sensors

MDPI

Article

Research on Efficient Reinforcement Learning for Adaptive Frequency-Agility Radar

Xinzhi Li * and Shengbo Dong

Beijing Institute of Remote Sensing Equipment, Beijing 100854, China; shbdong@aliyun.com
* Correspondence: lxzxin@foxmail.com; Tel.: +86-68761276

Abstract: Modern radar jamming scenarios are complex and changeable. In order to improve the adaptability of frequency-agile radar under complex environmental conditions, reinforcement learning (RL) is introduced into the radar anti-jamming research. There are two aspects of the radar system that do not obey with the Markov decision process (MDP), which is the basic theory of RL: Firstly, the radar cannot confirm the interference rules of the jammer in advance, resulting in unclear environmental boundaries; secondly, the radar has frequency-agility characteristics, which does not meet the sequence change requirements of the MDP. As the existing RL algorithm is directly applied to the radar system, there would be problems, such as low sample utilization rate, poor computational efficiency and large error oscillation amplitude. In this paper, an adaptive frequency agile radar anti-jamming efficient RL model is proposed. First, a radar-jammer system model based on Markov game (MG) established, and the Nash equilibrium point determined and set as a dynamic environment boundary. Subsequently, the state and behavioral structure of RL model is improved to be suitable for processing frequency-agile data. Experiments that our proposal effectively the anti-jamming performance and efficiency of frequency-agile radar.

Keywords: radar anti-jamming; reinforcement learning (RL); frequency-agility radar; Markov decision process (MDP); Markov game (MG)

Citation: Li, X.; Dong, S. Research on Efficient Reinforcement Learning for Adaptive Frequency-Agility Radar. *Sensors* **2021**, *21*, 7931. https://doi.org/10.3390/s21237931

Academic Editor: Ali Khenchaf

Received: 18 October 2021
Accepted: 24 November 2021
Published: 27 November 2021

1. Introduction

Frequency domain is an important field of radar electronic countermeasure (ECM). Because frequency agile radar has excellent anti-jamming performance and high range resolution, it is used in electronic countermeasures. With the advancement and progress of cognitive interference methods [1,2], the need to enhance the radar anti-jamming performance has increased. Frequency agile radar needs to have the ability to "intelligently" [3] change the radar frequency and adaptively select countermeasures [4] according to the tasks performed in the actual working environment. The design of the existing anti-jamming processing strategy relies on expert experience, and the working mode unable to change along with the environment and the target. When confronted with complex interference scenarios, artificially designed anti-jamming frequency-agility methods become cumbersome and difficult to implement, resulting in a decrease in radar anti-jamming performance. Therefore, it is an inevitable trend for frequency agile radar to have environmental perception and intelligent anti-jamming capabilities [5].

After Simon Haykin [6] first proposed the concept of cognitive radar in 2006, brain science [7] and artificial intelligence [8] were introduced into the radar system to adapt it to the complex and changing electromagnetic environment. Some studies [9,10] use support vector machine (SVM) [11] to identify and classify radar anti-jamming behavior to improve the anti-jamming effect of radar. Compared with traditional methods, this kind of methods improves the processing accuracy. Nevertheless, those methods depend on the prior knowledge to select the kernel function. In 2018, Liu et al. [12] proposed the use of deep learning (DL) [13] algorithm to evaluate the environment based on sample

data, which was used as a basis for decision-making and achieved good results. However, methods based on deep learning rely on a large number of sample data, and cannot obtain good generalization performance in scenarios, e.g., incomplete information games. In 2017, Deligiannis et al. [14] adopted a Bayesian game theory framework to maximize the signal-to-interference-plus-noise ratio (SINR) between the radar and the target, and verified the convergence of the algorithm model through simulation results. In 2020, Garnaev et al. [15] used the Bayesian game model to construct an incomplete information relationship between the radar and the jammer, and optimized the performance of communication and radar components. Although the Bayesian game model does not need to know the objective function of the agent, it needs to master the distribution function and determine the optimal strategy through a specific probability. In 2010, Li et al. [16] tried to use RL [17] algorithm in cognitive radio countermeasures. After that, reinforcement learning (RL) algorithms were introduced and applied to radar systems to enhance radar anti-jamming performance. For example, in 2017, Qiang et al. [18] used the classic RL Q-learning (QL) algorithm [19] to make optimal allocation decisions for radar transmit power. This algorithm quantifies the performance of radar anti-jamming decision-making by establishing QTable [20] to record the state of the model. In 2019, Li et al. [21] proposed a deep Q-network (DQN) [22] based frequency agile radar strategy design method, which uses a neural network to approximate the anti-interference decision value function to guide the radar to select the optimal pulse carrier frequency. In 2019, Aref et al. [23] used the improved Double Deep Q-network (DDQN) [24] to conduct autonomous anti-jamming research in the radar frequency domain, and solved the problem of local anti-jamming optimal decision-making. In 2020, Ak et al. [25] studied the cognitive radar anti-jamming technology under the POMDP model, and quantified the performance of the radar being jammed through the conventional signal-to-noise ratio (SNR). Then use DQN and long-short term memory (LSTM) [26] network to find the optimal solution for the two frequency hopping strategies of the radar, which improves the anti-jamming performance of the radar.

RL is an interactive trial-and-error learning method, which can be applied to real-time update scenarios, e.g., online learning [27]. Through continuous trial and error, rewards from the environment can be used to continuously adjust one's own behavior without requirement for a large amount of prior knowledge of the environment. This is an online learning method that can be applied to the actual environment, so it has been extensively studied [28–30]. However, existing RL algorithms rely on certain prerequisites, that is, the agent can conduct sufficient experiments in the environment, which is the experimental environment is Markov decision process (MDP) [31]. In this case, RL can ensure the convergence of the optimal strategy. The radar system does not satisfy MDP in two aspects: on one hand, the system contains multiple agents, e.g., radar and jammers, which is beyond the scope of application of RL based on MDP and violates the assumption of algorithm convergence. On the other hand, the decision action taken by the radar is not a typical sequence decision process [32], and there is the possibility of step changes. That is, from the current parameter action with weak anti-interference ability to the parameter action with best anti-interference ability, the relationship between each action is independent. It is also not satisfied with MDP. Therefore, it is difficult for algorithms based on existing RL models to obtain the correct evaluation of current strategies from the radar system environment. This will lead to poor convergence and low computational efficiency of the algorithm in the environment, which cannot meet the high-precision and high-real-time requirements.

In conclusion, existing intelligent algorithms cannot be directly applied to radar systems: both SVM and DL algorithms require a great quantity of prior knowledge and do not have online learning capabilities; the Bayesian game theory model relies on the distribution function; RL online learning can obtain global optimal decisions, but basic theoretical models and frameworks are difficult to apply to radar systems. In response to the above problems, combined with the characteristics of frequency-agility of frequency agile radar, this paper proposes an efficient and adaptive RL anti-jamming model for frequency-agile radar. Firstly, the Markov game (MG) [33] is introduced, which transforms

the radar-jammer detection model into a reinforcement learning model that can satisfy the MG theory. Secondly, the state and behavioral framework structure of the original RL model is improved, so that the new framework structure can be applied to the frequency agile data, and the convergence of the framework is demonstrated through the Bellman equation. In the experiment, based on the frequency agile radar system environment, this paper compared the classical RL algorithms in the ideal MDP framework (ideal model, IM) and the improved framework (AM) proposed in this paper. According to the quantitative analysis of mean square error (MSE) and convergence speed, compared with the basic RL model, the algorithm model proposed in this paper can improve the anti-jamming performance and processing efficiency of frequency agile radar.

2. Reinforcement Learning Modeling Based on Radar System

2.1. Radar Countermeasure System Model

The environment of modern radar system usually contains a complex interference environment, which prevents the detection radar from obtaining accurate target information. As shown in Figure 1, the system model consists of two agents (frequency agile radar for detecting targets and airborne jammer): the radar obtains the echo signal through a receiver, and evaluates the transmit detection signal through the effectiveness evaluation. Then, it makes a decision and launchs a new anti-jamming detection signal. The jammer receives the radar detection signal and performs real-time interference processing.

