Multi-agent Reinforcement Learning Based Cognitive Anti-jamming

Mohamed A. Aref, Sudharman K. Jayaweera and Stephen Machuzak
Communications and Information Sciences Laboratory (CISL)
Department of Electrical and Computer Engineering, University of New Mexico
Albuquerque, NM 87131-0001, USA
Email: {maref, jayaweera, smachuzak29}@unm.edu

***Abstract*—This paper proposes a reinforcement learning based approach to anti-jamming communications with wideband autonomous cognitive radios (WACRs) in a multi-agent environment. Assumed system model allows multiple WACRs to simultaneously operate over the same (wide) spectrum band. Each radio attempts to evade the transmissions of other WACRs as well as avoiding a jammer signal that sweeps across the whole spectrum band of interest. The WACR makes use of its spectrum knowledge acquisition ability to detect and identify the location (in frequency) of this sweeping jammer and the signals of other WACRs. This information and reinforcement learning is used to successfully learn a sub-band selection policy to avoid both the jammer signal as well as interference from other radios. It is shown, through simulations, that the proposed learning-based sub-band selection policy has low computational complexity and significantly outperforms the random sub-band selection policy.**

***Index terms*—** Anti-jamming, Markov decision process, multi-agent reinforcement learning, Q-learning, sub-band selection, wideband autonomous cognitive radios, wideband spectrum scanning.

I. Introduction

An early application of cognitive radio (CR) technology was to overcome the problem of inefficient spectrum utilization via dynamic spectrum sharing (DSS) in which unlicensed users are allowed to opportunistically access the spectrum of a licensed user. However, when viewed as an evolution of software-defined radios (SDRs), CRs may find much more applications than just DSS [1]. Indeed, the ability for spectrum and network awareness and to modify operating mode based on autonomous decisions, make them ideal for pursuing some of the original motivations for SDR technology including, for example, interoperability [1], [2]. Wideband autonomous cognitive radios (WACRs), equipped with real-time reconfigurable RF front-ends spanning hundreds of megahertz (MHz) to several gigahertz (GHz), are aimed at such broader applications rather than simply DSS. They may find increasing relevance in space, military and homeland security applications in addition to consumer wireless communications.

The key to cognitive operation is the radio's ability to sense its surrounding RF environment. This functionality is known as spectrum knowledge acquisition and, as shown in Fig. 1, can be divided in to three steps [1]: wideband spectrum scanning, spectral activity detection and signal classification and identification. Wideband spectrum scanning step involves the real-time sensing of a wide spectrum range overcoming the instantaneous sensing bandwidth limitations imposed by the hardware constraints. In the second step of the spectrum knowledge acquisition process, the WACR detects any, and all, spectrum activities that may exist in the sensed sub-band. Finally, a third step of signal classification is assumed in order to identify and associate the detected active signals with particular systems and origins.

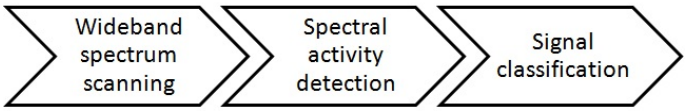

Fig. 1. Spectrum knowledge acquisition procedure.

A common situation in which cognitive communications can be a great asset is when malicious users launch jamming attacks to disrupt the reliable communications [3], [4]. In practice, this will result in a complicated multi-agent environment due to multiple WACRs simultaneously operating over the same wide spectrum band that is challenged by a malicious jammer. In this case, each WACR needs to avoid the jammer as well as transmissions of other WACRs. This paper addresses such an anti-jamming problem in a multi-agent environment with the goal of finding optimal anti-jamming and interference avoidance policies for the WACRs. However, direct computation of optimal decision policies can often be computationally too demanding. The use of machine learning may instead allow a WACR to learn an optimal, or at least an efficient, decision policy to adopt its transmission to avoid both the jammer attack and interference. Specifically, in this paper, we focus on a machine learning paradigm called reinforcement learning (RL) which could be well-suited when the underlying state dynamics are Markov. Indeed, RL has been applied in many CR applications involving both single-agent and multi-agent environments [5], [6]. For example, multi-agent reinforcement learning (MARL) based on Q-learning was proposed to let secondary users (SUs) select operating channels in the case of a two-user two-channel CR system in [7] and a multi-user multi-channel CR system in [8]. The performance objective in these earlier work, however, was to minimize the collisions among the SUs and primary users (PUs).

