A Q-Learning-Based Channel Selection and Data Scheduling Approach for High-Frequency Communications in Jamming Environment

Wen Li[1], Yuhua Xu[1(✉)], Qiuju Guo[2], Yuli Zhang[3], Dianxiong Liu[1], Yangyang Li[1], and Wei Bai[1]

[1] College of Communications Engineering, Army Engineering University of PLA, Nanjing 210000, China
wen-li13@outlook.com, yuhuaenator@gmail.com, dianxiongliu@163.com, 15651858962@163.com, baiweiaeu@163.com

[2] PLA 75836 Troops, Guangzhou 510000, China
dolly517@163.com

[3] National Innovation Institute of Defense Technology, Academy of Military Sciences PLA China, Beijing 100000, China
yulipkueecs08@126.com

Abstract. The existence of jammer and the limited buffer space bring major challenge to data transmission efficiency in high-frequency (HF) commuication. The data transmission problem of how to select transmission strategy with multi-channel and different buffer states to maximize the system throughput is studied in this paper. We model the data transmission problem as a Makov decision process (MDP). Then, a modified Q-learning with additional value is proposed to help transmitter to learn the appropriate strategy and improve the system throughput. The simulation results show the proposed Q-learning algorithm can converge to the optimal Q value. Simultaneously, the QL algorithm compared with the sensing algorithm has better system throughput and less packet loss.

Keywords: Anti-jamming · Dynamic spectrum access · Q-learning · High-frequency(HF) communication · Markov decision process (MDP)

1 Introduction

The high-frequency (HF) (3–30 MHz) communication which is mainly used in transmitting important telegram and low bit-rate speech and image, plays a significant role in military, disaster relief and long voyage [1–3]. The main challenge

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

X. B. Zhai et al. (Eds.): MLICOM 2019, LNICST 294, pp. 145–160, 2019.
https://doi.org/10.1007/978-3-030-32388-2_13

for data transmission in HF networks is to find appropriate channel selection strategy. This is for several reasons. First, the available HF spectrum resources are limited due to narrow bandwidth and multi-user accessing. Second, the time-varying characteristic which caused by the ionospheric variation makes the communication unstable. Third, there exists different kind of jamming including natural and malicious jamming with development cognitive radio [4,5] and the intelligent technologies [6,7], the jammer becomes more and more intelligent. In this paper, we mainly consider the data transmission problem in jamming environment.

The traditional anti-jamming methods in HF networks mainly include power control [8–10], frequency hopping (FH) [11,12] and automatic link establishment (ALE) [13,14]. The power control enhances communication performance by the game theory to find the optimal Nash Equilibrium(NE) point. The FH extensively used in real equipments switches in several frequencies to reduce the influence of fading and jamming. Now, the adaptive FH [11] and intelligent FH [12] technologies have attracted great attention. The power control and frequency hopping, however, are not able to deal with the intelligent jamming. The automatic link establishment (ALE) has been developed to the fourth generation, which aims to make link establishment more intelligent and faster. However, the ALE technology becomes weak, when the state of environment changes rapidly.

To cope with the complicated jamming environment, the reinforcement learning which can interact with environment and learns to get action by the reward, has attracted lots of attention [15–20]. [15] uses the Q-learning to fight against the sweep jamming considering the Markov channel model, and the simulation result shows that the agent can avoid jamming totally. [16,17] have settled the intelligent jamming by the reinforcement learning. However, the reinforcement learning is nor able to resist the complex and changeable jamming. The deep reinforcement learning (DRL) combining the deep learning and reinforcement learning is proposed to deal with above challenge [18–20]. In [18,19], the input of deep neural network uses the spectrum waterfall, and then acquires the optimal anti-jamming decision by training.

Most of existing anti-jamming studies only considered how to find the idle channels to avoid jamming and assume that the time of each data transmission is fixed. They ignored the transmission demand and the limited buffer space. In actual networks, the agent would send appropriate data packets according to the buffer and the environment state. The agent will send data packets as much as possible in the time gap between previous jamming and next jamming. Therefore, the agent should not only choose the idle channel, but also decide how many packets should send. Currently, many literatures [21,22] have studied data transmission problem in an unknown environment. However, they do not consider the existence of malicious jamming. Therefore, it is a meaningful task to solve the data transmission problem in jamming environment.

In this paper, we study the data transmission problem for high-frequency communication in the jamming environment using a modified Q-learning method. The problem is challenging due to following reasons: (1) the time-varying channels; (2)

the existence of malicious jamming; (3) the limited buffer space. The communication probability is proposed to deal with the time-varying characteristic in [23]. Motivated by [15,21], a modified Q-learning algorithm is proposed, which considers the balance of exploration and exploitation to optimize data transmission. Different from [15], the new system state is defined including the previous transmission channel, the current jamming channel, and the number of data packets in buffer. The state transmission is formulated as a markov decision process(MDP), which aims to maximize the throughput. Simulation results show that the modified Q-learning algorithm can avoid jamming effectively and data transmission compared with the sensing algorithm.

