Replicated Q-learning Based Sub-band Selection for Wideband Spectrum Sensing in Cognitive Radios

Mohamed A. Aref*, Stephen Machuzak*, Sudharman K. Jayaweera* and Steven Lane†
*Communications and Information Sciences Laboratory (CISL)
Department of Electrical and Computer Engineering, University of New Mexico
Albuquerque, NM 87131-0001, USA
†Air Force Research Laboratory, Kirtland Air Force Base, Albuquerque, NM
Email: *{maref, smachuzak29, jayaweera}@unm.edu, †steven.lane.1@us.af.mil

***Abstract*—Spectrum sensing is a key basic function in any wideband cognitive radio (CR) for detecting the presence of any spectral activities. However, due to hardware constraints, the instantaneous sensing bandwidth is limited to a single sub-band out of all sub-bands in the spectrum of interest. Hence, sub-band selection is an important step in wideband spectrum sensing. In this paper we develop a partially observable Markov decision process (POMDP) to model the sub-band dynamics and propose an efficient sub-band selection policy based on replicated Q-learning. It is shown through simulations that the proposed selection policy has reasonably low computational complexity and significantly outperforms the random sub-band selection policy.**

***Index terms*—** Cognitive radios, wide-band spectrum scanning, sub-band selection, partially observable Markov decision processes, Q-learning, replicated Q-learning.

I. Introduction

Cognitive radios (CRs) are widely believed to provide a promising solution to the problem of inefficient spectrum utilization driven by conventional static RF spectrum allocation schemes. CRs can achieve this via dynamic sharing of the limited RF spectrum resources [1]-[3]. However, this is only a narrow interpretation of capabilities of a CR. More generally, a CR can be equipped with learning and decision making abilities that may allow it to adapt to user needs or its RF environment. In order for a CR to operate in the best mode and detect spectrum opportunities, however, it must be able to observe and interpret its surrounding RF environment. This functionality is known as spectrum knowledge acquisition or spectrum awareness [4]. Besides self-learning and decision making, spectrum knowledge acquisition is a fundamental task for CRs to achieve the required awareness.

Spectrum knowledge acquisition consists of three stages as shown in Fig. 1 [4]. The first stage is the wideband spectrum scanning. Hardware constraints limit the instantaneous sensing bandwidth of most state-of-the-art software-defined radio (SDR) platforms to about 100MHz [5]. Hence, the challenge in this step is to design an efficient scheme to achieve real-time sensing over a wide spectrum range. After scanning a spectrum band of interest, the second step is to detect any spectrum activity present in a sensed spectrum sub-band [4]. The detection can be simply done by defining a threshold, where any power spectral activity above this threshold is considered as an active signal. In general, however, this may not provide sufficiently detailed information about the detected signal, as there may be many signals belonging to different radio systems. Thus, a third step of signal classification and identification may be needed after signal detection [4].

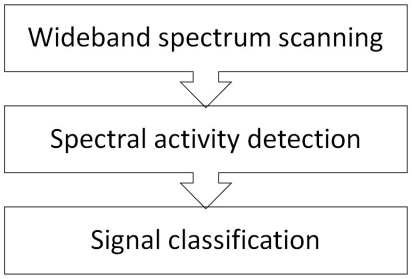

Fig. 1. Spectrum knowledge acquisition procedures.

The focus of this paper is on the first step of wideband spectrum scanning. In order to be able to scan a wide spectrum band in real-time,the spectrum of interest is first divided into a set of sub-bands. Each sub-band can be wide enough to contain multiple communication channels that possibly could belong to different systems. In this framework, the sub-band is the unit of the spectrum to be sensed and processed by a CR at any given time. Since the CR can sense only one sub-band at any given time, it needs to determine which one to be sensed at each time instant. This problem is known as sub-band selection problem in wideband spectrum sensing [4].

The sub-band selection problem can be considered as a decision making problem in which the system state can only be observed partially. Thus, if we assume that the underline system dynamics are Markov, the sub-band selection problem could be modeled as a partially observable Markov decision process (POMDP). However, with the definition of system state used, for example, in [4, 6] the number of states can become very large even for a few hundred MHz wide spectrum. This can lead to an impractically high computational demand in finding an optimal sub-band selection policy. Moreover, it needs to be computed in real time. This makes adaptive algorithms that are able to learn a decision policy from

the partially observable states an attractive alternative. Such algorithms are known as machine learning and in this paper, we specifically focus on a machine learning paradigm called reinforcement learning (RL) [7].

