A hidden anti-jamming method based on deep reinforcement learning

Yifan Wang[1], Xin Liu[1,2*], Mei Wang[1] and Yu Yu[3]
[1] Guilin University of Technology College of Information Science and Engineering Guilin, China
[2] Guangxi key Laboratory Fund of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin 541007, China
[3]31131 Troops, PLA, China
[e-mail: liuxin2017125@glut.edu.cn]
*Corresponding author: Xin Liu

Received October 16, 2020; revised August 4, 2021; accepted August 28, 2021; published September 30, 2021

Abstract

In the field of anti-jamming based on dynamic spectrum, most methods try to improve the ability to avoid jamming and seldom consider whether the jammer would perceive the user's signal. Although these existing methods work in some anti-jamming scenarios, their long-term performance may be depressed when intelligent jammers can learn user's waveform or decision information from user's historical activities. Hence, we proposed a hidden anti-jamming method to address this problem by reducing the jammer's sense probability. In the proposed method, the action correlation between the user and the jammer is used to evaluate the hiding effect of the user's actions. And a deep reinforcement learning framework, including specific action correlation calculation and iteration learning algorithm, is designed to maximize the hiding and communication performance of the user synchronously. The simulation result shows that the algorithm proposed reduces the jammer's sense probability significantly and improves the user's anti-jamming performance slightly compared to the existing algorithms based on jamming avoidance.

Keywords: Environmental Cognition, Anti-Intelligent Jamming, Deep Reinforcement Learning, Hidden Anti-Jamming

This research work was supported by the National Natural Science Foundation of China under Grant 61961010, 62071135，the Key Laboratory Found of Cognitive Radio and Information Processing, Ministry of Education (Guilin University of Electronic Technology) under Grant No. CRKL200204，the ‘Ba Gui Scholars’ program of the provincial government of Guangxi”.

http://doi.org/10.3837/tiis.2021.09.019 **ISSN : 1976-7277**

1. Introduction

Anti-jamming is a hot topic that aims at realizing continuous and stable communications. Although the mobility and openness of wireless communications bring many conveniences to people's lives, there are also many security issues, such as being susceptible to various types of jamming attacks [1][2][3]. With the development of artificial intelligence, jamming with environmental sense and learning capabilities has posed new challenges to anti-jamming technology. The anti-jamming technology urgently needs to be iteratively upgraded to more intelligent [4][5][6].

To confront intelligent jammers with environmental sense and learning capabilities, genetic algorithms [7], particle swarm algorithms [8], and artificial bee colony algorithms [9] have been used for making frequency, power, or coding action. However, such algorithms need different degrees of prior information, which limits their application in real scenarios. Considering the interaction between users and opposing jammer, game theory has been applied to analyzing communication strategies in anti-jamming [10][11][12][13][14]. Generally, the purpose of the communication user is to avoid jamming, while the jammer wants its frequency to be the same as the user. The utility of both parties is precisely the opposite. The "zero-sum game" is often used as a model for frequency immunity[15]. Considering the hierarchical decision-making of the user and the jammer, frequency domain anti-jamming can also be modeled using a Steinberg game[16]. However, decision algorithms based on game-theoretic models rely on many challenging assumptions in practice, i.e., the user and the jammer know their opponents' channel state information (CSI). To address the reality that the state of the environment is unknown and difficult to get, the anti-jamming technique based on reinforcement learning (RL) has been more widely used and achieved more satisfactory results. Examples include adaptive frequency hopping actions based on Q-learning [17], joint time-frequency anti-jamming communication [18], and intelligent anti-jamming relay systems based on RL [19]. With the improvement of jamming learning ability, users need to face an increasing amount of state space. Anti-jamming communication based on RL is hard to converge quickly, which affects the quality of communication. A sequential deep reinforcement learning (DRL) method that uses deep learning to classify complex environmental states, enabling subsequent use of RL for optimal action making, is proposed in Liu et al. [20]. Although it shortens the convergence time, it can only cope with the dynamic jamming mode. The anti-jamming ability effect would be weakened as the jammer adopts intelligent jamming strategies.

To counteract intelligent jammer, anti-jamming methods based on DRL have become a popular research direction today. For example, the methods proposed in the literature [21][22][23] use the time-frequency 2D information as the original input and apply the DRL technique against intelligent jammer. However, it makes optimal actions by considering only how to maximize jamming avoidance, i.e., just the SINR of the user is used as a basis for judging the immediate returns of the user. DRL is also used in the literature [24] to select the optimal policy. It is different from other articles because it uses a deep deterministic policy gradient (DDPG) instead of stochastic gradient descent (SGD) to update the network parameters. Although it shortens the convergence time, it only relates the immediate returns to the user's SINR when making optimal actions, i.e., it only considers whether the user has avoided jamming. To reduce the energy consumption of the system, [25] defines the environmental state as 3-dimensional information (time, frequency, and power) and also uses DRL algorithms for optimal action making (frequency and power). The system can guarantee

less energy consumption with increased throughput. But the core idea is still how to avoid jamming to the maximum extent possible.