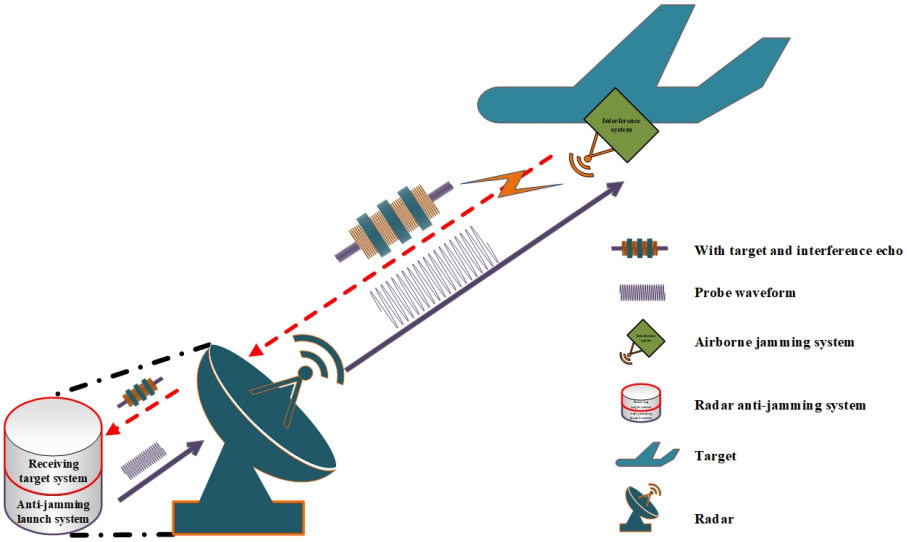
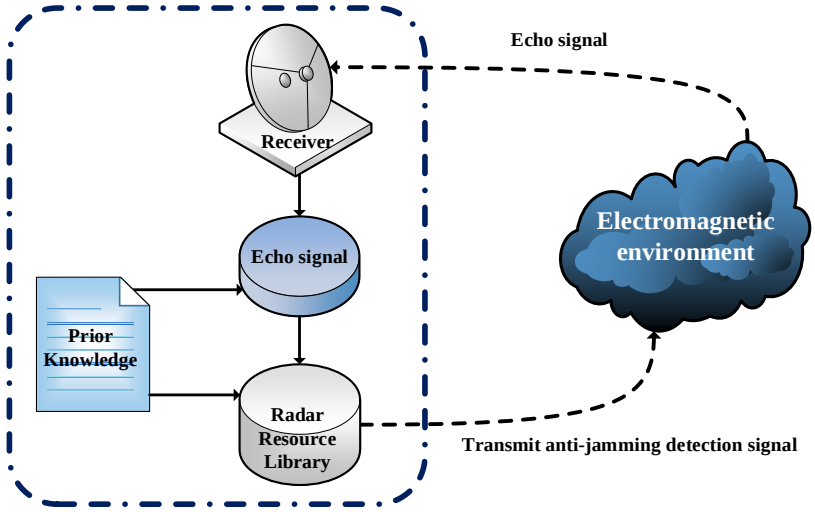


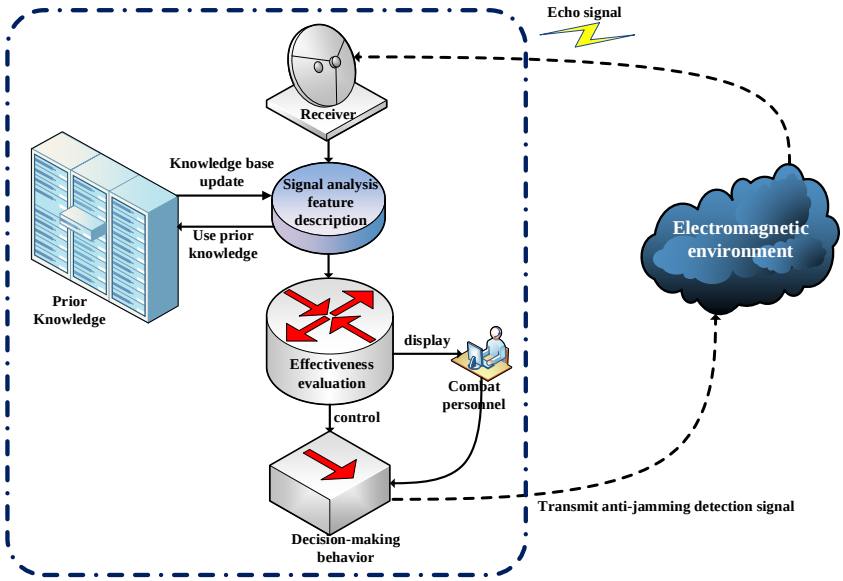
Figure 1. Anti-jamming principle diagram based on frequency agile radar radar and airborne jamming platform.

Traditional frequency agile radar usually adopts "cover pulse + tracking signal" composite detection waveform [34,35] to achieve anti-jamming measures. That is, the radar transmitter transmits two kinds of signals. One is a cover pulse with obvious spectral characteristics, which is used to guide the jammer to lock the frequency and waveform of the jamming signal to the cover pulse. The other is a tracking signal with low interception characteristics for real detection and tracking functions. The tracking signal gets staggered with the cover pulse and reduces the effective power of the jamming signal into the radar machine, so as to achieve the anti-jamming effect. The traditional radar anti-jamming system is shown in Figure 2a. First, the radar receiver receives echo signals from the electromagnetic environment to obtain specific measurement results of target and interference signal parameters. Then, based on the detected and existing prior knowledge base, e.g., pulse width, modulation method and repetition period of the target signal and the interference signal, which are used to determine the type of the interference signal. Finally, combine the target and interference signal measurement results to schedule anti-

interference resources, select effective anti-interference cover pulse and tracking signal, and transmit them through the transmitter.



(**a**)



(**b**)

Figure 2. Comparison chart of radar system anti-jamming process. (**a**) Traditional radar anti-jamming model, (**b**) Intelligent radar anti-jamming model.

According to the anti-interference characteristics of frequency agile radar, jammers usually use synchronous targeting jamming [36,37] in suppressive jamming for signal jamming. As shown in Figure 3, the jammer performs frequency targeting on each radar pulse received. Then, it interferes with the current radar signal and stops after ΔT_1 time. After the next pulse arrives, the jammer will perform frequency aiming again, and stop after jamming for a period of time ΔT_2, and so on.

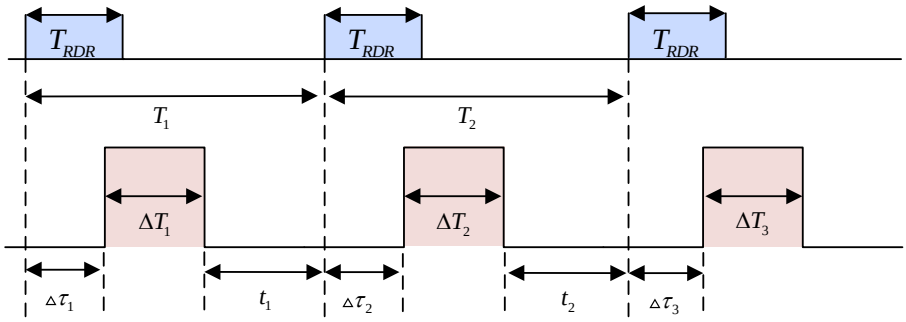


Figure 3. Synchronous aiming jamming timing diagram.

In Figure 3, the pulse repetition period of the T_1 and T_2 frequency agile radar. The distance between the radar and the target is determined by the propagation speed c and the waveform propagation time T_{RDR}. Here, T_{RDR} is used to describe the distance between the target and the radar. $\Delta\tau_1$, $\Delta\tau_2$ and $\Delta\tau_3$ are the delay time of each pulse aiming frequency. ΔT_1, ΔT_2 and ΔT_3 are the real-time jammers based on the aiming results Interference time; t_1 and t_2 are the stop interference time.

Synchronous targeting jamming technology can effectively interfere with frequency agile radar. Moreover, it is irrelevant to the frequency agile modulation method. When the frequency agile radar system detects that the interference signal parameters do not exist in the prior expert knowledge, it cannot dynamically adjust the anti-jamming detection radar signal pattern based on the anti-jamming effect feedback obtained from the environment. In this situation, the anti-jamming effect of radar deteriorates drastically, and the processing efficiency reduces greatly.

Compared with traditional radar system relying on a static prior knowledge base to make decisions, intelligent radar anti-jamming system has the ability to dynamically perceive the external environment and optimize the knowledge base. The system can dynamically adjust the anti-jamming measures based on the feedback of the anti-jamming measures from the external environment (whether the target is tracked, whether the jamming signal is locked to the shield pulse). The intelligent radar anti-jamming system is shown in Figure 2b. First, the echo signal is received from the receiver, and the specific measurement results of the target and interference signal parameters are obtained. Then, the interference pattern is analyzed according to the prior knowledge base. In this case, if the interference pattern does not exist in the prior knowledge base, it is necessary to identify and analyze the situation of the target and the interference signal, and evaluate the current anti-jamming effectiveness. The evaluation content includes the accuracy evaluation of the target information and whether the jamming signal is locked to the shielding pulse, etc. Finally, the anti-jamming method is adjusted according to the results of the effectiveness evaluation and launched. In addition, add the newly detected interference pattern feature parameter information and effectiveness evaluation results into the prior knowledge base. The intelligent radar anti-jamming system continuously interacts with the external environment to learn online anti-jamming strategies that are not in the prior knowledge base. The system can finally form a set of optimal strategies for a specific environment based on the accumulation of experience.

