Model-free Reinforcement Learning based Multi-stage Smart Noise Jamming

Yuanhang Wang, Tianxian Zhang*, Longxiao Xu, Tuanwei Tian, Lingjiang Kong and Xiaobo Yang

***Abstract*—In this paper, considering a multi-stage electronic countermeasure against a Fire Control Radar (FCR) with multiple working modes, an optimal multi-stage jamming power allocation for smart noise jamming is investigated with unknown environment model. The optimal jamming power allocation problem is solved by proposing a multi-stage jamming power allocation method based on model-free reinforcement learning. Firstly, we construct a Markov decision process (MDP) with unknown environment model for multi-stage jamming power allocation. To evaluate the performance of multi-stage smart noise jamming, we choose the expected of the search to lock-on time of the FCR as an evaluation criterion. Then, to overcome the challenge of the unknown environment model, a reinforcement learning framework for multi-stage jamming power allocation is formulated. Under the framework, a method of multi-stage jamming power allocation based on Q-learning is proposed. Finally, numerical results are provided to verify the validity of the proposed method.**

I. Introduction

Modern Fire Control Radar (FCR) has strong anti-jamming ability and multiple working modes [1]. Facing this kind of modern FCR, the performance of the conventional noise jamming is getting worse. That is because once the jamming is recognized, the FCR will take a lot of effective anti-jamming measures, such as frequency agility, pulse repetition period agility, etc. In this case, It is necessary to reduce the probability of jamming discovery. In these years, some researchers have focused on the smart noise jamming [2-7]. This jamming method makes jamming signal cover exactly real targets echo in time-domain. The generation of smart noise jamming waveforms is researched in [2]. To improve jamming performance, technology of smart noise jamming based on convolution modulation is researched in [4]. However, the above-mentioned works only focus the generation and design of jamming signal. They all ignore the allocation of jamming power. Actually, the probability of jamming discovery can be significantly reduced by continuous and effective dynamic power optimal allocation.

In real applications, a jammer is often faced with the multi-stage electronic countermeasure against a FCR with multiple working modes. In this case, the multi-stage jamming power optimal allocation is important for an effective smart noise jamming. As we know, the multi-stage jamming power allocation of smart noise jamming often relies on expert experience. However, the expert experience is often inaccurate, the inaccurate jamming power allocation will result in the loss of jamming performance: **1)** an overly small jamming power can not effectively degenerate the performance of the FCR. **2)** an overly large jamming power will increase the probability of jamming discovery. Therefore, an effective method of multi-stage jamming power allocation is necessary to be investigated for smart noise jamming. However, to the authors best knowledge, multi-stage jamming power allocation problem for smart noise jamming is not yet available in open literature. This is indeed the main topic of this paper.

In this paper, under the situation of multi-stage electronic countermeasure against a FCR with multiple working modes, an optimal multi-stage jamming power allocation for smart noise jamming is studied with unknown environment model. Firstly, the optimal multi-stage jamming power allocation problem is formulated as a Markov decision process with unknown environment model. To evaluate the performance of smart noise jamming, we choose the expected of the search to lock-on time of the FCR as a criterion. Then, in order to overcome the challenge of the unknown environment model, a method of multi-stage jamming power allocation based on model-free reinforcement learning is proposed. Finally, numerical results are provided to verify the validity of the proposed method.

The rest of this paper is organized as follows. In Section II, we construct the problem modeling and the evaluation criterion. In Section III, the principle of multi-stage smart noise jamming is introduced, the method of multi-stage jamming power allocation based on the model-free reinforcement learning is proposed. We show some simulation results in Section IV. Finally, In Section V we conclude this paper.

II. Problem Modeling

Assume that a targeted aircraft with a self-defence system is chased by an attacked aircraft with a FCR. The airborne self-defence system of the targeted aircraft contains a jammer and a radar warning receiver (RWR). The RWR can be used to intercept the signal parameters and recognize the working modes of the FCR [8]. The power of smart noise jamming can be adaptively adjusted based on the signal strength and the working modes of the FCR. Our concern is the policy of adaptive adjustment. The airborne FCR of the attacked aircraft has multiple working modes. Typical working modes

This work was supported by the Chang Jiang Scholars Program, the National Defense Science and Technology Innovation Special Zone Project, the Fundamental Research Funds of Central Universities under Grant ZYGX2018J009.