In summary, the core idea of the above methods is to maximize the probability of avoiding jamming. Although an excellent anti-jamming effect can be achieved, the user's past signal waveform and frequency action information may be exposed simultaneously. As the jamming continues to learn the above information, the effectiveness of anti-jamming may diminish. Therefore, coping with intelligent jamming while ensuring that the user's information doesn't leak as far as possible is also a direction that anti-jamming researchers need to explore in-depth. Intelligent jamming relies on the sense of the environment for learning actions, and environmental factors influence the jammer sense of the environment. If the user tries to select a frequency where the jammer does not sense well, it may avoid jamming and information leakage. For example, there is a high power station near the jammer, but far from the user's receiver, the transmitting frequency of this station is an excellent frequency for users to avoid sensing. Unfortunately, the above information is not available to the user in advance. However, the user can indirectly verify whether the user is avoiding being sensed by the jammer by analyzing the correlation between the jammer's action and the user's action. Therefore, an action relevance method is designed to measure the correlation between the jammer's action and the user's action and judge whether the user can avoid being sensed by the jammer. The action correlation measurement and DRL are combined to realize hidden anti-jamming communications. Finally, an anti-jamming deep reinforcement learning algorithm based on hiding strategy(ADRLH) is proposed.

The main contributions are summarized as follows:

•A hidden anti-jamming idea is proposed. The user selects the channels that are hard to detect by the jammer, so the jammer cannot analyze the user decision rules to carry out more targeted jamming.

•A measurement method for evaluating hide performance is designed in this paper. The hiding effect evaluation is complex because the jammer does not actively inform its sensing results. This paper analyzes the hiding effect of users from the perspective of action correlation between the user and the jammer to solve this problem. Supposing the jammer can effectively sense the user's actions, we can infer that the jammer's actions should be highly related to the user's actions, i.e., the reactive jamming can be modeled as a delayed function of the user's actions. Otherwise, if their actions are not relevant, it could be inferred that the jammer doesn't sense the action of the user effectively. Therefore, the correlation between the actions of the user and the jammer can evaluate the user's hidden effect.

•An anti-jamming learning approach aiming at hiding is proposed, avoiding jamming and reducing information exposure. Specifically, in addition to the throughput performance, the hidden performance is taken as a part of the immediate return. The user prefers the decision that can conceal its signal after learning. Therefore, even the highly intelligent jammer cannot make effective jamming strategies due to the lack of valuable sensing information.

The rest part of this paper is presented as follows. In Section 2, the anti-jamming system model is given. After that, an anti-jamming deep reinforcement learning algorithm based on a hiding strategy is presented in Section 3. Besides, the analysis of simulation results and the conclusion are given in Section 4 and Section 5, respectively.

2. Anti-jamming system model

The anti-jamming system model is schematically illustrated in **Fig. 1**. It is mainly composed of a transmitter, a receiver, a jammer, a sensor, and an environmental interference source. The transmitter sends the signals to the receiver. The receiver accepts signals from the environment and converts them into environmental information. It learns to make actions and feeds back to the transmitter whether or not to communicate. And it transmits communication frequencies through the next time slot. The sensor passes the signals obtained from the environment to the jammer. Then the jammer makes action and releases the jamming signal at the corresponding frequency point after learning from the above environmental signals.

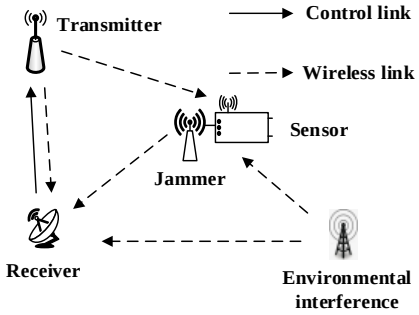


Fig. 1. Anti-jamming system model

The starting frequency and ending frequency of the communication band of both user and jammer is f_s and f_e, the number of channels is N, and the signal bandwidth of both antagonists is $b=\left(f_e-f_s\right)/N$. The whole available frequency set for user and jammer is defined as $\mathcal{F}=\left\{f_e+\frac{b}{2}, f_e+\frac{3b}{2},\cdots, f_e+\frac{(2N-1)b}{2}\right\}$ and $f_t^I, f_t^J \in \mathcal{F}$ represents the user's and jammer's selected frequencies at the moment t, respectively. The transmitting power of the user's signal is $P^I=\int_{-b/2}^{b/2}U\left(f\right)df$, where $U\left(f\right)$ is the baseband power spectral density (PSD) of the user's signal. The baseband PSD of the jamming waveform is $J\left(f\right)$, and the PSD of the environment interference signal is $E\left(f\right)$, and the PSD of the noise signal is $n\left(f\right)$. Let g^{TR} represent the channel gain of the transmitter-receiver link and g^{ER} represent the channel gain of the environmental interference to the receiver. The received SINR of the user can be expressed as:

$$\Phi_t^I=\frac{g^{TR}P^I}{\int_{f_t^I-b/2}^{f_t^I+b/2}\left\{n\left(f\right)+g^{JR}J\left(f-f_t^J\right)+g^{ER}E\left(f\right)\right\}df} \tag{1}$$

