A Cognitive Anti-jamming and Interference-Avoidance Stochastic Game

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

***Abstract*—This paper presents a design of a wideband autonomous cognitive radio (WACR) for anti-jamming and interference-avoidance. The proposed system model allows multiple WACRs to simultaneously operate over the same spectrum range producing a multi-agent environment. The target of each radio is to predict and evade a dynamic jammer signal as well as avoiding transmissions of other WACRs. The proposed cognitive framework is made of two operations: sensing and transmission. Each operation is helped by its own learning algorithm based on Q-learning, but both will be experiencing the same RF environment. The simulation results indicate that the proposed cognitive anti-jamming technique has low computational complexity and significantly outperforms non-cognitive sub-band selection policy while being sufficiently robust against the impact of sensing errors.**

***Index Terms*—Anti-jamming, multi-agent reinforcement learning, Q-learning, stochastic game, wideband autonomous cognitive radios.**

I. Introduction

Wideband autonomous cognitive radios (WACRs) are radios that have the ability of self learning and autonomous decision making. As a result, they can optimally self-reconfigure to adapt to the user needs and surrounding RF environment in real-time [1], [2]. The key to such autonomous operation is the radio's ability to sense and comprehend its operating environment. In general, it is desired that the radio can operate over a wide spectrum range that makes the problem of sensing all frequencies of interest to the radio in real-time a challenging problem. However, if this is achieved, such WACRs may find increasing relevance in aerospace, military and homeland security applications in addition to consumer wireless communications. One of the most challenging security threats in which WACRs can be a great asset is jamming attacks. Jamming is malicious signal transmissions generated by an outside source that aims to disrupt the reliable communications. In practice, however, there may be multiple WACRs simultaneously operating over the same spectrum range leading to a multi-agent environment in which each WACR will need to avoid both malicious jammer as well as the transmissions of other radios. This scenario may be modeled as a stochastic game, an extension of Markov Decision Processes (MDPs), in which interactions among different agents is considered [2]. In this context, a WACR may use multi-agent reinforcement learning (MARL) to solve the stochastic game by learning an optimal, or near-optimal, policy to keep its communication link unjammed [2], [3].

MARL has previously been proposed in the literature [4]−[7] for anti-jamming transmission in cognitive radio (CR) networks. For instance, the authors in [4] proposed a stochastic general-sum game for modeling the jammed control channels. The objective was to obtain an optimal control channel allocation strategy for CRs to avoid jamming attacks using Win-or-Learn-Fast (WoLF) principle [8]. The approach in [4] considered the effect of sensing errors, however it was limited only to control channels. The authors in [5] used minimax Q-learning to find anti-jamming policies for secondary users (SUs) in multi-channel CR systems. The CR and the jammer in [5] were treated as two equally knowledgeable learning agents. One of the drawbacks of the proposed algorithms in [5] is that it also assumed perfect sensing. The authors in [6] formulated a competing stochastic game by dividing the network into two sub-networks: the ally network and the enemy network. The objective of each of the two sub-networks is to achieve the maximum spectrum utilization while jamming the opponent transmission as much as possible. Several reinforcement learning techniques were proposed: Minimax-Q, Nash-Q and Friend-or-Foe Q-learning. This work was extended in [7] for the case of time-varying channel rewards. A new algorithm based on online convex programming was introduced in [7] to obtain an optimal strategy that achieves the best steady-state channel rewards.

Most recently, a cognitive anti-jamming stochastic game model was proposed in [9] to enable a WACR evading a jammer signal that sweeps across the spectrum of interest to the radio as well as transmissions of other radios. The advantage of the proposed model in [9] was that it enables the WACR to switch its operating sub-band before getting jammed, compared to previously proposed anti-jamming techniques that switch the operating sub-band only after getting jammed. Although the performance of the proposed learning policy was shown to be excellent in [9], the scenario was simplified. In particular, as with many other previous work, [9] also assumed perfect sensing. The purpose of this paper is two-fold: Extend the proposed cognitive framework of [9] by introducing a new definition of reward functions that may reduce the probability of getting jammed and examine the robustness of the proposed technique against sensing errors.

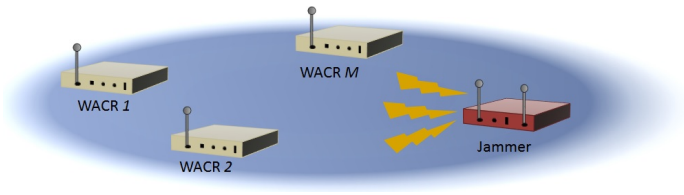

Figure 1. M WACRs operate in the same frequency range challenged by a dynamic jammer [9].

The rest of the paper is organized as follows: Section II describes the system model. Section III gives an overview of the reinforcement learning algorithm. Section IV discusses the implementation of the proposed cognitive stochastic game for anti-jamming and interference-avoidance. The simulation results are provided in Section V. Finally, concluding remarks are given in Section VI.