Y. Wang, T. Zhang, L. Kong and X.Yang are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (E-mail: tianxianzhang@gmail.com; txzhang@uestc.edu.cn).

(*Corresponding author: Tianxian Zhang.*)

978-1-7281-1679-2/19/$31.00 ©2019 IEEE

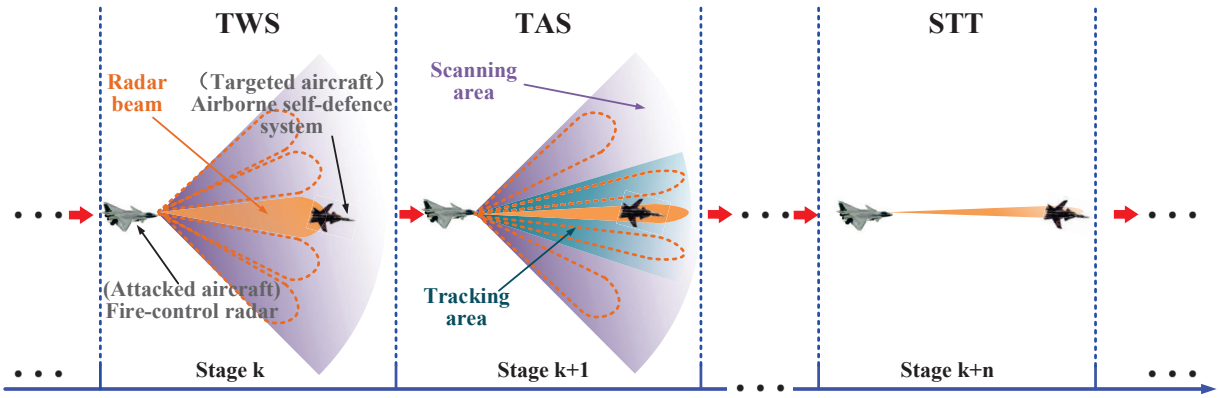

Fig. 1: The multi-stage electronic countermeasure between self-defence system and the FCR.

of airborne FCR contain track while scan (TWS) mode, track and search (TAS) mode, single target tracking (STT) mode, etc. During the electronic countermeasure, the FCR will switch its working mode as needed. For example, if the target is confirmed, the mode of the FCR will be transferred to the one that allocate more resources on the target tracking. If the target is lost, the mode of the FCR will be transferred to the initial mode. Anti-jamming measures will be taken in the FCR if the jamming is recognized.

As shown in Fig.1, the electronic countermeasure is divided into multiple stages due to the transition of working modes of the FCR. More specifically, the airborne self-defence system of the targeted aircraft and the FCR of the attacked aircraft interact at each of a sequence of stages, $k = 0, 1, 2, \cdots$, until the targeted aircraft is locked on by the FCR. At each stage k, the self-defence system receives the transmit signal of the FCR and recognizes the working modes $s_k \in \mathbf{S}$ of it, where $\mathbf{S}$ is the set of possible working modes of the FCR. Then, the jammer selects a jamming action $a_k \in \mathbf{A}$ corresponding to a jamming power and performs the smart noise jamming, where $\mathbf{A}$ is the set of jamming actions available. One stage later, the self-defence system receives a numerical evaluation t_k of the jamming performance in stage k, and recognizes the new working mode s_{k+1} of the FCR, where t_k is the time spent by the FCR in the kth stage.

