IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 23, 2019, accepted November 6, 2019, date of publication November 20, 2019, date of current version December 5, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954531

Pattern-Aware Intelligent Anti-Jamming Communication: A Sequential Deep Reinforcement Learning Approach

SONGYI LIU, YIFAN XU, XUEQIANG CHEN, XIMING WANG, MENG WANG, WEN LI, YANGYANG LI, AND YUHUA XU

College of Communications Engineering, Army Engineering University of PLA, Nanjing 210000, China

Corresponding author: Yifan Xu (yifanxu1995@163.com)

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

ABSTRACT This paper investigates the problem of anti-jamming communication in dynamic and intelligent jamming environment. A sequential deep reinforcement learning algorithm (SDRLA) without prior information is proposed, and raw spectrum information is used as the input of SDRLA. The proposed SDRLA algorithm mainly contains two parts: Firstly, deep learning and sliding window principle are introduced to identify jamming patterns; Secondly, reinforcement learning is carried out to make on-line channel selection based on recognized jamming patterns. In addition, channel switching cost is introduced for the purpose of formulating the trade-off relationship between throughput and overhead. Taking advantage of both deep learning and reinforcement learning, this method can realize rapid and effective anti-jamming channel selection with no need for modeling the jammer's characteristics. Simulation results show the convergence and effectiveness of the proposed SDRLA algorithm. Compared with single-mode reinforcement learning, our approach can reach better convergence performance and higher utility.

INDEX TERMS Pattern-aware, reinforcement learning (RL), deep learning (DL), anti-jamming communication, channel switching cost.

I. INTRODUCTION

Anti-jamming is an eternal topic in communications, especially for wireless communications. External jamming environment puts great threat on communication link quality. In recent years, with the development of artificial intelligence, combating intelligent jamming attacks is extremely challenging and interesting [1]–[3]. Facing complex external jamming environment and changeable jamming patterns, users can hardly deal with these new threats. In addition, due to the rapid change of jamming attacks, it is difficult for users to adopt anti-jamming strategy in real time.

Since jamming attacks pose serious threat to the security of wireless communication, anti-jamming researches have attracted more and more attention in recent years. However, most existing anti-jamming works need to obtain some prior information of the jammer before formulating relevant anti-jamming strategies [4]–[7]. In addition, some researches adopted machine learning algorithms to make online anti-jamming decisions [2], [3], [8]. However, those proposed algorithms can only be applied in high regularity jamming environment and converged slowly in the early stage. If the external jamming mode changed, these algorithms needed to relearn until they achieve convergence. Therefore, these algorithms have strong limitation when applied in practical scenarios.

To solve these problems and challenges mentioned above, this paper propose a pattern-aware intelligent anti-jamming approach to realize channel access in complex and changeable jamming environment. Moreover, motivated by [9], we define the raw spectrum information as the environment states to avoid the loss of jammer information. A sequential deep reinforcement learning algorithm (SDRLA) without prior information is proposed. Firstly, the raw historical spectrum information is stored and analyzed. Secondly, according to the characteristics of different jamming types, the jamming modes are classified using a convolutional neural network. Finally, on the basis of recognized jamming pattern,

The associate editor coordinating the review of this manuscript and approving it for publication was Zheng Yan.

a reinforcement learning (RL) algorithm is designed for achieving real-time anti-jamming channel access. The key contributions of the paper are as follows:

- A sequential deep reinforcement learning algorithm (SDRLA) without prior information is proposed. The proposed algorithm does not need to estimate and make assumptions about the pattern and utility function of the jammer, which has strong environmental adaptability and wide application range.
- A sliding window processing mechanism is designed to solve the transition problem of multiple and changeable intelligent jamming modes.
- The concept of channel switching cost is introduced, and the communication overhead is optimized by reducing the frequency of channel switching.

The rest of this paper is organized as follows. We review the related work in Section II. The system model and problem formulation are presented in Section III. Moreover, the sequential deep reinforcement learning approach is presented in Section IV and Section V. In detail, the multi-jamming pattern awareness algorithm is presented in Section IV, and the anti-jamming reinforcement learning algorithm based on recognized jamming mode is proposed in Section V. Channel switching cost is also introduced in Section V. In Section VI, simulation results and performance analysis are presented. In the end, conclusion is presented in Section VII.

II. RELATED WORK

Various studies with respect to the anti-jamming techniques have been proposed. However, traditional anti-jamming techniques like frequency hopping spread spectrum (FHSS) and Direct sequence spread spectrum (DSSS) have some drawbacks in dealing with new intelligent jamming attacks. For example, frequency hopping spread spectrum (FHSS) relied heavily on a predefined secret frequency hopping sequence [10]. Direct sequence spread spectrum (DSSS) relied on a local pseudo-random code [11]. Considering the interactions between users and malicious jammer, game theory is suitable for analyzing the communication strategy under jamming, and has been widely used in anti-jamming field, spectrum resource allocation and dynamic spectrum access [4]–[7], [12]–[19]. The anti-jamming problem under incomplete information was investigated in [20] and [21]. However, in those papers, model and prior information of jammer were preconditions for designing the anti-jamming algorithms [22], [23]. Thus, these anti-jamming methods have some limitations in practical application scenarios.

In general, collecting and analyzing of raw spectral information was an important way to obtain jamming information [3]. Therefore, to obtain malicious jammer's strategies, feature extraction was carried out in paper [1], [8], [24] to simply distinguish between interference and users. Moreover, in paper [25], feature extraction was used to identify different kinds of jamming patterns. However, when applied to anti-jamming problems, feature extraction has two disadvantages: i). Using preprocessed data for feature extraction may cause some loss of important information. ii). When the jammer can switch the jamming patterns fast enough, it is impossible for legitimate users to track and obtain the jamming information in real time. Hence, It is very important to design an anti-jamming method which can adapt to the dynamic environment.

Reinforcement learning is an effective method to solve the decision-making problems in dynamic jamming environment. As a kind of widely used reinforcement learning algorithm, Q-learning [26], [27] has been widely used in solving anti-jamming problems [4], [28], [29]. However, when facing complex jamming environment, Q-learning (QL) can not process the raw environment information directly. Thus, paper [8] introduced deep reinforcement learning into anti-jamming decision-making. By learning the real-time changes of the external jamming, anti-jamming decisions were made [30] rapidly. Moreover, [31] proposed a deep reinforcement learning based frequency-space anti-jamming mobile communication scheme, which can realize the optimal power distribution and mobile strategy of mobile devices. [32] proposed a deep reinforcement learning method to extract jamming features from raw time-frequency data, which provided a new idea for us to solve anti-jamming problem.

Different from most existing works, in this paper, the idea of combining deep learning with reinforcement learning is proposed to solve the communication anti-jamming channel access problem. The principle of sliding window and channel switching factor are introduced to improve the performance of the algorithm and reduce the system overhead.

III. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

As shown in Fig. 1, we consider a wireless communication scenario which includs one or several jammers and one user (transmitter-receiver pair). An agent which is installed at the receiving end can make real-time anti-jamming strategies, and then send strategies to the transmitter through reliable control link. Moreover, wide-band spectrum sensing can be conducted at the receiving end [8]. In the communication process, jammer can switch jamming patterns by changing the jamming period randomly. As different kinds of jamming patterns can be distinguished by jamming characteristics, we can identify those jamming patterns through observing local raw information and designing rational learning algorithm.

Under malicious jamming environment, it is impossible for a jammer to notify its jamming strategy to legitimate users. Hence, analyzing and effectively utilizing the stored raw jamming information is necessary for users to obtain jamming strategy. Moreover, the receiver can collect raw spectrum information by sensing the channel state.

The essence of analyzing the historical jamming information is feature extraction. That is to extract jamming behavior or jamming pattern from the observed environmental

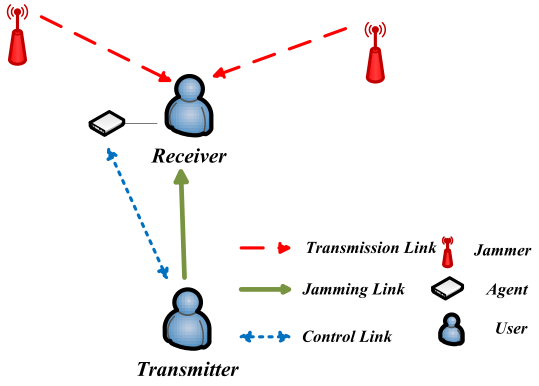


FIGURE 1. System model.

information. According to different jamming characteristics, the jamming pattern is classified by using labels. When the user receives the signal, the agent stores decision feedback through the perceptual spectrum, judges the current jamming type, and then makes anti-jamming channel decision under current jamming type.