II. Cognitive Radio System Model

The proposed system model includes M WACRs operating over the same spectrum range and challenged by a dynamic jammer as shown in Fig. 1 [9]. The spectrum of interest is divided into N_b sub-bands with equal bandwidth [1]. The stochastic game formulation includes a set of states and a collection of action sets denoted by $\mathcal{S}$ and $\mathcal{A}_1, \cdots, \mathcal{A}_M$, respectively. The players of this game are the WACRs. The game is played in a sequence of stages. The game moves from its current state to a new random state with transition probability determined by the current state and one action from each player $T : \mathcal{S} \times \mathcal{A} \times \cdots \times \mathcal{A}_{\mathcal{M}} \rightarrow PD(\mathcal{S})$. For simplicity, the state of each spectrum sub-band is assumed to be constant within a single game stage. The objective is to obtain optimal, or near-optimal, policy that enables each of the WACRs to switch the operating sub-band before getting jammed or interrupted by an interference.

The proposed cognitive framework consists of two operations: sensing and transmission [9]. Each operation will have its own learning algorithm with different targets, but they will be experiencing the same RF environment. The goal of sensing is to learn the pattern of jammer's behavior and transmissions of other radios. On the other hand, the cognitive objective of the transmission operation is to determine when to switch the operating sub-band and to where. Thus, at any stage of the game, there are two sub-bands associated with a given WACR: one for sensing and one for transmission. Essentially, if the sensing operation were to learn an optimal policy, the WACR would be able to accurately predict the jammed or interfered sub-bands. This will help the transmission operation as follows: if the current transmission sub-band is predicted to be jammed during the next time instant, the WACR will switch to another sub-band and may avoid getting jammed [9].

There is a different game state associated with each operation: $s_s[n] \in \mathcal{S}$ and $s_t[n] \in \mathcal{S}$ represent the indices of sensing and transmission sub-bands, respectively, at time n. The space of game states is then given by $\mathcal{S} = \{1, \cdots, N_b\}$. On the other hand, each sub-band can be in one of two possible states: state "0" and state "1". At any given time, if the sub-band is getting jammed or facing an interference, it is considered in state "0" (not-available). Otherwise, it is considered to be in state "1" (available). The set of sub-band states can be then given by $\mathcal{V} = \{0, 1\}$. Note the main difference between the two types of states: the sub-band state v refers to the availability of the sub-band, while the game state s refers only to the index of the operating sub-band apart from it is available or not.

The WACR has to detect the state of operating sub-bands for both sensing and transmission operations. In other word, it has to detect value of $v \in \mathcal{V}$ for sub-band index $s \in \mathcal{S}$. For the sensing operation, the WARC can perform *spectral activity detection* to detect any active signal in the sensed sub-band and hence identify the availability of the sub-band [1], [2]. On the other hand, during transmission operation, the link quality will determine if transmission over the current sub-band is acceptable or not. If it was acceptable, the sub-band is considered available (state "1") otherwise it is not available (state "0"). After determining the states of both operating sub-bands, the WACR will select and execute actions for both operations. We denote by $a_s[n]$ and $a_t[n]$ the new operating sub-bands (actions) for sensing and transmission, respectively, at time n. The action space for both processes thus defined as $\mathcal{A} = \{1, \cdots, N_b\}$.

III. Reinforcement learning approach

Computing an optimal policy for the proposed stochastic game in section II is complicated due to the computational complexity and the real-time computation requirements. Moreover, the model parameters are time-varying due to the dynamic nature of the wireless environment. As an alternative, we may use machine learning in which a WACR attempts to learn an optimal policy instead of computing one. This will allow the WACR to deal with any time-varying wireless environments. A particular type of machine learning approach, called reinforcement learning, could especially be suited when dealing with MDP and stochastic games [3]. Q-learning is one of the most widely used reinforcement learning approaches.

Algorithm 1 Q-learning-aided proposed cognitive anti-jamming approach

1: **Initialize:**
   α, γ, $\epsilon \in [0, 1]$
   $Q(s, a) \leftarrow 0 \ \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$
2: **for** each stage n **do**
3: Identify the state ($v \in \mathcal{V}$) of operating sub-band s
4: **if** sub-band state $v = 0$ **then**
5: Compute reward r for current state s and action a
6: Update Q-value $Q(s, a)$ as follow:
7: $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[r + \gamma \max_{a} Q(s', a)]$
8: Select new action $a' \in \mathcal{A}$ for the new state s' according to the following:
9: $$a' = \begin{cases} \arg\max_{a \in \mathcal{A}} Q(s', a) & \text{with probability } 1 - \epsilon, \\ \sim U(\mathcal{A}) & \text{with probability } \epsilon, \end{cases}$$

