A Hierarchical Learning Solution for Anti-Jamming Stackelberg Game With Discrete Power Strategies

Luliang Jia, *Student Member, IEEE*, Fuqiang Yao, Youming Sun, *Student Member, IEEE*, Yuhua Xu, *Member, IEEE*, Shuo Feng, *Student Member, IEEE*, and Alagan Anpalagan, *Senior Member, IEEE*

***Abstract*—This letter investigates the anti-jamming problem with discrete power strategies, and then a Stackelberg game is formulated to model the competitive interactions between the user and jammer. Specifically, the user acts as the leader, whereas the jammer is the follower. Based on their own utilities, the user and jammer select their power strategies and determine their respective optimal strategies. Also, a hierarchical power control algorithm (HPCA) is proposed to obtain the Stackelberg equilibrium, and the asymptotic convergence is analyzed. In addition, we consider the impact of the imperfect information due to the jammer's bounded rationality and inaccurate observation of the user's action. Finally, simulations are conducted to show the effectiveness of the proposed HPCA algorithm, and simulation results demonstrate that the jammer's bounded rationality and limited observation lead to the increase of the user's utility.**

***Index Terms*—Anti-jamming, Stackelberg game, power control, Q-learning.**

I. Introduction

THE THREAT of jamming attacks is a serious issue to the security of wireless networks. In particular, the jammers, who are able to learn the user's transmission strategies, can launch more threating and devastating attacks. To cope with jamming attacks, various techniques have been proposed, including both non-game theoretic methods [1] and game theoretic methods [4]–[8]. As a powerful mathematical tool, game theory [2] is adequate to analyze the interactions between the users and jammers. Among these game theoretic models, Stackelberg game stands out as a natural paradigm that can be used to analyze the hierarchical competition between the user and jammer, and to make a sequential decision-making. Timing channel [3] was exploited, and an attacker-defender Stackelberg game was formulated in [4]. In [5] and [6], a power control Stackelberg game was proposed in the presence of a jammer, which can learn the user's transmission strategies. Li *et al.* [7] investigated the anti-jamming power control problem in the cooperative wireless networks, and a Stackelberg game was employed to analyze the interactions between the transmitter and jammer. In [8], an anti-jamming Bayesian Stackelberg game with incomplete information was proposed, and the optimal strategies based on duality optimization theory were derived. The above studies focused on the continuous power strategies. However, the discrete power scenarios are practically appealing in current communication systems such as 3GPP LTE networks.

In this letter, we investigate the anti-jamming problem with discrete power, and subsequently formulate it as a Stackelberg game, in which the user acts as the leader, and the jammer is the follower. Compared with the continuous power strategies, the discrete power scenario brings about new challenges, and existing methods cannot be directly applied. Therefore, it is particularly necessary and challenging to develop new methods for the discrete power scenarios. To achieve the solution of the formulated game, we resort to learning technologies, which can obtain desirable solutions through repeated interactions. The Q-learning [9], [10], which is commonly used and that obtains decision policies through interaction with the environment, has been widely adopted in wireless communication systems. In this letter, a hierarchical learning framework is formulated, and a hierarchical power control algorithm (HPCA) based on Q-learning is proposed. Note that, in a Stackelberg game, the follower may suffer from bounded rationality and limited observation in real-world domains [11]. In other words, the jammer may deviate from the optimal choice due to inaccurate observation of the leader's strategy, and it may choose sub-optimal strategies due to bounded rationality as well. In this letter, the impact of the follower's bounded rationality and limited observation are considered.

The main contributions of this letter are given as follows.

- We develop a hierarchical learning solution for anti-jamming Stackelberg game with discrete power strategies.
- To obtain the solution of the formulated game, a hierarchical power control algorithm (HPCA) is proposed.
- The impact of the jammer's bounded rationality and inaccurate observation are analyzed.

Note that a hierarchical learning approach to anti-jamming channel selection can be found in our work [14]. The main differences are: i) the considered network scenarios are different, i.e., channel selection for scenario with multiple users and one jammer was considered in [14] while discrete power strategies for scenario with one user and one jammer was considered in this letter, and ii) the impact of jammer's bounded rationality and imperfect observations were considered in this letter.