RL has been adopted in the literature [8]-[10] because it does not require prior knowledge of the operating environment and it is highly adaptive to the channels dynamics. One of the most commonly used RL approach is the Q-learning algorithm. In the case of a Markov decision process (MDP), Q-learning is known to converge to an optimal policy [7]. In [11], for example, a multi-agent Q-learning algorithm was proposed in order to enable a secondary user (SU) to select an available channel for data transmission. Di Felice et al. [12] proposed a joint dynamic channel selection (DCS) and channel sensing scheme using a set of distinctive Q-functions. In this scenario, the SU determines the durations of channel sensing and data transmission in order to enhance QoS, delay and packet delivery rate performance. Another joint DCS and channel sensing scheme was proposed in [13] where it was shown to improve the overall spectrum utilization. The goal of this scheme was to enable the SU agents to select their respective operating channels for sensing and data transmission in which the collisions among the SUs and primary users (PUs) must be minimized.

While most of the previous works use the Q-learning algorithm for the purpose of SU performance enhancements, whilst minimizing interference to PUs in a dynamic spectrum sharing (DSS) scenario, in this paper we consider the sub-band selection problem in a wideband spectrum sensing. Our scenario could be applicable for many real-life CR applications (e.g: jamming/anti-jamming systems). However, the sub-band selection problem addressed in this paper is a POMDP problem. As a result, the Q-learning algorithm is not directly applicable and thus, in this paper, we focus on one of the extensions of Q-learning, called the replicated Q-learning [15], to obtain an optimal policy for the sub-band selection POMDP.

The rest of the paper is organized as follow. Section II describes our assumed spectrum dynamics model. The POMDP model for the sub-band selection problem is described in Section III. Section IV discusses the implementation of the proposed replicated Q-learning algorithm. Simulation results are provided in Section V, followed by concluding remarks in Section VI.

II. Spectrum dynamics model

The wideband spectrum of interest can be considered as made of N_b sub-bands. Each sub-band may include a different number of communication channels. We denote by M_i the number of channels in the ith sub-band. In our model, we assume a semi-infinite slotted time horizon with each time slot having an equal time length of T seconds. Let $\mathcal{K} \in \{1, 2, ...\}$ denote the set of time slot indices. For simplicity, we assume the channel state to be constant within a single time slot. At any given time, the channel state has two possibilities. It could be either occupied by another radio system (busy) or available to be used by a CR (idle). We assume that

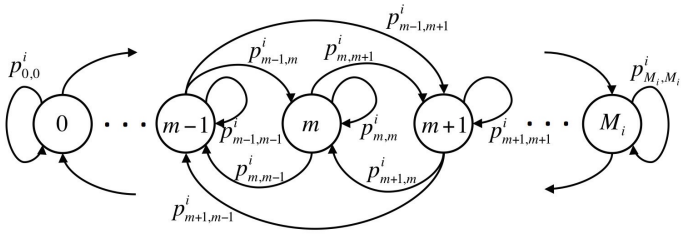

Fig. 2. Markov chain model for the ith sub-band when the state is defined to be the number of idle channels in the sub-band.

this idle/busy state of each channel evolves according to a two-state first order Markov chain. In [14], a Markov chain model for a single channel was proposed where states 0 and 1 correspond to busy and idle states, respectively. We denote the state of the jth channel in the ith sub-band at time k as $C_{i,j}[k] \in \{0, 1\}$, for $j \in \{1, ..., M_i\}$ and $i \in \{1, ..., N_b\}$. Note that, this channel Markov model can be characterized by using state transition probabilities. For simplicity, we assume time-invariant transition probabilities. The transition probability of the (i, j)th channel from state c to state c' is defined as

$$p^{i,j}_{c,c'} = \mathbf{Pr}\left\{C_{i,j}[k+1] = c' \mid C_{i,j}[k] = c\right\}, \quad \forall c, c' \in \{0, 1\}. \tag{1}$$

Sub-band selection decisions by a CR will depend on its performance objective. An application of CRs that has found wide interest is DSS. Here, the goal of the CR is to find idle spectrum opportunities for its own transmission needs. In so-called dynamic spectrum access (DSA) networks, this is achieved by CRs radios that perform channel-by-channel spectrum sensing.

