Intelligent Anti-Jamming Decision Algorithm of Bivariate Frequency Hopping Pattern Based on DQN With PER and Pareto

Jiasheng Zhu, Hangzhou Dianzi University, China*
Zhijin Zhao, School of Communication Engineering, Hangzhou Dianzi University, China
Shilian Zheng, Science and Technology on Communication Information Security Control Laboratory, China

ABSTRACT

To improve the anti-jamming performance of frequency hopping system in complex electromagnetic environment, a Deep Q-Network algorithm with priority experience replay (PER) based on Pareto samples (PPER-DQN) is proposed, which makes intelligent decisions for bivariate FH patterns. The system model, state-action space, and reward function are designed based on the main parameters of the FH pattern. The DQN is used to improve the flexibility of the FH pattern. Based on the definition of Pareto dominance, the PER based on the TD-error and immediate reward is proposed. To ensure the diversity of the training set, it is formed by Pareto sample set and several random samples. When selecting Pareto sample, the confidence coefficient is introduced to modify its priority. It guarantees the learning value of the training set and improves the learning efficiency of DQN. The simulation results show that the efficiency, convergence speed, and stability of the algorithm are effectively improved. And the generated bivariate FH pattern has better performance than the conventional FH pattern.

KEYWORDS

Bivariate Frequency Hopping Pattern, Complex Electromagnetic Environment, Deep Reinforcement Learning, Pareto, Priority Experience Replay

INTRODUCTION

Frequency hopping communication system, as a traditional anti-interference technology, has strong anti-jamming ability. In the past few decades, based on the traditional frequency hopping communication technology, differential frequency hopping, multi sequence frequency hopping, variable speed frequency hopping, cognitive frequency hopping and other technologies have been developed. It can be found that most of these technologies realize anti-interference by timely and appropriately adjusting the important parameters which have a great impact on the performance of frequency hopping system. However, under the influence of increasingly complex electromagnetic environment and gradually intelligent interference strategy, these frequency hopping communication technologies can no longer meet the communication needs.

DOI: 10.4018/IJITWE.297970

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.
As the most important parameter of frequency hopping communication system, the research on frequency hopping pattern has been promoted all the time. Researchers often screen frequency points on the basis of accurate spectrum sensing results, and generate frequency hopping patterns in various ways. However, these methods have limited effect under the complex electromagnetic environment. In order to ensure the quality of frequency hopping communication, the application of more intelligent anti-interference technology in the design of frequency hopping pattern is necessary.

With the development of machine learning technology, the anti-interference technology is becoming more and more intelligent. As an important branch of machine learning, reinforcement learning is an algorithm based on Markov decision process (MDP). Its essence is the process in which agents constantly interact with the environment. Reinforcement learning interacts with the environment through an agent with learning ability. The agent selects and executes a certain action according to the current state and the learned environmental experience, and transfer to a new state under the joint action of the action and the environment. At the same time, the environment will also feedback certain rewards or punishments according to the current state and the actions taken. The agent updates the cognition of the current environment through rewards or punishments, and makes subsequent decisions. After completing the learning, the agent will obtain an environment-state-action mapping relationship. It is used to guide the agent to select the appropriate action based on the state in the actual decision-making process and obtain the maximum cumulative reward value from the environment. Therefore, reinforcement learning is suitable for solving intelligent decision-making problems in complex environments.

The intelligent anti-jamming decision algorithm of frequency hopping communication system based on deep reinforcement learning in complex electromagnetic environment is studied in this paper. Through intelligent decision-making, frequency hopping patterns with variable hopping rate and frequency interval are generated, which are more suitable for the complex spectrum environment and can avoid or reduce the impact of interference on the overall communication process as much as possible. In order to make the occupied frequency band can appear at anywhere in available spectrum and improve the flexibility of frequency hopping pattern and the utilization rate of spectrum, DQN algorithm is used for intelligent decision-making. Considering the need to improve the efficiency of the algorithm in the complex electromagnetic environment, the improved Priority Experience Replay (PER) technology is applied to DQN. Because there are researches show that more effective examples can be sampled from the sequence with large cumulative reward which make DQN algorithm achieve the best strategy faster. The TD-error and immediate reward are both considered in the PER process of this paper to further improve the superiority and learning value of the training set. In order to screen samples based on two criteria at the same time, the priority experience replay based on Pareto samples is proposed. It ensures the diversity of training set, improves the sample utilization and the efficiency of experience replay, so as to improve the performance and convergence speed of the algorithm. In addition, the confidence parameter is introduced to measure the reliability of Pareto samples and the probability of selecting samples with high priority but long storage time is reduced. Cause the samples with long storage time come from the network a long time ago, their reliability is low. Through the above ameliorations, the overall superiority and learning value of training set and the learning efficiency of DQN algorithm can be improved and the Deep Q Network with Priority Experience Replay Based on Pareto Samples (PPER-DQN) is proposed. The simulation results show that in the intelligent decision-making problem of bivariate frequency hopping pattern, this algorithm can screen the experience pool more efficiently, improve the convergence speed of the algorithm effectively. And the bivariate frequency hopping pattern has better performance than the conventional frequency hopping pattern.

This paper studies the design method of bivariate FH pattern using deep reinforcement learning algorithm, and hopes to fundamentally improve the anti-interference performance of frequency hopping communication system by intelligently generating high-performance FH pattern. The remainder of this paper is organized as follows. Section 2 summarizes the work related to FH patterns and
deep reinforcement learning algorithms in recent years. Section 3 describes the model of bivariate frequency hopping communication system. The intelligent anti-jamming decision algorithm of bivariate frequency hopping pattern based on PPER-DQN is proposed in section 4. The simulations are performed and discussed in section 5. Finally, we conclude the paper in section 6.