Contrapose different radar system environments in the intelligent radar anti-jamming system, we have established a RL model suitable for a single radar system and a RL model suitable for a radar-jammer system environment. We analyze the reasons why the existing basic theoretical models and frameworks of RL are not suitable for the environment of radar-jammer system, and propose an efficient RL model suitable for the system.

2.2. Reinforcement Learning Model for Single Radar System

MDP is the theoretical derivation basis of single-agent RL [38], and solving RL problems depends on the framework. As shown in Figure 4a, the Markov decision process is a tuple <s, a, t, r> composed of four elements, where s represents the limited set of all states contained in the environment by the agent, a represents the limited set of all actions that the current agent will take in the environment, T represents the transformation equation, and r is the reward equation. The decision process sequence of the agent under the MDP framework can be expressed as $\{ s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, \ldots \}$, where r_t and s_{t+1} only depend on the previous state s_t and action a_t, and t represents time.

The frequency agile radar system based on the RL theory first launches the frequency agile detection signal source strategy a_t^{RDR}. Then the echo signal is received by the receiver. Finally, according to the current target state s_t^{RDR} and the detection result evaluation r_t^{RDR}, select and transmit the next anti-interference agile detection signal source strategy a_{t+1}^{RDR}. The radar system obtains target status information by continuously interacting with the environment. According to the evaluation result of the target status information, the transmitted frequency signal is changed to improve the radar's anti-jamming detection target ability in real time.

Early radar systems were simple and usually contained only a radar and a target. As shown in Figure 4b, the simple single radar-target environment system satisfies the MDP condition, which can be analogous to the RL theoretical model. The radar in the radar system which is regarded as an agent, and together with the target and its environment, it is regarded as a complete MDP environment. Its parameters are defined as shown in Table 1. During the interaction between the radar and the environment, the goal of the radar is to adjust the emitted signal source to maximize the detection of the target's position, speed and acceleration in the environment.

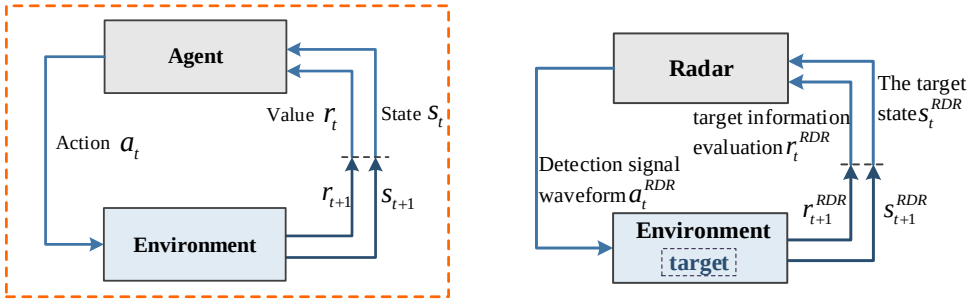


Figure 4. Comparison chart of basic RL theory model and single radar system model. (**a**) Basic reinforcement learning model framework, (**b**) Intelligent radar anti-jamming model.

Table 1. Parameter definition of RL model based on single radar system.

RL Parameters	Radar System Parameters
Agent	Radar.
Environment	Target environment.
State	The status of the detected target s_t^{RDR}, which contains target speed, angle and distance information.
Action	Frequency agile anti-jamming cover pulse and tracking signal source emitted by radar a_t^{RDR}, which contains radar amplitude A, pulse width τ and pulse repetition period (PRF) β.
Reward	Obtain target information accuracy evaluation r_t^{RDR}.

This paper gives the parameter definition of RL model based on single radar system: Assuming a discrete time sequence $t = 0, 1, 2, 3...$, at each time t, the radar can detect the state s_t^{RDR} of a target from the environment. Define the signal source taken by the radar at time t as action a_t^{RDR}. At the next moment, s_{t+1}^{RDR} is defined as the received value return r_{t+1}^{RDR} as the result of the radar taking action a_t^{RDR} to evaluate whether the detected target is accurate. At the next moment, s_{t+1}^{RDR} defines the result of the radar taking action a_t^{RDR} in the current state s_t^{RDR}, and assessing whether the detected target is accurate and other information is defined as taking action a_t^{RDR} to obtain value return r_{t+1}^{RDR}. At each moment, the radar completes the mapping from the state s_t^{RDR} to the selection probability of each possible signal action a_t^{RDR}. This mapping relationship becomes the radar strategy, denoted as π_t^{RDR}, $\pi_{s,a}^{RDR}$ which is the probability of selecting a_t^{RDR} when the state is s_t^{RDR}.

Through the above analysis, a single radar system can be analogous to a RL theoretical model. However, when there are multiple agents in the radar system environment (radar, jammer, and target), the MDP-based RL model framework cannot be directly applied. It is necessary to introduce new theories and conditional constraints to ensure the convergence of the algorithm model.

2.3. Reinforcement Learning Modeling for Radar-Jammer System

With the rapid development of information technology and its widespread application in the military field, the application environment of modern radars has become increasingly complex. Usually, the combat environment contains radar and jammer, and there is an interference effect between the two to prevent the detection radar from obtaining accurate information about the target. In the above scenario, the existing RL algorithm model cannot be directly applied to the radar-jammer countermeasure system. RL is developed based on the MDP theory. It is a strategy for a single agent to learn the possible delayed return signal maximization strategy in a random static environment, e.g., QL [18], DQN [21], DDQN [24], DQN+LSTM [25] and other classic algorithm models. This type of algorithm relies on certain prerequisites, that is, the agent can perform sufficient experiments in the environment, and the experimental environment is MDP, then RL can ensure the convergence of the optimal strategy.

The radar-jammer system does not satisfy the MDP theory in two aspects: On one hand, the rader-jammer system is beyond the scope of MDP theoretical application. The main reason is that in this kind of environmental system, the optimal strategy of the agent not only depends on the environment, but also on the strategies adopted by other agents, which violates the assumptions required to ensure convergence. When agents have opposing targets, there may not be a clear optimal strategy, resulting in frequent retrieval of the balance between multiple agents, making the environment unstable. On the other hand, the decision-making action taken by the radar is not a typical sequence decision-making process, and there is the possibility of step changes. That is, a step from the current parameter action with weak anti-interference to the parameter action with good anti-interference, and the relationship between each action is independent, which also does not satisfy the MDP theory. Therefore, in the complex electromagnetic environment of the radar system with multiple agents coexisting, it is difficult for the existing RL algorithm to obtain the correct evaluation of the current strategy from the environment, resulting in poor algorithm convergence and unable to satisfy the real-time requirements of military radars.

The combination of interdisciplinary game theory and artificial intelligence RL is the theoretical basis of solutions in the field of multi-agent reinforcement learning (MARL) [31]. The characteristic of a multi-agent system is the interaction of strategies between multiple agents. Each agent has independent goals and decision-making capabilities, and at the same time, the agent is affected by the behavioral decisions of other agents. When the radar is in a multi-agent environment, the value return of the radar is affected by the behavioral decisions of the jammer. According to MG theory, the radar and the jammer independently choose actions to form a joint action. If the learning model of multi-agents in the same environment is adopted, the process is shown in Figure 5, e.g., the actions a_1 and a_2 taken

by the radar and the jammer are regarded as shared actions, and joint actions $\bar{a}_t$ are adopted to obtain joint state $\bar{s}_t$ and value return $\bar{r}_t$.

