ISSN:2790-1688 Volume-15-(2025)

Intelligent Joint Optimization of Detection and Guidance Based on Convex Optimization Pre-training

Yuhang Guo [1], Chuanjun Li [1], and Jingquan Ma [1]

[1] School of Aerospace Engineering, Beijing Institute of Technology, Beijing, China;

Abstract. To address the challenges faced by traditional control methods in integrating the scheduling of detection and trajectory resources during the terminal guidance phase of hypersonic glide vehicles (HGVs), this paper proposes an intelligent joint optimization technique based on convex optimization pre-training. Using the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm, we introduce an adaptive detection-guidance weight distribution reward function. This design ensures that the vehicle meets guidance requirements during the terminal phase while allocating sufficient trajectory resources for radar detection. In interference suppression scenarios, the agent can avoid interference zones through trajectory resource scheduling and implement frequency hopping countermeasures via detection resource allocation. By jointly optimizing detection and trajectory resources, the method enhances terminal guidance confrontation performance. Additionally, to improve training efficiency, convex optimization-generated trajectory data is used for pre-training the agent, significantly reducing the convergence time. Simulation results show that this method effectively improves the performance of detection and counter-interference in the terminal guidance phase, providing new technical means for intelligent guidance in complex battlefield environments.

Keywords: deep reinforcement learning; TD3 algorithm; intelligent detection-guidance; convex optimization; counter-interference.

1. Introduction

Hypersonic glide vehicles (HGVs) play a critical role in modern warfare due to their high maneuverability and extended glide range [1]. However, during the terminal guidance phase, HGVs must overcome multiple challenges in highly dynamic and complex environments, including precise target engagement, the scheduling of detection resources, and effective jamming avoidance. Traditional control strategies often treat guidance and detection as separate issues, lacking a holistic view of their interdependencies and struggling to handle the coupling between trajectory and detection resources [2]. Moreover, advances in counter-interference technology have further complicated the terminal guidance process, necessitating more sophisticated coordinated optimization of both detection and guidance.

In recent years, deep reinforcement learning (DRL) has garnered considerable attention for its remarkable performance in high-dimensional decision-making problems [3]. The TD3 algorithm, an improved variant of the Deep Deterministic Policy Gradient (DDPG) method, incorporates dual Q-networks, delayed policy updates, and target policy smoothing to alleviate overestimation biases [4]. These enhancements yield greater policy stability and sample efficiency when optimizing continuous action spaces. In aerospace applications, TD3 has been successfully applied to flight attitude control and intelligent guidance tasks.

Motivated by these advances, this study addresses the joint optimization and allocation of detection and guidance resources during HGV terminal guidance. By embedding the TD3 algorithm into a coordinated optimization framework and designing a time-varying reward function for detection-guidance weight allocation, we achieve integrated objectives of trajectory optimization, detection resource scheduling, and interference countermeasures. Additionally, to boost training efficiency, convex optimization is employed to fit angle-of-attack curves, effectively narrowing the search space and accelerating convergence. Simulation results validate the advantages of the proposed method in terms of detection performance, jamming resistance, and training speed, thereby offering fresh insights into intelligent guidance technology for HGVs..

2. Modeling

2.1 Vehicle Model

2.1.1 Dynamics Model

The dynamics model of the HGV adopts the CAV-H model. For simplicity, the Earth's rotation is ignored [5]. The three-degree-of-freedom dynamic equations are established as follows:

$$\begin{cases} \dfrac{dr}{dt} = V\sin\gamma \\ \dfrac{d\theta}{dt} = \dfrac{V\cos\gamma\sin\psi}{r\cos\phi} \\ \dfrac{d\phi}{dt} = \dfrac{V\cos\gamma\cos\psi}{r} \\ \dfrac{dV}{dt} = -\dfrac{D}{m} - g\sin\gamma \\ \dfrac{d\gamma}{dt} = \dfrac{1}{V}\left[\dfrac{L\cos\sigma}{m} - \left(g - \dfrac{V^2}{r}\right)\cos\gamma\right] \\ \dfrac{d\psi}{dt} = \dfrac{1}{V}\left[\dfrac{L\sin\sigma}{m\cos\gamma} + \dfrac{V^2}{r}\cos\gamma\sin\psi\tan\phi\right] \end{cases}$$

Here, r is the distance from the Earth's center; θ and ϕ denote the longitude and latitude of the vehicle; V represents the velocity; γ and ψ are the flight path and heading angles; σ is the sideslip angle; m is the mass of the vehicle; g denotes gravitational acceleration; and L and D represent the lift and drag forces, respectively.