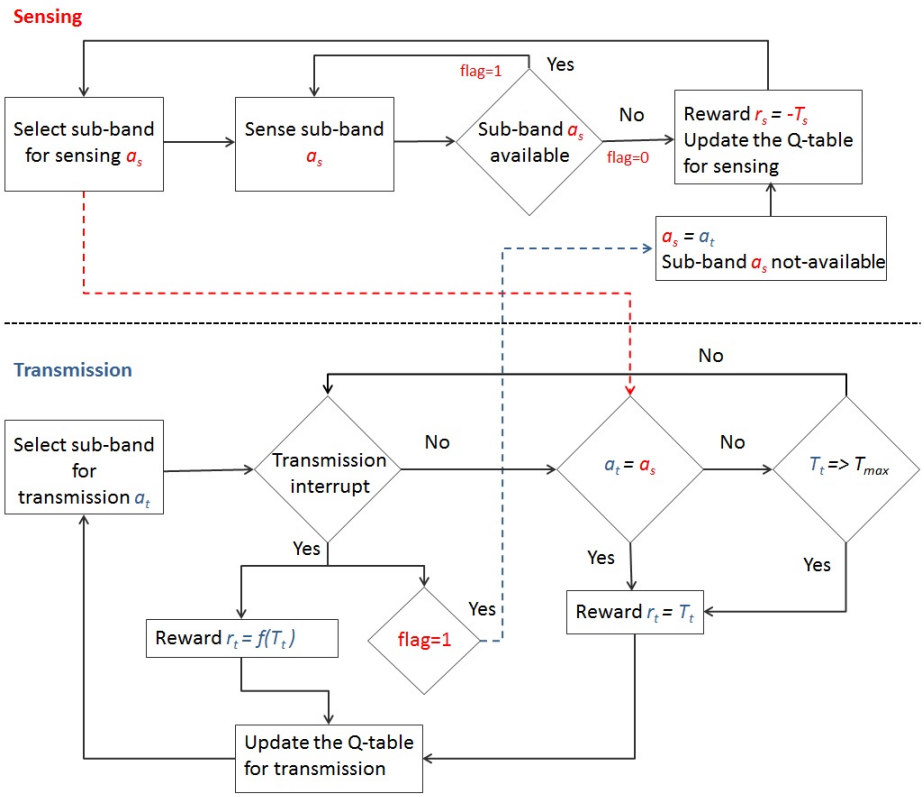

Figure 2. Proposed cognitive radio operations for sensing and transmission.

Although, Q-learning is a single-agent learning algorithm and our model is multi-agent, it is a common approach to apply a single-agent learning algorithm to a multi-agent domain for simplicity [10]. The advantage of Q-learning is that it does not require prior knowledge of the operating environment and is highly adaptive to the state dynamics [11].

The Q-learning algorithm uses a 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 when in state s [11]. The number of rows and columns of the Q-table corresponds to the number of possible states and possible actions, respectively. Thus, based on the game state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ definitions in the previous section, the dimension of the Q-table will be $N_b \times N_b$.

Algorithm 1 summarizes the Q-learning-aided proposed cognitive anti-jamming algorithm [9]. At any game stage n, the WACR has to identify the state v of the current operating sub-band s (lines 2-3). If the sub-band state is "0" (not-available), the WACR updates the Q-value $Q(s, a)$ of the current state s and action a based on a certain observed reward $r(s, a)$ (lines 4-7). On the other hand, if the sub-band state is "1" (available), the operating sub-band will remain the same without any changes. We denote by $\alpha \in (0, 1)$ the learning rate, while the parameter $\gamma \in [0, 1)$ represents the discount factor. Once the Q-table is updated, the WACR chooses a new action a' that represents the new operating sub-band according to line 9 of Algorithm I.

At any given state, the Q-learning algorithm reinforces the actions that lead to better rewards. However, unless the entire Q-table is updated, the Q-learning algorithm may get trapped in a sub-optimal policy. In order to mitigate this problem an exploration rate parameter $\epsilon \in (0, 1)$ is introduced. Choosing an appropriate value for exploration rate, the WACR may switch between selecting the action characterized by $\arg\max_{a \in \mathcal{A}} Q(s', a)$ or randomly selecting an action according to $U(\mathcal{A})$ where $U(\mathcal{A})$ denotes the uniform distribution over the action set $\mathcal{A}$. Note that, selecting a high exploration rate may help in updating all entries of the Q-table and avoid being trapped in a sub-optimal policy. On the other hand, a low exploration rate may help in exploiting an already learned policy that performs well-enough. Obtaining a policy with good performance requires the selection of an appropriate exploration rate that could strike a balance between the exploration and exploitation.

IV. Proposed Anti-Jamming Stochastic Game

In this section, we discuss how to obtain optimal or near-optimal policy for each WACR in order to predict and avoid the jammer and the transmissions of each other based on the stochastic game formulation and the state definition presented in section II. As mentioned earlier, the cognitive operation for a given WACR is divided in to two processes: sensing and transmission. Each process is aided by its own Q-learning

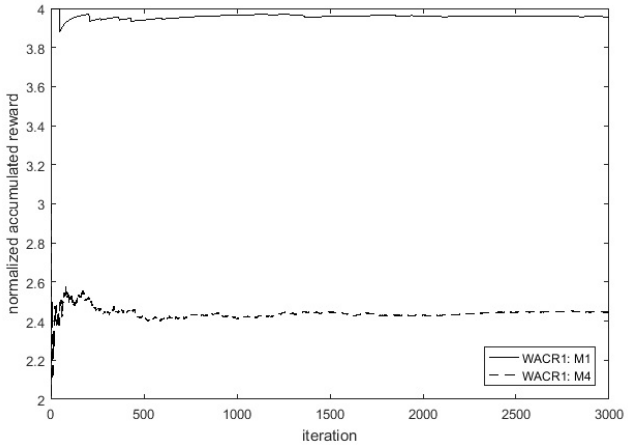

Figure 3. Test case 1 (1 Jammer, 5 Sub-bands, 1 WACR) with perfect sensing: Normalized accumulated reward values with proposed method 1 (M1) and random policy (M4).

