Opportunistic Bandwidth Sharing Through Reinforcement Learning

Pavithra Venkatraman, Bechir Hamdaoui, and Mohsen Guizani

ABSTRACT

As an initial step towards solving the spectrum shortage problem, FCC opens up for the so-called *opportunistic spectrum access* (OSA), which allows unlicensed users to exploit unused licensed spectrum, but in a manner that limits interference to licensed users. Fortunately, technological advances enabled cognitive radios, which have recently been recognized as the key enabling technology for realizing OSA. In this work, we propose a machine learning-based scheme that will exploit the cognitive radios' capabilities to enable effective OSA, thus improving the efficiency of spectrum utilization. Our proposed learning technique does not require prior knowledge of the environment's characteristics and dynamics, yet can still achieve high performances by learning from interaction with the environment.

I. INTRODUCTION

FCC's long term vision for solving the spectrum shortage problem [1, 2] is to promote the so-called *opportunistic spectrum access* (OSA), which allows unlicensed users (or *secondary users* (SUs)) to exploit unused licensed spectrum on an instant-by-instant basis, but in a manner that limits interference to licensed users (or *primary users* (PUs)) so as to maintain compatibility with legacy systems. The apparent promise of OSA has indeed created significant research interests, resulting in numerous research work ranging from protocol design [3–5] to performance optimization [6, 7], and from market-oriented access strategies [8, 9] to new management and architecture paradigms [10–13]. More recently, some work effort has also been given to the development of adaptive, learning-based approaches [14–26]. Zhao et al. [26] develops a model for predicting the dynamics of the OSA environment when periodic channel sensing is used. A simple two-state Morkovian model is assumed for activities of PUs on each channel. Using this model, Zhao et al. derive an optimal access policy that can be used to maximize channel utilization while limiting interference to PUs. In [20], Unnikrishnan et al. propose a cooperative, channel selection and access policy for OSA systems under interference constraints. In this work, the PUs' activities are assumed to be stationary Markovian, and the Markovian statistics are assumed to be known to all SUs. A centralized approach is considered, where all cooperating secondary users report their observations to a decision center, which makes decision regarding when and which channels to sense and access at each time slot. In [22], the authors develop channel decision policies for two SUs in a two-channel OSA system. PUs' activities are modeled as a discrete-time Markov chains. Liu et al. [23] considers the case of multiple, non-cooperative SUs in OSA system where SUs are assumed not to exchange information among themselves. The occupancy of primary channels is modeled as an i.i.d. Bernoulli process, and OSA is formulated as a multi-armed bandit problem where agents are not cooperative with each others. Chen et al. [24, 25] develop a cross-layer optimal access strategy for OSA that integrates physical-layer's sensing with MAC-layer's sensing and access policy. They establish a separation principle, meaning that physical-layer's sensing and MAC-layer's access policy can be decoupled from MAC-layer's sensing without losing optimality. The developed framework assumes that spectrum occupancy of PUs also follows a discrete-time ON/OFF Markov process.

In most of these works, the models developed for deriving optimal channel selection policies assume that PUs' activities follow the Markovian process model. Although analytically tractable, Markovian process may not accurately model the dynamics of PUs' activities. In fact, the OSA environment has very unique characteristics that make it too difficult to construct models that predict its dynamics, and it is therefore important to develop techniques that can achieve approximately optimal behaviors without requiring models of the environment's dynamics. Indeed, reinforcement learning (RL) [27] is a foundational idea built on the basis of learning from interaction without requiring models of the environment's dynamics, yet can still achieve approximately optimal behaviors. With this in mind, we propose in this paper an RL scheme for OSA that enables efficient spectrum utilization. Simulation results show that our scheme achieves high throughput performance by intelligently locating and exploiting spectrum opportunities without requiring prior knowledge of the environment's characteristics.

The paper is organized as follows. In Section II, we state the OSA problem and discuss its requirements. In Section III, we present our RL framework for efficient OSA. In Section IV, we evaluate the proposed approach. Finally, we conclude the paper in Section V.