There have been previous attempts at using RL specifically to achieve anti-jamming with cognitive radios. In [9], for example, the authors considered the jammer attacks on SUs in a CR network. While the SU's desire was to maximize spectrum utilization with a designed channel selection strategy, the jammer's objective was to decrease the spectrum utilization by strategic jamming. The state-action-reward-state-action (SARSA) and QV-learning, two different reinforcement learning algorithms, were used by the SUs to adapt their strategy on switching between control and data channels according to their observations about jammer's action, spectrum availability and channel quality. In [10] and [11], MARL algorithms based on minimax-Q and Win-or-Learn-Fast (WoLF) principles were applied, respectively, to find anti-jamming policies for SUs in multi-channel CR systems. The CR and the jammer, in [10] and [11], were treated as two equally knowledgeable learning agents. However, when the CR lacks sufficient knowledge about the jammer, these approaches may not lead to sufficient anti-jamming performance.

Most recently, a single-agent reinforcement learning (SARL) based on Q-learning was proposed in [12] to enable a WACR evade a jammer signal that sweeps across the whole spectrum of interest to the radio. Although the performance of the learning-based decision policy was shown to be excellent in [12], the scenario was too simplified to be useful in practice. The purpose of this paper is two-fold: Formalize the underlying Markov decision process (MDP) framework assumed in [12] and extend the RL based sub-band selection policy for anti-jamming to the scenarios in which there are multiple policy-learning WACRs operating in the same spectrum range challenged by a sweeping jammer. Thus, our performance objective is the combined anti-jamming defense and avoidance of interference from other WACRs. We formalize the underlying MDP model framework assumed in [12] by developing a new state definition for the spectrum. Note that, if the jammer is also equipped with cognitive radio technology, it will likely be able to adapt its jamming strategy in response to the strategies of the WACRs. In this paper, however, we assume a sweeping jammer that follows a fixed strategy leaving the above case for future research.

The remainder of the paper is organized as follows: Section II describes our assumed spectrum dynamics model and the proposed new definition for the state of a spectrum sub-band. The spectral activity detection framework is described in Section III. Section IV discusses the implementation of the proposed cognitive MARL algorithm for anti-jamming and interference avoidance. Simulation results are presented 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 [1]. Each sub-band may include a different number of communication channels. Let M_i denote the number of communication channels in the i-th sub-band. In our model, we assume having equal-length time slots, where each slot corresponds to a single sensing duration. For simplicity, we assume the sub-band state to be constant within a single time slot. Among the existing work defining the state of a sub-band, [1], [13] and [14] are the most relevant to our work. They defined the sub-band state as the number of idle channels available in a sub-band. However, this definition could result in a large total number of possible states leading to unacceptably high computational complexity.

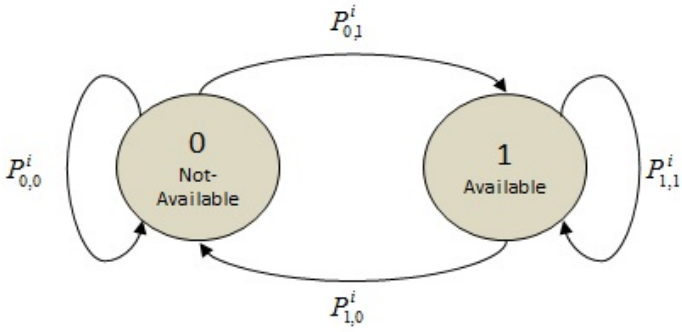

Fig. 2. Markov chain model for a single sub-band.