2.1.2 Aerodynamic Model

The aerodynamic model is derived by fitting empirical aerodynamic data as referenced in [5]. The aerodynamic parameters are computed as follows:

$$\begin{cases} D = 0.5\rho V^2 C_d S \\ L = 0.5\rho V^2 C_l S \\ C_d = 0.023 + 20.378\alpha + 0.398\,exp\,(-7.078 * 10^{-4} * V) \\ C_l = -0.234 + 9.466\alpha + 0.297\,exp\,(-3.393 * 10^{-4} * V) \\ \rho = 1.226\,exp\,(-1.3785 * 10^{-4} * H) \end{cases}$$

where H denotes the altitude, ρ is the atmospheric density, α is the angle of attack, S is the reference area, and C_l and C_d denote the lift and drag coefficients, respectively. D and L denote the lift and drag.

2.1.3 Constraint Conditions

Due to structural limitations, HGVs must satisfy certain rigid constraints during flight, including limits on heat flux, dynamic pressure, and overload [6]. These constraints are modeled as:

$$\begin{cases} \dot{Q} = k_Q\rho^{0.5}V^{3.15} \le \dot{Q}_{max} \\ n = \sqrt{L^2 + D^2}/(mg_0) \le n_{max} \\ q = \dfrac{1}{2}\rho V^2 \le q_{max} \end{cases}$$

where the terms represent the heat flux at a steady state ($\dot{Q}$), maximum allowable heat flux ($\dot{Q}_{max}$), heat flux model coefficient (k_Q, related to the vehicle's nose radius and heat shielding material), atmospheric density (ρ), overload (n, with a maximum limit), sea-level gravitational acceleration (g_0), dynamic pressure (q), and the maximum allowable dynamic pressure (q_{max}).

2.2 Radar Detection and Jamming Model

A simulation model of phased-array radar and jamming signals is constructed as part of the training environment [7]. This model accounts for factors such as radar transmit power (P_t), antenna gain (G_{radar}), target radar cross-section , target distance, background noise, jammer transmit power, center frequency, and bandwidth. It computes the received signal power (P_r), interference power ($P_{jamming}$), signal-to-noise ratio ($SNR_{radar\text{-}dB}$), and detection probability ($P_{detection}$). The detection probability is evaluated using the Marcum Q-function [8], defined as:

$$P_{\text{detection}} = Q_1\left(\sqrt{2 \cdot \text{SNR}_{\text{radar-linear}}}, \sqrt{2 \cdot \lambda_{\text{threshold}}}\right)$$

where $SNR_{radar\text{-}linear}$ is the linear value of the radar SNR and $\lambda_{threshold}$ is a threshold parameter in the detection process. Key formulas of the model are detailed below.

2.2.1 Radar Equation

The radar equation relates the received signal power to the target RCS [9]. According to the basic radar function, the received signal power is calculated considering the radar's transmit power (P_t), antenna gain (G_{radar}), target RCS (σ_{target}), propagation distance ($d_{radar-target}$) attenuation, and radar wavelength (λ_{radar}). Similarly, interference power is computed based on parameters such as jammer distance ($d_{jamming-radar}$), transmit power ($P_{jamming}$), bandwidth ($b_{jamming}$), and jamming wavelength ($\lambda_{jamming}$). The SNR is determined by the ratio of signal power to interference power, adjusted for background noise (N_0). Meanwhile, f represents the operating frequency. The calculation formulas are as follows:

$$P_r = P_t + 2G_{radar} + 10\log_{10}(\sigma_{target}) - 40\log_{10}(d_{radar-target}) + 20\log_{10}(\lambda_{radar}) - 20\log_{10}(4\pi)$$

$$P_{jamming} = P_{jamming} - 20\log_{10}(d_{jamming-radar}) + 20\log_{10}(\lambda_{jamming}) - 20\log_{10}(4\pi)$$

$$\text{SNR}_{\text{radar-dB}} = \begin{cases} P_{\text{signal}} - P_{\text{jamming}} - N_0 & |f_{\text{radar}}\text{-}f_{\text{jamming}}| < \dfrac{b_{\text{jamming}}}{2} \\ P_{\text{signal}} - N_0 & |f_{\text{radar}}\text{-}f_{\text{jamming}}| \geq \dfrac{b_{\text{jamming}}}{2} \end{cases}$$

2.2.2 Radar Jamming Parameter Settings

Based on [10], and adjust according to the scenario, the radar and jamming parameters are set as follows:

Table 1. Three Scheme comparing

Parameter	Value	Unit
Radar Frequency	7.50E+09 (time-varying)	Hz
Transmit Power	60 (time-varying)	dBm
Element Gain	25	dB
Target RCS	1	m^2
Target Distance	250000 (time-varying)	m
Background Noise	-100	dBm
Jamming Power	30	dBm
Jamming Frequency	3.30E+09 (time-varying)	Hz
Jamming Bandwidth	1.00E+08	Hz
Detection Threshold	5	Unitless
Number of Array Elements	128	Unitless

3. Algorithm Design

3.1 Convex Optimization Pre-training Technique

3.1.1 Distance-Domain Trajectory Planning Model

Due to the complex estimation of the remaining terminal guidance time, the time-domain model is transformed into a distance-domain model for convex optimization trajectory planning. In the original time-dependent model, the relative distance to the target is used as the independent variable [11]. The initial flight distance for trajectory planning is computed as:

$$S_{f_0} = arccos\left(cos\left(\phi_f\right) cos\left(\phi_0\right) cos\left(\theta_f - \theta_0\right) + sin\left(\phi_f\right) sin\left(\theta_0\right)\right)$$

Here, θ_0 and ϕ_0 denote the initial latitude and longitude coordinates, θ_f and ϕ_f represent the target's coordinates. With a given initial and target point, the trajectory planning problem is recast as one in which the relative distance decreases. Assuming an approximate conversion between unit distance dS and unit time dt, the relationship is given by:

$$\frac{dt}{dS} = -\frac{1}{V\,cos\left(\gamma\right) cos\left(\psi - \psi_p\right)/r}$$

Taking the initial heading angle of the target relative to the vehicle as a reference, the target's heading at the initial point is expressed as:

$$\psi_p = arcsin\left(\frac{sin\left(\theta_f - \theta_0\right) cos\left(\phi_f\right)}{sin\left(S_{f_0}\right)}\right)$$

The motion equations in the distance domain then become:

$$\begin{cases} \dfrac{dr}{ds} = \dfrac{-r\,sin\left(\gamma\right)}{cos\left(\gamma\right) cos\left(\psi - \psi_p\right)} \\ \dfrac{d\theta}{ds} = \dfrac{-sin\left(\psi\right)}{cos\left(\phi\right) cos\left(\psi - \psi_p\right)} \\ \dfrac{d\phi}{ds} = \dfrac{-cos\left(\psi\right)}{cos\left(\psi - \psi_p\right)} \\ \dfrac{dv}{ds} = \dfrac{Dr}{vm\,cos\left(\gamma\right) cos\left(\psi - \psi_p\right)} + \dfrac{Gm\,sin\left(\gamma\right)}{vr\,cos\left(\gamma\right) cos\left(\psi - \psi_p\right)} \\ \dfrac{d\gamma}{ds} = \dfrac{-L\,cos\left(\sigma\right) r}{mv^2\,cos\left(\gamma\right) cos\left(\psi - \psi_p\right)} - \dfrac{1}{cos\left(\psi - \psi_p\right)} + \dfrac{Gm}{rv^2\,cos\left(\psi - \psi_p\right)} \\ \dfrac{d\psi}{ds} = -\dfrac{rL\,sin\left(\sigma\right)}{mv^2\,cos^2\left(\gamma\right) cos\left(\psi - \psi_p\right)} - \dfrac{tan\left(\phi\right) sin\left(\psi\right)}{cos\left(\psi - \psi_p\right)} \end{cases}$$

where G is the standard gravitational parameter of the Earth. All kinematic variables are normalized according to the method detailed in [5].

3.1.2 Convex Optimization Training

The re-entry trajectory planning problem is modeled as a non-convex, infinite-dimensional optimal control problem, encompassing nonlinear dynamics, path constraints, and cooperative task constraints. To solve this problem via convex optimization, the model is convexified and discretized. The dynamic equations are linearized and discretized, with the state vector chosen as $X = [r, \theta, \phi, v, \gamma, \psi, \alpha, \sigma]$ and the control vector as $u = [\dot{\alpha}, \dot{\sigma}]$, so that the dynamics can be reformulated as:

$$\dot{X} = AX + BU + C$$

Similarly, the heat flux, overload, and dynamic pressure constraints are linearized and discretized. Using the MOSEK solver, the convex optimization problem is solved to obtain the trajectory data [12].