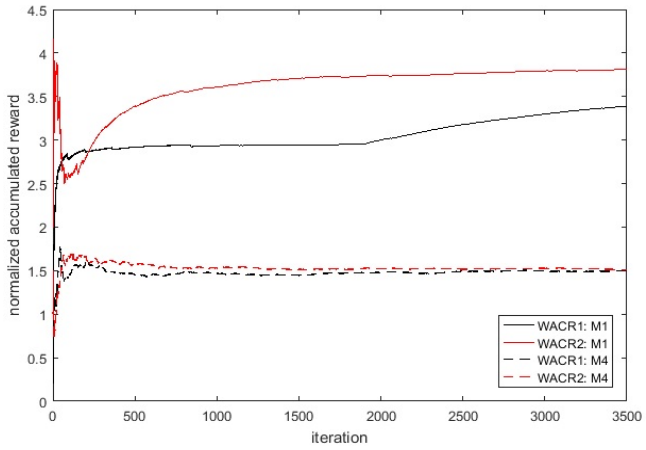

Figure 4. Test case 2 (1 Jammer, 6 Sub-bands, 2 WACRs) with perfect sensing: Normalized accumulated reward values with proposed method 1 (M1) and random policy (M4).

algorithm. Thus, there are two Q-tables that need to be updated at every game stage.

The sensing and transmission operations of a given WACR is described in Fig. 2. The goal of the sensing operation is to track the jammer and interference signals. Thus, the WACR uses the Q-learning algorithm in order to obtain an optimal policy that always senses the sub-band where the jammer or interference signals are present. In order to achieve this goal, the WACR computes the elapsed time until the jammer or interference signal arrives the newly selected sensing sub-band (sensing action), denoted by T_s. The reward corresponding to a sensing action is defined as the negative of this value:

$$r_s = -T_s. \tag{1}$$

Note that, the WACR will select a new sub-band for sensing if and only if the current sensing sub-band gets jammed or faces interference. The actions $a_s \in \mathcal{A}$ for the sensing operation are selected such that rewards are maximized. Thus, they are selected corresponding to the shortest time it takes until the operating sub-band gets jammed.

The objective of transmission protocol is to switch the operating sub-band before getting jammed or facing an interference. As can be seen from Fig. 2, the operating sub-band for transmission may be changed under two scenarios [9]: First is if the transmission is interrupted, implying that the current operating sub-band is facing either too much interference or a jamming attack. This is determined by monitoring the communication link quality. The second condition is when the sensing operation predicts that the current transmission sub-band to be the one that will get jammed/interfered in the next time instant. Since the objective is to switch the operating sub-band before getting jammed, effective learning must lead to switching always due to the second condition while avoiding the first. In order to achieve this objective, the first case is assigned a low reward (high penalty) while the second one is given relatively higher reward:

$$r_t = \begin{cases} f(T_t) & \text{if sub-band } a_t \text{ gets jammed} \\ T_t & \text{otherwise} \end{cases}, \tag{2}$$

where T_t is the transmission duration in sub-band a_t before switching to a new one and function $f(T_t)$ represents the penalty function for the undesirable case. In [9], this penalty function is given by $f(T_t) = -1$. In this paper, we define the penalty function as

$$f(T_t) = -N_b e^{-\lambda T_t} \tag{3}$$

where $\lambda > 0$ is a design parameter that may be optimized to obtain efficient learning. The action a_t for transmission operation is selected such that the reward is maximized. Thus, the selected new sub-band for transmission must remain available for the longest amount of time with high probability.

The framework in [9] is further extended by introducing a feedback branch from transmission operation to sensing as shown in Fig. 2. As mentioned earlier, the transmission operation may switch the operating sub-band if the transmission is interrupted. In such a case the sensing operation would have a false prediction about the location of the jammer or interference signal. Thus, the learning process for the sensing operation may be improved by using a feedback to notify the sensing operation with the current location of the jammer or interference signal as determined by the transmission link. A threshold T_{max} is defined, such that the transmission time over a certain sub-band cannot exceed it. This is an additional safe guard to improve the probability of getting jammed, especially in the presence of sensing errors.

V. Simulation Results

In this section, we use simulations to evaluate the performance of our proposed cognitive anti-jamming and interference-avoidance stochastic game. In the following, we

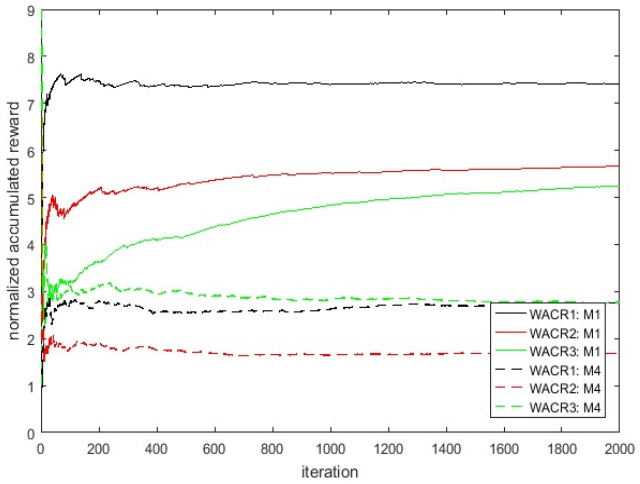

Figure 5. Test case 3 (1 Jammer, 12 Sub-bands, 3 WACRs) with perfect sensing: Normalized accumulated reward values with proposed method 1 (M1) and random policy (M4).

