Joint Power and Channel Selection for Anti-jamming Communications: A Reinforcement Learning Approach

Xufang Pei(✉), Ximing Wang, Lang Ruan, Luying Huang, Xingyue Yu, and Heyu Luan

College of Communications Engineering, Army Engineering University of PLA,
Nanjing 210000, China
peixufang@163.com, lgdxwxm@sina.com, ruanlangjyy@163.com,
15931876710@163.com, fistaon@163.com, 13201677586@163.com

Abstract. In this paper, the decision-making problem for anti-jamming communications is studied. Most of the existing anti-jamming researches mainly focus on the single-domain anti-jamming such as power domain or frequency domain, which has limited performance facing strong jamming. Therefore, to effectively deal with some jamming attack, this paper proposes a multi-domain joint anti-jamming scheme, and considers the power domain and the frequency domain jointly. By modeling the anti-jamming process as a Markov decision process (MDP), reinforcement learning (RL) is adopted to solve the MDP. Then, the multi-domain joint anti-jamming algorithm is proposed to find the optimal decision-making strategy. Moreover, the proposed algorithm is verified to converge to an effective strategy. Simulation results show that the proposed algorithm has better throughput performance than the sensing-based random selection algorithm.

Keywords: Multi-domain anti-jamming · Markov decision process (MDP) · Reinforcement learning

1 Introduction

Owing to the open nature of radio, wireless communication is badly threaten by jamming attacks [1–3]. Recently, with the fast advancement of artificial intelligence technologies, jamming technologies have become increasingly intelligent. Due to low spectrum utilization and fixed transmission patterns, traditional anti-jamming technologies such as spread spectrum and frequency hopping technologies [4,5] can not be able to meet the increasing anti-jamming

This work was supported by the National Natural Science Foundation of China under Grant No. 61771488, No. 61671473 and No. 61631020, in part by the Natural Science Foundation for Distinguished Young Scholars of Jiangsu Province under Grant No. BK20160034.

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 551–562, 2019.
https://doi.org/10.1007/978-3-030-32388-2_47

requirements. Intelligent anti-jamming technologies are required to enhance the jamming-resistance ability of wireless communication systems.

In the existing research works [6–9], game theory can well model the decision-making interaction process between players, and has been widely used in the field of wireless communication anti-jamming. For example, authors in [10,11] used the Stackelberg game to model the confrontational relationship between users and jammers, and obtained the anti-jamming decision by solving the equilibrium solution. Similarly, authors in [12,13] modeled the confrontational relationship through zero-sum game. However, all the above studies assume that both sides of the game (users and jammers) know each other's information, which is impractical in actual communication.

Reinforcement learning is an effective way to make real-time decision making in unknown environment [14–16]. Researchers have applied reinforcement learning to explore the optimal policy for dynamic spectrum access and anti-jamming problems. For instance, authors in [17] investigated the dynamic spectrum access problem in multi-user scenario. In [18], the dynamic spectrum anti-jamming problem in fading environment was studied. Considering the difference of channel transmission rate of actual channel, a reinforcement learning based channel selection scheme was proposed, which significantly improves the throughput performance of users compared with the random channel selection algorithm. In [14], authors proposed a modified Q-learning algorithm. When the cognitive agent learns the jamming mode of the jammer, it adopts a way of updating the Q value table in parallel, which improves the convergence speed of the algorithm. However, these studies mainly focus on solving the single-domain anti-jamming problems. It will fail when the power of the jamming is strong or the frequency band is very wide.

There are several studies that consider multi-domain anti-jamming. In [19], a multi-domain anti-jamming decision-making problem with unknown channel state was studied. Specifically, in the power domain, the user's transmit power was adjusted to confront the jamming. When the jamming was severe (the jamming power exceeds a certain threshold), the channel switching mode chose to avoid the jamming attack. However, this mechanism only considers switching between power domain and frequency domain, which has low energy efficiency. A joint optimization of power and frequency resources is in need.

Inspired by above studies, this paper studies the multi-domain joint anti-jamming problem. Modeling the anti-jamming decisions as MDP, our object is to maximize long-term cumulative throughput while considering transmission overhead. In order to find the optimal strategy, a multi-domain joint anti-jamming algorithm based on reinforcement learning is designed. In simulation results, the performance of the proposed algorithm is verified compared with the sensing-based random selection strategy algorithm.