II. PROBLEM STATEMENT

We assume that the spectrum is divided into m non-overlapping bands, and that each band is associated with a set of PUs. We denote η_j the primary-user traffic load on band b_j. In OSA, an agent is a group of two or more SUs who want to communicate together. We assume that all SUs are associated with a *home* band to which they have usage rights at all time. In order to communicate with each other, all SUs in the group must be tuned to the same band, being either their home band or another unused licensed band. While communicating on the home band, the secondary-user group may decide to seek for spectrum opportunities in another band. This typically happens when, for example, any of the SUs judge that the quality of their current band is no longer acceptable. This can be done by continuously assessing and monitoring the quality of the band via some quality metrics, such as signal-to-noise ratio (SNR), packet success rate, achievable data rate, etc. That is, when the monitored quality metric drops below a threshold that can be defined *a priori*, the secondary-user group is triggered to start seeking for spectrum opportunities. When a new opportunity is discovered on another band, the group switches to that band and starts communicating on it. Now suppose the group is currently using a licensed band, not the home band. Then, upon the return of PUs to their band and/or when the quality drops below the threshold, SUs must vacate the licensed band by either switching back to their home band or by searching for new opportunities. Hereafter,

P. Venkatraman and B. Hamdaoui are with the School of EECS at Oregon State University. Emails: venkatrp,hamdaoui@eecs.orst.edu. M. Guizani is with the CS Department at Western Michigan University. Email: mguizani@ieee.org.

we say that an *exploration event* is triggered when either (i) PUs return back to their licensed band, and/or (ii) the band's quality is degraded below the threshold. In the RL terminology, we therefore consider that the agent and the environment interact at each of a sequence of discrete time steps, each of which takes place at the occurrence of an exploration event.

III. RL for Opportunistic Spectrum Access

A. Markov Decision Process (MDP)

We formulate OSA as a finite MDP, defined by its state set $\mathcal{S}$, action set $\mathcal{A}$, transition function δ, and reward function r as follows:

State set. $\mathcal{S}$ consists of $m+1$ states, $\{s_0, s_1, \ldots, s_m\}$. The secondary-user group is said to be in state s_i when it is using band b_i at the current time step; i.e., no PUs are currently using band b_i. Note that state s_0 corresponds to when the group is communicating on its home band b_0. Throughout this section, the terms agent and secondary-user group will be used interchangeably to mean the same thing. The same also applies to the terms state and band.

Action set. At every time step (i.e., an exploration event), while in state s_i, the agent can either choose to *exploit* by switching back to its home band b_0, or choose to *explore* by searching for new spectrum opportunities. If a decision is made in favor of exploration, then the agent senses an ordered sequence of bands $\{b_{k_1}, b_{k_2}, \ldots, b_{k_n}\}$, where $n = 1, 2, \ldots, m$, on a one-by-one basis until it finds, if any, the first available band. If there is one available, the agent switches to and starts using it until the the next time step. If none are available, then the agent switches back to b_0 at the end of the search. At the next time step, the same exploration vs. exploitation process repeats again. We will refer to n as the exploration index as it balances between exploration and exploitation; i.e., the larger the n, the more the exploration. Now by letting a_0 denote the action of returning to the home band b_0, and $a_k = \{b_{k_1}, b_{k_2}, \ldots, b_{k_n}, b_0\}$ the action of exploring new opportunities, the set $\mathcal{A}$ of all actions is $\mathcal{A} = \{a_0, a_1, \ldots, a_p\}$, where $p = \frac{m!}{(m-n)!}$. The index n can be viewed as a design parameter to be set *a priori*.

Transition function. $\delta : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the transition function, specifying the next state the system enters provided its current state and the action to be performed. Given any state, s_j, for action a_0, the transition function $\delta(s_j, a_0)$ equals s_0, and for any action $a_k = \{b_{k_1}, b_{k_2}, \ldots, b_{k_n}, b_0\}$, $k = 1, 2, \ldots, p$, the transition function $\delta(s_j, a_k)$ equals

$$\delta(s_j, a_k) = \begin{cases} s_0 & \text{w/ prob. } \prod_{i=1}^{n} \eta_{k_i} \\ s_{k_1} & \text{w/ prob. } 1 - \eta_{k_1} \\ s_{k_l} & \text{w/ prob. } \prod_{i=1}^{l-1} \eta_{k_i}(1 - \eta_{k_l}) \\ & \text{for } l = 2, 3, \ldots, n \end{cases}$$

For example, when $n = 2$, and the secondary user is in state s_j. If action $a_k = \{b_2, b_3, b_0\}$ is taken, then the user ends up in state s_2 (i.e., band b_2) with probability $1 - \eta_2$ (i.e., b_2 is available), ends up in state s_3 (i.e., band b_3) with probability $\eta_2(1 - \eta_3)$ (i.e., b_2 is occupied and b_3 is not), or ends up in state s_0 (i.e., band b_0) with probability $\eta_2\eta_3$ (i.e., both bands are not available).

