including [44], [51], [159], [162]–[168]. Most of the applications of the SVM in cognitive radio context, however, has been in performing signal classifications.

In [165], for example, a MAC protocol classification scheme was proposed to classify contention based and control based MAC protocols in an unknown primary network based on SVMs. To perform the classification in an unknown primary network, the mean and variance of the received power are chosen as two features for the SVM. The SVM is embedded in a cognitive

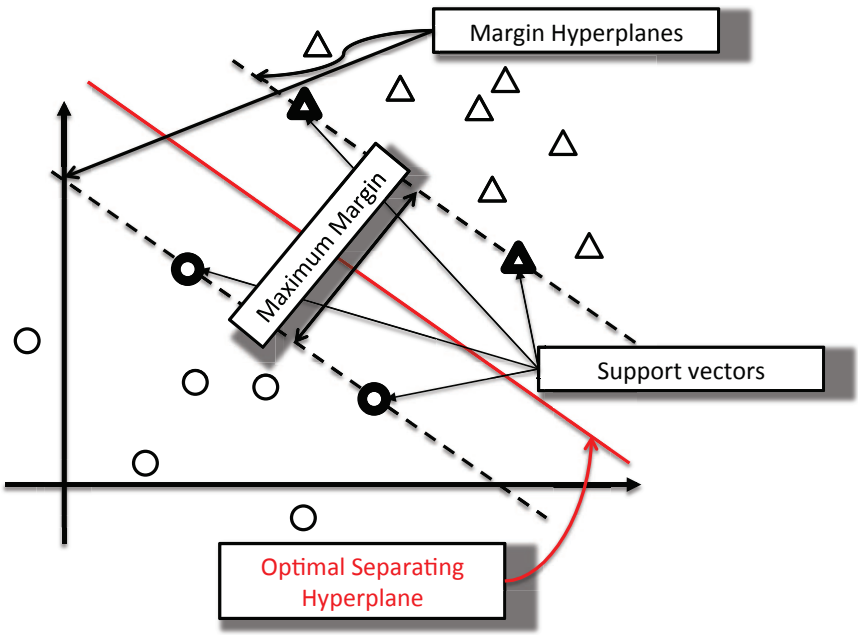

Fig. 8. A diagram showing the basic idea of SVM: optimal separation hyperplane (solid red line) and two margin hyperplanes (dashed lines) in a binary classification example; Support vectors are bolded.

radio terminal of the secondary network. A TDMA and a slotted Aloha network were setup as the primary networks. Simulation results showed that TDMA and slotted Aloha MAC protocol could be effectively classified by the cognitive radio terminal and the correct classification rate is proportional to the transmission rate of the primary networks, where the transmission rate for the primary networks is defined as the new packet generating/arriving probability in each time slot. The reason why the correct classification rate increases when the transmission rate increases is the following: for slotted Aloha network, the higher transmission rate brings the higher collision probability, and thus the higher instantaneous received power captured by a cognitive radio terminal; for TDMA network, however, there is no relation between transmission rate and instantaneous captured received power. Therefore, when the transmission rates of the primary networks both increase, it makes a cognitive radio terminal easier to differentiate TDMA and slotted Aloha.

SVM classifiers can not only be a binary classifier as shown its application in the previous exmaple, but also it can be easily used as multi-class classifiers by treating a K-class classification problem as K two-class problems. For example, in [166] the authors presented a study of multi-

class signal classification based on automatic modulation classification (AMC) through SVMs. A simulated model of an SVM signal classifier was implemented and trained to recognize seven distinct modulation schemes; five digital (BPSK, QPSK, GMSK, 16-QAM and 64-QAM) and two analog (FM and AM). The signals were generated using realistic carrier frequency, sampling frequency and symbol rate values, and realistic Raised- cosine and Gaussian pulse-shaping filters. The results show that the implemented classifier correctly classifies signals with high probabilities.

We summarize the discussed unsupervised learning techniques discussed in Section III and supervised learning techniques discussed in this section in the table shown in Fig. 9, with their suitable applications.

		Spectrum Sensing and MAC Protocols	Signal Classification and Feature Detection	Power Allocation and Rate adaptation	System Parameters Reconfiguration	Problem Formulation and Characteristics/Comments
Unsupervised learning techniques	Reinforcement learning (RL)	✔				Suitable for Markov Environments (i.e. MDP) Open problem in POMDP's, Non-Markov Environments and Decentralized Protocols.
	Non-parametric Learning: DPMM		✔			Feature classification based on the DPMM
	Game theory-based Learning	✔		✔		Stochastic games are Suitable for multi-agent systems with Markov states.
	Threshold Learning				✔	Suitable for controlling specific parameters, in particular threshold adaptation.
Supervised learning techniques	Artificial Neural Network (ANN)		✔			Requires training and supervision.
	Support Vector Machine (SVM)		✔			Requires training and supervision.

Fig. 9. A summary of the unsupervised and supervised learning techniques discussed in this survey with their common applications.

V. Centralized and Decentralized Learning in cognitive radio

Since noise uncertainties, shadowing, and multi-path fading effects limit the performance of spectrum sensing, when the received primary SNR is too low, there exists a SNR wall, below which reliable spectrum detection is impossible in some cases [169], [170]. If SU's cannot detect the primary transmitter, while the primary receiver is within the SU's transmission range, a hidden terminal problem occurs [171], [172], and the primary user's transmission will be interfered with. By taking advantage of diversity offered by multiple independent fading channels (multiuser diversity), cooperative spectrum sensing improves the reliability of spectrum sensing and the utilization of idle spectrum [173], [174], as opposed to non-cooperative spectrum sensing.

In centralized cooperative spectrum sensing [173], [174], a central controller collects local observations from multiple SU's, decides the spectrum occupancy by using decision fusion rules, and informs the SU's which channels to access. In distributed cooperative spectrum sensing [41], [175], on the other hand, SU's within a cognitive radio network exchange their local sensing results among themselves without requiring a backbone or centralized infrastructure. On the other hand, in the non-cooperative decentralized sensing framework, no communications are assumed among the SU's [176].

In [177], the authors showed how various centralized and decentralized spectrum access markets (where cognitive radios can compete over time for dynamically available transmission opportunities) can be designed based on a stochastic game (introduced in Section III-C) framework and solved using the proposed learning algorithm. The authors in [177] proposed a learning algorithm to learn the following information in the stochastic game: state transition model of other SU's, the state of other SU's, the policy of other SU's, and the network resource state. The proposed learning algorithm was similar to Q-learning. However, the main difference between this algorithm and Q-learning is that the former explicitly considers the impact of other SU actions through the state classifications and transition probability approximation. The computational complexity and performance are also presented in [177].