The main contributions of this paper are summarized as follow:

- The data buffer is considered and the time of each transmission is not fixed. The data transmission problem in HF jamming environment is formulated as a MDP, in which the new state contains the previous transmission channel, the current jamming channel, and the number of data packets in buffer is used.
- We proposed a modified Q-learning algorithm to solve data transmission problem. The modified Q-learning balances exploration and exploitation of action selection, which reduces the convergence time to optimal Q value.

The rest of the paper is organized as follows. In Sect. 2, the system model is introduced, and the data transmission problem is formulated as a MDP problem. The Q-Learning-based data transmission scheme which proves to converge to the optimal strategy is proposed in Sect. 3. Section 4 gives the simulation results and analysis. Finally, we draw a conclusion in Sect. 5.

2 System Model and Problem Formulation

2.1 System Model

As depicted in Fig. 1, we consider a high-frequency (HF) communication system which is composed of a transmitter, a receiver and a jammer. The point-to-point from the transmitter to receiver is considered and there are M available channels denoted by $\mathcal{M} = \{1, 2, \cdots, M\}$. The jammer intending to damage the point-to-point communication generates jamming signals in modes like comb, sweeping and intelligence. We assume that the transmitter and the jammer keep the transmitting power unchanged all the time. The data packets generated according to task demands are stored in the buffer. The maximum length of the buffer is L. We assume that the arriving data packets follow the Poisson distribution with the arrival rate λ. When the buffer is full, the packets arrive later will be lost. When one packet is jammed by the jammer, the receiver will not get this packet and then tell the transmitter to send it again. The transmission schedule is decided by the transmitter based on the channel and buffer state. The transmitter can get the current jamming channel by wide band spectrum sensing(WBSS). In each transmission, the transmitter selects a channel and sends several packets to the receiver. The transmitter must comprehensively consider the channel state and the number of packets in the buffer.

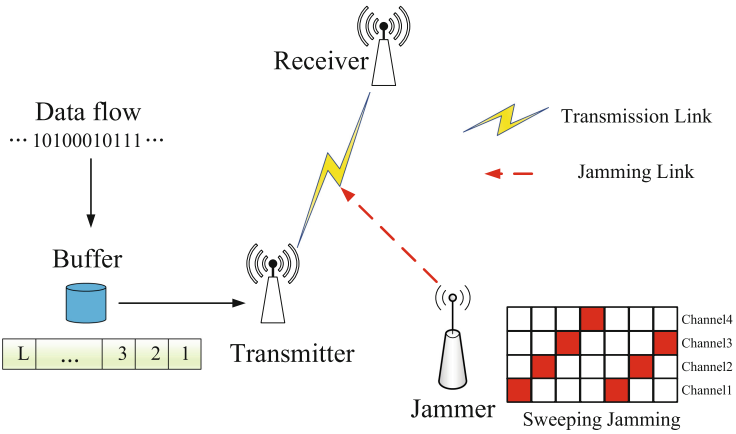

Fig. 1. System model.

Channel State. The HF communication achieves the long-distance depending on ionosphere to reflect signals. The ionosphere, however, is influenced by various factors like season, weather, location and solar activity [1]. The above factors make the HF channel time-varying and hard to predict. Therefore, the HF channel state is hard to be modeled as a Markov chain which is widely used in other literatures [24,25]. As shown in Fig. 2, the transmission will fail, when the channel is deep fading or jammed by jammer. For example, the channel 2 is unavailable in time-slot 5 with deep fading, 5 and 7 with jamming. Motivated by [23], the communication probability of channel is defined to describe the communication performance. It is a statistical concept, which can be calculated by long-time observation. The communication probability of the M available channels is denoted by $P = \{p_1, p_2, \cdots, p_M\}$, where p_i means that the data is transmitted successfully with p_i, when the transmitter chooses channel i.

Buffer State. The arriving packets follow the Poisson distribution with the arrival rate λ, which means that there are d_k arriving packets with probability $P(d_k) = e^{-\lambda T_k}(\lambda T_k)^{d_k}/d_k!$ in k-th transmission, where T_k is k-th transmission time. We assume the buffer length is l_k at the beginning of k-th transmission. If the number of transmitted packets is l_k^T, the number of arriving packets is l_k^A and the number of packets which are jammed or suffer deep fading is l_k^J, then the buffer length after k-th transmission is

$$l_{k+1} = \min(L, l_k + l_k^A + l_k^J - l_k^T), \tag{1}$$

where L is the maximum length of buffer. Since the buffer space is limited, if the number of packets is more than L, packet loss happens. It is noted that the buffer state of $k+1$-th transmission is only associated with the k-th buffer state. Thus, the buffer state is a Markov state and the transition probability is denoted

by $p(l_{k+1}|l_k, b_k)$, where b_k is the of number of packets selected to transmit in k-th transmission.

2.2 Problem Formulation

In this section, we formulate the data transmission problem in HF networks as a Markov decision process(MDP) and give the explanation of the state, the action and the utility.

