Anti-Jamming Strategy Based on Reinforcement Learning with Sequence Information

Myeong Ro Lee
School of Electronic Engineering
Soongsil University
Seoul 06978, Korea
Email: mrlee1102@soongsil.ac.kr

Yoan Shin†
School of Electronic Engineering
Soongsil University
Seoul 06978, Korea
Email: yashin@ssu.ac.kr

***Abstract*—In this paper, we present an anti-jamming strategy based on an actor-critic model of reinforcement learning, which uses sequence information as model input to improve learning performance compared to the conventional structure. Consequently, in scenarios with enemy's partial-band jamming, we demonstrate that frequency hopping for evasion can be effectively performed by determining the frequency channels based on the actor-critic model, thereby efficiently mitigating the effects of enemy's jamming.**

***Index Terms*—deep learning, reinforcement learning, actor-critic, partial-band jamming, anti-jamming, frequency hopping**

I. Introduction

Effectively mitigating the effects of partial-band jamming is important for the successful communication of allied forces in electronic warfare (EW) environments. In EW, the allied forces are faced with the problem of having no prior information about the patterns and timing of enemy's jamming [1]. In this paper, we apply Reinforcement Learning (RL) to develop an anti-jamming (AJ) policy that adapts to enemy's jamming patterns through interactive experiences in an EW environment, aiming to address the aforementioned problem. Additionally, we propose an AJ model that applies sequence information to improve adaptability to enemy's jamming patterns, and compare the AJ performance of the proposed method with the conventional method. Consequently, this paper demonstrates that RL-based AJ techniques can effectively develop strategies to counter the enemy's jamming patterns, while also providing insights into the application of RL methods.

II. Background

A. Reinforcement Learning

RL is a crucial aspect of machine learning where an agent learns to make decisions by interacting with an environment. Key components of RL include an agent, an environment, actions, states, and rewards. The agent learns a policy which is a strategy for choosing actions based on states, to maximize the total reward over time. This approach is also effective in scenarios where the environment model is unknown or complex [2]. Recent work has included the creation of a RL environment for EW scenarios and the application of deep RL to counter jamming by interacting with the environment [3], [4].

B. Actor-Critic Method

The actor-critic (AC) method is an important approach in the field of RL, which combines the concepts of value-based and policy-based strategies. This method involves two key components: the actor who is responsible for making decisions (choosing actions), and the critic who evaluates the actions taken by the actor. The actor updates its policy based on the feedback from the critic, allowing for more efficient and effective learning. This approach is particularly advantageous in complex decision-making scenarios where the action space is large and dynamic. The AC model's ability to simultaneously learn a policy (actor) and a value function (critic) provides a balanced mechanism for more resilient navigation and adaptation to new and complex environments [2], [5]. The process diagram of the AC method is shown in Fig. 1. In this diagram, $V_\pi(s)$ represents the value function which denotes the value of state s when following the current policy π.

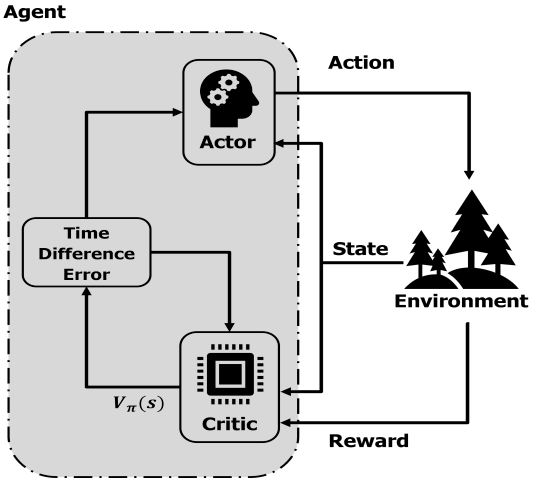

Fig. 1. Process diagram of actor-critic method

†Corresponding author.

This research was supported by the MSIT, Korea, under the ITRC support program (IITP-2023-RS-2023-00258639) supervised by the IITP.