In [104] the authors proposed and analyzed both a centralized and a decentralized decision-making architecture with reinforcement learning for the secondary cognitive radio network. In this work, a new way to encourage primary users to lease their spectrum is proposed: the SU's place bids indicating how much power they are willing to spend for relaying the primary

signals to their destinations. In this formulation, the primary users achieve power savings due to asymmetric cooperation. In the centralized architecture, a secondary system decision center (SSDC) selects a bid for each primary channel based on optimal channel assignment for SU's. In the decentralized cognitive radio network architecture, an auction game-based protocol was proposed, in which each SU independently places bids for each primary channel and receivers of each primary link pick the bid that will lead to the most power savings. A simple and robust distributed reinforcement learning mechanism is developed to allow the users to revise their bids and to increase their rewards. The performance results show the significant impact of reinforcement learning in both improving spectrum utilization and meeting individual SU performance requirements.

In [178], the authors considered dynamic spectrum access among cognitive radios from an adaptive, game theoretic learning perspective, in which cognitive radios compete for channels temporarily vacated by licensed primary users in order to satisfy their own demands while minimizing interference. For both slowly varying primary user activity and slowly varying statistics of fast primary user activity, the authors applied an adaptive regret based learning procedure which tracks the set of correlated equilibria of the game, treated as a distributed stochastic approximation. The proposed approach is decentralized in terms of both radio awareness and activity; radios estimate spectral conditions based on their own experience, and adapt by choosing spectral allocations which yield them the greatest utility. Iterated over time, this process converges so that each radio's performance is an optimal response to others' activity. This apparently selfish scheme was also used to deliver system-wide performance by a judicious choice of utility function. This procedure is shown to perform well compared to other similar adaptive algorithms. The results of the estimation of channel contention for a simple CSMA channel sharing scheme was also presented.

In [179], the authors proposed an auction framework for cognitive radio networks to allow SUs to share the available spectrum of licensed primary users fairly and efficiently, subject to the interference temperature constraint at each PU. The competition among SU's was studied by formulating a non-cooperative multiple-PU multiple-SU auction game. The resulting equilibrium was found by solving a non-continuous two-dimensional optimization problem. A distributed algorithm was also developed in which each SU updates its strategy based on local information

to converge to the equilibrium. The proposed auction framework was then extended to the more challenging scenario with free spectrum bands. An algorithm was developed based on the no-regret learning to reach a correlated equilibrium of the auction game. The proposed algorithm, which can be implemented distributively based on local observation, is especially suited in decentralized adaptive learning environments. The authors demonstrated the effectiveness of the proposed auction framework in achieving high efficiency and fairness in spectrum allocation through numerical examples.

There has always been a trade-off between the centralized and decentralized control for radio networks in general. This is also true for cognitive radio networks. While the centralized scheme ensures efficient management of the spectrum resources, it often suffers from signaling and processing overhead. On the other hand, a decentralized scheme can reduce the complexity of the decision-making in cognitive networks. However, radios that act according to a decentralized scheme adopt a selfish behavior and try to maximize their own utilities, at the expense of the sum utility of the network, leading to an overall network efficiency. This problem can become more severe especially when considering heterogeneous networks in which different nodes belong to different types of systems and have different objectives (usually conflicting objectives). To resolve this problem, [180] proposes a hybrid approach for heterogeneous cognitive radio networks where the wireless users are assisted in their decisions by the network center. At some states of the system, the network manager imposes his decisions on users in the network. In other states, the mobile nodes may take autonomous actions in response to the information sent by the network center. As a result, the model in [180] avoids the completely decentralized network, due to the inefficiency of the non-cooperative network. Nevertheless, a large part of the decision-making is delegated to the mobile nodes to reduce the processing overhead at the central node.

In the problem formulation of [180], the authors consider a wireless network composed of S systems that are managed by the same operator. The set of all serving systems is denoted by $\mathcal{S} = \{1, \cdots, S\}$ and it corresponds to different serving systems. Since the throughput of each serving system drops in function of the distance of between the mobile and the base station, the throughput of a mobile changes within a given cell. To capture this variation, each cell is split into N circles of radius d_n ($n \in \mathcal{N} = \{1, \cdots, N\}$). Each circle area is assumed to have the same radio characteristics. In this case, all mobile systems that are located in circle $n \in \mathcal{N}$ and are

served by system $s \in \mathcal{S}$ achieve the same throughput. The network state matrix is denoted by $\mathbf{M} \in \mathcal{F}$, where $\mathcal{F} = \mathbb{N}^{N \times S}$. The (n, s)-th element M_n^s of the matrix $\mathbf{M}$ denotes the number of users with radio condition $n \in \mathcal{N}$ which are served by system $s \in \mathcal{S}$ in the circle. The network is fully characterized by its state $\mathbf{M}$, but this information is not available to the mobile nodes when the radio resource management (RRM) is decentralized. In this case, by using the *radio enabler* proposed by IEEE 1900.4, the network reconfiguration manager (NRM) broadcasts to the terminal reconfiguration manager (TRM) an aggregated load information that takes values in some finite set $\mathcal{L} = \{1, \cdots, L\}$ indicating whether the load state at mobile terminals are either low, medium or high. The mapping $f : \mathbf{M} \mapsto \mathcal{L}$ specifies a macro-state $f(\mathbf{M})$ for each network micro-state $\mathbf{M}$. This state encoding reduces the signaling overhead, while satisfying the IEEE 1900.4 standards which state that *the network manager side shall periodically update the terminal side with context information* [181]. Given the load information $l = f(\mathbf{M})$ and the radio condition $n \in \mathcal{N}$, the mobile makes its decision $P_{n,l} \in \mathcal{S}$, specifying which system it will connect to, and the user's decision vector is denoted by $\mathbf{P}^l = [P_{1,l} \cdots, P_{N,l}] \in \mathcal{P}$.

The authors in [180] find the association policies by following three different approaches:

1) Global optimum approach.
2) Nash equilibrium approach.
3) Stackelberg game approach.

The global optimum approach finds the policy that maximizes the global utility of the network. However, since it is not realistic to consider that individual users will seek the global optimum, another policy (corresponding to the Nash equilibrium) is obtained such that it maximizes the users's utilities. Finally, a Stackelberg game formulation was developed for the operator to control the equilibrium of its wireless users. This leads to maximizing the operator's utility by sending appropriate load information $l \in \mathcal{L}$.

The authors analyzed the network performance under these three different association policies. They demonstrated by means of Stackelberg formulation, how the operator can optimize its global utility by sending appropriate information about the network state, while users maximize their individual utilities. The resulting hybrid architecture achieves a good trade-off between the global network performance and the signaling overhead, which makes it a viable alternative to be considered when designing cognitive radio networks.