We consider the data transmission with limited buffer space in jamming environment. In Fig. 2, let the available channels $M = 3$, the buffer length $L = 6$ and the jammer generates the sweeping jamming. We assume the time-slot of jamming denoted by T_J and the time to transmit each packet are fixed. However, the transmitter can choose different number of packets to transmit in each transmission according to the buffer state. As shown in Fig. 2, there is a gap between adjacent interference in the same channel. For the transmitter, it would like to send all its packets in the buffer if it can find the gap. However, the gap it choose may not be enough for all packets transmission, and then the jamming happens which makes it has to retransmit these packets. Thus, in each transmission, the transmitter has to choose a better channel which can support more packets be transmitted without being jammed.

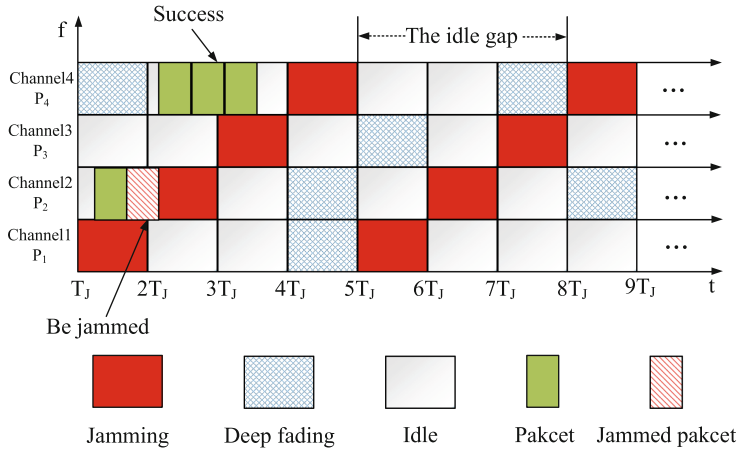

Fig. 2. The data transmission process in HF networks.

The system state in k-th transmission is defined as $s_k = (f_k^n, f_k^J, l_k)$, where f_k^n is the communication channel in last transmission, f_k^J is the jamming channel obtained by WBSS at the beginning of the transmission and l_k is the current buffer length. The process of data transmission in HF networks is actually a process of state transition. It is obvious that the next system state s_{k+1} is obtained, after the transmitter executes an action according the current state s_k. The next

state is only associated with the current state and previous states have no effect on it, which can be expressed as

$$p(s_{k+1}|s_k, s_{k-1}, \cdots, s_1) = p(s_{k+1}|s_k), \tag{2}$$

where $p(\cdot)$ is the transition probability. Therefore, we can model the problem of data transmission as a Markov Decision Process (MDP) [26].

Action Set. The transmitter has to select an action which contains the channel and the number of packets at the beginning of k-th transmission. The channel selection is denoted by $c_k \in \{1, 2, \cdots, M\}$ and the packets to transmit is $b_k \in \{0, 1, 2, \cdots, l_k\}$, where l_k is current buffer length. For easy analysis, we map the two actions to a new action a_k, i.e., $f : (c_k, b_k) \rightarrow a_k$. The map f is expressed as $a_k = f(c_k, b_k) = b_k \cdot M + c_k$. Therefore, we denote the action set as $\mathcal{A} = \{1, 2, \cdots, (L+1)M\}$.

System Utility. In this paper, our goal is to maximize the system throughput. In the state s_k, we assume the number of packets which are not jammed is n_{succ} after taking action $a_k = f(c_k, b_k)$. Since the channel is unstable, the average packets transmitted successfully is denoted by

$$N(s_k, a_k) = p(c_k) \times n_{succ}, \tag{3}$$

where $p(c_k)$ is the communication probability of channel c_k. more packets are transmitted successfully, the larger system throughput is. Thus, the system utility is proportional to the average packets.

The more packets are in the buffer, the arriving packet may be lost with larger probability because of the limited buffer space. We define the pressure value of the buffer as $f(s_k, a_k) = \exp(\theta \times l_k)$, where θ is the pressure coefficient [21]. The less pressure value means less packets loss. Therefore, the system utility is inversely proportional to the buffer pressure. At the same time, the number of jammed packets is denoted by n_{Jam}. We define the jamming value as $J(s_k, a_k) = \exp(\beta \times n_{Jam})$. Thus, combining the buffer pressure and the jamming degree, the loss of data transmission is expressed as

$$H(s_k, a_k) = f(s_k, a_k) \times J(s_k, a_k). \tag{4}$$

The system utility, which is related to the packets transmitted successfully and the transmission loss, is described as

$$u_k = u(s_k, a_k) = N(s_k, a_k)/H(s_k, a_k). \tag{5}$$

In this paper, we want to maximize the throughput performance of the HF networks by online learning method. The action a_k is related to the history data transmission strategies $\{a_1, a_2, \cdots, a_{k-1}\}$ and history utility information

($\{u_1, u_2, \cdots, u_{k-1}\}$). Our problem is to find the optimal data transmission strategy to maximize the cumulative expected throughput [15]

$$P : \max E[\sum_{i=1}^{k} u_i(a_i)], a_i \in \mathcal{A}. \tag{6}$$

According to previous analysis, the data transmission problem is modeled as a MDP problem. The reinforcement learning(RL) which interacts with the environment to find the optimal action is widely used for the MDP problem [15–19]. As the system state and the action are discrete, the Q-learning is suitable to solve the data transmission problem. In the next section, we will propose a modified Q-learning algorithm and prove the convergence of it.

