Optimal Jamming Frequency Selection for Cognitive Jammer based on Reinforcement Learning

Lulu Wang, Jinlin Peng, Zhidong Xie
Artificial Intelligence Research Center
National Innovation Institute of Defense Technology
Beijing, China
e-mail: wanglulunudt@163.com

Yi Zhang
State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, Luoyang, China
e-mail: zhangyinudt@nudt.edu.cn

***Abstract*—In this paper, a cognitive jammer is developed which can adaptively and optimally jam the radar and protect the target from being detected. The interaction of the cognitive jammer and the environment is modeled as a finite Markov Decision Process based on the framework of reinforcement learning. Q-learning algorithm is used to solve the optimal jamming frequency selection problem. After several interactions, the jammer can learn the radar's strategy and optimize its jamming frequency to achieve a larger jamming-plus-noise to signal ratio (JNSR) during the whole process. Numerical results are given to illustrate the effectiveness of the proposed method. Compared with a random jamming frequency selection method, the JNSR of the proposed method is significantly larger.**

Keywords-jamming frequency selection; reinforcement learning; Q-learning

I. INTRODUCTION

In modern electronic warfare, the competition between radar and jammer is increasingly intense [1]. The anti-jamming performance of modern radar has been greatly improved by various technologies. Therefore, there is an urgent need for the improvement of jamming technology. Recently development of machine learning algorithms and hardware has enabled the investigation and implementation of cognitive electronic warfare [2]. In cognitive electronic warfare, the intelligent jammer is able to learn and adapt to the unknown environment.

Jamming has traditionally been studied by using either optimization or game-theoretic or information theoretic principles [3]. In [1], optimal jamming techniques based on signal-to-interference plus noise ratio (SINR) and mutual information is studied. Zheng et al investigated the optimal jamming strategy towards MIMO radar using CRB as the optimization criterion [4]. In [5], Wang et al studied the jamming power allocation strategy for MIMO radar based on MMSE and mutual information. In [6-9], the interaction between a MIMO radar and jammer is modeled as a game. Optimal power allocation strategies are obtained at the equilibrium.

In this paper, we investigate the frequency selection strategy of a cognitive jammer, which can sense the environment and choose the jamming frequency to influence the environment so that the target can be protected. Reinforcement learning is an area of machine learning with connections to control theory, optimization, and cognitive sciences [10]. The reinforcement learning framework is a machine learning technique that allows an active learner, called agent, to learn through experience. Specifically, the agent learns how to choose the best action to achieve its goal by interacting with an unknown environment, without any preassigned control policy. The "learning loop" characterizing the reinforcement learning framework has strong similarity with the cognitive jammer iterative feedback control system.

Reinforcement learning has been widely used in policy making in both cognitive radio [11-12] and cognitive radar [13-14]. In [11], Q-learning algorithm is used to explain how a cognitive radio can exploit its ability of dynamic spectrum access and its learning capabilities to avoid jamming channels. In [12], deep reinforcement learning is investigated to improve the anti-jamming communication performance. Anti-jamming frequency hopping strategies for cognitive radar based on Q-learning and deep Q-network is investigated in [13]. In [14], adaptive waveform selection in cognitive radar is studied using Q-learning. For optimal jamming, S. Amuru et al used the multi-armed bandit (MAB) method to optimize the physical layer parameters of jamming [15-16] against the transmitter-receiver pair. The methods can also be categorized into the reinforcement learning framework. In this paper, we emphasize the problem of optimal jamming against radar under the framework of reinforcement learning.

The rest of the paper is organized as follows. Firstly, we introduce the signal model of a cognitive jammer and its environment. Then the reinforcement learning framework is given with some definitions and notations. Our problem is modeled under the reinforcement learning framework. The optimal jamming frequency selection is solved using Q-learning algorithm. At last, numerical examples are given to show the effectiveness of our algorithm.

II. SIGNAL MODEL

We consider the scenario where a radar, a target and a support jammer are present, which is shown in Fig.1. The radar is a frequency agile radar, the carrier frequency of which can change pulse by pulse in a coherent processing interval (CPI). The support jammer is cognitive, which is

978-1-7281-5102-1/19/$31.00 ©2019 IEEE

capable of sensing the environment and adaptively change its jamming frequency so that the target could be well protected.