In this work, we get around the complexity issue by introducing a new state definition for a sub-band. This definition depends on the availability of sufficiently large interference-free (idle) bandwidth to satisfy a specified minimum required bandwidth for transmission. To be specific, let β denote the minimum required bandwidth for transmission defined by the system (e.g. β= 20 MHz for IEEE 802.11g WiFi). Then, according to our new definition, each sub-band can only be in one of two possible states: state 0 and state 1 as shown in Fig. 2: At any given time, if the available idle bandwidth in the sub-band is greater than or equal to β then the sub-band is considered to be in state 1 (available). Otherwise, it is considered to be in state 0 (not-available). Let us denote the state of the i-th sub-band at time t by $S_i[t] \in \{0, 1\}$, for $i \in \{1, ..., N_b\}$. It can reasonably be argued that this state $S_i[t]$ is a discrete-time Markov process. Then, the transition probability of the i-th sub-band from state s to state s' can be written as

$$p^i_{s,s'} = Pr\left\{S_i[t+1] = s' \mid S_i[t] = s\right\}, \forall s, s' \in \{0, 1\}. \quad (1)$$

Most traditional communication systems transmit each signal only over a contiguous bandwidth. However, many emerging systems have the capability of transmission over non-contiguous bandwidths (e.g. carrier aggregation (CA) in LTE systems [15]). Thus, we may define two modes of operation for our WACRs: First is non-contiguous bandwidth mode in which the available bandwidth of a sub-band is calculated by adding up of all the interference-free frequencies in this sub-band regardless of whether they are contiguous or not. Second is the contiguous bandwidth mode in which the available bandwidth of a sub-band is defined as the maximum interference-free contiguous bandwidth in this sub-band. In this paper, for simplicity, the focus is on the contiguous bandwidth operation mode although the same approach may be extended to the non-contiguous bandwidth mode.

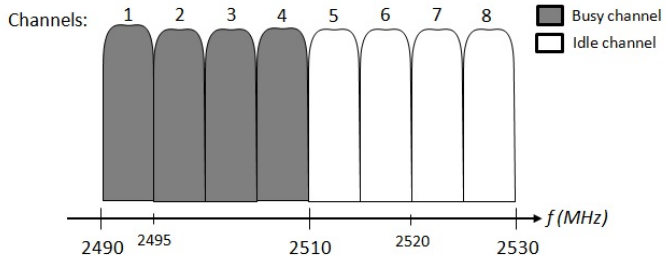

Fig. 3. An example of a sub-band made of 8 channels. At the present time, last 4 channels are idle. Hence, current state of this sub-band is 1 (available) if the minimum bandwidth parameter $\beta \leq 20$ MHz.

In order to determine the state of a sensed sub-band, the WACR should have the ability to detect any and all active signals in this sub-band and determine precisely at which frequencies these active signals exist. This will allow it to compute the amount of idle bandwidth available in the sensed sub-band. This process, known as spectral activity detection [1], is described briefly below in section III.

As an example, let us consider a sub-band formed of 8 channels of equal bandwidth as shown in Fig. 3. As can be seen from Fig. 3, only the last four channels are currently idle. Let us assume that the minimum required bandwidth β = 20 MHz. In this case, the sub-band shown in Fig. 3 will considered to be in state 1 (available). If this sub-band was selected for transmission, the cognitive engine (CE) of the WACR will then inform the SDR platform the center frequency of the largest available contiguous bandwidth in the sub-band. The SDR will then be able to up-convert the baseband signal to be transmitted to the corresponding carrier frequency as shown in Fig. 4.

With the above sub-band state definition, the overall spectrum state at time t can be defined as $\mathbf{S}[t] = (S_1[t], S_2[t], \cdots, S_{N_b}[t])$, in which $S_i[t]$ represents the (binary) state of the i-th sub-band at time t. Let us denote by $\mathcal{S}$ the set of all the possible states $\mathbf{S}[t]$ may take. The set $\mathcal{S}$ can take 2^{N_b} possible states. Note that, if, as in [1], [13] and [14], the number of idle channels in a sub-band was taken as the sub-band state definition, we will end up with $\prod_{i=1}^{N_b}(M_i+1)$ number of possible spectrum states which can be considerably larger than 2^{N_b} when $M_i > 1$, for $i = 1, \cdots, N_b$.

III. Spectral activity detection

The spectral activity detection procedure is described in Fig. 5. In order to determine the amount of available idle bandwidth in each sub-band, a detector based on the Neyman-Pearson (NP) criterion is used. This detector would allow the WACR to identify the carrier frequencies of all active signals in the sensed sub-band [1], [12].