Manuscript received July 21, 2017; revised August 18, 2017; accepted August 26, 2017. Date of publication August 31, 2017; date of current version December 15, 2017. This work was supported in part by the Natural Science Foundation for Distinguished Young Scholars of Jiangsu Province under Grant BK20160034, in part by the National Science Foundation of China under Grant 61771488, Grant 61631020, Grant 61671473, Grant 61401508, and Grant 61401505, in part by the Jiangsu Provincial Natural Science Foundation of China under Grant BK20130069, Grant BK20151450, and Grant BK20141071, and in part by the Open Research Foundation of Science and Technology in Communication Networks Laboratory. The associate editor coordinating the review of this paper and approving it for publication was S. De. *(Corresponding author: Yuhua Xu.)*

L. Jia, Y. Sun, and Y. Xu are with the College of Communication Engineering, PLA Army Engineering University, Nanjing 210007, China (e-mail: jiallts@163.com; sunyouming10@163.com; yuhuaenator@gmail.com).

F. Yao is with the Nanjing Telecommunication Technology Institute, Nanjing 210007, China (e-mail: yfq2030@163.com).

S. Feng is with the Cognitive Systems Laboratory, McMaster University, Hamilton, ON L8S 4L8, Canada (e-mail: fengs13@mcmaster.ca).

A. Anpalagan is with the Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: alagan@ee.ryerson.ca).

Digital Object Identifier 10.1109/LWC.2017.2747543

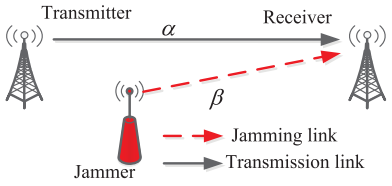

Fig. 1. System model.

II. System Model and Problem Formulation

A. System Model

As shown in Fig. 1, our model consists of one user (a transmitter-receiver pair) and one jammer [5]. Let α and β denote the channel gain of the transmission link and jamming link, respectively. The jammer can learn the user's transmission strategies and adjust its strategies accordingly. We assume that the user and jammer respectively select their strategies from their discrete power sets $\mathcal{P} = \{p_1, \ldots, p_m, \ldots, p_M\}$ and $\mathcal{J} = \{\varphi_1, \ldots, \varphi_n, \ldots, \varphi_N\}$. Similar to [5], based on Signal-to-Interference-plus-Noise Ratio (SINR), the utility of the user is $\mu_s(p_m, \varphi_n) = (\alpha p_m/(\delta_0^2 + \beta\varphi_n)) - c_s p_m$, where δ_0^2 is the noise power, c_s represents the user's transmission cost per unit power; p_m and φ_n denote the transmission power of the user and jammer, respectively. Similarly, the jammer's utility is $\mu_j(p_m, \varphi_n) = -(\alpha p_m/(\delta_0^2 + \beta\varphi_n)) - c_j\varphi_n$, where c_j denotes the jamming cost per unit power of the jammer.

B. Game Model

In this letter, a Stackelberg game is formulated. Specifically, the user acts as the leader, whereas the jammer is the follower. Let 1 and 2 denote the index of the user and jammer, respectively. Mathematically, the Stackelberg game is expressed as $\mathcal{G} = \{\mathcal{N}, \mathcal{P}, \mathcal{J}, \mu_s, \mu_j\}$, where $\mathcal{N} = \{1, 2\}$ denotes the player set, $\mathcal{P}$ and $\mathcal{J}$ respectively represent the strategy space of the user and jammer, μ_s and μ_j are the utility function of the user and jammer, respectively.

In a Stackelberg game, because of the follower's inaccurate observation and bounded rationality, the follower may deviate from its expected strategy. To describe the inaccurate observations, the observation factor $\gamma \in [0, 1]$ is introduced. That is, the larger the observation factor γ is, the higher probability of accurate observation it is. If $\gamma = 1$, it means that the jammer can perfectly observe the user's action.

To explicitly capture the limited observation, we model the observation error matrix as $H = [\phi_{g,l}]_{0\le g,l\le M}$, where $\phi_{g,l}$ denotes the probability that the power level p_g is to be taken as p_l. We assume $\phi_{g,g} = \gamma$ if $g = l$ holds, and otherwise, $\phi_{g,l} = (1-\gamma)/(M-1)$, where $\phi_{g,g}$ represents the probability that the jammer can accurately observe the user's power p_g.[1]

Considering the limited observation, the jammer's utility is:

$$\tilde{\mu}_j(p_m, \varphi_n) = \sum_{l=1}^{M} \phi_{m,l}\left(-\frac{\alpha p_l}{\delta_0^2 + \beta\varphi_n} - c_j\varphi_n\right). \tag{1}$$

From the perspective of the jammer, given the user's power strategy $p_m \in \mathcal{P}$, the optimization problem can be written as:

$$\max_{\varphi_n} \tilde{\mu}_j(p_m, \varphi_n). \tag{2}$$

[1]The basic observation error matrix used in this letter is only for the simplicity of illustration, other forms would also be feasible.