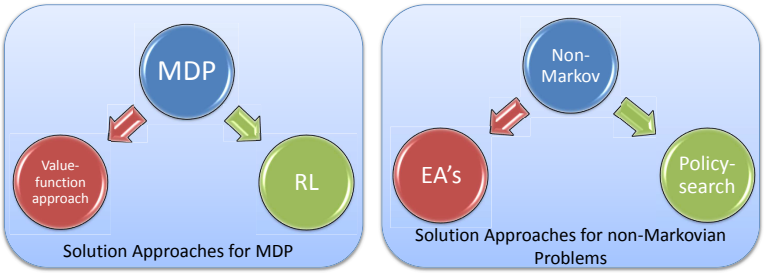

Fig. 10. Different approaches for solving Markovian and non-Markovian problems.

VI. Learning in non-Markovian environments

While reinforcement learning (RL) can lead to an optimal policy for the Markov decision process (MDP) problem, different studies have shown that evolutionary algorithms (EA's) can outperform the RL in non-Markovian environments [65], [68], compared to the *value-function* method [66], [67]. Non-Markovian environments arise in different situations, such as in the partially observable MDP (POMDP) problem. In addition, [65]–[67] suggested that methods that adopt *policy-search* algorithms also have higher advantage in non-Markovian tasks. These methods search directly for optimal policies in the policy space, without having to estimate the actual states of the systems [66], [67]. By adopting gradient search algorithms, these methods allow updating certain policy vector to reach optimality (might be local optima). Moreover, the value-function approach has several limitations: First, it is restricted to obtain deterministic policies. Second, any small changes in the estimated value of an action can cause that action to be, or not to be selected [66]. This would affect the optimality of the resulting policy since optimal actions might be eliminated due to an underestimation of their value functions. We illustrate in Fig. 10 the adequate solution methods that should be applied under each of the Markovian and non-Markovian frameworks discussed above.

To illustrate the policy-search approach, we give a brief overview of policy-gradient algorithms, as described in [67]. Consider a class of stochastic policies that are parameterized by $\theta \in \mathbb{R}^K$. By computing the gradient with respect to θ of the average reward, the policy could be improved by adjusting the parameters in the gradient direction. To be concrete, assume $r(X)$ to be a reward function that depends on a random variable X. Let $q(\theta, x)$ be the probability of the event

$\{X = x\}$. The gradient with respect to θ of the expected performance $\eta(\theta) = \mathbb{E}\{r(X)\}$ can be expressed as:

$$\nabla\eta(\theta) = \mathbb{E}\left\{ r(X) \frac{\nabla q(\theta, x)}{q(\theta, x)} \right\} \ . \tag{25}$$

An unbiased estimate of the gradient can be obtained via simulation by generating N independent identically distributed (i.i.d.) random variables $X_1, \cdots, X_N$ that are distributed according to $q(\theta, x)$. The unbiased estimate of $\nabla\eta(\theta)$ is thus expressed as:

$$\hat{\nabla}\eta(\theta) = \frac{1}{N} \sum_{i=1}^{N} r(X_i) \frac{\nabla q(\theta, X_i)}{q(\theta, X_i)} \ . \tag{26}$$

By the law of large numbers, $\hat{\nabla}\eta(\theta) \rightarrow \nabla\eta(\theta)$ with probability one. Note that the quantity $\frac{\nabla q(\theta, X_i)}{q(\theta, X_i)}$ is referred to as the *likelihood ratio* or *score function*. By having an estimate of the reward gradient, the policy parameter $\theta \in \mathbb{R}^K$ can be updated by following the gradient direction, such that:

$$\theta_{k+1} \leftarrow \theta_k + \alpha_k \nabla\eta(\theta) \ , \tag{27}$$

for some step size α_k.

Note that, the estimation of the gradient $\nabla\eta(\theta)$ is not straight-forward, especially in the absence of simulators that generate the X_i's. To resolve this problem, special algorithms can be designed to obtain reasonable approximations of the gradient. A straight-forward approach is to modify some elements in the parameter vector $\theta \in \mathbb{R}^k$ and to observe its effect on the reward $r(X)$. This is known as the Monte-Carlo method, but it is prohibitively inefficient for most of the problems.

VII. Conclusion

In this survey paper, we have characterized the learning problem in cognitive radios and stated the importance of machine learning in developing real cognitive radios. We have presented the state-of-the-art learning methods that are applied to cognitive radios and classified those methods under supervised and unsupervised learning. A description of the major learning algorithms was provided, and we presented their related applications in the cognitive radio domain. We also showed some of the challenging learning problems for cognitive radios and we showed their

possible solution methods.

REFERENCES

[1] J. Mitola III and G. Q. Maguire, Jr., "Cognitive radio: making software radios more personal," *IEEE Personal Communications*, vol. 6, no. 4, pp. 13 –18, Aug. 1999.

[2] J. Mitola, "Cognitive radio: An integrated agent architecture for software defined radio," Ph.D. dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden, 2000.

[3] T. Costlow, "Cognitive radios will adapt to users," *IEEE Intelligent Systems*, vol. 18, no. 3, p. 7, May-June 2003.

[4] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[5] FCC, "Report of the spectrum efficiency working group," FCC spectrum policy task force, Tech. Rep., Nov. 2002.

[6] ——, "ET docket no 03-322 notice of proposed rulemaking and order," Tech. Rep., Dec. 2003.

[7] N. Devroye, M. Vu, and V. Tarokh, "Cognetive radio networks," *IEEE Signal Processing Magazine*, vol. 25, pp. 12–23, Nov. 2008.

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

[9] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Sig. Proc. Magazine*, vol. 24, no. 3, pp. 79–89, May 2007.

[10] G. Zhao, J. Ma, Y. Li, T. Wu, Y. H. Kwon, A. Soong, and C. Yang, "Spatial spectrum holes for cognitive radio with directional transmission," in *2008 IEEE Global Telecommunications Conference (GLOBECOM 2008)*, Nov. 2008, pp. 1–5.

[11] A. Ghasemi and E. Sousa, "Spectrum sensing in cognitive radio networks: requirements, challenges and design trade-offs," *Communications Magazine, IEEE*, vol. 46, no. 4, pp. 32–39, April 2008.

[12] B. Farhang-Boroujeny, "Filter bank spectrum sensing for cognitive radios," *Signal Processing, IEEE Transactions on*, vol. 56, no. 5, pp. 1801–1811, May 2008.

[13] B. Farhang-Boroujeny and R. Kempter, "Multicarrier communication techniques for spectrum sensing and communication in cognitive radios," *Communications Magazine, IEEE*, vol. 46, no. 4, pp. 80–85, April 2008.

[14] C. R. C. da Silva, C. Brian, and K. Kyouwoong, "Distributed spectrum sensing for cognitive radio systems," in *Information Theory and Applications Workshop, 2007*, 29 2007-Feb. 2 2007, pp. 120–123.

[15] C. Cordeiro, M. Ghosh, D. Cavalcanti, and K. Challapali, "Spectrum sensing for dynamic spectrum access of tv bands," in *Cognitive Radio Oriented Wireless Networks and Communications, 2007. CrownCom 2007. 2nd International Conference on*, Aug. 2007, pp. 225–233.

