IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 3, 2018, accepted June 4, 2018, date of publication June 13, 2018, date of current version July 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2847045

Cross-Layer Anti-Jamming Scheme: A Hierarchical Learning Approach

CHEN HAN[1] **AND YINGTAO NIU**[2]
[1]College of Communications Engineering, Army Engineering University of PLA, Nanjing 210000, China
[2]Institute of Nanjing Telecommunication Technology, Nanjing 210007, China

Corresponding author: Yingtao Niu (niuyingtao78@hotmail.com)

This work was supported by the Natural Science Foundation of Jiangsu Province under Grant BK20151450.

ABSTRACT This paper investigates the cross-layer optimization for anti-jamming in the network and MAC layers, in which the jammer can adjust the jamming policies to maximize the jamming effectiveness. The joint problem of routing selection, channel allocation, and power control is formulated as a Stackelberg game. The jammer leads the game by choosing the optimal jamming power and channels. The user follows by selecting the optimal nodes and corresponding channels, and adjusts its transmitting power to meet the communication requirement. Then, based on Q-learning, a cross-layer anti-jamming learning algorithm is proposed to obtain the Stackelberg equilibrium. Finally, simulation results are presented to verify the effectiveness of the proposed algorithm.

INDEX TERMS Anti-jamming, Q-learning, Stackelberg game, cross-layer, routing selection, channel allocation, power control.

I. INTRODUCTION

The smart jammer can adjust the jamming policies to maximize the jamming effectiveness, which severely degrades the performance of wireless communication systems. It is important to achieve effective and reliable communication in the presence of jamming. Game theory can be employed to analyze the interactions between the jammers and users, in which both the jammers and users select their strategies independently and selfishly. In [1], Wang *et al.* considered cognitive jammer and proposed an anti-jamming stochastic game framework to learn the optimal policy for maximizing the spectrum-efficient throughput. In [2], a Bayesian game was formulated to analyze the distributed competitive interactions between a jammer and a secondary user. A two-party zero-sum game was formulated in [3] to demonstrate the strategic decision-making under hostile jamming. However, the above-mentioned game models never considered the hierarchical behaviors among players. In the anti-jamming field, the Stackelberg game is a sophisticated method to deal with the hierarchical interactions among players. In [4], a Stackelberg game approach for anti-jamming was proposed to determine the optimal transmission power.

The problems of routing selection and channel allocation for anti-jamming have been separately investigated in previous studies [5], [6]. In [7] and [8], Zhu *et al.* and Yao *et al.* proposed an anti-jamming game to learn the channel selection policies. In [9], an anti-jamming relay game was formulated to obtain the optimal relay strategy against smart jammer in the vehicular ad-hoc networks. The layered structure of the OSI model leads to longer delays and more signaling overhead when the communication network is threatened by jamming. It is difficult to counter smart jamming with fixed routing protocols and independent channel selection policies. The cross-layer design involving the network layer and MAC layer is necessary to improve the anti-jamming performance. Due to the unevenly distributed spectrum, the channels available to each communication node change over time and the links between nodes change with the channel allocation. Therefore, the problem of routing selection should be considered in unison with that of channel allocation to determine the next node in the communication path and corresponding channels for more effective and reliable communication. In [10], the conjoint design of channel allocation and network routing proved to improve the connection stability and end-to-end throughput. In [11], a cross-layer optimization solution was formulated to achieve cooperation between the MAC layer and network layer by hierarchical selection of routing and channels. In [12], a systematic layered Markov decision process (MDP) framework was proposed to optimize the cross-layer transmission policy.

However, the problem of cross-layer optimization in the jamming environment has not been considered in the above studies. Furthermore, power control is also an important method that can be used for anti-jamming. It has a significant impact on the quality of communication as well as the channel allocation. In [13], a Stackelberg game was formulated to analyze the anti-jamming problem with discrete power strategies. In [14] and [15], Xiao *et al.* proposed an anti-jamming game to learn the power control strategies. The above studies only focused on the power control policies without considering routing selection and channel allocation.

Reinforcement learning methods have been applied to learn the jamming rule, and obtain the optimal anti-jamming policy [16]. Then, wireless communication systems can adapt dynamically to the jamming strategy and achieve reliable communication. In [17], based on Q-learning algorithm, an cross-layer aware resource allocation algorithm was proposed to allow dynamic spectrum access users to effectively locate and exploit unused spectrum opportunities. In [18], Ghaffari proposed a real-time routing algorithm based on Q-Learning to solve the route instability problem caused by the movement of nodes. In the existing literature, Q-Learning algorithms have been widely applied to the field of wireless communications. However, to the best of our knowledge, work on cross-layer anti-jamming optimization using Q-learning has not been reported openly. This observation motivates the work in this paper.

The main contributions of this paper can be summarized as follows

- A hierarchical learning scheme for anti-jamming Stackelberg game is proposed to solve the joint selection problem of routing selection, channel allocation, and power control.
- To obtain the solution of the anti-jamming game, a cross-layer anti-jamming learning algorithm (CALA) is proposed.

Note that some relevant works can be found in [7]–[9], [13], and [15], which independently studied a certain aspect of anti-jamming techniques, including power selection, channel allocation, and relay message. The main difference is that we investigate the joint anti-jamming problem of routing selection, channel allocation, and power control.

The rest of this paper is organized as follows. The system model and problem formulation are established in Section II. The anti-jamming game is formulated in Section III. In Section IV, a cross-layer anti-jamming learning algorithm is proposed. In Section V, the simulation results are shown and the performance of proposed algorithm is analyzed. Section VI draws the conclusions.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

There are N channels whose bandwidths are W_{ch}. A power-limited jammer selects H channels to launch the jamming attacks. The jamming area encompasses some communication nodes. The jamming power set and jamming channel set are denoted as $P_j = [p_{j1}, p_{j2} \dots p_{jL}]$ and $F_j = [f_{j1}, f_{j2} \dots f_{jN}]$, respectively, where L is the number of available jamming power. In this paper, we assume that only one channel is jammed at a time to simplify the analysis. The jammer automatically adjusts its jamming power and jamming channel according to the jamming effect on the network to achieve smart jamming attacks. Every node implements the full-duplex strategy and there is no delay in packet forwarding. The available transmitting power set of the user is expressed as $P_u = [p_{u1}, p_{u2} \dots p_{uL}]$.