The jammer confirmed the user signal by capturing the synchronization sequence of the user. The lower the user's SINR sensed by the jammer, the lower the accuracy of determining the existence of the user signal. Let g^{TJ} indicates the channel gain from the transmitter to the

jammer and g^{EJ} indicates the channel gain from environmental interference to the jammer. The SINR of jammer's receiving single from the user can be represented as:

$$\Phi_{t,n}^{J}=\frac{g^{TJ}P^{I}}{\int_{(n-1)b}^{nb}\left\{\left\{n(f)+g^{RJ}J\left(f-f_t^{J}\right)+g^{EJ}E(f)\right\}df\right\}} \tag{2}$$

where n represents the number of the channel.

The receiving end of the user needs to obtain the environment status in real-time to make communication frequency decisions. Hence it continuously senses the whole communication band. Based on the anti-jamming system model described in Fig.1, the PSD of the signal at the receiving end can be expressed as:

$$R_t(f)=g^{TR}U\left(f-f_t^{I}\right)+g^{JR}J\left(f-f_t^{J}\right)+g^{ER}E(f)+n(f) \tag{3}$$

In the actual processing, the received spectrum is discretized as spectrum vector $\boldsymbol{s}_t=\left\{r_{1,t},r_{2,t},\cdots,r_{N,t}\right\}$, where $r_{n,t}=10\log\left[\int_{n\Delta f}^{(n+1)\Delta f}R_t(f+f_s)df\right]$, and the Δf is the spectral resolution.

3. An Anti-Jamming Deep Reinforcement Learning Algorithm based on Hiding strategy

Although the user does not know whether the jammer is sensing its actions successfully, it can infer some information from the jammer's actions. The action-making process for jammer is shown in **Fig. 2**.

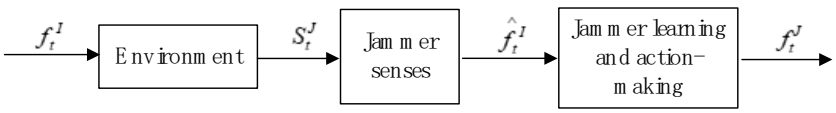


Fig. 2. The jammer's action-making process

From **Fig. 2**, it can be seen that at the moment t when the user frequency action f_t^I is entered into the environment, the jammer would sense the environmental state S_t^J and identify the user's frequency $\hat{f}_t^I$. The jammer then learns from $\hat{f}_t^I$ and decides for the corresponding frequency f_t^J. Assuming the jammer estimates the user's channel right, which is $\hat{f}_t^I=f_t^I$, the jammer's action sequence $\boldsymbol{F}_t^J=\left(f_t^J,f_{t-1}^J,\cdots,f_{t-T+1}^J\right)$ should be highly correlated with the user's action sequence $\boldsymbol{F}_t^I=\left(f_t^I,f_{t-1}^I,\cdots,f_{t-T+1}^I\right)$, where T is the observation length. Conversely, $\boldsymbol{F}_t^I$ has little or no correlation with $\boldsymbol{F}_t^J$, which means that $\hat{f}_t^I\neq f_t^I$, i.e., the jammer does not correctly sense the user's action information.

From the above, it is clear that the user has no direct access to whether its decisions are avoid being sensed by the jammer. However, it is possible to obtain disturbed decision

sequences by receiving environmental states. Comparing the correlation between the user's decision sequence $\boldsymbol{F}_t^J$ and the jammer's decision sequence $\boldsymbol{F}_t^I$ can indirectly verify whether the user's decision avoids being sensed by the jammer. To evaluate the correlation mentioned above, we proposed an action correlation function $\rho_t^{IJ}(\tau)$, which is defined as

$$\rho_t^{IJ}(\tau) = 1 - \frac{d(\boldsymbol{F}_t^J, \boldsymbol{F}_{t+\tau}^I)}{T} \tag{4}$$

where τ is the time offset, and $d(\boldsymbol{a},\boldsymbol{b})$ is the vector distance function, which is the sum distance of all elements. Assuming $\boldsymbol{a} = \{a_0, a_1, \cdots a_{T-1}\}$ and $\boldsymbol{b} = \{b_0, b_1, \cdots b_{T-1}\}$, the $d(\boldsymbol{a},\boldsymbol{b})$ can be expressed as

$$d(\boldsymbol{a},\boldsymbol{b}) = \sum_{n=0}^{T-1} \delta\left[(a_n - b_n) \neq C_{\boldsymbol{a},\boldsymbol{b}}\right] \tag{5}$$

where $\delta(x)$ is the indicator function, it is 1 when x is true and 0 when x is false. $C_{\boldsymbol{a},\boldsymbol{b}}$ is the most frequent element in the set $\{(a_n - b_n)\}, n \in [0, T-1]$, which represents the fixed bias of $\boldsymbol{a}$ and $\boldsymbol{b}$. Since the user doesn't know the delay between the jammer actions and user actions, action correlation functions with possible delay are calculated, and then the largest is selected as the correlation evaluation value, which is expressed as