[16] H. Chen, W. Gao, and D. G. Daut, "Signature based spectrum sensing algorithms for ieee 802.22 wran," in *Communications, 2007. ICC '07. IEEE International Conference on*, June 2007, pp. 6487–6492.

[17] Y. Zeng and Y. Liang, "Maximum-minimum eigenvalue detection for cognitive radio," in *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*, Sept. 2007, pp. 1–5.

[18] ——, "Covariance based signal detections for cognitive radio," in *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on*, April 2007, pp. 202–207.

[19] X. Zhou, Y. Li, Y. H. Kwon, and A. Soong, "Detection timing and channel selection for periodic spectrum sensing in cognitive radio," in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, Nov. 2008, pp. 1–5.

[20] Z. Tian and G. B. Giannakis, "A wavelet approach to wideband spectrum sensing for cognitive radios," in *Cognitive Radio Oriented Wireless Networks and Communications, 2006. 1st International Conference on*, June 2006, pp. 1–5.

[21] G. Ganesan and Y. Li, "Cooperative spectrum sensing in cognitive radio, part i: Two user networks," *Wireless Communications, IEEE Transactions on*, vol. 6, no. 6, pp. 2204–2213, June 2007.

[22] ——, "Cooperative spectrum sensing in cognitive radio, part ii: Multiuser networks," *Wireless Communications, IEEE Transactions on*, vol. 6, no. 6, pp. 2214–2222, June 2007.

[23] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2053 –2071, May 2008.

[24] S. Huang, X. Liu, and Z. Ding, "Opportunistic spectrum access in cognitive radio networks," in *The 27th Conference on Computer Communications. IEEE INFOCOM '08*, Phoenix, AZ, Apr. 2008, pp. 1427 –1435.

[25] K. Ben Letaief and W. Zhang, "Cooperative communications for cognitive radio networks," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 878 –893, May 2009.

[26] J. Ma, G. Y. Li, and B. H. Juang, "Signal processing in cognitive radio," *Proceedings of the IEEE*, vol. 97, no. 5, pp. 805–823, May 2009.

[27] W. Zhang, R. Mallik, and K. Letaief, "Optimization of cooperative spectrum sensing with energy detection in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 8, no. 12, pp. 5761 –5766, Dec. 2009.

[28] Y. M. Kim, G. Zheng, S. H. Sohn, and J. M. Kim, "An alternative energy detection using sliding window for cognitive radio system," in *10th International Conference on Advanced Communication Technology (ICACT '08)*, vol. 1, Gangwon-Do, South Korea, Feb. 2008, pp. 481 –485.

[29] J. Lunden, V. Koivunen, A. Huttunen, and H. Poor, "Collaborative cyclostationary spectrum sensing for cognitive radio systems," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4182 –4195, Nov. 2009.

[30] A. Dandawate and G. Giannakis, "Statistical tests for presence of cyclostationarity," *IEEE Transactions on Signal Processing*, vol. 42, no. 9, pp. 2355 –2369, Sep. 1994.

[31] B. Deepa, A. Iyer, and C. Murthy, "Cyclostationary-based architectures for spectrum sensing in ieee 802.22 wran," in *IEEE Global Telecommunications Conference (GLOBECOM '10)*, Miami, FL, Dec. 2010, pp. 1 –5.

[32] M. Gandetto and C. Regazzoni, "Spectrum sensing: A distributed approach for cognitive terminals," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 546 –557, Apr. 2007.

[33] J. Unnikrishnan and V. Veeravalli, "Cooperative sensing for primary detection in cognitive radio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 18 –27, Feb. 2008.

[34] T. Cui, F. Gao, and A. Nallanathan, "Optimization of cooperative spectrum sensing in cognitive radio," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 4, pp. 1578 –1589, May 2011.

[35] O. Simeone, I. Stanojev, S. Savazzi, Y. Bar-Ness, U. Spagnolini, and R. Pickholtz, "Spectrum leasing to cooperating secondary ad hoc networks," *IEEE J. Sel. Areas Commun.*, vol. 26, pp. 203–213, Jan. 2008.

[36] Q. Zhang, J. Jia, and J. Zhang, "Cooperative relay to improve diversity in cognitive radio networks," *IEEE Communications Magazine*, vol. 47, no. 2, pp. 111 –117, Feb. 2009.

[37] Y. Han, A. Pandharipande, and S. Ting, "Cooperative decode-and-forward relaying for secondary spectrum access," *IEEE Transactions on Wireless Communications*, vol. 8, no. 10, pp. 4945 –4950, Oct. 2009.

[38] L. Li, X. Zhou, H. Xu, G. Li, D. Wang, and A. Soong, "Simplified relay selection and power allocation in cooperative cognitive radio systems," *IEEE Transactions on Wireless Communications*, vol. 10, no. 1, pp. 33 –36, Jan. 2011.

[39] E. Hossain and V. K. Bhargava, *Cognitive Wireless Communication Networks*. Springer, 2007.

[40] "Special issue on cognitive radio," *Proceedings of the IEEE*, vol. 97, no. 4 and 5, 2009.

[41] B. Wang and K. J. R. Liu, "Advances in cognitive radio networks: A survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 1, pp. 5 –23, Feb. 2011.

[42] A. El-Saleh, M. Ismail, M. Ali, and J. Ng, "Development of a cognitive radio decision engine using multi-objective hybrid genetic algorithm," in *IEEE 9th Malaysia International Conference on Communications (MICC 2009)*, Dec. 2009, pp. 343 –347.

[43] L. Morales-Tirado, J. Suris-Pietri, and J. Reed, "A hybrid cognitive engine for improving coverage in 3g wireless networks," in *IEEE International Conference on Communications Workshops (ICC Workshops 2009).*, June 2009, pp. 1 –5.

[44] Y. Huang, H. Jiang, H. Hu, and Y. Yao, "Design of learning engine based on support vector machine in cognitive radio," in *International Conference on Computational Intelligence and Software Engineering (CiSE '09)*, Wuhan, China, Dec. 2009, pp. 1 –4.

[45] Y. Huang, J. Wang, and H. Jiang, "Modeling of learning inference and decision-making engine in cognitive radio," in *Second International Conference on Networks Security Wireless Communications and Trusted Computing (NSWCTC)*, vol. 2, Apr. 2010, pp. 258 –261.

[46] Y. Yang, H. Jiang, and J. Ma, "Design of optimal engine for cognitive radio parameters based on the duga," in *3rd International Conference on Information Sciences and Interaction Sciences (ICIS 2010)*, June 2010, pp. 694 –698.

[47] H. Volos and R. Buehrer, "Cognitive engine design for link adaptation: An application to multi-antenna systems," *IEEE Transactions on Wireless Communications*, vol. 9, no. 9, pp. 2902 –2913, Sep. 2010.

