Reinforcement Learning based Anti-jamming Frequency Hopping Strategies Design for Cognitive Radar

Li Kang, Jiu Bo, Liu Hongwei, Liang Siyuan
National Lab. Of Radar Signal Processing
Xidian University
Xi'an, China
xdlk@foxmail.com

***Abstract*—Frequency agile (FA) radar is capable of altering carrier frequency randomly, which is especially useful for radar anti-jamming designs. Obviously, random frequency hopping is not the best choice if the radar can learn the jammer' strategy. In this paper, a novel frequency hopping strategy design method is proposed for cognitive radar to defeat the smart jammer, in which the radar does not know the exact jamming model. Q-learning and deep Q-network (DQN) is utilized to solve this problem. By applying the reinforcement learning algorithm, the radar is able to learn the jammer's strategies through the interaction with environment and adopt the best action to obtain high reward. The learning performance of DQN is much better than that of Q-learning especially when the available frequencies are large. The proposed method can improve the signal-to-interference-plus-noise ratio (SINR) for the radar when the jamming model is not available. Numerical results are given to illustrate the effectiveness of the proposed method.**

Keywords—cognitive radar; frequency hopping; Q-learning; deep Q-network; jammer;

I. INTRODUCTION

Frequency agile radar can be able to change pulse-to-pulse carrier frequency randomly within a given bandwidth and has many advantages over the traditional radar [1-3], especially useful for electronic counter-countermeasures (ECCM) designs [4].

One of the biggest trends of radar is the cognitive radar which is proposed by Haykin. The cognitive radar enables the radar to interact with the surrounding environment and then takes advantage of the feedback information to improve the detecting or tracking performance [5]. That means the performance of ECCM may also be improved by adjusting its strategies (carrier frequency, transmission power, waveform parameters and so on) when facing the attack from a smart jammer. In this paper, we focus on the frequency hopping strategies for cognitive radar in a given bandwidth without knowing the exact jamming model.

Reinforcement learning is a branch of machine learning and it aims at making an agent learn to take proper actions to achieve a high reward [6]. Different from supervised learning, reinforcement learning does not have true labels telling an agent how to act to achieve the high reward. The agent must interact with its environment and figure out how to act to obtain high reward. The reinforcement learning problem is usually modelled as a Markov Decision Problem (MDP) and there are many algorithms to solve it. Q-learning is one of the most popular algorithms and has been applied in communication systems recently [7-9]. In [7], a two-dimensional anti-jamming communication scheme for cognitive radio networks is developed based on Q-learning and deep Q-network.

Inspired by the previous work in communication systems, Q-learning is utilized to design the frequency hopping strategies to defeat the smart jammer, which may also alter the frequency according to the radar's carrier frequency. When the available frequencies are large, it is difficult for Q-learning to give a good result because of the large state spaces. DQN is developed by Google DeepMind [10] recently. DQN uses a neural network to map the state to the corresponding state-action value so as to overcome the problem of high-dimensionality and accelerate the learning speed. Based on Q-learning, a DQN based algorithm is also applied in this paper.

In addition to the attack from the enemy, the radar may also receive interference signals from some unintended sources such as other radars, communication systems and so on. In this paper, intended and unintended attacks are both considered. Based on Q-learning and DQN, the cognitive radar interacts with the environment and adjusts its carrier frequency to achieve the best performance.

We formulate the paper as the following order. Firstly, the frequency agile model is introduced briefly. Then, the frequency hopping strategies design method based on Q-learning and DQN is presented in section III. Finally, numerical results and conclusions are given in section IV and V, respectively.

II. FREQUENCY AGILE MODEL

As shown in Fig. 1, the carrier frequency of the radar is randomly changed in a coherent processing interval (CPI) [4]. Here let f_c be the initial carrier frequency and Δf be the frequency step. f_n is the carrier frequency of the n th pulse and

This work is partially supported by the Nation Natural Science Foundation of China (61671351), the National Science Fund for Distinguished Young Scholars (61525105), the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 project) (No.B18039).

