An Anti-Jamming Stochastic Game for Cognitive Radio Networks

Beibei Wang, *Student Member, IEEE,* Yongle Wu, *Student Member, IEEE,* K. J. Ray Liu, *Fellow, IEEE,* and T. Charles Clancy, *Member, IEEE*

***Abstract*—Various spectrum management schemes have been proposed in recent years to improve the spectrum utilization in cognitive radio networks. However, few of them have considered the existence of cognitive attackers who can adapt their attacking strategy to the time-varying spectrum environment and the secondary users' strategy. In this paper, we investigate the security mechanism when secondary users are facing the jamming attack, and propose a stochastic game framework for anti-jamming defense. At each stage of the game, secondary users observe the spectrum availability, the channel quality, and the attackers' strategy from the status of jammed channels. According to this observation, they will decide how many channels they should reserve for transmitting control and data messages and how to switch between the different channels. Using the minimax-Q learning, secondary users can gradually learn the optimal policy, which maximizes the expected sum of discounted payoffs defined as the spectrum-efficient throughput. The proposed stationary policy in the anti-jamming game is shown to achieve much better performance than the policy obtained from myopic learning, which only maximizes each stage's payoff, and a random defense strategy, since it successfully accommodates the environment dynamics and the strategic behavior of the cognitive attackers.**

***Index Terms*—Security mechanism, spectrum management, cognitive radio networks, game theory, reinforcement learning.**

I. Introduction

IN RECENT years, cognitive radio technology [1] [2] [3] has been proposed as a promising communication paradigm to solve the conflict between the limited spectrum resources and the increasing demand for wireless services. By exploiting the spectrum in an opportunistic fashion, cognitive radio enables secondary users to sense which portion of the spectrum is available, select the best available channel, coordinate the spectrum access with other users, and vacate the channel when a primary user reclaims the spectrum usage right. In order to utilize the spectrum resources efficiently, various spectrum management approaches have been proposed in the literature, such as the pricing-based spectrum sharing approaches [5]–[13], where primary users lease the available spectrum bands to secondary users, and the opportunistic spectrum sharing approaches based on sensing and stochastic modeling about the primary user's access [14]–[16].

Manuscript received 1 December 2009; revised 18 May 2010.

B. Wang is with Corporate Research and Development, Qualcomm Incorporated, San Diego, CA 92121, USA (e-mail: beibei.bbwang@gmail.com).

Y. Wu is with Qualcomm Incorporated, San Diego, CA 92121, USA (e-mail: wuyongle@gmail.com).

K. J. R. Liu is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA (e-mail: kjrliu@umd.edu).

T. C. Clancy is with the Virginia Tech Hume Center for National Security and Technology, Alexandria, VA 22314, USA (e-mail: tcc@vt.edu).

Digital Object Identifier 10.1109/JSAC.2011.110418.

Although these proposed approaches have been shown to be able to improve the spectrum utilization or bring monetary gains for the primary users, most of them are based on the assumption that the users only aim at maximizing the spectrum utilization, either in a cooperative way where all users are coordinated by the same network controller and serve a common goal, or in a selfish manner where the autonomous secondary users want to maximize their own benefit. However, such an assumption does not hold when the secondary users are in a hostile environment, where there exist malicious attackers whose objective is to cause damage to the legitimate users and prevent the spectrum from being utilized efficiently. Therefore, how to secure spectrum sharing is of critical importance to the wide deployment of the cognitive radio technology.

Malicious attackers can launch various types of attacks in different layers of a cognitive radio network [4]. In [17], the authors studied the primary user emulation attack, where the cognitive attackers mimic the primary signal to prevent secondary users from accessing the licensed spectrum. Localization-based defense mechanism was proposed, which verifies the source of the detected signals by observing the signal characteristics and estimating its location. The work in [18] investigated the spectrum sensing data falsification attack, and proposed a weighted sequential probability ratio test to alleviate the performance degradation due to sensing error. Other possible security issues such as denial of service attacks in cognitive radio networks are discussed in [19] and [20]. However, most of these works [19][20] only provide qualitative analysis about the countermeasures, and have not considered the real dynamics in the spectrum environment and the cognitive attackers' capability to adjust their attacking strategy.

In this work, we focus on the jamming attack in a cognitive radio network and propose a stochastic game framework for anti-jamming defense design, which can accommodate the dynamic spectrum opportunity, channel quality, and both the secondary users and attackers' strategy changes. The jamming attack has been extensively studied in wireless networking, and existing anti-jamming solutions include physical layer defenses, such as directional antennas [22] and spread spectrum [23], link-layer defenses such as channel hopping [25][26][27][28], and network-layer defenses, such as spatial retreats [29]. However, they are not directly applicable to cognitive radio networks, since the spectrum availability keeps changing with the primary users returning/vacating the licensed bands. For instance, the work in [28] proposed to use error-correcting codes (n, m) to ensure reliable data

communications with a high throughput. However, this approach requires that at each time there are at least n channels available, which may not be satisfied if many licensed bands are occupied by primary users.

Moreover, most of the works assume that the attackers adopt a fixed strategy that will not change with time. However, if the attackers are also equipped with cognitive radio technology, it is highly likely that they will adapt their attacking strategy according to the environment dynamics as well as the secondary users' strategy. Therefore, in our work, we model the strategic and dynamic competition between the secondary users and the cognitive attackers as a zero-sum stochastic game. In order to ensure reliable transmission, we propose to reserve multiple channels for transmitting control messages, and the control channels should be switched with the data channels from time to time, according to the attackers' strategy. We define the spectrum availability, the channel quality, and the observation about the attackers' action as the state of the game. The secondary users' action is defined as how many control or data channels they should reserve and how to switch between the control and data channels, and their objective is to maximize the spectrum-efficient throughput, defined as the ratio between the expected achievable throughput over the total number of active channels used for transmitting control and data messages.

Using the minimax-Q learning algorithm, the secondary users can obtain the optimal policy, with a proved convergence. Simulation results show that when the channel quality is not good, the secondary users should reserve a lot data channels and a few control channels to improve the throughput. As the channel quality becomes better, they should reserve more control channels to ensure communication reliability. When the channel quality further increases, the secondary users should be more conservative by reserving less data channels to improve the spectrum-efficient throughput. At the states when some control or data channels are observed to be jammed, the secondary users should adopt a mixed strategy to avoid being severely jammed next time. When there are more than one licensed band available, the attackers' decision making becomes more difficult, and the secondary users can take more aggressive action by having more data channels. It is also shown that the secondary users can achieve a higher payoff using the stationary policy learned from the minimax-Q learning than using myopic learning and a random strategy.

The remaining of the paper is organized as follows. In Section II, we introduce the system model about the secondary user network and the anti-jamming defense. In Section III, we formulate the anti-jamming defense as a stochastic game by defining the states, actions, objective functions, and the state transition rules. In Section IV, we obtain the optimal policy of the secondary user network using the minimax-Q learning algorithm. In Section V we present the simulation results, followed by conclusions in Section VI.

II. System Model

In this section, we present the model assumptions about the secondary user network and the anti-jamming defense against the malicious attackers.

A. Secondary User Network

In this paper, we consider a dynamic spectrum access network where multiple secondary users equipped with cognitive radio are allowed to access temporarily-unused licensed spectrum channels that belong to multiple primary users. There is a secondary base station in the network, which coordinates the spectrum usage of all secondary users. In order to avoid conflict or harmful interference to the primary users, the secondary users need to listen to the spectrum before every attempt of transmission. We assume the secondary network is a time-slotted system, and at the beginning of each time slot, secondary users need to reserve a certain time to detect the presence of a primary user. Various detection techniques are available, such as energy detection, or feature detection if the secondary users know some prior information about the primary users' signal. In cooperative spectrum sharing such as a spectrum auction, secondary users can avoid harmful interference by listening to the primary users' announcement about whether they would share the licensed channels with the secondary users. To simplify analysis, we assume perfect sensing or cooperative spectrum sharing in this work. Therefore, the secondary user network can take every opportunity to utilize the currently unused licensed spectrum, and vacate the spectrum whenever a primary user reclaims the spectrum rights.

Due to the primary users' activity and channel variations, the spectrum availability and quality keep changing. In order to coordinate the spectrum usage and achieve efficient spectrum utilization, necessary control messages need to be exchanged between the secondary base station and the secondary users through dedicated control channels[1]. Control channels serve as a medium that can support high-level network functionality, such as access control, channel assignment, spectrum handoff, etc. If the control messages are not correctly received by the secondary users or base station, certain network functions will get impaired.

B. Anti-Jamming Defense in Cognitive Radio Networks

Radio jamming is a Denial of Service (DoS) attack which targets at disrupting communications at the physical and link layers of a wireless network. By keeping the wireless spectrum busy, e.g., constantly injecting packets to a shared spectrum [25], a jamming attacker can prevent legitimate users from accessing an open spectrum band. Another type of jamming is to inject high interference power around the vicinity of a victim [21] [29], so that the signal to noise ratio (SNR) deteriorates heavily and no data can be received correctly.

In a cognitive radio network, malicious attackers can launch jamming attack to prevent efficient utilization of the spectrum opportunities. In this paper, we assume that the characteristics of the transmitted signal by the primary users and the secondary users are distinguishable, and the attackers also listen to the licensed band when the secondary users are sensing the spectrum. The attackers will jam the secondary

[1]Many wireless networks employ control channels for sending system control information [38], e.g., the GSM cellular communication system has multiple control channels, which are located at very specific time slots and physical frequency band.

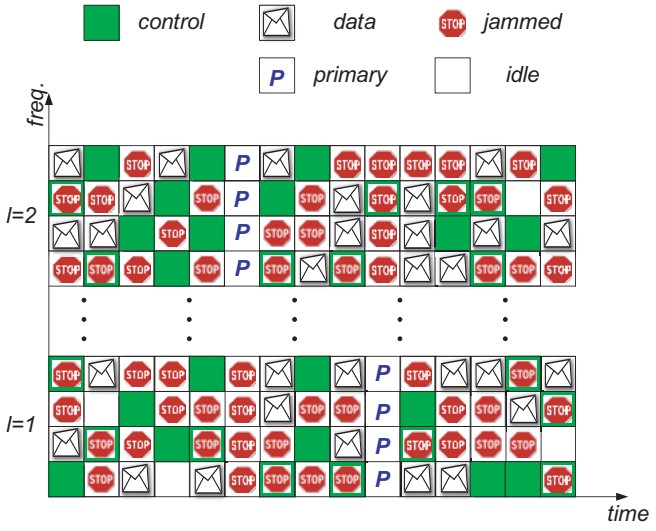


Fig. 1. Illustration of the anti-jamming defense.

users' transmission, while will not jam the licensed bands when the primary users are active, either because there may be a very heavy penalty on the attackers if their identities are known by the primary users, or because the attackers cannot get close to the primary users. Moreover, due to the limitation of the number of antennas and/or the total power, we assume that the attackers can jam at most $\bar{N}$ channels in each time slot. Then, the objective of the attackers is to cause the most damage to the secondary user network with the limited jamming capability.