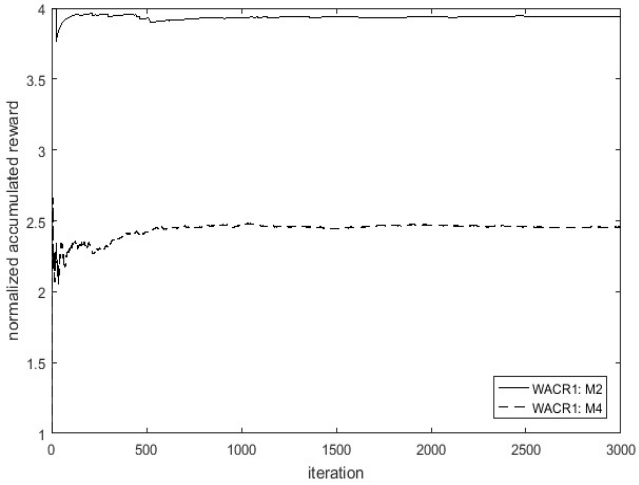

Figure 6. Test case 1 (1 Jammer, 5 Sub-bands, 1 WACR) with perfect sensing: Normalized accumulated reward values with proposed method 2 (M2) and random policy (M4).

consider the proposed penalty function $f(T_t)$ with two different λ values: $\lambda = 0.2$ and $\lambda = 1$, defined as proposed method 1 (M1) and proposed method 2 (M2), respectively. In addition, the proposed algorithm in [9], labeled as method 3 (M3), is also evaluated for comparison. Finally, a random sub-band selection policy, defined as method 4 (M4), in which all sub-bands are selected for transmission with equal probabilities is considered. We use the normalized accumulated reward corresponding to the transmission operation, defined as

$$R_N = \frac{1}{N} \sum_{n=1}^{N} r_t(s_t[n], a_t[n]), \tag{4}$$

as our performance metric where $r_t(s_t[n], a_t[n])$ represents the immediate reward of taking action $a_t[n]$ when in state $s_t[n]$ during the transmission operation and N denotes the number of iterations. Note that, the rewards in this case are

Table I
NORMALIZED ACCUMULATED REWARD VALUES FOR DIFFERENT SIMULATION SCENARIOS WITH PERFECT SENSING

Test case	Scenario	Reward upper bound	WACR 1	WACR 2	WACR 3	Average
1	1 WACR and 5 sub-bands	4	M1: 3.9 M2: 3.9 M3: 3.8 M4: 2.5			M1: 3.9 M2: 3.9 M3: 3.8 M4: 2.5
2	2 WACRs and 6 sub-bands	4	M1: 3.4 M2: 2.9 M3: 2.85 M4: 1.5	M1: 3.8 M2: 3.9 M3: 2.9 M4: 1.5		M1: 3.6 M2: 3.4 M3: 2.87 M4: 1.5
3	3 WACR and 12 sub-bands	9	M1: 7.4 M2: 5.5 M3: 5.3 M4: 2.8	M1: 5.6 M2: 8.7 M3: 5.5 M4: 2.3	M1: 5.2 M2: 5.5 M3: 6.1 M4: 2.8	M1: 6.06 M2: 6.56 M3: 5.63 M4: 2.63

Table II
PROBABILITIES OF GETTING JAMMED FOR DIFFERENT SIMULATION SCENARIOS WITH PERFECT SENSING

Test case	Scenario	WACR 1	WACR 2	WACR 3	Average
1	1 WACR and 5 sub-bands	M1: 0.45% M2: 0.75% M3: 0.86% M4: 1.8%			M1: 0.45% M2: 0.75% M3: 0.86% M4: 1.8%
2	2 WACRs and 6 sub-bands	M1: 2.18% M2: 2.2% M3: 2.6% M4: 46%	M1: 1.9% M2: 2.7% M3: 2.1% M4: 42%		M1: 2.04% M2: 2.45% M3: 2.35% M4: 44%
3	3 WACR and 12 sub-bands	M1: 6.7% M2: 9.1% M3: 8.2% M4: 51%	M1: 5.9% M2: 4.4% M3: 12.9% M4: 56%	M1: 3.8% M2: 4.8% M3: 6.24% M4: 51%	M1: 5.46% M2: 6.1% M3: 9.1% M4: 52.6%

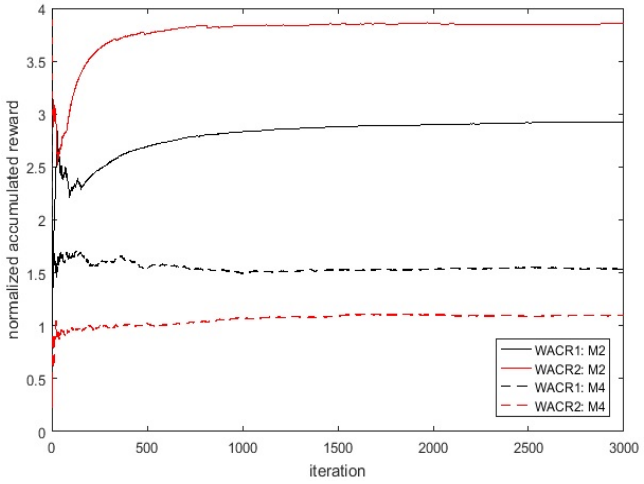

Figure 7. Test case 2 (1 Jammer, 6 Sub-bands, 2 WACRs) with perfect sensing: Normalized accumulated reward values with proposed method 2 (M2) and random policy (M4).

