A Novel Cognitive Anti-jamming 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
†Bluecom Systems and Consulting LLC, Albuquerque, NM.
Email: {maref, jayaweera}@unm.edu

***Abstract*—This paper proposes a new cognitive anti-jamming stochastic game model for multi-agent environments in which each wideband autonomous cognitive radio (WACR) attempts to predict and evade the transmissions of other radios as well as a dynamic jammer signal. The cognitive framework is divided into two operations: sensing and transmission. Each is helped by its own learning algorithm based on Q-learning. It is shown, through both analysis and simulations, that the proposed cognitive anti-jamming technique has low computational complexity and significantly outperforms non-cognitive sub-band selection policy.**

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

I. Introduction

Wideband autonomous cognitive radios (WACRs) can be useful in many applications including, for example, aerospace, military and consumer wireless communications [1]. A common situation in which WACRs can be a great asset is when malicious users launch jamming attacks to disrupt the reliable communications. In practice, there could be multiple WACRs simultaneously operating over the same spectrum band of interest, producing a complicated multi-agent environment. In this case, each WACR needs to avoid the jammer as well as transmissions of other WACRs. In this context, stochastic games can be exploited as a stochastic tool to model the WACR decision-making problem in the presence of both jamming and interference. A WACR may use multi-agent reinforcement learning (MARL) to solve the stochastic game and learn an optimal, or near-optimal, policy to keep its communication link unjammed [1], [2].

MARL has been adopted in the literature [3], [4] for anti-jamming communications in cognitive radio (CR) networks. In [3], a MARL algorithm based on Minimax-Q learning was proposed to find anti-jamming policies for secondary users (SUs) in multi-channel CR systems. There the CR and the jammer were treated as two equally knowledgeable learning agents. One of the drawbacks of the proposed algorithm in [3] is that it assumed perfect sensing. In [4], the authors formulated the competition for open spectrum access as a competitive mobile network game by dividing the network into two sub-networks: the ally network and the enemy network. The objective of each network was to achieve the maximum spectrum utilization while jamming the opponent transmission as much as possible. Thus, each network integrated anti jamming and jamming games to jointly solve for an optimal strategy. Several MARL techniques were proposed in [4], including Minimax-Q, Nash-Q and Friend-or-Foe Q-learning.

In this paper we address the cognitive anti-jamming problem in a multi-agent environment that is modeled as a general-sum stochastic game. The objective of this paper is two-fold: First, introduce new state, action and reward definitions for the proposed stochastic game. Second, obtain optimal, or near-optimal, anti-jamming and interference avoidance policies for each WACR using reinforcement learning (RL). The proposed RL algorithm is based on standard Q-learning algorithm. Although Q-learning is a single-agent learning algorithm, it is often applied in multi-agent problems due to simplicity [5]. One of the most interesting aspects of this work is that we introduce novel state, action and reward definitions that enable 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 [1].

The paper is structured as follows: First, the system model is described in Section II. Section III gives an overview of the Q-learning algorithm. In Section IV, we introduce the proposed cognitive stochastic game for anti-jamming and interference avoidance. The simulation results are presented in Section V. Finally, concluding remarks are given in Section VI.

II. System Model

The wideband spectrum of interest is considered as made of N_b sub-bands with equal bandwidth [6]. Assume M WACRs operating over the N_b sub-bands and challenged by a single jammer as shown in Fig. 1. Each WACR is considered a player in a stochastic game. The game includes a set of states and a collection of action sets denoted by $\mathcal{S}$ and $\mathcal{A}_1, \cdots, \mathcal{A}_M$, respectively. At each stage of the game, all players have to execute an action. 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 $T : \mathcal{S} \times \mathcal{A} \times \cdots \times \mathcal{A}_{\mathcal{M}} \rightarrow PD(\mathcal{S})$, where $PD(\mathcal{S})$ denotes the set of probability distributions defined over $\mathcal{S}$.

The objective of this framework is to obtain optimal or sub-optimal policy that enables each of the WACRs to switch the operating sub-band before getting jammed while also avoiding unnecessary transitions. Our framework consists of two operations: sensing and transmission. Each will have its own learning algorithm with different targets, but they both will experience the same RF environment. The objective of

978-1-5386-3988-7/17/$31.00 ©2017 IEEE

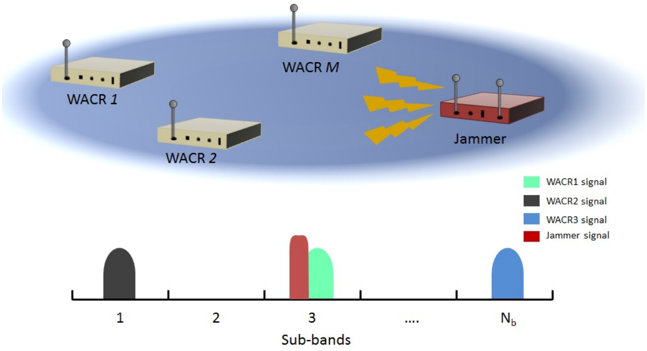

Figure 1. M WACRs operate in the same frequency range challenged by a dynamic jammer.

the sensing operation is to track the jammed sub-bands. On the other hand, the transmission policy determines when and where to switch the operating sub-band. Hence, at any time instant there are two operating 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 sub-bands. This will help the transmission operation as follows: if the current operating sub-band is predicted to be jammed during the next time instant by the sensing policy, the WACR will switch to another sub-band thereby avoiding the possibility of getting jammed.