In the following, we will assume that the goal of the CR is to find large chunks of idle spectrum opportunities for its own transmissions. Thus, we will set the performance objective in sub-band selection to be to sense the sub-band that has the largest number of idle channels. For that we may define a new state $S_i[k]$ denoting the number of the idle channels in ith sub-band at time k, where $S_i \in \{0, 1, ..., M_i\}$. If channel idle/busy dynamics were to be Markov, as assumed above, then the dynamics of this new state $S_i[k]$ will also be Markov. Figure 2 shows the sub-band Markov model with $(M_i + 1)$ possible states [6]. The transition probability of the ith sub-band from state s to state s' is defined as

$$p^i_{s,s'} = \mathbf{Pr}\left\{S_i[k+1] = s' \mid S_i[k] = s\right\}, \quad \forall s, s' \in \{0, 1, ..., M_i\}, \tag{2}$$

The overall spectrum state at time k can then be defined as $\mathbf{S}[k] = \{S_1[k], S_2[k], \cdots, S_{N_b}[k]\}$, in which $S_i[k]$ represents the state of the ith sub-band at given time k. Let us denote by $\mathcal{S}$ the set of all the possible states $S[k]$ may take. The set $\mathcal{S}$ can take Z possible states where

$$Z = \prod_{i=1}^{N_b} (M_i + 1). \tag{3}$$

The $Z \times Z$ transition probability matrix $\mathbf{P}$ corresponding to the overall spectrum state set $\mathcal{S}$ can be obtained, in which the (m, m')-th element represents the transition probability from the spectrum state $\mathbf{s}_m$ to state $\mathbf{s}_{m'}$

$$\mathbf{Pr}\left\{\mathbf{S}[k+1] = \mathbf{s}_{m'} \mid \mathbf{S}[k] = \mathbf{s}_m\right\} = p_{m,m'} \tag{4}$$

where $\mathbf{s}_m, \mathbf{s}_{m'} \in \mathcal{S}$ and $m, m' \in \{1, 2, \cdots, Z\}$.

III. POMDP MODEL FOR SUB-BAND SELECTION

The objective of the CR at any given time is to select one of the available N_b sub-bands for sensing. We may define the selection process at time k as taking an action $a[k] \in \mathcal{A}$ with the action space $\mathcal{A} = \{1, 2, \cdots, N_b\}$ representing the set of sub-band indices.

In many MDP applications, however, we may not have access to the system state but only an observation related to the actual state. These systems are known as POMDP systems. This, in fact, is true in the case of the sub-band selection problem since, at any given time, a CR can only observe a single sub-band out of the total N_b sub-bands. Essentially, at time k we can only observe $Y[k]$ corresponding to state $S_i[k]$ of the ith sub-band, but not the overall spectrum state $\mathbf{S}[k]$. Figure 3 shows a simple illustration of the POMDP framework over time. The action $a[k]$ at time k, which represents the selected sub-band for sensing during time $k+1$, is taken at the beginning of the time slot k. On the other hand, the observation $Y[k]$ of the state $\mathbf{S}[k]$ will not be available at the beginning of the time slot k so that action $a[k]$ will be selected before observing $Y[k]$ corresponding to the current state $\mathbf{S}[k]$ at time k. Instead, what is available to the CR is the history made of observations, actions and the associated rewards up to the current time k denoted by $h[k]$ where,

$$h[k] = (h[k-1], a[k-1], Y[k-1]). \tag{5}$$

Given all the available information up to time k, the *a posteriori* probability of state $\mathbf{S}[k]$ at time k can be defined as

$$b_m[k] = \mathbf{Pr}\left\{\mathbf{S}[k] = \mathbf{s}_m \mid h[k]\right\}, \tag{6}$$

which represents our *a posteriori* belief that the current state $\mathbf{S}[k]$ is $\mathbf{s}_m$. The set of all *a posteriori* probabilities corresponding to all possible states is called the belief state vector

$$\mathbf{b}[k] = \left[b_1[k], b_2[k], \cdots, b_Z[k]\right]^T, \tag{7}$$

with $b_m \in [0, 1]$ for $m = 1, \cdots, Z$.