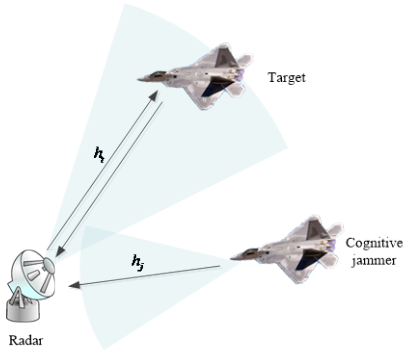

Figure 1. Radar and a target equipped with a jammer.

Suppose that the target can be regarded as a point target with radar cross section (RCS) σ. The transmission power of the radar and jammer are P_r and P_j, respectively. The channel gain from the radar to the target is h_t. The channel gain from the jammer to the radar is h_j. Because the jamming signal comes from the sidelobe, there will be a loss L. For the frequency agile radar, suppose that there are N pulses in a CPI. The carrier frequency of the nth pulse is $f_r^{(n)}$, with a bandwidth $B_r^{(n)}$. For the cognitive jammer, the transmit jamming signal is also changed pulse by pulse. The carrier frequency of the nth jamming signal is supposed to be $f_j^{(n)}$, with a bandwidth $B_j^{(n)}$. For simplicity, we suppose that if there is no overlap between the frequency band of the radar and the jammer, i.e., $f_r^{(n)} + \frac{B_r^{(n)}}{2} < f_j^{(n)} - \frac{B_j^{(n)}}{2}$ or $f_r^{(n)} - \frac{B_r^{(n)}}{2} > f_j^{(n)} + \frac{B_j^{(n)}}{2}$, the SINR of the radar is:

$$SINR_n = \frac{P_r h_t^2 \sigma}{P_j h_j L + \sigma_n^2} \tag{1}$$

where σ_n^2 is the noise power at the radar receiver. Note that if only part of the frequency bands of the jamming signal is received by the radar, there will be another loss. However, we don't take this into account in this paper.
Otherwise, the radar is not jammed, so

$$SINR_n = \frac{P_r h_t^2 \sigma}{\sigma_n^2} \tag{2}$$

Note that if the accumulative SINR in a CPI is high, the target detection performance of the radar is better. Therefore, to protect the target, the jammer should try to reduce the accumulative SINR of the radar.

III. Q-Learning Based Frequency Selection

In this section, we provide a full description of the Reinforcement learning based jamming frequency selection problem. The basic concepts and notations of reinforcement learning are recalled first. Then the jamming frequency selection problem is modeled in the reinforcement learning framework. Q learning algorithm is introduced to solve the policy decision problem.

A. Reinforcement Learning and Some Notations

Reinforcement learning is learning by interacting with the environment [17]. Refer to Fig.2, the intelligent *agent* can make observations from the *environment*. Based on the observations, the agent is able to sense the state of its environment to some extent and then decide which action to take so that the state of the environment should be influenced by this action. There is a reward signal which defines the goal of the problem. The reward can be regarded as an evaluative feedback that indicates how good the action taken was.

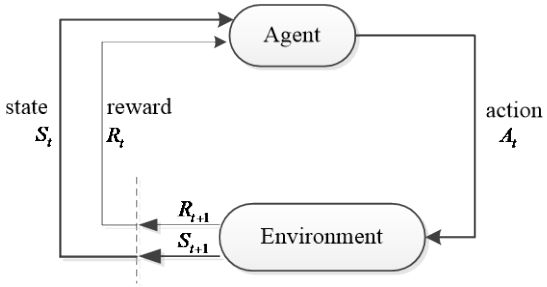

Figure 2. The agent and environment interaction.

More specifically, the agent and environment interact at each of a sequence of discrete time steps, $t = 0, 1, 2, \cdots$. At each time step t, the agent receives the representation of the environment's *state*, $S_t \in \mathcal{S}$, and selects an *action* $A_t \in \mathcal{A}(s)$ based on the state. One time step later, the agent receives a numerical *reward*, $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$, and the state goes to S_{t+1}. Usually, a (finite) Markov decision process (MDP) is used to describe the procedure. Formally, and MDP model is a tuple $\{\mathcal{S}, \mathcal{A}, P, \mathcal{R}\}$ with four essential components:

- $\mathcal{S}$ is the (finite) state space,
- $\mathcal{A}$ is a (finite) set of actions,
- $P: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the transition probability,
- $\mathcal{R}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function.