978-1-5386-7946-3/18/$31.00 ©2018 IEEE

it equals to $f_c + d_n \Delta f$, where n ranges from 0 to $N-1$ and N is the number of pulses in a CPI. d_n is the frequency modulation code and it is usually a random integer ranging from 0 to $M-1$, where M is the number of available frequencies.

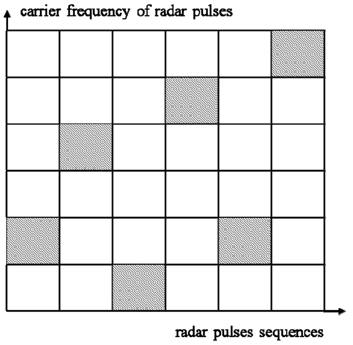


Fig. 1 Carrier frequency of frequency agile radar

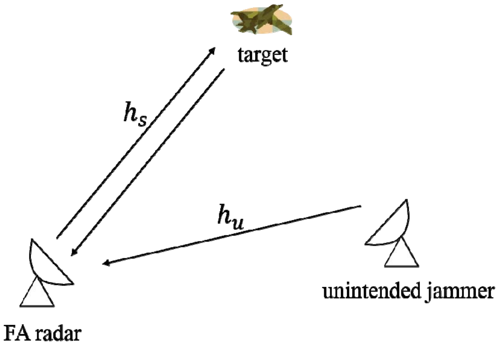


Fig. 2 FA radar and jammers

As shown in Fig. 2, an intended jammer is a plane equipped with a self-defense jammer. The unintended jammer is another radar in an alliance with the FA radar. Let σ denote the radar cross section (RCS) of the target. The channel gains from the FA radar to the target and the unintended jammer are h_s and h_u, respectively. f_n is the carrier frequency of the nth pulse of the FA radar. f_j and f_u are the frequency of the intended and the unintended jammer, respectively. The SINR for the nth pulse at the FA radar receiver can be defined as follows.

$$SINR_n = \frac{P_s h_s^2 \sigma}{n_0 + P_j h_s I\left(f_j = f_n\right) + P_u h_u I\left(f_u = f_n\right)} \qquad (1)$$

where P_s is the power of the FA radar, and n_0 is the FA radar receiver noise power. P_j and P_u are the power of the intended and unintended jammer, respectively. If ξ is true, $I(\xi)$ equals to 1. If else, $I(\xi)$ equals to 0.

III. Q-LEARNING AND DQN BASED FREQUENCY HOPPING STRATEGIES

A MDP is used to describe the state transition in the frequency hopping problem. Q-function, $Q(s,a)$, is utilized to evaluate the long term expected reward when the agent takes action a at the state s. The mathematical form of Q-function is defined as

$$\begin{aligned} Q^{\pi}(s,a) &= E_{\pi}\left\{R_t \mid \mathrm{s}_t = s, a_t = a\right\} \\ &= E_{\pi}\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid \mathrm{s}_t = s, a_t = a\right\}, \end{aligned} \qquad (2)$$

where $E_{\pi}\{\ \}$ denotes the expectation operation with respect to the policy π. t is any time step. γ is the discount factor and r_t is the reward that the agent received at time step t. s_t and a_t are the system state and the agent action at time step t, respectively. The solution to the reinforcement problem is as follows

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \quad . \qquad (3)$$

The agent in our problem is the FA radar and the action at nth pulse is the carrier frequency f_n. The SINR defined in (1) is chosen as the reward and the radar will receive a reward for each action. In this paper, it is assumed that the radar does not know the jamming model so the radar can only choose its action by the learned Q-function. The carrier frequency f_{t-1} and the SINR at time step $t-1$, $s_t = \left[f_{t-1}, SINR_{t-1}\right]$, are chosen as the state in the Q-function.