3.2 Twin Delayed Deep Deterministic Policy Gradient Algorithm

This study employs the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm as the core DRL method [13]. TD3 is particularly well-suited for high-dimensional continuous action spaces and effectively mitigates the overestimation issue inherent in DDPG by employing two Critic networks and selecting the smaller Q-value during updates. In addition, noise sampled from a normal distribution is added to the target action, smoothing the Q-value function update, and the policy network is updated with a delay to stabilize the Q-value estimation [14]. The TD3 architecture consists of six networks: an Actor network, two Critic networks (Critic0 and Critic1), and their corresponding Target networks. During offline training, the Actor network selects the policy while the two Critic networks evaluate the state-action pairs using normalized state inputs and outputs from the Actor, yielding expected cumulative rewards [15]. The Target networks, which mirror the structure of the main networks, are used to compute target values by taking the minimum of the two estimated Q-values. The network structure is shown in the figure below:

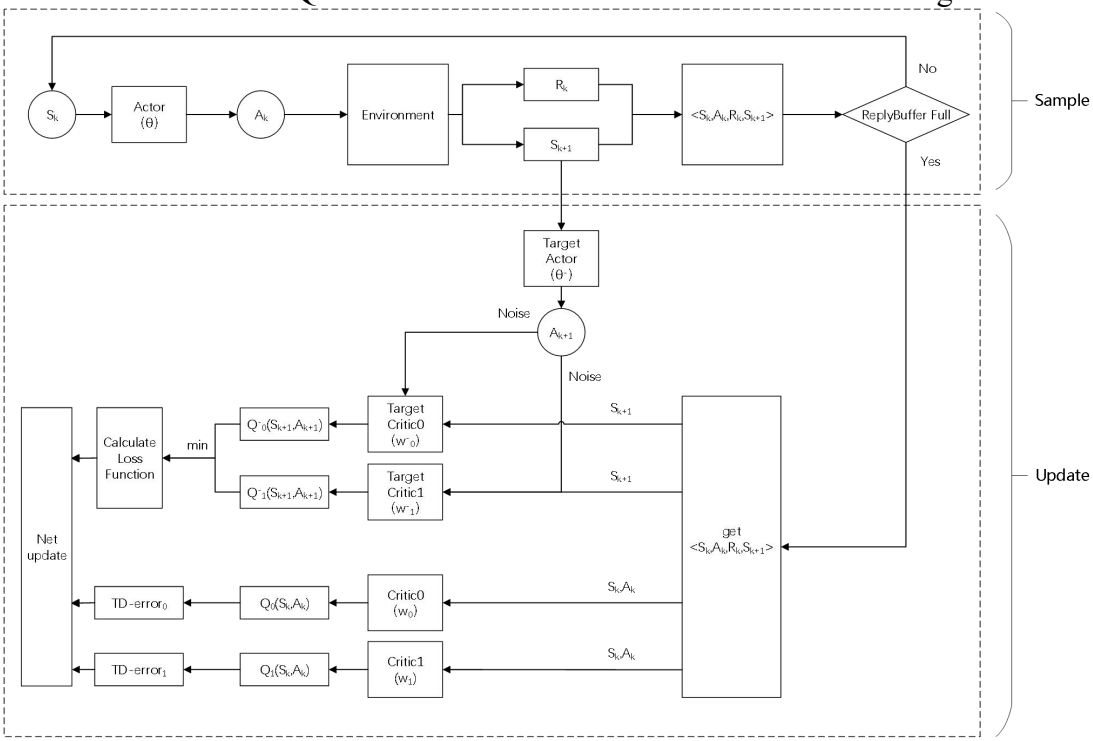

Fig. 1 TD3 Algorithm Structure

During offline training, the algorithm uses the Actor network to select policies and employs two Critic networks to evaluate state-action pairs. The normalized state information is fed into the Actor network as input, which then outputs actions in the action space. The Critic networks take both the state and action outputs as inputs and generate the expected cumulative reward $Q^{\pi}(s_t, a_t)$:

$$Q^{\pi}(s_t, a_t) = E_{\pi}\{r_t(s_t, a_t) + \gamma E_{\pi}[Q^{\pi}(s_{t+1}, a_{t+1})]\}$$

The three Target networks share the same structure as their non-Target counterparts and are primarily used to mitigate overestimation issues. When computing target values, the smaller value between the two is used to estimate the state-action value of the next state-action pair.

$$y = r + \gamma min_{0,1}\, Q'_i(s', a' \,|\, \theta_i^{Q'})$$

After training the reinforcement learning model using the TD3 algorithm to obtain the optimal policy function, this policy can be deployed in the online testing environment to validate the effectiveness of the agent's joint decision-making.

3.3 Intelligent Guidance Method with Jamming Countermeasures

3.3.1 Construction of a Jamming-Aware Environment

Under jamming scenarios, the DRL agent's interaction environment models both terminal guidance and dynamic jamming processes [16, 17]. The design is as follows:

(1) State Space

$$\boldsymbol{s}_t = [d_t, \phi_{vt}, \theta_{vt}, v_t, dQ_t, dn_t, dq_t, \mathrm{SNR}_t, J_t]^T$$

The state vector comprises the normalized relative distance (d_t) between the vehicle and the target, azimuth error (ϕ_{vt}), and pitch error (θ_{vt}); the vehicle's speed (v_t); the normalized differences between current heat flux (dQ_t), overload (dn_t), and dynamic pressure (dq_t) and their respective limits; as well as the normalized current SNR_t from detection and the jamming frequency (J_t) deviation.