It has been shown in [16] that the belief state vector is a sufficient statistic for optimal decision making in a POMDP. Thus, when making a decision, instead of taking into account all the history information $h[k]$, we may rely only on the belief state $\mathbf{b}[k]$. Moreover, the belief of the next state $\mathbf{S}[k+1]$ at time $k+1$, denoted by $\mathbf{b}[k+1]$, can be predicted from the knowledge of the current belief state vector $\mathbf{b}[k]$, the selected action $a[k]$ at time k and the resulting observation $Y[k]$ as shown in (8).

Let us denote by $r(a)$ the immediate reward from taking action $a \in \mathcal{A}$ when in state $\mathbf{s}_m$. As in [4], we define this reward to be the number of idle channels available in a-th sub-band at time $k+1$, if action a (i.e. the sub-band a) was chosen when in state $\mathbf{S}[k] = \mathbf{s}_m$ at time k. However, in the POMDP sub-band selection problem we cannot compute the immediate reward since the radio does not have access to the actual state $\mathbf{S}[k]$. We may instead make decisions based on the expected immediate reward given the belief state vector $\mathbf{b}$, computed as

$$\begin{aligned} r(\mathbf{b}, a) &= \mathbb{E}\left\{r(\mathbf{S}[k], a) \mid \mathbf{b}\right\} \\ &= \mathbf{r}(a)^T \mathbf{b} \end{aligned} \tag{9}$$

where $\mathbf{r}(a) = \left[r_1(a), r_1(a), \cdots, r_Z(a)\right]^T$, with $r_m(a) = r(\mathbf{s}_m, a)$ for $m \in \{1, 2, \cdots, Z\}$.

IV. REPLICATED Q-LEARNING ALGORITHM FOR SUB-BAND SELECTION

Finding an optimal policy for the sub-band selection POMDP, however, leads to many challenges. First, it may require high computational complexity due to the continuous state space of the belief state vector. Second, a policy needs to be computed in real-time. Moreover, we need the knowledge of sub-band Markov model parameters and, in particular, the transition probabilities of the model to be able to update the belief state vector as in (8). In addition, these model parameters may vary with time due to the dynamic nature of the wireless environment. These all make any attempt to directly compute an optimal policy complicated. As an alternative, we may use machine learning in which a CR may attempt to learn an optimal policy instead of computing one. This could help also in dealing with any time-varying RF environments.

A special machine learning approach, called reinforcement learning, could especially be suited when underlying state dynamics are Markov [7]. Q-learning is one of the most widely used reinforcement learning approaches. The basic idea of the Q-learning algorithm is to maintain a table, similar to what is shown in (10), that contains what is called Q-values. Assuming a completely observable Markov decision process (COMDP), the Q-value denoted by $Q(S, a)$ represents a measure of goodness resulting from taking an action a when in state S. In our system model, the action a refers to the index of the selected sub-band, with $a \in \{1, 2, \cdots, N_b\}$. Hence, if the selected sub-band contains a large number of idle channels this may lead to a high reward and, consequently, a high Q-value. In contrast, if the selected sub-band has a low number of idle channels the resultant Q-value is expected to be small. In (10), the column indices of the Q-table refer to the action set $\mathcal{A}$ (sub-band indices) while the rows of the Q-table represent all possible states of the spectrum enumerated in set $\mathcal{S}$.

Each time an action is selected in a given state, the Q-table is updated as shown in (11). Observe that, this updating can be performed without any knowledge of the Markov model parameters and involves only the selected actions and resulting rewards. Recall that, the reward $r(\mathbf{S}[k], a[k])$ represents the