Given the limited jamming capability, the attackers can adopt an attacking strategy that targets at as many data channels as possible to reduce the gain of the secondary user network by transmitting data. On the other hand, if the number of control channels is less than $\bar{N}$, while the number of data channels is greater than $\bar{N}$, the attackers can try to target at the control channels to make the attack even more powerful. If the secondary user network adopts a fixed channel assignment scheme for transmitting data and control messages, a cognitive attacker can capture such a pattern[2], distinguish between the data channels and control channels, and target at only the data or control channels and cause the highest damage.

Therefore, secondary users need to perform channel hopping/switching to alleviate the potential damage due to a fixed channel assignment schedule. As shown in Figure 1, the channels that are used for transmitting data/control messages in this time slot may no longer be data/control channels in the next time slot. By introducing randomness in their channel assignment, secondary users' access pattern becomes more unpredictable. Then, the attackers also have to strategically change the channels they will attack with time. Therefore, channel hopping is more resistant to the jamming attack than a fixed channel assignment.

When designing the channel hopping mechanism in a cognitive radio network, the secondary users need to take the following facts into consideration.

- *There is a tradeoff in choosing a proper number of control channels.* The secondary network functionality relies heavily upon the correct reception of control messages. Thus, it is more reliable to transmit duplicate control messages in multiple channels (i.e., control channels). However, if the secondary user network reserves too many control channels, the number of channels where data messages are transmitted (i.e., data channels) will be small, and the achievable gain through utilizing the licensed spectrum will be unnecessarily low. Therefore, a good selection should be able to balance the risk of having no control messages successfully received and the gain of transmitting data messages. To make the defense mechanism more general, we assume that the secondary user network can choose to transmit nothing in some channels even when the licensed band is available. This is because when the secondary base station believes it has reserved enough data or control channels under very severe jamming attack, allocating more channels for transmitting messages can only result in a waste the energy, and it will be better the leave some channels as idle, if the energy consumption is a concern of the secondary user network.
- *The channel hopping mechanism must be adaptive to the attackers' strategy.* This is because the attackers may also be equipped with cognitive radio technology and adjust their strategies based on the observation about the spectrum environment dynamics and the secondary users' strategy. Thus, the secondary users cannot pre-assume that the attackers will adopt a fixed attack strategy. Instead, they need to build a stochastic model that captures the dynamic strategy adjustment of the attackers, as well as the spectrum environment variations.

According to the above-mentioned assumptions about the system model and the jamming attack, we know that the secondary users aim at maximizing the spectrum utilization with carefully-designed channel switching schedules, while the malicious attackers want to decrease the spectrum utilization by strategic jamming. Therefore, they have opposite objectives and their dynamic interactions can be well modeled as a noncooperative (zero-sum) game[3]. As we assume that the spectrum access of all the secondary users is coordinated by the secondary base station, and the malicious users work together to cause the most damage to the secondary users, we can view all the secondary users in the network as one player, and all the attackers as another player. Moreover, considering that the spectrum opportunity, channel quality, and both the secondary users and malicious attackers' strategies are changing with time, the noncooperative game should be considered in a stochastic setting, i.e., the dynamic anti-jamming defense in the secondary user network should be formulated as a stochastic game.

[2]Since control messages may have distinguishable features from data messages, for instance, different lengths, headers, and acknowledgement, the attackers can determine whether the jammed channels are control or data channels after jamming for a number of time slots. Similar assumptions can be found in [39].

[3]Note that if the individual cost of the attacker, e.g. cost due to energy consumption of jamming, is a concern of the attackers, the payoff of the attackers will not be the negative of the secondary users' payoff, and the game is better modeled as a general-sum game. However, to simplify analysis, in this paper we only discuss the zero-sum stochastic game and case of general-sum games can be studied in a similar way.

III. Stochastic Anti-Jamming Game Formulation

Before we go into details of the stochastic anti-jamming game formulation, let us first introduce the stochastic game to get a general idea. A stochastic game [31][36][37] is an extension of Markov Decision Process (MDP) [32] by considering the interactive competition among different agents. In a stochastic game $\mathbb{G}$, there is a set of states, denoted by $\mathcal{S}$, and a collection of action sets, $\mathcal{A}_1, \cdots, \mathcal{A}_k$, one for each player in the game. The game is played in a sequence of stages. At the beginning of each stage the game is in some state. After the players select and execute their actions, the game then moves to a new random state with transition probability determined by the current state and one action from each player: $T : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_k \mapsto PD(\mathcal{S})$. Meanwhile, at each stage each player receives a payoff $R_i : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_k \mapsto \mathbb{R}$, which also depends on the current state and the chosen actions. The game is played continually for a number of stages, and each player attempts to maximize his/her expected sum of discounted payoffs, $E\{\sum_{j=0}^{\infty} \gamma^j r_{i,t+j}\}$, where $r_{i,t+j}$ is the reward received j steps into the future by player i and γ is the discount factor.

After introducing the concepts of a stochastic game, we next formulate the anti-jamming game by defining each component of the game.

A. States and Actions

We consider a spectrum pooling system, where the secondary user network can use the temporarily unused spectrum bands that belong to L primary users. As the bandwidth of different licensed bands may be different, we assume that each licensed band is divided into a set of adjacent channels with the same bandwidth. Then, there are N_l channels in primary user l's band, and we assume all of them will be occupied/released when primary user l reclaims/vacates the band. Then, we can denote primary user l's states in the l-th band at time t as P_l^t, whose value can be either $P_l^t = 1$, meaning primary user l is active at time t, or $P_l^t = 0$, meaning primary user l will not use the licensed band at time t and the secondary users can access the channels in the l-th band. According to some empirical studies on the primary users' access pattern [30], the states P_l^t can be modeled by a two-state Markov chain, where the transition probabilities are denoted by $p_l^{1\to1} = p(P_l^{t+1} = 1 | P_l^t = 1)$ and $p_l^{0\to1} = p(P_l^{t+1} = 1 | P_l^t = 0)$.

The secondary user network will achieve a certain gain by utilizing the spectrum opportunity on the licensed bands. The gain can be defined as a function of the data throughput, packet loss, delay, or other proper Quality of Service (QoS) measure, and is often an increasing function of the channel quality. Due to the channel variations on each licensed band, the channel quality may change from one time slot to another, so the gain of utilizing a licensed band also changes over time. We assume that the gain of each channel within the same licensed band l is identical at any time t, and it can take any value from a set of discrete values, i.e., $g_l^t \in \{q_1, q_2, \cdots, q_n\}$. Since the channel quality (in terms of SNR) is often modeled as a finite-state Markov chain (FSMC) [24], the dynamics of the l-th licensed band's gain g_l^t can also be expressed by an FSMC. Note that the achievable gain of utilizing the licensed bands also depends on the primary users' status, i.e., when the primary user is active in the l-th band ($P_l^t = 1$), the secondary users are not allowed to access band l, and thus $g_l^t = 0$. So the state of the FSMC should be able to capture the joint dynamics of both the primary users' access and the channel quality, which can be denoted by (P_l^t, g_l^t).

The transition probability of the FSMC with states (P_l^t, g_l^t) can be derived as follows. When the l-th licensed band is not available for two consecutive time slots, the transition depends only on the primary users' access pattern, so we have

$$p(P_l^{t+1} = 1, g_l^{t+1} = 0 | P_l^t = 1, g_l^t = 0) = p_l^{1\to1}. \quad (1)$$

When the l-th band becomes available with gain q_n at time $t+1$, we have

$$p(P_l^{t+1} = 0, g_l^{t+1} = q_n | P_l^t = 1, g_l^t = 0) = (1 - p_l^{1\to1}) p_{g_l}^{0\to n}, \quad (2)$$

where $p_{g_l}^{0\to n}$ denotes the probability that the gain of band l is q_n at time $t+1$, given that $P_l^t = 1$ and $P_l^{t+1} = 0$. When the l-th band is available for two consecutive time slots, we have the state transition probability as

$$p(P_l^{t+1} = 0, g_l^{t+1} = q_n | P_l^t = 0, g_l^t = q_m) = (1 - p_l^{0\to1}) p_{g_l}^{m\to n}, \quad (3)$$

where $p_{g_l}^{m\to n}$ is the probability that the gain transits from q_m at time t to q_n at time $t+1$. Finally, when the l-th band turns unavailable from time t to time $t+1$, the transition probability is

$$p(P_l^{t+1} = 1, g_l^{t+1} = 0 | P_l^t = 0, g_l^t = q_m) = p_l^{0\to1}, \quad (4)$$

since the transition does not depend on the the gain g_l^t at time t.

In the above, we have discussed the dynamics of primary users' returning/vocating the licensed bands and the gains of utilizing the licensed spectrum. Clearly, these dynamics will affect the secondary users' decisions about how to allocate the channels for transmitting control and data messages. For instance, in order to obtain higher utilization of the spectrum opportunities, the secondary users tend to allocate more channels with higher gains as data channels and those with lower gains as control channels. However, their channel allocation decisions should also depend on the observations about the malicious attackers' strategies, which can be conjectured from the channels that get jammed by the attackers. Thus, the secondary users should maintain a record about *which channels have been jammed by the attackers* and *what type of messages have been transmitted in the jammed channels*. Since the channels within the same licensed band are assumed to have the same gain, what matters to the secondary users is only the number and the type of the jammed channels. Based on these assumptions, the observations of the secondary user network are denoted by $\{J_{l,C}^t, J_{l,D}^t\}$, where $J_{l,C}^t$ and $J_{l,D}^t$ denotes the number of control and data channels that get jammed in the l-th band observed at time slot t, and $l \in \{1, 2, \cdots, L\}$. Such observation can be obtained when the secondary users do not receive a confirmation about message receipt from the receiver. The secondary users cannot tell whether an idle channel gets jammed or not, since no messages are transmitted in those

channels. Thus, the number of idle channels that get jammed is not an observation of the secondary users, and will not be considered in the state of the stochastic game. In summary, the state of the stochastic anti-jamming game at time t is defined by $\mathbf{s}^t = \{s_1^t, s_2^t, \cdots, s_L^t\}$, where $s_l^t = (P_l^t, g_l^t, J_{l,C}^t, J_{l,D}^t)$ denotes the state associated with the l-th band.