The set of channels available to every node is time-varying. Based on the current network topology, the minimal number of hops M from the source node to destination node is obtained via the minimum hop routing algorithm. Every node is capable of spectrum sensing and it is able to select the next routing node independently. The Signal-to-Jamming-plus-Noise Ratio (SJNR) of each channel is denoted as $SJNR = [SJNR_1, SJNR_2, \dots SJNR_N]$. The channel capacity is given by $C = W_{th}\log_2(1 + SJNR)$. The entire communication path consists of M transmission links. The maximum transmission rate of each link is expressed as

$$r_m = C_m, \quad 1 \le m \le M, \tag{1}$$

The final data throughput is limited by the transmission link with the minimal channel capacity

$$r = \min(r_1, r_2, \dots r_m, \dots r_M). \tag{2}$$

B. PROBLEM FORMULATION

The selection of the destination node and corresponding channel L_m in the m-th hop depends on the destination node and the channel selected in L_{m-1}. Therefore, the selection of the destination node and corresponding channel is regarded as a Markov decision process [19].

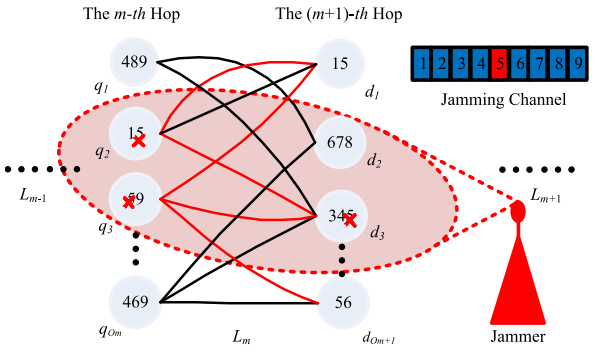

FIGURE 1. The joint selection of nodes and channels in L_m.

Consider the selection of destination nodes and the corresponding channels in L_m as shown in Fig. 1, where the circles represent the communication nodes, and the numbers in each circle are expressed as the set of channels available at each node in the current network topology. The links between adjacent nodes represent the available channels common to both nodes and these are the channels that can be used for communication. The dotted red line is expressed as the

jamming area. The solid black lines are the current available channels, while the solid red lines are the channels which are jammed. In Fig. 1, channel 5 is jammed at present.

S is denoted as the set of possible states

$$\{s_m \in S|S = [(q_1, p_1), \ldots, (q_1, p_N), \ \ldots \\ (q_2, p_1), \ldots (q_{O_m}, p_N)]\}, \quad (3)$$

where s_m is the state in the *m-th* hop, comprised of the node q_m and the corresponding channel p_m. O_m is the number of available nodes in the *m-th* hop and N is the number of channels.

A is denoted as the set of possible actions

$$\{a_m \in A|A = [(d_1, z_1), \ldots, (d_1, z_N), \ \ldots \\ (d_2, z_1), \ldots (d_{O_{m+1}}, z_N)]\}, \quad (4)$$

where a_m is comprised of the selected node d_{m+1} in the next *(m+1)-th* hop and corresponding channel z_{m+1}. O_{m+1} is the number of available nodes in the *(m+1)-th* hop.

For each state-action pair $\{a_m|s_m\}$, an immediate transmission reward is defined as $r_m = C(z_m)$. Based on [20], the objective of the user is to find an optimal strategy π^* which probabilistically maps state s to action a so that the final data throughput $r = \min(r_1, r_2, \ldots r_m, \ldots r_M)$ is maximized,

$$\pi^* = P(a_m|s_m). \quad (5)$$

According to [20, 21], Q-value is given by

$$Q(s_m, a_m) = R(s_m, a_m) \\ + \gamma \sum_{s_{m+1} \in S} P_{s_m, s_{m+1}}(a_m) Q(s_{m+1}, a_{m+1}), \quad (6)$$

where $R(s_m, a_m)$ is the mean value of r_m, and $\gamma \in (0, 1)$ is the discount factor which maps the future reward to the current state. The state transition probability from state s_m to state s_{m+1} with action a_m is expressed as $P_{s_m, s_{m+1}}(a_m)$. Then, the optimal policy π^* can be obtained as follows

$$\pi^* = \arg \max_{\pi} \{Q(s, a)\}, \quad (7)$$

However, it is usually difficult to obtain the values of $R(s_m, a_m)$ and $P_{s_m, s_{m+1}}(a_m)$. In this paper, based on Q-Learning algorithm, a cross-layer anti-jamming learning algorithm is proposed to determine the optimal policy π^* without a priori information on $R(s_m, a_m)$ and $P_{s_m, s_{m+1}}(a_m)$.

The user communicates on the optimal communication path once the optimal nodes and corresponding optimal channels in every hop are determined. However, the data throughput may still fall under the minimum communication requirement due to the action of the jammer. In this situation, the user has to adaptively adjust the transmission power and play a power game with the jammer to obtain the optimal transmission power.

$$r_m = W_{th} \log_2 \left(1 + \frac{p_{u,l}}{n_{0,m} + p_{j,v}}\right), \quad (8)$$

$$p^*_{u,l} = \arg \max_{p_{u,l}} \{\min(r_1, \ldots r_m, \ldots r_M)\}, \quad (9)$$

where $p_{u,l}$, $n_{0,m}$, and $p_{j,v}$ are the transmission power, channel noise, and jamming power in L_m, respectively.

III. ANTI-JAMMING GAME

A. ANTI-JAMMING ROUTING-CHANNEL SELECTION GAME

Routing selection and channel allocation for anti-jamming can be formulated as a Stackelberg game. Mathematically, the anti-jamming routing-channel selection game is denoted as $\wp = \{U_c, J_c, A, F_j, r_{uc}, r_{jc}\}$, where U_c is denoted as the user, J_c is expressed as the jammer, A and F_j represent the strategy space of the user and jammer, r_{uc} and r_{jc} are the utility functions of the user and jammer, respectively. To be specific, the jammer leads the game by choosing its jamming strategy first. The user then follows by detecting the jamming environment and adjusting its communication strategy.

The utility function of the user in routing-channel selection game can be given by

$$r_{uc} = r. \quad (10)$$

The utility function of the jammer is expressed as

$$r_{jc} = r_{max} - r, \quad (11)$$

where r_{max} is the maximal data throughput of the user without jamming attacks.

B. ANTI-JAMMING POWER GAME

If the data throughput is still under the minimum communication requirement, the user should adaptively adjust the power policy. Power control for anti-jamming is formulated as a Stackelberg game. Mathematically, the anti-jamming power game is denoted as $\Im = \{U_p, J_p, P_u, P_j, r_{up}, r_{jp}\}$, where U_p is expressed as the user, J_p is denoted as the jammer, P_u and P_j represent the power strategy space of the user and jammer, r_{up} and r_{jp} are the power utility functions of the user and jammer, respectively. Again, the jammer is the leader, while the user is the follower.