The remainder of the paper is organized as follows. In Sect. 2, the system model and problem formulations are investigated. In Sect. 3, a Q-learning based multi-domain anti-jamming communication scheme is proposed. In Sect. 4, the simulations and analysis are given. In the end, the paper is concluded in Sect. 5.

2 System Model

2.1 System Model

As Fig. 1 shows, in the system model, there exist one user (containing a transmitter and a receiver) and a malicious jammer. The available channel set is assumed to be $\mathcal{M} = \{1, 2, \ldots, M_i\}$, i.e., there are M available channels with B bandwidth. The user's available power set is defined as $\mathcal{P} = \{1, 2, \ldots, P_i\}$, and the jamming power is defined as a constant J. The transmitter transmits data information to the receiver through data link. The receiver is responsible for running the intelligent decision-making algorithm and returns the decision information to the transmitter through the control link. The jammer transmits the jamming signal to block the normal communication of the user. In order to facilitate calculation and intelligent decision making, we divide the transmission time into several equal-length slot units, and the transmission slot set is defined as $\{1, 2, \ldots, K\}$. The user is assumed to access only one channel in each time slot.

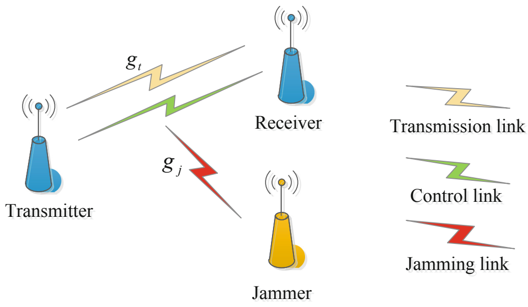

Fig. 1. System model.

Assuming that the channel has large-scale fading, the signal will have path loss during transmission. As shown in Fig. 1, g_t represents the link gain between user transmitter and receiver, Specifically, it can be defined as

$$g_t = (d_t)^{-\alpha} \varepsilon_t, \tag{1}$$

where d_t denotes the distance between the user transmitter and receiver, α represents the user's path fading factor, ε_t represents the user's instantaneous fading coefficient. Similarly, g_j represents the link gain from the jammer to the user receiver, which is specifically defined as:

$$g_j = (d_j)^{-\beta} \varepsilon_j, \tag{2}$$

where d_j denotes the distance from the jammer to the user receiver, β represents the path fading factor of the jammer, and ε_j represent the instantaneous fading coefficient of the jammer, both ε_t and ε_j obey the lognormal fading. Therefore, the user's signal-to-interference-plus-noise ratio can be denoted as follows:

$$SINR = \frac{g_t P(k)}{N_0 + g_j J \delta(f_t - f_j)}, \tag{3}$$

where $P(k)$ represents the transmission power selected by the user, N_0 represents the background noise power, J represents the power of jamming, f_t represents the channel of user signal, f_j represents the channel of jamming signal, and $\delta(\cdot)$ is the indication function. The indication function indicates the occupancy of the selected working channel of the user, and the specific definition is as follows:

$$\delta(f_t - f_j) = \begin{cases} 1, f_t = f_j, \\ 0, f_t \neq f_j. \end{cases} \tag{4}$$

That is, when $f_t = f_j$ indicates that the jamming is on the same channel as the user, the user collides with the jamming, otherwise the two are on different channels, that is, the user is not interfered.

2.2 Problem Modeling

In order to solve the problems mentioned above [20,21], user's anti-jamming decision process can be modeled as a Markov decision process (MDP). MDP is generally defined by a four-tuple, namely (S, A, O, R), whose core elements are defined as follows: S represents the state space, A represents the action space, O represents the state transition probability matrix, and R represents the reward value.

In the actual communication scenario, assuming that there are M_i available channel, the user has P_i power levels. Taking the $M_i = 5$, $P_i = 3$, and jamming modes to continuously apply the sweep jamming as an example, the jammer interferes with multiple channels simultaneously. The jamming variation pattern is shown in Fig. 2. In the figure, the yellow square denotes that the current channel is disturbed and the white square denotes that it is not disturbed. To define the k-th time slot user channel and power selection strategy as $a(k)$, the user's utility can be defined as:

$$U_k = \begin{cases} B\log_2(1 + SINR) - C_s P(k) + X, SINR \geq \Gamma \\ 0, \ otherwise \end{cases}, \tag{5}$$

where B represents the channel bandwidth and C_s represents the user unit power transmission cost. X represents a constant in case of the reward of the user being negative. Γ indicates the set threshold. If $SINR$ is higher than the given threshold Γ, the transmitted data packet can be successfully received. Otherwise, if $SINR$ is less than a given threshold Γ, the transmitted data packet fails to be received.