(2) Action Space

$$\boldsymbol{a}_t = [\sigma, \alpha, f_d, P_t]^T$$

The action space includes parameters for adjusting both the vehicle's trajectory and its detection settings. The trajectory control variables are the sideslip angle (σ) and angle of attack (α), while the detection control variables are the center frequency variation (f_d) and the percentage change in transmit power (P_t).

3.3.2 Construction of the Detection-Guidance Weight Allocation Reward Function

The reward structure is divided into sparse and non-sparse rewards. Sparse rewards assess the overall effectiveness of the guidance process, while non-sparse rewards guide the agent towards preferable detection-guidance strategies.

(1) Sparse Guidance Outcome Reward

$$R_{\text{guide}} = \begin{cases} k_{g1} \cdot e^{-5\Delta d} & ||\Delta \boldsymbol{x}|| \le 0.1 d_{\text{CEP}} \\ -k_{g2} & ||\Delta \boldsymbol{x}|| > d_{\text{CEP}} \\ -k_{g2} & Rigid\ restraint\ fails \end{cases}$$

where the terms represent the absolute value of the relative distance ($||\Delta x||$), the normalized terminal error ($\Delta d = \left|\left|\Delta x\right|\right|/d_{CEP}$), and a weighting coefficient (k_{gi}).

(2) Sparse Detection Outcome Reward

$$R_{\text{det}} = \begin{cases} k_{d1} \cdot (1 + I_{\text{track}}) \\ k_{d2} \cdot N_{\text{avoid}} \\ -k_{d3} \cdot T_{\text{jamming}}, t_{\text{jamming}} > t_{\text{th}} \end{cases}$$

where the reward is based on a binary indicator for successful target detection (I_{track}), the accumulated SNR, the number of successful jamming avoidances (N_{avoid}), total dwell time under jamming ($T_{jamming}$), and a weighting coefficient (k_{di}).

(3) Non-sparse Joint Detection-Guidance Optimization Reward

Two reward functions are designed according to task-specific requirements: one for guidance optimization and one for detection optimization:

$$R_{\text{guidance}} = -\lambda_1 \parallel \Delta v \parallel^2 - \lambda_2 |\gamma - \gamma_{\text{ref}}| + \lambda_3 I_{\text{stable}}$$

$$R_{\text{detection}} = \eta_1 \Delta f_{\text{eff}} - \eta_2 \frac{P_t}{P_{\max}} + \eta_3 I_{\text{lock}}$$

$$\Delta f_{\text{eff}} = \sum_{t=1}^{T} \left(1 - \frac{|f_t - f_{\text{jam}}|}{f_{\max}}\right) I_{\text{active}}$$

where the coefficients represent the reward scaling factors for guidance and detection(λ_i, η_i). Reference values such as the desired heading angle (γ_{ref}), stability indicators (I_{stable}), detection lock status(I_{lock}), frequency offset compensation reward (Δf_{eff}), environmental jamming frequency (f_{jam}),

and jamming activation flags (I_{active}) are incorporated. To address the time-varying nature of the weights, dynamic weighting coefficients for detection (ω_g) and guidance (ω_d) are set as functions of the relative distance. In the early stages of the trajectory, detection rewards are prioritized, whereas in the terminal phase, guidance rewards dominate. The final dynamic joint optimization reward obtained by the agent is given by:

$$\omega_g = \frac{1}{1 + e^{-k(J_{\text{norm}} - 0.5)}}$$

$$\omega_d = 1 - \omega_g$$

with the normalized jamming frequency deviation ($J_{norm} = J_t/J_{max}$) as one of the parameters. The dynamic weight reward obtained by the agent is as follows:

$$R_{\text{dense}} = \omega_g R_{\text{guidance}} + \omega_d R_{\text{detection}}$$

(4) Non-sparse Finite-State Machine Reward

Stage rewards are designed based on a finite-state machine corresponding to different jamming states, encouraging the agent to learn strategies to evade jamming zones.

$$R_{\text{FSM}} = \begin{cases} +2/\text{step}, & I_{\text{jamming}} = 1 \cap \Delta f_{\text{eff}} > 0.7 \\ +5, & Region\ switching \\ -1/\text{step}, & I_{\text{jamming}} = 1 \cap P_t < 0.2 P_{\max} \end{cases}$$