Due to the follower's bounded rationality, it may not strictly maximize its utility. Motivated by [11], in order to describe the uncertainty due to the bounded rationality, the rationality factor $\eta \in [0, 1]$ is introduced, which denotes the rationality degree. Larger rationality factor η means higher rationality degree. If $\eta = 1$, it means that the follower is perfectly rational.

To capture the bounded rationality, we define the rationality matrix as $\Gamma = [\psi_{w,k}]_{0\le w,k\le N}$, where $\psi_{w,k}$ represents the probability that the power level φ_w is regarded as φ_k. If $w = k$ holds, $\psi_{w,w} = \eta$ represents the rationality degree of the jammer; Otherwise, $\psi_{w,k} = (1-\eta)/(N-1)$.

Considering the bounded rationality, the user's utility is:

$$\tilde{\mu}_s(p_m, \varphi_w) = \sum_{k=1}^{N} \psi_{w,k}\left(\frac{\alpha p_m}{\delta_0^2 + \beta\varphi_k} - c_s p_m\right). \tag{3}$$

Thus, the user's optimization problem can be expressed as:

$$\max_{p_m} \tilde{\mu}_s(p_m, \varphi_n). \tag{4}$$

C. Stackelberg Game Solution

In anti-jamming field, the player with mixed strategy can fool its opponent due to randomness. In this letter, we consider the case where both the user and jammer adopt a mixed strategy, which denotes a probability distribution for all possible power strategies. It is assumed that $\mathbf{q}_1$ and $\mathbf{q}_2$ represent the mixed strategies of the user and jammer, respectively. The expected utility of the player i ($i \in \{s, j\}$) is given by $\hat{\mu}_i(\mathbf{q}_1, \mathbf{q}_2) = E[\mu_i|\mathbf{q}_1, \mathbf{q}_2]$. Motivated by [2], [8], and [13], the Stackelberg Equilibrium (SE) is defined as follows.

Definition 1: The policy profile $(\mathbf{q}_1^*, \mathbf{q}_2^*)$ constitutes the SE, if no player can improve its utility by deviating unilaterally in the hierarchical framework, and the following conditions hold.

$$\hat{\mu}_s(\mathbf{q}_1^*, \mathbf{q}_2^*) \ge \hat{\mu}_s(\mathbf{q}_1, \mathbf{q}_2^*), \tag{5}$$

$$\hat{\mu}_j(\mathbf{q}_1^*, \mathbf{q}_2^*) \ge \hat{\mu}_j(\mathbf{q}_1^*, \mathbf{q}_2). \tag{6}$$

Lemma 1: There exists a user's stationary strategy and a smart jammer's stationary strategy, which constitute a SE.

Proof: Inspired by [12]–[14], every finite strategy game has a mixed strategy equilibrium [2], that is, there exists a SE in the sense of stationary strategy in the formulated game.

According to Definition 1, the jammer aims to maximize its utility, and its best-response policy is given by:

$$\mathbf{q}_2^* = \underset{\mathbf{q}_2}{\operatorname{argmax}}\, \hat{\mu}_j(\mathbf{q}_1, \mathbf{q}_2). \tag{7}$$

The optimal policy of the user is:

$$\mathbf{q}_1^* = \underset{\mathbf{q}_1}{\operatorname{argmax}}\, \hat{\mu}_s(\mathbf{q}_1, \mathbf{q}_2(\mathbf{q}_1)). \tag{8}$$

Thus, $(\mathbf{q}_1^*, \mathbf{q}_2^*(\mathbf{q}_1^*))$ forms a stationary SE. ■

III. Hierarchical Learning Algorithm

A. Algorithm Description

In this section, based on Q-learning [9], [10], the HPCA algorithm is proposed. It is assumed that the user and jammer are intelligent agents. A user's mixed policy is represented as $\mathbf{q}_1(k) = (q_{1,1}(k), q_{1,2}(k), \ldots, q_{1,m}(k), \ldots, q_{1,M}(k))$, and $\sum_{m=1}^{M} q_{1,m}(k) = 1$. The user's policy $q_{1,m}(k)$ means the probability with which it chooses the power action p_m from the set $\mathcal{P}$. A jammer's mixed policy can be

denoted as $\mathbf{q}_2(t) = (q_{2,1}(t), q_{2,2}(t), \ldots, q_{2,n}(t), \ldots, q_{2,N}(t))$, and $\sum_{n=1}^{N} q_{2,n}(k) = 1$. The jammer's policy $q_{2,n}(t)$ is the probability that it chooses the power action φ_n from the set $\mathcal{J}$.