The utility function of the jammer in the power game is given by

$$r_{jp} = r_{jc} - \lambda_j \times p_{j,v}, \quad (12)$$

where λ_j represents the jamming cost per unit power of the jammer. The objective of the jammer is to maximize the jamming utility

$$a_{jp} = \arg \max_{a_{jp} \in P_j} \{r_{jp}\}. \quad (13)$$

The utility function of the user in the power game is denoted as

$$r_{up} = r_{uc} - \lambda_u \times p_{u,l}, \quad (14)$$

where λ_u represents the transmission cost per unit power. The user chooses the optimum power strategy to maximize its utility

$$a_{up} = \arg \max_{a_{up} \in P_u} \{r_{up}\}. \quad (15)$$

C. STACKELBERG GAME SOLUTION

The jammer and user adopt mixed policies, which define the probability distribution for all possible actions including the jamming channel and jamming power for the jammer, and the communication nodes, corresponding channels, and transmission power for the user. The mixed policies of the jammer and user are respectively denoted as π_j and π_u. The expected utility of the player e ($e \in \{j, u\}$) is obtained by $\hat{r}_e\left(\pi_j, \pi_u\right) = E\left[r_e|\pi_j, \pi_u\right]$. Motivated by [13] and [22], the Stackelberg Equilibrium (SE) is defined as follows.

Definition 1: The strategy profile $\left(\pi_j^*, \pi_u^*\right)$ constitutes the SE if the following conditions in (16) hold. Thus, each player cannot increase its own utility by deviating unilaterally within the game framework,

$$\begin{aligned} \hat{r}_j\left(\pi_j^*, \pi_u^*\right) &\geq \hat{r}_j\left(\pi_j, \pi_u^*\right) \\ \hat{r}_u\left(\pi_j^*, \pi_u^*\right) &\geq \hat{r}_j\left(\pi_j^*, \pi_u\right). \end{aligned} \tag{16}$$

Theorem 1: In the proposed game, there exist stationary policies of the user and jammer that constitute the SE.

Proof: Based on [13], [22]–[24], the finite strategic game has a mixed policy equilibrium. Thus, there exists a SE, in the meaning of stationary strategy, in the proposed game.

Based on *Definition 1* and considering the objective of the user is to maximize its communication utility, the optimal strategy is obtained by

$$\pi_u^* = \arg\max_{\pi_u} \hat{r}_u\left(\pi_j, \pi_u\right). \tag{17}$$

The optimal strategy of the jammer is given by

$$\pi_j^* = \arg\max_{\pi_j} \hat{r}_j(\pi_j, \pi_u(\pi_j)). \tag{18}$$

Therefore, $\left(\pi_j^*, \pi_u^*\left(\pi_j^*\right)\right)$ constitutes a stationary SE. ■

IV. CROSS-LAYER ANTI-JAMMING LEARNING ALGORITHM

A. ALGORITHM DESCRIPTION

In this section, based on Q-Learning, the cross-layer anti-jamming learning algorithm (CALA) is proposed. Denote $\pi_p^u(i) = \left[\pi_{p,1}^u(i), \pi_{p,2}^u(i) \ldots \pi_{p,L}^u(i)\right]$ as the transmission power mixed policy at time i, where $\sum_{l=1}^{L} \pi_{p,l}^u(i) = 1$. The policy $\pi_{p,l}^u(i)$ represents the probability of the user to choose the power action $p_{u,l} \in [p_{u1}, p_{u2} \ldots p_{uL}]$. Denote $\pi_n^u(t) = \left[\pi_{n,1}^u(t), \pi_{n,2}^u(t) \ldots \pi_{n,O_{m+1}}^u(t)\right]$ and $\pi_c^u(t) = \left[\pi_{c,1}^u(t), \pi_{c,2}^u(t) \ldots \pi_{c,N}^u(t)\right]$ as the mixed policies of the node selection and channel selection in the *m-th* hop at time t. Similarly, $\pi_p^j(k) = \left[\pi_{p,1}^j(k), \pi_{p,2}^j(k) \ldots \pi_{p,L}^j(k)\right]$ and $\pi_f^j(k) = \left[\pi_{f,1}^j(k), \pi_{f,2}^j(k) \ldots \pi_{f,N}^j(k)\right]$ are expressed as the mixed policies of the jamming power and jamming channel at time k, respectively.

At time t, the Q-value of the user in L_m is denoted as $Q_{m,t}(o, h, c_n)$, where c_n represents the current communication node in the *m-th* hop, o is expressed as the next node in the *(m+1)-th* hop, and h is the communication channel from the *m-th* node to the *(m+1)-th* node. $Q_{m,t}(o, h, c_n)$ is updated as follows

$$\begin{aligned} Q_{m,t}(o, h, c_n) = Q_{m,t-1}(o, h, c_n) \\ + \alpha\left(r_{m,t} - Q_{m,t-1}(o, h, c_n)\right). \end{aligned} \tag{19}$$

The learning rate $\alpha \in (0, 1)$ satisfies $\sum_{l=0}^{\infty} \alpha_l = \infty$, $\sum_{l=0}^{\infty} \alpha_l^2 < \infty$, which is updated according to

$$\alpha = \alpha_0 / \mu(s_m, a_m), \tag{20}$$

where $\mu(s_m, a_m)$ represents the times that the state-action pair (s_m, a_m) is visited, and α_0 is the initial step size.

The node selection policy $\pi_{n,o}^u(t)$ and channel selection policy $\pi_{c,h}^u(t)$ are given by

$$\pi_{n,o}^u(t+1) = \frac{\tau^{\max_{h'}(Q_{m,t}(o,h',c_n))/\zeta_u}}{\sum_{o''=1}^{O_{m+1}} \tau^{\max_{h'}(Q_{m,t}(o'',h',c_n))/\zeta_u}}, \tag{21}$$

$$\pi_{c,h}^u(t+1) = \frac{\tau^{Q_m(o,h,c_n)/\zeta_u}}{\sum_{h''=1}^{N} \tau^{Q_m(o,h'',c_n)/\zeta_u}}, \tag{22}$$

where ζ_u and τ are the Boltzmann coefficients.

$$\begin{aligned} \zeta_u &= \zeta_0 \tau^{(-\upsilon t)}, \zeta_u \geq \zeta_{final} \\ \zeta_u &= \zeta_{final}, \zeta_u < \zeta_{final}, \end{aligned} \tag{23}$$

where ζ_0 is the initial parameter, which represents the time of exploration, ζ_{final} is expressed as the ending condition in the exploration state, τ and υ are the Boltzmann coefficients which affect the transition from exploration to exploitation.