Under the jamming condition, the user starts to select the best transmission channel and transmission power according to its own strategy at each time

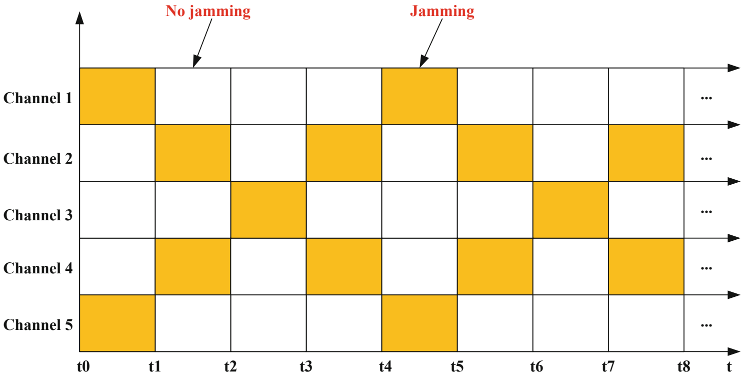

Fig. 2. Diagram of cross sweep jamming. (Color figure online)

slot. Under energy constraints, users must save transmission power to reduce transmission overhead while meeting minimum communication requirements. Therefore, the user's reward $R(k)$ is defined as:

$$R_k = U_k \frac{T_{\text{tran}}}{T_s}, \tag{6}$$

where T_s is the length of a time slot length, T_{tran} is the transmission time. The user's goal is to find the optimal policy to get the maximum cumulative rewards, which can be formulated as:

$$\pi^* = \arg\max_{\pi \in \Omega} E_\pi [\sum_{k=0}^{\infty} R(k)]. \tag{7}$$

This paper assumes that the jammer's strategy (jamming mode) remains unchanged. For the anti-jamming decision problem in such fixed jamming scenarios, the reinforcement learning method can be used to solve the MDP. Since reinforcement learning can learn the optimal strategy in the unknown environment without state transition probability, this paper uses Q-learning [22] algorithm to solve the power and channel selection optimization problem, it is one of the most widely used algorithms in reinforcement learning. Different from the definition of Q-learning action space in [18], this paper combines channel and power to make decision when selecting action, and proposes a Q-learning based multi-domain joint anti-jamming algorithm.

3 Q-learning Based Multi-domain Anti-jamming Communication Scheme

3.1 Algorithm Description

In the strategy selection process, considering the influence of power against jamming performance, this paper designs a multi-domain anti-jamming algorithm based on reinforcement learning to solve this problem. In the process of the user performing the reinforcement learning algorithm, the user evaluates the quality of different actions in each state by maintaining a Q values table. The Q value reflects the quality of different actions. The larger the Q value, the better the selected action. The algorithm calculates the immediate reward value obtained by taking action in each state, and update the Q values corresponding to each action in real time. The updating function of Q values [18] can be expressed as

$$Q_{k+1}(S_k, a_k) = Q_k(S_k, a_k) + \alpha(R_k + \gamma V_{k+1} - Q_k(S_k, a_k)), \tag{8}$$

where α represents the learning step, γ represents the discount factor, that is, the importance of future returns to the current selection action, $\alpha, \gamma \in (0, 1]$, R_k represents the immediate return value of the current S_k state, and V_{k+1} is the maximum Q values of all strategies in the S_{k+1} state. After the agent selects and executes the action a_k, it reaches the S_{k+1} state in the $(k+1)$-th time slot. The calculation formula of V_{k+1} is as follows:

$$V_{k+1} = \max Q_k(S_{k+1}, \tilde{a}), \forall \tilde{a} \in \mathcal{M} \times \mathcal{P}, \tag{9}$$

$\tilde{a}$ is an optional power and channel set under the state S_{k+1}. The update formula of the action selection probability vector $W(k) = (w_1(k), \ldots, w_c(k))$ denotes as [18]:

$$w_c(k+1) = \frac{\exp(\xi Q(S_k, c))}{\sum\limits_{c \in \mathcal{M} \times \mathcal{P}} \exp(\xi Q(S_k, c))}, \forall c \in \mathcal{M} \times \mathcal{P}, \tag{10}$$

where ξ represents the Boltzmann coefficient constant, $w_c(k+1)$ denotes the probability that the $(k+1)$-th time slot selects the power and channel combination strategy as c.