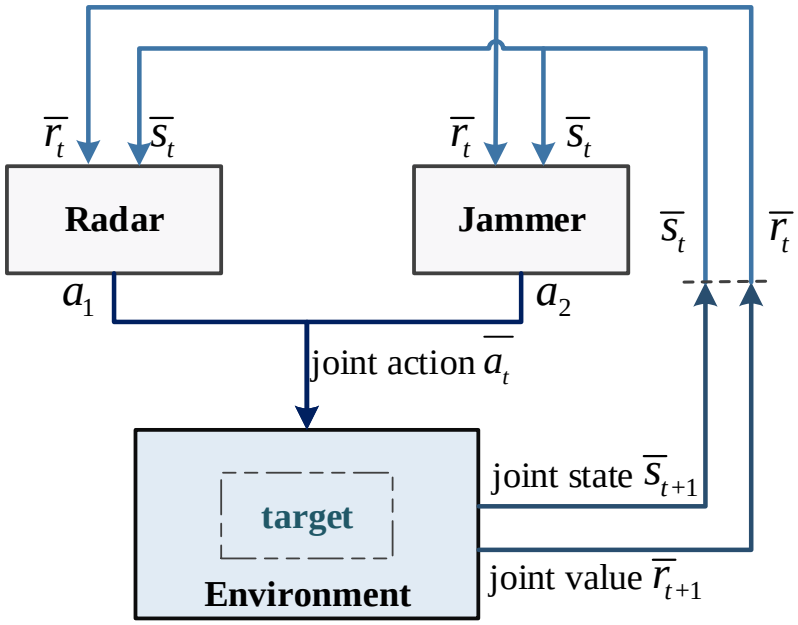


Figure 5. Environment model of radar system based on MARL.

However, in reality, the radar and the jammer cannot know the status and action strategy of the other side, and should regard the other side as part of the environment. As shown in Figure 6, this paper proposes a radar-jammer system environment model, which regards the radar and the jammer as two groups of agent environments: radar is environment1 (Envi1), jammer and target are composed of environment2 (Envi2).

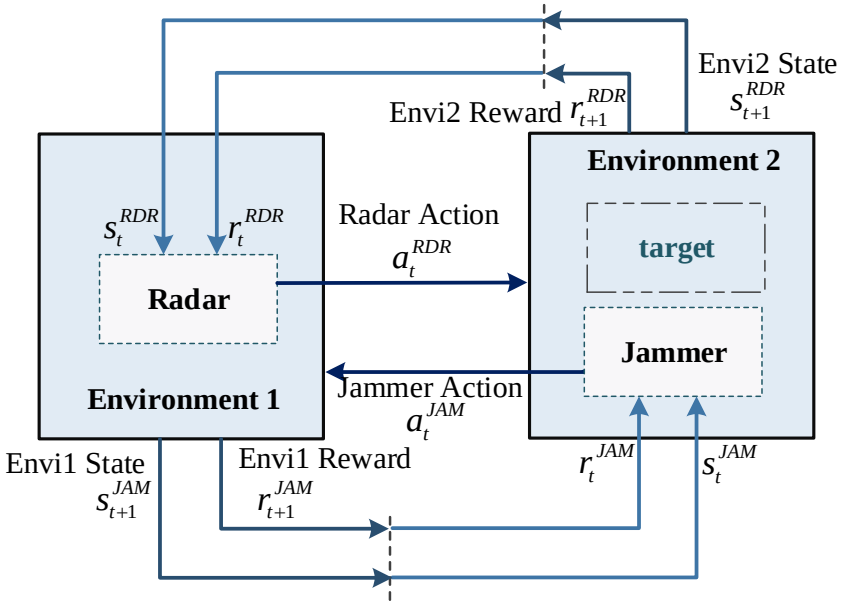


Figure 6. The environment model of radar-jammer system proposed in this paper.

The parameter definitions of the radar side and the jammer are shown in Table 2. For the radar side, the environment of the jammer and the target is regarded as Envi2. The radar sends a signal a_t^{RDR} to Envi2, from which a target state s_t^{RDR} with jamming

information and a return value r_t^{RDR} to evaluate the accuracy of the target are obtained. Then the transmitting signal behavior a_t^{RDR} of the radar is adjusted to detect the accurate target information.

Table 2. Definition of radar parameters of radar-jammer system environment model.

RL Parameters	Radar System Radar Side's Parameters
Agent	Radar.
Environment	Target and jamming radar environment.
State	Status of detected targets with interference information s_t^{RDR}, which contains target speed v, angle θ and distance d information, target error rate ρ_d.
Action	Frequency agile anti-jamming cover pulse and tracking signal source emitted by radar a_t^{RDR}, which contains radar amplitude A, pulse width τ and pulse repetition period (PRF) β.
Reward	Obtain target information accuracy assessment and whether the jamming signal locks the cover pulse r_t^{RDR}.

Where, the target error rate ρ_d refers to the false target probability that the target in the echo signal detected by the radar front-end detector is the jammer. The jammer uses the adaptive radar to interfere with the synchronized aiming jamming technology in Section 2.1. Nevertheless, it cannot optimize interference strategies based on environmental feedback.

This section proposes a radar-jammer system environment model suitable for the actual radar system environment. The theoretical basis of this model will be analyzed in Section 2.4.

2.4. Markov Game Inference Applicable to Radar-Jammer System Environment

MDP includes one agent and multiple states. Radar-jammer system is a typical special case of multi-agent systems, and MDP theory cannot be applied to the multi-agent systems. For a game with multiple agents and multiple states, Markov game (MG) is defined. MG uses a tuple to represent $\{ n, s, a_1, \ldots, a_n, T, t_1, \ldots, t_n \}$ where n represents the number of multi-agent in the current environment, T represents the transfer function, a_i $(i = 1, ..., n)$ represents the behavior set of the ith agent, and r_i represents the reward function of the agent i. The transfer function T in MG refers to the probability distribution of the next state when the current state and joint behavior of the agent are given. The reward function r_i $(s, a_1, ..., a_n, s')$ indicates that the agent i takes the joint in the state s. After the behavior $(a_1, ..., a_n)$, the reward is obtained in the next state s'.

Combining with MG, give the definition of a radar-jammer countermeasure system. Assuming that the system contains a radar and a jammer, a tuple can be expressed as $\{2, s^{RDR}, a^{RDR}, a^{JAM}, T, r^{RDR}, r^{JAM}\}$:

1. 2 refers to the number of agents in the radar-jammer system, including the radar and the jammer;
2. $s = \{s^{RDR}, s^{JAM}\}$ indicates the joint state of the radar-jammer system model, s^{RDR} indicates the current state of the radar side, and s^{JAM} indicates the state of the jammer;
3. a^{RDR} represents the action set of the radar, and a^{JAM} represents the action set of the jammer;
4. T represents the transfer function, which represents the probability of the radar side taking action a^{RDR} and the jammer taking action a^{JAM} under the current state and then transferring to the next state;

5. r^{RDR} represents the reward function of radar; r^{JAM} represents the reward function of jammer. Among them, the radar side and the jammer side have independent reward function R. Different agents can switch from the same state to different rewards.

For judging whether the MG is applicable to the model, the focus is to determine whether there is a Nash equilibrium in the model. The core of game theory is to establish a strategy interaction model for the game between multi-agent. Each game is equivalent to a mathematical model, which is used to describe the interaction results of each agent's reward strategy. Combined with a radar-jammer system, the process of target detection and recognition by the radar each time is a game. Each target detection is performed by selecting a group of signals a^{RDR} from the set of transmitted radar waveform signals (as actions a^{RDR}) by the radar.

Radar-jammer systems are different from games, e.g., Go [21]: the mutual gains of the radar and the jammer will become part of the other party's losses. However, in the actual situation, the two parties are not sure that they will bring losses to the other party, which is a general game and there is a Nash equilibrium. As shown in Table 3, taking two actions each of the two agents as an example, analyze the Nash equilibrium between the detection target radar and the jammer: The vertical radar side takes actions a_i^{RDR} $(i = 1, 2)$, and the horizontal jammer side takes actions a_i^{JAM} $(i = 1, 2)$. The two have their own optimal decision-making actions a_i^{RDR} and a_i^{JAM}, respectively.