At time i, the power Q-value of the user is expressed as $Q_{p,i}\left(p_{u,l}\right)$, and the transmission power selection policy is denoted as $\pi_{p,l}^u(i)$,

$$Q_{p,i}\left(p_{u,l}\right) = Q_{p,i-1}\left(p_{u,l}\right) + \alpha\left(r_{u,i} - Q_{p,i-1}\left(p_{u,l}\right)\right), \tag{24}$$

$$\pi_{p,l}^u(i+1) = \frac{\tau^{Q_{p,i}(p_{u,l})/\zeta_p}}{\sum_{l'=1}^{L} \tau^{Q_{p,i}(p_{u,l'})/\zeta_p}}. \tag{25}$$

Similarly, the Q-value of the jammer is denoted as $Q_{j,k}\left(c_f, x, p_{j,v}\right)$, where c_f is expressed as the current jamming channel, x represents the next jamming channel, and $p_{j,v}$ is the next jamming power. $Q_{j,k}\left(c_f, x, p_{j,v}\right)$ is updated as follows

$$\begin{aligned} Q_{j,k}\left(c_f, x, p_{j,v}\right) = Q_{j,k-1}\left(c_f, x, p_{j,v}\right) \\ + \alpha\left(r_{j,k} - Q_{j,k-1}\left(c_f, x, p_{j,v}\right)\right). \end{aligned} \tag{26}$$

Denote the jamming power selection policy and jamming channel selection policy as $\pi_{p,v}^j(k)$ and $\pi_{f,x}^j(k)$

$$\pi_{p,v}^j(k+1) = \frac{\tau^{Q_{j,k}(cf,x,p_{j,v})/\zeta_j}}{\sum_{v'=1}^{L} \tau^{Q_{j,k}(cf,x,p_{j,v'})/\zeta_j}} \tag{27}$$

$$\pi_{f,x}^j(k+1) = \frac{\tau^{Q_{J,k}(cf,x,p_{j,v})/\zeta_j}}{\sum_{x'=1}^{N} \tau^{Q_{j,k}(cf,x',p_{j,v})/\zeta_j}}. \tag{28}$$

The utility of the user and the jammer are respectively given by

$$r_u = r - I(r - R) \times \lambda_u \times p_{u,l}, \tag{29}$$

$$r_j = r_{max} - r - \lambda_j \times p_{j,v}, \tag{30}$$

where R is the minimum communication requirement, and $I(r-R)$ represents the sign function, which is equal to 1 if $r < R$ and 0 if $r > R$.

Each node detects the currently available channel set $F_m = [f_1, f_2, \ldots f_n], 0 \leq n \leq N$, where the common channel between adjacent nodes is denoted as $f_m \in F_m$. According to the current network topology, W routing paths $LK_w, 1 \leq w \leq W$, each of which consists of M links, can be obtained using the minimum hop routing algorithm, where $LK_w = [L_1, L_2, \ldots L_m, \ldots L_M]$ and $L_m = f_m$.

If $W = 1$, i.e. there is only one minimum hop routing path, the user chooses channels for each link according to the current channel selection policy $\pi_c^u(t)$. But, if $W > 1$, i.e. the minimum hop routing path is not unique, the user selects the routing path out of W routing paths and corresponding channels according to the node and channel selection policies $\pi_n^u(t)$ and $\pi_c^u(t)$, respectively.

The user measures the communication utility $r_m = C(f_m)$ for each link L_m and the data throughput $r = \min(r_1, r_2, \ldots r_m, \ldots r_M)$. If the data throughput r is under the minimum communication requirement R, i.e. $r < R$, the user selects transmission power according to the power selection policy $\pi_p^u(i)$ and plays a power game with the jammer. Finally, the user communicates on the optimal path with the optimal transmission power.

B. PERFORMANCE ANALYSIS

Motivated by [13], [22], and [25], the evolution of the Q-values can be described using the differential equation as follows

$$\frac{dQ(k+1)}{dk} = \alpha(r - Q(k)). \tag{31}$$

However, the evolution of the policies, rather than that of the Q-values, is of greater interest in this work. Using Equation (31) and differentiating Equation (27) with respect to k

$$\begin{aligned}\frac{d\pi_{p,v}^j(k)}{dk} &= \pi_{p,v}^j(k)\frac{\alpha_j}{\zeta_j}\left\{\left[r_{j,v}(k-1) - \sum_{v'=1}^{L}\pi_{p,v'}^j(k)\, r_{j,v'}(k-1)\right]\right. \\ &\quad \left. - \zeta_j \sum_{v'=1}^{L}\pi_{p,v'}^j(k)\ln\left[\left[\pi_{p,v}^j(k)/\pi_{p,v'}^j(k)\right]\right]\right\}. \end{aligned} \tag{32}$$

Based on [25], the steady jamming power selection strategy profile can be obtained by equating the right-hand side of Equation (32) to zero, which can be expressed as

$$\pi_{p,s}^{j*} = \frac{\tau^{r_{j,s}/\zeta_j}}{\sum_{s'=1}^{L}\tau^{r_{j,s'}/\zeta_j}}. \tag{33}$$

Algorithm 1 Cross-Layer Anti-Jamming Learning Algorithm

Step 1: Set $k = 0$ and initialize $Q_{j,k}\left(c_f, x, p_{j,v}\right) = 0$, $\pi_p^j(k) = 1/L$, $\pi_f^j(k) = 1/N$.

Step 2: The jammer selects jamming power and jamming channel according to $\pi_p^j(k)$ and $\pi_f^j(k)$.

Step 3: The user learns its optimal routing-channel selection and optimal transmission power.

(1) Routing-channel selection

1) Set $t = 0$ and initialize $Q_{m,t}(o, h, c_n) = 0$. Initialize $\pi_n^u(t) = 1/O_{m+1}$ and $\pi_c^u(t) = 1/|F_m|$.
2) If $W = 1$, choose channels for each link according to $\pi_c^u(t)$. If $W > 1$, select the routing path and corresponding channels according to $\pi_n^u(t)$ and $\pi_c^u(t)$.
3) The user measures r and updates $Q_{m,t}(o, h, c_n)$ and $\pi_c^u(t)$.
4) Update $t = t + 1$, and update α_u, ζ_u.
5) Go to 2) and repeat process until converge

(2) If $r < R$, the user plays a power game with the jammer.

1) Set $i = 0$, and initialize $Q_{p,i}\left(p_{u,l}\right) = 0$ and $\pi_p^u(i) = 1/L$.
2) The user chooses transmission power according to $\pi_p^u(i)$, and measures r_u.
3) Update $Q_{p,i}\left(p_{u,l}\right)$ and $\pi_p^u(i)$.
4) Set $i = i + 1$, and update α_p, ζ_p.
5) Go to 2) and repeat process until converge

Step 4: The jammer measures r_j. Then, update $Q_{j,k}\left(c_f, x, p_{j,v}\right)$, $\pi_f^j(k)$, and $\pi_p^j(k)$.

