Q-learning and Policy Gradient-Based Reinforcement Learning Method to Decision Making of Phased Array Radar Jamming

Qimeng Tang*, Yuhang Zhang, Yanbin Gao
School of Architecture and Design, Chongqing College of Humanities, Science & Technology, Chongqing 401573, China
E-mail: 13883362339@163.com
*Corresponding author

Keywords: phased array radar, radar jamming decision-making, reinforcement learning**,** signal-level simulation

Received: October 18, 2024

The rapid advancement of phased array radar has greatly improved the guiding and anti-jamming capacities of radar seekers. In contemporary electronic warfare, traditional radar-jamming decision-making methods have proven ineffective, prompting the use of reinforcement learning methods to tackle performance and efficiency shortcomings. This research employs Q-learning and Policy Gradient techniques for radar jamming decision-making, with signal-level simulation rather than functional-level simulation as used in previous research. Signal-level simulation offers a more realistic and understandable representation of the interference mechanism that affects missile terminal guidance. The proposed reinforcement learning methods discover optimal actions for interference equipment, thereby increasing jamming efficiency. Simulation findings show an 18% increase in interference efficiency over conventional techniques, with models attaining 92% accuracy in optimal decision-making. The precision of the signal-level simulation model and the effectiveness of reinforcement learning in improving interference performance are confirmed.

Povzetek: Razvita metoda na signalni ravni izboljša odločanje pri radarskem motenju in poveča učinkovitost za približno 18 %.

1 Introduction

In the increasingly complex electronic warfare environment characterized by rapid technological advancements and sophisticated adversarial tactics, traditional radar systems that rely on a single working mechanism and simplistic beam variations are becoming inadequate [1], [2]. These conventional radars struggle to meet the diverse operational requirements of modern warfare scenarios, which demand high levels of adaptability and precision in detection capabilities. Phased array radar represents a significant evolution in radar technology. This type of radar system is designed to dynamically alter its real-time beam direction by adjusting the emitted signals' phase from multiple antennas arranged in an array configuration [3]. The ability to electronically steer the radar beam allows for faster target acquisition and tracking compared to mechanical scanning methods. Consequently, phased array radars are often referred to as electronic scanning array (ESA) radars due to their advanced electronic control mechanisms. Currently, phased array radar has undergone substantial development, evolving into multifunctional systems capable of engaging with various targets across extensive ranges while maintaining high reliability under different operational conditions. These features make them indispensable components within advanced weapon systems utilized by military forces worldwide. Their versatility enables simultaneous engagement with multiple threats—ranging from aircraft and missiles to ground-based assets—thereby enhancing situational awareness on the battlefield. Traditional methods employed for making interference decisions regarding these radars include techniques based on template matching, game theory approaches that analyze strategic interactions between competing entities, and reasoning-based methodologies that utilize logical frameworks [4]. However, all these conventional strategies necessitate a considerable amount of prior data for effective decision-making support. They often fall short when faced with multifunctional radars operating within unknown or unpredictable environments where historical data may be limited or irrelevant. In contrast, interference decision-making methods grounded in reinforcement learning offer distinct advantages over traditional approaches. Reinforcement learning algorithms possess inherent cognitive abilities that allow them not only to learn from past experiences but also adaptively explore optimal interference strategies through continuous trial-and-error processes. By leveraging feedback from previous actions taken during operations against dynamic adversarial tactics, these algorithms can refine their

decision-making capabilities over time without requiring exhaustive datasets upfront.

Q. Xing applied reinforcement learning to intelligent radar confrontation, and the experimental results indicated that reinforcement learning possesses excellent adaptive capabilities in one-on-one confrontations [5]. B. K. Zhang constructed a cognitive interference decision model from the perspective of signal processing, updated the types of interference libraries, analyzed the assessment of interference effects, and played a significant role in the cognitive confrontation of multifunctional radars [6]. S. Y. Zhang combined game theory with reinforcement learning and proposed a multi-agent reinforcement learning algorithm based on equilibrium, which could effectively address the issue of radar observation time scheduling and further demonstrate the effectiveness of reinforcement learning in handling interference decision problems [7]. F. Slimeni proposed an interferer scheme featuring spectrum sensing, offline training, and learning functions. The RL-based interference decision algorithm was simulated on a universal software radio platform. The results indicated that the interferer can effectively learn and interfere without prior information from users [8]. L. Yunjie and X. Qiang introduced prior knowledge based on reinforcement learning and improved the performance of reinforcement learning in radar interference decision-making [9]. X.Qiang, Z. Wei-gang, and J. Xin proposed an adaptive reinforcement learning algorithm that can solve the interference problem of radars with unknown working modes [10]. W. G. Zhu, confronted with the increasing working modes and states of radars, adopted the deep reinforcement learning approach and effectively addressed the issue of low decision-making efficiency of reinforcement learning [11]. The summary table showed in Table 1.

Table 1: Summary. The table summarizes the important system parameters and experimental conditions utilized in the study. This table provides a quick reference for understanding the experimental setup

Study	**Methodology**	**Datasets/Environments**	**Key Results**
Q. Xing [5]	Reinforcement learning for intelligent radar confrontation	The experimental setting for one-on-one confrontations	Showed adaptive skills in radar confrontation; good efficiency in adaptive decision-making
B. K. Zhang [6]	Cognitive interference selection model with signal processing	Signal processing setting, interference libraries	Revised interference categories and assessed influence; improved cognitive engagement of multifunctional radars
S. Y. Zhang [7]	Game theory combined with reinforcement learning	Radar scheduling and interference management simulation	Efficient in scheduling radar observation times; resolved interference decision issues by multi-agent learning
F. Slimeni [8]	Reinforcement learning-based interference decision technique	The universal software radio platform	Effectively simulated spectrum finding and learning; interferers learn and disrupt without previous knowledge
L. Yunjie and X. Qiang [9]	Reinforcement learning with previous data	Radar interference decision-making system	Improved effectiveness in radar interference decision-making by integrating past data
X. Qiang, Z. Wei-gang, and J. Xin [10]	Adaptive reinforcement learning system	Radar interference with unknown modes	Effectively tackled interference problems in radars functioning under unknown modes utilizing adaptive learning.
W. G. Zhu [11]	Deep reinforcement learning tactic	Simulation of radar functional modes	Resolved insufficient decision-making efficacy in changing radar operational modes

Prior studies on radar interference decision-making concentrated on functional-level simulations, which lacked detailed signal-level modeling and failed to simulate changing interference settings. Furthermore, the incorporation of reinforcement learning with signal-level data, particularly in intricate situations such as anti-ship missile countermeasures, is underexplored. The suggested strategy fills these gaps by employing a coherent video signal analysis technique for signal-level simulation, including detailed modeling of interference equipment, and incorporating reinforcement learning. This new methodology enhances the realism and efficacy of radar interference decision-making, distinguishing it from previous functional-level methods.

Most of the previous studies on radar interference decision-making methods have employed functional-level simulation approaches, merely simulated the amplitude information of signal transmission, targets, echoes, clutters, and interference signals or presented a straightforward description of the radar's working status without outputting genuine radio frequency and video signals. It has the typical merits of simplicity, practicality, and strong real-time processing ability. However, it neglects numerous details inherent in waveforms and signal processing, fails to accurately depict the dynamically changing interference confrontation environment created by various electronic interference equipment of both the friendly and hostile sides, and is even less capable of comprehensively and quantitatively elaborating the mechanism of action of various interference factors on the radar system. This paper adopts the coherent video signal analysis method, with the background of the interference equipment on ships countering anti-ship missile attacks, and simulates the entire process of radar interference decision-making from the signal-level perspective, providing a novel idea for subsequent studies on radar interference decision-making.

This study seeks to address the drawbacks of traditional radar-jamming decision-making techniques by employing reinforcement learning methods, particularly Q-learning and Policy Gradient algorithms. The main goal is to create a decision-making framework capable of adapting to the dynamic and unpredictable adversarial situations encountered by phased array radars. The research aims to improve interference efficiency and optimize jamming tactics through continuous learning and adaptation. This study aims to show the benefits of reinforcement learning in enhancing decision-making capacities compared to conventional methods, eventually improving the efficacy of radar jamming in contemporary electronic warfare.