(5) Non-sparse Energy Management Reward

An energy management reward is defined to encourage energy savings during the jamming countermeasure phase, where coefficients balance the initial energy, current overload, and detection power. The final energy management reward is computed as a percentage of the initial energy standard.

$$R_{\text{energy}} = \nu \left(c_0 E_{\text{init}} - \int_0^t (c_1 \parallel n_t \parallel^2 + c_2 P_t) dt \right)$$

The overall reward structure, achieved through state-dependent dynamic weighting, enables the agent to autonomously balance guidance and detection. Layered sparse rewards guide key decision nodes, while an escape penalty enforces a "failed frequency compensation → power reduction escape → re-guidance" logic. The final rewards for the agent are:

$$R = k_1 R_{\text{guide}} + k_2 R_{\text{det}} + k_3 R_{dense} + k_4 R_{FSM} + k_5 R_{energy}$$

4. Simulation and Results Analysis

4.1 Analysis of Convex Optimization Trajectory Pre-training Results

After obtaining the initial trajectory data through convex optimization, the corresponding control commands serve as output to interact with the environment, yielding reward data for the convex optimization trajectory. Both the Actor and Critic networks are then trained via supervised learning [18]. The training loss curves for both networks are presented as follows:

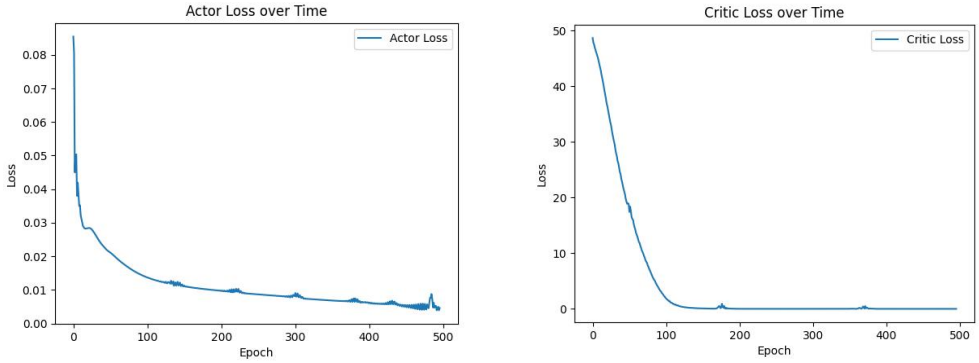

Fig. 2 Actor and Critic Network Loss Curve

Once the full reinforcement learning interaction begins, the reward curve demonstrates that, without pre-training, the reward curve struggles to converge. In contrast, the pre-trained network converges significantly faster during interactive training. As shown in the figure below, the pre-trained agent reaches convergence at around 2500 episodes—a speed improvement of at least 50%.

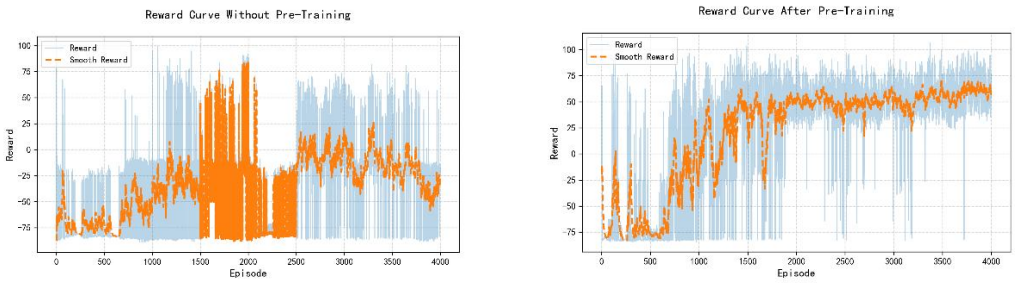

Fig. 3 Comparison of Training Efficiency

4.2 Frequency Hopping Countermeasure Strategy Test and Analysis

At an initial relative distance of 250 km and an initial speed of 5 Mach, in scenarios with jamming that permits frequency hopping, the agent opts for frequency hopping countermeasures without heavily adjusting trajectory resources. Test results showing the variation of the vehicle's radar operating frequency, environmental interference frequency in the current region, and time are as follows:

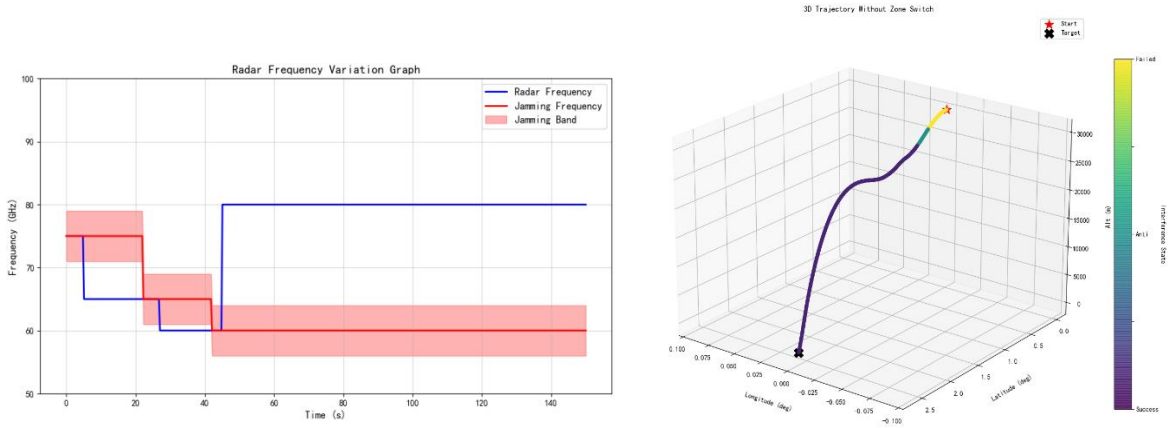

Fig. 4 Result Diagram of Frequency Hopping Countermeasure Strategy

The figure indicates that when the vehicle is within a counterable jamming zone, it adopts frequency hopping to counter the interference. Upon successful jamming countermeasures, the vehicle maintains its trajectory for attack; the corresponding guidance trajectory is depicted in the subsequent figure.

4.3 Test and Analysis of Silent Evasion Strategy in Jamming Domains

With an initial relative distance of 250 km and an initial speed of 5 Mach, an additional non-counterable jamming zone is defined. In this scenario, the agent chooses to adjust its trajectory to evade the jamming zone and simultaneously reduce radar transmit power to lower the probability of being detected. Once evasion is completed, the radar is reactivated. Test results showing the changes in radar transmit power over time are presented in the figure below:

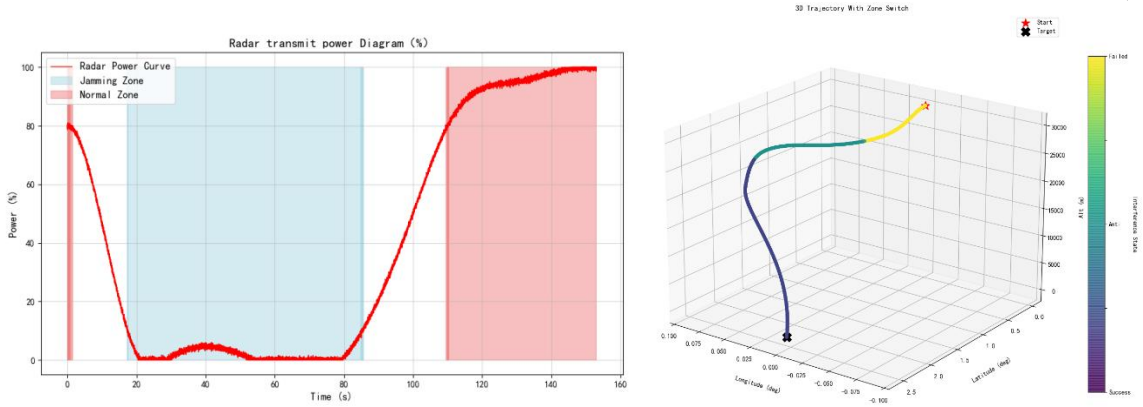

Fig. 5 Result Diagram of Silent Evasion Strategy

The figure demonstrates that when the vehicle is initially within a strong jamming zone, if frequency hopping fails to counter the interference over a certain duration, the agent reduces its radar transmit power and adjusts trajectory resources to fly around the jamming zone.

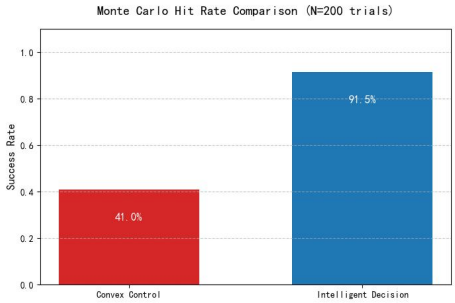

Fig. 6 Monte Carlo Shooting Confrontation Result Diagram

Monte Carlo tests comparing the agent's decisions with convex optimization trajectories reveal that [19], over 200 episodes (with 100 episodes under counterable jamming and 100 episodes under non-counterable jamming), the convex optimization method achieves only a 41% success rate, markedly lower than the 91.5% success rate obtained with intelligent decision-making.