RELATED WORK

Compared with the traditional communication system, the frequency hopping communication system has some unique parameters which affect system performance greatly, such as frequency set and hopping rate. According to these parameters, the frequency hopping can be realized and achieve the purpose of anti-interference and anti-interception during communication. However, in the traditional frequency hopping communication system, its parameters cannot change with the environment. So, its advantages are difficult to reflect under the influence of the increasingly complex electromagnetic environment and gradually intelligent hostile interference. In recent years, the intelligent anti-jamming decision-making technology of the frequency hopping communication system in complex electromagnetic environment has attracted the attention of scholars from various fields.

Excellent frequency hopping pattern, as one of the most important parameters of frequency hopping system, can avoid interference, improve the anti-interference performance and the communication quality effectively. At present, most of the researches was the design of frequency hopping patterns in the undisturbed frequency band which got by spectrum sensing. These researches all focused on various pseudo-random sequences and their improvement and encryption. Wang, X., Li, L., & Zhang, S. (2018) proposed a method for generating frequency hopping patterns with wide gap in discontinuous frequency band. According to the constraints of frequency hopping interval and uniformity frequency hopping pattern, the random frequency hopping pattern is generated based on the frequency point and uniformity compensation frequency distribution function. Wu, X. (2017), firstly, uses the improved AES algorithm to encrypt RS sequence. Then, they generate their frequency hopping pattern by this improve RS sequence. What’s more, they introduce wide gap technology into the generating process to ensure that the interval of adjacent frequency is not less than a specified value and improve its confidentiality.

However, in the complex electromagnetic environment, it is hard to ensure accurately sensing spectrum at all times and there are few undisturbed frequency bands. It will greatly limit the performance of frequency hopping patterns. In this case, the intelligent decision-making of various parameters of frequency hopping pattern will further improve the anti-interference performance of the system. Hopping rate and channel division interval are essential parameters of it. The higher the hopping rate is, the stronger the anti-interference performance is, but the higher performance the hardware are required. And a wider channel division interval is conducive to anti-interference and multi-path fading, but it is also more vulnerable to partial band interference. In this regard, on the one hand, Chen, G. & Li, F. (2016) analyzed the principle and main parameters of tracking interference and proposed the frequency hopping communication with variable hopping rate. They make users’ residence time change pseudo-randomly, and adopts the majority decision method. Then, Na, D. (2016) theoretically analyzed and verified that the frequency hopping pattern with variable hopping rate can reduce the bit error rate (BER) and improve the anti-jamming performance of the system by making the user’s residence time change pseudo-randomly.

On the other hand, Ren. W., Gao, X. & Wang, F. (2020) constructed a new class of wide-gap frequency hopping sequences based on the known prime sequences, interleaving technique and Chinese remainder theorem. These sequences have more flexible parameters and lose less information when interfered by single frequency narrowband interference and partial band blocking interference. Wang, Y., & Quan, H. (2020) a wide gap multi-pattern frequency hopping (WGMPFH) scheme to improve the anti-following jamming performance of frequency hopping system. WGMPFH has the data channel and the complementary channel and uses channels to represent messages. The data channel
is used to lure the following jamming and the complementary channel will be away from it. By this way, following jamming does not affect the complementary channel but increases the signal energy in the data channel, thus the effect of follow jamming is reduced. The experimental results show that WGMPFH shows superior jamming rejection performance under following jamming especially in severe signal-to-jamming ratio (SIR).

Yan, J. (2012) pointed out that for effectively interfering frequency hopping users, the jammer needs to quickly and accurately extract various characteristic information of the target users through signal reconnaissance, interception and analysis. Only after that, jammer can clarify the frequency band to be interfered and guide the equipment to concentrate power to this frequency band. In this case, fixed residence time and minimum frequency interval are obvious characteristics to jammer. Actually, the intelligent decision of the parameters of frequency hopping pattern has already been studied. By using intelligent decision algorithm, the parameters of frequency hopping pattern can be changed according to the characteristics of interference environment. These changes will increase the difficulty of signal analysis for the jammer and effectively improve the system performance of the counter tracking interference, comb interference, blocking interference and the counter interference. Further, they also indicated that the bivariate frequency hopping pattern had the potential to further improve its anti-jamming ability by combining with the technologies of frequency adaptive, power adaptive and so on. Obviously, the intelligent decision of the parameters of frequency hopping pattern has great significance to improve the anti-jamming performance of the communication system. However, at present, most of the research on frequency hopping pattern with variable parameter only involves one of the speed domain or frequency domain. The frequency hopping pattern with both variable hopping rate and frequency interval are less concerned.

As mentioned above, previous studies are all based on the accurate environment model. However, in the complex electromagnetic environment, the model of environment is always unknow or cannot be fully described. It leads to that the decision-making of traditional optimization algorithm is poor. As an important branch of machine learning, reinforcement learning can well adapt to the dynamic and complex environment, which relies on its characteristics of model free and non-supervision. Q-learning, which is the most common reinforcement learning algorithm, has been widely used in various decision-making problems. Huang, J. (2020) uses Q-learning to intelligently generate frequency hopping patterns in the environment with multitone sweeping interference. The results show Q-Learning can help the agent to learn the characteristics of complex environment and jamming effectively and avoid multitone sweeping interference timely. However, Q-learning is troubled by the problem of dimension explosion. Deep Reinforcement Learning (DRL) has effectively made up this problem and been widely used in various large-scale decision-making problems in complex environment. Han, C. & Huo, L. (2020) proposed a Deep Reinforcement Learning Based Routing Algorithm (DRLR) to obtain an available routing subset in the problem of the anti-jamming communication of the heterogeneous Internet of Satellites (IoS). Frikha, M. S., & Gammar, S. M. (2021) discussed and summarized the application of reinforcement learning and deep reinforcement learning in the field of Internet of Things (IoT). At present, reinforcement learning and deep reinforcement learning have shown good performance in resolving issues related to routing, scheduling, resource allocation, dynamic spectrum access, energy, mobility and caching. Huang, L., Xu, T. & Chen, X. (2021) solved the joint relay and channel selection problem based on a DRL approach in the multi-relay anti-jamming communication system. These researches are all troubled by the intelligent jamming and the high dynamics of environment.