3.2 Communication Process Description

As shown in Fig. 3, the user-jamming slot diagram in the anti-jamming decision process, wherein the length of the jamming slot is T_j, and the length of the user slot is T_s. The user's single time slot composition includes a transmission phase, a sensing phase, a learning phase, and an ACK feedback phase, and executed in this order.

- Transmission phase: The initial state of a given user is $S_0(f_t(0), f_j(0))$, and the user randomly selects a transmission power $P_t(0)$, that is, the user starts transmitting data on the given channel $f_t(0)$ with the power of the $P_t(0)$ at the 0-th slot, where $f_j(0)$ is obtained through wideband spectrum sensing. Simultaneously, the return value $R(0)$ of the current working channel $f_t(0)$ and transmission power $P_t(0)$ is calculated.
- Sensing phase: Detecting the occupancy of each channel in the current time slot through wideband spectrum sensing and obtaining the jamming channel $f_j(1)$;

- Learning phase: The reinforcement learning algorithm is executed to determine the transmission channel $f_t(1)$ and the transmission power $P_t(1)$ of the next time slot. Note that the Q-learning process time is ignored;
- ACK feedback phase: The selected policy (i.e., the working channel and transmission power of the next time slot) is fed back to the user transmitter through the control link, and the user status is updated to $S_1(f_t(1), f_j(1))$.

In the next $k-1$ time slots, the user goes through the same process to update the working channel and transmission power through the reinforcement learning decision. In particular, the Q values table of the 0-th time slot is an all-zero matrix, and in the subsequent time slots, the user updates the Q values of the selected action in the current state by reinforcement learning. The user cyclically executes the process to continuously enhance the awareness of the environment, and finally achieves a state of stable optimal strategy in a complex dynamic environment. The flow of multi-domain joint anti-jamming algorithm based on reinforcement learning is shown in Algorithm 1.

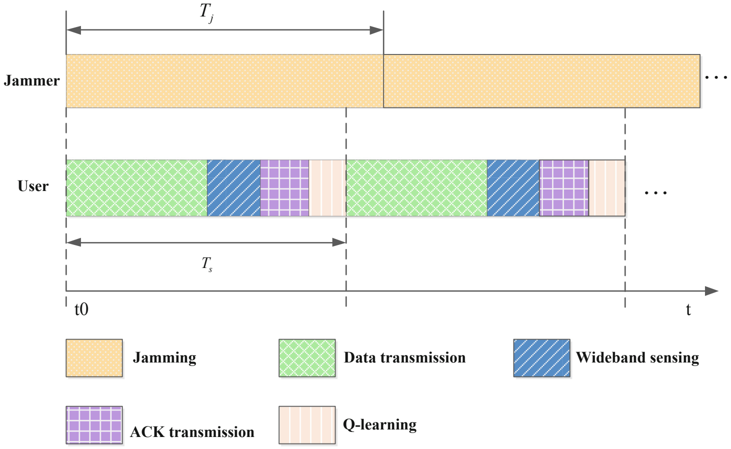

Fig. 3. Time slot structure.

4 Simulation Results and Discussions

This paper mainly studies how to choose the optimal strategy to effectively deal with jamming. The convergence performance of the algorithm is simulated and analyzed. In order to prove the validity of the proposed algorithm, this paper compares the proposed algorithm with the sensing-based random selection algorithm. The sensing-based random selection algorithm firstly implement wideband spectrum sensing at each time slot to obtain the location of the jamming, and then randomly selects one transmitting power and one channel for access. For