3 Q-Learning-Based Data Transmission Scheme

The Q-learning algorithm interacts with the environment and learn to obtain the optimal action in a online-learning way. The Q-value table is used to evaluate the performance of the action. In the state s_k, the agent takes an action a_k according to the Q-value table, then, it obtains instantaneous reward r_k and switch to next state s_{k+1}. At the same time, it updates the Q-value table. The more detailed explanation about Q-learning can be found in [27].

In the learning process, the agent interacts with the environment to find the optimal actions, considering the immediate reward and the future rewards. The discounted future rewards under a policy π is defined as

$$V^{\pi}(s_k) = \sum_{j=k}^{+\infty} \gamma^{j-k} u_j, \tag{7}$$

where $0 < \gamma < 1$ is the discount factor. Then, the corresponding Q value can be formulated as

$$Q(s_k, a_k) \leftarrow r_k + \gamma V^{\pi}(s_{k+1}). \tag{8}$$

Our goal is to maximize the discounted utility. According to the Bellman equation [18], the Q value by replacing the r_k and $V^{\pi}(s_{k+1})$ can be expressed as

$$Q(s_k, a_k) \leftarrow u_k + \gamma \max_{a_{k+1}} Q(s_{k+1}, a_{k+1}). \tag{9}$$

Different from [15] and [18] in which the number of available actions is small, the number of the actions in this paper is calculated as $M(L+1)$. Since the action set is large, the normal Q-learning may converge to the optimal Q value with large steps. Motivated by [21], we proposed a modified Q-learning algorithm in jamming environment. It is important to balance the exploration and exploitation of the large action set for Q-learning algorithm. In order to choose the action effectively, a additional value is added to find the optimal action quickly [21]. The additional value can reduce the convergence time by taking advantage of

history rewards and adjusting the explore range. The action a_k is selected by following equation,

$$a_k = \arg\max_a (Q(s_k, a) + Add(s_k, a)). \tag{10}$$

The additional value $Add(s_k, a)$ which can help to find the optimal with less learning steps is expressed as

$$Add(s_k, a) = C_p \sqrt{2 \ln k \times \min\{1/4, V_a(k)\}/T_a(k)}, \tag{11}$$

where C_p is a greater than zero [28], and $T_a(k)$ is the number of times that action a has been executed after k transmissions. $V_a(k)$ is the bias factor, which is defined as

$$V_a(k) = \sigma_a^2(k) + \sqrt{2 \ln k / T_a(k)}, \tag{12}$$

where $\sigma_a^2(k)$ is the utility variance. The variance can reflect the volatility of the action. It can be calculated by

$$\sigma_a^2(k) = \sum_{i=1}^{T_a(k)} u^2(s_i(a), a)/T_a(k) - (\sum_{i=1}^{T_a(k)} u(s_i(a), a)/T_a(k))^2, \tag{13}$$

where $s_i(a)$ is the i-th of states which have selected the action a. The above action selection method with the additional value makes the best of history rewards and chooses the action with larger reward, which is the exploitation characteristic of the system. Simultaneously, it will explore the action which is not selected or rarely selected, which reflects the exploration of the system.

When the transmitter chooses an action a_k according to the state s_k and the above method, it obtains the reward $r_k = u_k$, and then it updates the Q values as follow:

$$Q_{k+1}(s_k, a_k) = (1 - \alpha) Q_k(s_k, a_k) + \alpha(r_k + \gamma \max_a (Q_k(S_{k+1}, a)), \tag{14}$$

where $\alpha (0 < \alpha \leq 1)$ is the learning rate which is defined as $\alpha = 1/(1 + T_{a_k}(k))$, and γ is the discount factor.

In the jamming environment, the transmitter performs the modified algorithm to adjust the transmission strategy at the beginning of each transmission. The Fig. 3 shows the time-slot of the modified Q-learning algorithm. At the beginning of k-th transmission, the current state $s_k = (f_k^n, f_k^J, l_k)$ is obtained according the last transmitting channel, the jamming channel by WBSS and the buffer length. In the T_A period, the receiver feedbacks the jammed packets n_J to the transmitter by ACK. The current jamming channel f_{k+1}^J is obtained by WBSS in the T_W period. During the T_L, the transmitter observes the buffer length l_{k+1} and obtains the next state $s_{k+1} = (f_{k+1}^n, f_{k+1}^J, l_{k+1})$. At the same time, it updates the Q values according to (14). The detailed process of the modified QL algorithm is shown in Algorithm 1.