As the environment becomes more and more complex, the efficiency of deep reinforcement learning also needs to be further improved. Experience replay is one of the most important part of Deep Q-Network (DQN) which is the most common DRL approach. And the keys of the experience replay are the sampling, storing and replaying of samples. But the random sampling mechanism limits its efficiency. Manela, B., & Biess, A. (2021) introduce the improved hindsight experience replay (HER), a method for multi-goal reinforcement learning algorithm with sparse reward functions, into
the Deep Deterministic Policy Gradient (DDPG) to improve its performance. Based on HER, they prioritize the virtual goals from which the agent will learn more valuable information and reduce existing bias in HER by the removal of misleading samples. So that, they successfully improve the sample efficiency. Cao, X., Wan, H. (2019) classified the samples to several groups based on their temporal difference-error (TD-error) which represents the accuracy of the network’s prediction to the sample. Then, the samples are extracted from different groups by different proportion to improve the overall performance of the sample set and ensure its diversity. Shi, S. & Liu, Q. (2021) adopted the classified experience replay method and proposed the Deep Deterministic Policy Gradient with Temporal Difference-error Classification (TDC-DDPG) and the Deep Deterministic Policy Gradient with Reward Classification (RC-DDPG). In these algorithms, to indicate its importance, the new sample was classified according to the average TD-error or immediate reward of all samples in the main experience pool. And two deputy experience pools were used to store classified samples separately. During the sampling process, most of the training set were sampled from the deputy experience pool with high importance by randomly selecting and the less of it were sampled from another deputy experience pool by the same way. TDC-DDPG and RC-DDPG effectively improve the quality of training set samples and the performance of DDPG. However, in order to ensure the accuracy of classification, the two algorithms need to use the Q-network frequently to calculate TD-error of experience pool every time the new sample appears. This consumed a lot of computing resources. Wang, T. & Luo, Y. (2021) proposed a DDPG Algorithm Considering the State Distribution (DDPG-SD) was. It compared the current state data with the saved state data and selected different random policy parameters according to their similarity degree. At the same time, it improved the replay probability of scenes with fewer times to be replayed. This algorithm can keep enough exploration ability when it confronts the scene with large difference from the previous data distribution at the later stage of training. However, too strong exploration ability and its experience replay mechanism cause the unstable performance after convergence.

In order to reduce the cost of interaction between agent and environment more effectively and improve the efficiency of sample utilization and experience replay, Schaul et al. (2016) proposed the Priority Experience Replay (PER) first and modified the DQN algorithm. In PER, different priorities were given to samples according to their TD-error. Then the samples were selected based on the probabilities which are proportional to their priorities, so as to improve the quality of training set and the efficiency of algorithm.

Ye, Z. & Wang, Y. (2019) used PER and Dueling Deep Q Network (Dueling DQN) to solve the problem of multi-user dynamic power control strategy in cognitive radio. Cause the secondary users may access the primary users’ channels to perform transmission tasks without knowing the control policy or transmission power of the primary users. The received signal strength information collected by the micro base station is input to the Dueling DQN as environment state information. And the output of Dueling DQN is the dynamic power control policy of the secondary users. The experimental results show that the optimal power control policy can be obtained after training and learning. The secondary users can improve their behavior when environmental parameters are updated and the spectrum utilization efficiency is enhanced. In addition, Dueling DQN is also an effective and sample method to improve the performance of DQN and has been widely used. Sometimes, for an agent, the actions to be performed have little impact on the future state and environment. Based on this idea, Dueling DQN improves the efficiency of the algorithm by splitting Q-value into state flow and value advantage flow. Similarly, the Dueling Deep Q Network with Rank-based Priority Experience Replay was proposed by Zhou, Y. & Li, Y. (2020). Before being sampling, the TD-errors of samples are sorted. Different from PER, Rank-based Priority Experience Replay will give priority to samples based on their ranking. The lower the sort, the lower the priority. This mechanism can avoid the lack of diversity of training set and the problem of network over fitting caused by samples with exorbitant priorities. Hu, Z., & Gao, X. (2021) design a sampling method with double-screening, combine it with the DDPG algorithm and proposes the Relevant Experience Learning-DDPG (REL-
The REL-DDPG algorithm uses the PER mechanism to break the correlation of continuous experiences in the experience pool, finds the experiences most similar to the current state to learn according to the theory in human education, and expands the influence of the learning process on action selection at the current state. Compare with DDPG, REL-DDPG improves the convergence speed and the convergence result, but the theory in human education leads to the decline of its universality.