Algorithm 1. Q-learning based multi-domain joint anti-jamming algorithm.

```
Initialization: Set parameter $\alpha, \gamma$, the total simulation time slot to $K$ and the time index $k = 0$. Initialize Q values matrix $Q(S, a) = 0$. Set the initial state is represented as $S_0(f_t(0), f_j(0))$.
While $k < K$, do
The user receives data information on the $f_t(k)$ channel with the power of $P_t(k)$, updates the state to $S_k(f_t(k), f_j(k))$, and calculates the $SINR$ of the $f_t(k)$ channel according to Equation (3), and compares whether the $SINR$ is greater than the set threshold;
    if $SINR > \varGamma$
        $R_k = U_k \frac{T_{\mathrm{tran}}}{T_s}$, where $U_k$ can be obtained from Equation (5),
    else
        $R_k = 0$.
    end
    The current jamming channel $f_j(k+1)$ is found by wideband spectrum sensing;
    The action selection probability vector $W(k)$ is updated according to Equation (10), and the action $a(k) = (f_t(k+1), P_t(k+1))$ of the next time slot is selected according to the $P(k)$;
    update the Q value table according to Equation (8);
    Return ACK to the user transmitter, adjust the working channel $f_t(k+1)$ and transmit power $P_t(k+1)$ of the next time slot user;
    $k = k + 1$
End while
```

both algorithms, the performance of the system under different parameters is analyzed.

Considering that there is a user (including transmitter and receiver) and a malicious jamming (interferer) in the wireless communication system, the jammer applies two cross-sweep jamming signals. The system has 4 available channels and 3 power levels. Considering the fading characteristics of the channel, a lognormal fading model is established to reflect the channel quality [17,23], and the channel gain can be expressed as e^Z. Among them, Z represents a Gaussian variable with a mean of zero and a variance of η^2. The lognormal fading model can generally be expressed as $\eta = 0.1 \log(10)\eta_{dB}$. Assuming that the user accesses only one channel in one time slot, the jamming can interfere with two channels at the same time. The algorithm simulation specific parameter settings are shown in Table 1, where the time slot parameter setting refers to [18], and the channel parameter setting refers to [17,23].

The Q values variation curve and the selection probability curve of each action in the state $S(f_t = 2, f_j = 1, 4)$ (i.e., the user transmits data on the transmission channel-2, the jamming signals are in channel-1 and channel-4) are shown in Fig. 4, Fig. 5. The simulation results show that the Q values of each action is 0 at the beginning of the reinforcement learning, and the probability of selecting each action is equal. With the enhancement of users' cognition of the environment, the Q values table maintained by users is constantly updated.

Table 1. Simulation parameter setting.

Parameters	Value
Number of channels	$M_i = 4$
Number of user power levels	$P_i = 3$
The user's transmission power set	$\mathcal{P}_t = \{2\,\mathrm{W},\ 4\,\mathrm{W},\ 6\,\mathrm{W}\}$
Jamming constant power	$\mathcal{P}_j = 3.5\,\mathrm{W}$
Channel noise power spectral density	$N0 = -135\,\mathrm{dB/Hz}$
Channel bandwidth	$B = 1\,\mathrm{MHz}$
The distance between the transmitter and the receiver	$d_t = 5\,\mathrm{km}$
The distance between the jammer and the receiver	$d_j = 25\,\mathrm{km}$
Jamming time slot length	$T_{jam} = 4\,\mathrm{ms}$
Data transmission time	$T_d = 2\,\mathrm{ms}$
ACK transmission time	$T_A = 0.3\,\mathrm{ms}$
Wideband sensing time	$T_W = 0.6\,\mathrm{ms}$
Transmission time slot length	$T_s = T_d + T_A + T_W = 2.9\,\mathrm{ms}$
Learning step	$\alpha = (0, 1]$
Discount factor	$\gamma = 0.8$
Boltzmann coefficient	$\xi = 5–20$

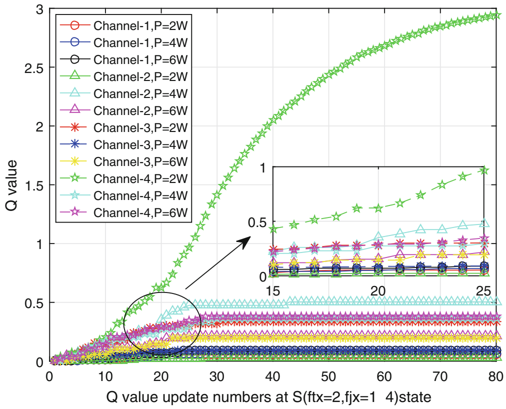

Fig. 4. Q value curves at $S(f_t = 2, f_j = 1, 4)$ state.