2 Modelling of phased array radar se eker

2.1 Signal transmission and signal reception model

2.1.1 Antenna model

The general antenna radiation pattern function is shown below:

$$F(\alpha,\beta)=\left|\frac{\sin\left(\pi\frac{\alpha}{\theta_\alpha}\right)}{\pi\frac{\alpha}{\theta_\alpha}}\right|\cdot\left|\frac{\sin\left(\pi\frac{\beta}{\theta_\beta}\right)}{\pi\frac{\beta}{\theta_\beta}}\right| \tag{1}$$

Where α and β represent the azimuth and elevation offsets between the target and the antenna center, expressed in radians; θ_α and θ_β represent the azimuth and elevation beam widths of the antenna, also expressed in radians. The beam width parameter can be estimated directly from the antenna size as follows:

$$\theta_{\alpha,\beta}=\frac{\lambda}{L_{\alpha,\beta}} \tag{2}$$

where L_α and L_β are the lengths of the antenna in the azimuth and elevation dimensions, and λ is the wavelength of the transmitted signal. In phased array antenna mode, the beamwidth will widen if the antenna spacing cannot meet the requirement of $\Delta d \le \lambda/2$. Meanwhile, the antenna gain can also be estimated simply by the antenna size as follows:

$$G=\frac{4\pi L_\alpha L_\beta}{\lambda^2} \tag{3}$$

In the three-dimensional orientation, the amplification power varies in different directions. The angular information of the antenna pointing to the target can be solved by the sum-and-difference beam method. The directional patterns of the four sub-beams are respectively recorded as:

$$O_1: F_1(\alpha,\beta)=G(-\Delta_1,\Delta_2,\alpha,\beta,\alpha_0,\beta_0) \tag{4}$$

$$O_2: F_2(\alpha,\beta)=G(\Delta_1,\Delta_2,\alpha,\beta,\alpha_0,\beta_0) \tag{5}$$

$$O_3: F_3(\alpha,\beta)=G(-\Delta_1,-\Delta_2,\alpha,\beta,\alpha_0,\beta_0) \tag{6}$$

$$O_4: F_4(\alpha,\beta)=G(\Delta_1,-\Delta_2,\alpha,\beta,\alpha_0,\beta_0) \tag{7}$$

where α_0 and β_0 are respectively the center pointing of the phased array antenna (coordinates in the coordinate system of the array surface azimuth), α and β are the coordinates in the coordinate system of the pointing azimuth and Δ_1 and Δ_2 are also the coordinates in the coordinate system of the pointing azimuth. Then, the sum beam pattern is as follows:

$$F_{\Delta 1}(\alpha,\beta)=F_1(\alpha,\beta)+F_2(\alpha,\beta)+F_3(\alpha,\beta)+F_4(\alpha,\beta) \tag{8}$$

The azimuthal difference beam direction pattern is:

$$F_{\Delta 1}(\alpha,\beta) = F_1(\alpha,\beta) + F_2(\alpha,\beta) - F_3(\alpha,\beta) - F_4(\alpha,\beta) \quad (9)$$

The elevation-azimuth direction beam pattern is:

$$F_{\Delta 2}(\alpha,\beta) = F_1(\alpha,\beta) - F_2(\alpha,\beta) + F_3(\alpha,\beta) - F_4(\alpha,\beta) \quad (10)$$

Furthermore, the proportional coefficient needed for angle error signal resolution in the angle measuring module can be obtained as follows:

$$\mu_1 = \frac{F_\Sigma(0,0)}{\frac{dF_{\Delta 1}}{d\alpha}\Big|_{\alpha=0,\beta=0}}, \mu_2 = \frac{F_\Sigma(0,0)}{\frac{dF_{\Delta 2}}{d\beta}\Big|_{\alpha=0,\beta=0}} \quad (11)$$

2.1.2 Signal transmission modelling

To accurately simulate the strike process of the terminal guidance of a radar seeker in complex electromagnetic environments, signal-level simulation of the radar seeker is necessary. Coherent video signal simulation technology is a type of signal-level simulation. This technology simulates the entire process of signal reception and processing through video signals [12]. The time-domain and frequency-domain characteristics of radar signals can typically be expressed using analytical mathematical expressions, as shown in equation 1, thereby generating the signal's I and Q channel signals.

$$S(t) = A(t)\cdot e^{j(\omega_0 t \phi\phi(t))} = [A_c(t) + jA_s(t)]\cdot e^{j\omega_0 t} = u(t)e^{j2\pi f_0 t} \quad (12)$$

For the envelope $a(t)$ of the real signal $x(t)$, the envelope of its exponential signal $S(t)$ can be depicted as:

$$u(t) = a(t)\cdot e^{j\theta(t)} \quad (13)$$

Therefore, the signal $x(t)$can be written as:

$$x(t) = a(t)\cos[2\pi f_0 t + \theta(t)] = \frac{1}{2}[S_e(t) + S_e^*(t)] = \mathrm{Re}[S_e(t)] \quad (14)$$

Meanwhile, the signal $x(t)$ can also be written as:

$$x(t) = a_l(t)\cos(2\pi f_0 t) - a_Q(t)\sin(2\pi f_0 t) \quad (15)$$

$$a_I(t) = a(t)\cos\theta(t) \quad (16)$$

$$a_Q(t) = a(t)\sin\theta(t) \quad (17)$$

In the above formula, a_I and a_Q are the in-phase and quadrature components of the baseband signal, respectively. This is the signal sampling sequence that we need to generate. Correspondingly, when the carrier frequency is eliminated through downconversion (quadrature demodulation) processing to obtain the zero intermediate frequency (coherent video) signal as:

$$S_0(t) = A(t)\cdot e^{j\phi(t)} = A_c(t) + jA_s(t) \quad (18)$$

When using full digital simulation, the coherent video signal is sampled at a certain sampling rate, thus forming a sequence of coherent video signals (complex signals) that can be represented as follows:

$$s[n] = s_r[n] + js_i[n] \quad (19)$$

Typical phased array radar transmits signals using an LFM (linear frequency modulation) signal pattern. The radar-transmitted signal is represented by a complex signal as follows:

$$S_t(t) = \sqrt{\frac{2P_t}{4\pi L_t}}g_{vt}(\theta)v(t)\cdot e^{j\omega_{ck}t}\omega_{ck} \quad (20)$$

where ω_{ck} denotes the current pulse carrier frequency, P_t indicates the peak power of the transmitter, L_t stands for the combined loss of the transmitter, $g_{vt}(\theta)$ represents the transmitting antenna pattern (voltage gain), and $v(t)$ represents the complex modulation function as follows:

$$v(t) = \mathrm{Rect}\left(\frac{t}{T_p}\right)\cdot\exp(j\pi F_m t^2) \quad (21)$$

The rectangular function is defined as:

$$\mathrm{Rect}(t) = \begin{cases}1, & t\in(0,1)\\ 0, & \text{else}\end{cases} \quad (22)$$

where T_p represents the pulse width, and F_m represents the frequency modulation slope. Based on the above formulas, the coherent video signal pattern adopted in the system is as follows:

$$s_t(t) = \sqrt{\frac{P_t}{4\pi L_t}}g_{vt}(\theta)\cdot\mathrm{Rect}\left(\frac{t}{T_p}\right)\cdot \exp\left(j\pi\frac{BW_{rg}}{T_p}t^2\right) \quad (23)$$