Furthermore, the agent observes the current state s_t of the environment at time t, and selects an action a_t to perform through the actor network, interacting with the environment. Based on the results of this interaction, a reward r_t and the next state s_{t+1} are obtained. The critic network is then used to estimate the value of the current state, and the temporal difference (TD) error is calculated by (1) with the estimated value of the next state s_{t+1}, which is used to adjust the policy of the actor network.

$$\delta = r_t + \gamma V(s_{t+1}) - V(s_t), \tag{1}$$

$$\Delta\theta = \alpha \cdot \delta \cdot \nabla_\theta \log \pi(a_t|s_t, \theta). \tag{2}$$

Here, γ is the discount factor, and (2) represents the policy gradient method for updating the policy of the actor network through the TD error of (1). Consequently, the AC approach represents the process of stable and effective policy learning by improving the policy of the actor network using the critic network [6].

III. Proposed AJ Scheme

The proposed RL-based AJ strategy derives adaptive and effective AJ results through interaction with the pre-configured RL environment. Therefore, this section describes the learning environment of the proposed technique and presents a method that exploits sequence information to improve performance in complex environments.

A. Environment of RL-based AJ

The architecture of the RL environment for training the AJ policy is as follows. It is assumed that the transmission frequency band is divided into N frequency channels, each of which can be in either a jammed or a clear state. Considering an environment where the enemy's jamming patterns are unknown, in the initial time steps, the enemy's jamming targets random channels out of N, and the direction of the sweeping pattern is also randomly determined. In this scenario, the goal of the allied forces is to observe the previous channel state and select a current channel that is clear. Therefore, the size of the environment state and the size of the agent's action space are both N, as expressed in (3) and (4).

$$s_t = \{(f_1, x), (f_2, x), \cdots, (f_n, x)\}, \tag{3}$$

$$a_t = \{f_1, f_2, \cdots, f_n\}, \tag{4}$$

$$r_t(s_t, a_t, s_{t+1}) = \begin{cases} -1, & \text{if } s_{t+1} = (f_n, 1), a_t = f_n. \\ +1, & \text{otherwise} \end{cases} \tag{5}$$

In (3) and (4), n ranges from 1 to N, f_n is separated frequency channels, $x \in \{1, 0\}$ represents the state of the frequency channel, where it is 1 in a jammed situation and 0 in a clear situation. and reward function is represented by (5). If $r_t = -1$, it indicates that AJ strategy is failed to avoid jamming, whereas $r_t = +1$ signifies a successful avoidance of jamming.

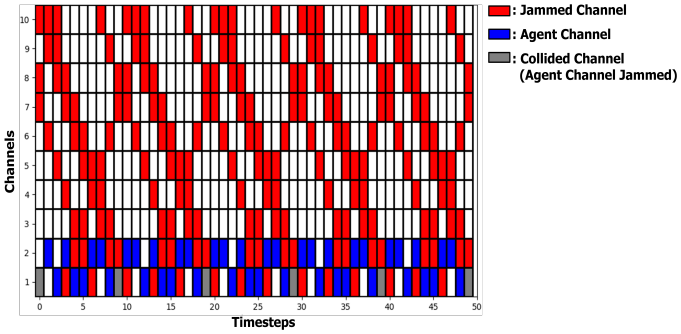

Fig. 2. Environment configuration

B. AC-based AJ

The configured jamming environment is shown in Fig. 2. With such an environment configuration, we train the AJ policy agent based on the AC method.

The multi-layered perceptron (MLP) structure of the actor network is shown in Fig. 3. The size of both input and output is equal to the size of the state space, N, and the hidden network consists of 24 nodes each. The nodes between networks are fully connected, and the activation function applied is rectified linear unit (ReLU). The output is the softmax result of N nodes, which represents the policy of the agent (i.e., the action probability density function). The MLP structure of the critic network is shown in Fig. 4. Similar to the actor network, the input size of the critic network is N, and the hidden network consists of 24 nodes each, using ReLU as the activation function. The output is the estimated value of the state information $V_\pi(s)$, which is the input to the critic network.