In addition, the problem of low credibility of the old samples in the experience pool also limits the performance of the experience replay. For solving it, Wang G. (2019) proposed the priority experience replay based on storage period. The priority is given according to the storage period of samples, not their Td-errors, which help the new samples to be sampled easier. But it also makes it difficult to guarantee the independence of training samples. Bai, C. & Liu, P. (2019) analyzed the relationship among storage priority, real priority and storage period, and proposed an Active Sampling Method Based on Adaptive TD-error Correction (ATDC-PER). Based on the linear regression model, the regularly updated prediction weight matrix, storage priority and storage period were used to predict the real priority, and realized the priority correction. It avoids the problem of frequently using network in TDC-DDPG and RC-DDPG algorithms. However, its correction mechanism is based on the linear relationship between the priority difference and the storage period, so it is not universal. What’s more, it involves matrix inversion, which consumes a lot of computing resources, and its existence cannot be guarantee.

MODEL OF BIVARIATE FH COMMUNICATION SYSTEM

Signal Model of Bivariate Frequency Hopping

In the conventional frequency hopping communication, the hopping rate and frequency interval of the frequency hopping pattern are fixed. So, the residence time of each hop is invariant and the frequency points in the frequency set are distributed with the integer multiple of the fixed minimum interval $\Delta f$. The conventional frequency hopping signal can be expressed as follows:

$$x(t) = A \sum_{l=1}^{L} \cos(2\pi f_{c} t + a_{l} 2\pi f_{l} t)p(t - l\tau)$$

where $f_{c}$ is the minimum hopping frequency, $a_{l}$ is the frequency control word which generated by the pseudo-random sequence and used to control the frequency change, $f_{l}$ is the minimum frequency interval which determined by the channel partition interval, $p(t)$ is the pulse function, and $\tau$ is user’s residence time in each hop which determined by the hopping rate.

To increase the randomness of frequency hopping pattern, the pseudo-random factors are introduced into the hopping speed and frequency interval of the bivariate frequency hopping pattern. The pseudo-random variables $\tau(a_{l})$ and $f_{l}(a_{l})$ are used to replace the residence time $\tau$ and the minimum frequency interval $f_{l}$, respectively. The equation (1) is modified to equation (2):

$$x(t) = A \sum_{l=1}^{L} \cos[2\pi f_{c} t + a_{l} 2\pi f_{l}(a_{l})]p(t - l\tau(a_{l}))$$

The strategy of variable hopping rate can ensure the communication quality and increase the difficulty of being reconnoitered. As shown in equation (3), if the hopping rate of the bivariate
frequency hopping pattern is $V \in [V_{\text{lower}}, V_{\text{upper}}]$, in the $k$’th hop, user’s hopping rate is $V_k$ and the corresponding residence time is $\tau_k$:

$$\frac{1}{V_{\text{upper}}} \leq \tau_k = \frac{1}{V_k} \leq \frac{1}{V_{\text{lower}}}$$ (3)

It can be seen from equation (3) that there is a nonlinear relationship between residence time and hopping rate. When the hopping rate is not very high, the variation range of residence time is large. The nonlinear relationship increases the difficulty and time-consuming of the jammer to detect and analyze the frequency hopping signal, so as to improve the anti-jamming performance of the communication system. When the hopping rate is high, fast hopping is the main way to resist the interference.

The strategy of variable interval makes the frequency set no longer have the original characteristic that the frequency points arranged with the integer multiple of the minimum frequency interval. It allows the occupied channel can be anywhere in the entire spectrum. An example of spectrum distribution is shown in Figure 1. The first line is the interference, the second line is conventional FH pattern and the third line is bivariate FH pattern. Because the frequency interval of a conventional frequency pattern is fixed, the spectrum distribution is detected and interfered more easily and seriously, five frequency points of conventional FH are interfered, while only one frequency point of bivariate FH is interfered. So, the flexibility of frequency interval improves the performance of the communication system to resist narrowband interference, multitone interference and comb interference.

Starting from the speed domain and frequency domain, the bivariate frequency hopping technology increases the difficulty of feature extraction and the time jammer spends for tracking and guiding. It can effectively improve the anti-tracking interference performance.

The Optimization Model for Bivariate Frequency Hopping System

Supposed that the users communicate in the spectrum $W_{\text{lower}} \sim W_{\text{upper}}$ based on the bivariate frequency hopping communication technology. Their transmission power is $P$, source rate is $C$, hopping rate is $V \in [V_{\text{lower}}, V_{\text{upper}}]$ and frequency interval is $D \in [D_{\text{lower}}, D_{\text{upper}}]$. Based on the predicted spectrum state in a short-time future $\Delta$, the bivariate frequency hopping pattern can be formed by intelligent anti-jamming decision. And the spectrum occupation of the $k$’th hop can be determined by frequency point $W_k$, hopping rate $V_k$ and frequency interval $D_k$.

Obviously, for frequency hopping communication system, the higher Signal to Interference plus Noise Ratio (SINR) the bands users occupied with and the shorter the users reside on these bands,
the better the overall communication quality is. Therefore, the communication quality can be improved by maximizing the SINR of the whole communication process. And the objective function of maximizing SINR in the short-time future $\Delta$ can be expressed by equation (4):

$$\max \sum_k \frac{C}{V_k} \text{SINR}_k = \max \sum_k \frac{1}{V_k} \text{SINR}_k = \max \sum_k \frac{10}{V_k} \Delta \log \left( \frac{P}{J_k + n_0} \right)$$  \hspace{1cm} (4)

where $1/V_k$, $\text{SINR}_k$ and $J_k$ are the residence time, the SINR and the received interference power of the $k$'th hop, respectively, and $n_0$ is the power of Gaussian white noise. The formula (5) indicates that $K$ hops has passed in $\Delta$.