The receiver continuously senses the spectrum of the entire communication band and stores the historical information of the spectrum in the background of the agent. We assume that the spectrum vector of communication band at time t is $\boldsymbol{P}_t = \left(p_{t,1}, p_{t,2}, \ldots, p_{t,i}, \ldots, p_{t,n}\right)$, where $p_{t,i} = 10\log\left[\int_{i\Delta f}^{(i+1)\Delta f} S\left(f + f_L\right) df\right]$ represents the spectral energy of frequency n at time t. Moreover, $S\left(f\right)$ is the power spectral density (PSD) function and Δf is the resolution of spectrum analysis. The time-frequency characteristic $\boldsymbol{S}_t$ is generated though the spectrum vector's correlation rules. The thermodynamic diagram of $\boldsymbol{S}_t$ is called spectral waterfall diagram [33] (or thermal chart), which represents the collected time and frequency domain information. Based on the collected historical spectrum vector, the thermal chart of the time-frequency characteristic matrix $\boldsymbol{S}_t$ can be expressed:

$$\boldsymbol{S}_t = \begin{bmatrix} \mathbf{P}_{t-1} \\ \mathbf{P}_{t-2} \\ \vdots \\ \mathbf{P}_{t-M} \end{bmatrix} = \begin{bmatrix} p_{t-1,1} & p_{t-1,2} & \cdots & p_{t-1,N} \\ p_{t-2,1} & p_{t-2,2} & \cdots & p_{t-2,N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{t-M,1} & p_{t-M,2} & \cdots & p_{t-M,N} \end{bmatrix}, \quad (1)$$

where $\boldsymbol{S}_t$ contains all historical spectral information before time t. As M approaches infinity, the state value becomes extremely large, which makes the optimization problem more difficult. In consequence, M needs an appropriate value to solve relevant problems, and it specific value needs to be determined according to the time-varying characteristics of the interfering environment. The thermodynamic diagrams of $\boldsymbol{S}_t$ matrix under several common jamming modes are given to illustrate the rationality of using $\boldsymbol{S}_t$ as the basis in anti-jamming decision-making. As shown in Fig. 2 and Fig. 3, we can exhibit the frequency range and intensity (color) of signal accurately and intuitively.

Different from the paper [2], $\boldsymbol{S}_t$ is defined as the environmental state. To reduce the complexity of its state set, we need to pretreat the $\boldsymbol{S}_t$. Firstly, channelization is performed on the range of transmitting frequency and received frequencies used by a user. Hence, the model can be closer to the actual communication situation and the state set is reduced. Secondly, pattern awareness in this paper is based on the change of external jamming environment to identify different kinds of jamming patterns, and the identified jamming pattern are tagged. Then, the reinforcement learning is carried out under different jamming modes. Above methods can reduce the complexity of the behavior decision set, and realize the fast convergence and real-time decision of the on-line learning algorithm. Specific pattern awareness methods and details of implementing reinforcement learning are presented in the next section.

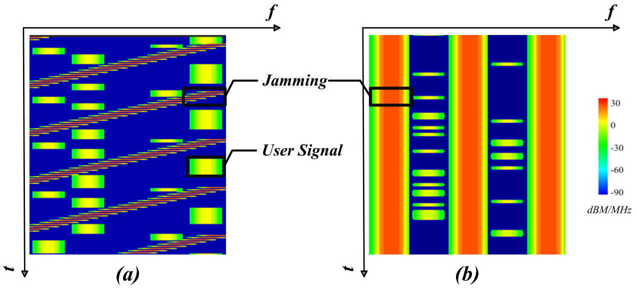


FIGURE 2. Thermodynamic chart of jamming pattern.

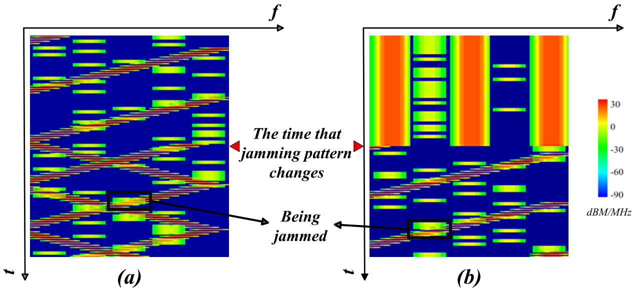


FIGURE 3. Graphical state under dynamic jamming.

Due to the actual scenario and the model set in this paper has certain differences, we make the following assumptions:

- The jamming mode can be switched randomly or periodically, and the historical information collected by the user must include all jamming patterns.
- We assume that there is a reliable link for real-time anti-jamming channel access strategy transmission.

B. PROBLEM FORMULATION

Assuming that the central frequency of communication is f_t, the jamming optional central frequency is f_j, the user's communication transmission bandwidth is b, the power spectral density (PSD) function of white Gaussian noise is $n(f)$, the PSD of the jamming signal is J_t, the transmission power p, the channel gain of transmission link is g_t, and the channel gain of the jamming link is g_j. Inspired by the paper [8], the signal-to-interference-plus-noise ratio (SINR) at the receiving end can be expressed as:

$$\eta\left(f_t, f_j\right) = \frac{g_t p}{\int_{f_t-b/2}^{f_t+b/2}\left\{n(f)+\sum_{j=1}^{J} g_j J_t\left(f-f_j\right)\right\} df}. \quad (2)$$

Considering the coexistence of user signals and jamming, the power spectral density (PSD) function of the receiving end is expressed as:

$$S_t(f) = g_u U\left(f-f_t\right)+\sum_{i=1}^{J} g_j J_t\left(f-f_j\right)+n(f). \quad (3)$$

Denote $\mu\left(f_t, f_j\right)$ as the indicator function for successful transmission, which can be expressed as follows [3]:

$$\mu\left(f_t, f_j\right) = \begin{cases} r_m, & \beta\left(f_t, f_j\right) \geq \beta_{th}, \\ 0, & \beta\left(f_t, f_j\right) < \beta_{th}, \end{cases} \quad (4)$$

where β_{th} is defined as the threshold of SINR and r_m is the feedback after a successful transmission ($r_m \geq 1$). When SINR $\beta\left(f_t\right) < \beta_{th}$, the transmission is seen as failed. The jamming frequency range is denoted by $\mathrm{B_j}$, and the user available frequency range is denoted by $\mathrm{B_u}$. To simplify the problem, we set $\mathrm{B_j} = \mathrm{B_u}$. Denote $\mathrm{u_b}$ as user's transmission band, then the number of user's strategy set can be calculated as $n = \frac{B_u}{b_u}$. The action set of user can be denoted by $A = \{a_1, a_2, a_3, \ldots a_n\}$. $a(t) \in A$ represents the channel selection of the user at time t. Moreover, the cost formula of user switching channel is expressed as:

$$W\left(f_t, f_j\right) = \begin{cases} 0, & a(t) = a(t-1), \\ c, & a(t) \neq a(t-1), \end{cases} \quad (5)$$

where c represents the channel switching factor. To sum up, the optimization objective of the user is:

$$\max_{f_t' \in A} U = \sum_{t=0}^{\infty} \gamma^t\left(\mu\left(f_t, f_j\right)-W\left(f_t, f_j\right)\right), \quad (6)$$

where γ is the discount factor and $\gamma \in (0, 1)$.

The goal is to select anti-jamming decisions in a fashion that maximizes the cumulative future reward. In this paper, the switching communication frequency is the key to realize anti-jamming communication. Hence, the switching cost is considered in the optimization objective.

IV. MULTI-JAMMING PATTERN AWARENESS ALGORITHM

The multi-jamming pattern awareness algorithm is presented in this section. Firstly, the basic principle of pattern awareness is introduced. Secondly, the pattern awareness algorithm adopted in this paper is demonstrated. Finally, the mechanism of sliding window is introduced.