After observing the state at each stage, both the secondary users and the attackers will choose their actions for the current time slot. The secondary users may no longer choose the previously jammed channels as control or data channels if they believe that the attackers will stay in the jammed channels until they detect no activity of the secondary users. On the other hand, if the attackers believe that the secondary users will hop away from the jammed channels, they will choose the previously un-attacked channels to jam; then for the secondary users, staying still in the previously jammed channels may be a better choice. When facing such uncertainty about each other's strategy, both the secondary users and the attackers should adopt a randomized strategy. The secondary users will still transmit control or data messages in part of the previously jammed channels in case that the attackers are more likely to jam the previously un-attacked channels, and start transmitting in part of the previously un-attacked channels in case that the attackers are more likely to keep jamming the previously jammed channels for a while. Similarly, the attackers will keep jamming some of the previously jammed channels and start to jam the channels that were not jammed in the previous time slot.

In addition, as discussed in Section II, the secondary users may need to perform channel switching to make their channel access pattern more unpredictable to the attackers and alleviate the potential damage due to jamming. Thus, at every time the secondary users can switch a control channel to a data or an idle channel, and vice versa. If so, when there are N_l channels in each licensed band l, the secondary users will have 3^{N_l} different actions to choose from on the l-th band and $\prod_{l=1}^{L} 3^{N_l}$ actions in total. This will complicate the decision making of the secondary users. To have the decision making computable in a reasonable time, we formulate the action set for both players as follows. Note that more complicated action modeling will only affect the performance, while not affecting the stochastic anti-jamming game framework.

Mathematically, the actions of the secondary users are defined as $\mathbf{a}^t = \{a_1^t, a_2^t, \cdots, a_L^t\}$, with $a_l^t = (a_{l,C_1}^t, a_{l,D_1}^t, a_{l,C_2}^t, a_{l,D_2}^t)$, where action a_{l,C_1}^t (or a_{l,D_1}^t) means that the secondary network will transmit control (or data) messages in a_{l,C_1}^t (or a_{l,D_1}^t) channels uniformly selected from the previously un-attacked channels, and action a_{l,C_2}^t (or a_{l,D_2}^t) means that the secondary network will transmit control (or data) messages in a_{l,C_1}^t (or a_{l,D_1}^t) channels uniformly selected from the previously jammed channels. Similarly, the actions of the attackers are defined as $\mathbf{a}_J^t = \{a_{1,J}^t, a_{2,J}^t, \cdots, a_{L,J}^t\}$, with $a_{l,J}^t = (a_{l,J_1}^t, a_{l,J_2}^t)$, where action a_{l,J_1}^t (or a_{l,J_2}^t) means that the attackers will jam a_{l,J_1}^t (or a_{l,J_2}^t) channels uniformly selected from the previously un-attacked (or attacked) channels at current time t. It can be seen that the above choice of actions has modeled the players' uncertainty about each other's strategy on the jammed and un-jammed channels, as well as the need for channel switching.

B. State Transitions and Stage Payoff

With the state and action space defined, we next discuss the state transition rule. We assume that the players choose their actions in each band independently, then the transition probability can be expressed by

$$p(\mathbf{s}^{t+1}|\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) = \prod_{l=1}^{L} p(s_l^{t+1}|s_l^t, a_l^t, a_{l,J}^t). \tag{5}$$

Since the dynamics of the primary users' activity and the channel variations are supposed to be independent of the players' actions, the transition probability $p(s_l^{t+1}|s_l^t, a_l^t, a_{l,J}^t)$ can be further separated into two parts, i.e.

$$\begin{aligned} p(s_l^{t+1}|s_l^t, a_l^t, a_{l,J}^t) = & p(J_{l,C}^{t+1}, J_{l,D}^{t+1}|J_{l,C}^t, J_{l,D}^t, a_l^t, a_{l,J}^t) \\ & \times p(P_l^{t+1}, g_l^{t+1}|P_l^t, g_l^t), \end{aligned} \tag{6}$$

where the first term on the right hand side of (6) represents the transition probability of the number of jammed control and data channels, and the second term represents the the transition of the primary user status and the channel condition. As the second term has been derived in (1)-(4), we only need to derive the first term for different cases.

Case 1: $P_l^t = 1$. As discussed in Section II, we assume that the attackers will not jam the licensed bands when the primary users are active; then, when the l-th band is occupied by the primary user at time slot t, i.e., $P_l^t = 1$, the action of the attackers will be $a_{l,J}^t = (0, 0)$, and the state variable $J_{l,C}^{t+1}$ and $J_{l,D}^{t+1}$ will be 0. Therefore, when $P_l^t = 1$, we have

$$\begin{aligned} p(s_l^{t+1}|s_l^t, a_l^t, a_{l,J}^t) = & p(P_l^{t+1}, g_l^{t+1}|P_l^t, g_l^t), \\ & \text{if } J_{l,C}^{t+1} = 0 \text{ and } J_{l,D}^{t+1} = 0. \end{aligned} \tag{7}$$

Case 2: $P_l^t = 0$. When the l-th band is available to the secondary users, according to the observation $J_{l,C}^t$ and $J_{l,D}^t$ at time t about the jammed channel status in the previous time slot, the secondary network will choose an action $a_l^t = (a_{l,C_1}^t, a_{l,D_1}^t, a_{l,C_2}^t, a_{l,D_2}^t)$, and the attackers choose an action $a_{l,J}^t = (a_{l,J_1}^t, a_{l,J_2}^t)$. As the jammed control (or data) channels at the next time slot $t+1$ include those control (or data) channels that the secondary network has selected from both the previously un-jammed and jammed channels, when deriving the transition $p(J_{l,C}^{t+1}, J_{l,D}^{t+1}|J_{l,C}^t, J_{l,D}^t, a_l^t, a_{l,J}^t)$, we need to consider all possible pairs of (n_{C_1}, n_{C_2}) and (n_{D_1}, n_{D_2}), where n_{C_1} (or n_{D_1}) denotes the number of jammed control (or data) channels that are previously un-jammed, n_{C_2} (or n_{D_2}) denotes the number of jammed control (or data) channels that are previously jammed, with $n_{C_1} + n_{C_2} = J_{l,C}^{t+1}$, and $n_{D_1} + n_{D_2} = J_{l,D}^{t+1}$. Given that the secondary users uniformly choose a_{l,C_1}^t (or a_{l,D_1}^t) channels as control (or data) channels out of the un-jammed $N_l - J_{l,C}^t - J_{l,D}^t$ channels, and the attackers uniformly jam a_{l,J_1}^t channels, the probability that n_{C_1} control channels and n_{D_1} data channels get jammed at time t can be written by

$$\begin{aligned} & p(n_{C_1}, n_{D_1}|J_{l,C}^t, J_{l,D}^t, a_l^t, a_{l,J}^t) \\ & = \frac{\binom{a_{l,C_1}^t}{n_{C_1}} \binom{a_{l,D_1}^t}{n_{D_1}} \binom{N_{l,1}^t - a_{l,C_1}^t - a_{l,D_1}^t}{a_{l,J_1}^t - n_{C_1} - n_{D_1}}}{\binom{N_{l,1}^t}{a_{l,J_1}^t}}, \end{aligned} \tag{8}$$

where $N_{l,1}^t = N_l - J_{l,C}^t - J_{l,D}^t$. Similarly, the transition probability of n_{C_2} and n_{D_2} is expressed as

$$p(n_{C_2}, n_{D_2}|J_{l,C}^t, J_{l,D}^t, a_l^t, a_{l,J}^t) = \frac{\binom{a_{l,C_2}^t}{n_{C_2}}\binom{a_{l,D_2}^t}{n_{D_2}}\binom{N_{l,2}^t - a_{l,C_2}^t - a_{l,D_2}^t}{a_{l,J_2}^t - n_{C_2} - n_{D_2}}}{\binom{N_{l,2}^t}{a_{l,J_2}^t}}, \quad (9)$$

where $N_{l,2}^t = J_{l,C}^t + J_{l,D}^t$ denotes the number of jammed channels. Then, the transition probability of $J_{l,C}^t$ and $J_{l,D}^t$ becomes

$$p(J_{l,C}^{t+1}, J_{l,D}^{t+1}|J_{l,C}^t, J_{l,D}^t, a_l^t, a_{l,J}^t) = \sum_{n_{C_1}+n_{C_2}=J_{l,C}^{t+1}} \sum_{n_{D_1}+n_{D_2}=J_{l,D}^{t+1}} \big[p(n_{C_1}, n_{D_1}|J_{l,C}^t, J_{l,D}^t, a_l^t, a_{l,J}^t) \times p(n_{C_2}, n_{D_2}|J_{l,C}^t, J_{l,D}^t, a_l^t, a_{l,J}^t)\big]. \quad (10)$$

Substituting (3)(4) and (10) into (6), we can get the state transition probability.

After the secondary users and the attackers choose their actions, the secondary users will transmit control and data messages in the selected channels, and attackers will jam their selected channels. In order to coordinate the spectrum access and simplify operation, we assume that the same control messages are transmitted in all the control channels, and one correct copy of control information at time t is sufficient for coordinating the spectrum management in the next time slot $t+1$. The gain of a channel can only be achieved when it is used for transmitting data messages and at least one control channel is not jammed by the attackers. Considering that it costs energy for the secondary users to transmit control and data messages and they may be energy-constrained, the objective of the secondary users is to achieve the highest gain with a limited energy. Therefore, the stage payoff of the secondary users can be defined as the expected gain per active channel. Another explanation of the stage payoff is that the secondary users want to maximize the spectrum-efficient gain.

Based on these assumptions, the stage payoff can be expressed by

$$r(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) = T(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) \times (1 - p^{block}(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)), \quad (11)$$

where $T(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ denotes the expected spectrum-efficient gain when not all control channels get jammed, and $p^{block}(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ denotes the probability that all control channels in all L bands are jammed.