It is important to reiterate that this function is only provided to generate samples of the OSA environment so as to evaluate our RL algorithm. That is, although in practice our RL technique will not need models to perform, we use models here to generate samples of the environment's behavior to mimic an OSA environment. For example, in the evaluation section, it is assumed that the primary user traffic follows a Poisson distribution, and hence, an ON/OFF renewal process model is used to mimic such an environment.

Reward function. $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ defines the reward function $r(s_i, a_k)$, specifying the reward the agent earns when taking action $a_k \in \mathcal{A}$ while in state $s_i \in \mathcal{S}$. The reward $r(s_i, a_k)$ also depends on the next state $s_j = \delta(s_i, a_k)$ the agent enters as a result of taking a_k while in state s_i. More specifically, the reward perceived by the agent when entering state s_j is a function of the quality level the secondary-user group receives when using the band it ends up selecting. We therefore assume that each band b_j is associated with a quality level q_j, which can be determined via metrics like SNR, packet success rate, data rates, etc, and let $\phi(q_j)$ denote the reward (without including the cost of exploration yet) resulting from receiving q_j.

It is important to note that exploration also comes with a price. Recall that secondary users are allowed to use any licensed band only if the band is vacant (no primary users are using it), and that discovery of opportunities is done through spectrum sensing. That is, secondary users periodically (or proactively) switch to and sense certain bands to find out whether any of them is vacant or not. Unfortunately, during the sensing process, the system incurs some "sensing overhead", which can be of multiple types: energy consumed to perform sensing, delays resulting from switching across bands, throughput reduced as a result of ceasing communication, etc. By letting c_{ij} denote the cost incurred as a result of exploring band b_j while in band b_i, and s_j denote the next state, $\delta(s_i, a_k)$, the reward function $r(s_i, a_k)$ can now be written as

$$r(s_i, a_k) = \begin{cases} \phi(q_{k_1}) - c_{ik_1} \\ \qquad \text{w/ prob. } 1 - \eta_{k_1} \\ \phi(q_{k_l}) - c_{ik_1} - \sum_{t=1}^{l-1} c_{k_t k_{t+1}}, l = 2, 3, \ldots, n \\ \qquad \text{w/ prob. } \prod_{t=1}^{l-1} \eta_{k_t}(1 - \eta_{k_l}) \\ -c_{ik_1} - \sum_{t=1}^{n-1} c_{k_t k_{t+1}} - c_{k_n 0} \\ \qquad \text{w/ prob. } \prod_{t=1}^{n} \eta_{k_t} \end{cases}$$

where $a_k = \{b_{k_1}, b_{k_2}, \ldots, b_{k_n}, b_0\}$, $k = 1, 2, \ldots, p$.

Consider a special scenario where the primary-user traffic load is the same and equal to η for all bands b_j. Suppose that $\phi(q_j) = q$ for all bands b_j, and that the cost c_{ij} incurred when switching from band b_i to band b_j is equal to c for all i, j. Let $\bar{E}$ denote the expected value of the reward function $r(s_i, a_k)$ normalized with respect to c (i.e., $\bar{E} = E[r(s_i, a_k)]/c$). One can now express $\bar{E}$ as

$$\bar{E} = (\frac{q}{c} - 1)(1 - \eta) + \frac{q}{c}(\eta - \eta^n) + \frac{\eta^{n+1} - 2\eta + \eta^2}{1 - \eta} \quad (1)$$

Using Eq. (1), one can easily see that the reward that the agent receives increases monotonically with the exploration index n when $\frac{q}{c} > \frac{\eta}{1-\eta}$ (or equivalently $\eta < \frac{q}{q+c}$), decreases monotonically with the index n when $\frac{q}{c} < \frac{\eta}{1-\eta}$ (or equivalently $\eta > \frac{q}{q+c}$), and is independent of the index n when $\frac{q}{c} = \frac{\eta}{1-\eta}$ (or equivalently $\eta = \frac{q}{q+c}$). Therefore, for a given primary-user traffic load, the optimal exploration index n that the agent should use so as to maximize its reward depends on the ratio q/c (or equivalently $\frac{q}{q+c}$).