$$R_t = \max\left\{\rho_t^{IJ}(\tau)\right\}, \tau \in [-T+1, T-1] \tag{6}$$

According to equations (4), (5), and (6), the larger the value of R_t is, the conspicuous the correlation between jammer actions and user actions is, which means that the jammer is likely to perceive the user's decision better. Therefore, if the user wants to hide its actions from the intelligent jammer, the immediate goal is to reduce R_t. At the same time, the user also needs to ensure communication quality, which means that it needs to maximize its SINR ratio Φ_t^I. Combining these two factors, we design the immediate reward r_t as

$$r_t = \log(1 + \Phi_t^I) + \alpha(1 - R_t) \tag{7}$$

where α is used to balance communication quality and hiding effect.

Similar to the deep learning method in [23], we adopt the spectrum waterfall as the input state for retaining sufficient raw information. Hence the environment state is defined as $\boldsymbol{S}_t = \{\boldsymbol{s}_t, \boldsymbol{s}_{t-1}, \cdots, \boldsymbol{s}_{t-T+1}\}$, where T denotes the duration of the waterfall. From the definition of the system model, the user's action is the frequency $f_t^I \in \boldsymbol{\mathcal{F}}$ selected at each time. Paper [22] has proved that the environment state based on spectrum waterfall is a Markov process. Hence the anti-jamming problem in this paper can also be modeled as MDP, where $\boldsymbol{S}_t \in \{\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \cdots\}$ is the set of environment state, $f_t^I \in \boldsymbol{\mathcal{F}}$ is the frequency action of the user, $P\left(\boldsymbol{S}_{t+1} \middle| \boldsymbol{S}_t, f_t^I\right)$ is the transition probability of the state changing from $\boldsymbol{S}_t$ to $\boldsymbol{S}_{t+1}$ after taking action f_t^I, and the $r_t = \log(1 + \Phi_t^I) + \alpha(1 - R_t)$ is the immediate reward.

After all the elements of the MDP are determined, the following learning process is the conventional DRL process, and the unique task is to design the deep network to fit the Q function, which is defined as

$$Q\left(\boldsymbol{S}_t, f_t^I\right)=E\left\{r_t+\gamma \max_{f_{t+1}^I} Q\left(\boldsymbol{S}_{t+1}, f_t^I\right)\middle|\boldsymbol{S}_t, f_t^I\right\} \tag{8}$$

where γ is the reward discount factor. Since the spectrum waterfall $\boldsymbol{S}_t$ is similar to 2D graphics, convolutional neural networks (CNN) are preferred. And the CNN has been verified in [24] that can effectively extract time-frequency features of jamming. After the Q function is perfectly fitted, it can be directly used for online decisions, i.e., the greedy decision can be formulated as $f_t^I=\arg\max_f Q\left(\boldsymbol{S}_t, f;\theta\right)$, where θ is the ideal network weights. So then the core problem is how to get the ideal θ.

In the DRL framework, the network weights, in essence, are trained by historical experience, including action, reward, and state transition, which should be recorded in detail throughout the training process. Specifically, the experience can be stored in a tuple $e_t=\left(\boldsymbol{S}_t, f_t^I, r_t, \boldsymbol{S}_{t+1}\right)$ at each time slot t, and the most recent experience is updated in the set $E=\left(e_1, e_2, \cdots, e_t\right)$. Hence, the set E is the auto-generated and dynamic data set for training. Assuming θ_i represents the network weighting factor at the i-th iteration, the target value can be estimated as $\eta_i=r_t+\gamma \max_{f_{t+1}^I} Q\left(\boldsymbol{S}_{t+1}, f_{t+1}^I;\theta_{i-1}\right)$, which can be seen as the output label of the input $\boldsymbol{S}_t$. The mean square error $L_i\left(\theta_i\right)=E\left(\eta_i-Q\left(\boldsymbol{S}_t, f_t^I;\theta_i\right)\right)^2$ is used as the loss function, and then the gradient of the loss function $L_i\left(\theta_i\right)$ is calculated as

$$\nabla_{\theta_i} L_i\left(\theta_i\right)=E\left[L_i\left(\theta_i\right)\nabla_{\theta_i} Q\left(\boldsymbol{S}_t, f_t^I;\theta_i\right)\right] \tag{9}$$

At last, the updated weight is updated by the θ_i and gradient $\nabla_{\theta_i} L_i\left(\theta_i\right)$ until getting convergence.