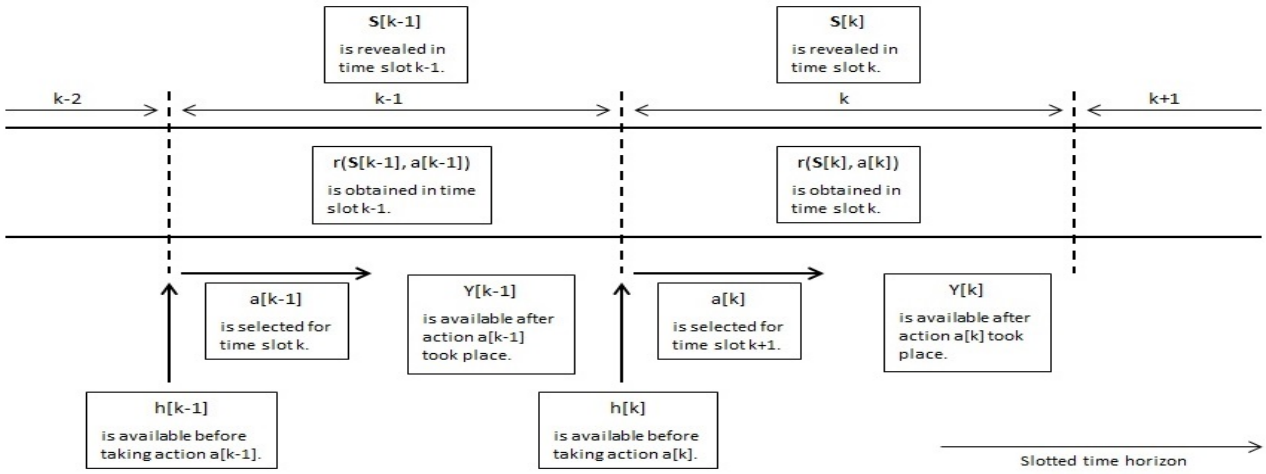

Fig. 3. An illustration of the POMDP procedure on the slotted time horizon.

$$\begin{aligned} b_m[k+1] &= \mathbf{Pr}\left\{\mathbf{S}[k+1] = \mathbf{s}_{m'} \mid h[k+1]\right\} \\ &= \sum_{m=1}^{Z} p_{m,m'} \mathbf{Pr}\left\{\mathbf{S}[k] = \mathbf{s}_m \mid h[k+1]\right\} \end{aligned} \tag{8}$$

$$Q\_table[k] = \begin{bmatrix} Q(\mathbf{S}_0[k], a_0[k]) & Q(\mathbf{S}_0[k], a_1[k]) & \cdots & Q(\mathbf{S}_0[k], a_{N_b}[k]) \\ Q(\mathbf{S}_1[k], a_0[k]) & Q(\mathbf{S}_1[k], a_1[k]) & \cdots & Q(\mathbf{S}_1[k], a_{N_b}[k]) \\ \vdots & \vdots & \ddots & \vdots \\ Q(\mathbf{S}_Z[k], a_0[k]) & Q(\mathbf{S}_Z[k], a_0[k]) & \cdots & Q(\mathbf{S}_Z[k], a_{N_b}[k]) \end{bmatrix}, \tag{10}$$

number of idle channels available in the selected sub-band. We denote by $\alpha \in (0, 1)$ the learning rate while the parameter $\gamma \in [0, 1)$ represents the discount factor.

Future actions (sub-band selections) will then be selected based on the updated Q-values:

$$a^* = \arg\max_{a \in \mathcal{A}} Q(\mathbf{S}, a). \tag{12}$$

The Q-learning algorithm attempts to reinforce the actions that lead to better outcomes from a given state. However, it may get trapped in a policy, that may not be the optimum, unless all entries the Q-table (corresponding to all (state, action) pairs) are updated consistently. In order to avoid this problem we may define a new parameter called exploration rate $\epsilon \in (0, 1)$. Depending on the exploration rate, the CR can switch between selecting the action characterized by (12) or just randomly selecting an action out of all possible actions:

$$a^* = \begin{cases} \arg\max_{a \in \mathcal{A}} Q(\mathbf{S}, a) & \text{with probability } 1 - \epsilon, \\ \sim U(\mathcal{A}) & \text{with probability } \epsilon, \end{cases} \tag{13}$$

where $U(\mathcal{A})$ denotes the uniform distribution over the action set $\mathcal{A}$. Choosing a high exploration rate may help in updating the entire Q-table and avoid being trapped in a sub-optimal policy. On the other hand, a low exploration rate will help in exploiting the already learned optimal actions. Thus, obtaining an optimal policy requires the selection of an appropriate exploration rate that could balance between the exploration and exploitation, as we will show in section V.