Intuitively, when the network is lightly loaded (η is small), the chances of finding available bands are high, and hence, it is rewarding to explore for more bands. This explains why for small η values (i.e., $\eta < \frac{q}{q+c}$), the higher the exploration index, the higher the reward. Now when the network is heavily loaded (η is large), the chances of finding empty bands are low, and hence, it is not

rewarding to explore for more bands. This explains why for high values of η (i.e., $\eta > \frac{q}{q+c}$), the lower the exploration index, the higher the reward. That is, the expected reward is not worth the exploration cost for high values of η. Note that as the cost c goes to zero, $\frac{q}{q+c}$ goes to 1. Therefore, when the cost is negligible, $\eta < \frac{q}{q+c}$ holds for all η since $\frac{q}{q+c} \approx 1$, and thus, the reward increases monotonically with the exploration index n regardless of the primary-user load η.

B. Learning-Based OSA Scheme

The goal of the agent is to learn a policy, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, for choosing the next action a_i based on its current state s_i that produces the greatest possible expected cumulative reward. A cumulative reward R is typically defined through a discount factor γ, $0 \leq \gamma < 1$, as $\sum_{t=0}^{\infty} \gamma^t r(s_{i+t}, a_{i+t})$. Because it is naturally desirable to receive rewards sooner than later, the reward is expressed in a way that future rewards are discounted with respect to immediate rewards.

A function, $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, is defined for each state-action (s_i, a_k) pair as the maximum discounted cumulative reward that can be achieved when starting from state s_i and taking action a_k according to the optimal policy. Hence, given the Q-function, it is possible to act optimally by selecting actions that maximize $Q(s_i, a_k)$ at each state. Q can be constructed recursively as follows. The Q-learning algorithm learns an estimate $\hat{Q}$ of the optimal Q-function by selecting actions and observing their effects. In particular, each step in the environment involves taking an action a_k in state s_i and then observing the following state and the resulting reward. Given this information, Q is updated via the following equation:

$$\begin{aligned}\hat{Q}(s_i, a_k) \leftarrow \quad & (1-\alpha_l)\hat{Q}(s_i, a_k) + \alpha_l\{E[r(s_i, a_k)] \\ & +\gamma \max_{k'} \hat{Q}(\delta(s_i, s_k), a_{k'})\}\end{aligned}$$

where $\alpha_l = 1/(1 + visits_l(s_i, a_k))$ and $visits_l(s_i, a_k)$ is the total number of times this state-action pair has been visited up to and including the lth iteration. This approximation algorithm is guaranteed to converge to the optimal Q-function in any MDP given the appropriate exploration during learning [27].

IV. Evaluation of the Proposed Approach

In this section, we study the proposed Q-learning scheme by evaluating and comparing its performance to a random access scheme. The random scheme will be used here as a baseline for comparison, and is defined as follows. Whenever an exploration event is triggered, the secondary-user group, using the random access approach, selects a spectrum band among all bands randomly. If the selected band is idle, then the group uses it until the return of a primary user. Otherwise, i.e., if the selected band happens to be busy, then the group goes back to its home band. This process repeats until an idle band is found.

A. Simulation Settings

We consider that the spectrum is divided into m non-overlapping bands, and that each band is associated with a set of primary users. We model primary users' activities on each band as a renewal process alternating between ON and OFF periods, which represent the time during which primary users are respectively present (ON) and absent (OFF). For each spectrum band b_j, we assume that ON and OFF periods are exponentially distributed with rates λ_j and μ_j, respectively. Note that the primary traffic load η_j on band b_j can be expressed as $\mu_j/(\mu_j + \lambda_j)$. Recall that the power of RL lies in its capability to converge to approximately an optimal behavior without needing prior knowledge of primary users' traffic behavior. The exponential distributions will, however, be used to generate samples so as to be able to evaluate our learning techniques using simulated interaction. Throughout this section, we characterize the primary-user traffic system load by $\bar{\eta} = \frac{1}{m}\sum_{i=1}^{m} \eta_i$ (denoted as `pbar` in figures) and $CoV = \sigma/\bar{\eta}$, which respectively denote the average and the coefficient of variation of primary-user traffic loads across all bands, where σ denotes the standard deviation of traffic loads.