A. BASIC PRINCIPLE

Pattern recognition is the foundation of artificial intelligence [30]. With the development of computer and artificial intelligence technology, pattern recognition applications becomes ever more extensive [35]. Pattern recognition can identify the external jamming patterns by processing collected information and classifying things with the same characteristics through classifiers according to certain rules. It is mainly used in many fields such as speech recognition and image processing. Pattern recognition is divided into supervised learning classification and unsupervised learning classification. The main difference between the two is whether the categories are already known. For supervised learning classification, feature extraction is carried out through a large number of sampled data, and previously extracted features are taken as labels for the classification of other new samples. The recognition effect of supervised learning classification is very good, but it has some limitations in solving practical problems if new samples do not belong any category that has been labeled [36]. Unsupervised classification is a direct classification method using the raw data, and does not need labels as the basis for classification. Considering the strong pattern recognition ability of Convolutional neural network (CNN), we adopt Convolutional neural network as the basis of multi-jamming pattern awareness algorithm. CNN is generally composed of convolution layer, pooling layer and full connection layer, and has achieved good results in machine learning applications. The following three levels of functions are introduced respectively:

- Convolutional Layer: The convolutional layer is mainly used for local feature extraction. Each layer has following parameters: map size, kernel size, and the number of maps. Moreover, a kernel can shift over the region of the input picture [35].
- Pooling Layer: The pooling layer plays two main roles. The first is to extract the main features. The second is to compress the feature map to simplify the network computational complexity.
- Classification Layer: The classification output layer allocates one neuron to every label in the classification task, and the classification layer is situated at the end of the CNN.

B. PATTERN AWARENESS ALGORITHM DESCRIPTION

In this section, we propose a deep learning algorithm for multi-jamming pattern awareness. This paper conducts jamming feature extraction and jamming classification based on the acquired external history information, so the supervised learning method is adopted. The user obtains the raw spectral information in real-time and generates the thermal chart. Due to the dynamic nature and uncertainty of the external environment, the historical information of the processing is large, and the number of samples of the spatial state set is large. Therefore, it is difficult to perform artificial feature extraction.

Convolution neural network can effectively solve the feature extraction and data classification for a large number of historical data sample set. Therefore, we consider putting historical data into convolutional neural networks for correlation processing. Due to the continuous historical data information is collected, the large number of image

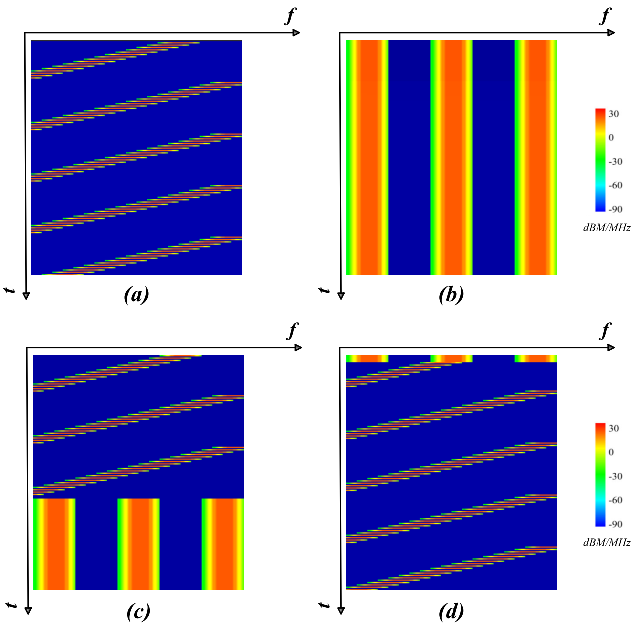


FIGURE 4. Special frame state comparison.

(thermodynamic chart) data are similar to video frame shots. The adjacent data sets collected by these data sets have a high degree of similarity. Later, we take the historical information in sweeping jamming mode as an example.

When switching between different jamming modes, there will be a special case of "jamming transition band". As is shown in Fig. 4 (a), we can judge that the current moment is sweep jamming according to the characteristics of thermal chart. Moreover, the comb jamming is shown in Fig. 4 (b). When the jamming transition zone is in critical condition as shown in Fig. 4 (c) and (d), we define the jamming pattern in Fig. 4 (c) as sweep jamming transition zone and Fig. 4 (d) as the comb jamming transition zone.

For convolutional neural networks, if the jamming modes are simply classified according to different labels, the transition band will be divided into corresponding jamming modes. If the above classification method is adopted for neural network fitting, over-fitting will occur and the accuracy of the fitted network decision will be greatly reduced. Therefore, in order to achieve the goal of improving classification accuracy, the transition bands in each mode are re-divided and defined as a new type of jamming mode in the convolutional neural network. If there are N types of current jamming patterns. The type of "jamming transition band" generated is $N \times (N-1)$ types. In conclusion, it is necessary to perform N^2 classification labels in the convolutional neural network. The flow chart of the multi-jamming pattern awareness algorithm (MPA) is shown in Algorithm 1.

For the purpose of implementing the MPA algorithm, a convolutional neural network (CNN) that can classify multi-mode jamming states needs to be built. According to the relevant design principles and processing requirements, a total of six hidden layers are designed. Fig. 5 shows the constructed convolutional neural network structure. The first three layers are composed of convolution layer and pooling layer. The last three layers are fully connected layers. The main function is equivalent to a "classifier" for label classification output.

Algorithm 1 Multi-Jamming Pattern Awareness Algorithm (MPA)

Initialize:

Preprocess historical information, set up relevant learning parameters, learning times and loss functions, and build the designed neural network.

Step 1) Classify the obtained historical jamming channel information according to the jamming characteristics, and classify the required I jamming modes;

Step 2) Tagged historical information is divided into training set, verification set and test set according to requirements;

Step 3) Input the training set data into the designed neural network for training;

Step 4) The validation set is used to determine whether the over-fitting phenomenon occurs in the training process, and the neural network with good test effect is obtained by adjusting the relevant parameters of the network;

Step 5) Conduct data test set test, verify and evaluate the final generalization model of the model;

Step 6) Save the final model and relevant parameters and directly introduce the results into the anti-jamming algorithm under real-time jamming identification;

End the algorithm

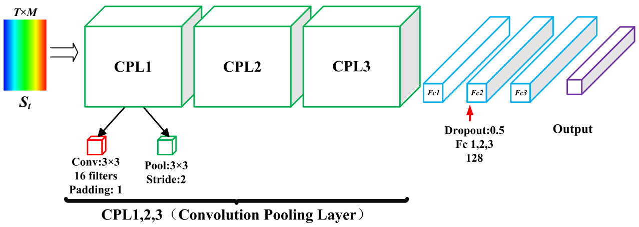


FIGURE 5. The network structure of CNN.

The convolution neural network described in this paper needs to calculate the gradient and update the weight. In the deep learning fitting process, loss function as shown in equation (7) is used for the *ith* iteration,

$$L_i(\theta_i) = E_e\left[(y_i - S_t)^2\right], \tag{7}$$

where θ_i represents the parameters of the deep learning convolutional neural network (CNN) in the *ith* iteration. According to the gradient descent method, the loss function is differentiated. The gradient of loss function is obtained, as shown in equation (8):

$$\nabla_{e_i} L_i(\theta_i) = E_e\left[(y_i - S_t)\nabla_{e_i} S_t\right], \tag{8}$$

where $L_i(\theta_i)$ represents the loss function, and ∇_{θ_i} represents the gradient operation.

The neural network designed and constructed in this paper is composed of three convolutional layers and three full connected layers. First layer of the network convolves input with

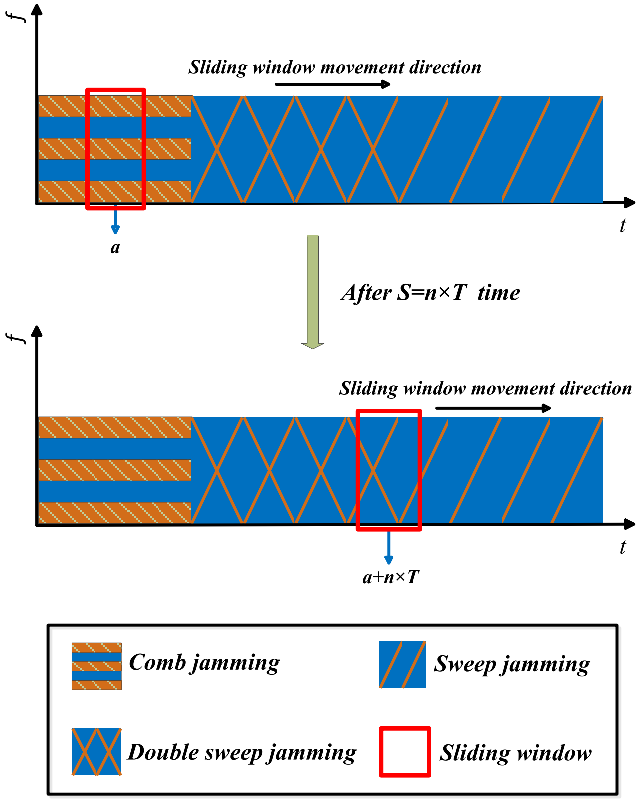


FIGURE 6. Schematic diagram of sliding window algorithm.