In the later stage of learning, the user selects the working channel-4 and the transmission power $2W$ with a probability close to 1.

In Fig. 6, we set the threshold $\Gamma = 3.8\,\mathrm{dB}$, the unit power transmission cost coefficient $C_s = 0.1$. We compared the system throughput performance of the multi-domain joint anti-jamming algorithm and the sensing-based random selection algorithm. In order to make the simulation results more clear, the throughput value of each time slot in the figure is calculated by averaging the throughput value of 50 consecutive time slots. The simulation results

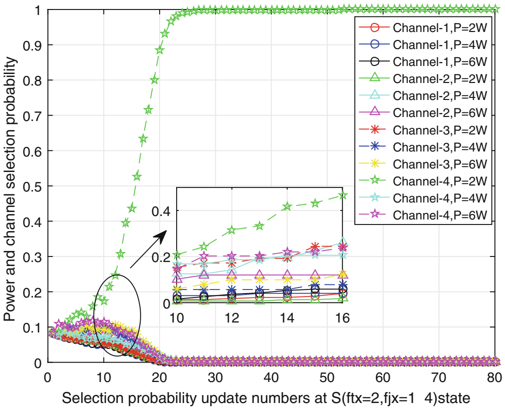

Fig. 5. Selection probability curves at $S(f_t = 2, f_j = 1, 4)$ state.

show that the throughput based on the sensing algorithm is about 0.5Mbps, while that the multi-domain joint anti-jamming algorithm based on Q-learning is about 0.88Mbps. Therefore, the proposed algorithm has better anti-jamming performance than the sensing-based random selection algorithm.

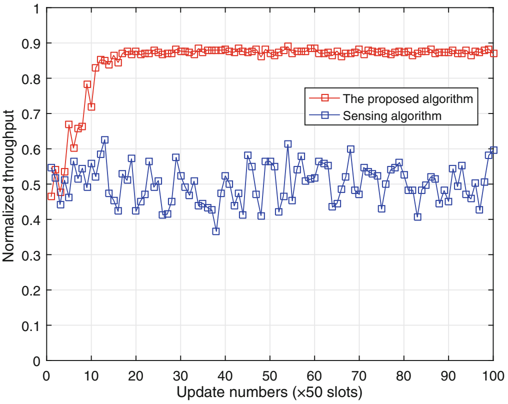

Fig. 6. Performance comparison of throughput for different algorithms.

5 Conclusions

Aiming at the problem of user's power and channel joint decision in jamming environment, the multi-domain anti-jamming decision process is modeled as a MDP. A Q-learning based multi-domain joint anti-jamming algorithm is proposed to execute decision-making. By the exploration and exploitation process of Q-learning, the algorithm can learn the jamming strategy based on the historical information. Simulation results show that the proposed algorithm can not only converge in the complex jamming environment, but also obtain the optimal power and channel selection in the continuous communication process. What's more, the throughput performance of the proposed algorithm is significantly improved compared with the sensing-based algorithm. On the basis of this paper, we will consider more complex jamming environment in the future research.