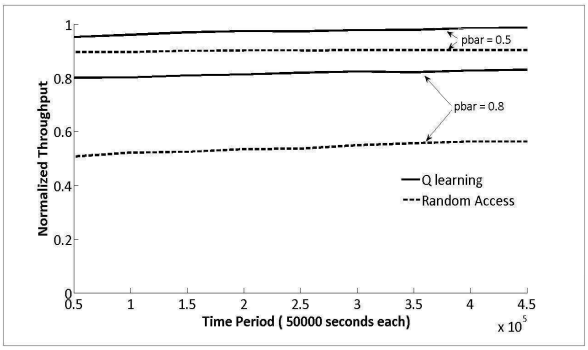

Fig. 1. Throughput behavior under two different primary-user traffic loads, pbar $\equiv \bar{\eta} = 0.5$ and 0.8, for $m = 7$ and $CoV = 0.5$

B. Effect of Primary-User Traffic Load

We begin by studying the effect of primary-user traffic load $\bar{\eta}$ on the achievable throughput. Fig. 1 plots the total throughput, normalized w.r.t. the maximal achievable throughput[1], that the secondary-user group achieves as a result of using our Q-learning and the random access schemes for two different primary-user traffic loads: $\bar{\eta} = 0.5$ and $\bar{\eta} = 0.8$. The measured throughput is based on what the secondary-user group receives from the m licensed bands only; i.e., not accounting for the home band. In this simulation scenario, CoV is set to 0.5, exploration index n is set to 3, and the total number of bands m is set to 7. First, as expected, note that the higher the $\bar{\eta}$, the lesser the achievable throughput under both schemes. However, regardless of the primary-user load, the Q-learning scheme always outperforms the random scheme. Also, note that the more loaded the system is, the higher the difference between the throughput achievable under Q-learning and that achievable under random access (e.g., the throughput gain is higher when $\bar{\eta} = 0.8$).

To further illustrate the effect of $\bar{\eta}$ on the performance of the proposed Q-learning scheme, we plot in Fig. 2 the throughput gain as a function of $\bar{\eta}$. Note that the throughput gain increases as the primary-user traffic load increases. In other words, the Q-learning scheme performs even better under heavily loaded systems. This can be explained as follows. When $\bar{\eta}$ is high; i.e., when spectrum opportunities are scarce, the learning capability of the Q-learning scheme allows the OSA agent to efficiently locate where the opportunities are, whereas random access leads to less throughput since it is accessing bands randomly. When $\bar{\eta}$ is small, on the other hand, the random access scheme is able to achieve high throughput since spectrum opportunities are too many to miss even when bands are selected unintelligently.

[1]The maximal/ideal achievable throughput corresponds to when the agent knows exactly where spectrum opportunities are; i.e., the agent always knows which bands are available, and thus, it exploits them without any cost.

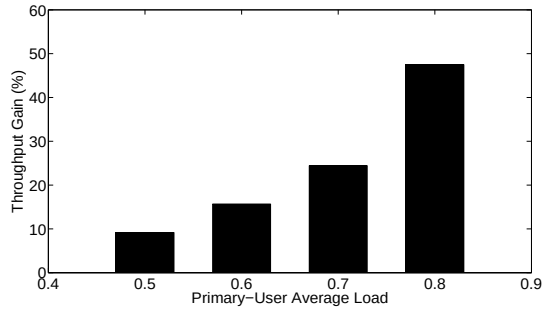

Fig. 2. Throughput gain as a function of the primary-user average loads, $\bar{\eta}$, for $m = 7$ and $CoV = 0.5$

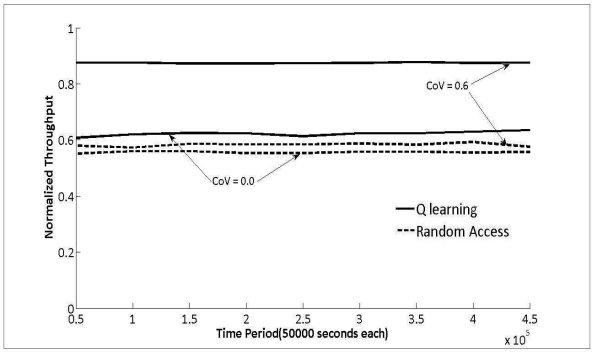

Fig. 3. Achievable throughput under Q-learning and random access schemes: $\bar{\eta} = 0.8$, $m = 7$, $n = 3$.