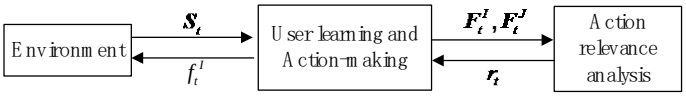


Fig. 3. Flowchart of action making on the user

As the immediate reward is related to the action sequence of both user and jammer, the action-making procedure is more complicated than normal DRL. To illustrate this more clearly, we present **Fig. 3** to explain the entire process. As shown in Fig.3, the user should sense the environment $\boldsymbol{S}_t$ first. Then it can estimate the action sequence $\boldsymbol{F}_t^J$ by analyzing $\boldsymbol{S}_t$ and $\boldsymbol{F}_t^I$ by recording its actions. And then, it is the action correlation analysis procedure, which estimates the correlation between $\boldsymbol{F}_t^J$ and $\boldsymbol{F}_t^I$. Combined with the estimation of the received SINR, the immediate reward can be obtained r_t. At last, the learning algorithm makes a decision f_t^I based on the input state $\boldsymbol{S}_t$ and the current network

weights θ related to the experience $e_t = \left(\boldsymbol{S}_t, f_t^I, r_t, \boldsymbol{S}_{t+1}\right)$. At the same time, the action f_t^I also trigger the state changing form $\boldsymbol{S}_t$ to $\boldsymbol{S}_{t+1}$, and the next period begins. The detailed algorithm of ADRLH is summarized in **Algorthim 1**.

Algorithm 1: An anti-jamming deep reinforcement learning algorithm based on hiding strategy

INIT $E=\varnothing$, $i=0$, θ_0 with random values, $S_1 = O(T \times N)$, $\xi = 1$, Training=True.

FOR $t = 1, 2, \cdots, \infty$ **DO**

Generate the random value $\varepsilon \in U(0,1)$

Sense the environment state $\boldsymbol{S}_t$

IF $\varepsilon > \xi$

Select the action $f_t^I = \arg\max_f Q\left(\boldsymbol{S}_t, f; \theta_i\right)$

ElSE

Select the action f_t^I randomly

END

Execute action f_t^I and sense $\boldsymbol{S}_{t+1}$

Estimate $\boldsymbol{F}_t^J$ and load $\boldsymbol{F}_t^I$ from the record

Compute $\rho_t^{IJ}(\tau), \tau \in [-T+1, T-1]$ and $R_t = \max_\tau \left\{\rho_t^{IJ}(\tau)\right\}$

Compute $r_t = \log\left(1+\Phi_t^I\right) + \alpha\left(1-R_t\right)$

Store $e_t = \left(\boldsymbol{S}_t, f_t^I, r_t, \boldsymbol{S}_{t+1}\right)$ and update set E

IF $Sizeof(E) > 1000$ **and** Training=True

Sample random mini-batch of transitions e_t from set E

Compute η_i $L_i(\theta_i)$, and $\nabla_{\theta_i} L_i(\theta_i)$

Update θ_i to θ_{i+1}

Update $i = i+1$ and decrease ξ

IF ($\frac{1}{K}\sum_{k=i-K+1}^{i} L_k(\theta_k) < L_{TH}$)

Training=False

END

END

END

4. Simulation Results and Analysis

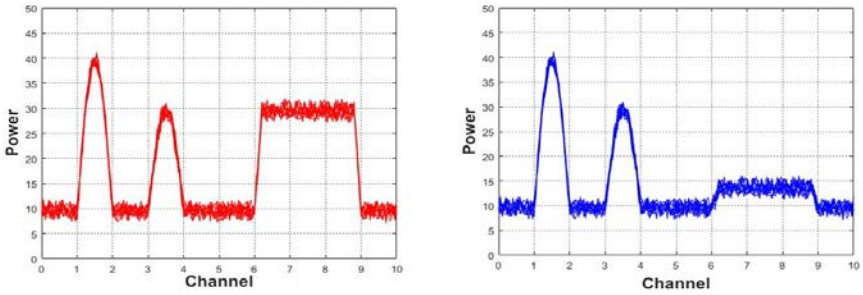
4.1 Simulation parameters

In this paper, the user and the jammer combat each other in a frequency band of 10MHz. And the number of channels $N=10$. The user and jammer perform spectral sensing once per $1ms$, and their center frequency can be changed once per $10ms$. Both user and jammer signals

are raised cosine waveforms with roll-off factor α = 0.5, in which the jammer signal power is 40dBm, and user signal power is 30dBm. The spectral resolution Δf is set to be $1kHz$, and the observation length is $T = 100$, so the environmental state $\boldsymbol{S}_t$ is a two-dimensional matrix 10×100. The balance coefficient is $\alpha = 2$, and the reward discount factor is $\gamma = 0.8$. A total of 10,000 iterations were carried out in the experiment. Due to the dynamic nature of immediate rewards, the original performance curve with the iteration is very confusing. To see the performance trend clearly, we use a 100-level smoothing filter with all coefficients of 0.01 to eliminate jitter.

Two kinds of jamming scenarios are considered for simulation, follower jamming and intelligent jamming based on reinforcement learning. ADRLH is compared with three methods, It contains Adaptive Frequency Hopping(AFH), Frequency Hopping Spread Spectrum(FHSS), and Anti-jamming Deep Reinforcement Learning Algorithm based on Avoiding strategy (ADRLA) proposed in [23].