16 filters of size 3×3 and stride 4. The second layer convolves output of the first layer with 32 filters of size 2×2 and stride 2. The third layer convolves output of the second layer with 64 filters of size 1 × 1 and stride 1. Then, a full-connected layer with 128 units is followed, and the last full-connected layer outputs the estimated long-term cumulative rewards for each action. In addition, activation function is used in each layer (tanh function is used as an activation function in this paper). At the same time, in order to improve the fitting accuracy of the training neural network and speed up the fitting speed, the gradient descent function of deep learning network is optimized.

C. DESIGN MECHANISM OF SLIDING WINDOW

The jamming types can be judged in real-time through pattern recognition. In order to control the input of pattern information, the concept of "sliding Windows" needs to be introduced. As shown in Fig. 6, the size of sliding window is set, and the distance each frame of the sliding window is defined. The sliding window forward sliding time of each frame is set to T. The moment of the current sliding window in the figure is a (the moment of the middle position in current sliding window). After $S = nT$ time, the sliding window moves to the moment $a + nT$ to intercept jamming information of the same amount of information. The width of the sliding window size and the moving speed of each frame depend on the processing speed of real-time information of the convolutional neural network. Different from traditional sliding window protocol, the proposed sliding window algorithm neither requires the transmitter to acknowledge ACK, nor needs to change window size in real-time. According to the mathematical model in this paper, the real-time information obtained by sliding window generates the relevant jamming diagram (thermal chart). Then, we put it into the trained convolutional network for pattern awareness and judge the current jamming type. Reinforcement learning is carried out based on current jamming types and relevant channel information.

V. MULTI-PATTERN REINFORCEMENT LEARNING BASED ON RECOGNIZED JAMMING MODES

This section mainly gives the multi-pattern reinforcement learning algorithm based on the identification of the current jamming mode. Firstly, the basic principle of reinforcement learning is introduced. Secondly, the detail of proposed algorithm is given. Finally, for the purpose of reducing the network cost, channel switching factor is introduced.

A. BASIC PRINCIPLE

As mentioned above, due to the different characteristics of each jamming mode, it is the best method adopt different anti-jamming strategies for specific jamming modes. For anti-jamming strategies, we mainly adopt the reinforcement learning on-line real-time judgment method. Reinforcement learning, also known as motivational learning, generally consisted of the following elements to form a general framework for reinforcement learning [38]. The three most basic elements are state space (S), action space (A), and immediate reward ($R = r(s, a), (s \in S, a \in A)$). In order to make these three elements intrinsically linked to promote the intensive learning operation, it is also necessary to introduce the transition probability space (P), the state-action value function (S-A value function: $q_\pi(s, a)$)and the adopted policy set ($\pi(s) \in A$). For reinforcement learning, the external environment is affected by the current state and the actions made, it is generally modeled as a Markov decision process. By introducing the immediate reward function of status-action, the Q-learning algorithm is driven based on the action policy of the maximization of long-term reward.

Q-learning process generally obtains information through the external environment and defines it as the current state. The action can be selected randomly in the current action space, or selected based on the relevant data in the current policy set. After the action is made, the state will be changed, and feedback is received when the state changes. According to the feedback, the calculation of S-A value function is given to update the set of strategies we need. The agent aims to find the optimal policy $\pi^*(s)$ to maximization of long-term reward $q_\pi(s, a)$ for each environmental state s. The state transition probability ($p_{ss'}$) is the probability of moving from the current state to the next after the action is made. The formula (9) given by the S-A value function shows that reinforcement learning usually adopts a strategy of maximizing long-term return value. Since long-term reward is related to both current and future reward, we can express long-term reward in a recursive form as:

$$q_{\pi^*}(s,a) = r(s,a) + \eta \sum_{s\in S} p_{ss'}(a) \max_{a\in A} q_{\pi^*}\left(s',a'\right) \tag{9}$$

where $\eta \in [0,1]$ is the effect of the subsequent action reward on the current reward, called the discount factor. $\left(s',a'\right)$ is the next state-action pair after the QL algorithm executes the action a at the state s. Then, the optimal policy is

$$\pi^*(s) = \arg\max_{a\in A} \left[q_{\pi^*}(s,a)\right], \forall s \in S \tag{10}$$

The classical Q-learning algorithm in reinforcement learning is adopted in this paper. Since the Q-learning algorithm has its own Q value table, it is equivalent to the above policy set as the basis for selecting strategies. The Q value table size is $|S| \times |A|$, which stores the cumulative reward (long-term reward) by executing the action a in the environment (state s). Moreover, the update principle of the Q value table is as follows:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left[r(s,a) + \eta \max_{a\in A} Q\left(s',a'\right)\right], \tag{11}$$

where a is the learning rate. $Q\left(s',a'\right)$ is the next state operation a' and the next environmental state s', after Q-learning to perform the current operation. In order to achieve a balance between the exploration of best actions and the exploitation of experiences, Q-learning introduces the ϵ-greedy algorithm to select actions for each state. The ϵ-greedy algorithm means that the probability $1-\epsilon$ performs the action $a = \arg\max_{a\in A} Q(s,a)$, and the probability ϵ randomly selects the action set. Here ϵ is the trade-off factor for making an action. So we can optimize the ϵ to achieve the fastest speed convergence of the optimal decision.

B. MULTI-PATTERN REINFORCEMENT LEARNING ALGORITHM DESCRIPTION

Due to the influence of the dynamics and uncertainty of the environment, the state set is large and the corresponding behavior strategy table ($\boldsymbol{\pi}$) is complex. If only use the independent reinforcement learning (single mode reinforcement learning) [34], the corresponding behavioral strategy table cannot adapt to the dynamics of the environment. If a neural network is used to fit the behavior strategy table, the deep reinforcement learning requires higher regularity of jamming. It takes a long time for deep reinforcement learning to learn and extract the characteristics of jamming, which leads to a slower convergence. When the exterior jamming is random or switch the jamming mode quickly, the algorithm used is difficult or even impossible to converge. The reinforcement learning algorithm for anti-jamming is different from the single mode reinforcement learning algorithm. First we define the state space, the action space, and the immediate reward function:.

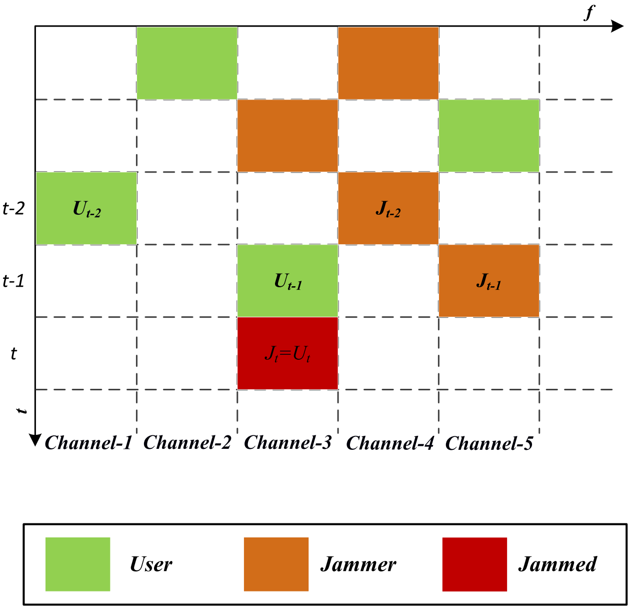


FIGURE 7. Time-frequency schematic diagram of user and jammer.

- State space: We define the current state of all channels as s. If the i channel has jamming at the current time, the channel is defined as 1. If the i channel has no jamming at the current time, the channel is defined as 0. Therefore, the channel state can be represented by a so-called binary number. Assuming that the current channel has a total of n, the state set size is 2^n. For example, assume that there are 5 available channels, when the second channel and the fourth channel are jammed, the state is expressed as [01010].
- Action space:Let A_n be the channel n policy selected by the user, then A denotes all the strategies for the user to select the channel, and its state space is $A = \{a_1, a_2, a_3, \ldots, a_n\}$
- Immediate reward function: We call it the instant reward function $r(s,a,t)$, which represents the feedbacks that performs action a in state s at time slot t. As shown in Fig. 7, the relation of state and action is demonstrated, and the vertical axis and the horizontal axis represent time and frequency respectively. For example, at time slot $t-2$, environment state is [00010], and the user's action is $a(t-2) = a_1$, hence, the user is not attacked by the jammer, and user switches from channel a_5 to a_1. Hence, the user can obtain reward $r(s,a,t-2) = r_m - C$.