References

1. Zou, Y., Zhu, J., Wang, X., Hanzo, L.: A survey on wireless security: technical challenges, recent advances, and future trends. Proc. IEEE **104**(9), 1727–1765 (2016)
2. Sagduyu, Y.E., Berry, R.A., Ephremides, A.: Jamming games in wireless networks with incomplete information. IEEE Commun. Mag. **49**(8), 112–118 (2011)
3. Sharma, R.K., Rawat, D.B.: Advances on security threats and countermeasures for cognitive radio networks: a survey. IEEE Commun. Surv. Tutor. **17**(2), 1023–1043 (2015)
4. Pelechrinis, K., Iliofotou, M., Krishnamurthy, S.V.: Denial of service attacks in wireless networks: the case of jammers. IEEE Commun. Surv. Tutor. **13**(2), 245–257 (2011)
5. Worthen, A., Stark, W.: Interference mitigation in frequency-hopped spread-spectrum systems. In: 2000 IEEE Sixth International Symposium on Spread Spectrum Techniques and Applications, vol. 1, pp. 58–62 (2000)
6. Yu, L., Li, Y., Pan, C., Jia, L.: Anti-jamming power control game for data packets transmission. In: 2017 IEEE 17th International Conference on Communication Technology (ICCT), Chengdu, pp. 1255–1259 (2017)
7. Xiao, L., Chen, T., Liu, J., Dai, H.: Anti-jamming transmission stackelberg game with observation errors. IEEE Commun. Lett. **19**(6), 949–952 (2015)
8. Zhang, Y., et al.: A multi-leader one-follower stackelberg game approach for cooperative anti-jamming: no pains, no gains. IEEE Commun. Lett. **22**(8), 1680–1683 (2018)
9. Gao, Y., Xiao, Y., Wu, M., Xiao, M., Shao, J.: Game theory-based anti-jamming strategies for frequency hopping wireless communications. IEEE Trans. Wirel. Commun. **17**(8), 5314–5326 (2018)
10. Jia, L., Yao, F., Sun, Y., Xu, Y., Feng, S., Anpalagan, A.: A hierarchical learning solution for anti-jamming stackelberg game with discrete power strategies. IEEE Wirel. Commun. Lett. **6**(6), 818–821 (2017)
11. Xu, Y., Ren, G., Chen, J., Jia, L., Xu, Y.: Anti-jamming transmission in UAV communication networks: a Stackelberg game approach. In: 2017 IEEE/CIC International Conference on Communications in China (ICCC), Qingdao, pp. 1–6 (2017)
12. Chen, T., Liu, J., Xiao, L., Huang, L.: Anti-jamming transmissions with learning in heterogenous cognitive radio networks. In: 2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), New Orleans, pp. 293–298 (2015)
13. Wang, B., Wu, Y., Liu, K.J.R., Clancy, T.C.: An anti-jamming stochastic game for cognitive radio networks. IEEE J. Sel. Areas Commun. **29**(4), 877–889 (2011)

14. Slimeni, F., Scheers, B., Chtourou, Z., Le Nir, V.: Jamming mitigation in cognitive radio networks using a modified Q-learning algorithm. In: 2015 International Conference on Military Communications and Information Systems (ICMCIS), Cracow, pp. 1–7 (2015)
15. Slimeni, F., Chtourou, Z., Schaeers, B., Nir, V.L., Attia, R.: Cooperative Q-learning based channel selection for cognitive radio. Wirel. Netw. **4**, 1–11 (2018)
16. Xu, N., Zhang, H., Xu, F., Wang, Z.: Q-learning based interference-aware channel handoff for partially observable cognitive radio ad hoc networks. Chin. J. Electron. **26**(4), 856–863 (2017)
17. Wu, Q., Xu, Y., Wang, J., Shen, L., Zheng, J., Anpalagan, A.: Distributed channel selection in time-varying radio environment: interference mitigation game with uncoupled stochastic learning. IEEE Trans. Veh. Technol. **62**(9), 4524–4538 (2013)
18. Kong, L., Xu, Y., Zhang, Y., et al.: A reinforcement learning approach for dynamic spectrum anti-jamming in fading environment. In: IEEE 18th International Conference on Communication Technology (ICCT), Chongqing, pp. 51–58 (2018)
19. Jia, L., Xu, Y., Sun, Y., Feng, S., Yu, L., Anpalagan, A.: A multi-domain anti-jamming defense scheme in heterogeneous wireless networks. IEEE Access **6**, 40177–40188 (2018)
20. Singh, S., Trivedi, A.: Anti-jamming in cognitive radio networks using reinforcement learning algorithms. In: 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN), Indore, pp. 1–5 (2012)
21. Cavazos-Cadena, R., Fernandez-Gaucherand, E.: Markov decision processes with risk-sensitive criteria: dynamic programming operators and discounted stochastic games. In: Proceedings of the 40th IEEE Conference on Decision and Control, vol. 3, pp. 2110–2112 (2001)
22. Machuzak, S., Jayaweera, S.K.: Reinforcement learning based anti-jamming with wideband autonomous cognitive radios. In: 2016 IEEE/CIC International Conference on Communications in China (ICCC), Chengdu, pp. 1–5 (2016)
23. Stuber, G.: Principles of Mobile Communications. Kluwer Academic Publishers, Dordrecht (2001)