[48] C. Clancy, J. Hecker, E. Stuntebeck, and T. O'Shea, "Applications of machine learning to cognitive radio networks," *IEEE Wireless Communications*, vol. 14, no. 4, pp. 47 –52, Aug. 2007.

[49] A. N. Mody, S. R. Blatt, N. B. Thammakhoune, T. P. McElwain, J. D. Niedzwiecki, D. G. Mills, M. J. Sherman, and C. S. Myers, "Machine learning based cognitive communications in white as well as the gray space," in *IEEE Military Communications Conference. (MILCOM '07)*, Orlando, FL, Oct. 2007, pp. 1 –7.

[50] X. Dong, Y. Li, C. Wu, and Y. Cai, "A learner based on neural network for cognitive radio," in *12th IEEE International Conference on Communication Technology (ICCT '10)*, Nanjing, China, Nov. 2010, pp. 893 –896.

[51] M. M. Ramon, T. Atwood, S. Barbin, and C. G. Christodoulou, "Signal classification with an SVM-FFT approach for feature extraction in cognitive radio," in *SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC '09)*, Belem, Brazil, Nov. 2009, pp. 286 –289.

[52] B. Hamdaoui, P. Venkatraman, and M. Guizani, "Opportunistic exploitation of bandwidth resources through reinforcement learning," in *IEEE Global Telecommunications Conference (GLOBECOM '09)*, Honolulu, HI, Dec. 2009, pp. 1 –6.

[53] K.-L. A. Yau, P. Komisarczuk, and P. D. Teal, "Applications of reinforcement learning to cognitive radio networks," in *IEEE International Conference on Communications Workshops (ICC), 2010*, Cape Town, South Africa, May 2010, pp. 1 –6.

[54] 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.

[55] Y. Reddy, "Detecting primary signals for efficient utilization of spectrum using q-learning," in *Fifth International Conference on Information Technology: New Generations (ITNG '08)*, Las Vegas, NV, Apr. 2008, pp. 360 –365.

[56] M. Li, Y. Xu, and J. Hu, "A q-learning based sensing task selection scheme for cognitive radio networks," in *International Conference on Wireless Communications Signal Processing (WCSP '09)*, Nanjing, China, Nov. 2009, pp. 1 –5.

[57] J. Lunden, V. Koivunen, S. Kulkarni, and H. Poor, "Reinforcement learning based distributed multiagent sensing policy for cognitive radio networks," in *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '11)*, Aachen, Germany, May 2011, pp. 642 –646.

[58] Y. Yao and Z. Feng, "Centralized channel and power allocation for cognitive radio networks: A q-learning solution," in *Future Network and Mobile Summit*, Florence, Italy, June 2010, pp. 1 –8.

[59] P. Venkatraman, B. Hamdaoui, and M. Guizani, "Opportunistic bandwidth sharing through reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 6, pp. 3148 –3153, July 2010.

[60] T. Jiang, D. Grace, and P. Mitchell, "Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing," *IET Communications*, vol. 5, no. 10, pp. 1309 –1317, 1 2011.

[61] Z. Han, R. Zheng, and H. Poor, "Repeated auctions with bayesian nonparametric learning for spectrum access in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 10, no. 3, pp. 890 –900, Mar. 2011.

[62] T. Clancy, A. Khawar, and T. Newman, "Robust signal classification using unsupervised learning," *IEEE Transactions on Wireless Communications*, vol. 10, no. 4, pp. 1289 –1299, Apr. 2011.

[63] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 38, no. 2, pp. 156 –172, march 2008.

[64] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Madison, WI, Jul. 1998, pp. 746–752.

[65] G. D. Croon, M. F. V. Dartel, and E. O. Postma, "Evolutionary learning outperforms reinforcement learning on non-markovian tasks," in *Workshop on Memory and Learning Mechanisms in Autonomous Robots, 8th European Conference on Artificial Life*, Canterbury, Kent, UK, 2005.

[66] R. Sutton, D. Mcallester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proceedings of the 12th conference on Advances in Neural Information Processing Systems (NIPS '99)*. Denver, CO: MIT Press, 2001, pp. 1057–1063.

[67] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *Journal of Artificial Intelligence Research*, vol. 15, pp. 319–350, 2001.

[68] D. E. Moriarty, A. C. Schultz, and J. J. Grefenstette, "Evolutionary algorithms for reinforcement learning," *Journal of Artificial Intelligence Research*, vol. 11, pp. 241–276, 1999.

[69] F. Dandurand and T. Shultz, "Connectionist models of reinforcement, imitation, and instruction in learning to solve complex problems," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 2, pp. 110 –121, Aug. 2009.

[70] Y. Xing and R. Chandramouli, "Human behavior inspired cognitive radio network design," *IEEE Communications Magazine*, vol. 46, no. 12, pp. 122 –127, Dec. 2008.

[71] M. van der Schaar and F. Fu, "Spectrum access games and strategic learning in cognitive radio networks for delay-critical applications," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 720 –740, Apr. 2009.

[72] B. Wang, K. Ray Liu, and T. Clancy, "Evolutionary cooperative spectrum sensing game: how to collaborate?" *IEEE Transactions on Communications*, vol. 58, no. 3, pp. 890 –900, Mar. 2010.

[73] "Dictionary.com," 2011, retrieved Nov. 9, 2011 from http://www.dictionary.com.

[74] R. S. Michalski, "Learning and cognition," in *World Conference on the Fundamentals of Artificial Intelligence (WOCFAI '95)*, Paris, France, July 1995, pp. 507–510.

[75] S. K. Jayaweera and C. G. Christodoulou, "Radiobots: Architecture, algorithms and realtime reconfigurable antenna designs for autonomous, self-learning future cognitive radios," University of New Mexico, Technical Report EECE-TR-11-0001, Mar. 2011. [Online]. Available: http://repository.unm.edu/handle/1928/12306

[76] J. Burbank, A. Hammons, and S. Jones, "A common lexicon and design issues surrounding cognitive radio networks operating in the presence of jamming," in *IEEE Military Communications Conference (MILCOM '08)*, San Diego, CA, Nov. 2008, pp. 1 –7.

[77] E. L. Thorndike, "Animal intelligence: An experimental study of the associative processes in animals," Ph.D. dissertation, Columbia University, 1898.

[78] ——, *Animal Intelligence*. Hafner, Darien, CT, 1911.

[79] A. L. Samuel, "Some studies in machine learning using the game checkers," *IBM Journal on Research and Development*, vol. 3, pp. 211–229, 1959.

[80] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, 1992.

[81] C. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Cambridge University, 1989.

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

[83] L. Busoniu, R. Babuska, and B. De Schutter, "Multi-agent reinforcement learning: A survey," in *9th International Conference on Control, Automation, Robotics and Vision (ICARCV '06)*, Grand Hyatt, Singapore, Dec. 2006, pp. 1 –6.

