



REGULAR PAPER

Adaptive reinforcement learning control for a class of missiles with aerodynamic uncertainties and unmodeled dynamics

X. Ning^{1,2,3}, S. Cao^{1,2,4}, B. Han⁵, Z. Wang^{1,6,7,8}  and Y. Yin^{1,3} 

¹National Key Laboratory of Aerospace Flight Dynamics, Northwestern Polytechnical University, Xi'an, China, ²Science and Technology on Electromechanical Dynamic Control Laboratory, Xi'an, China, ³School of Astronautics, Northwestern Polytechnical University, Xi'an, China, ⁴Xi'an Institute of Electromechanical Information Technology, Xi'an, China, ⁵Xi'an Aeronautics Computing Technique Research Institute, Xi'an, China, ⁶Research Center for Unmanned System Strategy Development, Northwestern Polytechnical University, Xi'an, China, ⁷Northwest Institute of Mechanical and Electrical Engineering, Xianyang, China and ⁸Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an, China
Corresponding author: Z. Wang; Email: wz_nwpu@126.com

Received: 8 June 2022; **Revised:** 7 February 2023; **Accepted:** 8 April 2023

Keywords: straight air compound missile system; unmodeled dynamics; reinforcement learning; super-twisting disturbance observer

Abstract

In this paper, a super-twisting disturbance observer (STDO)-based adaptive reinforcement learning control scheme is proposed for the straight air compound missile system with aerodynamic uncertainties and unmodeled dynamics. Firstly, neural network (NN)-based adaptive reinforcement learning control scheme with actor-critic design is investigated to deal with the tracking problems for the straight gas compound system. The actor NN and the critic NN are utilised to cope with the unmodeled dynamics and approximate the cost function that are related to control input and tracking error, respectively. In other words, the actor NN is used to perform the tracking control behaviours, and the critic NN aims to evaluate the tracking performance and give feedback to actor NN. Moreover, with the aid of the STDO disturbance observer, the problem of the control signal fluctuation caused by the mismatched disturbance can be solved well. Based on the proposed adaptive law and the Lyapunov direct method, the eventually consistent boundedness of the straight gas compound system is proved. Finally, numerical simulations are carried out to demonstrate the feasibility and superiority of the proposed reinforcement learning-based STDO control algorithm.

Nomenclature

α, β	attack angle and slide angle
ω_z, ω_y	angle velocities
J_z, J_y	rotational inertia
δ_z, δ_y	output signals of the elevator and rudder
m	mass of the missile
V	velocity of the missile
S	reference area of the missile
L	reference length of the missile
$C_{(i)}^{(j)}$	coefficients of the aerodynamic forces
$m_{(i)}^{(j)}$	coefficients of the aerodynamic moments
$d_i (i = \omega_z, \omega_y)$	unknown disturbances
$\chi_i (i = z, y)$	uncertainties caused by the unmodeled dynamics
$\eta(t)$	unmodeled dynamics

$y_d(t)$	desired tracking signal
$x_{2c}(t)$	inner loop virtual signal
$e_1(t), e_2(t)$	tracking errors
k_0, k_1, k_2	positive control gains
$r(t)$	dynamic auxiliary signal
Δf	system unknown nonlinear term
W_a, W_c	weights of actor NN and critic NN
$\Phi_a(Z_a), \Phi_c(Z_c)$	activation functions of actor NN and critic NN
$\varepsilon_{W_a}, \varepsilon_{W_c}$	estimation errors of actor NN and critic NN
ε_v	unknown upper bound of the total disturbance $D(t)$
$u(t)$	designed controller
$J(t)$	integral penalty function
$e_c(t)$	error variable of critic NN
$E_c(t)$	error objective function of critic NN
$\hat{\cdot}$	estimation value of \cdot
$\tilde{\cdot}$	estimation error of \cdot and $\tilde{\tau} = \hat{\tau} - \tau$