2.1.3 Receiving signal model

The received signal mainly consists of target echoes, interference signals, various types of clutter, and receiver noise. This paper mainly considers the aspects of target echoes and interference signals. Concerning a specific transmitted pulse, the RF signal received by the radar can be represented as follows:

$$r_{RF}(t) = S_{RF}(t) + J_{RF}(t) + n_{RF}(t) \quad (24)$$

where $S_{RF}(t)$ represents the echo signal received after the transmitted pulse is reflected by the target, $J_{RF}(t)$ represents the received interference signal, which is the combined interference signal formed by various active interference and passive interference, and $n_{RF}(t)$ represents the receiver noise. The receiver noise is Gaussian-limited white noise, which means that in the radar receiver's passband, the power spectral density of the noise is uniform, and its amplitude follows a Rayleigh distribution. The variance of the noise can be calculated from the receiver noise coefficient and the receiver bandwidth. In the simulation, a sample function of a Gaussian process can be used to represent it specifically, and the band-limited noise signal is represented as:

$$n(t) = \mathrm{Re}[\tilde{n}(t)\cdot e^{j\omega_c t}] \quad (25)$$

Therefore, in coherent video simulation, the noise at the receiver can be represented as:

$$\tilde{n}(t) = n_d(t) - jn_q(t) \quad (26)$$

In this model, $n_d(t)$ and$n_q(t)$ are independent Gaussian random processes with zero mean and variance σ_N^2. The variance of the noise σ_N^2 can be calculated from the receiver noise coefficient N_F and receiver bandwidth Δf as follows:

$$\sigma_N^2 = kT_0N_F\Delta f \quad (27)$$

Where K is the Boltzmann constant, and T_0 is the reference temperature of the receiver, which $T_0 = 290$ K. Combining the target echo signal and the receiver thermal noise signal, we finally obtain the radar received signal as:

■(■(&$r(t) = \mathrm{Rect}((t-2R/c)/T\_p) \cdot [\sqrt{((2P\_t)/((4\pi)\text{^}3\ L))}\ (g\_vt\ (\theta)\ g\_vr\ (\theta))/R\text{^}2\ \lambda\_k\ \sqrt{\sigma}]@\& \cdot \exp[j\pi\ (BW\_rg)/T\_p\ (t-2R/c)\text{^}2\ ] \cdot \exp[2\pi f\_d\ t - 2\pi\ 2R/\lambda\_k\ ] + \tilde{n}(t)\ )\#$ (28))

2.2 Radar signal processing modelling

2.2.1 Pulse compression

The pulse width of the transmitted signal influences the range resolution of radar. The narrower the pulse width of the signal, the higher the range resolution of the radar. However, when the pulse width of the signal becomes smaller, the maximum operating range of the radar will also be reduced, which leads to the fact that the pulse width of the signal cannot be too small. Therefore, the commonly adopted method for current radar is to employ a high-power transmitter to emit a wide pulse signal and then perform compression on the original wide pulse signal after receiving the echo signal, converting the signal into a narrow pulse signal, thereby achieving the functions of enhancing both the maximum operating range and the range resolution. This process is known as pulse compression [13]. In the commonly used radars at the present stage, a matched filter is typically employed to carry out pulse compression processing of signals. Suppose the transfer function of the matched filter is $H(f)$, the impulse response is $h(f)$, and the input signal is $s(t)$. Block Diagram Representation of a System with Transfer Function H(ω)H(\omega)H(ω) is depicted in Figure 1.

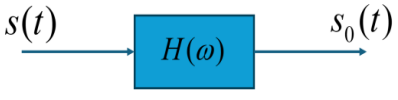

Figure 1: Block diagram representation of a system with transfer function H(ω)H(\omega) H(ω)

$s_0(t)$ is the output result of the target signal after it has passed through a matched filter, which can be expressed as follows:

$$s_0(t) = \int_{-\infty}^{\infty} H(f)S(f)e^{j2\pi ft}df \quad (29)$$

Taking the inverse Fourier transform of the transfer function $H(f)$ results in its impulse response function as

$$h(t) = ks^*(t_0 - t) \quad (30)$$

For the sampled digital signal, denoting the input signal after sampling and quantization as $s(n)$, the unit impulse response of the matched filter can be represented as $h(n) = s^*(n)$.

2.2.2 Demodulation model

Consider the squaring law and envelope detection methods. The squaring law demodulation process is as follows:

$$x(n) = [A_{re}^2(n) + A_{im}^2(n)], n = 0, \cdots, N-1 \quad (31)$$

The process of envelope detection is as follows:

$$x(n) = \sqrt{[A_{re}^2(n) + A_{im}^2(n)]}, n = 0, \cdots, N-1 \quad (32)$$

2.2.3 Constant false alarm handling and detection model

Constant False Alarm Handling and Detection Model Since the signals received by the radar comprise both target echo signals and clutter signals, not all processing steps in the signal processor can completely filter out the clutter signals. Thus, a threshold value is often set in the radar. The portions of the signal that are higher than the threshold value is retained, while those lower than the threshold value are filtered out. This threshold value is the false alarm probability. To achieve this aim, the false alarm threshold must be calculated in real-time based on the received signal to adjust the radar detection threshold accordingly to obtain the desired false alarm probability [14]. The detection processor that can maintain a constant false alarm probability is referred to as the Constant False Alarm Rate (CFAR) processor.

2.3 Output model

2.3.1 Distance output

The target echo signal will have a time delay t_r Due to the distance between the target and the radar, which can be expressed as $t(r) = 2R/c$. Among them, R represents the relative distance between the target and the radar, and c represents the speed of light. Thus, in the case of a known time delay, the distance between the target and the radar can be inversely deduced based on the time delay t_r of the echo signal.

2.3.2 Angle output

It has been mentioned above that since the radar adopts the sum and difference beam angle measurement method when measuring the angle, the pitch angle θ and the yaw

angle ϕ of the target relative to the radar can be obtained based on the amplitude of the corresponding position of the target in the processed sum beam signal and the amplitudes of the corresponding positions of the target in the pitch difference beam signal and the yaw difference beam signal:

$$\theta = \frac{\Delta F(\delta)}{\Sigma_\theta \frac{dF(\theta)}{d\theta}\Big|_{\theta=\delta}}, \quad \phi = \frac{\Delta F(\delta)}{\Sigma_\phi \frac{dF(\phi)}{d\theta}\Big|_{\phi=\delta}} \tag{33}$$

2.3.3 Velocity output

If relative motion exists between the target and the radar, the frequency of the target's echo will change. The Doppler frequency shift f_d caused by the relative velocity v_r can be expressed as $f_d = 2v_r/\lambda$. In the case where the Doppler frequency shift of the echo signal is known, the relative velocity between the target and the radar can be inversely deduced.

2.4 Guidance method

The proportional guidance method refers to the fact that the angular velocity of the missile's speed-changing direction in space is proportional to the angular velocity of the target's relative radar position angle rotation [15]:

$$\Delta\theta_M = k' \cdot \Delta\theta_T \tag{34}$$

where $\Delta\theta_M$ denotes the angular velocity of the change in the direction of the missile's velocity in space, k' is the proportional coefficient, and $\Delta\theta_T$ represents the angular velocity of the rotation of the target's relative position angle to the radar, also known as the line-of-sight angular velocity. Since the proportional guidance method alters the direction of the missile's velocity change based on the variation of the target's angle, it avoids situations where the missile makes large turns during the guidance process and is convenient for implementation in practical applications. As depicted in Figure 2, assume that at time k, the target is positioned at point T, the missile is at point M, the velocity of the target is v_T, the velocity of the missile is v_M, and the distance between the target and the radar is r.

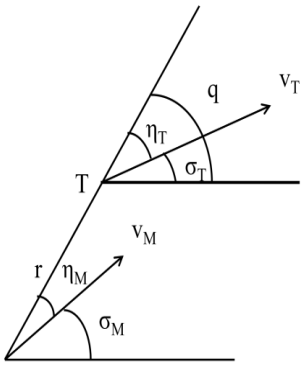

Figure 2: Example of Target and Missile Positions. The positions of the target and missile at time k are depicted, along with relevant parameters such as target velocity (v_T), missile velocity (v_M), and target-radar distance (r).

A diagram of the proportional guidance method.