Algorithm 1 Q-learning based frequency hopping strategies

```
Initialize $\gamma$, $P_s, P_j, P_u$, $Q(s,a) = 0, \forall s, \alpha$
Repeat (for each CPI):
  Given $SINR_0$ and $a_0 = f_c$ to initialize the state $s_1 = [a_0, SINR_0]$
  For $t = 1, 2, ..., N$:
    Choose action $a_t = f_c + d_t \Delta f, d_t \in \{0, 1, ..., M-1\}$ using ϵ-greedy algorithm
    Take action $a_t$ and observe the reward $r_t = SINR_t$
    Obtain $s_{t+1} = [\mathrm{a}_t, SINR_t]$
    Update the Q-function according to (4)
$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_t + \gamma \max_{a'} \mathrm{Q}(s_{t+1}, a') - Q(s_t, a_t)\right]$ (4)
```

Q-learning is used to solve the reinforcement problem mentioned above. When the radar is at state s_t, it takes actions according to the learned Q-function $Q(s_{\mathrm{t}}, a)$. ϵ-greedy strategy is used to balance the exploration and exploitation . On the one hand, the radar takes the optimal action which maximizes the Q-function at state s_t with the probability 1- ϵ. On the other hand, the radar takes action randomly to with the probability ϵ to find the potential actions that may benefit the future reward. The Q-learning based frequency hopping strategies design method proposed in this paper is shown in Algorithm 1.

DQN, which combines with Q-learning with neural network, is proposed [10]. By using a neural network like convolutional neural network (CNN) or fully connected neural network to approximate the Q-function, the high dimensional state problem can be overcome. Moreover, the neural network has higher nonlinear fitting capacity so the convergence speed can also be improved.

Algorithm 2 DQN based frequency hopping strategies

```
Initialize the relay memory D to capacity E
Initialize action-value function $Q$ with random weights $\theta$
Initialize target action-value function $\hat{Q}$ with weights $\theta^{-}=\theta$
Initialize $\gamma, P_s, P_j, P_u$
Repeat (for each CPI):
  Given $SINR_0$ and $a_0=f_c$ to initialize the state $s_1=[a_0, SINR_0]$
  For $t=1,2,...,N$ :
    Choose action $a_t=f_c+d_t\Delta f, d_t\in\{0,1,...,M-1\}$ using ϵ-greedy algorithm
    Take action $a_t$ and observe the reward $r_t=SINR_t$
    Obtain $s_{t+1}=[\mathrm{a}_t, SINR_t]$
    Store transition $(s_t, a_t, r_t, s_{t+1})$ in D
    Sample random minibatch of transitions $(s_j, a_j, r_j, s_{j+1})$ from D
    Set $y_j=r_j+\gamma\max_{a'}\hat{Q}(s_{j+1}, a';\theta^{-})$
    Perform a gradient descent step on $(y_j-Q(s_j, a_j;\theta))^2$ with respect to the network parameter $\theta$
    Every C steps reset $\hat{Q}=Q$
```

In this paper, three fully connected (FC) layers are used to approximate the Q-function $Q(s,a;\theta)$, where θ represents the parameters of the neural network. Note that two tricks are considered in DQN to improve the stability of the algorithm. The first is to set a new network $\hat{Q}(s,a;\theta^{-})$ as the target network. $\hat{Q}$ is used to generate the target value $r_t+\gamma\max_{a'}\mathrm{Q}(s_{t+1},a')$ for the parameter update of the network Q. Generally, we remain the parameter of $\hat{Q}$ unchanged for C steps and then clone the parameter of Q to $\hat{Q}$ [10]. The second trick is called experience replay. As the name shows, there is a memory pool D to store the agent's experience, denoted as $e_t=(s_t,a_t,r_t,s_{t+1})$, at each time step [10]. When training the network Q, sample a minibatch of memories randomly in D.

As a benchmark, we propose a DQN based frequency hopping strategies as shown in Algorithm 2.

The detail of the network is illustrated in Fig. 3. The network consists of three hidden layers and each layer includes 20 units activated by a ReLU activation function.