Generally, the search to lock-on time of the FCR can be used as a criterion to evaluate the performance of jamming. The jammer's goal is to increase the search to lock-on time of the targeted aircraft as far as possible by establishing the optimal jamming actions, roughly speaking, is to maximize the expected value of the search to lock-on time of the FCR over the long run, which can be described as

$$\max_{\chi} \mathrm{E}\left(T\right) = \max_{\chi} \mathrm{E}\left(t_0 + t_1 + t_2 + \cdots\right), \tag{1}$$

where, T is the search to lock-on time of the FCR, $\mathrm{E}\left(\cdot\right)$ denotes the the expected value, $\chi = [a_0, a_1, a_2, \cdots]$ is the set of selected jamming actions at a sequence of stages, which can be called a jamming power allocation policy χ. It is not difficult to know that T is closely related to policy χ, but the relationship between them is affected by many factors, such as jamming recognition and anti-jamming measures of the FCR, noise, measurement errors, the rule of working mode transition of the FCR, etc. These factors are unknown to jammer, which can be called the envirnmental factors. Without loss of generality, the relationship between T and χ can be formulated as

$$T = \mathrm{F}\left(\chi, \text{envirnmental facters}\right). \tag{2}$$

Obviously, a bigger T means a better jamming performance. Therefore, to obtain bigger T, we need to find a better jamming power allocation policy $\hat{\chi}$, which can be described as

$$\hat{\chi} = \arg\max \mathrm{E}\left(\mathrm{F}\left(\chi, \text{envirnmental facters}\right)\right). \tag{3}$$

Generally, the working mode transition probability of the FCR is unknown, and the numerical evaluation of the jamming performance in each stage is uncertain, too. The optimal multi-stage jamming power allocation problem can be formulated as a Markov decision process with unknown environment model. There are two difficulties in solving such problem: **1)** the number of stages of the problem is uncertain, and the objective function is difficult to obtain. **2)** the environment model is unknown. Based on the above two points, traditional optimization algorithms such as evolutionary algorithms and dynamic programming can not solve the power allocation problem. Fortunately, model-free reinforcement learning is an effective tool for this kind of optimization problem, but the allocation of multi-stage jamming power based on this algorithm remains unexplored. The method of smart noise jamming based on model-free reinforcement learning will be proposed in the following section.

III. The Method of Smart Noise Jamming Based on Model-free Reinforcement Learning

This section contains two main parts: **1)** the principle of multi-stage smart noise jamming is introduced. **2)** the method of multi-stage jamming power allocation based on the model-free reinforcement learning is proposed.

978-1-7281-1679-2/19/$31.00 ©2019 IEEE

A. The principle of multi-stage smart noise jamming

Our proposed method of multi-stage smart noise jamming can be used for a variety of jamming signals, the power of smart noise jamming is allocated based on the signal strength and the working modes of the FCR. Without loss of generality, we assume that the transmitted power of the FCR is P_{tk}, the antenna gain is G_{tk} in stage k , the effective aperture is A_r. Therefore, the transmitted power density received by RWR is

$$D_k = \frac{P_{tk}G_{tk}}{4\pi R^2}, \tag{4}$$

where, R is the distance between the FCR and target. After the signal of the FCR is received, the working mode s_k is recognized. Based on D_k and s_k, the power of smart noise jamming for each pulse of the kth stage is determined as

$$\begin{aligned} P_{gk} &= a_k D_k \sigma_u \\ &= \frac{a_k P_{tk} G_{tk} \sigma_u}{4\pi R^2}, \end{aligned} \tag{5}$$

where, $a_k \in \mathbf{A}$ is a coefficient selected by the jammer, which is considered as the jamming action of the kth stage. σ_u is the average RCS of the target, which is considered as the estimate of the target's real-time RCS. We assume that the real-time RCS is $\sigma \sim U(\sigma_{\min}, \sigma_{\max})$. Actually, σ is unknown to jammer. For each pulse, the total return power received by the FCR can be described as

$$\begin{aligned} P_{rk} &= P_{rgk} + P_{rsk} + P_n \\ &= \frac{P_{tk}G_{tk}a_k\sigma_u A_r}{(4\pi R^2)^2} + \frac{P_{tk}G_{tk}\sigma A_r}{(4\pi R^2)^2} + P_n \\ &= \frac{P_{tk}G_{tk}(a_k\sigma_u + \sigma)A_r}{(4\pi R^2)^2} + P_n, \end{aligned} \tag{6}$$