As explained in Section III-A, we assume that the attackers uniformly select a_{l,J_1}^t channels from the previous $N_{l,1}^t$ unattacked channels to jam, and select a_{l,J_2}^t channels from the previous $N_{l,2}^t$ attacked channels to jam. Then, the probability that a channel will not be jammed at time t can be represented by $(1 - \frac{a_{l,J_1}^t}{N_{l,1}^t})$ and $(1 - \frac{a_{l,J_2}^t}{N_{l,2}^t})$, respectively. Given the gain of the channels g_l^t and assuming that different data is transmitted in different channels, we have the expected gain of using band l as $[a_{l,D_1}^t(1 - \frac{a_{l,J_1}^t}{N_{l,1}^t}) + a_{l,D_2}^t(1 - \frac{a_{l,J_2}^t}{N_{l,2}^t})]g_l^t$. Then, we can express $T(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ as (12), where the denominator denotes the total number of control and data channels. Thus, (12) reflects the spectrum-efficient gain.

Only when all the control channels in each licensed band l are jammed can the secondary network be blocked. Therefore, the blocking probability $p^{block}(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ can be expressed as

$$\begin{aligned} &p^{block}(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) \\ &= \prod_{l=1}^{L} \frac{\binom{a_{l,C_1}^t}{a_{l,C_1}^t}\binom{N_{l,1}^t - a_{l,C_1}^t}{a_{l,J_1}^t - a_{l,C_1}^t}}{\binom{N_{l,1}^t}{a_{l,J_1}^t}} \times \frac{\binom{a_{l,C_2}^t}{a_{l,C_2}^t}\binom{N_{l,2}^t - a_{l,C_2}^t}{a_{l,J_2}^t - a_{l,C_2}^t}}{\binom{N_{l,2}^t}{a_{l,J_2}^t}} \\ &= \prod_{l=1}^{L} \frac{\binom{N_{l,1}^t - a_{l,C_1}^t}{a_{l,J_1}^t - a_{l,C_1}^t}}{\binom{N_{l,1}^t}{a_{l,J_1}^t}} \times \frac{\binom{N_{l,2}^t - a_{l,C_2}^t}{a_{l,J_2}^t - a_{l,C_2}^t}}{\binom{N_{l,2}^t}{a_{l,J_2}^t}}, \end{aligned} \quad (13)$$

where the first (or second) term in the product represents the probability that all the control channels uniformly selected from the previously un-jammed (or jammed) channels in the l-th band get jammed at time t.

Substituting (12) and (13) back into (11), we can obtain the stage payoff for the secondary users, and the attackers' payoff is the negative of (11).

IV. Solving Optimal Policies of the Stochastic Game

Based on the stochastic anti-jamming game formulation in the previous section, in this section, we discuss how to come up with the optimal strategy, i.e., the optimal defending policy of the secondary users.

In general, the secondary users have a long sequence of data to transmit, and the energy of the attackers can afford to jam the secondary network for a long time given that the number of the jammed channels at each stage will not exceed $\bar{N}$. Thus, we can assume that the anti-jamming game is played for an infinite number of stages. Moreover, the secondary users treat the payoff in different stages differently, e.g., delayed messages usually have less value in delay-sensitive applications, and a recent payoff should weigh more than a payoff that will be received in the faraway future. Then, the secondary users' objective is to derive an optimal policy that maximizes the expected sum of discounted payoffs

$$\max \quad E\{\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)\}, \quad (14)$$

where γ is the discount factor of the secondary user network. A *policy* in the stochastic game refers to a probability distribution over the action set at any state. Then, the policy of the secondary network is denoted by $\pi : \mathcal{S} \rightarrow PD(\mathcal{A})$, and the policy of the attackers can be denoted by $\pi_J : \mathcal{S} \rightarrow PD(\mathcal{A}_J)$, where $\mathbf{s}^t \in \mathcal{S}$, $\mathbf{a}^t \in \mathcal{A}$, and $\mathbf{a}_J^t \in \mathcal{A}_J$. Given the current state $\mathbf{s}^t$, if the defending policy π^t (or jamming policy π_J^t) at time t is independent of the states and actions in all previous time slots, the policy π (or π_J) is said to be *Markov*. If the policy is further independent of time, i.e., $\pi^t = \pi^{t'}$, given that $\mathbf{s}^t = \mathbf{s}^{t'}$ the policy is said to be *stationary*.

It is known [33] that every stochastic game has a non-empty set of optimal policies, and at least one of them is stationary. Since the game between the secondary network and the attackers is a zero-sum game, the equilibrium of each stage game is the unique minimax equilibrium, and thus the optimal policy will also be unique for each player. In order to solve

$$T(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) = \frac{\sum_{l=1}^{L} \left[a_{l,D_1}^t (1 - \frac{a_{l,J_1}^t}{N_{l,1}^t}) + a_{l,D_2}^t (1 - \frac{a_{l,J_2}^t}{N_{l,2}^t}) \right] g_l^t}{\sum_{l=1}^{L} (a_{l,C_1}^t + a_{l,D_1}^t + a_{l,C_2}^t + a_{l,D_2}^t)}, \tag{12}$$

the optimal policy, we can use the minimax-Q learning method [33]. Here, the Q-function $Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ at stage t is defined as the expected discounted payoff when the secondary users take action $\mathbf{a}^t$, the attackers take action $\mathbf{a}_J^t$, and both of them follow their stationary policies thereafter. Since the Q-function is essentially an estimate of the expected total discounted payoff which evolves over time, in order to maximize the worst-case performance, at each stage the secondary users should treat the $Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ as the *payoff* of a *matrix game*, where $\mathbf{a}^t \in \mathcal{A}$ and $\mathbf{a}_J^t \in \mathcal{A}_J$. Given the payoff $Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ of the game, the secondary users can find the minimax equilibrium and update the Q-value with the value of the game [33]. Therefore, the value of a state in the anti-jamming game becomes

$$V(\mathbf{s}^t) = \max_{\pi(\mathbf{a}^t)} \min_{\pi_J(\mathbf{a}_J^t)} \sum_{\mathbf{a}^t \in \mathcal{A}} Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) \pi(\mathbf{a}^t), \tag{15}$$

where $Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ is updated by

$$Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) = r(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) + \gamma \sum_{\mathbf{s}^{t+1}} p(\mathbf{s}^{t+1} \mid \mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) V(\mathbf{s}^{t+1}). \tag{16}$$

In order to avoid the complexity of estimating the state transition probability, we can modify the value iteration and the Q-function is updated according to [34] [35]

$$\begin{aligned} Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) =& (1 - \alpha^t) Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) \\ &+ \alpha^t \left[r(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t) + \gamma V(\mathbf{s}^{t+1}) \right], \end{aligned} \tag{17}$$

where α^t denotes the learning rate decaying over time by $\alpha^{t+1} = \mu \alpha^t$, with $0 < \mu < 1$, and $V(\mathbf{s}^{t+1})$ is obtained by (15). In the modified update in (17), the current value of a state $V(\mathbf{s}^{t+1})$ is used as an approximate of the true expected discounted future payoff, which will be improved during the value iteration; and the estimate of $Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ is updated by mixing the previous Q-value with a correction from the new estimate at a learning rate α^t that decays slowly over time. It is shown that [34] the minimax-Q learning approach converges to the true Q and V values and hence the optimal policy, as long as each action is tried in every state for infinitely many times.

Then, the minimax-Q learning for the secondary users to obtain the optimal policy is summarized in Table I. Since no secondary user (or attacker) will transmit in (or jam) a licensed band when the primary user is active, when the primary users' status are different in various states, the corresponding action spaces of the players at these states are also different. Thus, the action space depends on the state. At the beginning of each stage t, the secondary users check whether they have observed state $\mathbf{s}^t$ before: if not, they will add $\mathbf{s}^t$ to the observation history about every state $\mathbf{s}_{hist}$, and initialize the variables used in the learning algorithm, Q, V, and policy $\pi(\mathbf{s}^t, \mathbf{a})$. If $\mathbf{s}^t$ already exists in the history $\mathbf{s}_{hist}$, the secondary users just call the corresponding action sets and function values. Then, the secondary users will choose an action $\mathbf{a}^t$: with a certain probability p_{exp}, they choose to explore the entire action space $\mathcal{A}(\mathbf{s}^t)$ and return an action uniformly. With probability $1 - p_{exp}$, they choose to take action $\mathbf{a}^t$ that is drawn according to the current $\pi(\mathbf{s}^t)$. After the attackers take action $\mathbf{a}_J^t$, the secondary users receive the reward, and the game transits to the next state $\mathbf{s}^{t+1}$. The secondary users update the Q and V function values, update policy $\pi(\mathbf{s}^t)$ at state $\mathbf{s}^t$, and decay the learning rate. The value iteration will continue until $\pi(\mathbf{s}^t)$ approaches the optimal policy, and we will demonstrate the convergence of the minimax-Q learning in the simulation results.

TABLE I
MINIMAX-Q LEARNING FOR THE ANTI-JAMMING STOCHASTIC GAME

1. At state $\mathbf{s}^t$, $t = 0, 1, \cdots$

- if state $\mathbf{s}^t$ has not been observed previously, add $\mathbf{s}^t$ to $\mathbf{s}_{hist}$,
  - generate action set $\mathcal{A}(\mathbf{s}^t)$, and $\mathcal{A}_J(\mathbf{s}^t)$ of the attackers;
  - initialize $Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J) \leftarrow 1$, for all $\mathbf{a} \in \mathcal{A}(\mathbf{s}^t)$, $\mathbf{a}_J \in \mathcal{A}_J(\mathbf{s}^t)$;
  - initialize $V(\mathbf{s}^t) \leftarrow 1$;
  - initialize $\pi(\mathbf{s}^t, \mathbf{a}) \leftarrow 1/|\mathcal{A}(\mathbf{s}^t)|$, for all $\mathbf{a} \in \mathcal{A}(\mathbf{s}^t)$;
- otherwise, use previously generated $\mathcal{A}(\mathbf{s}^t)$, $\mathcal{A}_J(\mathbf{s}^t)$, $Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)$, $V(\mathbf{s}^t)$, and $\pi(\mathbf{s}^t)$;

2. Choose an action $\mathbf{a}^t$ at time t:

- with probability p_{exp}, return an action uniformly at random;
- otherwise, return action $\mathbf{a}^t$ with probability $\pi(\mathbf{s}^t, \mathbf{a})$ under current state $\mathbf{s}^t$.