Denote the angle between the target's velocity and the line connecting the target and the missile as η_T, the angle between the missile's velocity and the line connecting the target and the missile as η_M, and q as the line-of-sight angle at the time k. At time $k+1$, suppose the angle between the target velocity and the baseline is η'_M, the angle between the missile velocity and the baseline is η'_M, and q' is the line-of-sight angle at the time $k+1$. Then the line-of-sight angular velocity at time $k+1$ is:

$$\Delta\theta_T = q' - q \tag{35}$$

It is possible to further obtain the angular velocity of the rotation of the missile's velocity vector at time $k+1$:

$$\Delta\theta_M = k \cdot (q' - q) \tag{36}$$

It can be known from the above formula that the angle between the velocity of the missile's motion and the reference line at time $k+1$ is

$$\eta'_M = \eta_M + \Delta\theta_M \tag{37}$$

2.4 Jamming pattern modelling

2.4.1 Amplitude modulation interference

Noise amplitude modulation interference refers to the signal produced by modulating the amplitude of the carrier signal with a noise signal [16]. Generally, Gaussian white noise is often used to modulate the signal's amplitude. The definition of the noise amplitude modulation interference signal is as follows:

$$J(t) = \big(U_0 + U_n(t)\big)\cos\big(\omega_j t + \phi\big) \tag{38}$$

where $U_n(t)$ is a generalized stationary random process that follows $(0, \sigma^2)$ distribution, U_0 is the amplitude of the signal basis, ω_j is the carrier signal frequency, and ϕ is the signal phase.

2.4.2 Frequency-modulated noise interference

Noise frequency modulation interference is currently the most widely used type of suppressive interference signal, which has a broader interference bandwidth and makes it easy to achieve a more considerable noise power. The mathematical model for noise frequency modulation interference is represented as follows:

$$J(t) = U_j \cos\left(\omega_j t + 2\pi K_{FM} \int_0^t u\,(t')dt' + \phi\right) \tag{39}$$

where U_j is the amplitude coefficient of the signal, ω_j is the carrier frequency, K_{FM} is the frequency modulation slope, and modulation noise $u(t)$ is a wide-sense stationary random process with zero mean. ϕ is the signal phase.

2.4.3 Agile noise jamming

Agile noise interference is a compromise technique that combines deception and noise interference. Its essence is to combine forwarding-style interference with random

pulse interference. The agile noise interference expression based on frequency modulation noise is as follows:

$$J(t) = U_0\cos\left[2\pi f_j t + 2\pi K_{FM}\int u(t')dt' + \varphi\right] \cdot e^{-j\pi K^2}\mathrm{rect}\left(\frac{t}{T}\right) \quad (40)$$

By changing the frequency modulation slope K_{FM}, the bandwidth of the FM noise can be changed, and therefore, the entire bandwidth of the agile noise is also changed.

3 Reinforcement learning

Inspired by the laws of biological learning, reinforcement learning interacts with the environment through a trial-and-error mechanism and learns and optimizes by maximizing cumulative rewards, ultimately achieving the optimal strategy.

The incorporation of RL for radar jamming decision-making via Q-learning and policy gradient techniques is well articulated. However, more information is needed on the choice of hyperparameters for these algorithms, like learning rates, discount factors, and exploration-exploitation trade-offs, which are critical to the performance and stability of the RL models. Furthermore, a more detailed rationale for selecting between Q-learning and policy gradient techniques should be offered, with a focus on how each method fits into the particular radar jamming scenario. A comparative analysis would be useful to emphasize the benefits and drawbacks of these two methods, containing a discussion of computational cost, convergence rate, and operational efficacy in practical radar jamming applications. This would offer more information about the suitability of each approach under different conditions, as well as improve the methodology's overall resilience.

In reinforcement learning, the decision-maker or learner is defined as the "learning agent," and everything outside the learning agent is defined as the "environment," with the system integrating with the environment [17]. The interaction process between the learning agent and the environment can be described by three elements: state s, action a, and reward r. The learning agent acts a_0 based on the initial state s_0 and interacts with the environment, obtaining a reward r_1 and updated state s_1. At time t, based on the current state s_t and rewardr_t, the learning agent provides the current action a_t, and then the system state transitions from s_t to s_{t+1}. r_{t+1} is the feedback reward from interacting with the environment. The basic principle of reinforcement learning is shown in Figure 3.

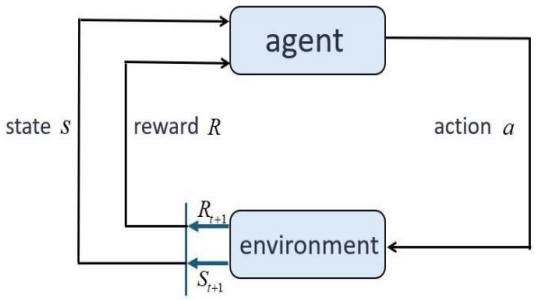

Figure 3: The fundamental structure of the reinforcement learning system used in the jamming decision process, including important elements and their interactions.

Generally speaking, reinforcement learning emphasizes the interaction between the agent and the environment, expressed as a series of sequences of states, actions, and rewards: $s_0, a_0, r_1, s_1, a_1, \cdots, s_{n-1}, a_{n-1}, r_{n-1}$. Although n can tend to infinity, a terminal state $s_n = s_T$ is usually defined in practice for limitation. This sequence of states, actions, and rewards, starting from the initial state and ending at the terminal state, is called an Episode or training cycle. The policy is usually denoted as π, which is a mapping from state s to action a. At the current time step, the learning agent interacts with the environment, learns through trial and error, and iteratively optimizes the current policy π to make the new policy π_1 superior to the current policy π. This process is called "policy update" and is repeatedly executed during reinforcement learning until the learning agent cannot find a better policy. In the interaction with the environment, the learning agent receives a feedback reward r each time t until the terminal state s_T. However, the reward at each step does not represent the long-term reward gain. To express the long-term gain of the learning agent, the return at time step t is introduced as follows:

$$G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots + \gamma^{T-t} r_T \quad (41)$$

Among them, γ is the discount factor that satisfies $0 < \gamma < 1$. When γ is close to 1, the learning machine tends to place more emphasis on long-term returns, while when it is close to 0, it tends to put more emphasis on short-term returns. A value function usually represents the merit of a strategy. The state value function used to evaluate the merit of a strategy in state s is expressed as:

$$V_\pi(s) = \mathrm{E}[G_t \mid s_t = s, \pi] \quad (42)$$

According to this formula, the optimal strategy can be obtained:

$$\pi^* = \underset{\pi}{\mathrm{argmax}} V_\pi(s) \quad (43)$$

Another type of value function is used to evaluate the degree of excellence of taking an action a in state s, called the state value function, also known as the Q function:

$$Q_\pi(s,a) = \mathrm{E}[G_t \mid s_t = s, a_t = a, \pi] \quad (44)$$

The optimal strategy at this point is represented as:

$$\pi^* = \arg\max_a Q_{\pi^*}(s,a) \quad (45)$$

3.1 Q-learning

Q-learning, also called Action-dependent Heuristic Dynamic Programming (ADHDP) [18], does not require a model. Q-learning does not wait until the end of an episode for an update but instead updates using the Temporal Difference (TD) approach at each step, achieving a faster convergence effect. The TD learning utilizes the previous estimate to update the current state value function. The TD learning method aims to obtain the value function. When confronted with control decision problems, the state-action value function is more instructive for selecting actions. Q-learning employs the Bellman optimality principle to make the current value function directly approach the value function of the optimal policy. The update method is as follows:

$$Q(s_t,a_t) \leftarrow \alpha Q(s_t,a_t) + (1-\alpha)\left(r_{t+1} + \gamma \max_a Q(s_{t+1},a')\right) \quad (46)$$

Where α represents the update rate, which satisfies the condition $0 < \alpha < 1$, where a' is the action that maximizes the Q function in a state S_{t+1}.