Step 5: Update $k = k + 1$, and update α_j, ζ_j.

Step 6: Repeat algorithm starting from Step 2, until maximal iteration number is reached.

For the channel selection policies of both the user and jammer, and the power and node selection policies of the user, the evolution of these policies can be similarly studied.

According to [13], [22], and [26], the policy profile of the jammer and user at time t can be denoted as $\pi(t) = \left(\pi_j(t), \pi_u(t)\right)$. An ordinary differential equation (ODE) can be used to determine the convergence of $\pi(t)$. The right-hand side of Equation (32) can be denoted as $f(\pi)$. As $\alpha \to 0$, $\pi(t)$ weakly converges to $\left(\pi_j^*, \pi_u^*\left(\pi_j^*\right)\right)$, which is the solution of $\frac{d\pi}{dt} = f(\pi)$, with any initial condition $\pi(0) = \pi_0$.

Theorem 2: The proposed algorithm can converge to a mixed strategy SE.

Proof: Inspired by [27], the Q-learning algorithm can converge to the true optimal value of the state-action pair, provided it satisfies $\sum_{l=0}^{\infty}\alpha_l = \infty$, $\sum_{l=0}^{\infty}\alpha_l^2 < \infty$, and all actions in every state are accessed with non-zero probability. As the algorithm iterates, each state is visited a sufficient number of

TABLE 1. Simulation parameters.

Parameters	Value
Minimum communication requirement R	0.8Mbit/s
Jamming cost per unit power λ_j	0.2
Transmission cost per unit power λ_u	0.2
Channel bandwidth W_{ch}	1.2MHz
Iteration count K_j, K_u, K_p	1000
Initial learning rate α_0	1
Boltzmann model coefficients $\tau, \upsilon_j, \upsilon_u, \zeta_0, \zeta_{final}$	1.5, -0.05, -0.1, 0.1

times, and α is reduced to 0. Therefore, the CALA algorithm will converge to the steady policy.

According to [13], the mixed strategy SE can be proved by contradiction. Assuming that the learning process can converge to a non-SE point. However, based on [26, Th. (3.1)], the learning process converges to a stable point, which is the solution of the ordinary differential equation. Therefore, the non-SE point is stable. ■

V. SIMULATION RESULTS

In this section, the simulation results are presented to demonstrate the performance of the proposed CALA algorithm. In this paper, we focus on selecting the optimal routing path and corresponding channels from the minimal hop routing paths. There are $N = 5$ channels available for information transmission. The number of the minimal hop routing paths $W = 4$, including 6 communication nodes. Each communication path consists of $M = 3$ links. The other simulation parameters are shown in Table 1.

A. PERFORMANCE OF ROUTING-CHANNEL SELECTION GAME

In this simulation, the user plays a routing-channel selection game with the jammer. The jammer chooses the optimal channel to jam, and the user learns the optimal selection of routing path and corresponding channels. The equivalent noise power in each channel of the network topology A is expressed as $N^A_{\text{mW}} = [0.05, 0.1, 0.25, 1.25, 3]$. The jamming power $p_j = 10$mW, and the transmission power of user $p_u = 5$mW. The simulated network topology A without jamming is shown in Fig. 2, where $g_{m,i}$ is expressed as the *i-th* node in the *m-th* hop. The solid black lines represent the

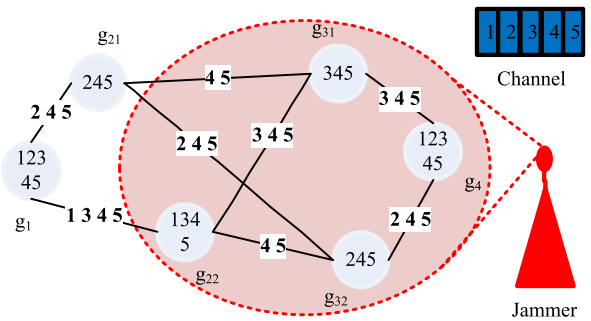

FIGURE 2. Network topology *A* without jamming.

available channel sets. The first hop between g_1 and $g_{2,i}$ can select from 2 nodes. The available channel sets are expressed as $(g_1, g_{21}) = [2, 4, 5]$ and $(g_1, g_{22}) = [1, 3, 4, 5]$, respectively. Similarly, the available channel sets in the second hop are given by $(g_{21}, g_{31}) = [4, 5]$, $(g_{21}, g_{32}) = [2, 4, 5]$, $(g_{22}, g_{31}) = [3, 4, 5]$, and $(g_{22}, g_{32}) = [4, 5]$. The available channel sets in the third hop are expressed as $(g_{31}, g_4) = [3, 4, 5]$ and $(g_{32}, g_4) = [2, 4, 5]$.

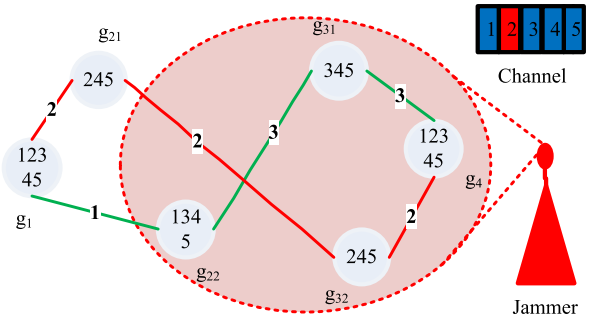

FIGURE 3. Network topology *A* with jamming.

The simulated network topology A with jamming is shown in Fig. 3. The red solid line in Fig. 3 represents the optimal communication path with the maximal data throughput in the current network before the jamming is launched. The green solid line is expressed as the communication path selected by our proposed algorithm under the condition of smart jamming.

Fig. 4 shows the channel selection probability of the jammer. At the beginning of the simulation, the jammer launches jamming attacks on the 5 channels with equal probability. As the algorithm iterates, the selection probability of channel 2 converges to 1 in about 850 iterations, and those of the other channels converge to 0 as expected. The optimal communication path without jamming is now disrupted.

Fig. 5 - Fig. 7 show the convergence of the channel selection probabilities in the first hop, the second hop, and the third hop, respectively. The node and channel selection policies of the user converge to node 2 (i.e. g_{22}) and channel 1 in the first hop, node 1 (i.e. g_{31}) and channel 3 in the second hop, and channel 3 in the third hop. The user selects the sub-optimal communication path as shown by the green line in Fig. 3.