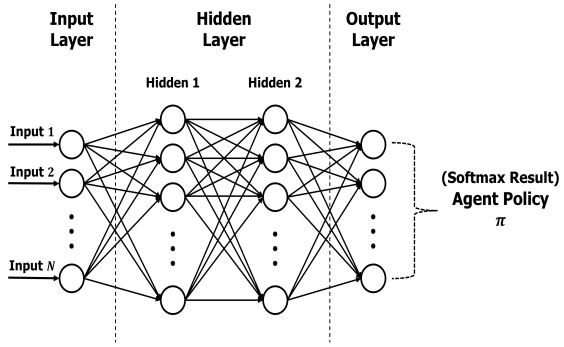

Fig. 3. Actor network architecture

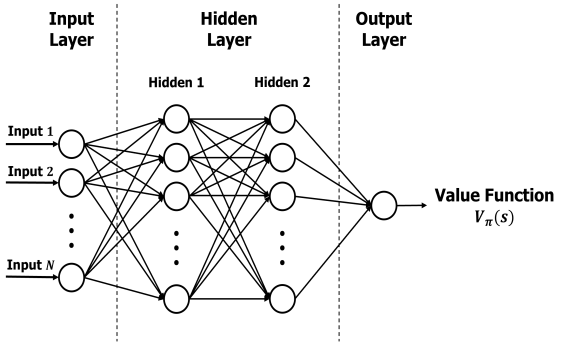

Fig. 4. Critic network architecture

C. Improved Model through Sequence Information

In this paper, we not only propose an AJ strategy based on the AC method, but also configure the model to have better adaptability to changes in jamming patterns by incorporating sequence information at the input layer of the network. The existing model uses the state of N channels as input data to the AC model, which adapts to the enemy's jamming patterns over episodes to select effective AJ channels. However, when the enemy's jamming patterns change frequently over a short period of time, it becomes difficult to improve the performance of the proposed model. Considering this problem, to improve the adaptability of the AJ model, the input data is configured and applied as sequence information consisting of bundles of previous time steps. Sequence information stores state information from the most recent time steps up to a given window size. While the effect of sequence information is less significant at time steps smaller than the window size, as time progresses it becomes more effective in smoothly representing state changes through sequence information.

IV. Simulation Results

In this section, we present the performance of the proposed AJ technique in an environment where the enemy's partial-band jamming ratio is 50%, as well as the performance results with the application of the sequence information. The simulation parameters are given in Table I.

TABLE I
SIMULATION PARAMETERS

Parameters	Values
Number of channels, N	10
Window size of sequence information	5
Number of episodes	100
Number of time steps	200
Learning rate (actor)	15×10^{-4}
Learning rate (critic)	15×10^{-4}
Discount factor, γ	0.99
Smoothing factor, α	0.2

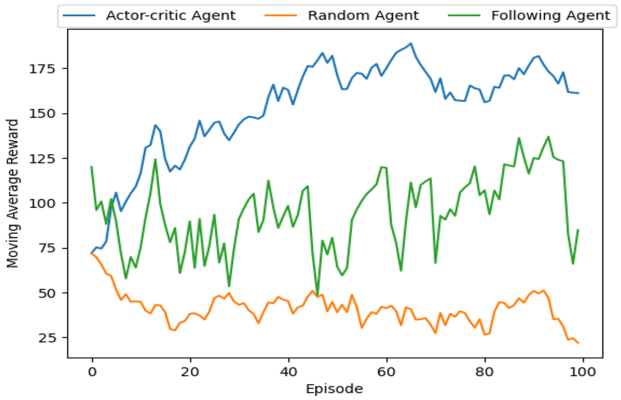

Fig. 5. Moving average reward of AJ policies

The AC based-AJ models present reward results through the exponential moving average (EMA) with a smoothing factor of α, and the EMA at time t is represented by (6).

$$EMA_t = \alpha \cdot R + (1 - \alpha) \cdot EMA_{t-1} \tag{6}$$

Here, R is the reward of each episode (i.e., instant reward of the simulation). The performance of the proposed AJ model in an environment with multiple sweeping jammers is shown in Fig. 5. Here, the random AJ policy selects an arbitrary AJ action from among N actions. The following AJ policy is one which monitors the jamming pattern changes of a particular jammer and adapts the AJ actions according to the pattern changes. The proposed AC-based AJ model shows superior moving average reward scoring results compared to other AJ policies, even when the number of jammers is 50% of the total. It also shows a tendency to reach the maximum reward value of 200 as the episodes progress. Therefore, it can be seen that this is an effective AJ strategy model in a multi-jammer environment.