The *policy* $\pi: \mathcal{S} \to \mathcal{A}$ is to determine which action to be taken at each state. The critical point of this learning procedure is finding the *optimal policy* π_*.

To estimate how good a policy is, we introduce the *state-value function* v_π for policy π.

$$v_\pi(s) \doteq E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s\right] \tag{3}$$

where $E_\pi[\cdot]$ denotes the expected value given that the agent follows policy π. $\gamma \in (0,1]$ is the discount rate, which means that long-term reward are considered. Small values of

γ emphasizing near-term gain and larger values giving significant weight to later rewards. Then the optimal policy is $\pi_*(s) = \arg\max_\pi v_\pi(s)$. The *action-value function* for policy π is defined as

$$q_\pi(s,a) \doteq E_\pi\left[\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\middle|S_t = s, A_t = a\right]\right] \quad (4)$$

Although there may be more than one optimal policies, they all share the same state-value function and action value function, called *optimal state-value function* v_* and *optimal action-value function* q_*, respectively.

The key idea of reinforcement learning is the use of value functions to organize and structure the search for good policies. Once q_* is obtained, the optimal policy at the state s is simply $\pi_*(s) = \arg\max_{a\in\mathcal{A}} q_*(s,a)$.

There are many methods to solve the reinforcement learning problem. Here we introduce the Q-learning algorithm, which provides a recursive way to learn the optimal action-value function. Define the learned action-value function as $Q(S_t, A_t)$, and its update rule:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right] \quad (5)$$

where $\alpha \in (0,1]$ is the learning rate which shows how much error will be learnt in the updating rule. Note that Q converges to q_* with probability 1.

B. Q-Learning Method for Jamming Frequency Selection

In this paper, the agent under investigation is the cognitive jammer. The environment contains the radar, target, noise and so on, which is all the electromagnetic field without the jammer. Suppose that the jammer has the knowledge of the target RCS, noise PSD, channel gain, and its loss at the radar receiver. However, it doesn't know how the radar frequency changes. At time step t, the jammer can sense from the environment the previous transmission of the radar $S_t = f_r^{(t-1)}$ and then choose the current jamming frequency $A_t = f_j^{(t)}$, hoping to jam the current radar transmission. The action is the jamming frequency. The reward is the estimated jamming-plus-noise to signal ratio (JNSR) at the receiver, which is the reciprocal of SINR. Note that the estimated JNSR should be obtained in the next time step $t+1$. When the jammer has sensed the radar frequency $S_{t+1} = f_r^{(t)}$, it can estimate the JNSR at time step t at the radar receiver. i.e.,

$$R_{t+1} = JNSR(s_{t+1}, a_t) = JNSR\left(f_r^{(t)}, f_j^{(t)}\right) = \begin{cases} \dfrac{P_j h_j L + \sigma_n^2}{P_r h_t^2 \sigma}, & jammed \\ \dfrac{\sigma_n^2}{P_r h_t^2 \sigma}, & not\ jammed \end{cases} \quad (6)$$

The Q-learning algorithm for optimal jamming frequency selection is shown as follows. Note that for each time step, the jammer chooses the frequency based on the Q values of the current state using ε-greedy algorithm. It chooses the action which has the largest Q value with the probability of 1-ε . And it randomly chooses the actions with the probability of ε . This is one of the common ways to balance the exploration and exploitation. In this paper, we let $\varepsilon = e^{-0.05k}$ which reduces the value with the learning process. And thus, the jammer will explore more at the beginning when the environment is not well learned and will exploit more after several interactions with the environment.

TABLE I. Q-LEARNING ALGORITHM

Q-learning algorithm for jamming frequency selection	
1:	Initialize $\gamma, \alpha, Q(s,a), f_r^{(0)}, CPI_{num}, N$
2:	For $k = 1, \cdots, CPI_{num}$:
3:	$s_1 = f_r^{(0)}$
4:	For $t = 1, \cdots, N$:
5:	Choose action $a_t = f_j^{(t)}$ according to $Q(s_t, a)$ using ε-greedy with $\varepsilon = e^{-0.05k}$
6:	Observe the next state $s_{t+1} = f_r^{(t)}$, and obtain reward R_{t+1}
7:	Update $Q(s,a)$ according to (5)
8:	state ← next state
9:	End For
10:	End For