1.0 Introduction

The missiles are a class of weapons that are equipped with guidance and control equipment to achieve precision flight and strike missions. The typical features of long range, high accuracy, great power, and strong defense penetration capabilities make the missiles an important research area. Recently, the high-precision control problem of missiles has become a relatively important research topic and has been extensively studied by scholars at home and abroad. In order to tackle this problem, several approaches have been proposed [1]. For example, a robust control scheme based on the quaternion feedback is proposed for the attitude control problem of the missile [2]. In Ref. [3], the robust control method based on disturbance observer is proposed, which not only ensures the robustness of the nonlinear system, but also solves the problem of mismatched disturbance by using the observer, and has been successfully applied to the nonlinear missile systems with various uncertain relations and external disturbances. In addition to the robust control method, sliding mode control is also favoured by scholars because of its unique advantages. Considering the condition that the attitude of the missile is affected by the rapid and large parameter variations and the partial instability in the boost phase, a multi-sliding surface attitude controller based on high-order sliding mode and traditional sliding mode is proposed to control the attitude of the missile [4]. Although sliding mode control theory has many advantages, it is prone to chattering during the design process, which will cause harm to the system. In light of this situation, two novel smooth sliding mode control methods were proposed, and successfully achieved the fast finite time convergence of the system [5]. Backstepping decomposes a complex nonlinear system into several subsystems, which has the ability to deal with mismatching uncertainties. Introducing the backstepping method into the missile control system makes the system more flexible and robust. By utilising the backstepping method, the guidance and control law is divided into a guidance loop and a control loop for design, and the state observer is used for online estimation and compensation of the aerodynamic parameter changes in the model [6]. The above-mentioned literatures adopt the single control method, compared with this, the compound control method can more effectively improve the performance of the system. Therefore, by a combination of backstepping and sliding mode control methods, the attitude controller is devised for a rotating missile with two moving masses inside [7]. Moreover, in order to depress the peaking phenomenon and chasing of backstepping sliding mode controller, filtering technology is introduced [8].

However, traditional aerodynamically controlled missiles have the problem of long overload response time. At present, by increasing the direct lateral force to form the aerodynamic/reaction-jet compound control system, the dynamic performance of the missile control system can be improved. In Ref. [9], a robust controller of the aerodynamic/reaction-jet compound control missile is designed, and in order to ensure the robustness when the jet factor changes, the parameter space method is used to design

the equivalent steering gear system. Moreover, by a combination of robust trail tracking control and dynamic inverse control, a blended robust control method is devised to deal with the blended attitude control with lateral thrust and aerodynamic force [10]. To realise the coordinated use of direct force and aerodynamic force and fast and accurate tracking of overload, a compound control strategy based on fixed time convergence sliding mode control theory and dynamic control allocation technology is proposed [11]. For the case of parameter perturbation and large external disturbances, a nonsingular fast terminal sliding mode control method is proposed, which improves the convergence speed of the system [12]. In Refs. [13, 14], backstepping method is used to design virtual control law to complete the control problem of compound missile.