By using Q-learning, it is possible to learn an optimal policy for the sub-band selection problem without any knowledge of the transition probabilities. However, for this the state should be completely observable. In section III, we have seen that the sub-band selection problem is in fact a POMDP problem. A modification of Q-learning, known as replicated Q-learning algorithm [15], can be used to deal with POMDP problems. Since in a POMDP, it is possible what we have is a belief with a certain probability of what the state could be, we may define the MDP in terms of the belief state vector $\mathbf{b}[k]$ and the Q-table updating rule (11) can be rewritten for a POMDP case as in (14). Recall that, the m-th element b_m of the belief state vector $\mathbf{b}$ represents the probability of true state being the m-th state. We define $Q(\mathbf{b}, a)$ as the average of the Q-values when taking action a from all possible states given the belief state $\mathbf{b}$,

$$Q(\mathbf{b}, a) = \mathbb{E}\{Q(\mathbf{S}, a) \mid \mathbf{b}\} = \sum_m b_m Q(s_m, a). \tag{15}$$

$$Q(\mathbf{S}[k-1],a[k-1]) \leftarrow Q(\mathbf{S}[k-1],a[k-1]) + \alpha \left[r(\mathbf{S}[k-1],a[k-1]) + \gamma \max_{a} Q(S[k],a) - Q(\mathbf{S}[k-1],a[k-1]) \right]. \quad (11)$$

$$Q(\mathbf{s}_m,a[k-1]) \leftarrow Q(\mathbf{s}_m,a[k-1]) + \alpha\, b_m[k-1] \left[r(\mathbf{s}_m,a[k-1]) + \gamma \max_{a} Q(\mathbf{b}[k],a) - Q(\mathbf{s}_m,a[k-1]) \right]. \quad (14)$$

Now, the action selection rule in terms of the belief state vector $\mathbf{b}[k]$ at time k becomes

$$a^*[k] = \begin{cases} \arg\max_{a\in\mathcal{A}} Q(\mathbf{b}[k],a) & \text{with probability } 1-\epsilon, \\ \sim U(\mathcal{A}) & \text{with probability } \epsilon. \end{cases} \quad (16)$$

In case of sufficient observations that makes the CR always certain about its state (i.e. $b_j[k] = 1$ at each time k), the replicated Q-learning converges to the standard Q-learning algorithm [15].

V. Simulation Results

In this section, we use simulations to evaluate the performance of our proposed replicated Q-learning algorithm for sub-band selection. We will compare its performance with four other algorithms. First, we use an algorithm called the upper-bound performance, assumes that the CR may observe the exact state at time k before selecting action $a[k]$. This is obviously an upper bound even for the performance of the associated MDP problem because in practice the CR can only obtain an observation after selecting a sub-band for sensing. Second, we use the performance of the optimal sub-band selection policy obtained by solving the Bellman-optimality equation [4]. This, in other words, is the optimal performance of the associated MDP problem. Third, we use a Q-learning algorithm under the assumption that the states are completely observable. Fourth, and finally, we use the performance of a random sub-band selection scheme in which all sub-bands are selected with equal probabilities.

In the spectrum model, we assume that there are $N_b = 3$ sub-bands. The total number of channels in the spectrum is 8 channels in which the second sub-band contains 2 channels and the remaining 6 channels divided equally in the first and the third sub-bands. All these channel are assumed to have the same bandwidth. In addition, the dynamics of these channels are independent of each other. Table I summarizes the simulation settings. Each simulation involved 10,000 iterations. We observed that about 1,500 iterations were needed for the Q-table to be considered as converged.

Figure 4 compares the performance of the replicated Q-learning with the other four methods mentioned above. As our performance metric, we use the normalized accumulated reward, defined as

$$R_N = \frac{1}{N} \sum_{k=1}^{N} r(\mathbf{S}[k],a[k]), \quad (17)$$

where $r(\mathbf{S}[k],a[k])$ represents the immediate reward of taking action a when in state $\mathbf{S}[k]$ at time k and N is the number of iterations. Unless noted otherwise, a discount factor of $\gamma = 0.2$ was used. In addition, initially we allowed a high exploration rate of $\epsilon = 0.8$ and a learning rate of $\alpha = 0.4$. A convergence check for the Q-table was performed by computing the difference between old and updated Q-values and comparing against a certain threshold. However, a minimum of 1500 iterations were always allowed before terminating the learning period. After convergence, we reduced the learning rate and the exploration rate to $\alpha = 0.1$ and $\epsilon = 0.01$, respectively. As shown in Fig. 4, there is only a very small difference between the performance achieved by the optimal solution given by the solution to the Bellman optimality equation and the aforementioned upper-bound performance. The random sub-band selection policy can only achieve about a 68% of the of the optimal policy. As one would expect, the performance of both Q-learning and replicated Q-learning lie