Then, the user's Q value is updated according to:

$$Q_{1,m}(k+1) = (1-\kappa_1^k)Q_{1,m}(k) + \kappa_1^k r_{1,m}(k), \tag{9}$$

where $\kappa_1^k \in [0,1)$ is the learning rate, satisfying $\sum_{k=0}^{\infty}\kappa_1^k = \infty$, $\sum_{k=0}^{\infty}(\kappa_1^k)^2 < \infty$, and $r_{1,m}(k) = \tilde{\mu}_s(p_m, \varphi_n)$ denotes the observed reward. The user's policy updates as:

$$q_{1,m}(k+1) = \frac{\exp\left[Q_{1,m}(k)/\tau_0\right]}{\sum_{w\in\mathcal{P}}\exp\left[Q_{1,w}(k)/\tau_0\right]}, \tag{10}$$

where τ_0 controls the tradeoff of exploration-exploitation.

Similarly, the jammer's Q value is updated as:

$$Q_{2,n}(t+1) = (1-\kappa_2^t)Q_{2,n}(t) + \kappa_2^t r_{2,n}(t), \tag{11}$$

where $\kappa_2^t \in [0,1)$ is the learning rate, satisfying $\sum_{t=0}^{\infty}\kappa_2^t = \infty$, $\sum_{t=0}^{\infty}(\kappa_2^t)^2 < \infty$, and $r_{2,n}(t) = \tilde{\mu}_j(p_m, \varphi_n)$. The jammer's policy is updated according to:

$$q_{2,n}(t+1) = \frac{\exp\left[Q_{2,n}(t)/\tau_0\right]}{\sum_{r\in\mathcal{J}}\exp\left[Q_{2,r}(t)/\tau_0\right]}. \tag{12}$$

The proposed HPCA is given in Algorithm 1, and the user and jammer update their policies at different time scales. The stop criterion is when either the maximum iteration number is reached, or the probability vector of $t+1$ is the same as t.

Algorithm 1 Hierarchical Power Control Algorithm (HPCA)

Step 1: Set t=0, k=0 and initialize the mixed policy $q_i(t)$ and Q values $Q_i(\cdot)$, $i \in \mathcal{N} = \{1, 2\}$.

Step 2: In the kth epoch, the user selects its power p_m from the discrete power set $\mathcal{P}$ according to its policy $q_1(k)$.

Step 3: The jammer's learning process.

(1) In the tth slot, the jammer selects its jamming power φ_n from the set $\mathcal{J}$ according to its policy $q_2(t)$.

(2) The jammer measures its utility $r_{2,n}(t)$.

(3) The jammer updates its Q values according to (11) and policy according to (12).

(4) Update t=t+1, and until the stopping criterion holds.

Step 4: The user measures its utility $r_{1,m}(k)$.

Step 5: The user updates its Q values according to (9) and policy according to (10).

Step 6: Go to step 2, and until the stopping criterion holds.

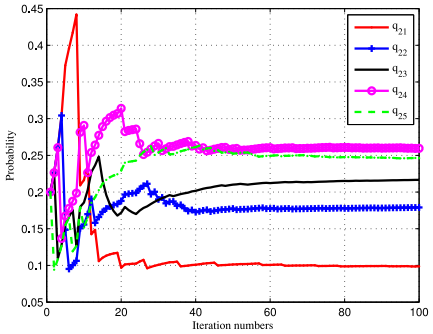

Fig. 2. The learning process of the jammer in the first epoch.