[84] T. Darrell, "Reinforcement learning of active recognition behaviors," *Interval Research Technical Report 1997-045*, 1997.

[85] U. Berthold, F. Fu, M. van der Schaar, and F. Jondral, "Detection of spectral resources in cognitive radios using reinforcement learning," in *3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '08)*, Chicago, IL, Oct. 2008, pp. 1 –5.

[86] A. Ben Hadj Alaya-Feki, B. Sayrac, A. Le Cornec, and E. Moulines, "Semi dynamic parameter tuning for optimized opportunistic spectrum access," in *IEEE 68th Vehicular Technology Conference (VTC '08-Fall)*, Calgary, AB, Canada, Sep. 2008.

[87] 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, April 2007.

[88] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: John Wiley and Sons, 1994.

[89] S. Sanner and C. Boutilier, "Practical solution techniques for first-order MDPs," *Artificial Intelligence Journal (AIJ)*, vol. 173, no. 5-6, pp. 748–788, Mar. 2009.

[90] H. Li, "Multi-agent q-learning of channel selection in multi-user cognitive radio systems: A two by two case," in *IEEE International Conference on Systems, Man and Cybernetics (SMC '09)*, San Antonio, TX, Oct. 2009, pp. 1893 –1898.

[91] M. Bkassiny, S. K. Jayaweera, and K. A. Avery, “Distributed reinforcement learning based mac protocols for autonomous cognitive secondary users,” in *Wireless and Optical Communications Conference*, Newark, NJ, April 2011.

[92] C. Watkins, “Learning from delayed rewards,” Ph.D. dissertation, University of Cambridge, United Kingdom, 1989.

[93] B. Xia, M. Wahab, Y. Yang, Z. Fan, and M. Sooriyabandara, “Reinforcement learning based spectrum-aware routing in multi-hop cognitive radio networks,” in *4th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM ’09)*, Hannover, Germany, June 2009, pp. 1 –5.

[94] K.-L. Yau, P. Komisarczuk, and P. Teal, “Performance analysis of reinforcement learning for achieving context-awareness and intelligence in cognitive radio networks,” in *IEEE 34th Conference on Local Computer Networks (LCN ’09)*, Zurich, Switzerland, Oct. 2009, pp. 1046 –1053.

[95] E. W. Weisstein. Stochastic Approximation. From MathWorld–A Wolfram Web Resource. [Online]. Available: http://mathworld.wolfram.com/StochasticApproximation.html

[96] H. Robbins and S. Monro, “A stochastic approximation method,” *The Annuals of Mathematical Statistics*, vol. 2, pp. 400 –407, 1951.

[97] 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.

[98] R. D. Smallwood and E. J. Sondik, “The Optimal Control of Partially Observable Markov Processes over a Finite Horizon,” *Operations Research*, vol. 21, no. 5, pp. 1071 –1088, Sept.-Oct. 1973.

[99] G. E. Monahan, “A survey of partially observable markov decision processes: Theory, models, and algorithms,” *Management Science*, vol. 28, no. 1, pp. pp. 1–16, 1982. [Online]. Available: http://www.jstor.org/stable/2631070

[100] E. A. Hansen, “Dynamic programming for partially observable stochastic games,” in *IN PROCEEDINGS OF THE NINETEENTH NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, 2004, pp. 709–715.

[101] J. Unnikrishnan and V. 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.

[102] M. Bkassiny, S. K. Jayaweera, and K. A. Avery, “Distributed reinforcement learning based mac protocols for autonomous cognitive secondary users,” in *20th Annual Wireless and Optical Communications Conference (WOCC ’11)*, Newark, NJ, Apr. 2011, pp. 1 –6.

[103] 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.

[104] S. Jayaweera, M. Bkassiny, and K. Avery, “Asymmetric cooperative communications based spectrum leasing via auctions in cognitive radio networks,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 8, pp. 2716 –2724, Aug. 2011.

[105] B. Wang, Y. Wu, K. Liu, and T. Clancy, “An anti-jamming stochastic game for cognitive radio networks,” *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 877 –889, Apr. 2011.

[106] M. L. Littman, “Markov games as a framework for multi-agent reinforcement learning,” in *IN PROCEEDINGS OF THE ELEVENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING*. Morgan Kaufmann, 1994, pp. 157–163.

[107] D. Dey and P. Muller, *Practical Nonparametric and Semiparametric Bayesian Statistics*, D. Sinha, Ed. New York: Springer-Verlag, 1998.

[108] D. Freedman, “On the asymptotic behavior of bayes estimates in the discrete case,” *Annals of Mathematical Statistics*, vol. 34, pp. 615–629, 1963.

[109] T. Ferguson, “Prior distributions on spaces of probability measures,” *The Annals of Statistics*, vol. 2, pp. 615–629, 1974.

[110] ——, “A bayesian analysis of some nonparametric problems,” *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[111] D. Blackwell and J. MacQueen, “Ferguson distribution via polya urn schemes,” *The Annals of Statistics*, vol. 1, pp. 353–355, 1973.

[112] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, “Hierarchical dirichlet processes,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.

[113] Y. W. Teh, “Dirichlet processes,” in *Encyclopedia of Machine Learning*. New York: Springer, 2007.

[114] M. Jordan. (2005) Dirichlet processes, Chinese restaurant processes and all that. [Online]. Available: http://www.cs.berkeley.edu/ jordan/nips-tutorial05.ps

[115] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of State Calculations by Fast Computing Machines,” *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953. [Online]. Available: http://dx.doi.org/10.1063/1.1699114

[116] M. D. Escobar, “Estimating normal means with a dirichlet process prior,” *Journal of the American Statistical Association*, vol. 89, no. 425, pp. 268–277, Mar. 1994. [Online]. Available: http://www.jstor.org/stable/2291223

[117] N. Tawara, S. Watanabe, T. Ogawa, and T. Kobayashi, “Speaker clustering based on utterance-oriented dirichlet process mixture model,” in *12th Annual Conference of the International Speech Communication Association (ISCA ’11)*, Florence, Italy, Aug. 2011, pp. 2905–2908.

[118] S. Geman and D. Geman, “Stochastic relaxation, gibbs distributions, and the bayesian restoration of images,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-6, no. 6, pp. 721 –741, nov. 1984.

[119] A. E. Gelfand and A. F. M. Smith, “Sampling-Based Approaches to Calculating Marginal Densities,” *Journal of the American Statistical Association*, vol. 85, no. 410, pp. 398–409, 1990. [Online]. Available: http://dx.doi.org/10.2307/2289776

[120] G. Casella and E. I. George, “Explaining the gibbs sampler,” *The American Statistician*, vol. 46, no. 3, pp. pp. 167–174, 1992. [Online]. Available: http://www.jstor.org/stable/2685208

[121] N. Shetty, S. Pollin, and P. Pawelczak, “Identifying spectrum usage by unknown systems using experiments in machine learning,” in *IEEE Wireless Communications and Networking Conference (WCNC ’09)*, Budapest, Hungary, Apr. 2009, pp. 1 –6.