4.2 Simulation results and analysis



(a) receiving spectrum of the jammer (b)receiving spectrum of the user

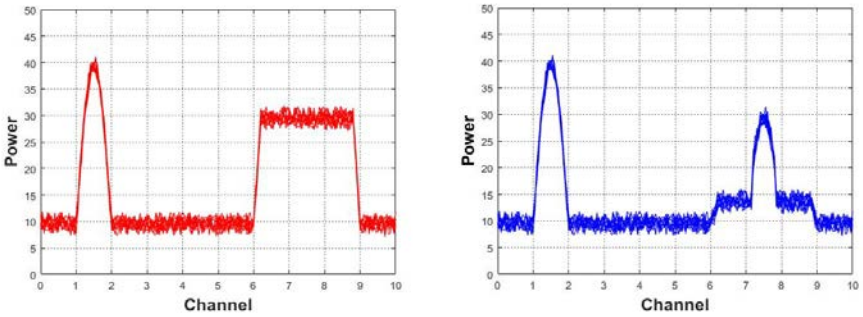
Fig. 4. Receiving spectrum of the jammer and the user based on ADRLA

For illustration and presentation, we first compare the receiving spectrum of the jammer and the user when the user adopts ADRLA or ADRLH, respectively. The comparison of receiving spectrum when the user adopts ADRLA is shown in **Fig. 4**, where (a) is the receiving spectrum of the jammer and (b) is that of the user. The result shows that when the user adopts ADRLA, the communication channel is selected by only considering communication quality. Although the user's actions can avoid jamming, the user's information may be obtained by the jammer.

The comparison of receiving spectrum when the user adopts ADRLH is shown in **Fig. 5**. Comparing (a) and (b), it is clear that the user prefers the frequencies at which the jammer itself is being interfered. Although the quality of communication may be slightly affected, it avoids the user's information leakage.

Secondly, the sensing probability of jammer with the iterations is shown in **Fig. 6**. When the user adopts the AFH method, the probability of being sensed by the jammer is higher than that of the other three methods. At the beginning of the iteration, there is no apparent difference between the probability being sensed by the jammer when the user adopts ADRLH, ADRLA, or FHSS. As the number of iterations increases, the sensing probability of jammer decreases if the user adopts ADRLH but increases when adopting ADRLA. Compared with the other three methods, the proposed algorithm reduces the sensing probability of jammer by 82.7%, 59.2%, and 75.8%, respectively. **Fig. 7** depicts the relationship between the action correlation coefficient and the number of iterations. Comparing **Fig. 6** with **Fig. 7**, it can be seen that no

matter which method the user adopts, as the number of iterations increases, the changing trends of the correlation between the actions of the user and the jammer and the sensing probability of jammer are consistent. In summary, the action correlation measurement method can verify whether the user's actions avoid being sensed by the jammer.



(a) receiving spectrum of the jammer (b) receiving spectrum of the user

Fig. 5. Receiving spectrum of the jammer and the user based on ADRLH

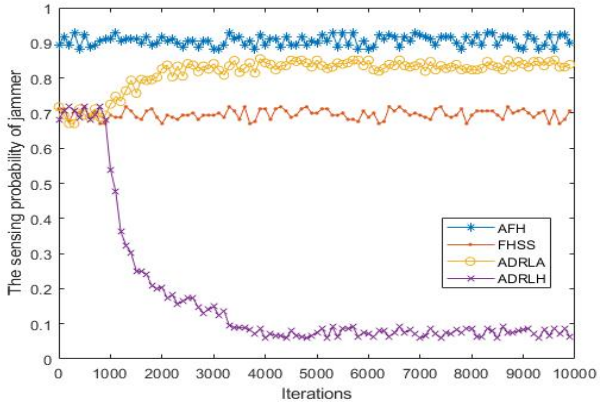


Fig. 6. The sensing probability of jammer

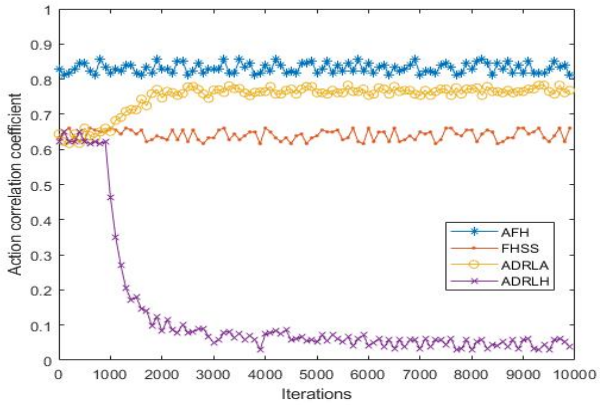


Fig.7. Relationship between the action correlation coefficient and iterations

Thirdly, the normalized throughput of the system with the iterations is shown in **Fig. 8** At the beginning of the learning procedure, the normalized throughput performance of ADRLH, FHSS, and ADRLA are close to each other and are superior to AFH. As the iteration progresses, the normalized throughput of ADRLH and ADRLA gradually increases while AFH remains unchanged, and FHSS gradually decreases. When steady converging, compared with other methods, the normalized throughput of ADRLH increases by 0.124, 0.332, and 0.687, respectively.