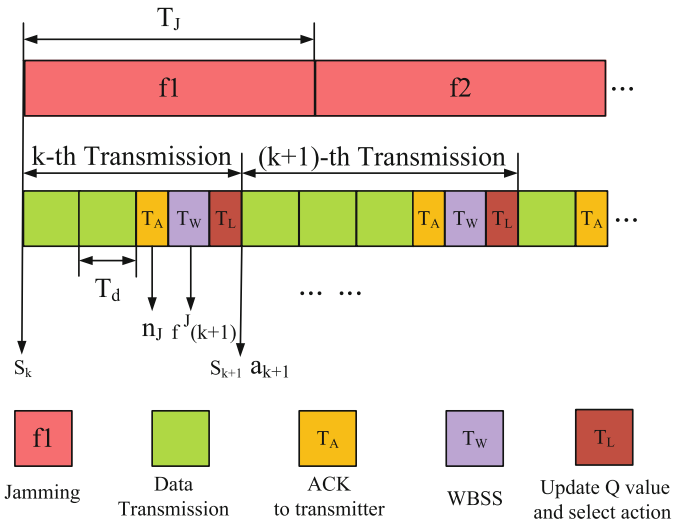

Fig. 3. The time-slot structure of the modified QL algorithm.

Algorithm 1. The modified Q-learning-based HF data transmission algorithm

1: Set parameter γ, simulation time K, and the time index $k = 0$.
2: Initialize the action recording vector $T_k = 0$ and the Q value $Q(s, a) = 0$.
3: Initialize the jamming channel f_0^J by WBSS and the buffer length $l_0 \leq L$, choose the initial transmitting channel f_0^n, and acquiring the initial state $s_0 = (f_0^n, f_0^J, l_0)$.
4: While $k < K$, do
5: if $k < 1000$
6: Select an action a_k randomly.
7: else
8: Calculate the additional value
$$Add(s_k, a) = C_p \sqrt{2 \ln k \times \min\{1/4, V_a(k)\}/T_a(k)}.$$
9: Select action according to (10)
$$a_k = \arg \max_a (Q(s_k, a) + Add(s_k, a)).$$
10: end
11: Upgrade the action recording vector $T_{a_k}(k) = T_{a_k}(k) + 1$.
12: Execute a_k, and obtain the reward r_k based on (5).
13: Receive the ACK from the receiver, calculate the buffer length l_{k+1}.
14: Obtain f_{k+1}^n according to a_k and the f_{k+1}^J by WBSS, Then the next state is $s_{k+1} = (f_{k+1}^n, f_{k+1}^J, l_{k+1})$.
15: Calculate $\alpha = 1/(1 + T_{a_k}(k))$.
16: Update $Q_k(s_k, a_k)$
$$Q_{k+1}(s_k, a_k) = (1 - \alpha) Q_k(s_k, a_k) + \alpha(r_k + \gamma \max(Q_k(S_{k+1}, a_{k+1})).$$
17: $k = k + 1$
18: End while

4 Simulation Results and Discussion

In this section, we define the parameters of the HF network and study the performance of the proposed algorithm. In the simulation, a network containing a jammer, a receiver and a transmitter is considered. The length of buffer in the transmitter is $L = 7$. The number of available HF channels is $M = 4$. We assume the jammer generates the sweeping jamming. The data transmission performance is compared with the sensing algorithm. The detailed simulation parameters are shown in Table 1.

Table 1. Simulation parameters.

Parameters	Value
Number of available channels	$M = 4$
Buffer length	$L = 7$
Channel communication probability	$P = [0.8, 0.85, 0.9, 0.95]$
Jammer time-slot $/ms$	$T_{jam} = 2$
Transmission time of each packet $/ms$	$T_d = 0.8$
ACK transmission time $/ms$	$T_{ACK} = 0.1$
WBSS time $/ms$	$T_{WBSS} = 0.2$
Simulation steps	$K = 25000$
Buffer pressure coefficient	$\theta = 0.5$
Jammed pakcet press coefficient	$\beta = 0.5$
Arrive rate	$\lambda = [0.6, 0.7, \cdots, 1.3]$
Learning rate	$\alpha = (0, 1]$
Discount factor	$\gamma = 0.8$
Index weight	$C_p = 1/\sqrt{2}$
Transmission power of each packet $/mw$	$P_{signle} = 0.3$

Figures 4 and 6 show the time-frequency diagram of the transmitter and the jammer at the initial and convergent stage, respectively, in which the red squares represent the sweeping jamming, the green squares are the data transmission and the blue squares are the WBSS and ACK transmission. As shown in Fig. 4, at the initial stage, there are mass of the overlapping squares which represent data transmission being jammed. At the same time, Figs. 5 and 7 show the buffer length after each transmission. It can be noted that the pressure of the buffer is large because of the jammed packets. However, it is noted that the transmitter can choose right transmission action to avoid the jamming after the Q-learning stage, which is depicted in Fig. 6. At the same time, The pressure of buffer is small after the learning phase which is shown in Fig. 7.