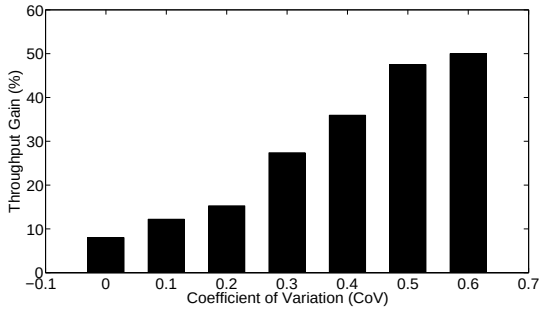

Fig. 4. Throughput gain as a function of primary-user load variability: $\bar{\eta} = 0.8$, $CoV = 0.2$, $m = 7$, $n = 3$.

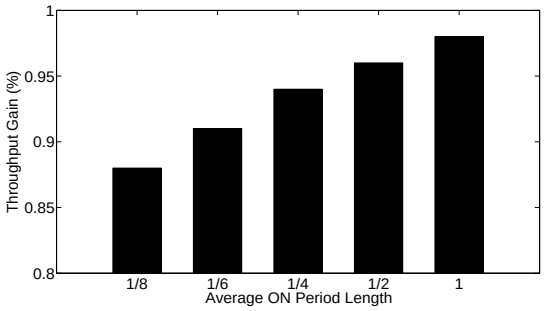

Fig. 5. Throughput gain as a function of ON/OFF period lengths: $\bar{\eta} = 0.8$, $CoV = 0.5$, $m = 7$, $n = 3$.

To summarize, these obtained results show that the proposed Q-learning scheme is capable of achieving between 80% to 95% of the maximal achievable throughput by learning from experience, and without prior knowledge of the environment. The results also show that the scheme achieves high throughput performance even under heavy traffic loads.

C. Effect of Primary-User Load Variability

Fig. 3 plots the total throughput that the secondary-user group achieves under our proposed Q-learning and the random access schemes for two different primary-user load variations: $CoV = 0$ and $CoV = 0.6$. (Recall that CoV reflects the variation of loads across different bands; i.e., the higher the CoV, the higher the variation.) Note that when the $CoV = 0.6$, the Q-learning scheme achieves about 90% of the maximal/ideal throughput by simply locating and exploiting unused opportunities through learning from experience, whereas the random access scheme achieves only about 60%. When $CoV = 0$ (i.e., all bands experience identical loads), the Q-learning and the random access achieve approximately about 64% and 55%, respectively. As expected, the throughput gain increases with the coefficient of variation. That is, and as shown in Fig. 3, the gain is higher when $CoV = 0.6$ than when $CoV = 0$. More insights on this are provided in the next paragraph.

To further illustrate the effect of primary-user load variability on the achievable throughput, we show in Fig. 4 the throughput gain for different values of CoVs. The CoV is varied from 0 to 0.6. The average primary-user traffic load, $\bar{\eta}$, is set to 0.8 (which implies that only 20% of the spectrum is available for the secondary-user group). The total number of bands is set to $m = 7$ and the exploration index is taken to be $n = 3$. Observe that the higher the variation of primary-user loads across different bands, the higher the throughput gain; i.e., the higher the throughput the agent/group can achieve when compared with that achievable under the random access scheme. This can be explained as follows. When the average of primary-user traffic loads is kept the same, a high variation in the loads across different bands increases the likelihood of finding highly available spectrum bands. This, on the other hand, also increases the likelihood of finding spectrum bands with less opportunities. With experience, the Q-learning scheme learns about, and starts exploiting, these more available bands, yielding then more throughput. When the load variation is low, on the other hand, the learning algorithm achieves less throughput because all bands are equally-loaded, and hence, there is no special (i.e., more available) bands that the algorithm can learn about. This explains why both the Q-learning and the random access achieve similar performances when all bands have identical loads. The gain can, however, reach up to 50% when bands have different loads (e.g., $CoV = 0.6$), as shown in Fig. 4.