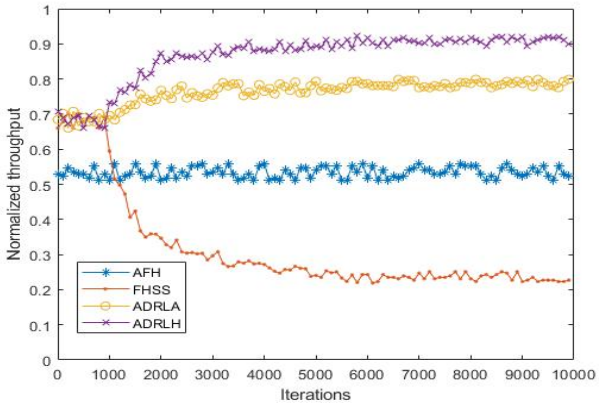


Fig. 8. Normalized throughput of the system with iterations

Table 1. Probability of jamming

Jamming methods / Anti-Jamming methods	Follower jamming	Q-Learning
AFH	45.7648%	25.6359%
FHSS	8.6523%	76.8216%
ADRLA	1.0867%	23.4368%
ADRLH	0.9489%	8.7475%

Finally, **Table 1** shows the probability of the user being jammed when different anti-jamming methods deal with follower jamming and intelligent jamming based on reinforcement learning. It can be seen from *Table* 1 that the performance of the hidden anti-jamming method proposed in this paper is better than that of AFH and FHSS when dealing with follower jamming or intelligent jamming. ADRLH reduced the probability of being jammed by 14.7% compared to ADRLA when fighting against intelligent jamming.

5. Conclusion

This paper proposes a hidden anti-jamming method based on reducing the sensing probability of the jammer. A deep reinforcement learning framework is designed, and an action correlation measurement method is designed to measure the action correlation between the user and the jammer. By analyzing the correlation between the actions of the user and the jammer, the jammer's sensing probability is indirectly obtained, which can be used as reference information for hiding frequency decision-making. Combined with the deep reinforcement learning method, Anti-jamming Deep Reinforcement Learning Algorithm

based on the hiding strategy(ADRLH) is proposed, aiming to avoid the information leakage of the user under the premise of ensuring communication quality. Simulation results show that ADRLH improves the anti-jamming performance in avoiding jamming and decreases the probability of being sensed by the jammer, compared with traditional anti-jamming methods.

References

[1] A. Kavianpour and M. C. Anderson, "An Overview of Wireless Network Security," in *Proc. of 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*, pp. 306-309, June26-28, 2017. Article (CrossRef Link)

[2] J. Li, Z. Feng, Z. Feng and P. Zhang, "A survey of security issues in Cognitive Radio Networks," *China Communications*, vol. 12, no. 3, pp. 132-150, Mar. 2015. Article (CrossRef Link)

[3] M. Bkassiny, Y. Li and S. K. Jayaweera, "A Survey on Machine-Learning Techniques in Cognitive Radios," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1136-1159, Third Quarter 2013. Article (CrossRef Link)

[4] X. Wang et al., "Dynamic Spectrum Anti-Jamming Communications: Challenges and Opportunities," *IEEE Communications Magazine*, vol. 58, no. 2, pp. 79-85, February 2020. Article (CrossRef Link)

[5] J. Wang, G. Ding, Q. Wu, et al., “Spatial-temporal spectrum hole discovery: a hybrid spectrum sensing and geolocation database framework,” *Chinese ence Bulletin*, 2014(16), 1896-1902, 2014. Article (CrossRef Link)

[6] G. Ding, J. Wang, Q. Wu, Y. Yao, F. Song and T. A. Tsiftsis, "Cellular-Base-Station-Assisted Device-to-Device Communications in TV White Space," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 107-121, Jan. 2016. Article (CrossRef Link)

[7] J. Xu and N. Wang, "Optimization of ROV Control Based on Genetic Algorithm," in *Proc. of 2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, pp. 1-4, May28-31, 2018. Article (CrossRef Link)

[8] Eryong Yang, Jianzhong Chen and Yingtao Niu, "Anti-jamming communication action engine based on Particle Swarm Optimization," in *Proc. of 2011 Second International Conference on Mechanic Automation and Control Engineering*, Hohhot, pp. 3913-3916, July15-17, 2011. Article (CrossRef Link)

[9] Xianyang Hui, Yingtao Niu and Mi Yang, "Decision method of anti-jamming communication based on binary artificial bee colony algorithm," in *Proc. of 2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, pp. 987-991, Dec.19-20, 2015. Article (CrossRef Link)

[10] B. Wang, Y. Wu, K. J. R. Liu and T. C. Clancy, "An anti-jamming stochastic game for cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 877-889, April 2011. Article (CrossRef Link)

[11] L. Jia, Y. Xu, Y. Sun, S. Feng, L. Yu and A. Anpalagan, "A Multi-Domain Anti-Jamming Defense Scheme in Heterogeneous Wireless Networks," *IEEE Access*, vol. 6, pp. 40177-40188, June27, 2018. Article (CrossRef Link)