Table 3. Radar and jammer matrix game example.

	a_1^{JAM}	a_2^{JAM}
a_1^{RDR}	(1, 1)	(1, 2)
a_2^{RDR}	(2, 1)	(2, 2)

Combining the Nash equilibrium theorem, as shown in Equation (1), set $\pi = (\pi_{RDR}, \pi_{JAM})$ to be the joint strategy of radar-jammer system, π_{RDR} to indicate the radar strategy, and π_{JAM} to indicate the jammer strategy.

$$r_{RDR}(\pi_{JAM}, \pi_{RDR}^*) \geq r_{RDR}(\pi_{JAM}, \pi_{RDR}')\forall\pi_{RDR}' \in \mu(a^{RDR}), \tag{1}$$

where π_{RDR}^* represents the strategy implemented by the jammer in the system except for the radar side, r_{RDR} represents the radar return function, $\mu(a^{RDR})$ represents the action set of the radar, and represents the probability distribution set on the action set a^{RDR} of the radar. When Equation (1) holds, it means that strategy $\pi_{RDR}^* \in \mu(a^{RDR})$ is the optimal strategy of the radar side. Here (π_{JAM}, π_{RDR}') indicates that when the jamming strategy π_{JAM} is fixed, the radar adopts an arbitrary strategy $\forall\pi_{RDR}' \in \mu(a^{RDR})$. (π_{JAM}, π_{RDR}^*) means that when the jamming strategy π_{JAM} is fixed, the strategy π_{RDR}^* adopted by the radar is the optimal strategy. In the game process, when the radar side makes the optimal decision, the interference side keeps its strategy unchanged, the current radar side cannot further improve its return, and the Nash equilibrium is reached at this time. According to this, we give the definition of the Nash equilibrium of the radar-jammer system: in the game model $G = \{\pi_{RDR}, \pi_{JAM}: \mu(a^{RDR}), \mu(a^{JAM})\}$, if the strategy combination of radar and jamming $(\pi_{RDR}^*, \pi_{JAM}^*)$, the strategy π_i^* $(i = RDR, JAM)$ of either party is the best strategy for the other party's strategy π_j^* $(j \neq i)$. It is said that $(\pi_{RDR}^*, \pi_{JAM}^*)$ is a Nash equilibrium of the radar-jammer system. Because the relationship between radar and jammer is confrontation, the two parties cannot adopt the optimal strategy at the same time. Therefore, the (a_2^{RDR}, a_2^{JAM}) in Table 3 does not exist in the radar-jamming system. Regarding this point, we set (a_2^{RDR}, a_1^{JAM}) to be the Nash equilibrium state of the radar's successful anti-interference, (a_1^{RDR}, a_2^{JAM}) is the Nash equilibrium state of the jammer's successful interference.In the process of the game between the radar and the jammer, there

exists a Nash equilibrium of one side's optimal strategy. At this time, the MG theory can be used in the radar-jamming system environment model proposed in this paper.

3. An Improved RL Framework for Radar-Jammer System

The frequency signal of frequency agile radar changes in steps rather than continuously. As shown in Figure 7, the jump target of the radar is to take action a_0 from initial state s_0 to final state s_7 where the definition of radar state s_i $(i = 1, ..., 7)$ is shown in Table 4. When the radar takes different actions a_i $(i = 0, 1, 2, ..., 13)$, it will jump from the current state s to other states s' with a certain probability. However, the difference from the Markov chain is that certain existing states and behaviors are not necessarily paths, as shown by the six paths in Equations (2)–(7). Because in the radar system environment, the echo signal is affected by the interference signal. At this point, according to the complete Markov decision theory, it can be obtained from the value echo r of the environment that taking behavior a_0 in the s_0 state is the best strategy. However, the radar-jammer system environment is based on the MG theory, and the radar and the jammer affect the behavior strategy. It is difficult for the radar side to implement behavior a_0.

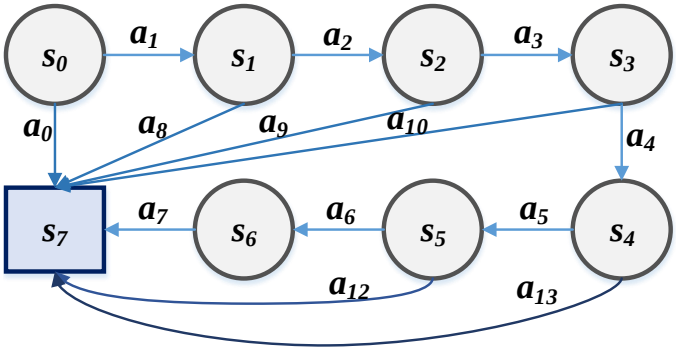


Figure 7. Radar system state-action transition diagram.

$$s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_3 \xrightarrow{a_4} s_4 \xrightarrow{a_5} s_5 \xrightarrow{a_6} s_6 \xrightarrow{a_7} s_7 \tag{2}$$

$$s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_3 \xrightarrow{a_4} s_4 \xrightarrow{a_5} s_5 \xrightarrow{a_{12}} s_7 \tag{3}$$

$$s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_3 \xrightarrow{a_4} s_4 \xrightarrow{a_{13}} s_7 \tag{4}$$

$$s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_3} s_3 \xrightarrow{a_{10}} s_7 \tag{5}$$

$$s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_2} s_2 \xrightarrow{a_9} s_7 \tag{6}$$

$$s_0 \xrightarrow{a_1} s_1 \xrightarrow{a_8} s_7 \tag{7}$$

Table 4. Radar status description.

s_0	Radar initial state, which is the first radar echo signal, is described by target error rate ρ_d, $\rho_d \geq threshold$.
s_1–s_6	Radar intermediate state, $\rho_d \geq threshold$.
s_7	Radar final state, $\rho_d < threshold$.

For the above problems, this paper improves from two aspects. Firstly, the state and behavior frame structure of the original RL model is improved to make the new frame structure suitable for the frequency agile data. Then, it is proved that the state value function satisfies the Bellman equation in the frame structure.

3.1. Improved RL Framework

Value evaluation is a method used to connect optimal criteria and strategies in RL. The agent needs to calculate the optimal strategy based on the value evaluation obtained from the environment feedback. In the application of RL based on radar system, J/R [39,40] is usually used to evaluate the value of the state. At this time, the following two problems exist in the continuous state of RL to build the model. On one hand, in the radar-jammer system environment, the jammer, which making the radar status and the strategy adopted cannot be directly correlated, affects the radar status and the status is difficult to quantitatively evaluate. On the other hand, the radar signal changes step by step and needs to be evaluated based on the radar echo signal, which has time-delay. To solve the above problems, the state transition model of RL is improved in this paper. Sets the echo signal received by the radar to state s, removes the intermediate state, and only retains the initial state and the final state. As shown in Figure 8, in the initial state s_0 of the radar, the final state s_7 is reached only when the action a_0 is taken, and the initial state s_0 is reached when other actions a_i ($i = 1, 2, ..., 13$) are taken; each state generated in the intermediate process is regarded as the current initial state s_0.

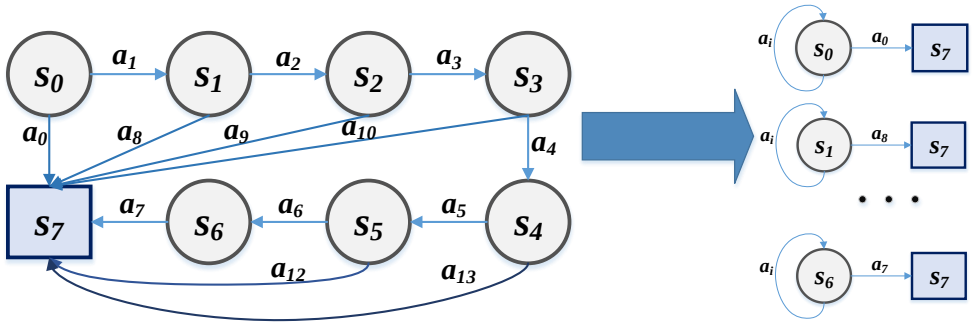


Figure 8. Suitable for frequency agile scenario state-action transition diagram.

The proposed method can effectively reduce the influence of the jammer and avoid the strategic association with the irrelevant state of the middle part. It is suitable for radar frequency-agility scenarios. This framework satisfies the MDP, reduces the data noise interference contained in the intermediate state, effectively enhances the robustness of the model, and greatly improves the training efficiency and inference performance of the model.

3.2. Bellman Equation in Frequency-Agility Scenario

The foundation and core of RL is Bellman equation. In the process of obtaining the optimal strategy, the fundamental goal of RL is to optimize the strategy function $\pi(s)$, thereby maximizing the value function. The state value function is one of the criteria for evaluating the quality of the strategy function. In the basic model of RL, each state s ($s \in S$, where S is the set of all states) can have multiple choices of action a ($a \in A$ where A is the set of all actions). When action a is performed, the system will transition to another state s' according to a certain probability $P^a_{ss'}$. At this point, in this framework, each state s is related through action a, and there is a connection between the states. Under this precondition, the state value function can effectively evaluate the strategy function and form a recursive form, as shown in the following Equation (8):

$$\begin{aligned} V^{\pi}(s) &= E^{\pi}[R_t|s_t = s] \\ &= E^{\pi}\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s\} \\ &= E^{\pi}\{r_{t+1} + \gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_t = s\} \\ &= \sum_a \pi(s,a)\sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma E^{\pi}\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_{t+1} = s'\}] \\ &= \sum_a \pi(s,a)\sum_{s'} P^a_{ss'}[R^a_{ss'} + \gamma V^{\pi}(s')], \end{aligned} \tag{8}$$

where, $V^{\pi}(s)$ represents the state value function, and $P^a_{ss'}$ represents the probability of state s transitioning to state s' when action a is selected. $R^a_{ss'}$ represents the reward for the transition from state s to next state s' when action a is selected, which is replaced by R_t and expressed by the following Equation (9):

$$R_t = \sum_{k=0}^{\infty}\gamma^k r_{t+k+1}, \tag{9}$$

where, r_{t+1} represents the value return from state s to state s', and γ is the loss factor, value range is (0,1). R_t determines the probability distribution at the end of each round of training.