D. Effect of Primary-User Load ON/OFF Period

In this section, we study the effect of ON/OFF period lengths on the performance of the Q-learning scheme. We vary the lengths of ON and OFF periods while keeping the primary-user traffic loads, η_i, the same for all i. Since the primary-user load is kept the same, an increase in OFF periods leads to an increase in ON periods as well, and vice versa. The normalized throughput that the Q-learning scheme achieves is shown in Fig. 5 for different values of ON period lengths. Here, CoV is set to 0.2, $\bar{\eta}$ is set to 0.5, n is set to 3, and m is set to 7.

Note that the higher the length of ON/OFF periods, the higher the throughput gain. Note also that having short ON/OFF periods forces the agent to make frequent transitions so as to find available spectrum bands. Whereas, when ON/OFF periods are long, the transitions are not that often, thus leading to less switching overhead, which yields more achievable throughput. Put differently, when the length of ON/OFF periods increases, the secondary-user group can possess available spectrum bands for longer periods of time. When

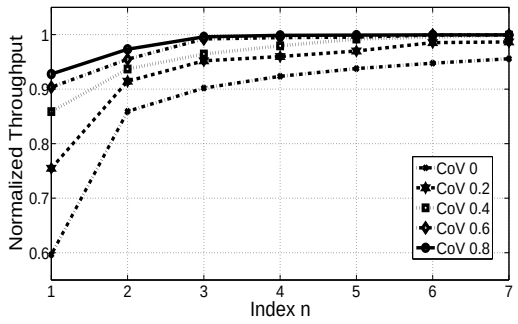

Fig. 6. Effect of index n on throughput: $\bar{\eta} = 0.8$, $m = 7$.

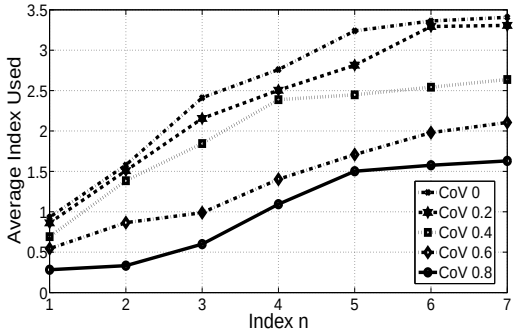

Fig. 7. Index used as a function of index n: $\bar{\eta} = 0.8$, $m = 7$.

the lengths of ON/OFF periods are low, the secondary-user group has the spectrum band available to it only for a short period of time, leading to frequent transitions across different bands.

E. Q-learning Optimality: Exploration Index n

In this section, we study the effect of the exploration index n on the behavior of the Q-learning scheme. Recall that the index n is a design parameter to be chosen and set *a priori*, which can take on any number less than or equal to the number of available bands m. This parameter balances between two conflicting objectives: the desire of increasing the chances of finding available bands (i.e., by increasing n), and the desire to reduce the incurred overhead/cost due to scanning (i.e., by decreasing n).

Fig. 6 plots the normalized throughput as a function of n for different values of CoV. Note that as the index n increases, the achievable throughput first increases with n, then flattens out. This means that increasing the number of scanned/searched bands beyond a certain threshold does not necessarily yield more achievable throughput.

To further study this behavior, for each index n scenario, we measured the average number of bands that are actually scanned before finding one available band. We refer to this number as *average index used*. Fig. 7 shows the average index used for finding available bands as a function of the exploration index n for different values of CoV. Note that as n increases, the average index used to find an available band first increases then flattens out. This means that even when the secondary-user group is allowed to scan all bands, it ends up visiting only a few before finding an available one as a result of using its learning capabilities. The figure also shows that the higher the CoV, the smaller the actual index used to find an available band. Therefore, the learning capabilities allow to find spectrum opportunities quickly, thus limiting the incurred exploration overhead.

V. Conclusion

Technological advances enabled cognitive radios, which have recently been recognized as the key technology for realizing OSA. Cognitive radios are viewed as intelligent systems that can self-learn from their surrounding environments, and auto-adapt their operating parameters in real-time to improve spectrum efficiency. In this paper, we developed a reinforcement learning-based framework that exploits the cognitive radios' capabilities to enable effective OSA, thus improving the efficiency of spectrum utilization. The proposed learning technique does not require prior knowledge of the environment's characteristics and dynamics, yet can still achieve high performance by learning from interaction with the environment.