TABLE I
SIMULATION PARAMETERS

Total # of sub-bands	3
# of channels in each sub-band	[3 2 3]
Total # of channels	8
# of states	48
# of simulation time steps	10,000
Minimum # of time steps for Q-learing convergence	1,500

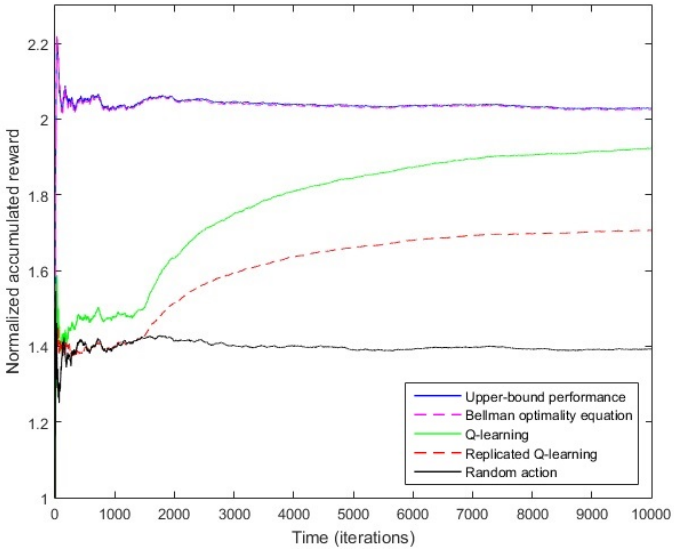

Fig. 4. Comparison of normalized accumulated reward of sub-band selection policies ($\epsilon = 0.01$, after convergence).

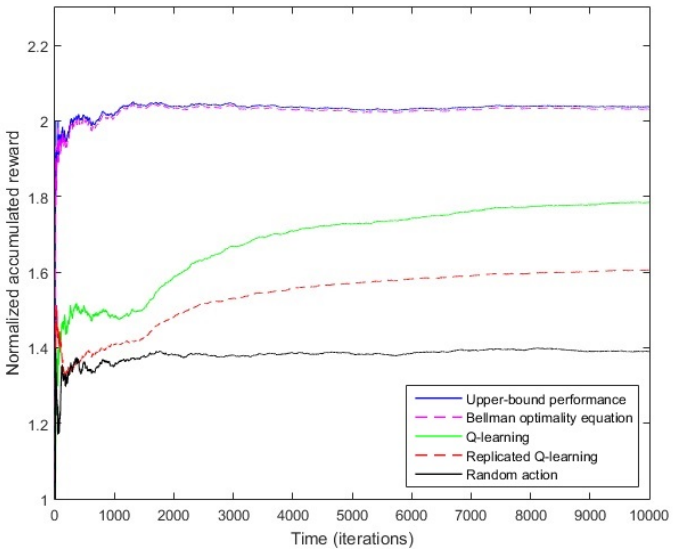

Fig. 5. Comparison of normalized accumulated reward of sub-band selection policies ($\epsilon = 0.3$, after convergence).

somewhere between the optimal and random-action policies. It can be seen from Fig. 4 that Q-learning converges about 95% of the performance achieved by the optimal policy. On the other hand, the replicated Q-learning algorithm achieves about 84% of the performance of the optimal policy. This is significant in three ways: First, it shows that the replicated Q-learning can indeed provide noticeably better performance than simply selecting random sub-bands for sensing. Second, its performance is not that far from that of the optimal sub-band selection policy that requires complete state observability. In fact, the difference is only about 16%. Third, and final, is the fact that replicated Q-learning achieves about 88% of the performance of the Q-learning which is a better performance upper-bound for comparison.

Recall that the choice of ϵ is a trade-off between the exploration and exploitation. In general, it may make sense to have a relatively larger ϵ value at the beginning and reduce it after convergence is achieved. Figure 5 shows the effect of using a relatively larger value of $\epsilon = 0.3$ after the convergence compared to Fig. 4. As can be seen from the figure, performance of both Q-learning and replicated Q-learning has degraded. The Q-learning achieves 88% of the optimal performance, while replicated Q-learning achieves only about 79% of the optimal performance. The reason is that the higher exploration rate leads to too much exploration. The CR selects random actions more often than in Fig. 4 as opposed to exploiting the already learned better actions.