The RL framework based on the radar system environment proposed in this paper also satisfies the state value function of the Bellman equation: Set $\gamma = 1$ when the initial state s is transferred to the final state s' by taking action a. In other cases $\gamma = 0$. This situation is equivalent to the special equation solution of Bellman equation. As shown in the following Equation (10):

$$V^{\pi}(s) = \begin{cases} \sum_a \pi(s,a)\sum_{s'} P^a_{ss'} \times R^a_{ss'}, & s \to s \\ \sum_a \pi(s,a)\sum_{s'} P^a_{ss'}[R^a_{ss'} + V^{\pi}(s')], & s \to s' \end{cases} \tag{10}$$

According to theoretical analysis, the model framework proposed in this paper is a special equation solution of Bellman equation, which demonstrates the effectiveness of the enhanced model under the radar system environment. In this framework model, the radar system model that satisfies the MG theory is converted into an environment model that satisfies the MDP, which ensures the convergence of the algorithm.

4. Experiments

4.1. Setup

In order to verify the availability of the proposed model in the radar-jammer system environment, the radar side and the jammer side simulation environment models were built on the TensorFlow platform. The four algorithms QL [18], DQN [21], DDQN [24], DQN+LSTM [25] are compared in the IM (Ideal Model, IM) and the proposed AM (Advanced Model, AM). The four RL methods are briefly described as follows:

- QL (Q-learning) [18]: Q-learning (QL) is a classic reinforcement learning algorithm, which can efficiently uses guided experience replay sampling to update the strategy, and selects the action that can get the most benefit based on the Q value.
- DQN (deep Q-network) [21]: Deep Q-learning replaces the regular Q-table with a neural network. Rather than mapping a state-action pair to a q-value, a neural network maps input states to (action, Q-value) pairs. DQN algorithm relies on limited discrete

empirical data storage memory and complete observation information, effectively solving the "dimension disaster" of reinforcement learning.

- DDQN (double deep Q-network) [24]: DDQN is an optimization of the DQN algorithm, using two independent Q-value estimators, each of which is used to update the other. Using these independent estimators, an unbiased Q-value estimation can be performed on actions selected using the opposite estimator. Therefore, the update can be separated from the biased estimate in this way to avoid maximizing bias.
- DQN+LSTM (deep Q-network + long-short term memory) [25]: DQN+LSTM is proposed to solve two problems of DQN: (1) The memory for storing empirical data is limited. (2) Need complete observation information. DQN+LSTM replaces the fully connected layer in DQN with an LSTM network. In the case of changes in observation quality, it has stronger adaptability.

To test the convergence of the algorithm, the experiment is set to 6000 rounds, including 8 effective radar states, of which s_7 is the effective final state. Radar behavior a^{RDR} is regarded as a behavior group, which contains 14 behaviors. Each a_i^{RDR} $(i = 0, 1, ..., 13)$ behavior sets a set of discrete reference values according to the real radar signal. The behavior of the jammer, a^{JAM}, was regarded as the behavior group, and 28 behaviors were selected. A set of discrete reference values was set for each behavior of a_i^{JAM} $(i = 0, 1, ..., 27)$ according to the real jammer signals. The reward values of the radar side and the jammer are shown in Table 5. The reason for setting the reward value of the radar side and the jammer is as follows: because the radar side does not know the information of the jammer at the beginning of the confrontation, the initial value of the reward value is set to 0. In order to effectively improve the ability of radar online anti-jamming learning, high score rewards for effective anti-jamming measures and low scores for general anti-jamming measures are given based on environmental feedback. Since the jammer can interfere with the radar aiming frequency at the initial stage, the initial value is set to 30. However, this interference method cannot evaluate the interference effect based on the radar status to optimize the strategy. Therefore, the jammer is set to score according to the radar score, and the reward range in the positive and negative directions is less than that of the radar. The termination condition for each round is that the two parties first reach 200 points or the number of confrontations reaches 2000. Each experiment runs 100 times repeatedly, and the processing time was the average of 6000 rounds of experimental data.

Table 5. The reward value of the radar side and the jammer.

Target Error Rate	Radar's Rewards	Jammer's Reward
$\rho_d < 5\%$	+5	−4
$5\% \leq \rho_d < 15\%$	+4	−2
$15\% \leq \rho_d < 35\%$	+1	−1
others	−4	+3

4.2. Results and Analyss

4.2.1. Convergence Analysis

The experimental results are evaluated based on the anti-jamming effect of the radar. As shown in Figure 9, it is the experimental results of the radar and jammer countermeasures under IM and AM with four RL algorithms of QL, DQN, DDQN, and DQN+LSTM. In the figure, the horizontal axis represents the number of rounds, and the vertical axis represents the respective confrontation scores of the radar and the jammer. The red curve represents the jammer, and the blue curve represents the radar side.

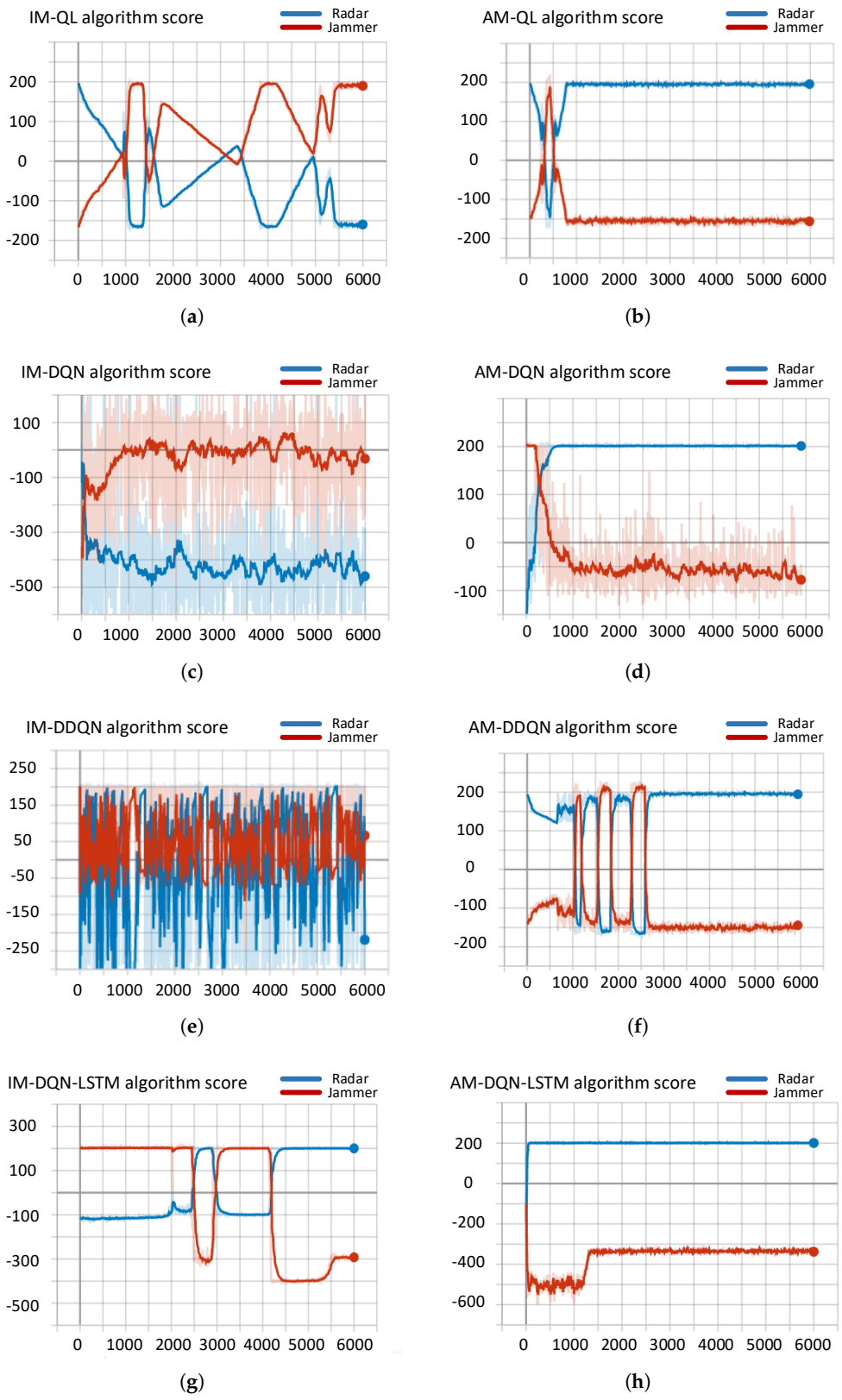


Figure 9. Comparisonof anti-jamming effects of classic RL algorithms under IM and AM models. (**a**) IM-QL, (**b**) AM-QL, (**c**) IM-DQN, (**d**) AM-DQN, (**e**) IM-DDQN, (**f**) AM-DDQN, (**g**) IM-DQN+LSTM, (**h**) AM-DQN+LSTM.