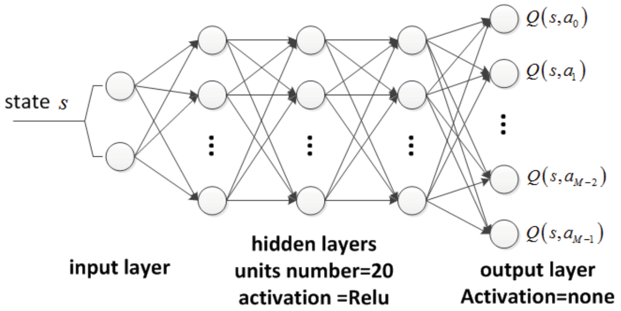


Fig. 3 Structure of fully connected neural network

IV. Experiment Results

In this section, simulations are performed to demonstrate the effectiveness of the proposed method. The carrier frequency is random stepped from $3GHz$ to $5GHz$ with a frequency step $\Delta f=1MHz$. As a result, the number of available frequencies M is 2000 and the action of the FA radar a belongs to $\{0,1,...,1999\}$ corresponding to the carrier frequency from $3GHz$ to $5GHz$. The number of pulses in a CPI is $N=100$.

We set $P_s=1000$ against the intended jammer with $P_j=500$ and the unintended jammer with $P_u=500$. Let the channel gains $h_s=h_u=0.1$. The RCS of the target is $\sigma=1$ and the noise power is $n_0=1$. At time step t, we assume that the jammer will choose action f_{t-1} with probability 1-ϵ otherwise choose action randomly. For the unintended jammer, we assume that the frequencies of the unintended jammer distribute over the intervals $[3GHz, 3.5GHz]$ and $[4GHz, 4.5GHz]$. That means when the radar's carrier frequency is belonging to these frequency intervals, it will be jammed.

Fig. 4 shows the performances of three different methods. Y axis represents the average SINR over one CPI. As shown in Fig. 4, the DQN based frequency hopping strategies design method outperforms two other methods. It is obvious that random choice is the worst method since the FA radar is always jammed by the unintended jammer. For the Q-learning based method, its performance is better than the random choice method but worse than the DQN based method. This is because the state space is so enormous that we can nearly not find an optimal policy even in the limit of inifinite time and data.

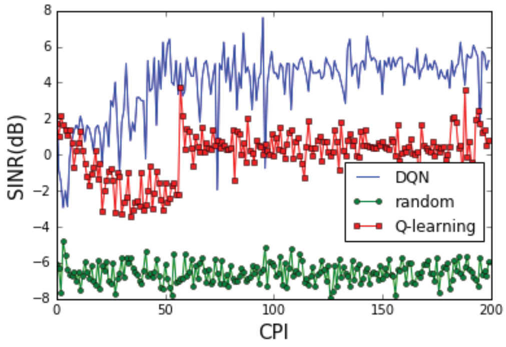
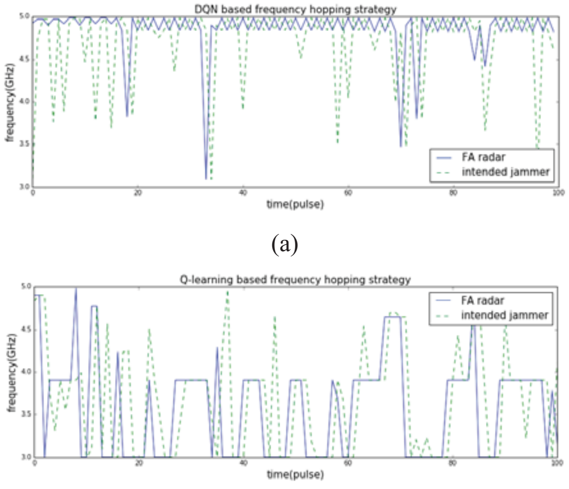


Fig. 4 Performace comparision among different methods



(a)

(b)

Fig. 5 Strategies of FA radar and intended in one CPI

Fig. 5 shows the strategies of the FA radar and the intended jammer in one CPI based on DQN and Q-learning, respectively. From Fig.5 (a) we can find that on the one hand, the radar keeps changing its carrier frequency in order to avoid being jammed by the intended jammer; on the other hand, the radar will not let its carrier frequency fall into the intervals $\left[3GHz,3.5GHz\right]$ and $\left[4GHz,4.5GHz\right]$ so as not to be jammer by the unintended jammer. From Fig. 5 (b), we can find that the radar is always jammed by intended and unintended jammers leading to a worse performance than DQN based method.