During initialization, the noise floor of each sub-band is estimated and is used to compute the required NP threshold for detecting spectral activity subject to a given false-alarm probability [1]. Next, the power spectral density (PSD) corresponding to the sensed sub-band signal is estimated. The locations of active signals in the sensed sub-band are identified by extracting the frequencies at which the power spectrum exceeds the NP threshold. We assume that the spectral activity detection is based on the periodogram power spectral density estimator, which is suitable when there is no *a priori* knowledge available on possible signals in the sub-band:

$$\hat{S}_y(F) = \frac{1}{N}\left|\sum_{n=0}^{N} y[n]e^{-j2\pi Fn}\right|^2 = \frac{1}{N}\left|Y(F)\right|^2, \quad (2)$$

where $y[n]$ is the N length time-domain sensed signal of the sub-band of interest and $Y(F)$ is the discrete-time Fourier transform (DTFT) of $y[n]$ with $-1/2 \leq F \leq 1/2$ denoting the normalized frequency.

The periodogram, however, is known to suffer from high noise fluctuations. This may result in erroneous spectral activity detector decisions, as at some frequency locations the PSD may exceed the NP threshold while it should not and vice versa. To reduce the effect of such noisy fluctuations on spectral activity detection, we may apply frequency-domain smoothing to the periodogram estimate of the sub-band spectrum. Assume the DTFT of the sensed signal is computed at a set of discrete frequency points $F_k = \frac{k}{N}$ for $k = 0, \ldots, N-1$, so that $Y[k] = Y(F_k)$. The decision statistic at frequency k is then obtained by smoothing the periodogram using a rectangular window of length L (assumed to be odd) centered at frequency k [1]:

$$T_k(\mathbf{Y}) = \frac{1}{LN}\sum_{l=-(L-1)/2}^{l=(L-1)/2}\left|Y[k+l]\right|^2, \quad (3)$$

where $\mathbf{Y} = (Y[0], Y[1], \cdots, Y[N-1])$.

The NP threshold is applied to the smoothed periodogram in (3) so that the WACR may detect the locations of the idle frequency bands within the sensed sub-band. These are next used to compute the maximum available contiguous bandwidth. The state of the sub-band is determined by comparing this to the minimum required bandwidth β for transmission.

IV. Cognitive MARL anti-jamming communications

The objective of the proposed cognitive MARL anti-jamming algorithm is to avoid both deliberate jamming and unintentional interference. Thus, at each time instant t, the WACR should make a decision on whether to continue transmission on the current sub-band or to switch to a new sub-band. To be effective, the WACR should be able to predict which sub-band will most likely meet the performance objectives of the user. This sub-band selection problem can be formulated as a partially observable Markov decision process (POMDP) since, at each time step, only the state of the sensed sub-band is knowable by the WACR. The complete state $\mathbf{S}[t]$ of the RF spectrum may not be fully observable due to hardware and signal processing limitations.