As we can see from Fig. 3 - Fig. 7, before the jamming is launched, channel 2 provided the maximum transmission rate. Thus, the data throughput of the entire communication

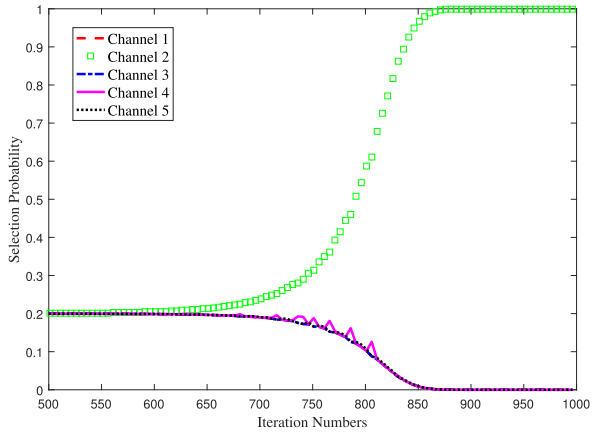

FIGURE 4. The channel selection probability of the jammer.

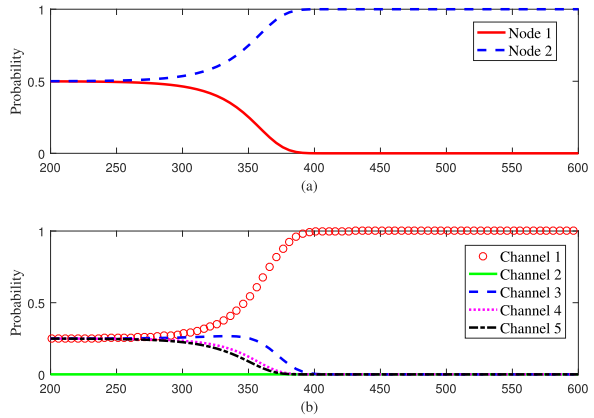

FIGURE 5. The node and channel selection probabilities in the first hop.

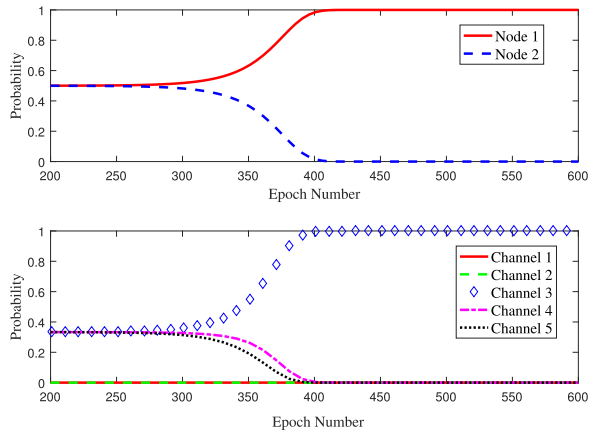

FIGURE 6. The node and channel selection probabilities in the second hop.

path is limited by the channel capacity of channel 2. The jammer then launches jamming attacks on channel 2 to disrupt the optimal communication path which is shown by the red solid line in Fig. 3. However, the user reselects the current sub-optimal communication path which is shown by the green line in Fig. 3. Now, the transmission rate is limited by the channel capacity of channel 3.

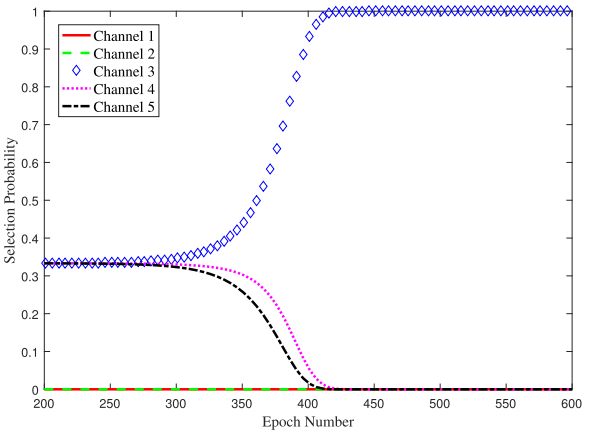

FIGURE 7. The channel selection probability in the third hop.

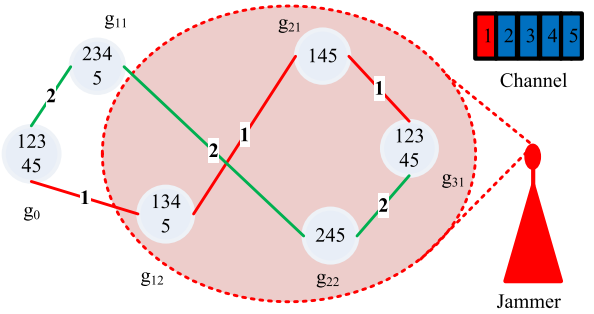

FIGURE 8. Network topology *B* with jamming.

The available channel sets of each node change over time. Consider the original network topology A changing to network topology B as shown in Fig. 8. After the routing-channel selection game, the channel selection policy of the jammer converges to channel 1 to disrupt the optimal communication path (the red solid line in Fig. 8) prior to jamming. The user then reselects the sub-optimal communication path (the green solid line in Fig. 8) under the condition of jamming. Thus, node 1 and channel 2 are selected in the first hop, node 2 and channel 2 are selected in the second hop, and channel 2 is selected in the third hop.

B. PERFORMANCE OF THE ROUTING-CHANNEL AND POWER GAME

If the wireless communication system is threatened by both smart jamming and other fixed jamming, switching channel may not be sufficient to meet the minimum communication requirement. Then, the routing-channel selection game and the power game are formulated to provide effective and reliable communication to the user. The equivalent channel noise of the network topology B is $N_{\mathrm{mW}}^{B} = [0.5, 2, 4, 6, 8]$. The jamming power options can be expressed as $p_{j,\mathrm{mW}} = [1, 3, 5, 7, 9]$. The initial transmission power of the user is $p_{u0} = 1\mathrm{mW}$, and the transmission power options can be denoted as $p_{u,\mathrm{mW}} = [1, 3, 5, 7, 9]$. The current network topology is shown in Fig. 8.

As indicated in Fig. 9, the selection policies of the jamming power and jamming channel converge to power 2 (3mW) and

TABLE 2. The comparison of power selection policies.