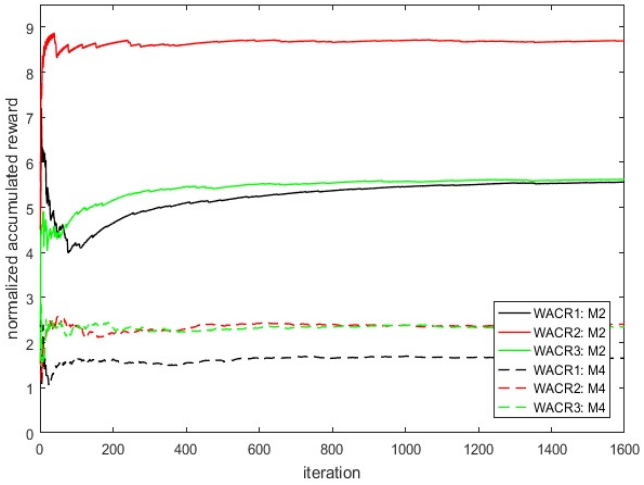

Figure 8. Test case 3 (1 Jammer, 12 Sub-bands, 3 WACRs) with perfect sensing: Normalized accumulated reward values with proposed method 2 (M2) and random policy (M4).

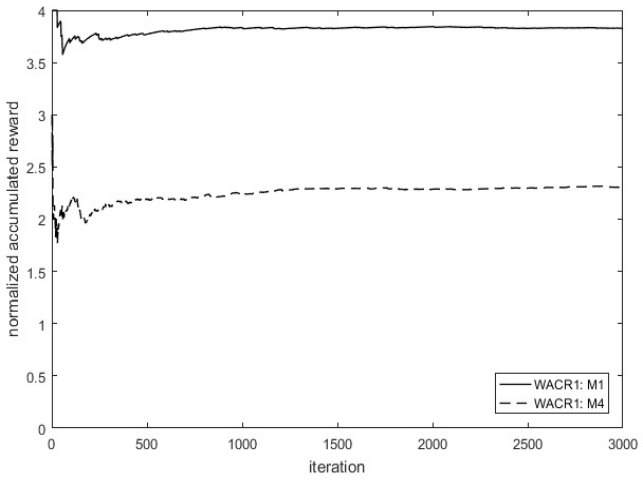

Figure 9. Test case 1 (1 Jammer, 5 Sub-bands, 1 WACR) with sensing errors: Normalized accumulated reward values with proposed method 1 (M1) and random policy (M4).

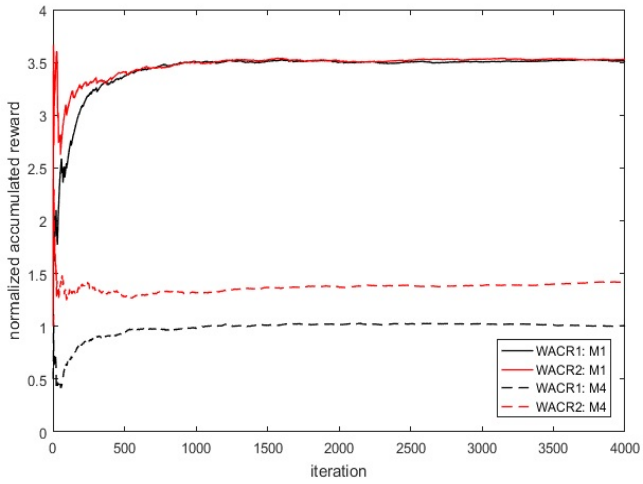

Figure 10. Test case 2 (1 Jammer, 6 Sub-bands, 2 WACRs) with sensing errors: Normalized accumulated reward values with proposed method 1 (M1) and random policy (M4).

those that achieved after the convergence of the Q-table. In all simulations, the current operating sub-band is excluded from the decision making choices for both transmission and sensing operations.

Three test cases are considered with different numbers of WACRs and different numbers of sub-bands. Test case 1, includes 1 WACR and 5 possible operating sub-bands. In test case 2, there are 2 WACRs and 6 sub-bands. Finally, test case 3, includes 3 WACRs and 12 possible operating sub-bands.

A. Perfect sensing

Tables I and II summarize the obtained values of normalized accumulated reward and probability of getting jammed for all test cases with perfect sensing, respectively. Figures 3-5 show the performance of proposed method 1 for the three different test cases, respectively. The performance analysis of proposed method 2 is shown in Figures 6-8 for the three different test cases. For test case 1, the maximum possible reward for a single WACR should be 4 time steps since there are 5 sub-bands in the system. Figures 3 and 6 show that proposed methods 1 and 2 achieve about 97% of this maximum possible reward. As can be seen from Table I, method 3 [9] also achieves the same reward as that of proposed methods 1 and 2, while the random policy can achieve only about 62% of the maximum reward. Figures 4 and 7 show that the achieved accumulated reward of proposed methods 1 and 2 for test case 2 lies somewhere between 72% to 97% of the maximum possible reward. Method 3 [9], on the other hand, achieves only about 72% of the maximum possible reward as shown in Table I, while the random policy can achieves only about 37%. From Table II, proposed methods 1 and 2 and method 3 [9] have very low probabilities of getting jammed, while