B. Performance Analysis

In the following, we investigate the convergence of the proposed HPCA algorithm. Motivated by [10], [12], and [13], to describe the evolution of the Q values of the user, we have

$$\frac{dQ_{1,m}(k+1)}{dk} = \kappa_1^k\left(r_{1,m}(k) - Q_{1,m}(k)\right). \tag{13}$$

However, we would like to investigate the evolution of the strategies compared to the Q values. By differentiating (10) with respect to k and using (13), we have

$$\frac{dq_{1,m}(k)}{dk} = q_{1,m}(k)\frac{\kappa_1^k}{\tau_0}\left\{\left[r_{1,m}(k-1) - \sum_{\rho\in\mathcal{P}} q_{1,\rho}(k)r_{1,\rho}(k-1)\right] - \tau_0\sum_{\rho\in\mathcal{P}} q_{1,\rho}(k)\ln\left(\frac{q_{1,m}(k)}{q_{1,\rho}(k)}\right)\right\}. \tag{14}$$

According to [10], the steady policy $q_1^s(k)$ is expressed as:

$$q_1^s(k) = \frac{\exp\left[r_{1,m}(k)/\tau_0\right]}{\sum_{v\in\mathcal{P}}\exp\left[r_{1,v}(k)/\tau_0\right]}. \tag{15}$$

For the jammer, we have similar conclusions.

Motivated by [12]–[14], the strategy profile of all players can be denoted as $\mathbf{q}(t) = (\mathbf{q}_1(t), \mathbf{q}_2(t))$. In order to capture the asymptotic convergence of $\mathbf{q}(t)$, we resort to an ordinary differential equation (ODE) [15]. The right-hand side of (14) can be expressed as $f(\mathbf{q})$. As $\kappa_i^t \to 0$, $\mathbf{q}(t)$ can converge weakly to $(\mathbf{q}_1^*, \mathbf{q}_2^*(\mathbf{q}_1^*))$, which is the solution of $\frac{d\mathbf{q}}{dt} = f(\mathbf{q})$, with initial condition $\mathbf{q}(0) = \mathbf{q}_0$.

Lemma 2: The proposed HPCA algorithm can converge to the optimal strategy.

Proof: Inspired by [9] and [13], as long as the condition $\sum_{k=0}^{\infty}\kappa_i^k = \infty$, $\sum_{k=0}^{\infty}(\kappa_i^k)^2 < \infty$ holds, the convergence is guaranteed. For brevity, we omit the proof, and readers can refer to [9] and [13] for detailed proof. ■

Proposition 1: The proposed HPCA can discover a SE.

Proof: Motivated by [12] and [14], it is proved by contradiction. We assume that the learning process converges to a non-SE point. According to [15, Th. 3.1], the learning process converges to the stationary point, which is the solution of the ODE. Therefore, the non-SE points are stationary, which contradicts the Lemma 1. ■

IV. Numerical Results and Discussions

In this section, simulation results are presented. Referring to [5] and [8], the channel gain α and β are set in the interval $[0.3, 0.9]$. The set $\mathcal{P} = \mathcal{J} = [0.5W, 1.0W, 1.5W, 2.0W, 2.5W]$. Each epoch contains T=100 time slots. Other parameters are given as: $c_j = c_s = 0.2$, $\delta_0^2 = 0.1$.

The update of the jammer's selection probability in the first epoch is presented in Fig. 2. Fig. 3 shows the user's convergence behavior over epoch numbers. The selection probability of the user (or jammer) converges to a stationary mixed strategy in about 60 iterations (or epoch numbers).

To evaluate the proposed HPCA algorithm, it is compared with the random selection algorithm (RSA), in which the user

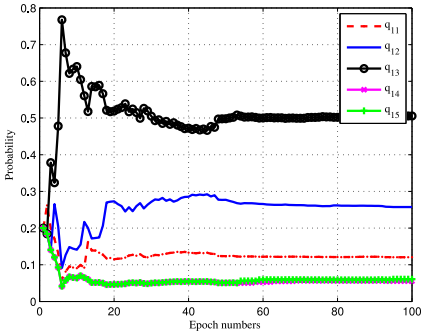

Fig. 3. The learning process of the user over epoch numbers.

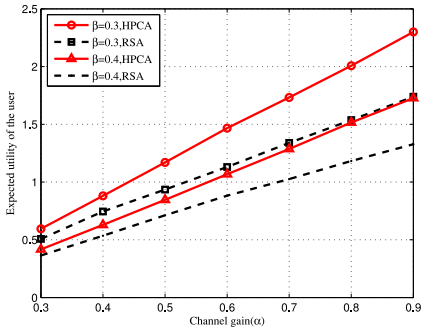

Fig. 4. Performance comparison of the user's utility for different solutions.