where, P_{rgk} is the jamming power received by the FCR, P_{rsk} is the power of the real target echo, P_n is background noise power. For the smart noise jamming, the jamming cover exactly real target echo. As a result, the real target echo and jamming are always mixed so that the FCR can not separate them. Generally, the FCR judges the smart noise jamming by the total power of each pulse echo. The FCR can estimate the total power of pulse echo, which can be defined as

$$P_{erk} = \frac{P_{tk}G_{tk}A_r\sigma_{\max}}{\left(4\pi\tilde{R}^2\right)^2}, \tag{7}$$

where, $\tilde{R}$ is the estimate of R. The FCR judges the jamming by comparing the ratio P_{rk} to P_{erk} and the threshold γ_k. γ_k is determined by the FCR, which is unknown to the jammer. Then, the SINR of each pulse echo in the kth stage can be formulated as

$$SINR_k = \begin{cases} SINR_{k1}, & P_{rk}/P_{erk} \le \gamma_k \\ SINR_{k2}, & P_{rk}/P_{erk} > \gamma_k \end{cases}, \tag{8}$$

where,

$$SINR_{k1} = \frac{P_{rsk}}{P_{rgk} + P_n} = \frac{P_{tk}G_{tk}\sigma A_r}{P_{tk}G_{tk}a_k\sigma_u A_r + P_n(4\pi R^2)^2}, \tag{9}$$

$$SINR_{k2} = \lambda SINR_{k1}, \tag{10}$$

where, $\lambda > 1$.

Obviously, the search to lock-on time of the FCR T is determined by $SINR_k$ and the rule of working mode transition of the FCR. Assume that the rule is fixed, if a_k is too small to effectively degenerate $SINR_k$, T will be small. Conversely, if a_k is overly large, the FCR will take anti-jamming measures. As a result, $SINR_k$ will rise, T will be small, too. So, the performance of the jamming can be improved by establishing the optimal jamming actions. However, the relationship between T and a_k is difficult to obtain. As a consequence, it is important for us to investigate the multi-stage jamming power allocation with unknown environment model. To overcome the challenge of the unknown environment model, the model-free reinforcement learning is adopted. The method of multi-stage jamming power allocation based on the model-free reinforcement learning will be proposed in the next subsection.

B. The method of multi-stage jamming power allocation based on the model-free reinforcement learning

In this subsection, we first formulate a reinforcement learning framework for the jammer to determine its jamming power allocation policy. After that, we describe the model-free reinforcement learning algorithm under the framework, and adopt Q-learning algorithm to solve the multi-stage jamming power allocation problem.

With the help of [9], the reinforcement learning framework for the multi-stage jamming power allocation problem can be described by a tuple $\langle \mathbf{S}, \mathbf{A}, \Theta, \Psi, \delta \rangle$, where $\mathbf{S}$ is the finite state space, which corresponds to the set of working modes of the FCR in this problem. $\mathbf{A}$ is the finite action space, which corresponds to the set of jammer's jamming actions available. Θ is the the state transition function. It equals to the working mode of the FCR in the next stage for the given jamming action and the working mode of the FCR in the current stage. Ψ represents a reward function, which corresponds to the time spent by the FCR in the current stage. δ is the discount factor, which indicates jammer's evaluation of the rewards that obtained in the future (or in the past). Under the reinforcement learning framework for the multi-stage jamming power allocation problem, Q-learning algorithm can be adopted. It is a kind of model-free reinforcement learning and provides a way for the agent to learn how to act optimally [10]. The algorithm learns the optimal state-action value function Q^*, which then defines the optimal policy. More specifically, the agent maintains a table containing its current estimates of $Q^*(s,a)$. It observes the current state s and selects the action a that maximizes $Q(s,a)$ with some exploration strategies. In the multi-stage jamming power allocation problem, the Q-learning algorithm can be adopted as follows. For jammer, upon receiving a reward t_k, after the end of the current stage and observing the next state s_{k+1}, the table of Q-values is

978-1-7281-1679-2/19/$31.00 ©2019 IEEE

updated as follows:

$$Q(s_k,a_k) \leftarrow Q(s_k,a_k) + \alpha\left[t_k + \delta \max_a Q(s_{k+1},a) - Q(s_k,a_k)\right], \tag{11}$$

where, $\alpha \in (0,1)$ is the learning rate. As the number of episodes increases, the performance of the algorithm increaseing and eventually converges. Summary of the Q-learning algorithm based the multi-stage jamming power allocation is given in Algorithm 1.