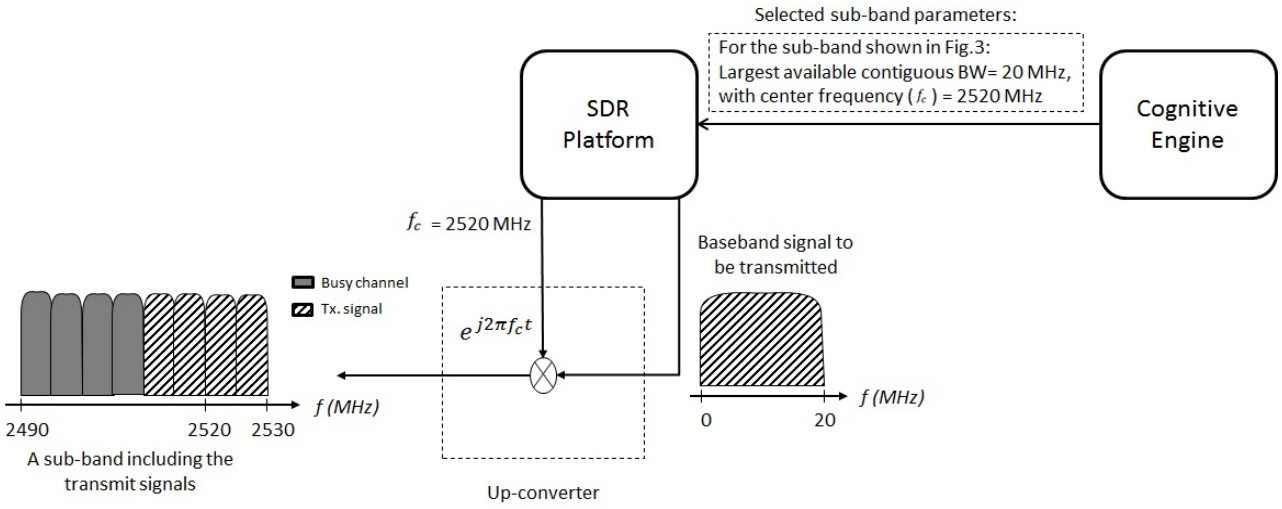

Fig. 4. Cognitive radio operation: The SDR maps the baseband signal to the RF frequency informed by the cognitive engine.

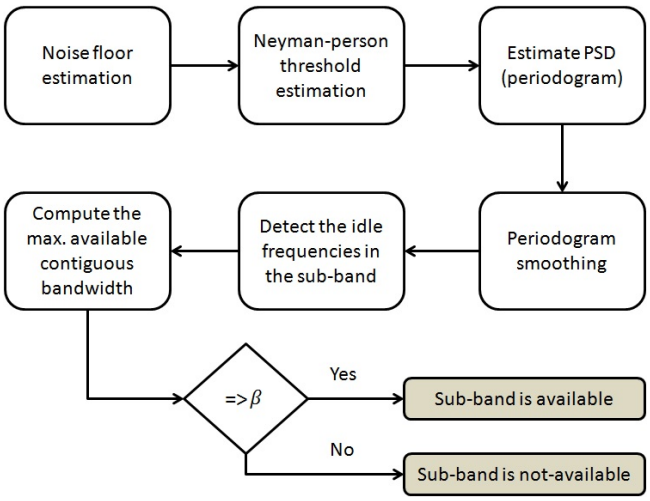

Fig. 5. Spectral activity detection procedure.

Computing an optimal policy for a POMDP, however, may lead to impractically high computational demand. Alternatively, machine learning can be used to learn an optimal, or at least a sub-optimal but reasonably good, sub-band selection policy. As mentioned earlier, a particular machine learning approach called reinforcement learning can especially be suited when the underlying state dynamics are Markov, as assumed in our system. 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, known as the Q-table, that contains what are called the Q-values denoted by $Q(S, a)$ representing a measure of goodness of taking the action $a \in \mathcal{A}$ when in state S [1], [16]. Since the action space $\mathcal{A} = \{1, 2, \cdots, N_b\}$ in our scenario is the set of sub-band indices, taking action a corresponds to selecting the a-th sub-band.

After each execution of an action, the WACR updates the Q-table, based on a certain observed reward, as shown in (4) where $\alpha \in (0, 1)$ is the learning rate and $\gamma \in [0, 1)$ is a discount factor. In our approach, we define a reward function $r(S, a)$ that depends on the amount of time it takes for the jammer or interference signals to interfere with a WACR transmission once it has switched to the a-th sub-band. Future actions (sub-band selections) are selected based on the updated Q-values:

$$a^* = \arg\max_{a \in \mathcal{A}} Q(\mathbf{S}, a). \tag{5}$$

The Q-learning algorithm, however, may get trapped in a non-optimum policy unless all entries of the Q-table are updated consistently [16]. This effect can be mitigated by introducing an exploration rate $\epsilon \in (0, 1)$. Depending on the exploration rate, the WACR may switch between selecting the action characterized by (5) or just randomly selecting an action out of all possible actions:

$$a^* = \begin{cases} \arg\max_{a \in \mathcal{A}} Q(\mathbf{S}, a) & \text{with probability } 1 - \epsilon, \\ \sim U(\mathcal{A}) & \text{with probability } \epsilon, \end{cases} \tag{6}$$

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 an already learned policy that performs well-enough. Thus, obtaining a policy with good performance requires the selection of an appropriate exploration rate that could strike a balance between the exploration and exploitation.