Above controllers are designed based on modern control theory, nevertheless, pure modern control theory cannot solve many problems of missile engineering application. With the development of artificial intelligence technology, many scholars try to combine artificial intelligence and control theory to solve the problem of missile control. The fuzzy control theory is introduced into the control design of compound missile, and the overload command is tracked by designing fuzzy controller [15]. Moreover, the combination of artificial intelligence integral controller and fuzzy controller is helpful to improve the stability of missile guidance system [16]. In Ref. [17], variable universe fuzzy control is introduced to solve the influence of aerodynamic parameters on missile control system. The typical features of parallel processing, distributed storage, high fault tolerance and nonlinear operation make artificial neural network particularly popular. The neural network reference model method is used to design the equivalent steering gear of composite missile [18]. Genetic algorithm has great advantages in optimising the control system. For instance, in Ref. [19], the gain matrix of missile controller is optimised by genetic algorithm, and simulation results show that the optimisation effect of this method cannot be achieved by traditional optimisation methods. With the rapid development of computer technology, adaptive technology has attracted more and more attention from scholars. The combination of adaptive control and intelligent control algorithm is an important direction of missile control research. In Ref. [20], fuzzy adaptive proportional–integral–derivative (PID) control is designed to solve the problem that PID controller cannot adjust parameters. Based on the robust adaptive controller, a fuzzy adaptive disturbance observer is used to compensate the disturbances in the linear velocity and angular velocity dynamics of the missile [21]. The control performance of hypersonic missile in cruise phase is improved by the combination of fuzzy control and adaptive sliding mode control [22]. In Ref. [23], a robust adaptive neural network state feedback control based on backstepping is proposed for missile systems with unknown parameters and unknown delay inputs, and an approximator based on neural network is used to compensate the uncertainty caused by unknown delay. For the missile with random disturbances and non-affine aerodynamic characteristics, the neural network is used to deal with the non-affine aerodynamic characteristics in the system, and the adaptive term is used to solve the problem of unknown target manoeuvre [24]. An improved adaptive genetic algorithm is proposed to solve the nonlinear integer programming model of large-scale missile firepower allocation. Compared with the traditional genetic algorithm, the crossover probability and mutation probability automatically adjusted by the adaptive rule significantly improve the search ability of the algorithm, so as to improve the accuracy of the model [25]. In view of the large jet interference of missile with lateral jets and aerodynamic surfaces, the control allocation algorithm is designed by using adaptive genetic algorithm to meet the real-time requirements of the algorithm; then the variable universe adaptive fuzzy control is used to design the ignition algorithm of attitude control engine, which overcomes the influence of jet interference and solves the problem of low precision of conventional fuzzy control [26].

Motivated by the above discussions, a STDO-based adaptive reinforcement learning control method is proposed for the straight air compound missile system with aerodynamic uncertainties and unmodeled dynamics. The main contributions of this paper can be summarised as follows.

- To deal with the tracking problems for the straight gas compound system, adaptive control with actor-critic design is investigated in this paper: the critic part is used to obtain the cost function to evaluate the tracking performance, and the actor part generates the control policy of the actuator according to the results from the critic part.

- To improve the control performance, reinforcement learning and neural networks are adopted in the actor-critic design: the critic neural network and the actor neural network are utilised to approximate the cost function and cope with the unmodeled dynamics, respectively.
- Considering that the negative impacts of the control signal fluctuation caused by the disturbances of the straight gas compound system, the STDO disturbance observer is used to solve the problem.

2.0 Problem formulation and preliminaries

2.1 Problem statement

Ignoring the roll channel, the attitude dynamic model of a direct force and aerodynamic force compound missile can be modeled as

$$\begin{aligned}
 \dot{\alpha} &= \omega_z - \frac{QS(C_y^\alpha \alpha + C_y^{\delta_z} \delta_z)}{mV} - \frac{F_y}{mV} \\
 \dot{\beta} &= \omega_y + \frac{QS(C_z^\beta \beta + C_z^{\delta_y} \delta_y)}{mV} + \frac{F_z}{mV} \\
 \dot{\omega}_z &= \frac{QSL}{J_z}(m_z^\alpha \alpha + m_z^{\delta_z} \delta_z + m_z^{\bar{\omega}_z} \bar{\omega}_z) + \frac{l_z}{J_z} F_y + d_{\omega_z} + \chi_z(\omega_z, \eta) \\
 \dot{\omega}_y &= \frac{QSL}{J_y}(m_y^\beta \beta + m_y^{\delta_y} \delta_y + m_y^{\bar{\omega}_y} \bar{\omega}_y) + \frac{l_y}{J_y} F_z + d_{\omega_y} + \chi_y(\omega_y, \eta)
 \end{aligned} \tag{1}$$

where α is the angle-of-attack, β is the slide angle, ω_z, ω_y are the angle velocities. J_z, J_y denote the rotational inertia. V and m are the velocity and the mass of the missile, respectively. S and L represent the reference area and length. δ_z, δ_y are the output signals of the elevator and rudder. $C_{(\cdot)}^{(\cdot)}$ denote the coefficients of the aerodynamic forces, while $m_{(\cdot)}^{(\cdot)}$ represent the coefficients of the aerodynamic moments. $d_i (i = \omega_z, \omega_y)$ are the unknown disturbances. η is the unmodeled dynamics, and $\chi_i (i = z, y)$ represents the uncertainties caused by the unmodeled dynamics.