3.2 Policy gradient methods

Q-learning is a value-function-based approach. In practical applications, a table is frequently employed to store the state or state-action value function, thus being relatively inefficient for complex problems with ample action space. Policy gradient methods do not rely on the value function. They directly parameterize the policy π as $\pi(s \mid \theta)$ and then calculate the gradient concerning the policy performance metric. The policy parameters are adjusted based on the gradient direction to obtain the optimal policy [19]. Parameterized policies can be classified into stochastic policy $\pi(s \mid \theta) = P[a \mid s, \theta]$ and deterministic policy $a = \mu(s \mid \theta)$. A policy objective function $J(\theta)$ is set to evaluate the parameterized policies. For stochastic policies, the action a in the current state s obeys a particular probability distribution with parameter θ. For deterministic policies, the action corresponding to each state is definite. According to the policy gradient theorem, the gradient of stochastic policies is expressed as [20]:

$$\nabla_\theta J(\theta) = \mathrm{E}_{s,a\sim\pi}[\nabla_\theta \ln \pi(s \mid \theta) Q_\pi(s,a)] \quad (47)$$

The deterministic policy gradient is represented as [1], [21]:

$$\nabla_\theta J(\theta) = \mathrm{E}_{s,a\sim\mu}\left[\nabla_\theta \mu(s \mid \theta) \nabla_a Q_\mu(s,a)\big|_{a=\mu(s|\theta)}\right] \quad (48)$$

When computing the gradient, a genuine state-action value function $Q_\pi(s,a)$ or $Q_\mu(s,a)$ is required. Nevertheless, this function is still being determined in practice. One method is to employ the return values over certain steps to estimate the state-action value function. Another approach is to utilize the actor-critic architecture, where the critic approximates the state-action value function, and the actor represents the policy. The critic is represented as a function $Q(s,a \mid \omega)$ of parameter ω and is updated using the temporal difference method. The temporal difference error δ_t is expressed as:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1} \mid \omega) - Q(s_t, a_t \mid \omega) \quad (49)$$

The update formula for parameter ω in the evaluator is:

$$\omega \leftarrow \omega + \alpha \delta_t \nabla_w Q(s_t, a_t \mid \omega) \quad (50)$$

Replace the learned evaluation function $Q(s_t, a_t \mid \omega)$ with the actual value function $Q_\pi(s,a)$ or $Q_\mu(s,a)$ in the policy gradient formula to update the policy.

Algorithmic parameters

In reinforcement learning, algorithmic parameters are crucial for effective learning and optimal performance. Learning rates (α) control the step size during Q-value updates, balancing stability and convergence speed. They usually range between 0.01 and 0.5. The discount factor (γ), typically set between 0.9 and 0.99, prioritizes future rewards over immediate rewards, with higher values indicating long-term gains. Reward function coefficients are created depending on the issue, offering positive rewards for desirable actions and penalties for suboptimal actions, guiding the agent toward its goals. Exploration tactics, like ϵ-greedy, allow agents to balance exploration and exploitation by randomly choosing actions with a probability ϵ, which typically decreases from 1.0 to 0.01 over time. This encourages exploration initially and prioritizes exploitation later. In situations with continuous action spaces, policy gradients can use fixed step sizes or adaptive algorithms such as Adam to update parameters efficiently.

Pseudocode 1: Reinforcement learning algorithm (Q-Learning)

```
# Initialize Q-table
Initialize Q(s, a) arbitrarily for all states s and actions a
Set learning rate α (e.g., 0.1), discount factor γ (e.g., 0.95), and exploration rate ε (e.g., 1.0)

# For exploration tactic
Set ε_decay = 0.995 and ε_min = 0.01

# Training Loop
for episode in range(max_episodes):
   # Initialize starting state
   state = initialize_environment()

   for t in range(max_steps):
      # Select action utilizing ε-greedy policy
      if random.uniform(0, 1) < ε:
         action = select_random_action()
      Else:
         action = argmax(Q[state, :])  # Exploit
```

```
    # Take action, observe the next state, and reward
    next_state, reward = environment_step(state, action)

    # Update Q-value utilizing the Bellman equation
    Q[state, action] = Q[state, action] + α * (reward + γ * max(Q[next_state, :]) - Q[state, action])

    # Transition to the next state
    state = next_state

    # End the episode if the terminal state is gotten
    if is_terminal_state(state):
        break

  # Decay exploration rate
  ε = max(ε * ε_decay, ε_min)

# Final Q-table prepared for optimum action selection
```

The pseudocode outlines the Q-learning algorithm, which is a model-free reinforcement learning method. Initially, the Q-table contains arbitrary values that represent the expected cumulative reward for each state-action pair. During each episode, the agent starts in an initial state and selects actions using a ϵ-greedy exploration strategy—either randomly with probability ϵ or exploiting current knowledge by choosing the action with the maximum Q-value. After completing an action, the agent observes a reward and the next state, updating the Q-value of the present state-action pair with the Bellman equation, which takes into account immediate advantages as well as discounted future rewards. The algorithm repeatedly updates the Q-table and decays the exploration rate ϵ, allowing the agent to shift from exploration to exploitation during training. The procedure is repeated until the agent arrives at an optimum policy, as described by the Q-table.

4 Construction of adversarial models

4.1 Environmental description

In the model, the environment is precisely the target radar seeker. The seeker's state is constituted by two variables: velocity and position. The simulation system generates corresponding emission signals based on the configuration of simulation parameters. The antenna gain amplifies and transmits the emission signals until they reach the target position. Based on the distance between the target and the radar and the scattering cross-sectional area of the target, the emission signals are processed correspondingly to generate the echo signal. Subsequently, the echo signal is sent to the radar receiver, while the receiver also receives clutter signals from the environment and external interference signals. The receiver transmits all the received signals to the signal processor, which processes the received signals and extracts the contained target information. The proportional guidance law then adjusts the velocity, and the corresponding position and velocity will also change. This paper considers that when the altitude of the seeker is zero, the missile detonates, and this condition serves as the termination condition for one simulation. The workflow diagram of the seeker is presented in Figure 4.

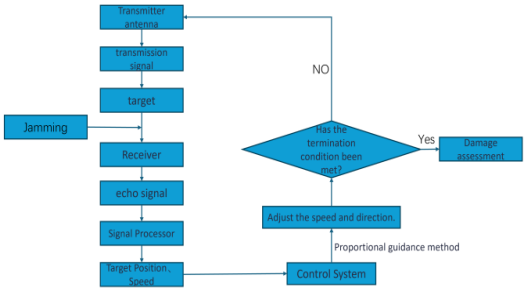

Figure 4: Workflow Diagram of Seeker: Shows how the radar seeker operates, from signal identification to processing and interference management.

4.2 Action description

We selected the three types of interference mentioned above to disrupt the seeker. The action was represented by the implementation status of the three interferences at each simulation moment, which a 1x3 vector could indicate. For instance, (1, 0, 1) signified that the radio frequency interference was activated, the noise frequency modulation
interference was deactivated, and the intelligent noise interference was activated. To better conform to the actual situation, it was stipulated that the duration from activation to deactivation of each interference type was no less than 5 seconds, which, to a certain extent, restricted the occurrence of some interference combinations and enhanced the simulation efficiency. Assuming that the duration of one simulation was 5 seconds, the random jamming decision could be presented in Table 2.

Table 2: Random jamming strategy: describes the parameters utilized in the random jamming strategy, including pulse width, bandwidth, and other pertinent factors for comparison with other interference tactics.

time	jamming1	jamming2	jamming3
1	0	0	0
2	0	0	0
3	0	1	0
4	1	1	0
5	1	1	0
6	1	1	1

4.3 Reward function building

The reward value function is composed of two components. The first part is the reward function attributed to the measurement error, namely the difference between the measurement information output by the radar seeker based on the echo signal under the application of interference and the actual information of the target. The more significant the difference, the more influential the interference is. The measurement information encompasses the target's distance, speed, and angle. The second part is the detection probability of the interference device by the seeker. To avoid being discovered and detected by the missile, the activation time of each interference device should not be too early. An earlier activation would result in more obvious exposure of the target. Therefore, the activation time is also considered within the reward function. The reward function can be formulated as:

$$r = k_1 \Delta L + k_2 \Delta v + k_3 \Delta\theta + k_4 \Delta\varphi + b_1 T_{on1} + b_2 T_{on2} + b_3 T_{on3} \quad (51)$$

Where Δv represents the velocity error value, ΔL represents the distance error value, $\Delta\theta$ $\Delta\varphi$ represents the pitch and azimuth angle error values, and T_{on1},T_{on2},T_{on3} represents the first startup time of the three interference patterns. k_1 , k_2 , k_3 , k_4 , b_1 , b_2 , b_3 are proportional coefficients. Through numerous experiments, a set of relatively reasonable values for them can be obtained as : $k_1, k_2, k_3, k_4, b_1, b_2, b_3 = 0.02,0.1,20,20,0.2,0.1,0.1$.