3. Learn:

Assume the attackers take action $\mathbf{a}_J^t$, after receiving reward $r(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ for moving from state $\mathbf{s}^t$ to $\mathbf{s}^{t+1}$ by taking action $\mathbf{a}^t$

- Update Q-function $Q(\mathbf{s}^t, \mathbf{a}^t, \mathbf{a}_J^t)$ according to (17);
- Update the optimal strategy $\pi^*(\mathbf{s}^t, \mathbf{a})$ by $\pi^*(\mathbf{s}^t) \leftarrow \arg\max_{\pi(\mathbf{s}^t)} \min_{\pi_J(\mathbf{s}^t)} \sum_{\mathbf{a}} \pi(\mathbf{s}^t, \mathbf{a}) Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)$;
- Update $V(\mathbf{s}^t) \leftarrow \min_{\pi_J(\mathbf{s}^t)} \sum_{\mathbf{a}} \pi^*(\mathbf{s}^t, \mathbf{a}) Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)$;
- Update $\alpha^{t+1} \leftarrow \alpha^t * \mu$;
- Go to step 1 until converge.

Note that in order to obtain the value of a state $V(\mathbf{s}^t)$, the secondary users need to solve the equilibrium of a matrix game, where the payoff is $Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)$, for all $\mathbf{a} \in \mathcal{A}(\mathbf{s}^t)$, and $\mathbf{a}_J \in \mathcal{A}_J(\mathbf{s}^t)$. Assume the attackers form the row player, whose strategy is denoted by vector $\pi_J(\mathbf{s}^t)$, and the secondary users form the column player, whose strategy is denoted by vector $\pi(\mathbf{s}^t)$. Then, the value of the game can be expressed by

$$\max_{\pi(\mathbf{s}^t)} \min_{\pi_J(\mathbf{s}^t)} \pi_J(\mathbf{s}^t)^T Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J) \pi(\mathbf{s}^t), \tag{18}$$

which cannot be solved directly. If we assume the secondary users's strategy $\pi(\mathbf{s}^t)$ is fixed, then the problem in (18) becomes

$$\min_{\pi_J(\mathbf{s}^t)} \pi_J(\mathbf{s}^t)^T Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J) \pi(\mathbf{s}^t). \tag{19}$$

Since $Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)\pi(\mathbf{s}^t)$ is just a vector, and $\pi_J(\mathbf{s}^t)$ is a probability distribution, the solution of (19) is equivalent to $\min_i [Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)\pi(\mathbf{s}^t)]_i$, i.e., finding the minimal element of $Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)\pi(\mathbf{s}^t)$. Then, the problem in (18) is simplified as

$$\max_{\pi(\mathbf{s}^t)} \min_i [Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J) \pi(\mathbf{s}^t)]_i. \tag{20}$$

Define $z = \min_i [Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)\pi(\mathbf{s}^t)]_i$, we have $[Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)\pi(\mathbf{s}^t)]_i \geq \min_i [Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)\pi(\mathbf{s}^t)]_i = z$.

Therefore, the original problem (18) becomes

$$\begin{aligned} \max_{\pi(\mathbf{s}^t)} \quad & z \\ \text{s.t.} \quad & [Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J)\pi(\mathbf{s}^t)]_i \geq z, \\ & \pi(\mathbf{s}^t) \geq \mathbf{0}, \\ & \mathbf{1}^T \pi(\mathbf{s}^t) = 1, \end{aligned} \tag{21}$$

where $\pi(\mathbf{s}^t) \geq \mathbf{0}$ means that each probability element in $\pi(\mathbf{s}^t)$ must be non-negative. By treating the objective z also as a variable, (21) can be turned to the following

$$\begin{aligned} \max_{\pi'} \quad & \mathbf{0}_{aug}^T \pi' \\ \text{s.t.} \quad & Q'\pi' \leq \mathbf{0}, \\ & \pi(\mathbf{s}^t) \geq \mathbf{0}, \\ & \mathbf{1}_{aug}^T \pi' = 1, \end{aligned} \tag{22}$$

where $\pi' = [\begin{smallmatrix} \pi(\mathbf{s}^t) \\ z \end{smallmatrix}]$, $Q' = ([\mathbf{O} \quad \mathbf{1}] - [Q(\mathbf{s}^t, \mathbf{a}, \mathbf{a}_J) \quad \mathbf{0}])$, $\mathbf{1}_{aug}^T = [\mathbf{1}^T \quad 0]$, and $\mathbf{0}_{aug}^T = [\mathbf{0}^T \quad 1]$. Problem (22) is a linear program, so the secondary users can easily obtain the value of the game z from the optimizer π'.

V. Simulation Results

In this section, we conduct simulations to evaluate the secondary user network's performance under the jamming attack. We first demonstrate the convergence of the minimax-Q learning algorithm, and analyze the strategy of the secondary users and attackers for several typical states. Then, we compare the achievable performance when the secondary users adopt different strategies. For illustrative purpose, we focus on examples with only one or two licensed bands to provide more insight; however, similar policies can be observed when there are more licensed bands available.

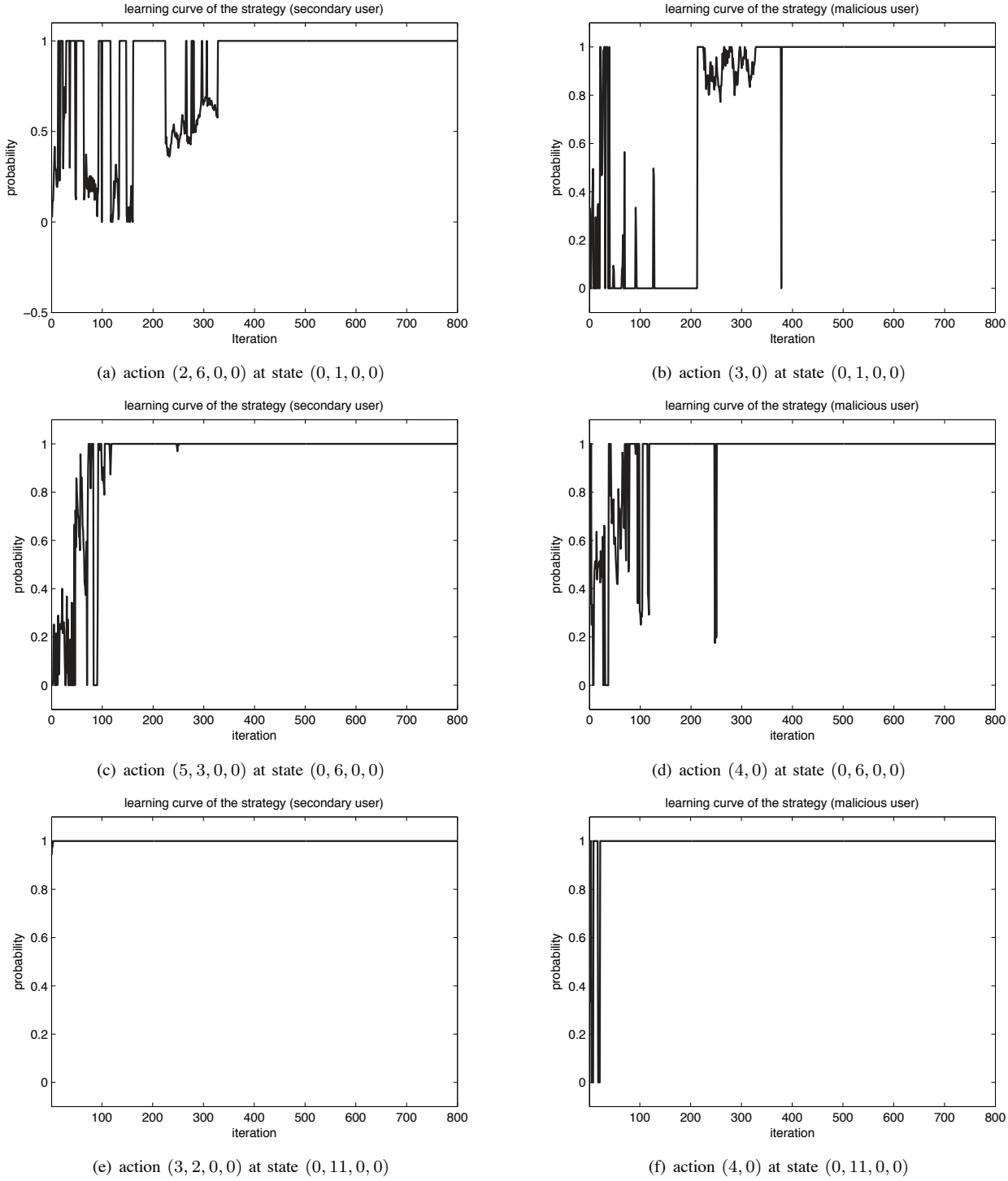
A. Convergence and Strategy Analysis

1) Anti-Jamming Defense in One Licensed Band: We first study the case when there is only one licensed band available to the secondary users, i.e., $L = 1$. There are eight channels in the licensed band, among which the attackers can at most choose four channels to jam at each time. The gain of utilizing each channel in the licensed band g_l^t can take any value from $\{1, 6, 11\}$, and the transition probability of the gain from any q_j to q_i is $p_{g_l}^{j\to 1} = p_{g_l}^{j\to 2} = 0.4$, $p_{g_l}^{j\to 3} = 0.2$, for $j = 1, 2, 3$, as well as for $j = 0$ when the primary user becomes inactive. The transition probabilities about the primary user's access are given by $p_l^{1\to 1} = 0.5$ and $p_l^{0\to 1} = 0.5$. The length of a time slot is 2 ms.