In our scenario, the goal for each WACR is to learn the pattern of behavior of the jammer and other WACRs in its vicinity by using the above Q-learning algorithm. Each time, the WACR will select a sub-band that has a contiguous idle

$$Q(\mathbf{S}[t-1], a[t-1]) \leftarrow Q(\mathbf{S}[t-1], a[t-1]) + \alpha \left[r(\mathbf{S}[t-1], a[t-1]) + \gamma \max_a Q(\mathbf{S}[t], a) - Q(\mathbf{S}[t-1], a[t-1]) \right]. \quad (4)$$

bandwidth of at least β. The selected new sub-band must have low interference for the longest amount of time with high probability. Once the desired idle bandwidth condition is violated in the current sub-band due to an interferer or a jammer, the WACR will select another sub-band according to the decision policy (6).

V. Simulation results

In this section, we use simulations to evaluate the performance of our proposed MARL based sub-band selection framework for anti-jamming. We will compare its performance with a random sub-band selection scheme in which all sub-bands are selected with equal probabilities. As our performance metric, we use the normalized accumulated reward, defined as

$$R_T = \frac{1}{T} \sum_{t=1}^{T} r_t(S_t, a_t), \quad (7)$$

where $r_t(S_t, a_t)$ represents the immediate reward of taking action a_t when in state S_t and T is the number of iterations. Note that, the rewards in (7) are those that achieved after the convergence of the Q-table. In all simulation cases, the currently occupied sub-band is excluded form the decision making choices.

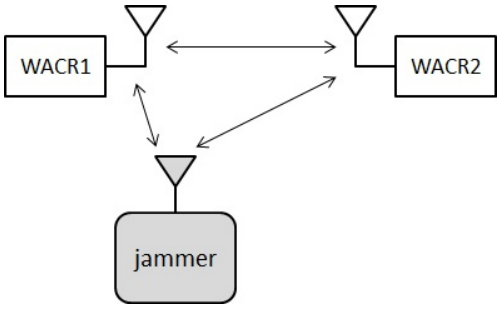

Fig. 6. Test case 1: Two WACRs operate in the spectrum range 2.0 GHz to 2.2 GHz. The jammer sweeps this 200MHz wide spectrum from low to high frequency.

In our simulations we considered 2 test cases. The first case assumes two WACRs and a sweeping jammer as shown in Fig. 6. The operating frequency band is taken to be from 2.0 to 2.2 GHz. This gives a total of 5 sub-bands each with a bandwidth of 40 MHz. In the second case, we assume three WACRs besides the sweeping jammer as shown in Fig. 7. The spectrum of interest in this case is taken to be from 2.0 to 2.4 GHz. This gives 10 sub-bands each with a bandwidth of 40 MHz. In both cases, the WACRs and the jammer are arranged randomly. For any 2 units, having a short distance in-between, implies that the transmission of one will be received by the other with a high signal strength causing high interference impact if both are operating on the same sub-band. We have

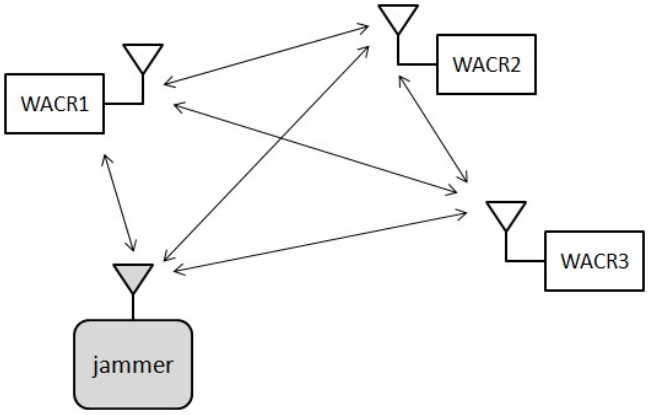

Fig. 7. Test case 2: Three WACRs operate in the spectrum range 2.0 GHz to 2.4 GHz. The jammer sweeps this 400MHz wide spectrum from low to high frequency.

used a continuous signal that sweeps the spectrum of interest from the lower to the higher frequency as the jammer. For simplicity, we have set the jammer to sweep a single sub-band within each sensing duration of 0.25 msec.