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 jammed or faces interference, it is considered to be 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 $\mathcal{V} = \{0, 1\}$. For the game state, we choose a simple definition for both sensing and transmission operations, where $s_s[n] \in \mathcal{S}$ and $s_t[n] \in \mathcal{S}$ represent the index of selected sub-bands for sensing and transmission, respectively, at time n. Hence, the number of possible states for each process is N_b, where N_b is the total number of sub-bands.

At any time instant, the state of operating sub-bands for both sensing and transmission (the value of $v \in \mathcal{V}$ for sub-band index $s \in \mathcal{S}$) has to be identified. During sensing operation, the WARC will perform spectral activity detection to detect any active signals in the sensed sub-band and hence identify whether the sub-band is available or not [6], [1]. During transmission operation, the link quality will determine if transmission over the current operating sub-band is acceptable. After determining the states of both operating sub-bands, the WACR will select and execute actions for both operations. We define actions $a_s[n]$ and $a_t[n]$ as the new indices of the new operating sub-bands for sensing and transmission, respectively, at time n. The action space can thus be defined as $\mathcal{A} = \{1, \cdots, N_b\}$.

The objective of this stochastic game is to find the optimal or sub-optimal policies for WACRs to predict and avoid jamming attacks as well as interference from other radios. The players in this game are the WACRs, in which they are competing to maximize their rewards for 2 different operations: sensing and transmission.

III. Q-Learning-Aided Cognitive Anti-Jamming Algorithm

The basic idea of the Q-learning algorithm is to maintain a table that contains what are called Q-values denoted by $Q(s, a)$ representing a measure of goodness of taking the action a when in state s [7]. 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-tables for both sensing and transmission operations will be $N_b \times N_b$.

A summary of the Q-learning-aided proposed cognitive anti-jamming algorithm is listed in Algorithm 1. At any given time n, the WACR has to identify the state of the current operating sub-band (lines 2-3). If the sub-band state is “1” (available), no further action is required. If the sub-band state is “0” (not-available), the WACR updates the Q-table, based on a certain observed reward (lines 4-7), where $\alpha \in (0, 1)$ is the learning rate and $\gamma \in [0, 1)$ is a discount factor. Note that the sub-band state will be “0” if the sub-band is getting jammed or an interference signal is present. Once the Q-table is updated, the WACR selects a new action a' representing the new operating sub-band according to line 9. The exploration parameter $\epsilon \in (0, 1)$ allows the WACR to switch between selecting the action characterized by $\underset{a\in\mathcal{A}}{\arg\max}\ Q(s', a)$ or randomly selecting an action according to function $U(\mathcal{A})$ where $U(\mathcal{A})$ denotes the uniform distribution over the action set $\mathcal{A}$.

IV. Proposed anti-jamming stochastic game

As mentioned earlier, each WACR performs two operations: sensing and transmission. Each of these operations has its own Q-learning algorithm. Thus, there are two Q-tables to be updated at every iteration. Figure 2 shows both sensing and transmission operations for a given WACR. In our approach, the goal for sensing operation is to learn the behavior of

Algorithm 1 Q-learning-aided cognitive anti-jamming communications algorithm

```
1: Initialize:
       $\alpha,\ \gamma,\ \epsilon \in [0, 1]$
       $Q(s, a) \leftarrow 0\ \forall s \in \mathcal{S},\ \forall a \in \mathcal{A}$
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:         $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha[r + \gamma \underset{a}{\max} Q(s', a)]$
8:         Select new action $a' \in \mathcal{A}$ for the new state $s'$ according to the following:
9:         $a' = \begin{cases} \underset{a\in\mathcal{A}}{\arg\max}\ Q(s', a) & \text{with probability } 1 - \epsilon, \\ \sim U(\mathcal{A}) & \text{with probability } \epsilon, \end{cases}$
```