Algorithm 1: The Q-learning algorithm for multi-stage smart noise jamming power allocation.

Input: Learning ratio sequence ($\{\alpha_n\} \in (0,1)$;
Exploration ratio sequence($\{\varepsilon_n\} \in (0,1]$;

Initialize the table of Q-values $Q(s,a) = 0$;
for *each episode* n **do**
  Initialize the working mode of the FCR s;
  while *target has not been locked on by the FCR* **do**
    With probability ε_n, choose the best jamming action a for s based on the table of Q-values, or with probability $1-\varepsilon_n$, randomly choose an available action for exploration;
    Perform jamming action a;
    Observe the time spent by the FCR t and the next working mode of the FCR s';
    Update the table of Q-values: $Q(s,a) \leftarrow Q(s,a) + \alpha_n\left[t + \delta \max_{a'} Q(s',a') - Q(s,a)\right]$;
    Update the state $s \leftarrow s'$ for the next stage;

IV. Simulation Results

In this section, we demonstrate the advantages and validity of the proposed multi-stage jamming power allocation method. Firstly, we assume that the FCR has 4 working modes (TWS, TAS1, TAS2, STT). We set that the FCR confirm a target (or lost a target) from 16 consecutive data frames: the target is considered lost if no target is detected in all data frames. The target is confirmed if the target is detected in M/N data frames, and the mode of the FCR will be transferred to the one that allocate more resources on the target tracking. The anti-jamming measure of the FCR is frequency agility. We set that the frequency agility lasts 9 data frames when the ratio P_{rk} to P_{erk} exceeds the threshold. We do not consider the loss brought by the frequency agility, and set that the jamming is invalid during the frequency agility. The detection probability is calculated using the Marcum Q function [11]. The coherent processing interval (CPI) of the FCR is 1 data frame. Transmit power of the FCR is 10kW, $A_r = 1\mathrm{m}^2$. The specific parameters of the FCR in each mode is shown in TABLE I, where L is the amount of pulses per CPI, G_t is mainbeam gain, Dr is tracking data rate, M/N is confirmation criterion, R_e is distance estimation error. In addition, we set that the jammer can select 13 jamming action $\mathbf{A} = [5.8, 6, 6.2, \cdots, 8.2]$, the interval of adjacent action is 0.2. We set $\sigma \sim U\left(10\mathrm{m}^2, 30\mathrm{m}^2\right)$, $\sigma_u = 20\mathrm{m}^2$. $P_n = 1.56 \times 10^{-13}\mathrm{W}$. To obtain the statistical performance, we repeat 400000 episodes, and average every 4000 episodes, which can be called a group of episodes. In each episode, we set the initial mode is TWS, $R \sim U(10\mathrm{km}, 30\mathrm{km})$.

TABLE I: The parameters of the FCR in each mode.

	L	G_t	Dr	M/N	R_e
TWS	24	40dB	8	3/8	2.5%
TAS1	30	52dB	20	4/8	1.5%
TAS2	35	58dB	25	4/8	0.75%
STT	50	69dB	100	5/8	0.15%

In the traditional method, the allocation of the multi-stage jamming power relies on expert experience and ignore the multi-stage electronic countermeasure. **The performance of the traditional method reaches the upper limit when the frequency agility threshold γ is known. However, in real applications, γ is difficult to obtain.** To prove the effectiveness of the proposed algorithm, we employ random allocation and the upper limit of the traditional method with known γ as control groups.

The simulations are divided into two different groups. For the first group of simulations, we set that the threshold of frequency agility γ is same for all 4 modes.