In conclusion, the optimization model of bivariate frequency hopping system is shown in equation (6):

$$\max \sum_k \frac{1}{V_k} \Delta \text{SINR}_k$$

s.t. \hspace{0.5cm} $\text{SINR}_k = 10 \log \left( \frac{P}{J_k + n_0} \right)$

$$\sum_k \frac{1}{V_k} = \Delta$$

$$V \in [V_{lower}, V_{upper}], D \in [D_{lower}, D_{upper}]$$

**INTELLIGENT DECISION MAKING BASED ON DQN WITH PRIORITY EXPERIENCE REPLAY AND PARETO SAMPLE**

**Deep Q-Network**

Q-learning is not suitable for solving large-scale discrete domain problems or continuous domain problems. As the most common deep reinforcement learning algorithm, DQN improves Q-learning in the following three aspects:

1. **Deep neural network model**: Used to instead of Q-table and fit the state-action function. It avoids the problem of excessive computation and low efficiency caused by dimension explosion.

2. **Experience replay**: The experience pool is established to store the sample generated by the agent. In each iteration, a part of the experience is randomly sampled from the experience pool and used to train the network. By random sampling, the independence of samples is ensured and the efficiency of network training is improved.

3. **Dual network structure**: The valuation Q-network and the target Q-network are established respectively for action selection and calculating target Q-value. After the valuation Q-network is updated several times, its parameters are assigned to the target Q-network. In this way, the update of the target Q-network will lag behind that of the valuation Q-network, and the stability of the algorithm will be improved.
Priority Experience Replay Based on Pareto Samples

Priority experience replay can give priorities to the samples and determine their probability of being sampled according to priorities. Compared with the uniform random sampling method of the traditional experience replay, sampling with priorities can improve the overall superiority of the training set. The key of priority experience replay is how to judge the importance of samples. TD-error is usually used as the criterion to judge the priority, and its form is shown in equation (7):

$$\delta^i_t = r_t + \gamma \max_a Q(s_{t+1}, a; \theta_{\text{val}}) - Q(s_t, a_t; \theta_{\text{val}})$$

(7)

where $\delta^i_t$ is the TD-error of the sample $i$ which generated in the iteration $t$, $r_t$ is the immediate reward that the agent is in the state $s_t$ and performs the action $a_t$ in the iteration $t$, $\gamma$ is the discount factor which represents the importance of future rewards, $Q(s_t, a; \theta_{\text{val}})$ and $Q(s_t, a; \theta_{\text{tar}})$ are the Q values obtained by the valuation Q-network and the target Q-network respectively, $\theta_{\text{val}}$ and $\theta_{\text{tar}}$ are the parameters of the two networks separately. The closer $\delta^i_t$ is to 0, the smaller the prediction accuracy of the network increase, and the smaller the improvement of network can get by replaying this sample. Replaying more samples whose $\delta^i_t$ are far away from 0 will improve the efficiency of training network and make the algorithm converge more easily.

On the basis of equation (7), the common priority definition are TD-error-based priority and rank-based priority whose forms are shown in equations (8) and (9) respectively. Correspondingly, the probability of being selected for sample is shown in equation (10):

$$p_i = (|\delta^i_t| + \sigma)^\alpha$$

(8)

$$p_i = 1 / \text{rank}_i^\alpha$$

(9)

$$P_i = p_i / \sum_{j=1}^{N} p_j$$

(10)

where $p_i$ is the priority of the sample $i$, $\sigma$ is a small positive number which used to ensure the sample with a very small $|\delta^i_t|$ also has a certain degree of priority, $\alpha \in [0, 1]$ is the coefficient which controls the use degree of priority, when $\alpha = 1$, the uniform random sampling method is adopted, $\text{rank}_i$ is the ranking of the sample $i$ in all samples according to the descending order of $|\delta^i_t|$. $P_i$ is the selected probability of the sample $i$, $N$ is the capacity of experience pool. Compared with equation (8), the definition of equation (9) is less sensitive to outliers and more robust. However, its priorities decay too fast, which makes the priority gap between the top samples is too large and the discrimination is insufficient for the backward samples. It is not of advantage to maintain the diversity of training set when there is no sample with exorbitant priority.

The $\delta^i_t$ of conventional experience replay only represents the evaluation error, but cannot judge whether the action is the best action in this state. As a result, the superiority of the actions in the replayed samples cannot be guaranteed, which limits the learning effect. After experiment and analysis,
Zhao, Y. & Liu, P. (2019) has proposed that more effective actions can be sampled from the sequence with large cumulative reward which make DQN algorithm achieve the best strategy faster. Therefore, in this paper, TD-error and immediate reward are both considered in the priority experience replay to further improve the superiority and learning value of the training set, so as to improve the performance and convergence speed of the algorithm.

In order to sample more efficiently, Pareto Dominance is defined as follows.

For sample $e_1$ and $e_2$, if and only if all $f_u(e_1)$ are better than $f_u(e_2), u = 1, 2, \cdots, U$, $e_2$ is dominated by $e_1$, otherwise, $e_1$ and $e_2$ do not dominate each other. Where $f_u(\cdot)$ is the $u$’th performance function of the sample.

According to the above definition, the priority experience replay based on Pareto samples is proposed in this paper. In this method, Pareto samples are selected from two aspects of TD-error and immediate reward. Considering that the interference power of different frequency bands is discrepant, it is unfavorable to compare the samples in different frequency bands. It will affect the diversity of training set and the superiority of network selection action. Therefore, the spectrum is divided into $G$ segments, and the Pareto samples are selected from the samples in the same segment. It ensures the Pareto sample set can contain the samples of each frequency band.