Initially, the Q-learning parameters are set to be $\gamma = 0.9$, $\alpha = 0.4$ and $\epsilon = 0.8$. Once the Q-table is considered to be converged, we reduced the learning rate and the exploration rate to $\alpha = 0.1$ and $\epsilon = 0.01$, respectively.

Figure 8 shows the normalized accumulated reward achieved by the first and second WACR (WACR1 and WACR2) with the proposed MARL based policy (6) and random action policy in test case 1. Note that, since there are 5 available sub-bands, the maximum immediate reward possible in this case is 1 msec. For example, assume that the transmission of a WACR in the 3rd sub-band is interrupted by a jammer. If it is the only transmitter in the system, then the WACR should choose sub-band 2 in order to avoid the jammer for the longest possible amount of time [12]. In this case, the jammer will spend 1 msec to sweep over 4 sub-bands until it reaches the sub-band 2 again. However, if we consider the interference caused by the transmission from other WACRs, it could affect the above maximum possible reward. From Fig. 8, the performance of the MARL policy lies somewhere between 75% to 90% of the above maximum possible reward of 1 msec. On the other hand, the random selection policy achieves only about 60% of the above maximum possible performance. Indeed, with random sub-band selection, a WACR could receive a reward of 0.25, 0.5, 0.75, or 1 msec, resulting in an average reward of 0.6 msec. These results show that the MARL policy can indeed provide noticeably better performance than simply selecting random sub-bands.

Next, we apply our proposed MARL anti-jamming algorithm to the second test case in which there are 3 WACRs operating over 10 sub-bands. In this case, the maximum

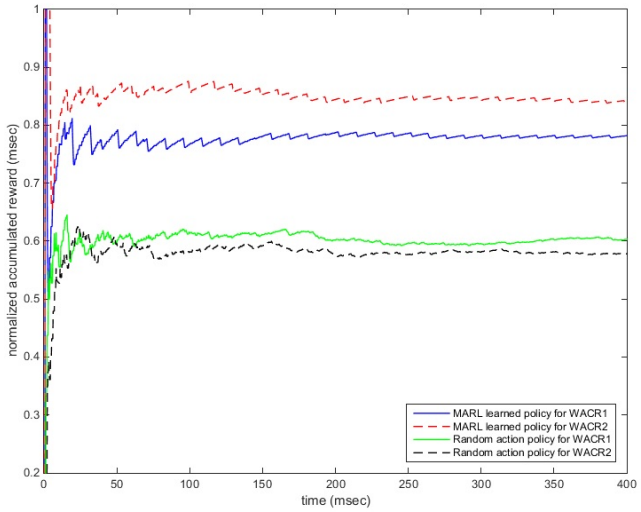

Fig. 8. Test case 1: Normalized accumulated reward of WACR1 and WACR2.

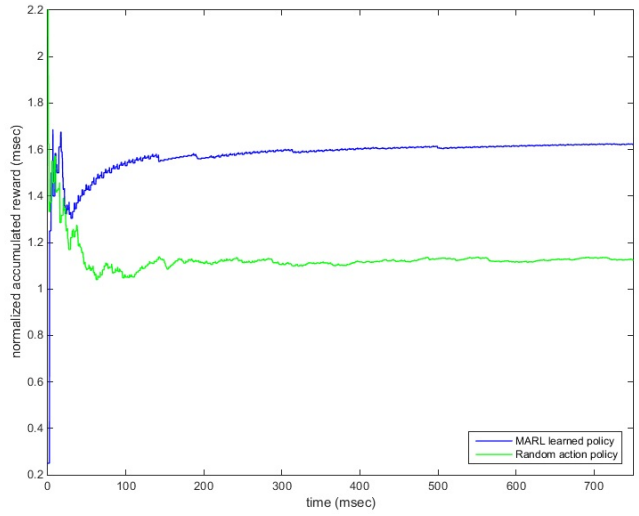

Fig. 9. Test case 2: Normalized accumulated reward of WACR1.

possible reward for a single WACR should be 2.25 msec since there are 10 sub-bands in the system. Figure 9 compares the performance of the first WACR (WACR1) with MARL and random selection policies in the test case 2. From Fig. 9 we observe that the proposed MARL policy can achieve about 73% of the above mentioned maximum possible performance while the random selection policy can achieve only about 48%. Clearly, these results show that the proposed MARL based sub-band selection policy can be an effective cognitive anti-jamming and interference avoidance protocol.