[12] Y. Xu, J. Wang, Q. Wu, J. Zheng, L. Shen and A. Anpalagan, "Dynamic Spectrum Access in Time-Varying Environment: Distributed Learning Beyond Expectation Optimization," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5305-5318, Dec. 2017. Article (CrossRef Link)

[13] Y. Xu, J. Wang, Q. Wu, A. Anpalagan and Y. Yao, "Opportunistic Spectrum Access in Unknown Dynamic Environment: A Game-Theoretic Stochastic Learning Solution," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1380-1391, April 2012. Article (CrossRef Link)

[14] L. Jia, Y. Xu, Y. Sun, S. Feng and A. Anpalagan, "Stackelberg Game Approaches for Anti-Jamming Defence in Wireless Networks," *IEEE Wireless Communications*, vol. 25, no. 6, pp. 120-128, December 2018. Article (CrossRef Link)

[15] J. Parras, J. del Val, S. Zazo, J. Zazo and S. V. Macua, "A new approach for solving anti-jamming games in stochastic scenarios as pursuit-evasion games," in *Proc. of 2016 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 1-5, June26-29, 2016. Article (CrossRef Link)
[16] Z. Feng et al., "Power Control in Relay-Assisted Anti-Jamming Systems: A Bayesian Three-Layer Stackelberg Game Approach," *IEEE Access*, vol. 7, pp. 14623-14636, January18, 2019. Article (CrossRef Link)
[17] Y. Wang, Y. Niu, J. Chen, F. Fang and C. Han, "Q-Learning Based Adaptive Frequency Hopping Strategy Under Probabilistic Jamming," in *Proc. of 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1-7, Oct.23-25, 2019. Article (CrossRef Link)
[18] X. Pei et al., "Joint Time-frequency Anti-jamming Communications: A Reinforcement Learning Approach," in *Proc. of 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1-6, Oct.23-25, 2019. Article (CrossRef Link)
[19] Z. Zhang, Q. Wu, B. Zhang and J. Peng, "Intelligent Anti-Jamming Relay Communication System Based on Reinforcement Learning," in *Proc. of 2019 2nd International Conference on Communication Engineering and Technology (ICCET)*, pp. 52-56, June03, 2019. Article (CrossRef Link)
[20] S. Liu et al., "Pattern-Aware Intelligent Anti-Jamming Communication: A Sequential Deep Reinforcement Learning Approach," *IEEE Access*, vol. 7, pp. 169204-169216, November20, 2019. Article (CrossRef Link)
[21] G. Han, L. Xiao and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," in *Proc. of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2087-2091, June19, 2017. Article (CrossRef Link)
[22] X. Liu, Y. Xu, Y. Cheng, Y. Li, L. Zhao and X. Zhang, "A heterogeneous information fusion deep reinforcement learning for intelligent frequency selection of HF communication," *China Communications*, vol. 15, no. 9, pp. 73-84, Sept. 2018. Article (CrossRef Link)
[23] X. Liu, Y. Xu, L. Jia, Q. Wu and A. Anpalagan, "Anti-Jamming Communications Using Spectrum Waterfall: A Deep Reinforcement Learning Approach," *IEEE Communications Letters*, vol. 22, no. 5, pp. 998-1001, May 2018. Article (CrossRef Link)
[24] W. Li, J. Wang, L. Li, G. Zhang, Z. Dang and S. Li, "Intelligent Anti-Jamming Communication with Continuous Action Decision for Ultra-Dense Network," in *Proc. of ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pp. 1-7, July15, 2019. Article (CrossRef Link)
[25] Y. Li, Y. Xu, X. Wang, W. Li and W. Bai, "Power and Frequency Selection optimization in Anti-Jamming Communication: A Deep Reinforcement Learning Approach," in *Proc. of 2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, pp. 815-820, Dec.6-9, 2019. Article (CrossRef Link)

Yifan Wang received his B.S. degree in communication Engineering and M.S. degree in computer Science and Technology from Guilin University of Technology in 2018 and 2021 respectively. His research interests focus on anti-jamming communication and deep reinforcement learning.

Xin Liu received his B.S. degree in Communications Engineering, M.S. degree in Communications and Information Systems, and Ph.D. degree in Communications and Information Systems from College of Communications Engineering, PLA University of Science and Technology, in 2004, 2008 and 2011 respectively. He has been with College of Information Science and Engineering, Guilin University of Technology since 2017, and currently as an Associate Professor. His research interests focus on anti-jamming communication, deep reinforcement learning, game theory, and soft defined radio. He has published several papers in international conferences and reputed journals in his research area.

Mei Wang, female, professor and doctoral supervisor at Guilin University of Technology. Her research interests include position sensing and collaborative positioning, sensor networks, and energy efficiency optimization.

Yu Yu received her B.S. degree in Communications Engineering, M.S. degree in Communications and Information Systems from the College of Communications Engineering, PLA University of Science and Technology, in 2005, 2008 respectively. Her current research interests include wire communication and wireless communication.