In the early stage, jamming has been pattern-identified and classified, so the external jamming signal in the anti-jamming reinforcement learning (RL) are more obvious. The agent algorithm can select a channel with a better channel quality for communication based on the characteristics of the current jamming. In the frame structure shown Fig. 8, it can be known that the user can obtain the channel status (whether there is jamming or not) from the transmitter to the receiver and the current jamming type through channel perception at the beginning of each frame.

We assume that the user's transmitter and receiver have achieved the communication synchronization. In a single

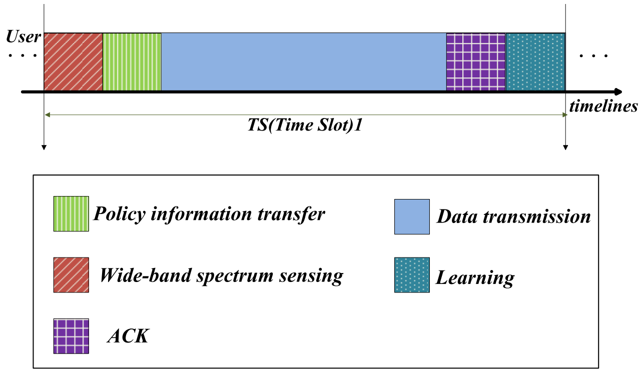


FIGURE 8. Q-learning time slot structure design.

communication time slot, the receiver performs wide-band spectrum sensing in the early phase of the time slot. According to the the current Q table information, the channel access strategy is selected. The channel access selection strategy is sent to the transmitter through the control link, and data transmission is carried out by transmitter. If the data transmission is success, the receiver sends an ACK to the transmitter. The receiver determines the immediate reward value of the current time slot according to the result of the transmission data, and then sends back ACK to the transmitter. The agent performs an iterative update of the Q value table according to the ACK sent by the control link of the receiver under the current time slot.

On the other hand, the optimal channel selection scheme is related to the optimal behavior selection rules of the state. Hence, the key point of this algorithm is updating Q-values. The algorithm flow is shown in the Fig. 9 and the proposed algorithm is shown in Algorithm 2.

Algorithm 2 Multi-Pattern Reinforcement Learning (MRL)

```
Initialize:
    Set the simulation start-end time, the relevant learning parameters and initialize the Q value table in the state of each jamming mode.
Loop:
    The user observes the current state of the environment to obtain real-time jamming information. The obtained jamming state thermal chart is input into the trained convolutional neural network. The current jamming pattern is $n$. Turn into Q-learning algorithm with current jamming type $n$.
    Q-learning algorithm for the current jamming type $n$ :
    Step 1) The user observes current state and selects the channel according to the following rules:
        1. Independent channel strategy $\pi_n^*(s)$ is selected randomly with probability $\epsilon$ ;
        2. The channel strategy with the maximal Q value of current state is selected with probability $1-\epsilon$;
    Step 2) The agent transfers the channel selection strategy to the transmitter. In the process of user communication, the immediate reward $r_n(s, a, t)$ for $n$th jamming pattern is obtained.
    Step 3) Update user's Q value table $Q_n$ for $n$th jamming pattern;
End the loop
```

C. ANTI-JAMMING MRL ALGORITHM INTRODUCING CHANNEL SWITCHING FACTOR

In previous subsection, the multi-pattern reinforcement learning algorithm for anti-jamming channel selection is introduced. In this subsection, considering that the rapid change of selected channel may cause heavy system overhead, we introduce the channel switching factor for the purpose of achieving maximal network utility [37].

In fact, there may be some channel switchers that do not change the channel when the current dynamic jamming environment changes. Therefore, there is no significant impact on long-term rewards. For example, at time $t-1$, the selected action is $a(t-1) = \arg\max_{a \in A} Q(s(t-1), a)$. In addition, at time t, the selected action is $a(t) = \arg\max_{a \in A} Q(s(t), a)$. If $a(t)$ is different from $a(t-1)$, channel switching occurs. However, $Q\left(s(t), a(t)\right)$ and $Q\left(s(t), a(t-1)\right)$ may have no effect on long-term rewards in channel switching. Therefore, based on the above principle, the purpose of reducing system overhead is achieved via avoiding frequent channel switching. Hence, it is necessary to re-plan and design the immediate reward function, and the incentive function is expressed as follows:

$$r(t) = \begin{cases} r_m, & \text{if } a(t) = a(t-1) \text{ and successful,} \\ r_m - c, & \text{if } a(t) \neq a(t-1) \text{ and successful,} \\ 0, & \text{if } a(t) = a(t-1) \text{ and failed,} \\ -c, & \text{if } a(t) \neq a(t-1) \text{ and failed.} \end{cases} \tag{12}$$

The agent achieves an ideal balance between system overhead and payback. By adjusting the switching cost factor, the desired balance between anti-jamming communication and system overhead can be achieved more flexibly. From the comparison between (a) and (b), (c) and (d) in Fig. 10, we can clearly observe that the number of channel switching is significantly reduced after introducing channel switching factor. Furthermore, the simulation results also verify the influence of channel switching factor.

VI. SIMULATION RESULTS AND ANALYSIS

A. SIMULATION SETTINGS

In the simulation, we assume that the number of available channels in the system is 5. The user and the jammer are confronted in 20MHz frequency band. It is assumed that the frame length of each frame is 5ms, and the transmission information time of each frame is 4ms. The total time of wide-band spectrum sensing (WBSS), ACK, intelligent reinforcement learning and anti-jamming Policy information transfer (PIT) is 1ms. The sliding window retains the spectrum data within 200ms. Hence, the thermal chart pixel of the collected jamming information is 200 × 200. Each slot allows

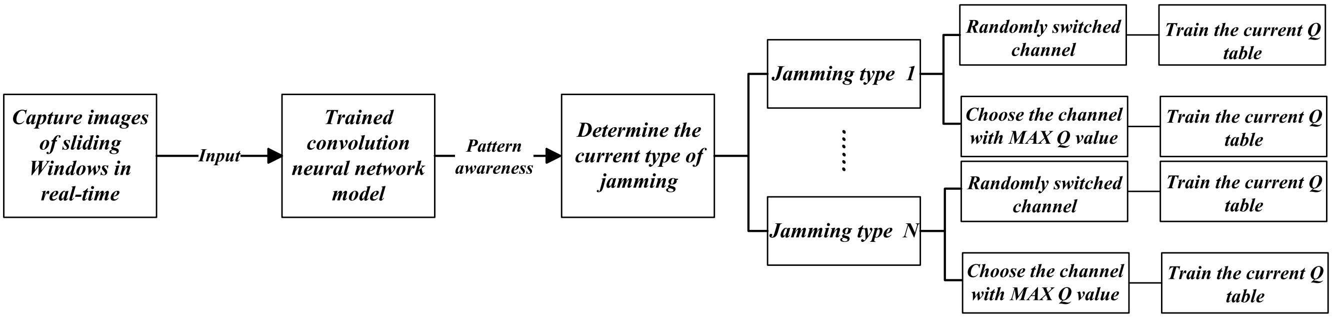

Capture images of sliding Windows in real-time
Input
Trained convolution neural network model
Pattern awareness
Determine the current type of jamming
Jamming type 1
Jamming type N
Randomly switched channel
Train the current Q table
Choose the channel with MAX Q value
Train the current Q table
Randomly switched channel
Train the current Q table
Choose the channel with MAX Q value
Train the current Q table

FIGURE 9. Algorithm flow chart.