We can see from Fig.3 that the performance of the proposed algorithm is far better than the random allocation, and is better than the upper limit of the traditional method with known γ, but the advantage is not obvious enough. With the increase of γ, the advantage is more obvious. That's because the tracking data rate of TWS is far smaller than other working modes, the FCR spent most of the time in TWS. when γ is small, the time spent by other working mode is so little that the best jamming action for TWS plays a major role. But with the increaseing of γ, the importance of other models increase. As a result, the advantage is more obvious. In fact, an overly low threshold γ can not be set in practical applications. That is because the anti-jamming measures will always be taken if γ is overly low, which will result in unnecessary loss of SNR and loss of the target speed information.

For the second group of simulations, we set that the threshold of frequency agility γ is different for 4 modes. we consider two typical cases. case1: $\gamma^{\mathrm{TWS}} = 5.6$, $\gamma^{\mathrm{TAS1}} = 5.8$, $\gamma^{\mathrm{TAS2}} = 5.8$, $\gamma^{\mathrm{STT}} = 6$, where γ^{TWS} indicates the frequency agility threshold of TWS, other symbols are similar. case2: $\gamma^{\mathrm{TWS}} = 6$, $\gamma^{\mathrm{TAS1}} = 5.8$, $\gamma^{\mathrm{TAS2}} = 5.8$, $\gamma^{\mathrm{STT}} = 5.6$.

We can see from Fig.4 that the performance of the proposed algorithm is better than the random allocation and the upper limit of the traditional method with known γ, and the advantage is obvious when the threshold γ is set according to case1. That's because the frequency agility threshold of TWS is lower than other modes, which decrease the importance of TWS and

978-1-7281-1679-2/19/$31.00 ©2019 IEEE

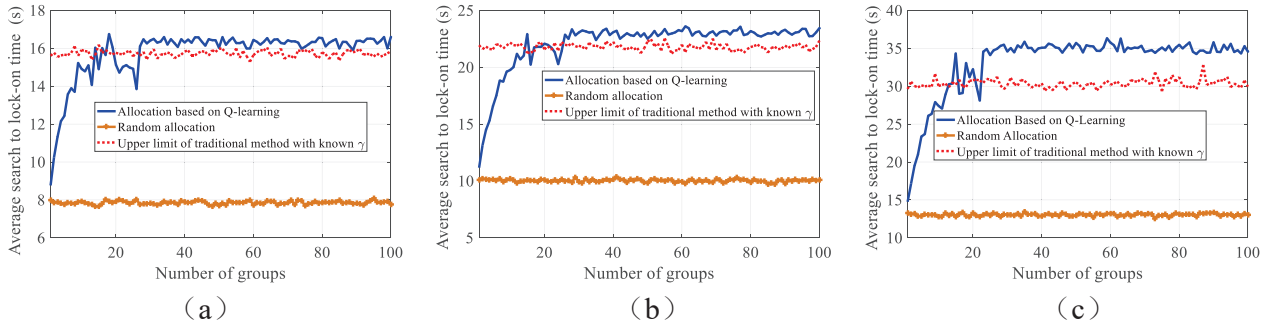

Fig. 2: Performance of different method when γ is same for all modes. (a) $\gamma = 5.6$. (b) $\gamma = 5.8$. (c) $\gamma = 6$

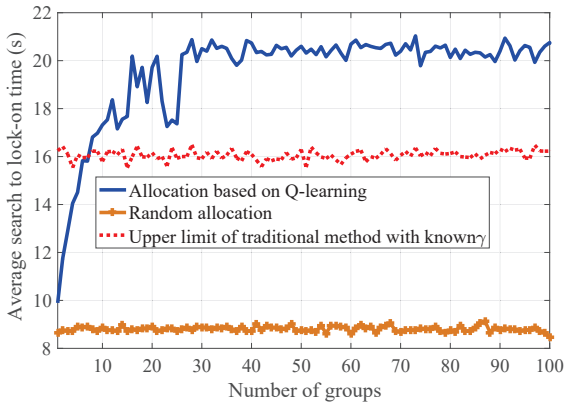

Fig. 3: Performance of different method when the threshold γ is set according to case1.