	CALA	RPSP	1mW	3mW	5mW	7mW	9mW
Jamming Channel	1	1	1	1	1	1	1
Jamming Power (mW)	3	3	1	3	3	3	5
Normalized Utility	1	0.81	0.57	0.82	0.97	1	0.95
Data Throughput (Mbit/s)	2.6039	2.0490	0.8844	1.5863	2.1688	2.6039	2.9513

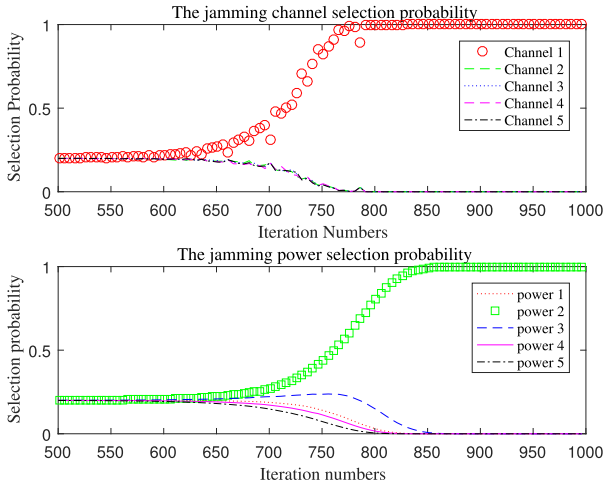

FIGURE 9. The selection probabilities of the jamming channel and jamming power.

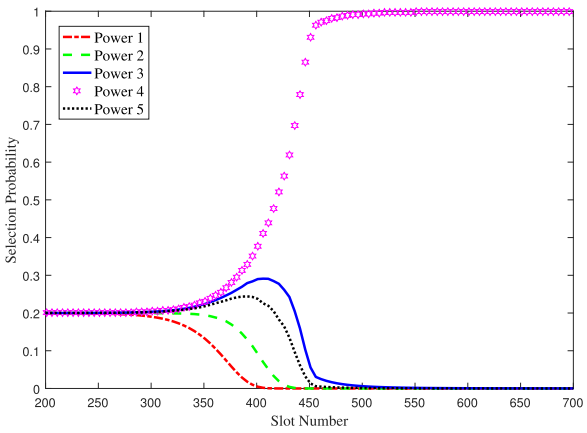

FIGURE 10. The power selection probability of the user.

channel 1, respectively, to disrupt the optimal communication path (the red solid line in Fig. 8) prior to jamming.

As can be seen from Fig. 10, the transmission power selection policy of the user converges to power 4 (7mW) which is optimal in terms of communication utility. The user reselects the sub-optimal communication path under jamming which is shown by the green solid line in Fig. 8.

C. PERFORMANCE COMPARISON

As the cross-layer anti-jamming methods for smart jamming in the field of wireless communication are not proposed so far, the proposed algorithm is compared with the fixed selection policy and random selection algorithm. The performance

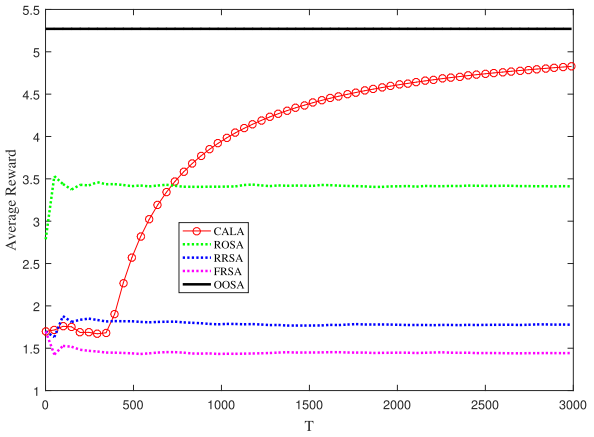

FIGURE 11. The comparison of average reward.

comparison of the cross-layer anti-jamming approach and fixed-independent anti-jamming methods in the network and MAC layers is shown in Fig. 11. The optimal-routing and optimal-channel selection algorithm (OOSA) is able to make the ideal decision with perfect information but it cannot be implemented in practice. In the fixed-routing and random-channel selection algorithm (FRSA), we choose a fixed routing path and randomly select the corresponding channels from the available channel set. Similarly, in the random-routing and optimal-channel selection algorithm (ROSA), we randomly choose a routing path and select the best channels with the maximum channel capacity. The random-routing and random-channel selection algorithm (RRSA) is used to randomly select the routing path and corresponding channels.

The average utility $\bar{R}$, as determined by (34), of the OOSA algorithm, FRSA algorithm, ROSA algorithm, RRSA algorithm and the proposed algorithm (CALA) can be compared in this section.

$$\bar{R} = \frac{\sum_{t=1}^{T} r_t}{T}. \tag{34}$$

Under the condition of jamming, the proposed algorithm eventually yields a higher average reward than the FRSA, ROSA, and RRSA algorithms. Furthermore, the average reward of the proposed algorithm generally converges to that of the OOSA algorithm. The communication utility and data throughput of the proposed power selection policy, the random power selection policy (RPSP), and the fixed power policy are shown in Table 2, and Fig. 12 - Fig. 13.

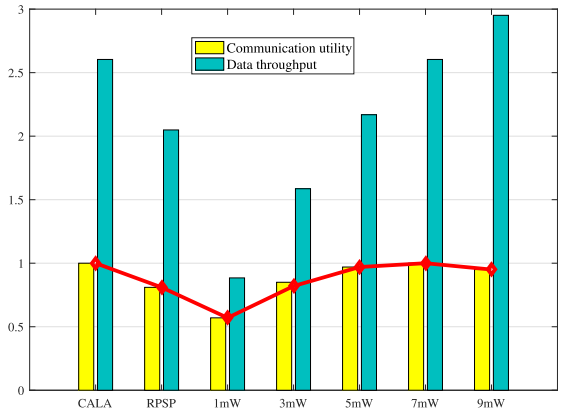

FIGURE 12. The comparison of the normalized communication utility and data throughput.

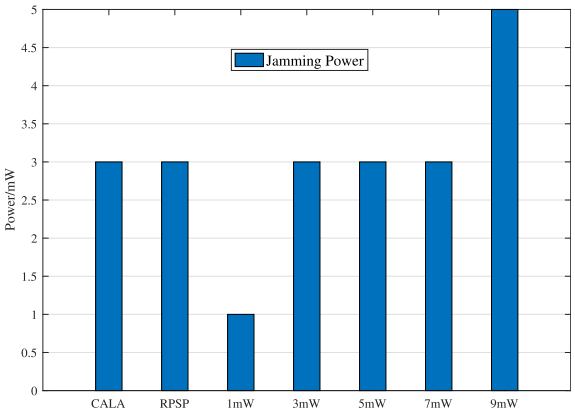

FIGURE 13. The corresponding jamming power of the jammer.

As shown in Fig. 12, the power selection policy of the proposed CALA algorithm yields higher communication utility and data throughput than the RPSP. Compared with the fixed power policies of 1mW, 3mW, 5mW, 7mW, and 9mW, it can be seen that the transmission power (7mW) selected by CALA algorithm is optimal as it yields the maximal normalized communication utility which makes a trade-off between data throughput and power consumption. Fig. 13 shows the corresponding jamming power of the jammer. It can be seen that the jammer can adjust its jamming power policy to maximize the jamming utility with its limited power.