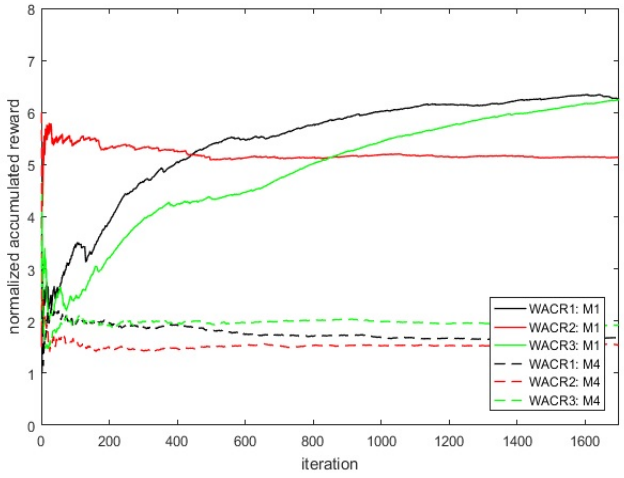

Figure 11. Test case 3 (1 Jammer, 12 Sub-bands, 3 WACRs) with sensing errors: Normalized accumulated reward values with proposed method 1 (M1) and random policy (M4).

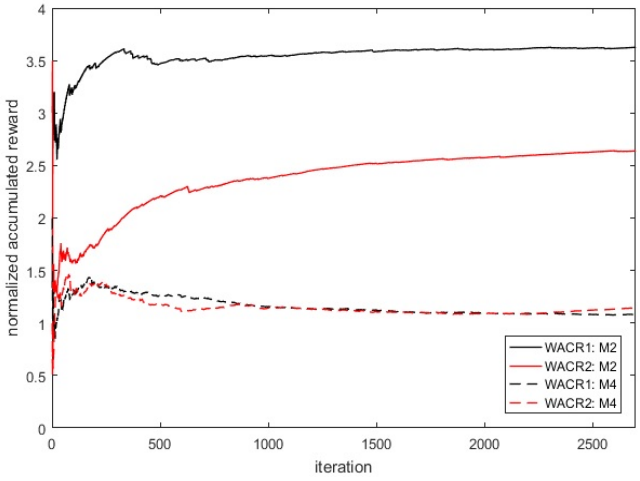

Figure 13. Test case 2 (1 Jammer, 6 Sub-bands, 2 WACRs) with sensing errors: Normalized accumulated reward values with proposed method 2 (M2) and random policy (M4).

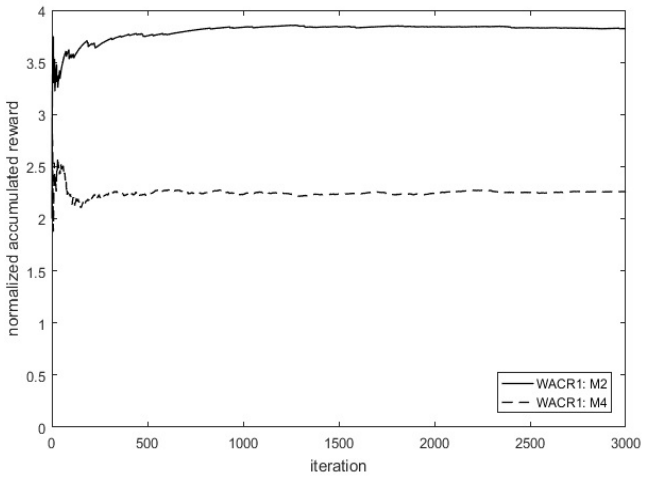

Figure 12. Test case 1 (1 Jammer, 5 Sub-bands, 1 WACR) with sensing errors: Normalized accumulated reward values with proposed method 2 (M2) and random policy (M4).

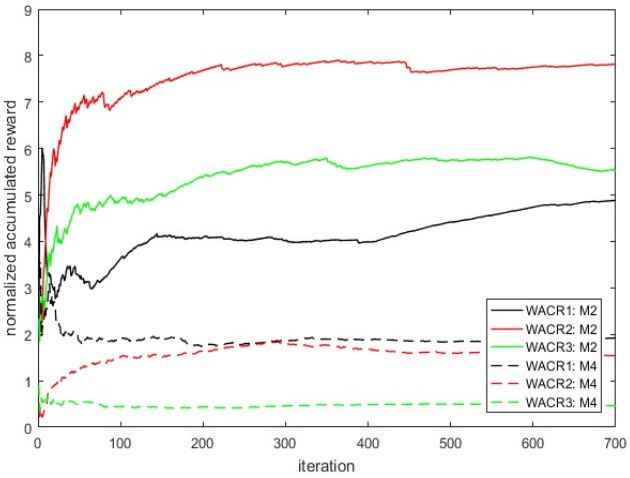

Figure 14. Test case 3 (1 Jammer, 12 Sub-bands, 3 WACRs) with sensing errors: Normalized accumulated reward values with proposed method 2 (M2) and random policy (M4).

the random policy has a 44% probability of getting jammed. Similarity, the results of test case 3 show that the proposed methods and method 3 [9] significantly outperform the random policy in terms of both jamming probability and accumulated reward. Moreover, there is a slight improvement in the results with proposed methods 1 and 2 compared to those with method 3 [9].

B. Effect of sensing errors

Sensing errors such as false-alarm and miss detection can have a major impact on the performance of the proposed algorithm. Note that, a false-alarm here corresponds to when a WACR mistakenly declares the availability of the operating sub-band. On the other hand, in miss detection case, WACR incorrectly detects the presence of spectral activity in the operating sub-band thereby missing the availability of the sub-band. Figures 9-14 show the normalized accumulated rewards for three different test cases for proposed methods 1 and 2, respectively, in the presence of sensing errors. In all test cases, we assumed false-alarm and miss detection probabilities of 0.02. Tables III and IV, respectively, summarize the obtained normalized accumulated reward and probability of getting jammed with these sensing errors. These simulation results show that the proposed methods have significantly improved the jammed probability in the presence of sensing errors compared to those with M3 and random policy.