To solve the problem that the credibility of samples, which has stored for a long time in the experience pool, will decline, the confidence parameter is introduced in this paper to measure the samples. When the dominant relationship between samples is judged, the confidence parameter is normalized, as shown in equation (11), and is used to adjust the priorities of samples:

$$
\mu = \frac{\xi - t_{store}}{z} \tan(\frac{\xi - t_{store}}{z})
$$

(11)

where $\mu$ is the confidence parameter, $t_{store}$ is the storage period of the sample in the experience pool, $\xi$ and $z$ is used to control the rate of $\mu$ decreasing with the storage period. The samples with long storage period are given small confidence parameters to reduce their priorities appropriately. The selected probabilities of samples with low reliability are also reduced further. And the training efficiency can be improved.

The samples with high priority are easier to be selected and replayed frequently, so the diversity of training set is reduced, and the network is easy to overfit. The importance sampling weight is used to correct this situation. Equations (12) and (13) are the importance sampling weights corresponding to TD-error-based and rank-based priorities, respectively. Equation (14) indicates the target value for training the parameter of network. Equation (15) indicates the corrected network loss function:

$$
 w_i = (NP_i)^{-\beta}
$$

(12)

$$
 w_i = \left( \frac{P_i}{P_{\min}} \right)^{-\beta}
$$

(13)

$$
 y_i = \begin{cases} 
 r_i, & \text{if episode end} \\
 r_i + \gamma \max_a Q(s_{t+1}, a; \theta_{se}), & \text{else} 
\end{cases}
$$

(14)
\[ L(\theta_{val}) = \sum_i w_i \left[ y_i - Q(s_i, a_i; \theta_{val}) \right]^2 \]  

(15)

where \( w_i \) is the importance sampling weight of sample \( i \), \( \beta \) is the parameter of the degree of \( w_i \) corrected, \( y_i \) is the target value of network which got from sample \( i \), the current episode ends when the state is traversed, \( r_i \) is the immediate reward of the sample \( i \), \( Q(s_{t+1}, a_{t}; \theta_{val}) \) and \( Q(s_t, a_t; \theta_{val}) \) represent the target Q value and the valuation Q value of the sample \( i \) respectively.

Sampling with some certain probabilities from the experience pool will greatly consume computing resources and affect the overall efficiency of the algorithm, so, as shown in Figure 2, Sumtree structure is introduced in this paper.

Through the Sumtree, the samples with higher priority can be selected more easily, and the time complexity of sampling process is low. Assuming that the experience pool has \( N \) samples and there are \( M \) samples need to be selected, the time complexity is \( O(NM) \).

In order to ensure the diversity of the training set and avoid overfitting, \( \eta \% \) of the samples in the training set are selected from Pareto sample set, and the rest are from non-Pareto sample set. Since the selected probabilities of Pareto samples are much higher than that of non-Pareto samples, the formula (12) should be improved. When the sample is a Pareto sample, the highest priority in the current experience pool is given to it, otherwise, the original priority is used. The improved importance sampling weights are shown in equations (16) and (17):

\[ w_i = \left( NP'_i \right)^{-\beta} \]  

(16)

\[ P'_i = \begin{cases} 
\max_{j} P_j, & j \in [1,N], \text{if } e_i \text{ is Pareto Sample} \\
 P_i, & \text{else} 
\end{cases} \]  

(17)

**State-Action Space, Action Selection Strategy and Reward Function**

\( \varepsilon \)-greedy mechanism is the most common action selection strategy in DQN. During the action selecting process, a random number is generated first. When the random number is less than \( \varepsilon \), the agent randomly selects an action as the best action and executes it. Otherwise, the agent compares the Q value of each action and selects action with the largest Q value as the best action and executes it. Because the \( \varepsilon \) of the conventional \( \varepsilon \)-greedy mechanism is fixed or decreases linearly with the iterations, the agent still has a certain probability to select and execute the random action in the middle.
and late iterations. The algorithm converges slowly. Therefore, as shown in equations (18) and (19), nonlinear factor is introduced into the ε-greedy mechanism in this paper:

\[
\pi(s_t) = \begin{cases} 
a_{random}, & \text{rand} < \varepsilon \\
\max_a Q(s_t, a), & \text{else}
\end{cases}
\] (18)

\[
\varepsilon = 1 / \left[ \hat{\lambda} t / \bar{T} \right]
\] (19)

Among them, \(\pi(s_t)\) is the best action selected by the agent in the state \(s_t\) and iteration \(t\), \(a_{random}\) is a randomly selected action, \(\text{rand} \in [0,1]\) is a random number, \(\lambda\) is used to control the rate of decrease of \(\varepsilon\), \(\bar{t}\) is the current training epoch number of the network, \(\bar{T}\) is the total number of training epoch. This strategy ensures that \(\varepsilon\) is large enough at the beginning of training, and make the agent explore actively. Then \(\varepsilon\) decreases rapidly, which makes the agent pay more attention to utilization.

The state of the agent is defined as frequency set and the action is the joint allocation of hopping rate \(V\) and frequency interval \(D\). In order to maximize the SINR, the reward function is defined as the SINR obtained by the agent in the hop:

\[
r_t = \text{SINR}_t = 10 \log_{} \left( \frac{P}{J_t + n_0} \right) = 10 \log_{} \left( \frac{P}{\int_{f_t-D/2}^{f_t+D/2} \left( n_0(f) + \sum_i J_i (f - f'_i) \right) df} \right)
\] (20)

where \(r_t\) is the immediate reward obtained in iteration \(t\), \(J_t\) is the received interference power in iteration \(t\), which caused by the common influence of several kinds of interference, \(J_i (f)\) and \(f'_i\) are the power spectral density function and the center frequency of the interference \(i\), \(n_0(f)\) is the power spectral density function of Gaussian white noise.