[122] G. Yu, R. Huang, and Z. Wang, “Document clustering via dirichlet process mixture model with feature selection,” in *Proceedings of the 16th ACM SIGKDD International conference on Knowledge Discovery and Data mining (KDD ’10)*. New York, NY, USA: ACM, 2010, pp. 763–772. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835901

[123] D. Fugenberg and J. Tirole, *Game Theory*. MIT Press, 1991.

[124] P. Zhou, W. Yuan, W. Liu, and W. Cheng, “Joint power and rate control in cognitive radio networks: A game-theoretical approach,” in *Proc. IEEE International Conference on Communications (ICC’08)*, May 2008, pp. 3296 – 3301.

[125] A. R. Fattahi, F. Fu, M. V. D. Schaar, and F. Paganini, “Mechanism-based resource allocation for multimedia transmission over spectrum agile wireless networks,” *IEEE Journ. Select Areas Commun.*, vol. 3, no. 25, pp. 601–612, Apr. 2007.

[126] O. Ileri, D. Samardzija, and N. B. Mandayam, “Demand responsive pricing and competitive spectrum allocation via a spectrum server,” in *New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005. 2005 First IEEE International Symposium on*, Nov. 2005, pp. 194–202.

[127] Y. Zhao, S. Mao, J. Neel, and J. Reed, “Performance evaluation of cognitive radios: Metrics, utility functions, and methodology,” *Proceedings of the IEEE*, vol. 97, no. 4, pp. 642–659, April 2009.

[128] J. Neel, R. M. Buehrer, B. H. reed, and R. P. Gilles, "Game theoretic analysis of a network of cognitive radio," in *45th Midwest Symp. on Circuits and Systems*, vol. 3, Aug. 2002, pp. III–409–III–412.

[129] M. R. Musku and P. cotae, "Cognitive radio: Time domain spectrum allocation using game theory," in *IEEE Int. Conf. on System and Systems Engineering (SoSE)*, April 2007, pp. 1–6.

[130] W. Wang, Y. Cui, T. Peng, and W. Wang, "Noncooperative power control game with exponential pricing for cognitive radio network," in *IEEE 65th Vehicular Technology Conf. (VTC)-Spring*, April 2007, pp. 3125–3129.

[131] J. Li, D. Chen, W. Li, and J. Ma, "Multiuser power and channel aloocation algorithm in cognitive radio," in *Int. Conf. on Parallel Processing, (ICPP)*, Sept. 2007, pp. 72–72.

[132] Z. Ji and K. J. R. Liu, "Cognitive radios for dynamic spectrum access- dynamic spectrum sharing: A game theoretical overview," *IEEE Communications Magazine*, vol. 45, no. 5, pp. 88–94, May 2007.

[133] N. Nie and C. Comaniciu, "Adaptive channel allocation spectrum etiquette for cognitive radio networks," in *1st IEEE Int. Symp. on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, Nov. 2005, pp. 269–278.

[134] R. G. Wendorf and H. Blum, "A channel-change game for multiple interfering cognitive wireless networks," in *Military Communi. Conf. (MILCOM)*, Oct. 2006, pp. 1–7.

[135] V. Krishnamurthy, "Decentralized specturm access amongst cognitive agents - An interacting multivariate global games approach," Nov. 2008, accepted.

[136] K. Akkarajitsakul, E. Hossain, D. Niyato, and D. I. Kim, "Game theoretic approaches for multiple access in wireless networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 13, no. 3, pp. 372 –395, quarter 2011.

[137] A. He, K. K. Bae, T. Newman, J. Gaeddert, K. Kim, R. Menon, L. Morales-Tirado, J. Neel, Y. Zhao, J. Reed, and W. Tranter, "A survey of artificial intelligence for cognitive radios," *Vehicular Technology, IEEE Transactions on*, vol. 59, no. 4, pp. 1578 –1592, may 2010.

[138] E. Rasmusen, "Games and information, fourth edition an introduction to game theory," *Quality*, no. November, p. 559, 2007. [Online]. Available: http://books.google.com/books?id=5XEMuJwnBmUC

[139] J. Li, D. Chen, W. Li, and J. Ma, "Multiuser power and channel allocation algorithm in cognitive radio," in *International Conference on Parallel Processing (ICPP '07)*, XiAn, China, Sep. 2007, p. 72.

[140] X. Zhang and J. Zhao, "Power control based on the asynchronous distributed pricing algorithm in cognitive radios," in *IEEE Youth Conference on Information Computing and Telecommunications (YC-ICT '10)*, Beijing, China, Nov. 2010, pp. 69 –72.

[141] L. Pillutla and V. Krishnamurthy, "Game theoretic rate adaptation for spectrum-overlay cognitive radio networks," in *IEEE Global Telecommunications Conference (GLOBECOM '08)*, New Orleans, LA, Dec. 2008, pp. 1 –5.

[142] H. Li, Y. Liu, and D. Zhang, "Dynamic spectrum access for cognitive radio systems with repeated games," in *IEEE International Conference on Wireless Communications, Networking and Information Security (WCNIS '10)*, Beijing, China, June 2010, pp. 59 –62.

[143] L. Chen, S. Iellamo, M. Coupechoux, and P. Godlewski, "An auction framework for spectrum allocation with interference constraint in cognitive radio networks," in *IEEE INFOCOM '10*, San Diego, CA, Mar. 2010, pp. 1 –9.

[144] G. Iosifidis and I. Koutsopoulos, "Challenges in auction theory driven spectrum management," *Communications Magazine, IEEE*, vol. 49, no. 8, pp. 128 –135, august 2011.

[145] L. S. Shapley, "Stochastic Games," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 39, no. 10, pp. 1095–1100, 1953. [Online]. Available: http://dx.doi.org/10.2307/88799

[146] F. Fu and M. van der Schaar, "Stochastic game formulation for cognitive radio networks," in *3rd IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN '08)*, Chicago, IL, Oct. 2008, pp. 1 –5.

[147] S. Gong, W. Liu, W. Yuan, W. Cheng, and S. Wang, "Threshold-learning in local spectrum sensing of cognitive radio," in *IEEE 69th Vehicular Technology Conference (VTC Sp. '09)*, Barcelona, Spain, Apr. 2009, pp. 1 –6.

[148] S. S. Haykin, *Neural networks : A Comprehensive Foundation*, 2nd ed. Prentice Hall, Jul. 1999.

[149] N. Baldo and M. Zorzi, "Learning and adaptation in cognitive radios using neural networks," in *Consumer Communications and Networking Conference, 2008. CCNC 2008. 5th IEEE*, jan. 2008, pp. 998 –1003.