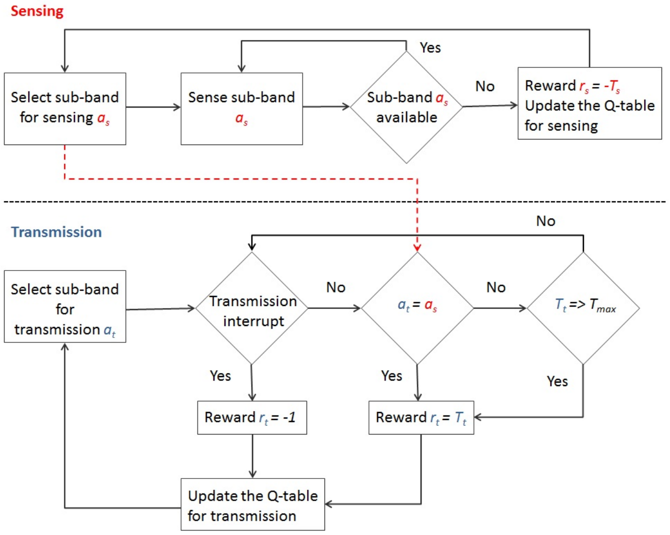

Figure 2. Proposed cognitive radio operations for sensing and transmission.

jammer and other WACRs in its vicinity by using the Q-learning. Ideally, the WACR should predict and sense the sub-band where the jammer or interference signals are located with the highest probability. For every new selected action (new sub-band), the WACR computes the time it takes until the jammer or interference signal arrives, denoted by T_s. The reward is defined as the negative of this value $r_s = -T_s$. The actions are selected such that rewards are maximized. Hence, they are selected corresponding to the shortest time it takes until the operating sub-band gets jammed. Note, during sensing operation, the WACR will switch the operating sub-band if and only if it gets jammed or faces interference.

In the transmission operation, changing the operating sub-band maybe triggered by two possible conditions. First is if the transmission is interrupted meaning that the current operating sub-band is either jammed or facing an interference signal. This is the most undesirable situation since our objective is to switch the operating sub-band before getting jammed. Hence, we assign a reward of $r_t = -1$ for this scenario. The second condition is when the sensing operation predicts that the current operating sub-band for transmission will be most likely getting jammed. The reward for this case is defined as the time that WACR kept transmitting over the sub-band before switching to a new one, denoted by $r_t = T_t$. The action is then selected such that the reward is maximized. Thus, the selected new sub-band for transmission must have low interference for the longest amount of time with high probability. In order to further reduce the probability of getting jammed, we set a threshold denoted by T_{max} such that the transmission time over a certain sub-band cannot exceed T_{max}.

V. Simulation Results

In this section, we will compare performance to a non-cognitive (random) sub-band selection scheme in which all sub-bands are selected for transmission with equal probabilities. As our performance metric, we use the normalized accumulated reward, defined as

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

where $r_t(s_t[n], a_t[n])$ represents the immediate reward of taking action $a_t[n]$ when in state $s_t[n]$ for transmission operation and N is the number of iterations. Note that, the rewards in this case are those that achieved after the convergence of the Q-table.

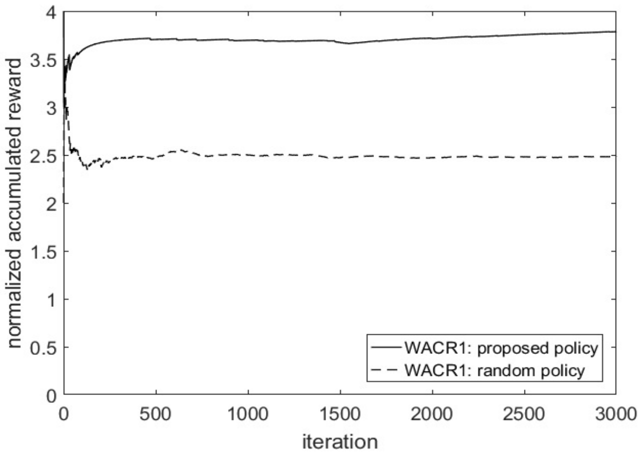

Figure 3. Normalized accumulated reward values for test case 1.

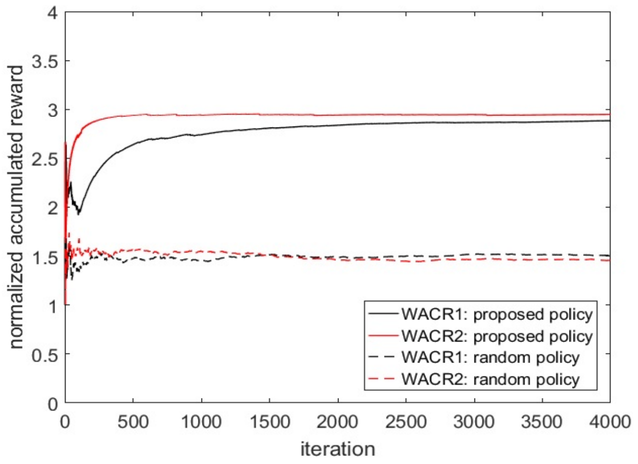

Figure 4. Normalized accumulated reward values for test case 2.