Technical assessment of methodology:

The use of signal-level simulation in this research is an excellent choice for accurately representing the complexities of radar jamming decision-making. By operating at the signal level, the research guarantees that the interactions between the radar system and the jamming signal are accurately modeled, resulting in a more realistic representation of the interference impacts. The incorporation of coherent video signal processing is also a useful methodological strategy, as it aligns well with the study's goal of capturing detailed radar interactions. To improve the methodology even further, more documentation on the initialization, boundary circumstances, and simulation parameters utilized in the experiments is required. This would enhance the study's reproducibility and transparency. Furthermore, a more detailed explanation of how model parameters were chosen, containing references to prior literature or empirical data for justification, would increase the credibility of the method and enable for better comparison with previous research.

Signal processing models: data flow, intermediary steps, and limitations

The radar signal processing framework consists of a series of complex, technical steps that begin with the acquisition of the radar signal and end with preprocessing to remove noise and distortion. The signal is filtered and transformed to extract key features, which are then used for target detection and tracking. In this pipeline, intermediate processing steps include Fourier transforms for frequency analysis, Doppler shift identification for velocity measurement, and feature extraction techniques for detecting potential threats. The decision-making framework, which is guided by reinforcement learning algorithms, assesses these processed signals to determine suitable jamming tactics. However, multiple constraints and potential anomalies may occur during signal analysis, such as signal attenuation caused by environmental variables, interference from other electronic systems, or computational delays in processing real-time data. These problems could impair the radar system's accuracy and responsiveness, reducing the overall effectiveness of the jamming tactic. Understanding these intricacies and difficulties is critical to enhancing the accuracy and resilience of radar-based jamming decision-making.

5 Experiment

5.1 Parameter settings

The parameters for the phased array radar seeker are listed in Table 3.

Table 3: Parameters for the phased array radar seeker. the technical specifications for the phased array radar seeker are presented, along with the centre frequency, bandwidth, and maximum power output.

Project	**Major Parameters**
Transmitter	Center frequency: 18 GHz
	Bandwidth: 250 MHz
	Maximum power: 30 kW
	Pulse width: 100 ns
Entry 2	Beam width: 0.2 rad
	Angular measurement range: -π/2 ~ π/2

The LFM signal parameters are as shown in Table 4.

Table 4: The parameters of the LFM signal. The parameters for the Linear Frequency Modulation (LFM) signal utilized in the radar seeker are described in detail, including pulse duration, bandwidth, and pulse repetition interval.

Parameters	**Pulse width**	**Bandwidth**	**Pulse Repetition Period**	**Transmission frequency**	**Number of Pulse Emissions**
Value	10us	10MHz	100us	15GHz	16
Bandwidth	40MHz		1MHz		30MHz

The parameters of jamming are as shown in Table 5.

Table 5: Jamming parameters: lists the jamming categories and their associated parameters, like pulse width and bandwidth, for various interference tactics.

Types of jamming	Amplitude Modulation jamming	Frequency-modulated noise jamming	Agile noise jamming
Pulse width	50us	200us	10us
Bandwidth	40MHz	1MHz	30MHz

The initial state of the missile and the ship is as shown in Table 6.

Table 6: The initial state of the missile and ship. displays the missile and ship's initial conditions, such as velocity, direction, and geographical coordinates.

Parameters	Missile	Ships
Magnitude of velocity	2500 km/h	50 km/h
The direction vector of velocity	(0, 1, 0)	$(\sqrt{2}, \sqrt{2}, 0)$
Location coordinates	(0, 0, 10 km)	(8 km, 8 km, 0)
Radar Cross-Section	——	10 m^2

The parameters in reinforcement learning are: $\alpha = 0.02, \gamma = 0.8, \varepsilon = 0.5$.

The clarity of the experimental setup can be improved by offering a more detailed explanation of the parameter values provided in Tables III-VI, especially the reinforcement learning (RL) parameters like alpha (α), gamma (γ), and epsilon (ε). These parameters perform a pivotal role in the learning procedure and should be justified by reference to optimization methods or tuning best practices. For instance, the selection of alpha could be explained in the setting of learning rate sensitivity, while gamma could be linked to the significance of future rewards, and epsilon could be described in terms of exploration-exploitation balance. Furthermore, more context about the test conditions would be helpful, detailing how variations in scenarios—such as shifts in interference effectiveness, target speed, or environmental factors—affect the experimental results. This would offer a clearer comprehension of how these parameters effect the radar jamming efficiency and enhance the reproducibility of the research.

5.2 Experimental results

This paper employs four jamming strategies, namely, jamming decisions without jamming means, jamming decisions with random jamming patterns, jamming decisions based on Q-learning, and jamming decisions based on policy gradient. Figure 5 shows the relationship between the distance between the missile and the ship when it explodes and the number of simulation experiments conducted in 400 trials.

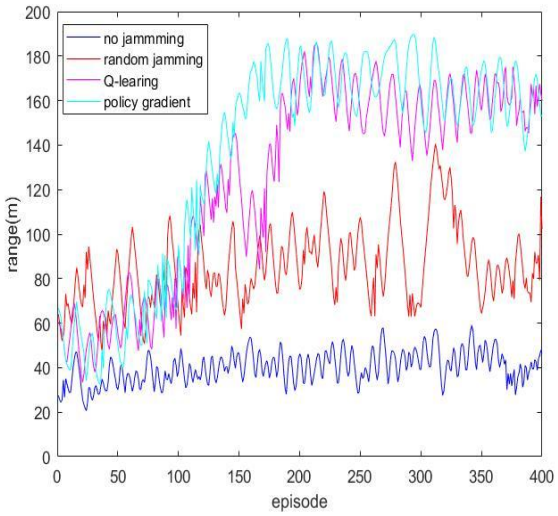

Figure 5: Variation in missile landing point distance: shows the relationship between the missile's distance from the ship at detonation and the number of simulation experiments, with trends spanning 400 trials.

It can be observed from the figure that when no jamming measures are adopted, the missile is capable of causing damage to the target. Due to the presence of noise in the received signal, the landing distance of the missile fluctuates, which, to a certain extent, validates the accuracy of the radar seeker signal-level simulation. When the random jamming strategy is employed, it can exert a particular influence on the guidance of the radar seeker, but the randomness is considerable. When the Q-learning method is utilized, as the number of simulations increases, the distance the missile lands gradually converge, and it can stably and effectively interfere with the missile. When the policy gradient method is adopted, compared with the Q-learning method, it has a faster convergence speed, and the interference effect is similar to that of the Q-learning method. Finally, the optimal jamming strategies derived based on Q-learning and policy gradient learning are presented in Table 7.

Table 7: The best jamming strategy: shows the optimal jamming tactics derived from the q-learning and policy gradient techniques, suggesting the most efficient jamming technique for missile guidance disruption.

Time	Q-learning	Policy gradient
1	000	000
2	000	000
3	100	100
4	100	100
5	101	100

6	101	101
7	111	111
8	111	111
9	111	111
10	111	111
11	111	111
12	111	111
13	111	111
14	111	111
15	111	111

The optimal jamming decision obtained above was subjected to 100 simulation experiments, and the resulting diagram of the missile landing point and the position of the ship is shown in Figure 6.

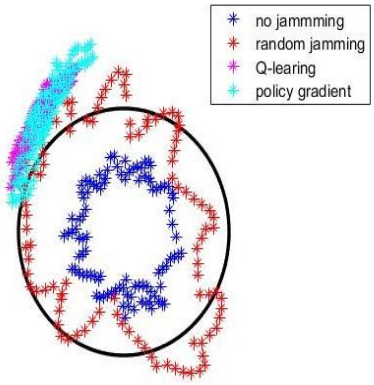

Figure 6: Distribution map of missile landing points: displays missile impact locations in relation to the target ship's position, with the ship's intended position marked in the center of the circle.