[150] N. Baldo, B. Tamma, B. Manojt, R. Rao, and M. Zorzi, "A neural network based cognitive controller for dynamic channel selection," in *Communications, 2009. ICC '09. IEEE International Conference on*, june 2009, pp. 1 –5.

[151] Y.-J. Tang, Q.-Y. Zhang, and W. Lin, "Artificial neural network based spectrum sensing method for cognitive radio," in *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on*, sept. 2010, pp. 1 –4.

[152] V. Tumuluru, P. Wang, and D. Niyato, "A neural network based spectrum prediction scheme for cognitive radio," in *Communications (ICC), 2010 IEEE International Conference on*, may 2010, pp. 1 –5.

[153] M. I. Taj and M. Akil, "Cognitive radio spectrum evolution prediction using artificial neural networks based multivariate time series modeling," *Wireless Conference 2011 - Sustainable Wireless Technologies (European Wireless), 11th European*, pp. 1 –6, april 2011.

[154] J. Popoola and R. van Olst, "A novel modulation-sensing method," *Vehicular Technology Magazine, IEEE*, vol. 6, no. 3, pp. 60 –69, sept. 2011.

[155] J. Connor, R. Martin, and L. Atlas, "Recurrent neural networks and robust time series prediction," *Neural Networks, IEEE Transactions on*, vol. 5, no. 2, pp. 240 –254, mar 1994.

[156] M. Han, J. Xi, S. Xu, and F.-L. Yin, "Prediction of chaotic time series based on the recurrent predictor neural network," *Signal Processing, IEEE Transactions on*, vol. 52, no. 12, pp. 3409 – 3416, dec. 2004.

[157] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

[158] ——, *Statistical Learning Theory*. New York: Wiley, 1998.

[159] T. Atwood, "RF channel characterization for cognitive radio using support vector machines," Ph.D. dissertation, University of New Mexico, Nov. 2009.

[160] S. Abe, *Support Vector Machines for Pattern Classification*. London, UK: Springer-Verlag London Limited, 2005.

[161] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[162] G. Xu and Y. Lu, "Channel and modulation selection based on support vector machines for cognitive radio," in *Wireless Communications, Networking and Mobile Computing, 2006. WiCOM 2006.International Conference on*, sept. 2006, pp. 1 –4.

[163] H. Hu, J. Song, and Y. Wang, "Signal classification based on spectral correlation analysis and svm in cognitive radio," in *Advanced Information Networking and Applications, 2008. AINA 2008. 22nd International Conference on*, march 2008, pp. 883 –887.

[164] L. Hai-Yuan and J.-C. Sun, "A modulation type recognition method using wavelet support vector machines," in *Image and Signal Processing, 2009. CISP '09. 2nd International Congress on*, oct. 2009, pp. 1 –4.

[165] Z. Yang, Y.-D. Yao, S. Chen, H. He, and D. Zheng, "Mac protocol classification in a cognitive radio network," in *Wireless and Optical Communications Conference (WOCC), 2010 19th Annual*, may 2010, pp. 1 –5.

[166] M. Petrova, P. Ma andho andnen, and A. Osuna, "Multi-class classification of analog and digital signals in cognitive radios using support vector machines," in *Wireless Communication Systems (ISWCS), 2010 7th International Symposium on*, sept. 2010, pp. 986 –990.

[167] D. Zhang and X. Zhai, "Svm-based spectrum sensing in cognitive radio," in *Wireless Communications, Networking and Mobile Computing (WiCOM), 2011 7th International Conference on*, sept. 2011, pp. 1 –4.

[168] T. D. Atwood, M. Martnez-Ramon, and C. G. Christodoulou, "Robust support vector machine spectrum estimation in cognitive radio," in *Proceedings of the 2009 IEEE International Symposium on Antennas and Propagation and USNC/URSI National Radio Science Meeting*, 2009.

[169] Z. Sun, G. Bradford, and J. Laneman, "Sequence detection algorithms for phy-layer sensing in dynamic spectrum access networks," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 1, pp. 97 –109, feb. 2011.

[170] D. Cabric, "Addressing feasibility of cognitive radios," *Signal Processing Magazine, IEEE*, vol. 25, no. 6, pp. 85 –93, november 2008.

[171] Z. Han, R. Fan, and H. Jiang, "Replacement of spectrum sensing in cognitive radio," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 6, pp. 2819 –2826, june 2009.

[172] S. Jha, U. Phuyal, M. Rashid, and V. Bhargava, "Design of omc-mac: An opportunistic multi-channel mac with qos provisioning for distributed cognitive radio networks," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 10, pp. 3414 –3425, october 2011.

[173] Y. Li, S. Jayaweera, M. Bkassiny, and K. Avery, "Optimal myopic sensing and dynamic spectrum access in centralized secondary cognitive radio networks with low-complexity implementations," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, may 2011, pp. 1 –5.

[174] 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 Vehicular Technology Conference (VTC-spring'2011)*, Budapest, Hungary, May 2011, accepted.

[175] B. Wang, K. Liu, and T. Clancy, "Evolutionary game framework for behavior dynamics in cooperative spectrum sensing," in *IEEE Global Telecommunications Conference (IEEE GLOBECOM '08)*, Dec. 2008, pp. 1 –5.

[176] E. C. Y. Peh, Y.-C. Liang, Y. L. Guan, and Y. Zeng, "Power control in cognitive radios under cooperative and non-cooperative spectrum sensing," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 12, pp. 4238 –4248, december 2011.

[177] M. van der Schaar and F. Fu, "Spectrum access games and strategic learning in cognitive radio networks for delay-critical applications," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 720 –740, april 2009.

[178] M. Maskery, V. Krishnamurthy, and Q. Zhao, "Decentralized dynamic spectrum access for cognitive radios: cooperative design of a non-cooperative game," *Communications, IEEE Transactions on*, vol. 57, no. 2, pp. 459 –469, february 2009.

[179] L. Chen, S. Iellamo, M. Coupechoux, and P. Godlewski, "An auction framework for spectrum allocation with interference constraint in cognitive radio networks," in *INFOCOM, 2010 Proceedings IEEE*, march 2010, pp. 1 –9.

[180] M. Haddad, S. Elayoubi, E. Altman, and Z. Altman, "A hybrid approach for radio resource management in heterogeneous cognitive networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 831 –842, Apr. 2011.

[181] S. Buljore, M. Muck, P. Martigne, P. Houze, H. Harada, K. Ishizu, O. Holland, A. Mikhailovic, K. A. Tsagkariss,

O. Sallent, M. S. G. Clemo, V. Ivanov, K. Nolte, and M. Stametalos, “Introduction to IEEE p1900.4 activities,” *IEICE Transactions on Communications*, vol. E91-B, no. 1, 2008.