References

[1] *FCC, Spectrum Policy Task Force (SPTF), Report of the Spectrum Efficiency WG, Report ET Docet no. 02-135, November*, 2002.

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

[3] A. Ghasemi and E. S. Sousa, "Interference aggregation in spectrum-sensing cognitive wireless networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 41–56, Februray 2008.

[4] Z. Quan, S. Cui, and A. H. Sayed, "Optimal linear cooperation for spectrum sensing in cognitive radio networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 28–40, Februray 2008.

[5] H. Su and X. Zhang, "Cross-layer based oppotrunitic MAC protocols for QoS provisionings over cognitive radio wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 118–129, January 2008.

[6] C.-T. Chou, S. Shankar, H. Kim, and K. G. Shin, "What and how much to gain by spectrum agility," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 3, pp. 576–588, April 2007.

[7] S. Srinivasa and S. A. Jafar, "Cognitive radio networks: how much spectrum sharing is optimal?," in *Proceedings of IEEE GLOBECOM*, November 2007, pp. 3149–3153.

[8] Z. Ji and K. J. R. Liu, "Belief-assisted pricing for dynamic spectrum allocation in wireless networks with selfish users," in *Proceedings of IEEE SECON*, September 2006, vol. 1, pp. 119–127.

[9] Z. Ji and K. J. R. Liu, "Multi-stage pricing game for collusion-resistant dynamic spectrum allocation," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 182–191, January 2008.

[10] S. Delaere and P. Ballon, "Flexible spectrum management and the need for controlling entities for reconfigurable wireless systems," in *Proceedings of IEEE DySPAN*, April 2007, pp. 347–362.

[11] Y. T. Hou, Y. Shi, and H. D. Sherali, "Spectrum sharing for multi-hop networking with cognitive radios," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 1, pp. 146–155, January 2008.

[12] A. Ginsberg, J. D. Poston, and W. D. Horne, "Toward a cognitive radio architecture: intergrating knowledge representation with software defined radio technologies," in *Proceedings of IEEE MILCOM*, October 2006, pp. 1–7.

[13] S. Yarkan and H. Arslan, "Exploiting location awareness toward improved wireless system design in cognitive radio," *IEEE Communications Magazine*, vol. 46, pp. 128–136, January 2008.

[14] Z. Han, C. Pandana, and K. J. R. Liu, "Distributive opportunistic spectrum access for cognitive radio using correlated equilibrium and no-regret learning," in *Proceedings of IEEE WCNC*, March 2007, pp. 11–15.

[15] K. E. Nolan, P. Sutton, and L. E. Doyle, "An encapsulation for reasoning, learning, knowledge representation, and reconfiguration cognitive radio elements," in *Proceedings of Int'l Conference on Cognitive Radio Oriented Wireless Networks and Communications*, June 2006.

[16] H. Kim and K. G. Shin, "Fast discovery of spectrum oppotunities in cognitive radio networks," in *Proceedings of IEEE DySPAN*, October 2008, pp. 1–12.

[17] H. Kim and K. G. Shin, "Efficient discovery of spectrum oppotunities with MAC-layer sensing in cognitive radio networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 5, pp. 533–545, May 2008.

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

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

[20] J. Unnikrishnan and V. V. Veeravalli, "Dynamic spectrum access with learning for cognitive radio," in *Proc. of Asilomar Conference on Signals Systems and Computers*, Oct. 2008.

[21] J. Unnikrishnan and V. 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, February 2008.

[22] H. Liu, B. Krishnamachari, and Q. Zhao, "Cooperation and learning in multiuser opportunistic spectrum access," in *Proceedings of IEEE ICC*, 2008.

[23] K. Liu and Q. Zhao, "Distributed learning in cognitive radio networks: multi-armed brandit with distributed multiple players," in *Submitted to IEEE Int. Conf. on Acousitcs, Speech, and Signal Processing*, 2010.

[24] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access," in *Proceedings of the SPIE Conf. on Advanced Signal Processing Algorithms, Architectures, and Implementations*, August 2007.

[25] 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.

[26] Q. Zhao, S. Geirhofer, L. Tong, and B. M. Sadler, "Opportunistic spectrum access via periodic channel sensing," *IEEE Transactions on Signal Processing*, vol. 2, no. 56, pp. 459–469, February 2008.

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