5. Summary

This paper addresses the challenge of joint optimization of detection and trajectory resource allocation during the terminal guidance phase of hypersonic glide vehicles. An intelligent joint optimization technique that combines convex optimization pre-training with deep reinforcement learning is proposed, yielding the following major contributions:

(1) Pre-training Acceleration Mechanism: By pre-training the agent with high-quality initial trajectory data generated through convex optimization, the convergence of the TD3 algorithm is accelerated from over 5000 episodes to approximately 2500 episodes—enhancing training efficiency by at least 50% and mitigating the cold-start problem in complex dynamic constraints.

(2) Multi-Objective Dynamic Reward Formulation: A dynamic reward system for joint detection and guidance optimization is developed based on prior knowledge. This system simultaneously considers the vehicle's inherent constraints and guides the agent to converge through a layered sparse reward mechanism.

(3) Generalization of Jamming Countermeasure Strategies: The agent is capable of autonomously choosing between frequency hopping and trajectory evasion strategies in response to different jamming scenarios. Experimental results show that, in Monte Carlo tests, the joint intelligent scheduling of power and trajectory resources enhances the countermeasure success rate by 50.5%.

References

[1] Xu, Hui;Cai, GuangbinCAa;Cui, Yalong;Hou, Mingzhe;Yao, Erliang.Reentry trajectory optimization method of hypersonic glide vehicle[J].Harbin Gongye Daxue Xuebao/Journal of Harbin Institute of Technology,2023,Vol.55(4): 44-55.

[2] Yibo D, Xiaokui Y U E, Guangshan C, et al. Review of control and guidance technology on hypersonic vehicle[J]. Chinese Journal of Aeronautics, 2022, 35(7): 1-18.

[3] Barto A G. Reinforcement Learning: An Introduction. By Richard's Sutton[J]. SIAM Rev, 2021, 6(2): 423.

[4] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C]//International conference on machine learning. PMLR, 2018: 1587-1596.

[5] Li C, Ma J, Liang X, et al. A segmented trajectory planning and guidance method for hypersonic glide vehicles considering target detection performance[J]. Aerospace Science and Technology, 2024, 153: 109461.

[6] Phillips T H. A common aero vehicle (CAV) model, description, and employment guide[J]. Schafer Corporation for AFRL and AFSPC, 2003, 27: 1-12.

[7] Zhao Z, Yuan J, Li M. Research on adaptive waveform optimization design of anti-jamming radar[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1650(2): 022111.

[8] Marcum J. A statistical theory of target detection by pulsed radar[J]. IRE Transactions on Information Theory, 1960, 6(2): 59-267.

[9] Richards M A. Fundamentals of Radar Signal Processing, 2nd edMcGraw-Hill[J]. New York, 2014.

[10] Bo Liu. Array Millimeter-wave Phased Antenna Radar Seeker Technology Research [D]. SiChuan: University of Electronic Science and Technology of China,2018.

[11] Chai R, Tsourdos A, Savvaris A L, et al. Trajectory planning for hypersonic reentry vehicle satisfying deterministic and probabilistic constraints[J]. Acta Astronautica, 2020, 177: 30-38.

[12] Ma J, Chen H, Wang J, et al. Real-Time Trajectory Planning for Hypersonic Entry Using Adaptive Non-Uniform Discretization and Convex Optimization[J]. Mathematics, 2023, 11(12): 2754.

[13] Li J, Song S, Shi X. Deep reinforcement learning based trajectory real-time planning for hypersonic gliding vehicles[J]. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering, 2024, 238(16): 1665-1682.

[14] Lillicrap T P. Continuous control with deep reinforcement learning[J]. arXiv preprint arXiv:1509.02971, 2015.

[15] Peters J, Schaal S. Reinforcement learning of motor skills with policy gradients[J]. Neural networks, 2008, 21(4): 682-697.

[16] Mannion P, Devlin S, Duggan J, et al. Reward shaping for knowledge-based multi-objective multi-agent reinforcement learning[J]. The Knowledge Engineering Review, 2018, 33: e23.

[17] Gu S, Yang L, Du Y, et al. A review of safe reinforcement learning: Methods, theory and applications[J]. arXiv preprint arXiv:2205.10330, 2022.

[18] Cunningham P, Cord M, Delany S J. Supervised learning[M]//Machine learning techniques for multimedia: case studies on organization and retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 21-49.

[19] James F. Monte Carlo theory and practice[J]. Reports on progress in Physics, 1980, 43(9): 1145.