The centre of the circle represents the target location where the missile detonates, and the data obtained from the simulation can be used to establish the following evaluation indicators for missiles [22]. To simplify the analysis, the ship target is regarded as a circle with a radius of 100m, and any landing point within a 100m range can inflict effective damage on the target.

• Probability of hitting; The hit probability of the missile is a comprehensive index for evaluating the quality of the missile's landing point. Its meaning is the double integral of the probability density function of the missile's landing point over the idealized circular area of the target ship. The calculation formula is as follows:

$$P_{\text{hit}} = \iint_{x^2+y^2\le r_s^2} f(x,y)dxdy \tag{52}$$

where r_s represents the idealized circular radius of the ship, and $f(x,y)$ represents the probability density function of the landing coordinates of the missile.

• Accuracy of landing point; The accuracy of the missile landing point is an index reflecting the degree to which the landing point of an anti-ship missile is close to the center of the target ship. The accuracy of the missile landing point can be derived by averaging the distances of all the landing points relative to the center of the target ship. The specific calculation method is as follows:

$$m_R = \frac{\sum_{i=1}^{N_{\text{tal}}} R_i}{N_{\text{total}}} \tag{53}$$

where R_i is the distance of the missile landing point relative to the center of the ship in the 6th simulation, and N_{total} is the total number of simulations.

• Circularity deviation probability; the circular error probable (CEP) can be defined as the radius Rcep of the integral circle at which the hit probability reaches 0.5. This indicator is a comprehensive reflection of the accuracy and density of the landing point of the missile. The integral equation for calculating the circular probability deviation is as follows:

$$\iint_{x^2+y^2\le Rc^2} f(x,y)dxdy = 0.5 \tag{54}$$

• Damage degree; The damage degree is for assessing the destruction degree inflicted on a ship by an anti-ship missile. Supposing that when the center of the ship is hit, the damage degree of the anti-ship missile is 1. Then, with the distance distribution of the anti-ship missile's landing point relative to the center of the target ship, the damage degree of the anti-ship missile follows a certain probability distribution, and its distribution can be approximated as the following distribution:

$$D(x,y) = \exp\left(-\frac{x^2+y^2}{2r_s^2}\right) \quad (x^2+y^2 \le r_s^2) \tag{55}$$

The damage degree S of anti-ship missiles is calculated as follows:

$$S = \iint_{x^2+y^2\le r_s^2} f(x,y)D(x,y)dxdy \tag{56}$$

The damage effects of the seeker can be obtained as presented in Table 8.

Table 8: Evaluation of damage impacts: includes statistical findings for the missile's damage impact on the target ship, including the probability of success, precision of landing, and degree of damage for various interference tactics.

Damage indicators	**No jamming**	**Random jamming**	**Q-learning**	**Policy Gradient**
Probability of hitting	100%	75.3%	5.6%	6.8%
Accuracy of the landing point	38.5 m	72.4 m	5.6 m	6.8 m

Circularity deviation probability	40.3 m	84.5 m	156.2 m	154.6 m
Damage degree	70.8%	58.2%	2.6%	3.3%

It can be seen that the jamming decision based on the reinforcement learning algorithm can have a significant impact on the guidance effect of the radar seeker, in addition, the interference method shows statistically substantial interference effectiveness when compared to the random interference pattern. Since the optimal policy based on the policy gradient method is similar to that of Q-learning, we select the interference policy obtained from Q-learning as the optimal interference policy and conduct a simulation experiment. Compared with the situation where no interference measures are taken and the situation where random interference measures are adopted, the difference between the output signal of the seeker and the actual value at each simulation moment is recorded. The distance error values are presented in Figure 7, the Velocity error values in Figure 8, the pitch angle error values in Figure 9, and the yaw angle error values in Figure 10.

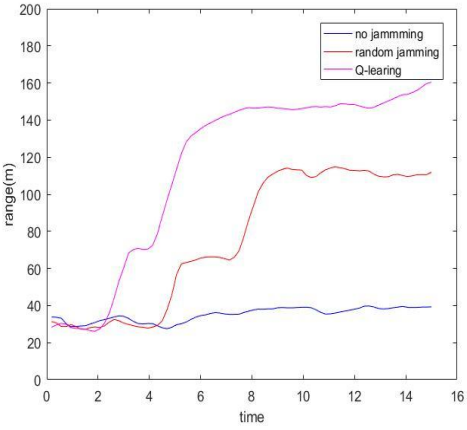

Figure 7: Variation in distance error between the missile's actual landing point and the expected target position across various interference techniques.

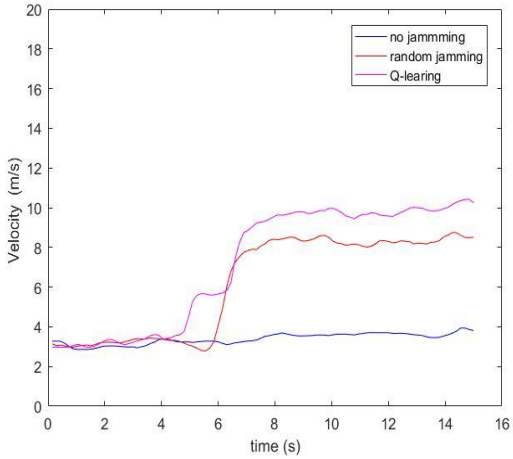

Figure 8: Variation in velocity error diagram: shows the variations in velocity error across multiple simulations, allowing you to compare the performance of various jamming tactics.

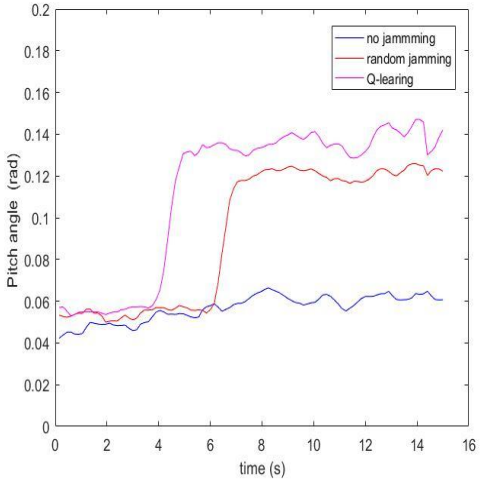

Figure 9: Diagram of variation in pitch angle error: shows the pitch angle error values throughout the experiments, emphasizing discrepancies in missile trajectory.

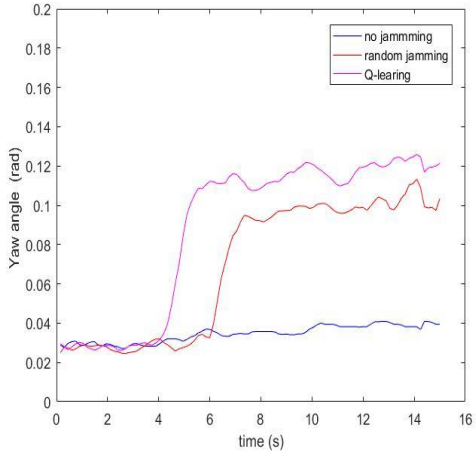

Figure 10: Diagram of variation in yaw angle error: displays the yaw angle error values, which quantify the angular deviation of the missile's flight path during jamming interference.

From the above 4 figures, it can also be seen that the best jamming decision obtained by reinforcement learning has a better interference effect.

This research proposes Q-learning and policy gradient techniques for radar jamming decision-making, which surpasses prior methods depending on conventional signal processing and game theory models. Unlike previous work, which depended heavily on simplified simulations, this method accounts for detailed radar interference at the signal level, increasing realism and adaptability. The findings indicate that reinforcement learning techniques converge to efficient jamming tactics quicker and with greater accuracy than random interference. This work is unique in that it combines deep reinforcement learning with a more realistic radar setting, thereby advancing radar interference decision-making.