VI. CONCLUSION

In this paper, the joint problem of routing-channel selection and power control in smart jamming environment was formulated as a Stackelberg game. Then, a cross-layer anti-jamming learning algorithm (CALA) was proposed to learn the optimal communication path and transmission power. Finally, the simulation results showed that the joint optimization in the network layer and MAC layer for anti-jamming had better performance than the fixed and independent anti-jamming methods. The user was able to determine the optimal anti-jamming strategy for effective and reliable communication in the dynamic jamming environment. In the future, we will investigate the challenging issues which are caused by the sharp increase in strategy space due to the smarter jammer and the more complex communication systems.

REFERENCES

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

[2] R. El-Bardan, S. Brahma, and P. K. Varshney, "Strategic power allocation with incomplete information in the presence of a jammer," *IEEE Trans. Commun.*, vol. 64, no. 8, pp. 3467–3479, Aug. 2016.

[3] T. Song, W. E. Stark, T. Li, and J. K. Tugnait, "Optimal multiband transmission under hostile jamming," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 4013–4027, Sep. 2016.

[4] D. Yang, G. Xue, J. Zhang, A. Richa, and X. Fang, "Coping with a smart jammer in wireless networks: A Stackelberg game approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4038–4047, Aug. 2013.

[5] L. Liu, G. Han, S. Chan, and M. Guizani, "An SNR-assured anti-jamming routing protocol for reliable communication in industrial wireless sensor networks," *IEEE Commun. Mag.*, vol. 56, no. 2, pp. 23–29, Feb. 2018.

[6] Y. Wu, B. Wang, K. J. R. Liu, and T. C. Clancy, "Anti-jamming games in multi-channel cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 4–15, Jan. 2012.

[7] Q. Zhu, H. Li, Z. Han, and T. Basar, "A stochastic game model for jamming in multi-channel cognitive radio systems," in *Proc. IEEE ICC*, May 2010, pp. 1–6.

[8] F. Yao, L. Jia, Y. Sun, Y. Xu, S. Feng, and Y. Zhu, "A hierarchical learning approach to anti-jamming channel selection strategies," *Wireless Netw.*, pp. 1–13, Jul. 2017, doi: 10.1007/s11276-017-1551-9.

[9] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhuang, "UAV relay in VANETs against smart jamming with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4087–4097, May 2018.

[10] G. Cheng, W. Liu, Y. Li, and W. Cheng, "Joint on-demand routing and spectrum assignment in cognitive radio networks," in *Proc. IEEE ICC*, Jun. 2007, pp. 6499–6503.

[11] Q. Wang and H. Zheng, "Route and spectrum selection in dynamic spectrum networks," in *Proc. CCNC*, Jan. 2006, pp. 625–629.

[12] Y. Zhang, F. Fu, and M. V. D. Schaar, "On-line learning and optimization for wireless video transmission," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3108–3124, Jul. 2010.

[13] L. Jia *et al.*, "A hierarchical learning solution for anti-jamming Stackelberg game with discrete power strategies," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 818–821, Dec. 2017.

[14] L. Xiao, T. Chen, J. Liu, and H. Dai, "Anti-jamming transmission Stackelberg game with observation errors," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 949–952, Jun. 2015.

[15] L. Xiao, J. Liu, Q. Li, N. B. Mandayam, and H. V. Poor, "User-centric view of jamming games in cognitive radio networks," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 12, pp. 2578–2590, Dec. 2015.

[16] S. Machuzak and S. K. Jayaweera, "Reinforcement learning based anti-jamming with wideband autonomous cognitive radios," in *Proc. IEEE/CIC ICCC*, Jul. 2016, pp. 1–5.

[17] M. B. Ghorbel, B. Hamdaoui, M. Guizani, and B. Khalfi, "Distributed learning-based cross-layer technique for energy-efficient multicarrier dynamic spectrum access with adaptive power allocation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1665–1674, Mar. 2016.

[18] A. Ghaffari, "Real-time routing algorithm for mobile ad hoc networks using reinforcement learning and heuristic algorithms," *Wireless Netw.*, vol. 23, no. 3, pp. 703–714, 2017.

[19] J. Nie and S. Haykin, "A Q-learning-based dynamic channel assignment technique for mobile communication systems," *IEEE Trans. Veh. Technol.*, vol. 48, no. 5, pp. 1676–1687, Sep. 1999.

[20] W. Wang, A. Kwasinski, D. Niyato, and Z. Han, "A survey on applications of model-free strategy learning in cognitive wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1717–1757, 3rd Quart., 2016.

[21] Y.-S. Chen, C.-J. Chang, and F.-C. Ren, "Q-learning-based multirate transmission control scheme for RRM in multimedia WCDMA systems," *IEEE Trans. Veh. Technol.*, vol. 53, no. 1, pp. 38–48, Jan. 2004.

[22] Y. Sun, H. Shao, X. Liu, J. Zhang, J. Qiu, and Y. Xu, "Traffic offloading in two-tier multi-mode small cell networks over unlicensed bands: A hierarchical learning framework," *KSII Trans. Internet Inf. Syst.*, vol. 9, no. 11, pp. 4291–4310, 2015.
[23] X. Chen, H. Zhang, T. Chen, and M. Lasanen, "Improving energy efficiency in Green femtocell networks: A hierarchical reinforcement learning framework," in *Proc. ICC*, Jun. 2013, pp. 2241–2245.
[24] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[25] A. Kianercy and A. Galstyan, "Dynamics of Boltzmann *Q* learning in two-player two-action games," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 85, no. 4, p. 041145, 2012.
[26] P. S. Sastry, V. V. Phansalkar, and M. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," *IEEE Trans. Syst., Man and*, vol. 24, no. 5, pp. 769–777, May 1994.
[27] R. Li, Z. Zhao, X. Chen, J. Palicot, and H. Zhang, "TACT: A transfer actor-critic learning framework for energy saving in cellular radio access networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2000–2011, Apr. 2014.

CHEN HAN received the B.S. degree in electronic information engineering from Beihang University, Beijing, China, in 2016. He is currently pursuing the M.S. degree with the College of Communication Engineering, Army Engineering University of PLA. His research interests include learning theory and communication anti-jamming technology.

YINGTAO NIU received the M.S. degree from PLA Commanding Communication Academy, China, in 2005, and the Ph.D. degree from the Institute of Communication Engineering, PLA University of Science and Technology Institute, China. He has authored over 30 journal and conference papers. His main research interests are spread-spectrum communication, cognitive radio theory and techniques, with particular emphasis on algorithms of wireless communication signal processing and decision-making in cognitive radio systems.