Figure 8 shows the Q value changing curve in the learning process at the state $s(1, 2, 4)$ for different actions. From the figure, we can see that the Q value

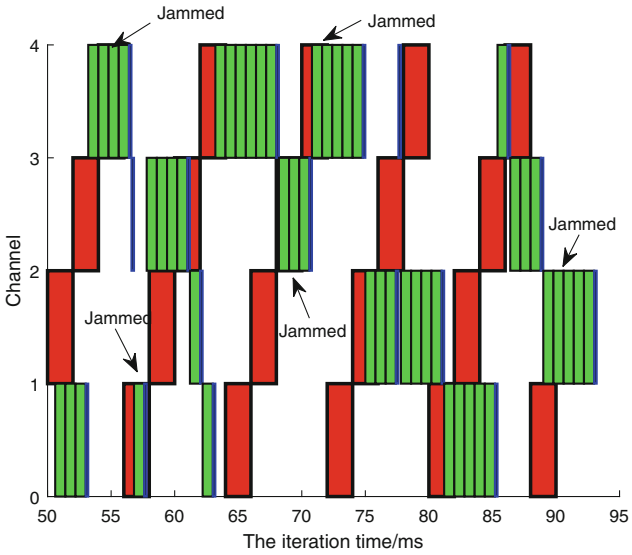

Fig. 4. Time-frequency diagram at initial stage. (Color figure online)

curve converges to a stable value. At the sate in which last transmission is in 1-th channel, the jammer is in 2-th channel and the buffer length is 4, the transmitter learns the jamming pattern and acquires the optimal action. As shown in figure, the transmitter will select the action $a = (4, 4)$ which sends 4 packets in the

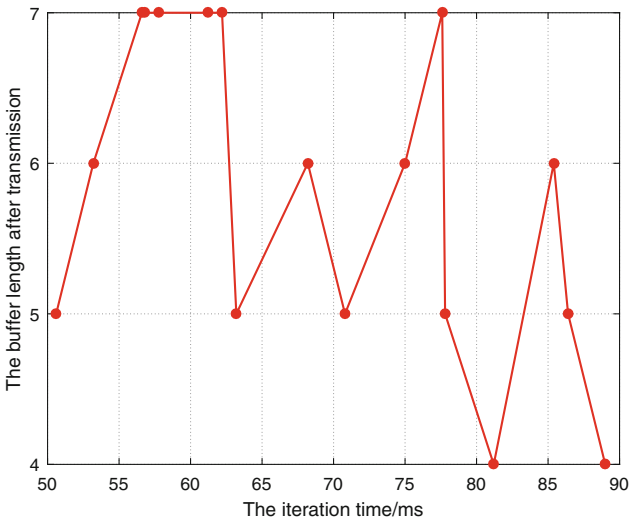

Fig. 5. The buffer state at initial stage.

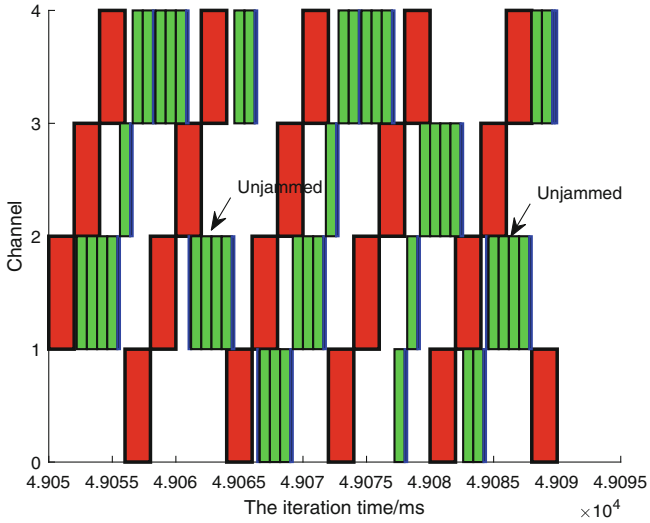

Fig. 6. Time-frequency diagram at convergent stage. (Color figure online)

channel 4. It finds the optimal action in the state $s(1, 2, 4)$. The convergence of the algorithm is verified.

Figure 9 shows the system throughput under two different algorithms containing the proposed QL algorithm and the sensing algorithm. The sensing algorithm is that each transmission randomly selects an action according to the

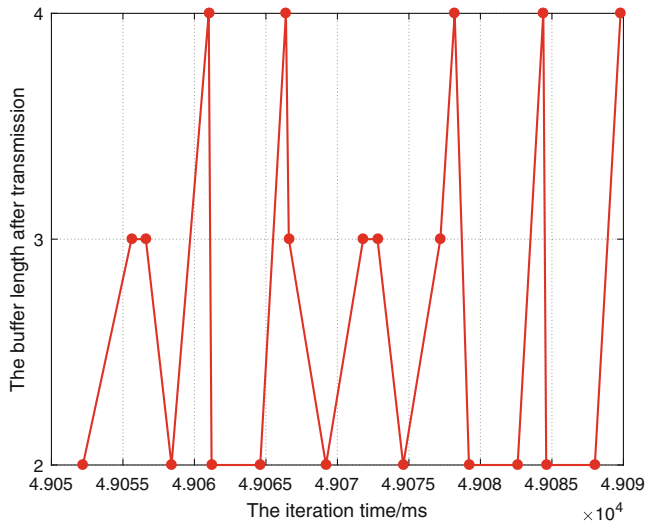

Fig. 7. The buffer state at convergent stage.