It can be analyzed from the experimental results in Figure 9 that the classic RL algorithms QL and DQN under the IM model cannot achieve effective anti-jamming measures. The result is that the jammer wins the confrontation. In the DDQN and DQN+LSTM algorithms, the radar anti-jamming effect is unstable, and effective anti-jamming measures cannot be implemented. Under the AM framework, the same algorithms enable the frequency agile radar to quickly search the optimal anti-jamming strategy, implement effective interference countermeasures and achieve the ultimate victory. This experiment proves that the AM framework proposed in this paper can effectively solve the problem of the algorithm cannot converge in the radar-jammer system.

4.2.2. Performance Analysis

Figure 10 shows the number of confrontations per round between the radar side and the interferer under the four algorithms of QL, DQN, DDQN, and DQN+LSTM under IM and AM. The horizontal axis represents the number of rounds, and the vertical axis represents the number of confrontations per round. The blue curve represents the IM model, and the red curve represents the AM model.

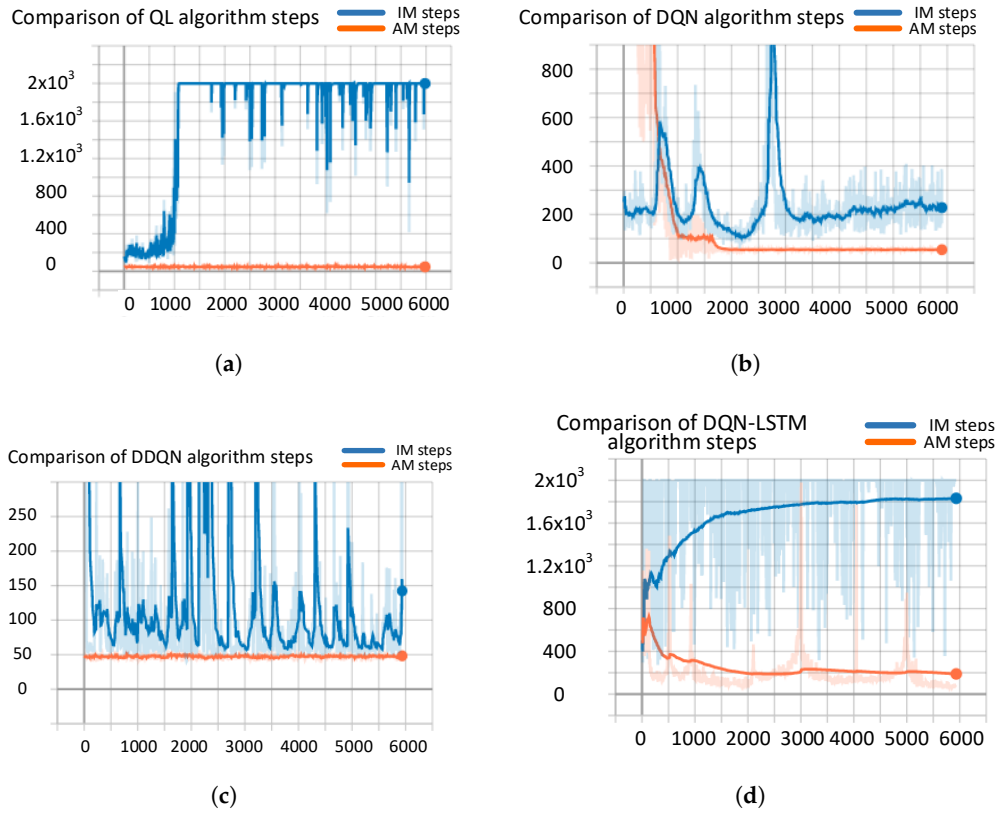


Figure 10. Convergence comparison chart of classical RL algorithms under IM and AM models. (**a**) QL algorithm comparison, (**b**) DQN algorithm comparison, (**c**) DDQN algorithm comparison, (**d**) DQN+LSTM algorithm comparison.

Table 6 is the online processing time and cumulative mean square error (MSE) analysis of the four algorithms QL, DQN, DDQN, DQN+LSTM under IM and AM, respectively.

Table 6. Experimental comparison of processing time of classic RL algorithms under IM and AM.

Mode	Convergence Time (Time/μs)	Single-Round Decision Time After Convergence (Time/μs)	6000 Rounds Total Running Time (Time/μs)	MSE	
				Before Convergence	After Convergence
IM_QL [18]	101.576 ± 0.12%	12.792 ± 0.88%	837,581.728 ± 0.69%	6.502 ± 1.10%	1.901 ± 2.97%
AM_QL	24.971 ± 0.33%	1.554 ± 4.01%	18,122.910 ± 1.78%	7.151 ± 0.84%	1.411 ± 5.72%
IM_DQN [21]	210.774 ± 1.04%	21.552 ± 5.19%	1,083,521.021 ± 0.72%	8.694 ± 0.23%	1.907 ± 1.41%
AM_DQN	117.863 ± 0.42%	2.016 ± 0.82%	159,102.375 ± 0.38%	8.967 ± 2.25%	1.433 ± 2.16%
IM_DDQN [24]	191.861 ± 0.23%	19.613 ± 0.97%	106,707.731 ± 3.61%	8.631 ± 3.14%	2.072 ± 0.55%
AM_DDQN	89.465 ± 0.52%	1.976 ± 2.49%	9153.983 ± 1.08%	8.985 ± 4.52%	1.361 ± 4.62%
IM_DDQN+LSTM [25]	285.373 ± 2.63%	30.184 ± 0.12%	2,550,491.971 ± 0.85%	6.781 ± 1.70%	1.923 ± 5.14%
AM_DQN+LSTM	132.926 ± 7.01%	7.622 ± 4.53%	283,561.026 ± 2.37%	8.012 ± 2.92%	1.876 ± 3.85%

Among them, the convergence time refers to the selection strategy from the beginning to the stable selection strategy in a single round of the algorithm. The single-round decision time after convergence refers to the running time from each turn to the optimal strategy after stable strategy selection. The cumulative mean square error analysis is the cumulative mean square error of the probability of selecting the optimal strategy from the first round. The pre-convergence MSE is the MSE of the probability of selecting the optimal strategy accumulated from the first round to the round before the convergence. After convergence, MSE is the MSE of the probability of selecting the optimal strategy from the start of the 3000th round to the end of the experiment. The cumulative mean square error can effectively analyze the convergence of the strategy and the stability of the algorithm.

According to Figure 10 and Table 6, it can be concluded that the convergence time of the classic reinforcement learning algorithm under AM is 2–5 time faster than the processing rate of the IM framework. The AM framework algorithm can quickly find the optimal anti-jamming countermeasures in the single-round processing time period after convergence, which is 4–10 times faster than the IM framework with the same algorithm processing speed. From the analysis of MSE, QL, DQN, DDQN and DQN+LSTM have greater error oscillations before convergence under AM than under IM, but the oscillations of errors after convergence are all smaller tan IM.

Comprehensive analysis, compared to IM, the AM model proposed in this paper is more suitable for radar-jammer system. It has better convergence and higher processing efficiency.

5. Conclusions

This paper analyzes and studies the modern radar-jammer system, using reinforcement learning (RL) algorithms to improve the frequency-agile radar anti-jamming decision-making effect. The analysis shows that RL algorithms are suitable for the environment of radar-jammer system, but frequency-agile radar system has two characteristics of environmental instability and step changes in decision-making actions, which does not conform the complete Markov decision theory. As a result, the existing classical RL algorithms as QL [18], DQN [21], DDQN [24] and DQN+LSTM [25], which are applied to this scenario, have problems, e.g., low training efficiency and large error oscillation. For the above two problems, this paper proposes an efficient RL anti-jamming model for frequency-agile radar: transforming the radar-jammer detection model into a model that satisfies the MG theory, and is suitable for step change data. Experimental results show that the proposed model can improve the anti-jamming performance and efficiency of frequency-agile radar.