**Algorithm Summary**

To sum up, the specific steps of the intelligent decision algorithm based on PPER-DQN are as follows. The state of the agent is frequency set and the action is the joint allocation of hopping rate \(V\) and frequency interval \(D\):

**Step 1:** According to the sensed spectrum state, the main parameters of interference are estimated, and the spectrum state in a short-time future \(\Delta\) is predicted.

**Step 2:** Initialize the valuation Q-network, target Q-network, experience pool and Sumtree. Set the learning rate of the network \(\eta\), the update period of the target Q-network \(T_{tar}\), the capacity of training set \(M\), the discount factor \(\gamma\), the parameters \(\alpha\), \(\beta\), \(\lambda\), \(\eta\), \(\xi\), \(z\), \(G\) and the total number of training epoch \(\bar{T}\). Initialize state \(s_0\) randomly.

**Step 3:** For the current state \(s_t\), the best action is selected and executed according to formula (18), (19) and valuation Q-network. The next state \(s_{t+1}\) is obtained. The immediate reward \(r_t\) is calculated by equation (20).
Step 4: Judge whether \( s_{t+1} \) satisfies the condition of terminating the current epoch. The samples are stored in the experience pool in the form of current state \( s_t \), action \( a_t \), next state \( s_{t+1} \), immediate reward \( r_t \) and whether the ending conditions are met. The biggest priority in current experience pool is given to the new sample and the Sumtree is updated.

Step 5: When the experience pool is not filled, jump to step 11, otherwise, continue.

Step 6: If the iteration number \( t \) reaches the update period of the target Q-network \( T_{tar} \), the target Q-network is updated, otherwise it is not updated.

Step 7: According to the priority experience replay based on Pareto samples, the confidence is calculated by equation (11) and the sample priority is adjusted. Screening Pareto samples from different segments in experience pool, the training set is formed by the samples collected from Pareto sample set and non-Pareto sample set based on Sumtree and priority \( p_i \).

Step 8: The TD-error of the training set and the target value \( y_i \) are calculated based on the valuation Q-network, the target Q-network and equations (7) and (14).

Step 9: The importance sampling weight \( w_i \) of each sample is calculated according to equations (16) and (17). The loss function \( L(\theta_{val}) \) is calculated according to equation (15). And the parameters of Q-network are updated by gradient back propagation of neural network.

Step 10: Update the TD-error of training set, the priority of experience pool and Sumtree by equations (8) and (10).

Step 11: \( s_t \leftarrow s_{t+1} \).

Step 12: If the number of network training epoch reaches \( T \), exit the algorithm, otherwise, return to Step 3.

EXPERIMENT RESULTS AND PERFORMANCE ANALYSIS

The parameters of signal and noise in simulation are set as follows: the total bandwidth \( W=200\)MHz, the transmission power of user \( P =100\)mW, the power of Gaussian white noise \( n_0 = 10^{-7} \) mW, the optional hopping rate set is \([500,1000,2000,4000]\) hop/s, the optional frequency interval set is \([1,2,3,4]\) MHz. The parameters of the proposed algorithm are set as follows: the capacity of experience pool \( N=2000 \), the capacity of training set \( M=50 \), the learning rate of network \( lr=0.0001 \), the update cycle of target Q-network \( T_{tar} =50 \), the total number of training epoch \( T =500 \), the discount factor \( \gamma =0.9 \) and \( \alpha=0.6 \), \( \beta=0.4 \), \( \lambda=20 \), \( \eta=75 \), \( \xi =500 \), \( z=150 \), \( G=20 \). As shown in Figure 3, to describe the complex electromagnetic environment more clearly, the spectrum information with temporal features, which is known as spectrum waterfall, is introduced. It indicates that there are several interferences which may change over time in the spectrum. The following simulation results are obtained under the interference environment in Figure 3, which includes wideband interference, narrowband interference, swept frequency interference and Gaussian white noise. The darker the color is, the greater the interference power is. The unit is mW.

The Performance of Screening From Experience Pool

In order to verify the superiority of sample selection of the Priority Experience Replay Based Pareto Sample (PPER) proposed in this paper, the PER, the Rank-based Priority Experience Replay (abbreviated as RankPER) proposed by Zhou Y. (2020), the Temporal Difference-error Classification (TDC) and the Reward Classification (RC) proposed by Shi, S. (2021) are used respectively to screen samples from the experience pool which generated randomly. The distribution of the samples selected by these methods is shown in Figure 4, where ‘*’ is the samples in the experience pool and the circle is the selected sample.
Figure 4 Shows the results of the screening samples.

It can be seen from Figure 4(a) that most of the samples screened by PPER are distributed in the areas with high immediate reward and priority. Because some of the samples are obtained by uniform random sampling method, there are also some samples located in other regions. It ensures the diversity of sampling result.

The PPER-DQN algorithm proposed in this paper, the DQN algorithm with Rank-based Priority Experience Replay (abbreviated as Rank-DQN), the Deep Q Network with Temporal Difference-error Classification (TDC-DQN) and the Deep Q Network with Reward Classification (RC-DQN) can be obtained by applying the above four sample selection methods into DQN algorithm. Their sample
selection situations during training are obtained and the number of samples of each frequency segment in the training set is computed. Then, the average priority, the average immediate reward and the distribution of training set are gotten. The results are shown in Table 1 and Figure 5.

Figure 5 shows the distribution of the training set.