IV. SIMULATION AND RESULTS

In this section, numerical experiments are performed to demonstrate the effectiveness of the proposed method. Suppose that the radar and jammer frequency bands are all 20MHz. The frequency can change from 3GHz to 5GHz, with a step of the bandwidth 20MHz. Therefore, the total jamming frequency strategies are 100, i.e., $|\mathcal{A}| = 100$. And the number of states is 100 too, i.e., $|\mathcal{S}| = 100$. The environment parameters are supposed to be known by the jammer. $P_r = 500W$, $P_j = 1000W$, $L = -10dB$, $h_t = h_j = 0.1$, $\sigma = 1$, $\sigma_n^2 = 1$, $N = 100$. The frequency agile radar is supposed to sweep its center frequency from 3GHz to 5GHz, with a step of 20MHz in a CPI. Note that is assumption is given to better illustrate the results that the jammer can learn the changing rule of the radar frequency. In practice, the radar may randomly change its transmitted frequency. However, the changing rule is able to be learned if the interaction of the radar and the jammer continues for a long time.

Let $\alpha = 0.01$, $\gamma = 0.8$. 200 CPI is considered. Within each CPI, there are $N = 100$ number of pulses or time steps. We apply the Q-learning algorithm in Table 1 to find the optimal jamming frequency during the interaction procedure of the radar and jammer. The average JNSR of a CPI is

calculated and compare with the random jamming frequency selection method. The result is shown in Fig.3. It is illustrated that the JNSR of using Q-learning is much greater than that of using random jamming frequency selection method. There is an improvement of about 5.5dB. Therefore, if using the jamming frequency selection method based on Q-learning, the target could be better protected owing to the reduced detection performance of the radar.

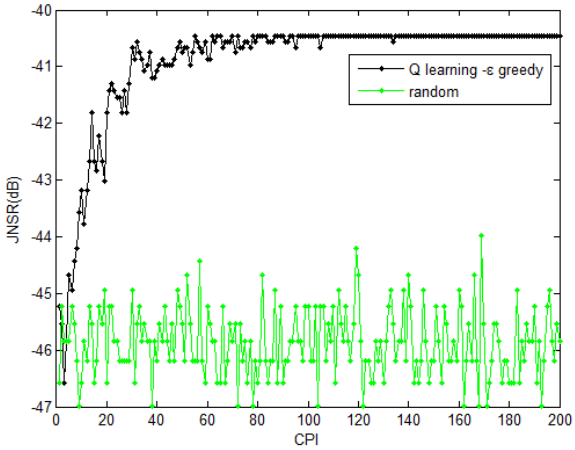

Figure 3. Average JNSR comparison.

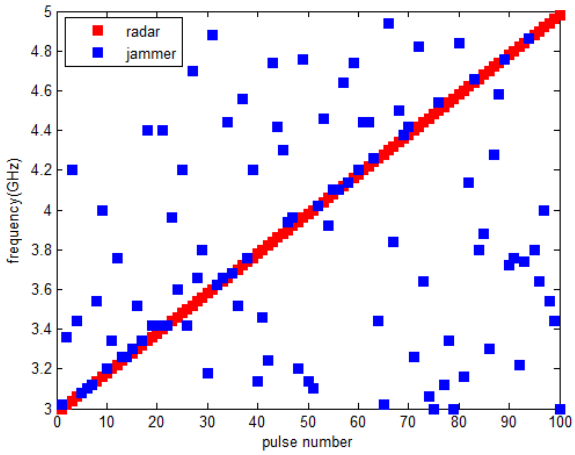

Figure 4. Jamming and radar frequency in a CPI.

Then we use the learned action-value function to shown the jamming frequency selection. Fig.4 shows the radar frequency and the jamming frequency. We can see that the jamming frequency is able to predict the radar frequency and choose the similar frequency bands to improve the jamming effect. Although there are still many pulses that the jamming signal cannot jam, we can see a predicted line from the left bottom to the right top. Therefore, the overall JNSR in this CPI is improved.

V. CONCLUSION

In this paper, the interaction between a cognitive jammer and a radar is modeled as a finite MDP based on the framework of reinforcement learning. The jammer could sense the environment and adaptively change its jamming frequency to protect the target from being detected. We model the jammer behavior in the environment as a Markov decision process. Q-learning algorithm is used to solve the optimal jamming frequency selection algorithm. Experiments show that the jamming-plus-noise to signal ratio (JNSR) is improved by 5.5dB of using the proposed algorithm compared with the random frequency selection algorithm. From the comparison of the radar frequency and jammer frequency, we can conclude that the jammer has the ability of learning the frequency changing rule of the radar.