Then, by defining

$$\begin{aligned}
 x_1(t) &= [\alpha \ \beta]^T, x_2(t) = [\omega_z \ \omega_y]^T, \\
 d_1(t) &= \begin{bmatrix} -\frac{QS(C_y^\alpha \alpha + C_y^{\delta_z} \delta_z)}{mV} - \frac{F_y}{mV} \\ \frac{QS(C_z^\beta \beta + C_z^{\delta_y} \delta_y)}{mV} + \frac{F_z}{mV} \end{bmatrix}, d_2(t) = \begin{bmatrix} \frac{QSL}{J_z}(m_z^\alpha \alpha + m_z^{\bar{\omega}_z} \bar{\omega}_z) + d_{\omega_z} \\ \frac{QSL}{J_y}(m_y^\beta \beta + m_y^{\bar{\omega}_y} \bar{\omega}_y) + d_{\omega_y} \end{bmatrix} \\
 B &= \begin{bmatrix} \frac{QSL}{J_z} m_z^{\delta_z} & 0 & \frac{l_z}{J_z} & 0 \\ 0 & \frac{QSL}{J_y} m_y^{\delta_y} & 0 & \frac{l_y}{J_y} \end{bmatrix}, u(t) = [\delta_z \ \delta_y \ F_y \ F_z]^T
 \end{aligned} \tag{2}$$

the equivalent model of Equation (1) can be given as

$$\begin{aligned}
 \dot{x}_1(t) &= x_2(t) + d_1(t) \\
 \dot{x}_2(t) &= Bu(t) + d_2(t) + \chi(x_2(t), \eta(t))
 \end{aligned} \tag{3}$$

Thus, the design objective is to develop a reinforcement learning-based STDO control scheme to maintain the desired trajectory tracking for the straight air compound missile system given in Equation (3) subjected to aerodynamic uncertainties and unmodeled dynamics.

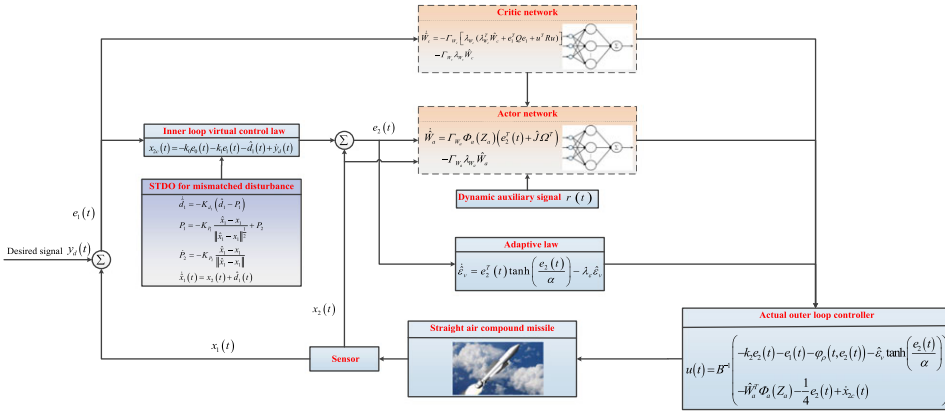


Figure 1. The structure of the proposed reinforcement learning-based STDO control algorithm.

2.2 Assumptions and lemmas

The following assumptions and lemmas are necessary.