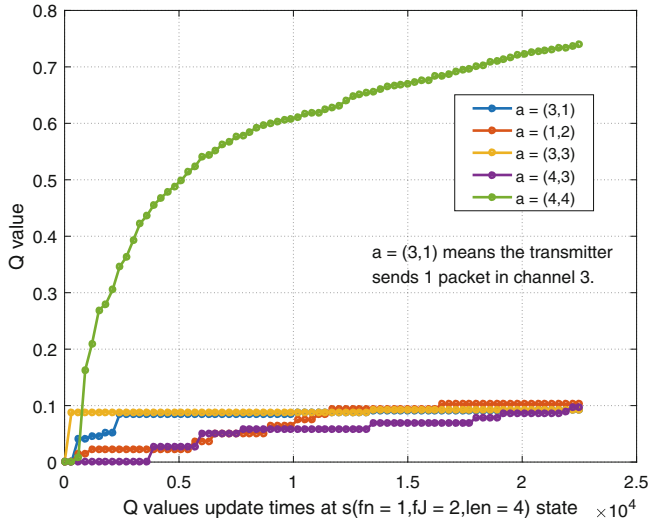

Fig. 8. Q value changing curve at the state $s = (1, 2, 4)$ with different actions.

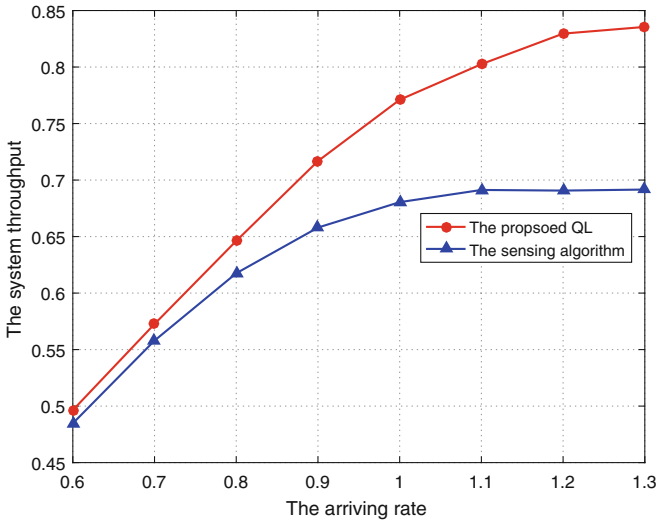

Fig. 9. Throughput comparison of different algorithms.

current sensing result. The throughput is calculated after each 100 transmissions, which is the ratio between sum of packets transmitted successfully and the total time. From the Fig. 9, we can find that the two algorithm have almost equivalent system throughput with small packet arriving rate. Because the pressure of the buffer is small, there are not enough packets to transmit. With the packet arriving rate increasing, the number of packets in the buffer gradually

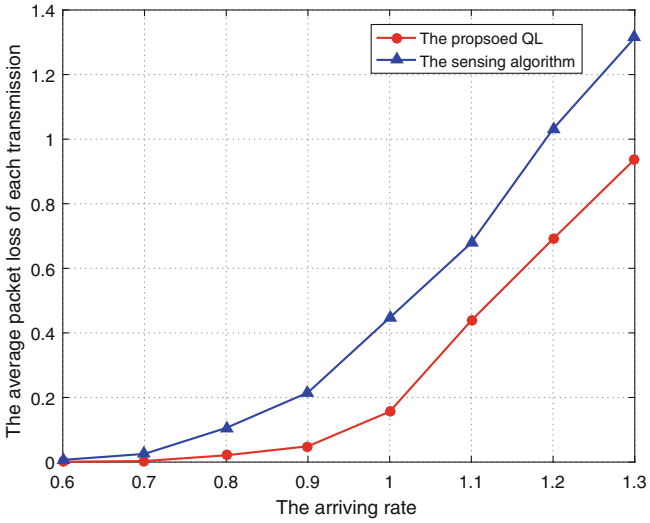

Fig. 10. Average packet loss of different algorithms.

increases which brings larger system throughput. Since the QL algorithm can learn to select appropriate action by interacting with the jamming environment, it has a better system throughput performance in the unknown and jamming environment.

Since the buffer space is limited, more packets will be lost, when the arriving rate of packet is large. As shown in Fig. 10, the average packet loss of each transmission is compared with different algorithms. With the arriving rate increasing, the packet loss is growing and it is almost linear. Because the sensing algorithm chooses the action randomly according to current sensing result, there are more jammed packets which make more packet loss than the QL algorithm.

5 Conclusion

In this paper, we considered the data transmission problem with jamming environment in the HF environment. To cope with the unstable characteristic of HF channel, the communication probability was used. A modified Q-learning algorithm has been proposed to optimize the strategy selection and achieve better communication performance. The data transmission problem in the jamming environment which was formulated as a MDP problem, was solved by the proposed algorithm. The proposed algorithm adding the additional value could balance the exploration and exploitation of action. The simulation results confirmed the convergence of the proposed QL algorithm and indicated that the QL had higher system throughput and less packet loss than the sensing algorithm. This paper only considered sweeping jamming. In the next step, the intelligent jammer will be further studied.