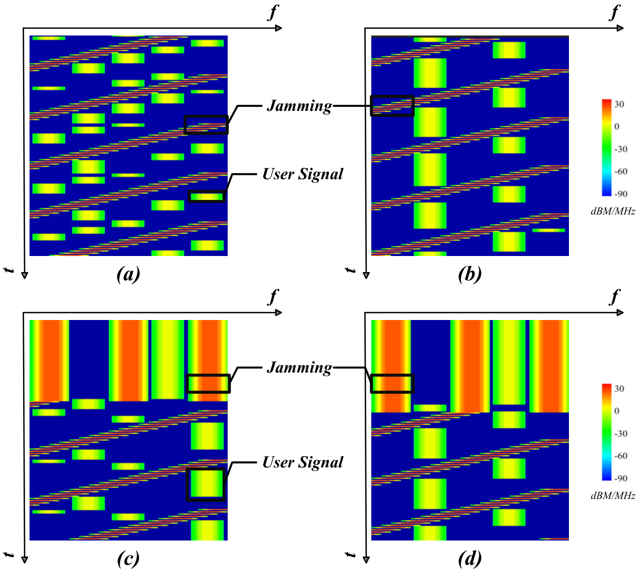


FIGURE 10. Convergence contrast graph with switching factor added.

TABLE 1. Simulation parameters.

Vectors	Initial
Number of channels	$M = 5$
The user's bandwidth	$N = 3.9MHz$
The thermal chart size	200×200
Comb jamming frequency band	$R = 2MHz, 10MHz, 18MHz$
Frequency sweep jamming rate	$F = 450MHz/s$
Jammer timeslot	$T_{jam} = 4ms$
Transmission timeslot	$T_s = 5ms$
Data transmission time	$T_d = 4ms$
ACK transmission time	$T_{ACK} = 0.2ms$
WBSS time	$T_{WBSS} = 0.3ms$
Learning time	$T_L = 0.3ms$
PIT time	$T_{PIT} = 0.2ms$
Learning rate	$\alpha = 0.1$
Reward decrement value	$\eta = 0.8$
Learning greedy factor	$\epsilon = 0.9$
Initial learning rate of CNN	0.8
Number of iterations	$max - step = 10000$
Number of test data volume	$batch - size = 25$
Average SNR	$\bar{r} = 30dB$
Average signal power	$\bar{p} = -5dB$
Feedback stimulus	$r_m = 1$
Frequency band	$b = 20MHz$

the user to perform channel switching every frame ($5ms$). The user's transmission bandwidth is $3.8MHz$. Inspired by work [40]–[42], the transmission signal is rising cosine waveforms. The demodulation threshold (th) is 10 dB. The jamming power is $30dB$, and the transmission power is $-5dB$. The feedback (r_m) of the set action is 1, and the system switching factor (c) is set as 0.6. In addition, the greedy factor is set as $\epsilon = 0.9$, and the learning rate a is 0.1. Furthermore, the discount factor $\eta = 0.8$. The number of channels is set to 5, so there are 32 states and the number of alternative action is 5. The input of CNN is 208 × 208, the number of test data volume ($batch - size$) is 25 (The amount of data entered each time is 200kb), and the number of iterations ($max - step$) is 10000. The initial learning rate of the convolutional neural network is 0.8.

Simulation mainly consider two kinds of jamming scenarios: i). Dynamic jamming including two kinds jamming patterns (Randomly switching between sweep jamming and comb jamming). ii). Dynamic jamming including three kinds jamming patterns (Randomly switching between sweep jamming, comb jamming and double sweep jamming). In detail, comb jamming is in three fixed frequency bands ($0 - 4MHz$, $8 - 12MHz$, $16 - 20MHz$), sweep jamming and double sweep jamming is set to sweep with rate $450MHz/s$. The detailed simulation parameters are described in Table. 1.

B. SIMULATION ANALYSIS

In this part, we mainly evaluate the effect of the algorithm under two kinds of dynamic jamming patterns. Fig. 11, shows the thermodynamic contrast diagram of the first jamming scenario as illustrated in previous subsection. It indicates that in the first interference scenario, the user successfully avoids all jamming. Fig. 12 shows the thermodynamic comparison diagram of the second jamming scenario. In detail, (a)-(f) shows the sliding window of double sweep jamming, comb-sweep jamming and sweep-double sweep jamming respectively, where (b) and (e), (c)and (f) are the case when the user is in the jamming transition zone.

For the purpose of calculating throughput of user, we define PN as the number of packets transmission per updating. In the simulation, we define $PN = 10$, which

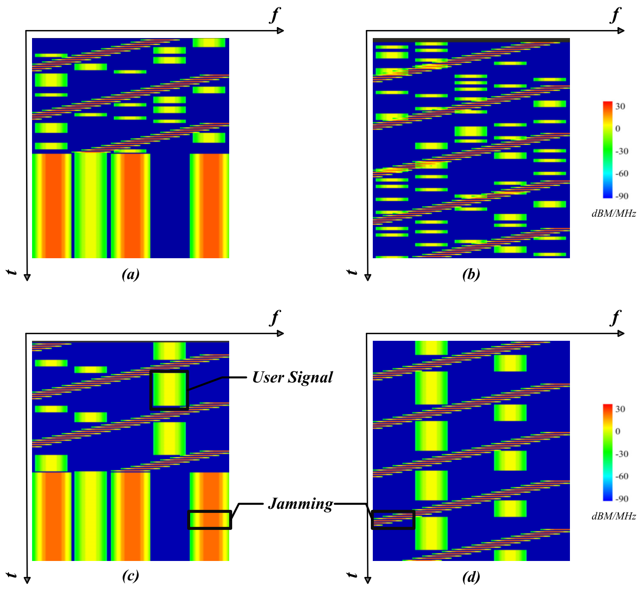


FIGURE 11. Comparison diagram of algorithm (two kinds of jamming patterns).

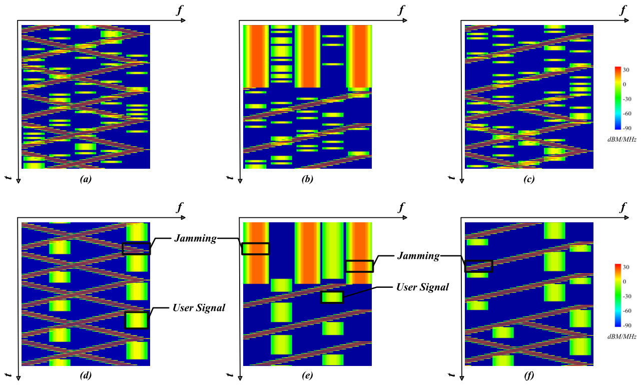


FIGURE 12. Comparison diagram of algorithm (three kinds of jamming patterns).

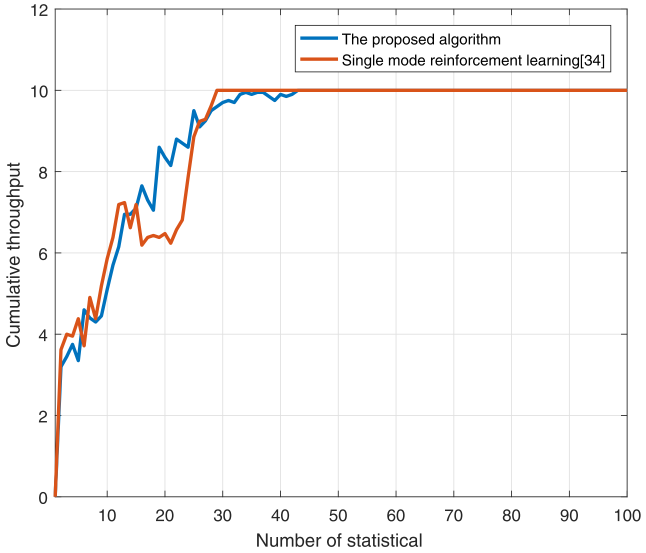


FIGURE 13. Performance comparison under a kind of jamming pattern.

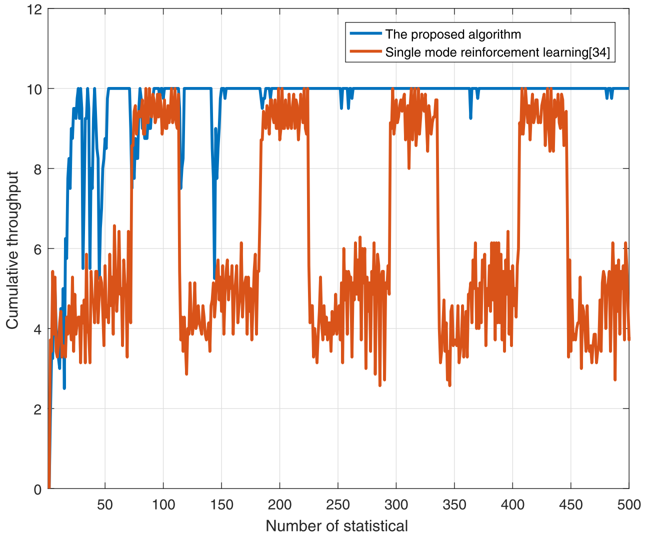


FIGURE 14. Performance comparison under period switching of two kinds of jamming pattern.

indicates that the user's maximal cumulative throughput for each updating is 10 (Each update time is 50ms). The updating number is set to be 500, and the average value of 50 monte carlo experiments is taken as the experimental result.