Assumption 1. The disturbance moments caused by structural uncertainty are bounded, that is, there exists the constants \bar{d}_1, \bar{d}_2 such that $\|d_1(t)\| \leq \bar{d}_1, \|d_2(t)\| \leq \bar{d}_2$.

Assumption 2. The desired tracking signal of the system $y_d(t)$ is smooth and twice differentiable.

Lemma 1. For any constant $\varepsilon > 0$ and vector $\xi \in R^n$, we have

$$\|\xi\| < \frac{\xi^T \xi}{\sqrt{\xi^T \xi} + \varepsilon^2} + \varepsilon \tag{4}$$

Lemma 2. [27] Given any constant $\varepsilon > 0$ and any variable $z \in R$, the following inequality holds

$$0 \leq |z| - z \tanh\left(\frac{z}{\varepsilon}\right) \leq \kappa \varepsilon \tag{5}$$

where κ is a constant satisfying $\kappa = e^{-(\kappa+1)}$, i.e. $\kappa = 0.2785$.

3.0 Stdo-based adaptive reinforcement learning control

In this section, as shown in Fig. 1, a STDO-based adaptive reinforcement learning control method is proposed.

Defining the control expected output signal as $y_d(t)$ and the inner loop virtual signal as $x_{2c}(t)$, then the tracking errors of $x_1(t)$ and $x_2(t)$ can be expressed as

$$\begin{aligned} e_1(t) &= x_1(t) - y_d(t) \\ e_2(t) &= x_2(t) - x_{2c}(t) \end{aligned} \tag{6}$$

Combining with Equation (3), one has

$$\begin{aligned} \dot{e}_1(t) &= x_{2c}(t) + e_2(t) + d_1(t) - \dot{y}_d(t) \\ \dot{e}_2(t) &= Bu(t) + d_2(t) + \chi(x_2(t), \eta(t)) - \dot{x}_{2c}(t) \end{aligned} \tag{7}$$

Then, we can design the inner loop virtual signal as

$$x_{2c}(t) = -k_0 \int_0^t e_1(\tau) d\tau - k_1 e_1(t) - \hat{d}_1(t) + \dot{y}_d(t) \tag{8}$$

where $\hat{d}_1(t)$ is the adaptive estimate of $d_1(t)$, k_0 and k_1 are the control gains.

In order to compensate and suppress the influence of the unknown mismatched disturbance $d_1(t)$, a second-order STDO is designed as follows

$$\begin{aligned} \dot{\hat{d}}_1(t) &= -K_{d_1}(\hat{d}_1(t) - P_1) \\ P_1 &= -K_{P_1} \frac{\hat{x}_1 - x_1}{\|\hat{x}_1 - x_1\|^{\frac{1}{2}}} + P_2 \\ \dot{P}_2 &= -K_{P_2} \frac{\hat{x}_1 - x_1}{\|\hat{x}_1 - x_1\|} \\ \hat{\dot{x}}_1(t) &= x_2(t) + \hat{d}_1(t) \end{aligned} \tag{9}$$

The dynamic signal $r(t)$ is introduced, which is defined by

$$\dot{r}(t) = -\gamma_0 r(t) + \rho(x_1(t), x_2(t)), r(0) = r_0 \tag{10}$$

where $\gamma_0 \in (0, \gamma_1)$.

The coupling uncertainty is assumed to satisfy the following inequality

$$e_2^T \chi(x_2(t), \eta(t)) \leq \|e_2^T(t)\| (\varphi_1(x_2(t)) + \varphi_2(\eta(t))) \tag{11}$$