VI. Conclusion

In this paper, we have considered the problem of sub-band selection for wideband spectrum sensing in a cognitive radio. The sub-band selection problem was first modeled as a partially observable Markov decision process (POMDP), in which only a single sub-band can be sensed at any given time out of all available sub-bands in the spectrum of interest. This model was then used to develop an effective, low-complexity policy to select the sub-bands based on the replicated Q-learning algorithm. Simulation results showed that the proposed replicated Q-learning method can provide a substantial improvement over the random sub-band selection policy. We also showed that it is better in practice to use a relatively larger exploration rate at the beginning so that fast learning can be achieved. However, after the Q-table convergence it is better to reduce the exploration rate to reap the benefits of the already learned actions.

Acknowledgment

This work was funded in part by the Air Force Research Laboratory, Space Vehicles Directorate, under grant FA9453-15-1-0314.

References

[1] B. Wang and K. J. R. Liu, "Advances in cognitive radio networks: A survey," IEEE J. Sel. Topics Signal Process., vol. 5, pp. 5-23, Feb. 2011.
[2] S. Haykin, "Cognitive radio: brain-empowered wireless communications," IEEE J. Sel. Areas Commun., vol. 23, no. 2, pp. 201-220, Feb 2005.
[3] J. Mitola, "Cognitive radio architecture evolution," Proc. IEEE, vol. 97, no. 4, pp. 626-641, Apr. 2009.
[4] S. K. Jayaweera, "Signal processing for cognitive radios," John Wiley & Sons, 2015.
[5] S. K. Jayaweera and C. G. Christodoulou, "Radiobots: architecture, algorithms and realtime reconfigurable antenna designs for autonomous, selflearning future cognitive radios," University of New Mexico, Technical Report EECE-TR-11-0001, Mar. 2011.
[6] Y. Li, S. K. Jayaweera, M. Bkassiny, and C. Ghosh, "Learning-aided sub-band selection algorithms for spectrum sensing in wide-band cognitive radios," IEEE Transactions on wireless communications, vol. 13, no. 4, pp. 2012-2024, April 2014.
[7] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," MIT Press, 1998.
[8] K. L. A. Yau, P. Komisarczuk, P. D. Teal, "Applications of reinforcement learning tocognitive radio networks," IEEE International Conference on Communications Workshops (ICC), May, 2010.
[9] J. Lunden, V. Koivunen, S. R. Kulkarni, and H. V. Poor "Reinforcement learning based distributed multiagent sensing policy for cognitive radio networks," in Proceedings of the IEEE International SymposiumonDynamic SpectrumAccessNetworks (DySPAN'11), pp. 642-646, May 2011.
[10] A. Galindo-Serrano, L. Giupponi, "Distributed Q-learning for aggregated inter-ference control in cognitive radio networks," IEEE Trans. Veh. Technol. vol. 59, no. 4, pp. 1823-1823, 2010.
[11] H. Li, "Multi-agent Q-Learning of Channel Selection in Multi-user Cognitive Radio Systems A Two by Two Case," in IEEE Conf. on System, Man and Cybernetics, San Antonio, Texas, USA, pp. 1893-1898, October 2009.
[12] M. Di Felice, K. R. Chowdhury, W. Meleis, and L. Bononi, "To sense or to transmit: a learning-based spectrum management scheme for cognitive radio mesh networks," in Proceedings of the 5th Annual IEEE Workshop on Wireless Mesh Networks (WiMesh'10), pp. 19-24, June 2010.
[13] M. Bkassiny, S. K. Jayaweera, and K. A. Avery, "Distributed Reinforcement Learning based MAC protocols for autonomous cognitive secondary users," in Proceedings of the 20th Annual Wireless and Optical Communications Conference (WOCC'11), Newark, NJ, USA, pp. 1-6, April 2011.
[14] E. N. Gilbert, "Capacity of a burst-noise channel," Bell System Technical Journal, vol. 39, pp. 1253-12653, Sept. 1960.
[15] M. L. Littman, A. R. Cassendra, and L. P. Kaelbling, "Learning polices for partially observable environment: Scaling up," in Proceedings of the 12th international conference on machine learning, Tahoe City, CA, pp. 362-370, 1955.
[16] R. D. Smallwood and E. J. Sondik,"The optimal control of partially observable Markov decision processes over a finite horizon," Operations research, vol. 21, pp. 1071-1088, 1973.