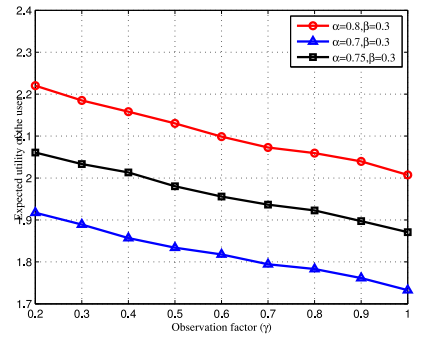

Fig. 5. The influence of the observation factor γ.

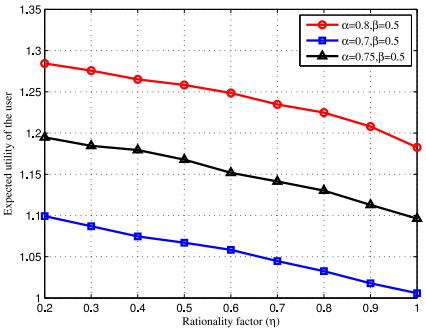

Fig. 6. The influence of the rationality factor η.

randomly selects one power action at each time. Fig. 4 indicates that the proposed HPCA algorithm yields higher utility of the user than the RSA algorithm. The reason is that the proposed HPCA algorithm can converge to a desirable solution, whereas the RSA algorithm is an instinctive approach.

Fig. 5 and Fig. 6 respectively show the impact of the observation factor γ and rationality factor η on the user's utility. As can be seen from Fig. 5, because of the inaccurate observation, the jammer will deviate from its optimal strategy, and the limited observation leads to the increase of the user's utility. Also, the improvement of the user's utility degrades with the growth of observation factor γ. As indicated in Fig. 6, due to the bounded rationality, the jammer may not respond with the maximum utility, which results in the improvement of the user's utility. Moreover, increasing rationality factor η degrades the improvement of the utility of the user.

V. Conclusion

In this letter, the power control problem with discrete power strategies had been investigated. We formulated an anti-jamming Stackelberg game to analyze the competitive interactions between the user and jammer. Then, a hierarchical power control algorithm (HPCA) was proposed. Moreover, we considered the impact of the bounded rationality and limited observation. Finally, the simulations were conducted, and the results have shown that the user will benefit from the jammer's bounded rationality and limited observation.

References

[1] S. D'Oro, E. Ekici, and S. Palazzo, "Optimal power allocation and scheduling under jamming attacks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 3, pp. 1310–1323, Jun. 2017.

[2] Z. Han *et al.*, *Game Theory in Wireless and Communication Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[3] W. Xu, W. Trappe, and Y. Zhang, "Anti-jamming timing channels for wireless networks," in *Proc. 1st ACM Conf. Wireless Netw. Security*, Alexandria, VA, USA, 2008, pp. 203–213.

[4] S. D'Oro *et al.*, "Defeating jamming with the power of silence: A game-theoretic analysis," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2337–2352, May 2015.

[5] 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.

[6] 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.

[7] Y. Li, L. Xiao, J. Liu, and Y. Tang, "Power control Stackelberg game in cooperative anti-jamming communications," in *Proc. GAMENETS*, Beijing, China, 2014, pp. 1–6.

[8] L. Jia, F. Yao, Y. Sun, Y. Niu, and Y. Zhu, "Bayesian Stackelberg game for antijamming transmission with incomplete information," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 1991–1994, Oct. 2016.

[9] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[10] 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, pp. 1–10, Apr. 2012.

[11] J. Pita, M. Jain, M. Tambe, F. Ordóñez, and S. Kraus, "Robust solutions to Stackelberg game: Addressing bounded rationality and limited observations in human cognition," *Artif. Intell.*, vol. 174, no. 15, pp. 1142–1171, Oct. 2010.

[12] X. Chen, H. Zhang, T. Chen, and M. Lasanen, "Improving energy efficiency in green femtocell networks: A hierarchical reinforcement learning framework," in *Proc. IEEE ICC*, Budapest, Hungary, 2013, pp. 2241–2245.

[13] Y. Sun *et al.*, "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, Nov. 2015.

[14] F. Yao *et al.*, "A hierarchical learning approach to anti-jamming channel selection strategies," *Wireless Netw.*, pp. 1–13, Jul. 2017, doi: 10.1007/s11276-017-1551-9.

[15] P. S. Sastry, V. V. Phansalkar, and M. A. L. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 5, pp. 769–777, May 1994.