As is shown in Fig. 13, it describes the performance comparison between our algorithm (SDRLA) and single reinforcement learning under single sweep jamming [34]. We can conclude that the convergence speed and performance of the SDRLA is similar to the single-mode reinforcement learning algorithm.

Moreover, in Fig. 14, the cumulative throughput comparison between our algorithm (SDRLA) and single reinforcement learning under the first jamming scenario is illustrated. This scenario including two kinds periodic jamming patterns, and the jamming cycle switching time is 30 seconds. By comparing the two algorithms, it can be found that the proposed algorithm needs to be re-converge in the early stage when the jamming mode is switched. As a result, the simulation curve of the proposed algorithm will have a large fluctuation in the initial stage, but will realize convergence in the later stage. At the later stage of the algorithm, different Q value tables have been fitted, so the trained Q value table can be directly invoked according to the current jamming mode. In the single-mode reinforcement learning algorithm, the convergence state only appears under specific jamming category, which means it can only suits single jamming pattern. When channel is randomly selected by the user, the throughput is only 30%-50% of the maximum throughput.

As is shown in Fig. 15, it describes the the cumulative throughput comparison between SDRLA algorithm and single reinforcement learning under the second jamming scenario. The duration of each jamming pattern in this scenario is set to randomly switch from 2 seconds to 60 seconds. For single mode reinforcement learning, it can hardly converge

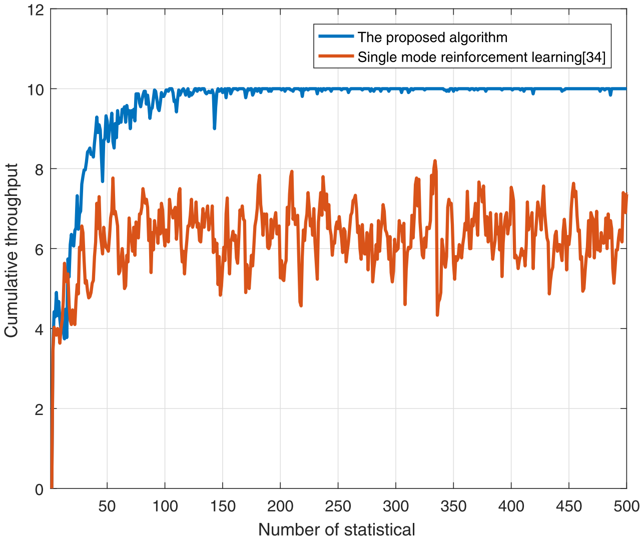


FIGURE 15. Performance comparison under random switching of three kinds of jamming patterns.

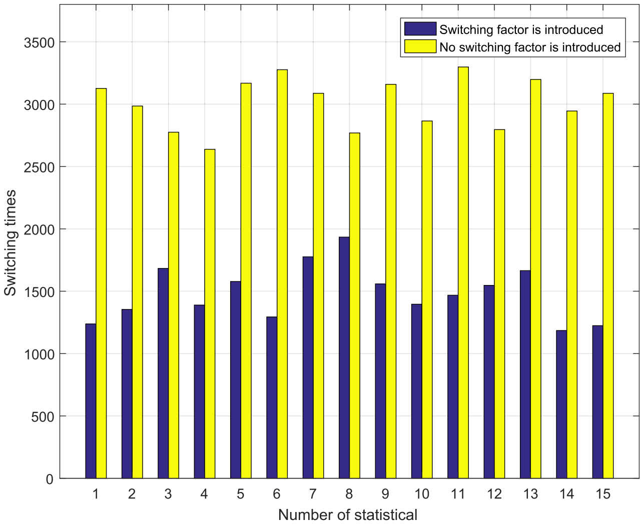


FIGURE 16. Comparison of algorithm after introducing switch factor.

in a dynamic jamming environment, and the performance is poor, which means it is difficult to converge in the dynamic jamming environment. Moreover, the channel blocking probability of the three kinds of jamming patterns are 20%-40% (Sweep jamming), 40%-80% (Double sweep jamming), and 60% (Comb jamming). Hence, the throughput of single-mode reinforcement learning fluctuates between 4-7. The algorithm proposed in this paper has slight fluctuations after convergence, which means our approach can achieve good channel selection performance. As the online learning time gets longer, the simulation curve of the proposed algorithm tends to be stable in the later stage. Moreover, the simulation results demonstrate that the proposed algorithm converges quickly and has a higher utility. It is also proved that the proposed algorithm can effectively avoid malicious jamming and achieve anti-jamming communication.

Fig. 16 shows the comparison between the SDRLA algorithm with and without channel switching factor. The simulation environment is the second jamming scenario, and keep other setting parameters unchanged. 15 monte carlo experiments are conducted, and the channel switching number in 5000 iterations are counted. As shown in the Fig. 16, it is obvious that the switching frequency of the algorithm which introduces the channel switching factor is significantly lower than the original algorithm. The switching frequency is reduced to 50%-70% of the original one. When channel switching cost is introduced, the system overhead reduces significantly.

VII. CONCLUSION

This paper studied the intelligent anti-jamming communication under dynamic jamming environment, and proposed a sequential deep reinforcement learning algorithm (SDRLA) without prior information. After conducting the SDRLA algorithm, the user learned and decided the best anti-jamming channel selection strategies when facing randomness or dynamic jamming scenarios. Furthermore, channel switching factor was introduced to improve the utility of anti-jamming channel selection. Simulation results verified the effectiveness and practicability of the proposed anti-jamming communication method, showing that our algorithm had strong environmental adaptability and wide application range.

REFERENCES

[1] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.

[2] X. Liu, Y. Xu, L. Jia, Q. Wu, and A. Anpalagan, "Anti-jamming communications using spectrum waterfall: A deep reinforcement learning approach," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 998–1001, May 2018.

[3] Y. Li, X. Wang, D. Liu, Q. Guo, X. Liu, J. Zhang, and Y. Xu, "On the performance of deep reinforcement learning-based anti-jamming method confronting intelligent jammer," *Appl. Sci.*, vol. 9, no. 7, p. 1361, 2019.

[4] 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.

[5] M. K. Hanawal, M. J. Abdel-Rahman, and M. Krunz, "Joint adaptation of frequency hopping and transmission rate for anti-jamming wireless systems," *IEEE Trans. Mobile Comput.*, vol. 15, no. 9, pp. 2247–2259, Sep. 2016.

[6] 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.

[7] L. Xiao, *Anti-Jamming Transmissions in Cognitive Radio Networks*. Cham, Switzerland: Springer, 2015.

[8] 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 Commun.*, vol. 15, no. 9, pp. 73–84, 2018.

[9] H. He and H. Jiang, "Deep learning based energy efficiency optimization for distributed cooperative spectrum sensing," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 32–39, Jun. 2019.

[10] H. Zhu, C. Fang, Y. Liu, C. Chen, M. Li, and X. S. Shen, "You can jam but you cannot hide: Defending against jamming attacks for Geo-location database driven spectrum sharing," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2723–2737, Oct. 2016.

[11] L. Zhang, Z. Guan, and T. Melodia, "United against the enemy: Anti-jamming based on cross-layer cooperation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5733–5747, Aug. 2016.

[12] 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.

[13] 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, 2018.

[14] Y. Xu, G. Ren, J. Chen, Y. Luo, L. Jia, X. Liu, Y. Yang, and Y. Xu, "A one-leader multi-follower Bayesian-Stackelberg game for anti-jamming transmission in UAV communication networks," *IEEE Access*, vol. 6, pp. 21697–21709, 2018.

[15] 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 Trans. Commun.*, vol. 65, no. 12, pp. 5305–5318, Dec. 2017.

[16] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1380–1391, Apr. 2012.

[17] L. Jia, Y. Xu, Y. Sun, S. Feng, and A. Anpalagan, "Stackelberg game approaches for anti-jamming defence in wireless networks," *IEEE Wireless Commun.*, vol. 25, no. 6, pp. 120–128, Dec. 2018.