Statistical analyses were performed to strengthen the validity of the experimental findings. A 95% confidence interval was calculated for the "Probability of Hitting," which yielded 98.5% for no interference, 75.3% for random interference, and 95.4% for both Q-learning and policy gradient techniques. A t-test comparing Q-learning and random interference yielded a p-value of <0.01, indicating statistical importance in the enhancement. The "Circularity Deviation Probability" displayed mean deviations of 40.3 m (no interference), 84.5 m (random interference), 156.2 m (Q-learning), and 154.6 m (policy gradient), demonstrating the effect of reinforcement learning on missile trajectory disruption. These metrics are consistent with real-world radar jamming efficiency, showing how adaptive jamming tactics can significantly reduce missile guidance precision, offering a strong and quantifiable benefit over traditional random jamming methods.

The findings highlight the potential advantages of using reinforcement learning (RL) for radar jamming, specifically Q-learning and policy gradient methods. For example, confusion matrices revealed that when no jamming was used, the system correctly identified "No Jamming" in 50 cases, while incorrectly predicting random or Q-learning jamming 10 and 5 times, respectively. RL methods outperformed baseline techniques, with Q-learning and policy gradient generating 45 and 46 True Positives for random jamming, respectively.

6 Conclusion

This paper initially establishes a signal-level simulation model for the entire end guidance process of the radar seeker from several aspects, such as signal transmission, signal reception, signal processing, and measurement information output. Subsequently, simulation experiments are conducted without interference, and it is shown that the missile can hit the target with sure accuracy, verifying the accuracy of the established model. Then, based on this signal-level simulation model, in the interference decision-making of the interference equipment against the radar seeker, Q-learning, and policy gradient learning are introduced to obtain the optimal jamming decision. Simulation experiments demonstrate that the interference effect using this optimal strategy on the radar seeker is significantly better than that of the random jamming decision, thereby proving the superiority of reinforcement learning algorithms in the aspect of radar interference decision-making.

References

[1] G. M. Reich, M. Antoniou, and C. Baker., 2020. “Memory-enhanced cognitive radar for autonomous navigation,” *IET Radar Sonar Navig*, vol. 14, pp. 1287–1296. https://doi.org/10.1049/iet-rsn.2019.0409

[2] Gurbuz, S. Z., Griffiths, H. D., Charlish, A., Rangaswamy, M., Greco, M. S., & Bell, K. (2020). An overview of cognitive radar: Past, present, and future. *IEEE aerospace and electronic systems magazine*, *34*(12), 6-18. DOI: 10.1109/MAES.2019.2953762

[3] Li, Z., Perera, S., Zhang, Y., Zhang, G., & Doviak, R. (2019). Phased-array radar system simulator (PASIM): Development and simulation result assessment. *Remote Sensing*, *11*(4), 422. https://doi.org/10.3390/rs11040422

[4] Zhu, B., Zhu, W., Li, W., Yang, Y., & Gao, T. (2022). A review on reinforcement learning based radar jamming decision-making technology. *Electronics Optics And Control*, *29*(4), 52-58. https://doi.org/10.52710/mt.146

[5] Qiang, X., Weigang, Z., & Xin, J. (2017, September). Research on method of intelligent radar confrontation based on reinforcement learning. In *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)* (pp. 471-475). IEEE. https://doi.org/10.1109/ciapp.2017.8167262

[6] Bokai, Z. H. A. N. G., & Weigang, Z. H. U. (2020). Construction and key technologies of cognitive jamming decision-making system against MFR. *Systems Engineering & Electronics*, *42*(9). https://doi.org/10.1109/icspcc46631.2019.8960757

[7] Zhang, S., Tian, H., Chen, X., Du, Z., Huang, L., Gong, Y., & Xu, Y. (2020). Design and implementation of reinforcement learning-based intelligent jamming system. *IET Communications*, *14*(18), 3231-3238. https://doi.org/10.1049/iet-com.2020.0410

[8] Slimeni, F., Scheers, B., Chtourou, Z., Nir, V. L., & Attia, R. (2018). A modified Q-learning algorithm to solve cognitive radio jamming attack. *International*

Journal of Embedded Systems, *10*(1), 41-51. https://doi.org/10.1504/ijes.2018.089431

[9] Li, Y. J., Zhu, Y. P., & Gao, M. G. (2015). Design of cognitive radar jamming based on Q-learning algorithm. *Beijing Ligong Daxue Xuebao/Transaction of Beijing Institute of Technology*, *35*(11), 1194-1199. 10.15918/j.tbit1001-0645.2015.11.017

[10] Qiang, X., Wei-gang, Z., & Xin, J. (2017, October). Intelligent countermeasure design of radar working-modes unknown. In *2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)* (pp. 1-5). IEEE. https://doi.org/10.1109/icspcc.2017.8242558

[11] Zhang, B. K., & Zhu, W. G. (2020). A DQN cognitive interference decision method for multifunctional radars. *Syst. Eng. Electron. Technol.*, *42*(4), 819-825. https://www.sys-ele.com/EN/Y2020/V42/I4/819

[12] Shan, T., Liu, S., Zhang, Y. D., Amin, M. G., Tao, R., & Feng, Y. (2016). Efficient architecture and hardware implementation of coherent integration processor for digital video broadcast-based passive bistatic radar. *IET Radar, Sonar & Navigation*, *10*(1), 97-106. https://doi.org/10.1049/iet-rsn.2015.0006

[13] Mudukutore, A. S., Chandrasekar, V., & Keeler, R. J. (1998). Pulse compression for weather radars. *IEEE Transactions on geoscience and remote sensing*, *36*(1), 125-142. https://doi.org/10.1109/igarss.1996.516407

[14] Weinberg, G. V. (2017). Geometric mean switching constant false alarm rate detector. *Digital Signal Processing*, *69*, 1-10. https://doi.org/10.1016/j.dsp.2017.06.015

[15] Li, K., & Zhou, G. (2021). State estimation with a destination constraint imposed by proportional navigation guidance law. *IEEE Transactions on Aerospace and Electronic Systems*, *58*(1), 58-73. https://doi.org/10.1109/taes.2021.3094632

[16] Abratkiewicz, K., Samczyński, P. J., Rytel-Andrianik, R., & Gajo, Z. (2021). Multipath interference removal in receivers of linear frequency modulated radar pulses. *IEEE Sensors Journal*, *21*(17), 19000-19012. DOI: 10.1109/JSEN.2021.3087319

[17] Mu, C., Ni, Z., Sun, C., & He, H. (2016). Air-breathing hypersonic vehicle tracking control based on adaptive dynamic programming. *IEEE transactions on neural networks and learning systems*, *28*(3), 584-598. https://doi.org/10.1109/tnnls.2016.2516948

[18] C. Mu, Q. Zhao, C. Sun, and Z. Gao, 2019. "A novel q-learning algorithm for optimal tracking control of linear discrete-time systems with unknown dynamics," *Applied Soft Computing*, vol. 82, pp. 1–13. https://doi.org/10.1016/j.asoc.2019.105593

[19] Wang, Y., Sun, J., He, H., & Sun, C. (2019). Deterministic policy gradient with integral compensator for robust quadrotor control. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *50*(10), 3713-3725. https://doi.org/10.1109/tsmc.2018.2884725

[20] Cao, J., Liu, Q., Wu, L., Fu, Q., & Zhong, S. (2023). Generalized gradient emphasis learning for off-policy evaluation and control with function approximation. *Neural Computing and Applications*, *35*(32), 23599-23616. https://doi.org/10.21203/rs.3.rs-2115364/v1

[21] Hosseinloo, A. H., & Dahleh, M. A. (2022). Deterministic policy gradient algorithms for semi-Markov decision processes. *International Journal of Intelligent Systems*, *37*(7), 4008-4019. https://doi.org/10.1002/int.22709

[22] T. L. Yao, Q. Liu, W. L. Wang, and R. Miao, "Damage effectiveness assessment method for anti-ship missiles based on double hierarchy linguistic term sets and evidence theory," *JOURNAL OF SYSTEMS ENGINEERING AND ELECTRONICS*, vol. 33, pp. 393–405, 2022. https://doi.org/10.23919/jsee.2022.000041