References

1. Wang, J.: Research and Development of HF Digital Communications. Science Press, Beijing, China (2013)
2. Hanson, R.: Military applications of HF communications. Mil. Technol. **10**, 70–77 (2010)
3. Xu, K., Jiang, B., Su, Z., et al.: High frequency communication network with diversity: system structure and key enabling techniques. China Commun. **15**(9), 46–59 (2018)
4. Haykin, S.: Cognitive radio: brain-empowered wireless communications. IEEE J. Sel. Areas Commun. **23**(2), 201–220 (2015)
5. Yucek, T., Arslan, H.: A survey of spectrum sensing algorithms for cognitive radio applications. IEEE Commun. Surv. Tutor. **11**(1), 116–130 (2009)
6. Bkassiny, M., Li, Y., Jayaweera, S.K.: A survey on machine learning techniques in cognitive radios. IEEE Commun. Surv. Tutor. **15**(3), 1136–1159 (2013)
7. Luong, N.C., Hoang, D.T., Gong, S., et al.: Applications of deep reinforcement learning in communications and networking: a survey. arXiv:1810.07862
8. Yang, D., Xue, G., Zhang, J., Richa, A., Fang, X.: Coping with a smart jammer in wireless networks: a Stackelberg game approach. IEEE Trans. Wirel. Commun. **12**(8), 4038–4047 (2013)
9. Jia, L., Xu, Y., Sun, Y., Feng, S., Anpalagan, A.: Stackelberg game approaches for anti-jamming defence in wireless networks. IEEE Trans. Wirel. Commun. **25**(6), 120–128 (2018)
10. Feng, Z., Ren, G., Chen, J.: Power control in relay-assisted anti-jamming systems: a Bayesian three-layer stackelberg game approach. IEEE Access **7**, 14623–14636 (2019)
11. Michael, R.M., et al.: Adaptive coding for frequency-Hop transmission over fading channels with partial-Band interference. IEEE Trans. Commun. **59**(3), 854–862 (2011)
12. Duan, R.J., et al.: Research on spectrum allocation of HF access network based on intelligent frequency hopping. In: 8th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, pp. 295–300 (2015)
13. Johnson, E., et al.: Third-Generation and Wideband HF Radio Communications. Artech House, Norwood (2013)
14. Wang, J., Ding, G., Wang, H.: HF communications: past, present, and future. China Commun. **15**(9), 1–9 (2018)
15. Kong, L., Xu, Y., Zhang, Y., et al.: A reinforcement learning approach for dynamic spectrum anti-jamming in fading environment. In: IEEE 18th International Conference on Communication Technology (ICCT), Chongqing, pp. 51–58 (2018)
16. Slimeni, F., Chtourou, Z., Schaeers, B., Nir, V.L., Attia, R.: Cooperative Q-learning based channel selection for cognitive radio. Wirel. Netw. **4**, 1–11 (2018)
17. Xiao, L., Lu, X., Xu, D., Tang, Y., Wang, L., Zhuang, W.: Uav relay in vanets against smart jamming with reinforcement learning. IEEE Trans. Veh. Technol. **67**(5), 4087–4097 (2018)
18. Liu, X., Xu, Y., Jia, L., et al.: Anti-jamming communications using spectrum waterfall: a deep reinforcement learning approach. IEEE Commun. Lett. **22**(5), 998–1001 (2018)
19. Liu, X., Xu, Y., Cheng, Y., et al.: A heterogeneous information fusion deep reinforcement learning for intelligent frequency selection of HF communication. China Commun. **15**(9), 73–84 (2018)

20. Chen, Y., Li, Y., Xu, D., Xiao, L.: DQN-based power control for IOT transmission against jamming. In: IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, pp. 1–5 (2018)
21. Zhu, J., Song, Y., Jiang, D., Song, H.: A new deep-Q-learning-based transmission scheduling mechanism for the cognitive internet of things. IEEE Internet Things J. **5**(4), 2375–2385 (2018)
22. Lin, X., Tan, Y., Zhang, J.: A MDP-based energy efficient policy for wireless transmission. Commun. Netw. **36**(7), 1433–1438 (2014)
23. Li, W., Ruan, L., Xu, Y., et al.: Exploring channel diversity in HF communication systems: a matching-potential game approach. China Commun. **15**(9), 60–72 (2018)
24. Wang, H.S., Moayeri, N.: Finite-state Markov channel-a useful model for radio communication channels. IEEE Trans. Veh. Technol. **44**(1), 163–171 (1995)
25. Zhang, Q., Kassam, S.A.: Finite-state Markov model for Rayleigh fading channels. IEEE Trans. Commun. **47**(11), 1688–1692 (1999)
26. Monahan, G.E.: State of the art-a survey of partially observable Markov decision processes: theory, models, and algorithms. Manag. Sci. **28**(1), 1–16 (1982)
27. Luong, N.C., Hoang, D.T., Gong, S., et al.: Applications of deep reinforcement learning in communications and networking: a survey (2018). arXiv:1810.07862
28. Browne, C.B., Powley, E., Whitehouse, D., et al.: A survey of Monte Carlo tree search methods, **4**(1), 1–43 (2012)