Fig. 6 shows the total number of times jammed by unintended jammer for each CPI. It can be discovered that number of times jammed by the unintended jammer based on Q-learning is much more than the times based on DQN. Based on DQN, the agent gradually learns the unitended jammer's strategies so it can avoid being jammed in the following CPI. However, the agent in Q-learning based method can not learn well because of the large state space, however, its performance is still better than that of the random choice method.

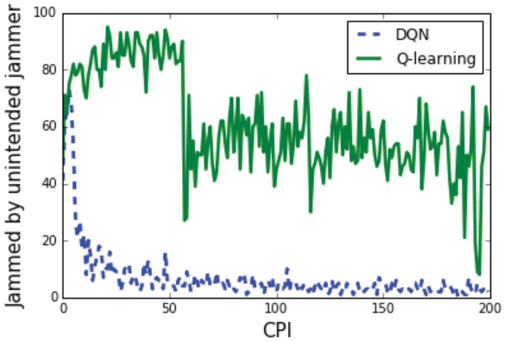


Fig. 6 Number of times jammed by unintended jammer

V. CONCLUSION

In this paper, a novel frequency hopping strategies design method is proposed for cognitive radar without knowing the jamming model. Q-learning and DQN are utilized to solve the problem. By comparing Q-learning based method with DQN based method, we can conclude that they are both much more efficient than the random choice method, and DQN based design method outperforms the Q-learning based method. The proposed method based on reinforcement learning enables the radar to learn the jammer's strategies according to its own experience and then adopt an optimal strategy to avoid being jammed.

REFERENCES

[1] S. R. J. Axelsson, "Analysis of Random Step Frequency Radar and Comparison With Experiments," IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 4, pp. 890-904, April 2007.

[2] T. Huang, Y. Liu, H. Meng and X. Wang, "Cognitive random stepped frequency radar with sparse recovery," IEEE Transactions on Aerospace and Electronic Systems, vol. 50, no. 2, pp. 858-870, April 2014.

[3] T. Huang, Y. Liu, G. Li and X. Wang, "Randomized stepped frequency ISAR imaging," 2012 IEEE Radar Conference, Atlanta, GA, 2012, pp. 0553-0557.

[4] T. Huang and Y. Liu, "Compressed sensing for a frequency agile radar with performance guarantees," 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), Chengdu, 2015, pp. 1057-1061.

[5] S. Haykin, "Cognitive radar: a way of the future," IEEE Signal Processing Magazine, vol. 23, no. 1, pp. 30-40, Jan. 2006.

[6] R. S. Sutton, A. G. Barto, Reinforcement Learning: An Introduction, London: MIT Press, 1998, pp.216–224.

[7] G. Han, L. Xiao and H. V. Poor, "Two-dimensional anti-jamming communication based on deep reinforcement learning," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2087-2091.

[8] L. Xiao, Y. Li, J. Liu, and Y. Zhao, “Power control with reinforcement learning in cooperative cognitive radio networks against jamming,” Journal of Supercomputing, pp. 3237–3257, Apr. 2015.

[9] Y. Gwon, S. Dastangoo, C. Fossa and H. T. Kung, "Competing Mobile Network Game: Embracing antijamming and jamming strategies with reinforcement learning," 2013 IEEE Conference on Communications and Network Security (CNS), National Harbor, MD, 2013, pp. 28-36.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski, “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, Jan. 2015.