[18] Z. Feng, G. Ren, J. Chen, X. Zhang, Y. Luo, M. Wang, and Y. Xu, "Power control in relay-assisted anti-jamming systems: A Bayesian three-layer Stackelberg game approach," *IEEE Access*, vol. 7, pp. 14623–14636, 2019.

[19] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in cognitive radio networks: Global optimization using local interaction games," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 2, pp. 180–194, Apr. 2012.

[20] 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.

[21] L. Jia, F. Yao, Y. Sun, Y. Xu, S. Feng, and A. Anpalagan, "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.

[22] Y. Xu, G. Ren, J. Chen, X. Zhang, L. Jia, and L. Kong, "Interference-aware cooperative anti-jamming distributed channel selection in UAV communication networks," *Appl. Sci.*, vol. 8, no. 10, p. 1911, 2018.

[23] 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.

[24] E. Ohbuchi, "Low power AI hardware platform for deep learning in edge computing," in *Proc. IEEE CPMT Symp. Japan (ICSJ)*, Kyoto, Japan, Nov. 2018, pp. 89–90.

[25] L. Tingpeng, W. Manxi, P. Danhua, and Y. Xiaofan, "Identification of jamming factors in electronic information system based on deep learning," in *Proc. IEEE 18th Int. Conf. Commun. Technol. (ICCT)*, Chongqing, China, Oct. 2018, pp. 1426–1430.

[26] H. R. Berenji, "Fuzzy Q-learning for generalization of reinforcement learning," in *Proc. IEEE 5th Int. Fuzzy Syst.*, vol. 3, New Orleans, LA, USA, Sep. 1996, pp. 2208–2214.

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

[28] S. Machuzak and S. K. Jayaweera, "Reinforcement learning based anti-jamming with wideband autonomous cognitive radios," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Chengdu, China, Jul. 2016, pp. 1–5.

[29] 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.

[30] G. Han, L. Xiao, and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 2087–2091.

[31] L. Xiao, D. Jiang, D. Xu, H. Zhu, Y. Zhang, and H. V. Poor, "Two-dimensional antijamming mobile communication based on reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9499–9512, Oct. 2018.

[32] Y. Yuan, L. Mou, and X. Lu, "Scene recognition by manifold regularized deep learning architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2222–2233, Oct. 2015.

[33] W. Chen and X. Wen, "Perceptual spectrum waterfall of pattern shape recognition algorithm," in *Proc. 18th Int. Conf. Adv. Commun. Technol. (ICACT)*, Pyeongchang, South Korea, Jan./Feb. 2016, pp. 382–389.

[34] F. Slimeni, Z. Chtourou, B. Scheers, V. L. Nir, and R. Attia, "Cooperative Q-learning based channel selection for cognitive radio networks," *Wireless Netw.*, vol. 24, no. 7, pp. 4161–4171, 2019.

[35] P. Bezak, "Building recognition system based on deep learning," in *Proc. 3rd Int. Conf. Artif. Intell. Pattern Recognit. (AIPR)*, Lodz, Poland, Sep. 2016, pp. 1–5.

[36] X. Kong, B. Xin, Y. Wang, and G. Hua, "Collaborative deep reinforcement learning for joint object search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1695–1704.

[37] L. Zhang, J. Tan, Y. Liang, G. Feng, and D. Niyato, "Deep reinforcement learning-based modulation and coding scheme selection in cognitive heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3281–3294, Jun. 2019.

[38] Z. Zhang, Q. Wu, B. Zhang, and J. Peng, "Intelligent anti-jamming relay communication system based on reinforcement learning," in *Proc. 2nd Int. Conf. Commun. Eng. Technol. (ICCET)*, Nagoya, Japan, Oct. 2019, pp. 52–56.

[39] M. Kang and K.-S. Hong, "Automatic bird-species recognition using the deep learning and Web data mining," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Jeju, South Korea, Oct. 2018, pp. 1258–1260.

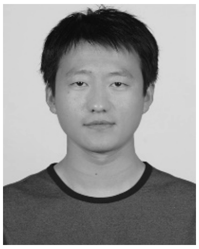
[40] A. Das, S. C. Ghosh, N. Das, and A. D. Barman, "Q-learning based co-operative spectrum mobility in cognitive radio networks," in *Proc. IEEE 42nd Conf. Local Comput. Netw. (LCN)*, Singapore, Oct. 2017, pp. 502–505.

[41] T. Ishida, I. Nitta, D. Fukuda, and Y. Kanazawa, "Deep learning-based wafer-map failure pattern recognition framework," in *Proc. 20th Int. Symp. Qual. Electron. Design (ISQED)*, Santa Clara, CA, USA, Mar. 2019, pp. 291–297.

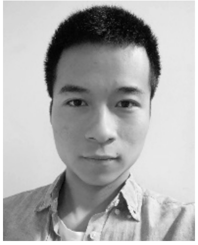
[42] X. Gao, J. Zhang, and Z. Wei, "Deep learning for sequence pattern recognition," in *Proc. IEEE 15th Int. Conf. Netw., Sens. Control (ICNSC)*, Zhuhai, Mar. 2018, pp. 1–6.

SONGYI LIU received the B.S. degree in electronic and information engineering from Xidian University, Shanxi, China, in 2018. He is currently pursuing the M.S. degree with the College of Communication Engineering, Army Engineering University of PLA. His current research interests include learning theory, dynamic optimization, and communication anti-jamming.

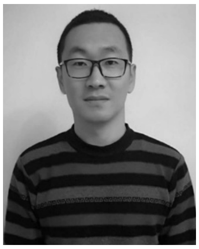
YIFAN XU received the B.S. degree in communication engineering from the Beijing Institute of Technology, Beijing, China, in 2016, and the M.S. degree in communications engineering and information system from the Army Engineering University of PLA, Nanjing, China, in 2018, where he is currently pursuing the Ph.D. degree with the College of Communication Engineering. His current research interests include game theory, learning theory, and communication anti-jamming technology.



XUEQIANG CHEN received the B.S. degree in electronics and communications engineering from the Qilu University of Technology, Jinan, China, in 2008, and the Ph.D. degree from the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He is currently a Lecturer with the College of Communication Engineering, Army Engineering University. His research is focused on cognitive radio, opportunistic spectrum access, and game theory.

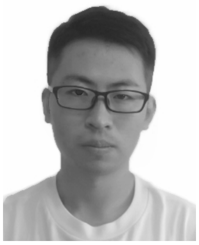


XIMING WANG was born in China, in 1993. He received the M.S. degree in communications and information systems from the College of Communication Engineering, Army Engineering University, in 2017, where he is currently pursuing the Ph.D. degree. His current research interests include anti-jamming communications, MIMO techniques, game theory, and reinforcement learning.

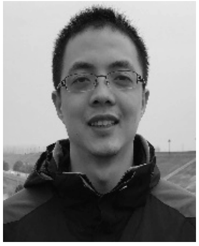


MENG WANG received the B.S. and M.S. degrees in instructional technology from Nanjing Normal University, in 2009. His current research interests include wireless network security, game theory, and intelligent anti-jamming technology.

WEN LI received the B.S. degree in electronic and information engineering from Tsinghua University, Beijing, China, in 2017. He is currently pursuing the M.S. degree with the College of Communication Engineering, Army Engineering University of PLA. His current research interests include game theory, dynamic optimization, and HF communication systems.



YANGYANG LI received the B.S. degree from the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, in 2014. He is currently pursuing the M.S. degree with the College of Communication Engineering, Army Engineering University of PLA. His current research interests include deep reinforcement learning, deep learning, and anti-jamming communications.



YUHUA XU received the B.S. degree in communications engineering and the Ph.D. degree in communications and information systems from the College of Communications Engineering, PLA University of Science and Technology, in 2006 and 2014, respectively. He is currently an Associate Professor with the College of Communications Engineering, Army Engineering University of PLA. His research interests focus on UAV communication networks, opportunistic spectrum access, learning theory, and distributed optimization techniques for wireless communications. He has published several articles in international conferences and reputed journals in his research area. He received Certificate of Appreciation as Exemplary Reviewer for the IEEE Communications Letters, in 2011 and 2012, respectively. He was selected to receive the IEEE Signal Processing Society 2015 Young Author Best Paper Award and the Funds for Distinguished Young Scholars of Jiangsu Province, in 2016.