Based on the current work, further research will be carried out. First, the existing model have good experimental results for the finite radar state. When the radar state is infinite, the model needs to be further verified and optimized. Second, it is necessary to further verify the universality of the model and study whether the new RL algorithm is also applicable to the model framework proposed in this article.

Author Contributions: Conceptualization, X.L. and S.D.; data curation, X.L.; methodology, X.L. and S.D.; software, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Xiangyang Cui for their help in this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, L.; Peng, J.; Xie, Z.; Zhang, Y. Optimal jamming frequency selection for cognitive jammer based on reinforcement learning. In Proceedings of the 2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP), Weihai, China, 28–30 September 2019; pp. 39–43.
2. Wang, Y.; Zhang, T.; Xu, L.; Tian, T.; Kong, L.; Yang, X. Model-free reinforcement learning based multi-stage smart noise jamming. In Proceedings of the 2019 IEEE Radar Conference (RadarConf), Boston, MA, USA, 22–26 April 2019; pp. 1–6.
3. Shang, Z.; Liu, T. Present situation and trend of precision guidance technology and its intelligence. In Proceedings of the LIDAR Imaging Detection and Target Recognition 2017, Changchun, China, 23–25 July 2017; p. 1060540.
4. Carotenuto, V.; De Maio, A.; Orlando, D.; Pallotta, L. Adaptive radar detection using two sets of training data. *IEEE Trans. Sign. Process.* **2017**, *66*, 1791–1801. [CrossRef]
5. Young, J.R.; Narayanan, R.M.; Jenkins, D.M. Modified transmitted reference technique for multi-resolution radar timing and synchronization. In Proceedings of the Radar Sensor Technology XXIII, Baltimore, MD, USA, 14–18 April 2019; Volume 11003, p. 110031A.
6. Haykin, S. Cognitive radar: A way of the future. *IEEE Sign. Process. Mag.* **2006**, *23*, 30–40. [CrossRef]
7. Edelman, G.M. *Second Nature: Brain Science and Human Knowledge*; Yale University Press: New Haven, CT, USA, 2006.
8. Russell, S.; Norvig, P. Artificial Intelligence: A Modern Approach, 2002. Available online: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/27702.pdf (accessed on 1 January 2020).
9. Bu, F.; He, J.; Li, H.; Fu, Q. Radar seeker anti-jamming performance prediction and evaluation method based on the improved grey wolf optimizer algorithm and support vector machine. In Proceedings of the 2020 IEEE 3rd International Conference on Electronics Technology (ICET), Chengdu, China, 8–12 May 2020; pp. 704–710.
10. Chu, Z.; Xiao, N.; Liang, J. Jamming effect evaluation method based on radar behavior recognition. *J. Phys. Conf. Ser.* **2020**, *1629*, 012001. [CrossRef]
11. Steinwart, I.; Christmann, A. *Support Vector Machines*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
12. Liu, M.; Qi, H.; Wang, J.; Liu, L. Design of intelligent anti-jamming system based on neural network algorithm. *Comput. Meas. Control* **2018**, *26*, 155–159.
13. Haykin, S. *Neural Networks and Learning Machines, 3/E*; Pearson Education India: Chennai, India, 2010.
14. Deligiannis, A.; Lambotharan, S. A bayesian game theoretic framework for resource allocation in multistatic radar networks. In Proceedings of the 2017 IEEE Radar Conference (RadarConf), Seattle, WA, USA, 8–12 May 2017; pp. 0546–0551.
15. Garnaev, A.; Petropulu, A.; Trappe, W.; Poor, H.V. A power control problem for a dual communication-radar system facing a jamming threat. In Proceedings of the 2020 IEEE Radar Conference (RadarConf20), Florence, Italy, 21–25 Septembe 2020; pp. 1–6.
16. Li, H.; Han, Z. Dogfight in spectrum: Combating primary user emulation attacks in cognitive radio systems, part i: Known channel statistics. *IEEE Trans. Wirel. Commun.* **2010**, *9*, 3566–3577. [CrossRef]
17. Wu, B. Hierarchical macro strategy model for moba game ai. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 1206–1213. [CrossRef]
18. Qiang, X.; Weigang, Z.; Xin, J. Research on method of intelligent radar confrontation based on reinforcement learning. In Proceedings of the 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, China, 8–11 September 2017; pp. 471–475.
19. Watkins, C.J.; Dayan, P. Technical note: Q-learning. *Mach. Learn.* **1992**, *6*, 279–292. [CrossRef]
20. Melo, F.S. *Convergence of q-Learning: A Simple Proof*; Institute of Systems and Robotics: Lisbon, Portugal, 2001.
21. Li, K.; Jiu, B.; Liu, H. Deep q-network based anti-jamming strategy design for frequency agile radar. In Proceedings of the 2019 International Radar Conference (RADAR), Toulon, France, 23–27 September 2019; pp. 1–5.
22. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
23. Aref, M.A.; Jayaweera, S.K. Spectrum-agile cognitive interference avoidance through deep reinforcement learning. In Proceedings of the International Conference on Cognitive Radio Oriented Wireless Networks, Poznan, Poland, 11–12 June 2019; pp. 218–231.
24. Van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double q-learning. *Proc. AAAI Conf. Artif. Intell.* **2016**, *30*, 2094–2100.

25. Ak, S.; Brüggenwirth, S. Avoiding jammers: A reinforcement learning approach. In Proceedings of the 2020 IEEE International Radar Conference (RADAR), Washington, DC, USA, 28–30 April 2020; pp. 321–326.
26. Olah, C. Understanding LSTM Networks. 2015. Available online: https://web.stanford.edu/class/cs379c/archive/2018/class_messages_listing/content/Artificial_Neural_Network_Technology_Tutorials/OlahLSTM-NEURAL-NETWORK-TUTORIAL-15.pdf (accessed on 1 November 2019).
27. Anderson, T. *The Theory and Practice of Online Learning*; Athabasca University Press: Athabasca, AB, Canada, 2008.
28. Laskin, M.; Lee, K.; Stooke, A.; Pinto, L.; Abbeel, P.; Srinivas, A. Reinforcement learning with augmented data. *arXiv* **2020**, arXiv:2004.14990.
29. Wang, Z.; Hong, T. Reinforcement learning for building controls: The opportunities and challenges. *Appl. Energy* **2020**, *269*, 115036. [CrossRef]
30. Kostrikov, I.; Yarats, D.; Fergus, R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv* **2020**, arXiv:2004.13649.
31. Van Otterlo, M.; Wiering, M. Reinforcement learning and markov decision processes. In *Reinforcement Learning*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 3–42.
32. Whitehead, S.D.; Lin, L.-J. Reinforcement learning of non-markov decision processes. *Artif. Intell.* **1995**, *73*, 271–306. [CrossRef]
33. Littman, M.L. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings*; Elsevier: Amsterdam, The Netherlands, 1994; pp. 157–163.
34. Pardhasaradhi, A.; Kumar, R.R. Signal jamming and its modern applications. *Int. J. Sci. Res.* **2013**, *2*, 429–431.
35. Peiqiang, W.; Zhang, J. Analysis and countermeasures of radar radio frequency-screen signal. *Aerosp. Electron. Warf.* **2013**, *29*, 47–50.
36. Jin, L.; Zhan, L.; Xiao, J.; Shen, Y. A syn-aim jamming algorithm against concatenated code. *J. Telem. Track. Command* **2014**, *35*, 37–42
37. Zhang, L.; She, C. Research into the anti-spot-jamming performance of terminal guidance radar based on random frequency hopping. *Shipboard Electron. Countermeas.* **2020**, *43*, 17–19+24.
38. Dayan, P.; Daw, N.D. Decision theory, reinforcement learning, and the brain. *Cognit. Affect. Behav. Neurosci.* **2008**, *8*, 429–453. [CrossRef] [PubMed]
39. Li, K.; Jiu, B.; Liu, H.; Liang, S. Reinforcement learning based anti-jamming frequency hopping strategies design for cognitive radar. In Proceedings of the 2018 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Qingdao, China, 14–16 September 2018; pp. 1–5.
40. Kozy, M.; Yu, J.; Buehrer, R.M.; Martone, A.; Sherbondy, K. Applying deep-q networks to target tracking to improve cognitive radar. In Proceedings of the 2019 IEEE Radar Conference (RadarConf), Boston, MA, USA, 2–26 April 2019; pp. 1–6.