Three experiments are considered with different numbers of WACRs and different numbers of sub-bands. In all experiments a sweeping jammer, that sweeps the spectrum of interest from the lower to the higher frequency is considered. Tables I and II summarize the obtained values of normalized accumulated reward and probability of getting jammed for all test cases, respectively. The first experiment includes 1 WACR that operates over 5 sub-bands. The maximum possible reward in this case should be 4 time steps since there are 5 sub-bands in the system. Figure 3 shows that the proposed policy achieves

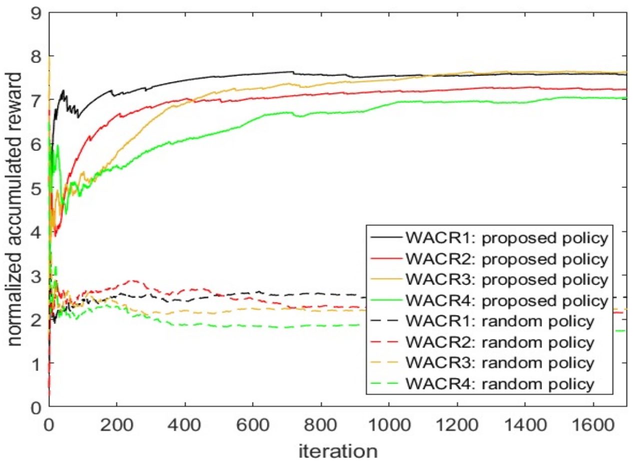

Figure 5. Normalized accumulated reward values for test case 3.

about 97% of the maximum possible reward, while the random policy achieves only about 62%.

Figure 4 shows the normalized accumulated reward for the second experiment in which 2 WACRs and 6 sub-bands are considered. The achieved accumulated reward of the proposed policy for both WACRs lies somewhere between 70% to 75% of the maximum possible reward. On the other hand, the random selection policy achieves only about 36% of the maximum possible performance. From Table II, we may observe that the proposed algorithm results in a very low probability of getting jammed, while the random policy has a 47% of probability of getting jammed. Finally, the third experiment includes 4 WACRs operating over 16 sub-bands. From Fig. 5 and Table I, the proposed algorithm achieves an acceptable normalized accumulated reward value between 58% to 63% of the maximum possible reward. The random policy on the other hand, achieves in average only 18% of the maximum possible reward. Moreover, the proposed policy significantly outperforms the random policy in terms of the probability of getting jammed as shown in Table II.

VI. Conclusion

In this paper we have proposed a novel cognitive anti-jamming stochastic game based on Q-learning that allows each WACR to predict and avoid jamming attacks as well as interference from other radios. Each WACR has to perform two operations: sensing and transmission. The objective of the sensing operation is to track the jammed sub-bands. On the other hand, the transmission operation determines when and where to switch the operating sub-band. When compared with random sub-band selection policy, simulation results showed that the proposed cognitive protocol has a very low probability of getting jammed and acceptable value for accumulated reward.

Acknowledgment

This work was funded in part by the Air Force Research Laboratory, Space Vehicles Directorate, under the grant FA9453-16-1-0052 and in part by the NASA STTR Phase II contract NNX17CC01C.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

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

Table I
Normalized accumulated reward values for different simulation scenarios

Test case	Scenario	Reward upper bound	WACR 1	WACR 2	WACR 3	WACR 4	Average
1	1 WACR and 5 sub-bands	4	Proposed:3.8 Random: 2.5				Proposed:3.8 Random: 2.5
2	2 WACRs and 6 sub-bands	4	Proposed:2.8 Random: 1.5	Proposed:3 Random: 1.4			Proposed:2.9 Random: 1.45
3	4 WACR and 16 sub-bands	12	Proposed:7.5 Random: 2.5	Proposed:7.2 Random: 2.2	Proposed:7.5 Random: 2.2	Proposed:7 Random: 1.8	Proposed:7.3 Random: 2.17

Table II
Probabilities of getting jammed for different simulation scenarios

Test case	Scenario	WACR 1	WACR 2	WACR 3	WACR 4	Average
1	1 WACR and 5 sub-bands	Proposed: 0.86% Random: 1.8%				Proposed: 0.86% Random: 1.8%
2	2 WACRs and 6 sub-bands	Proposed: 2.6% Random: 47.2%	Proposed: 2.1% Random: 48%			Proposed: 2.35% Random: 47.6%
3	4 WACR and 16 sub-bands	Proposed: 6.4% Random: 64.8%	Proposed: 7.6% Random: 66.3%	Proposed: 12.4% Random: 66.3%	Proposed: 12.3% Random: 72.6%	Proposed: 9.6% Random: 67.5%