We first study the strategy of the secondary users and the attackers at those states when the primary user is inactive and no channels are observed to be successfully jammed in the previous stage. Recall that the state of the stochastic anti-jamming game with $L = 1$ is denoted by $\mathbf{s}^t = \{P_1^t, g_1^t, J_{1,C}^t, J_{1,D}^t\}$, where $J_{1,C}^t$ and $J_{1,D}^t$ represent the number of jammed control and data channels observed from the previous stage, then three such states are $(0, 1, 0, 0)$, $(0, 6, 0, 0)$, and $(0, 11, 0, 0)$. We show the learning curve of the secondary users' strategy in these states in the left column of Figure 2, and the learning curve of the attackers' strategy in the right column. We see from Figure 2 that using the minimax-Q learning, the strategies of the secondary users and the attackers both converge within less than 400 time slots (0.8 s), and the optimal strategy for each player is a pure strategy. Recall that the action of the secondary users on the l-th band is denoted by $(a_{l,C_1}^t, a_{l,D_1}^t, a_{l,C_2}^t, a_{l,D_2}^t)$, and the action of the attackers is $(a_{l,J_1}^t, a_{l,J_2}^t)$. Then, in Figures 2(a) and 2(b) for state $(0, 1, 0, 0)$, we see that the optimal strategy of the secondary users finally converges to $(2, 6, 0, 0)$, meaning that the secondary users uniformly choose 2 channels as control channels, and 6 channels as data channels; and the attackers' optimal strategy converges to $(3, 0)$, meaning uniformly choose 3 channels to jam. This is because the gain of each channel in this state is only 1, and the secondary users choose to reserve a lot channels for transmitting data messages and a few channels for control messages, in hope of obtaining a higher gain while at a higher risk of having all the control channels jammed. When the gain increases to 6 per channel, as shown in Figures 2(c) and 2(d), the secondary users become more risk-averse by reserving 5 control channels and 3 data channels, and the attackers become more aggressive by attacking the maximal number of channels they can. This is because the gain of each channel is higher, and the secondary users want to ensure a certain gain by securing at lease one control channel from being jammed. When the gain further increases to 11 (Figures 2(e) and 2(f)), the secondary users become even more conservative by only having 2 data channels and 3 control channels. This is because the objective of the secondary users is defined as the spectrum-efficient gain as in (12), and leaving more channels as idle may probably increase the payoff.

Next, we observe how the players' strategy will change when some of the state variables are different, for instance, some control or data channels are jammed by the attackers in the previous stage. We only choose two states for illustration, state $(0, 6, 2, 0)$ and state $(0, 6, 0, 2)$, to compare with the strategy at state $(0, 6, 0, 0)$.

In Figure 3, we demonstrate the learning curve of the secondary users and the attackers at state $(0, 6, 2, 0)$, where 2 control channels are jammed in the previous stage. We see that both players' strategies converge within 50 time slots (0.1 s), and the optimal policies of both players at this state are mixed strategies. Since in the previous stage, the attackers successfully jam 2 control channels, it is highly likely that most of the remaining un-jammed channels are data channels. Thus, the attackers tend to jam the previously un-jammed channels with a relatively high probability, as shown by actions $(1, 0)$, $(2, 0)$, $(3, 0)$, $(2, 1)$ in Figure 3(b), the total probability of which is very high at the beginning. Then, the secondary users tend to reserve most of the previously jammed channels as data channels, as shown by those actions where $a_{l,D_2}^t \geq 1$ with a total probability greater than 0.9; and reserve only a few of the previously un-jammed channels as data channels, as shown by actions where $a_{l,D_1} \leq 3$ with a total probability greater than 0.8. Moreover, since the attackers will attack less than 3 channels from the previously un-jammed channels, the secondary users only reserve at most 3 control channels there to ensure reliable communications. The attackers generally jam less than 4 channels. If they choose to jam 4 channels, the secondary users facing the high chance of being attacked



(a) action $(2, 6, 0, 0)$ at state $(0, 1, 0, 0)$

(b) action $(3, 0)$ at state $(0, 1, 0, 0)$

(c) action $(5, 3, 0, 0)$ at state $(0, 6, 0, 0)$

(d) action $(4, 0)$ at state $(0, 6, 0, 0)$

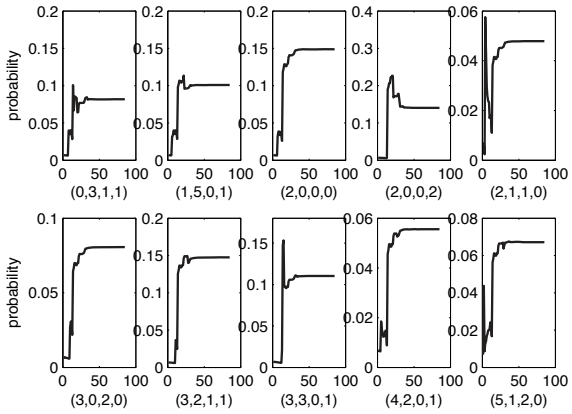
(e) action $(3, 2, 0, 0)$ at state $(0, 11, 0, 0)$

(f) action $(4, 0)$ at state $(0, 11, 0, 0)$

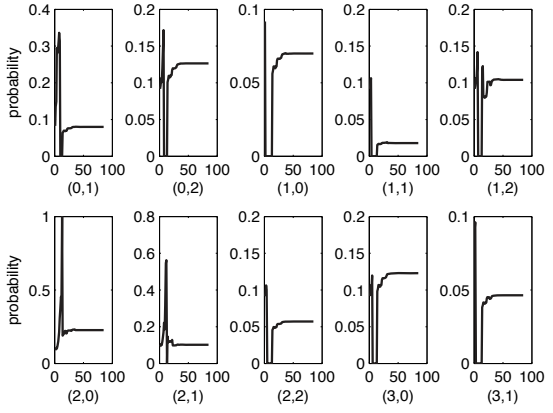
Fig. 2. Learning curve of the secondary users (left column) and the attackers (right column).

will leave more channels as idle. This may in return increase the secondary users' expected payoff, and thus the attackers at most jam 3 channels.

Both players' strategies at state $(0, 6, 0, 2)$ are shown in Figure 4. Since 2 data channels are successfully jammed in the previous stage, the secondary users tend to reserve less than 1 channel that are previously jammed as data channels to avoid "second jammed", as shown by actions $(5, 0, 1, 1)$ and $(5, 1, 1, 0)$ with a total probability greater than 0.7. Considering that the attackers will probably attack the previously un-jammed channels, the secondary users reserve most un-jammed channels as control channels to ensure reliability, again as shown by actions $(5, 0, 1, 1)$ and $(5, 1, 1, 0)$ where 5 un-jammed channels are selected as control channels. In response to the secondary users' strategy, the attackers will keep attacking the previously jammed channels, as shown by actions $(0, 2)$, $(1, 2)$, $(2, 2)$ with a total probability greater than 0.94, where $a_{l,J_2} = 2$. Comparing Figure 4 and Figure 3, we find that when the attackers successfully jam some data channels, more information about the secondary users'

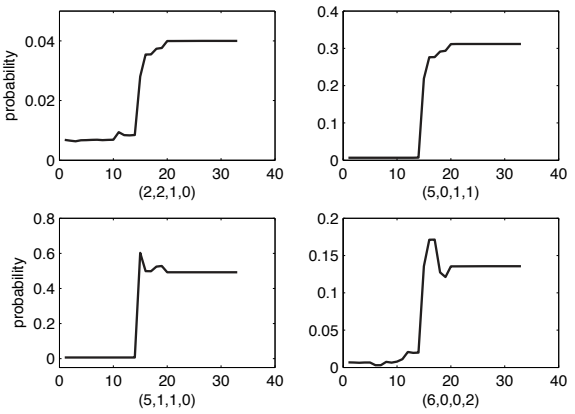


(a) Learning curve of the secondary users at state $(0, 6, 2, 0)$

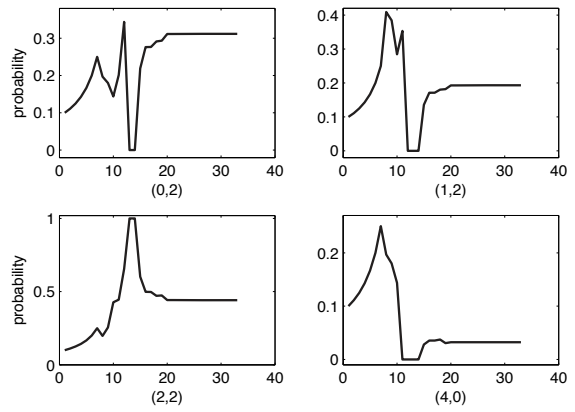


(b) Learning curve of the attackers at state $(0, 6, 2, 0)$

Fig. 3. Learning curve of the secondary users and the attackers at state $(0, 6, 2, 0)$.



(a) Learning curve of the secondary users at state $(0, 6, 0, 2)$



(b) Learning curve of the attackers at state $(0, 6, 0, 2)$

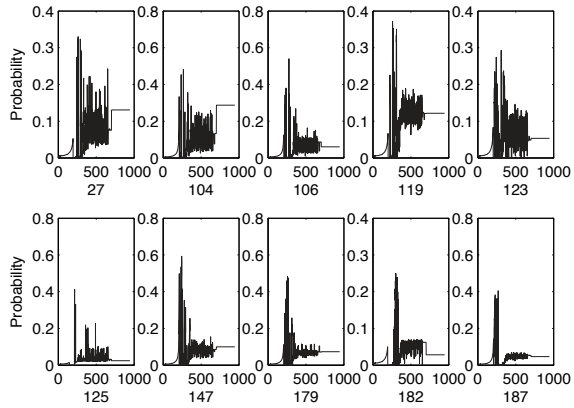
Fig. 4. Learning curve of the secondary users and the attackers at state $(0, 6, 0, 2)$.

strategy (on locating the data channels) is revealed, the damage of the jamming attack will be more severe, and the secondary users have to reserve more channels for control use, which leads to a reduced payoff.

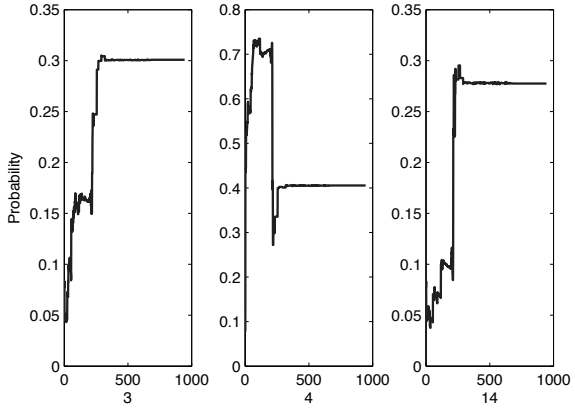
2) Anti-Jamming Defense in Two Licensed Bands: We now discuss the strategy of the secondary users and attackers when there are two licensed bands available, i.e. $L = 2$. There are four channels within each band, and the gain of the channels in each band still takes value from $\{1, 6, 11\}$, with the same transition probability as that in the one-band case. The transition probability about the primary user's access on the first band is $p_1^{1\to1} = p_1^{0\to1} = 0.5$, while the transition probability about the second band is $p_2^{1\to1} = p_2^{0\to1} = 0.2$, meaning that the the probability of the second band being available is higher than that of the first band. The attackers can jam at most four channels at each time.