In future, we will use more practical model and add other jamming strategies into the action set to further investigate the decision-making method of cognitive jammer. Furthermore, the radar is assumed to be a traditional frequency agile radar in this paper. Future work is on going which takes into account that both the radar and the jammer are cognitive and could learn to change their actions.

ACKNOWLEDGMENT

This work is funded by National Natural Science Foundation of China, grant number 61701502, China Postdoctoral Science Foundation, grant number 45649 and Natural Science Foundation of Hunan Province, grant number 2019JJ40340.

REFERENCES

[1] L. Wang, H. Wang, K. Wong, and P. Brennan. “Minimax robust jamming techniques based on signal-to-interference-plus-noise ratio and mutual information criteria”, IET Communications, vol. 8, no.10. pp. 1895-1867. 2013.

[2] B. Manz. “Cognition: EW gets brainy”, Journal of Electronic Defense. vol. 35, no. 10, pp. 32-39. 2012.

[3] S. Amuru, C. Tekin, M. van der Schaar, R. M. Buehrer. “Jamming bandits—A novel learning method for optimal jamming”. IEEE Transactions on Wireless Communications. Vol. 15, no. 4, pp. 2792-2808, 2016.

[4] G. Zheng, S. Na, T. Huang, L. Wang. “A barrage jamming strategy based on CRB maximization against distributed MIMO radar”. Sensors, 2019, 19, 2453; doi:10.3390/s19112453.

[5] L. Wang, L. Wang, Y. Zeng, M. Wang. “Jamming power allocation strategy for MIMO radar based on MMSE and mutual information”. IET Radar, Sonar & Navigation, vol. 11, no. 7, pp. 1081-1089, 2017.

[6] X. Song, P. Willett, S. Zhou, P.B. Luh. “The MIMO radar and jammer games”. IEEE Transactions on Signal Processing. Vol. 60, no. 2, pp. 687-699. 2012.

[7] H. Gao, J. Wang, C. Jiang, X. Zhang. “Equilibrium between a statistical MIMO radar and a jammer”, IEEE Radar Conference, Arlington, USA, pp. 461-466. 2015.

[8] A. Deligiannis,, G. Rossetti,, A. Panoui, S. Lambotharan. “Power allocation game between a radar network and multiple jammers”, IEEE Radar Conference, Philadelphia, USA. pp. 1-5. 2016.

[9] L. Wang, Y. Zhang. “MIMO radar and jamming power allocation game based on MMSE”, International Radar Symposium. ULM Germany. 2019.

[10] L. Wang, S. Fortunati, M. S. Greco, F. Gini. “Reinforcement learning-based waveform optimization for MIMO multi-target detection”, Asilomar Conference on Signals, Systems and Computers ,pp. 1329-1333. 2018

[11] F. Slimeni, B. Scheers, Z. Chtourou, V. Le Nir, R. Attia. “Cognitive radio jamming mitigation using markov decision process and

reinforcement learning". International Conference on Advanced Wireless, Information, and Communication Technologies, 2015.

[12] G. Han, L. Xiao, H. V. Poor. "Two-dimensional anti-jamming communication based on deep reinforcement learning". International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2019, pp. 2087-2091.

[13] K. Li, B. Jiu, H. Liu, S. Liang. "Reinforcement learning based anti-jamming frequency hopping strategies design for cognitive radar". 2018 IEEE International Conference on Signal Processing, Communications and Computing.

[14] B. Wang, J. Wang, X. Song, F. Liu. "Q-learning based adaptive waveform selection in cognitive radar". International Journal of Communications, Network and System Sciences, 2009, 7, pp. 669-674.

[15] S. Amuru, C. Tekin, M. van der Schaar, R. M. Buehrer. "Jamming bandits – A novel learning method for optimal jamming". IEEE Transactions on Wireless Communications, vol. 15, no. 4, 2016, pp. 2792-2808.

[16] S. Amuru, R. M. Buehrer. "Optimal jamming using delayed learning". IEEE Military Communications Conference. Baltimore, MD, USA, 2014, pp. 1528-1533.

[17] R. S. Sutton, A. G. Barto. Reinforcement learning: An introduction. Second Edition. MIT Press, Cambridge, MA, 2018.