It can be seen from Table 1 that in terms of priority, the average normalized priority of training set of PPER-DQN is significantly higher than that of experience pool, and the difference of the
average normalized priority between them is the largest. The average normalized priority difference between the training set and the experience pool of Rank-DQN is the second. And that of TDC-DQN and RC-DQN is the smallest. In terms of immediate reward, the average immediate reward of the training set of PPER-DQN is also significantly greater than that of the experience pool. The training set and experience pool of Rank-DQN and TDC-DQN only have a little difference in immediate reward. RC-DQN has great advantage in the gap between average immediate reward of training set and experience pool, but it has no advantage in priority.

As can be seen from Figure 5, samples distribution of the training set of PPER-DQN is more evenly distributed in each frequency band. Some samples are selected from non-Pareto samples based on priority, so the number of samples in some frequency bands is slightly more than that in other frequency bands. However, the training set samples of other algorithms distribute unevenly and are lack of diversity. According to Figure 3, it can be found that most of the samples of these training sets are selected from the frequency band with less interference power. This may make the network easier to overfit, and the performance of decision results in some frequency bands is poor.

To sum up, compared with other algorithms, the proposed PPER-DQN has great advantages in priority, immediate reward and sample distribution. It can screen samples with high priority and high immediate reward and improve the efficiency of the DQN algorithm effectively.
Performance Comparison

Under the same interference environment, five intelligent decision-making experiments of generating bivariate frequency hopping pattern are carried out respectively by using the PPER-DQN proposed in this paper, Rank-DQN, TDC-DQN, RC-DQN and basic DQN. The average SINR curve of each algorithm is shown in Figure 6.
Figure 6 shows the average SINR curve of five algorithms.

It can be seen from Figure 6 that the performance of PPER-DQN is the best. After convergence, its average SINR is higher than that of the other algorithms, and it has basically converged in the 160th training epoch. The convergence speed of Rank-DQN and TDC-DQN is faster than that of the basic DQN, and basically converges in the 220th and 200th training epoch respectively. But their average SINR still fluctuates after convergence. The convergence speed of RC-DQN is the slowest.
Performance Comparison of Frequency Hopping Patterns

Under the same interference environment, the average SINRs of the conventional frequency hopping patterns based on serval combinations of different hopping rates and frequency intervals and the
bivariate frequency hopping patterns generated by PPER-DQN (abbreviated as PDBi-FH Pattern) and random selected method (abbreviated as RBi-FH Pattern) are shown in Table 2.

It can be seen from Table 2 that the average SINRs of the three bivariate FH patterns generated by the PPER-DQN are higher and more balanced than that of the conventional FH patterns and the RBi-FH patterns. Because the frequency bands occupied by different patterns and their interferences are both not same, the performance of frequency hopping patterns generated by fixed hopping rate and frequency interval is quite different. Fixed parameters are not suitable to all frequency bands. The parameters of the RBi-FH patterns are selected randomly, so their performances are only better than most of that of conventional FH patterns, not all of them. It is hard to ensure the parameters are suitable enough all the time. And the PPER-DQN can make intelligent decision, generate excellent bivariate frequency hopping pattern, and its parameters have better adaptability to different frequency bands.

**CONCLUSION**

This paper mainly has studied the intelligent decision-making algorithm of using deep reinforcement learning to general bivariate frequency hopping pattern in frequency hopping communication system under complex electromagnetic environment. The priority experience replay based on Pareto samples has been used to select samples according to their TD-error and immediate reward. The training set has been formed based on Pareto sample set and random samples at the same time. It ensures the diversity of training set and improves the efficiency of sample utilization and experience replay. The confidence parameter has been proposed and works in the process of getting Pareto samples to avoid

<table>
<thead>
<tr>
<th>Type</th>
<th>Hopping Rate /hop/s</th>
<th>Frequency Interval /MHz</th>
<th>Average SINR/dB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pattern1</td>
</tr>
<tr>
<td>Conventional FH pattern</td>
<td>500</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>13.3723</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>13.4364</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>13.5136</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>13.6714</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>7.1293</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>7.5968</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>7.9104</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>8.1296</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2000</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>8.4468</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>8.6988</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>8.8345</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>8.9420</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4000</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>8.1848</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>8.2550</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>8.3811</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>8.5231</td>
</tr>
<tr>
<td>RBI-FH Pattern</td>
<td>[500,1000,2000,4000]</td>
<td>[1,2,3,4]</td>
<td>10.2183</td>
</tr>
<tr>
<td>PDBi-FH Pattern</td>
<td>[500,1000,2000,4000]</td>
<td>[1,2,3,4]</td>
<td>13.8534</td>
</tr>
</tbody>
</table>
the sample with too long storage period is selected as Pareto samples. It improves the superiority, learning value of the whole training set and the learning efficiency of DQN algorithm. The simulation results show that this method can selected samples more efficiently, improve the convergence speed of the algorithm. And the performance of the bivariate frequency hopping pattern is good. However, it is also more complex than the traditional frequency hopping pattern which will increase the difficulty of synchronization between receiver and sender. The receiver with the bivariate frequency hopping pattern need to be further studied.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundations of China under grant nos. U19B2016 and Key Lab of Data Storage and Transmission Technology of Zhejiang Province, Hangzhou Dianzi University.

FUNDING INFORMATION

The publisher has waived the Open Access Processing fee for this article.
REFERENCES


Shilian Zheng received the B.S. degree in telecommunication engineering and the M.S. degree in signal and information processing from Hangzhou Dianzi University, Hangzhou, China, in 2005 and 2008, respectively, and the Ph.D. degree in communication and information system from Xidian University, Xi’an, China, in 2014. He is currently an Associate Researcher with Science and Technology on Communication Information Security Control Laboratory, Jiaxing, China. His research interests include cognitive radio, spectrum management, and deep learning-based radio signal processing.