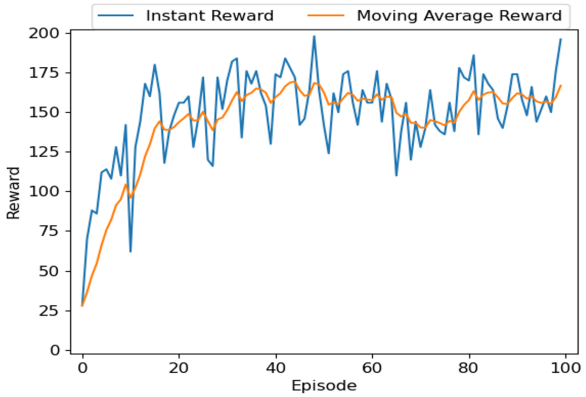

Fig. 6. Performance of the proposed AJ model in complex environment

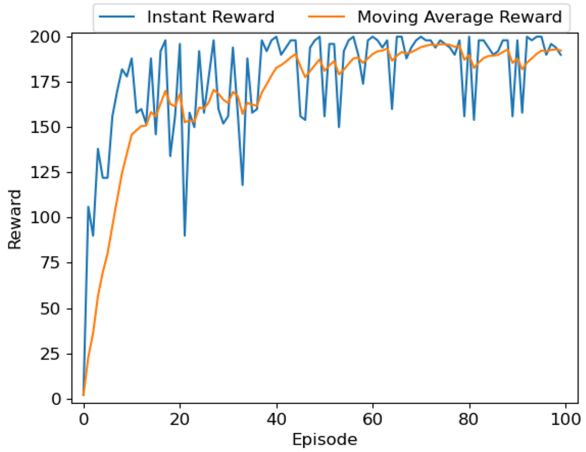

Fig. 7. Performance of the proposed AJ model with sequence information in complex environment

To highlight the performance by adopting the sequence information in the proposed model, a dynamic environment was constructed where the jammer's jamming patterns change

over time steps. This represents a situation that is more difficult to adapt to than the original environment, and takes into account the increased complexity of the jamming. Figure 6 shows the reward performance in the dynamic jamming environment, and Fig. 7 shows the results of the model with the sequence information. Due to the dynamic jamming characteristics of the environment, the moving average result (orange line) in Fig. 6, which represents the result of the simple proposed model, shows a lower score in the multiple jammer environment. On the other hand, Fig. 7 which includes the sequence information in the proposed model, shows that it achieves higher rewards in fewer episodes (i.e. faster learning of the correct policy). In addition, the simple proposed model shows unstable reward results during learning, while the model incorporating the sequence information shows more stable learning. Therefore, incorporating the sequence information into model training can improve the speed and the stability of learning.

V. Conclusion

In this paper, we proposed an RL-based AJ policy that derives AJ strategies against enemy's frequency channel jamming in EW environments. We have compared the proposed model with some AJ strategies and confirmed its effectiveness in avoiding the jamming in a sweeping jamming environment. Considering more complex environments, we used the sequence information as input to the AJ model to ensure performance against dynamic jammers. The AJ model incorporating sequence information demonstrated resilience against dynamic jammers, confirming that the model proposed in this paper is superior and can adopt optimal AJ strategies. In future work, we plan to combine the proposed model with Long Short-Term Memory (LSTM) to enhance adaptability to changes in jamming patterns. This approach will discuss the application of RL and LSTM models in EW environments, presenting advancements towards a more secure wireless communication environment for allied forces.

References

[1] Y. Wu, B. Wang, K. J. R. Liu, and T. C. Clancy, "Anti-jamming games in multi-channel cognitive radio networks," *IEEE Jour. Selected Areas Commun*., vol. 30, no. 1, pp. 4-15, Jan. 2012.

[2] R. S. Sutton, and A. G. Barto, *Reinforcement Learning: An Introduction*, Ch. 11, The MIT Press, 2018.

[3] A. S. Ali, W. T. Lunardi, L. Bariah, M. Baddeley, M. A. Lopez, J.-P. Giacalone, and S. Muhaidat, "Deep reinforcement learning based anti-jamming using clear channel assessment information in a cognitive radio environment," *Proc. CommNet 2022*, pp. 1–6, Virtual Conference, Dec. 2022.

[4] J. Qi, H. Zhang, X. Qi, and M. Peng, "Deep reinforcement learning based hopping strategy for wideband anti-jamming wireless communications," *IEEE Trans. Veh. Technol*., pp. 1-12, Early Access, Oct. 2023.

[5] O. Dogru, K. Velswamy, and B. Huang, "Actor-critic reinforcement learning and application in developing computer-vision-based interface tracking," *Engineering*, vol. 7, no. 9, pp. 1248-1261, Sept. 2021.

[6] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev*., vol. 42, no. 6, pp. 1291-1307, Nov. 2012.