To compare with the one-band case, we first study the strategy of both players at state $((0, 6, 0, 0), (0, 6, 0, 0))$, where both bands are available, with gain $g_1^t = g_2^t = 6$, and no control or data channels have been jammed in the previous stage. We show the learning curve of both players in Figure 5, where the number below each plot denotes the index of the action shown in that plot. We see that the secondary users' strategy converges to the optimal policy within 800 time slots (1.6 s), while the attackers' strategy converges within 400 time slots (0.8 s). Under the optimal policy, the secondary users mostly take action $((1, 3, 0, 0), (1, 3, 0, 0))$ indexed as 104, action $((0, 2, 0, 0), (2, 2, 0, 0))$ indexed as 27, action $((2, 0, 0, 0), (1, 3, 0, 0))$ indexed as 119, and action $((2, 2, 0, 0), (1, 1, 0, 0))$ indexed as 147; the attackers mostly take action $((0, 0), (3, 0))$ indexed as 3, action $((0, 0), (4, 0))$ indexed as 4, and action $((4, 0), (0, 0))$ indexed as 14. Since the availability of the second band is higher, the attackers tend to jam the channels in the second band (with a total probability 0.7 of action 3 and 4). But there is still a chance that they will attack the first band, indicating that the attackers' strategy is random. Compared to the equivalent state $(0, 6, 0, 0)$ in the one-band case, where the secondary users' policy is $(5, 3, 0, 0)$, the secondary users' policy in the two-band case is more aggressive, as seen from the fact that the secondary users assign more data channels and less control channels in total. This is because there are two available bands, the attackers' strategy becomes more random, and thus an aggressive policy can bring a higher gain to the secondary users.

Then, we study the strategy at state $((0, 1, 0, 0),(0, 6, 0, 0))$, where $g_1^t = 1$, and $g_2^t = 6$. The learning curves are shown in Figure 6. Since the second band has higher gain and is also more likely to be available in the next slot, the attackers tend

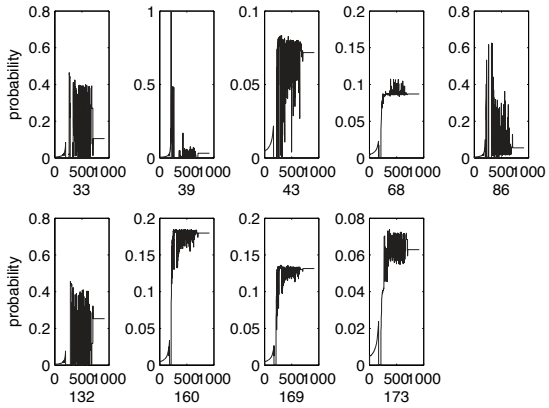


(a) Learning curve of the secondary users

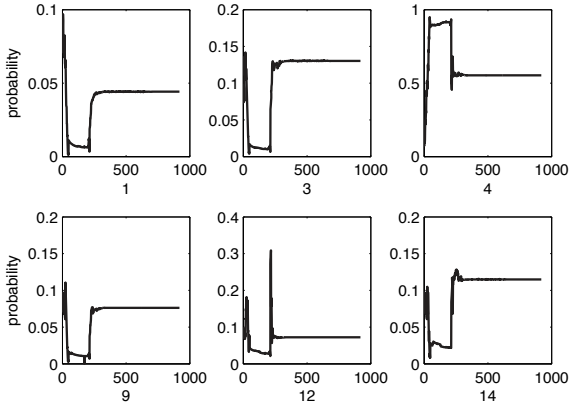


(b) Learning curve of the attackers

Fig. 5. Learning curve at state $((0,6,0,0),(0,6,0,0))$ when $L=2$.



(a) Learning curve of the secondary users



(b) Learning curve of the attackers

Fig. 6. Learning curve at state $((0,1,0,0),(0,6,0,0))$ when $L=2$.

to jam the second band, as seen from the probability of action $((0,0),(3,0))$ indexed as 3 and action $((0,0),(4,0))$ indexed as 4. In response to the attackers' strategy, the secondary users tend to reserve more control channels in the first band since it is less likely to be attacked, and more data channels in the second band since it has a higher gain for each channel, as seen from the probability of action $((2,1,0,0),(1,1,0,0))$ indexed as 132 and action $((3,0,0,0),(0,4,0,0))$ indexed as 160.

B. Comparison of Different Strategies

We also compare the performance of the secondary users when they adopt the stationary policy obtained from the minimax-Q learning with other policies to evaluate the proposed stochastic anti-jamming game and the learning algorithm. We assume the attackers use their optimal stationary policy that is trained against the secondary users who adopt the minimax-Q learning. We then consider the following three scenarios with different strategies for the secondary users.

- The secondary users adopt the stationary policy obtained by the minimax-Q learning (denoted by "proposed").
- The secondary users adopt a stationary policy obtained by myopic learning. By *myopic*, we mean that they care more about the immediate payoff than the future payoffs. In the considered myopic policy, we assume that the secondary users ignore the effect of their current action on the future payoffs, so it is the extreme case where $\gamma=0$ (denoted by "myopic").
- The secondary users adopt a fixed strategy which draws an action uniformly from the action space $\mathcal{A}(\mathbf{s}^t)$ for each $\mathbf{s}^t$ (denoted by "fixed").

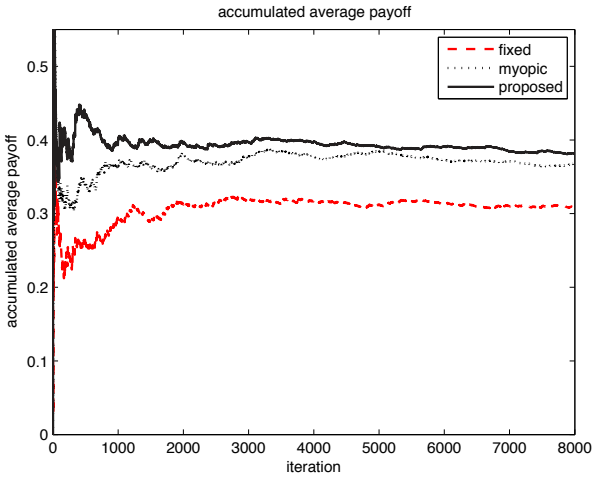


Fig. 7. Average payoff of different strategies.

In Figure 7, we compare the accumulated average payoff at each iteration t', calculated by

$$\bar{r}(t') = \frac{1}{t'} \sum_{t=1}^{t'} r(\mathbf{s}(t), \mathbf{a}(t), \mathbf{a}_J(t)). \quad (23)$$

We see that, since the proposed strategy and the myopic strategy maximize the worst-case performance, while the fixed strategy only uniformly picks any action regardless of the attackers' strategy, the former two strategies have a higher average payoff than the fixed strategy. Moreover, as shown in Figure 8, since the proposed strategy also considers the future payoff when optimizing the strategy at the current stage, it achieves the highest sum of discounted payoff (15% more than that of the myopic strategy and 42% more than that of the fixed strategy). Therefore, when the secondary users face a group of intelligent attackers that can adapt their strategy to the environment dynamics and the opponent's strategy, adopting the minimax-Q learning in the stochastic anti-jamming game modeling achieves the best performance.

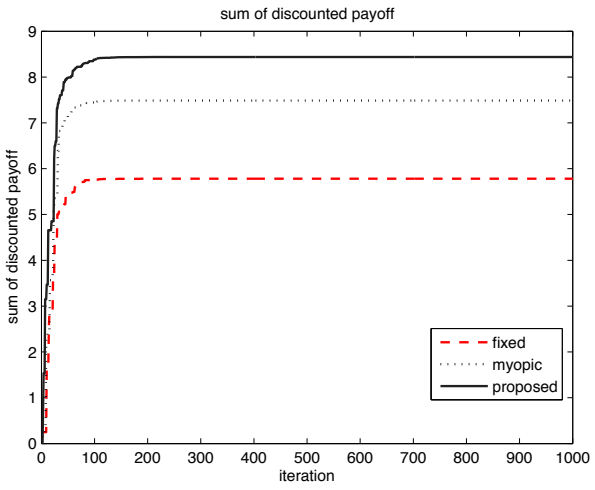


Fig. 8. Sum of discounted payoff of different strategies.

VI. Conclusion

In this paper, we have studied the design of anti-jamming defense mechanism in a cognitive radio network. Considering the spectrum environment is time-varying, and the cognitive attackers are able to use an adaptive strategy, we model the interactions between the secondary users and the attackers as a stochastic zero-sum game. The secondary users adapt their strategy on how to reserve and switch between control and data channels, according to their observation about the spectrum availability, channel quality and the attackers' actions. Simulation results show that the optimal policy obtained from the minimax-Q learning in the stochastic game can achieve much better performance in terms of spectrum-efficient throughput, compared to the myopic learning policy which only maximizes the payoff at each stage without considering the environment dynamics and the attackers' cognitive capability, and a random defense policy. The proposed stochastic game framework can be generalized to model various defense mechanisms in other layers of a cognitive radio network, since it can well model the different dynamics due to the environment as well as the cognitive attackers.

References

[1] J. Mitola, "Cognitive radio: an integrated agent architecture for software defined radio," Ph.D. dissertation, KTH Royal Institute of Technology, Stockholm, Sweden, 2000.

[2] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[3] B. Wang and K. J. Ray Liu, "Advances in cognitive radio networks: a survey," *IEEE J. Sel. Topics Signal Process., vol. 5, no. 1, pp. 5–23, Feb. 2011.*

[4] K. J. R. Liu and B. Wang, *Cognitive Radio Networking and Security: A Game-Theoretic View*, Cambridge University Press, 2010.

[5] M. M. Halldorson, J. Y. Halpern, L. Li, and V. S Mirrokni, "On spectrum sharing games," *Proc. ACM on Principles of distributed computing*, pp. 107–114, 2004.