VI. Conclusion

In this paper we have proposed a multi-agent reinforcement learning (MARL) algorithm, based on Q-learning, for WACRs to avoid a sweeping jammer signal as well as unintentional interference from other WACRs. Moreover, we have developed a new definition for the sub-band spectrum state to reduce the computational complexity of learning a decision policy. When the WACR's transmission faces interference, it switches to a new spectrum sub-band that will lead to the longest possible uninterrupted transmission as learned through Q-learning. Simulation results showed that the proposed MARL anti-jamming protocol can provide a substantial improvement over the random sub-band selection policy.

Acknowledgment

This work was funded in part by the Air Force Research Laboratory, Space Vehicles Directorate, under grants FA9453-15-1-0314 and FA9453-16-1-0052 and in part by a subcontract under the NASA STTR Phase I contract NNX15CC80P. The authors would like to thank the Communications & Intelligent Systems Division at NASA GRC for useful discussions.

References

[1] S. K. Jayaweera,"Signal Processing for Cognitive Radio," John Wiley & Sons, Hoboken, NJ, USA. ISBN: 978-1-118-82493-1, 2014.

[2] 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.

[3] R. Di Pietro and G. Oligeri, "Jamming mitigation in cognitive radio networks," IEEE Network, vol. 27, no. 3, pp. 10-15, May/June 2013.

[4] A. Sampath, H. Dai, H. Zheng and B. Y. Zhao, "Multi-channel jamming attacks using cognitive radios," Proc. of 16th International Conference on Computer Communications and Networks (ICCCN 2007), Honolulu, HI, USA, pp. 352-357, Aug. 2007.

[5] M. Bkassiny, Y. Li and S. K. Jayaweera, "A survey on machine-learning techniques in cognitive radios," IEEE Communications Surveys and Tutorials, vol. 15, no. 3, pp. 1136-1159. Third Quarter 2013

[6] 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, pp. 1-6, May 2010.

[7] H. Li, "Multi-agent Q-Learning of Channel Selection in Multi-user Cognitive Radio Systems A Two by Two Case," in IEEE Conference on System, Man and Cybernetics, San Antonio, Texas, USA, pp. 1893-1898, Oct. 2009.

[8] H. Li, "Multi-agent Q-learning for competitive spectrum access in cognitive radio systems," in IEEE Fifth Workshop on Networking Technologies for Software Defined Radio Networks, Boston, MA, USA, June 2010.

[9] S. Singh and A. Trivedi, "Anti-jamming in cognitive radio networks using reinforcement learning algorithms," in 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN), Indore, India, pp. 1-5, Sep. 2012.

[10] B. Wang, Y. Wu, K. R. Liu, and T. C. Clancy, "An anti-jamming stochastic game for cognitive radio networks," IEEE J. Sel. Areas Commun., vol. 29, no. 4, pp. 877-889, 2011.

[11] B. F. Lo and I. F. Akyildiz,"Multiagent jamming-resilient control channel game for cognitive radio ad hoc networks," in Proc. IEEE ICC, London, UK, June 2012.

[12] S. Machuzak and S. K. Jayaweera,"Reinforcement learning based anti-jamming with wideband autonomous cognitive radios," IEEE/CIC International Conference on Communications in China (ICCC), Chengdu China, July 2016.

[13] 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 Trans. on wireless communications, vol. 13, no. 4, pp. 2012-2024, April 2014.

[14] M. A. Aref, S. Machuzak, S. K. Jayaweera and S. Lane, "Replicated Q-learning based sub-band selection for wideband spectrum sensing in cognitive radio," IEEE/CIC International Conference on Communications in China (ICCC), Chengdu China, July 2016.

[15] Z. Shen, A. Papasakellariou, J. Montojo, D. Gerstenberger, and F. Xu, "Overview of 3GPP LTE-advanced carrier aggregation for 4G wireless communications," IEEE Commun. Mag., vol. 50, pp. 122-130, Feb. 2012.

[16] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction," MIT Press, 1998.