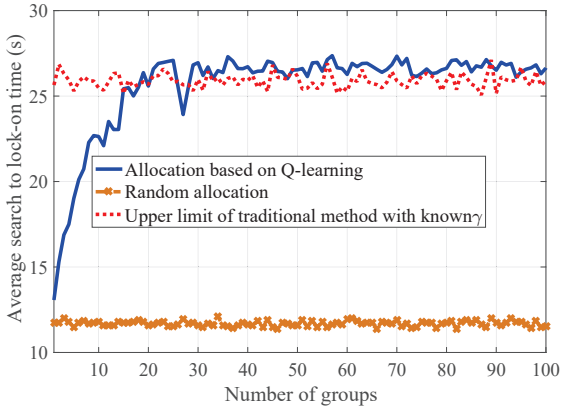

Fig. 4: Performance of different method when the threshold γ is set according to case2.

increase the importance of other working modes, so the upper limit of the traditional method with known γ degenerate. In contrast, when the threshold γ is set according to case2, the frequency agility threshold of TWS is higher than other modes, which increase the the importance of TWS and decrease the importance of other working modes, the upper limit of the traditional method with known γ improve as shown in Fig.5. But the performance of the proposed algorithm is still better.

In summary, the performance of the multi-stage jamming power allocation method based on the model-free reinforcement learning is better than control groups when the threshold γ is set according to above cases. In fact, the environment model can be learned using the model-free reinforcement learning. So, no matter how the threshold is set, a better jamming power allocation policy can be found by the proposed method.

V. Conclusion

In this paper, considering a multi-stage electronic countermeasure against a FCR with multiple working modes, an optimal multi-stage jamming power allocation for smart noise jamming has been studied with unknown environment model. The optimization goal is to obtain the optimal jamming performance by establishing the optimal jamming actions in multiple stages. We have proposed a method of multi-stage jamming power allocation based on Q-learning. Q-learning could be used to overcome the challenge of the unknown environment model. Finally, the simulation results show that the proposed method can achieve a more effective multi-stage smart noise jamming.

References

[1] A. D. Martino, *Introduction to modern EW systems*. Artech house, 2012.

[2] C. J. Watson, *A Comparison of DDS and DRFM Techniques in the Generation of Smart Noise Jamming Waveforms*. NAVAL POSTGRADUATE SCHOOL MONTEREY CA, 1996

[3] Z. Zhang, Y. Wu, J. Ren, K. Dong, "Smart noise jamming suppression by using atomic decomposition," *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, 2017, pp. 1377-1380.

[4] H. Hao, D. Zeng, P. Ge, "Research on the method of smart noise jamming on pulse radar." *Instrumentation and Measurement, Computer, Communication and Control (IMCCC), 2015 Fifth International Conference on*. IEEE, 2015.

[5] G. F. Stott, "Digital modulation for radar jamming." *Signal Processing in Electronic Warfare, IEE Colloquium on.* IET, 1994.

[6] D. DiFilippo, G. Geling, G. Currie, "Simulator for advanced fighter radar EPM development." *IEE Proceedings-Radar, Sonar and Navigation* 148.3 (2001): 139-146.

[7] D. C. Schleher, *Electronic warfare in the information age*. Artech House, Inc., 1999.

[8] D. L. Adamy, *EW 104: Electronic Warfare Against a New Generation of Threats*. Artech House, 2015.

[9] E. Yang, D. Gu, *Multiagent reinforcement learning for multi-robot systems: A survey*. tech. rep, 2004.

978-1-7281-1679-2/19/$31.00 ©2019 IEEE

[10] R. S. Sutton, A. G. Barto, *Introduction to reinforcement learning*. Vol. 135. Cambridge: MIT press, 1998.
[11] Y. Yang, T. Zhang, W. Yi, L. Kong, X. Li, X. Yang. "Multi-static radar power allocation for multi-stage stochastic task of missile interception." *IET Radar, Sonar & Navigation* 12.5 (2018): 540-548.

978-1-7281-1679-2/19/$31.00 ©2019 IEEE