[6] O. Ileri, D. Samardzija, and N. B. Mandayam, "Demand responsive pricing and competitive spectrum allocation via a spectrum server," *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN'05)*, pp. 194–202, Baltimore, Nov. 2005.

[7] J. Huang, R. Berry, and M. L. Honig, "Auction-based spectrum sharing," *ACM/Springer Mobile Networks and Apps.*, vol. 11, no. 3, pp. 405–418, June 2006.

[8] C. Kloeck, H. Jaekel, and F. K. Jondral, "Dynamic and local combined pricing, allocation and billing system with coginitive radios," *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN'05)*, pp. 73–81, Baltimore, Nov. 2005.

[9] S. Gandhi, C. Buragohain, L. Cao, H. Zheng, and S. Suri, "A general framework for wireless spectrum auctions," *IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN'07)*, pp. 22–33, Dublin, Apr. 2007.

[10] Z. Ji and K. J. R. Liu, "Belief-assisted pricing for dynamic spectrum allocation in wireless networks with selfish users," in *Proc. IEEE Int'l Conference on Sensor, Mesh, and Ad Hoc Communications and Networks (SECON)*, pp. 119–127, Reston, Sep. 2006.

[11] Z. Ji and K. J. R. Liu, "Multi-stage pricing game for collusion-resistant dynamic spectrum allocation," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 182–191, Jan. 2008.

[12] Y. Wu, B. Wang, K. J. R. Liu, and T. C. Clancy, "A Scalable Collusion-Resistant Multi-Winner Cognitive Spectrum Auction Game," *IEEE Trans. Commun.*, vol. 57, no. 12, pp. 3805–3816, Dec. 2009.

[13] B. Wang, Y. Wu, and K. J. R. Liu, "Game theory for cognitive radio networks: an overview," *Computer Networks, vol. 54, no. 14, pp. 2537–2561, Oct. 2010.*

[14] Y. Xing, R. Chandramouli, S. Mangold, and S. Shankar, "Dynamic spectrum access in open spectrum wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 626-637, Mar. 2006.

[15] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589-600, Apr. 2007.

[16] B. Wang, Z. Ji, K. J. R. Liu, and C. Clancy, "Primary-prioritized Markov approach for efficient and fair dynamic spectrum allocation," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1854–1865, Apr. 2009.

[17] R. Chen, J. Park, and J. H. Reed, "Defense against primary user emulation attacks in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 25–37, Jan. 2008.

[18] R. Chen, J. Park, and K. Bian, "Robust Distributed Spectrum Sensing in Cognitive Radio Networks," *IEEE 27th Conference on Computer Communications (INFOCOM)*, pp. 31-35, Phoenix, AZ, Apr. 2008.

[19] T. X. Brown and A. Sethi, "Potential cognitive radio denial-of-service vulnerabilities and protection countermeasures: a multi-dimensional analysis and assessment," *Mobile Networks and Applications*, vol. 13, no. 5, pp. 516–532, Oct. 2008.

[20] T. C. Clancy and N. Goergen, "Security in cognitive radio networks: threats and mitigation," *3rd International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom)*, pp. 1-8, Singapore, May 2008.

[21] A. Wood and J. Stankovic, "Denial of service in sensor networks," *IEEE Computer*, 35(10):54-62, Oct. 2002.

[22] G. Noubir, "On connectivity in ad hoc network under jamming using directional antennas and mobility," *International Conference on Wired /Wireless Internet Communications*, pp. 186–200, 2004.

[23] R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, "Theory of spread

spectrum communications-a tutorial," *IEEE Trans. Commun.*, vol. 20, no. 5, pp. 855-884, May 1982.
[24] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688-1692, Nov. 1999.
[25] V. Navda, A. Bohra, S. Ganguly, and D. Rubenstein, "Using channel hopping to increase 802.11 resilience to jamming attacks," *IEEE 26th Conference on Computer Communications (INFOCOM)*, pp. 2526-2530, Anchorage, AK, Apr. 2007.
[26] R. Gummadi, D. Wetherall, B. Greenstein, and S. Seshan, "Understanding and mitigating the impact of RF interference on 802.11 networks," *ACM SIGCOMM Computer Communication Review*, vol. 37, no. 4, pp. 385-396, 2007.
[27] A. D. Wood, J. A. Stankovic, and G. Zhou, "DEEJAM: defeating energy-efficient jamming in IEEE 802.15.4-based wireless networks," *Proc. 4th IEEE Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, pp. 60-69, San Diego, CA, June 2007.
[28] S. Khattab, D. Mosse, and R. Melhem, "Modeling of the channel-hopping anti-jamming defense in multi-radio wireless networks," *Proc. 5th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services (MobiQuitous)*, pp. 1-10, Dublin, Ireland, July 2008.
[29] W. Xu, T. Wood, W. Trappe, and Y. Zhang, "Channel surfing and spatial retreats: defenses against wireless denial of service," in *Proc. 3rd ACM Workshop on Wireless Security (WiSe)*, pp. 80-89, Philadelphia, PA, Oct. 2004.
[30] S. Geirhofer, L. Tong, and B. M. Sadler, "Cognitive medium access: constraining interference based on experimental models," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 95-105, Jan. 2008.
[31] L.S. Shapley, "Stochastic games," *Proc. Nat. Acad. Sci. USA*, vol. 39, no. 10, pp. 1095–1100, 1953.
[32] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*, Springer-Verlag, 1997.
[33] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," *Proc. 11th International Conference on Machine Learning*, pp. 157–163, 1994.
[34] M. L. Littman and C. Szepesvari, "A generalized reinforcement-learning model: Convergence and applications," *Proc. 13th International Conference on Machine Learning*, pp. 310–318, 1996.
[35] J. Hu and M. P. Wellman, "Multiagent reinforcement learning: Theoretical framework and an algorithm," *Proc. 15th International Conference on Machine Learning*, pp. 242-250, 1998.
[36] A. Neyman and S. Sorin, *Stochastic Games and Applications*, Kluwer Academic Press, 2003.
[37] J. F. Mertens and A. Neyman, "Stochastic Games," *International Journal of Game Theory*, vol. 10, pp. 53-66, 1981.
[38] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice Hall, 1996.
[39] A. Chan, X. Liu, G. Noubir, and B. Thapa, "Control channel jamming: resilience and identification of traitors," *Proc. ISIT*, 2007.

Beibei Wang (S'07) received the B.S. degree in electrical engineering (with the highest honor) from the University of Science and Technology of China, Hefei, in 2004, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park in 2009. From 2009 to 2010, she was a research associate at the University of Maryland. Currently, she is a senior engineer with Corporate Research and Development, Qualcomm Incorporated.

Her research interests include wireless communications and networking, signal processing, and game theory with current focus on cognitive radios, dynamic spectrum allocation and management, network security, and multimedia communications. Dr. Wang was the recipient of the Graduate School Fellowship, the Future Faculty Fellowship, and the Dean's Doctoral Research Award from the University of Maryland, College Park. She is a coauthor of *Cognitive Radio Networking and Security: A Game-Theoretic View*, Cambridge University Press, 2010.

Yongle Wu (S'08) received the B.S. (with highest honor) and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from University of Maryland, College Park in 2010. Currently, he is a senior engineer with Qualcomm Incorporated.

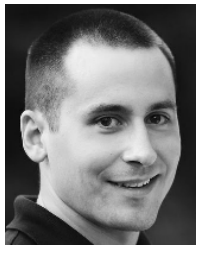
His current research interests are in the areas of wireless communications and networks, including cognitive radio techniques, dynamic spectrum access, and network security. Mr. Wu received the Graduate School Fellowship from the University of Maryland in 2006, the Future Faculty Fellowship from A. James Clark School of Engineering, University of Maryland in 2009, and the Litton Industries Fellowship from A. James Clark School of Engineering, University of Maryland in 2010.

K. J. Ray Liu (F'03) is named a Distinguished Scholar-Teacher of University of Maryland, College Park, in 2007, where he is Christine Kim Eminent Professor of Information Technology. He is Associate Chair of Graduate Studies and Research of Electrical and Computer Engineering Department and leads the Maryland Signals and Information Group conducting research encompassing broad aspects of wireless communications and networking, information forensics and security, multimedia signal processing, and biomedical engineering.

Dr. Liu is the recipient of numerous honors and awards including IEEE Signal Processing Society Technical Achievement Award and Distinguished Lecturer. He also received various teaching and research recognitions from University of Maryland including university-level Invention of the Year Award; and Poole and Kent Senior Faculty Teaching Award and Outstanding Faculty Research Award, both from A. James Clark School of Engineering. An ISI Highly Cited Author in Computer Science, Dr. Liu is a Fellow of IEEE and AAAS.

Dr. Liu is President-Elect and was Vice President - Publications of IEEE Signal Processing Society. He was the Editor-in-Chief of IEEE Signal Processing Magazine and the founding Editor-in-Chief of EURASIP Journal on Advances in Signal Processing.

His recent books include *Cognitive Radio Networking and Security: A Game-Theoretic View*, Cambridge University Press, 2010; *Behavior Dynamics in Media-Sharing Social Networks*, Cambridge University Press (to appear); *Handbook on Array Processing and Sensor Networks*, IEEE-Wiley, 2009; *Cooperative Communications and Networking*, Cambridge University Press, 2008; *Resource Allocation for Wireless Networks: Basics, Techniques, and Applications*, Cambridge University Press, 2008; *Ultra-Wideband Communication Systems: The Multiband OFDM Approach*, IEEE-Wiley, 2007; *Network-Aware Security for Group Communications*, Springer, 2007; *Multimedia Fingerprinting Forensics for Traitor Tracing*, Hindawi, 2005.



T. Charles Clancy (M'05-SM'10) is a faculty member at Virginia Tech where he is the Associate Director of the Hume Center for National Security and Technology, and leads the university's development efforts in cybersecurity research and education. Prior to joining Virginia Tech, Dr. Clancy was a senior advisor to the US military in Baghdad, Iraq, where he led successful efforts to establish Baghdad's first commercial international fiber-optic connectivity. Prior to Iraq, Dr. Clancy was a senior scientist with the Laboratory for Telecommunications Sciences, a federal research lab at the University of Maryland, where he led programs in RF and signal processing research. He received his MS in Electrical Engineering from the University of Illinois, and PhD in Computer Science from the University of Maryland. His research interests focus around cybersecurity issues related to wireless spectrum for next-generation communication systems.