VI. Conclusion

In this paper, we addressed the cognitive anti-jamming and interference-avoidance problem in a multi-agent environment. The cognitive framework was divided into two operations: sensing and transmission. Both used reinforcement learning approach based on Q-learning to obtain an optimal or sub-optimal policy. The objective of the sensing operation was to

Table III
NORMALIZED ACCUMULATED REWARD VALUES FOR DIFFERENT SIMULATION SCENARIOS WITH SENSING ERRORS

Test case	Scenario	Reward upper bound	WACR 1	WACR 2	WACR 3	Average
1	1 WACR and 5 sub-bands	4	M1: 3.8			M1: 3.8
			M2: 3.8			M2: 3.8
			M3: 3.5			M3: 3.5
			M4: 1.9			M4: 1.9
2	2 WACRs and 6 sub-bands	4	M1: 3.5	M1: 3.5		M1: 3.5
			M2: 3.6	M2: 2.6		M2: 3.1
			M3: 2.7	M3: 2.6		M3: 2.65
			M4: 1.2	M4: 1.4		M4: 1.3
3	3 WACR and 12 sub-bands	9	M1: 6.3	M1: 5.2	M1: 6.3	M1: 5.93
			M2: 4.9	M2: 7.8	M2: 5.5	M2: 6.06
			M3: 4.9	M3: 5.2	M3: 4.3	M3: 4.8
			M4: 1.9	M4: 1.4	M4: 1.9	M4: 1.73

Table IV
PROBABILITIES OF GETTING JAMMED FOR DIFFERENT SIMULATION SCENARIOS WITH SENSING ERRORS

Test case	Scenario	WACR 1	WACR 2	WACR 3	Average
1	1 WACR and 5 sub-bands	M1: 2.8%			M1: 2.8%
		M2: 2.9%			M2: 2.9%
		M3: 3.8%			M3: 3.8%
		M4: 8.1%			M4: 8.1%
2	2 WACRs and 6 sub-bands	M1: 10%	M1: 10.5%		M1: 10.25%
		M2: 8.7%	M2: 9.9%		M2: 9.3%
		M3: 8.1%	M3: 13.1%		M3: 10.6%
		M4: 51%	M4: 47%		M4: 49%
3	3 WACR and 12 sub-bands	M1: 19.7%	M1: 15.9%	M1: 18.2%	M1: 17.93%
		M2: 19.9%	M2: 21.3%	M2: 26.5%	M2: 22.56%
		M3: 21.6%	M3: 28.9%	M3: 26.5%	M3: 25.6%
		M4: 60%	M4: 68%	M4: 63%	M4: 63.3%

track the jammed sub-bands while the goal of the transmission operation was to switch the operating sub-band just before it is jammed. The simulation results showed that the proposed cognitive algorithm has a considerably low probability of getting jammed while maintaining reasonable accumulated reward values even under the impact of sensing errors.

ACKNOWLEDGMENT

This work was funded in part by the Army Research Laboratory under the award W911NF-17-1-0035.

REFERENCES

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

[2] M. A. Aref, S. K. Jayaweera and S. Machuzak, "Multi-agent Reinforcement Learning Based Cognitive Anti-jamming", IEEE Wireless Communications and Networking Conference (WCNC'17), San Francisco, CA, Mar. 2017.

[3] H. M. Schwartz,"Multi-Agent Machine Learning: A Reinforcement Approach," John Wiley & Sons, ISBN: 978-1-118-36208-2, 2014.

[4] B. F. Lo and I. F. Akyildiz, "Multiagent Jamming-Resilient Control Channel Game for Cognitive Radio Ad Hoc Networks," IEEE International Conference on Communications (ICC'12), London, UK, June 2012.

[5] B. Wang, Y. Wu, K. Liu, and T. Clancy, "An anti-jamming stochastic game for cognitive radio networks," IEEE Journal on Selected Areas in Communications, vol. 29, no. 4, Apr. 2011.

[6] Y. Gwon, S. Dastangoo, C. Fossa, and H. T. Kung,"Competing mobile network game: Embracing anti-jamming and jamming strategies with reinforcement learning," IEEE Conference in Communications and Network Security (CNS'13), National Harbor, MD, Oct. 2013.

[7] Y. Gwon, S. Dastangoo, C. Fossa, and H. T. Kung, "Fast online learning of antijamming and jamming strategies," 2015 IEEE Global Communications Conference (GLOBECOM'15), San Diego, CA, Dec. 2015.

[8] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," Artificial Intelligence, vol. 136, no. 2, Apr. 2002.

[9] M. A. Aref and S. K. Jayaweera, "A Novel Cognitive Anti-jamming Stochastic Game," IEEE Cognitive Communications for Aerospace Applications Workshop (IEEE-CCAA'17), Cleveland, OH, Jun. 2017.

[10] M. Bowling and M. Veloso, "Rational and Convergent Learning in Stochastic Games," 17th international joint conference on Artificial intelligence (IJCAI'01), Seattle, WA, Aug. 2001.

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