According to Lemma 1 and Young’s inequality, Equation (11) can be rewritten as

$$\begin{aligned} \|e_2^T(t)\| \varphi_1(x_2(t)) &\leq e_2^T(t) \bar{\varphi}_1(e_2(t), x_2(t)) + \varepsilon_1 \\ \|e_2^T(t)\| \varphi_2(\eta(t)) &\leq e_2^T(t) \bar{\varphi}_2(e_2(t), r(t)) + \varepsilon_2 + \frac{1}{4} e_2^T(t) e_2(t) + \varepsilon_3 \end{aligned} \tag{12}$$

where $\varepsilon_1, \varepsilon_2 > 0$ are arbitrary constants,

$$\begin{aligned} \bar{\varphi}_1(e_2(t), x_2(t)) &= \frac{\varphi_1(x_2(t)) e_2^T(t) \varphi_1(x_2(t))}{\sqrt{[e_2^T(t) \varphi_1(x_2(t))]^2 + \varepsilon_1^2}} \\ \bar{\varphi}_2(e_2(t), r(t)) &= \frac{\varphi_2 \circ \alpha_1^{-1}(2r(t)) e_2^T(t) \varphi_2 \circ \alpha_1^{-1}(2r(t))}{\sqrt{[e_2^T(t) \varphi_2 \circ \alpha_1^{-1}(2r(t))]^2 + \varepsilon_2^2}} \\ \varepsilon_3 &= [\varphi_2 \circ \alpha_1^{-1}(2\varepsilon_r)]^2 \end{aligned} \tag{13}$$

Next, we define

$$\Delta f = \bar{\varphi}_1(e_2(t), x_2(t)) + \bar{\varphi}_2(e_2(t), r(t)) \tag{14}$$

Since $\bar{\varphi}_1(e_2(t), x_2(t))$ and $\bar{\varphi}_2(e_2(t), r(t))$ are the functions that change irregularly in the control dynamic process, the actor NNs are introduced to approximate the unknown nonlinear term Δf , the actor NN structures of the optimal control Δf and the actual control $\hat{\Delta f}$ are designed as follows

$$\begin{aligned} \Delta f &= W_a^T \Phi_a(Z_a) + \varepsilon_{W_a} \\ \hat{\Delta f} &= \hat{W}_a^T \Phi_a(Z_a) \end{aligned} \tag{15}$$

where $W_a, \hat{W}_a \in \mathbb{R}^{p_1 \times n}$, $\Phi_a(Z_a) \in \mathbb{R}^{p_1 \times 1}$, $\Phi_a(Z_a) = e^{-\frac{(Z_a - \mu)^2}{2\sigma^2}}$, $Z_a = [e_2(t), x_2(t), r(t)]^T$, and there exists an upper bound of the estimation error such that $\|\varepsilon_{W_a}\| \leq \bar{\varepsilon}_{W_a}$. Thus, we can obtain that

$$e_2^T(t) \chi(x_2(t), \eta(t)) \leq e_2^T W_a^T \Phi_a(Z_a) + e_2^T \varepsilon_{W_a} + \frac{1}{4} e_2^T(t) e_2(t) + \sum_{i=1}^3 \varepsilon_i \tag{16}$$

Then, the matched disturbance $d_2(t)$ and actor NN estimation error ε_{W_a} of the controlled system need to be considered and compensated. Firstly, the total disturbance $D(t)$ can be constructed in the

following form:

$$D(t) = d_2(t) + \varepsilon_{w_a} \tag{17}$$

Thanks to Assumption 1, the following inequality is satisfied

$$|D(t)| = |d_2(t) + \varepsilon_{w_a}| \leq \varepsilon_v \tag{18}$$

where ε_v is an unknow positive constant. According to Lemma 2, we can easily obtain

$$e_2^T(t) D(t) \leq |e_2(t)| \varepsilon_v \leq \varepsilon_v e_2^T(t) \tanh\left(\frac{e_2(t)}{\alpha}\right) + \kappa \alpha \varepsilon_v \tag{19}$$

Based on the above analysis, the outer loop controller can be designed as

$$u(t) = B^{-1} \begin{pmatrix} -k_2 e_2(t) - e_1(t) - \varphi_p(t, e_2(t)) - \hat{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) \\ -\hat{W}_a^T \Phi_a(Z_a) - \frac{1}{4} e_2(t) + \dot{x}_{2c}(t) \end{pmatrix} \tag{20}$$

where k_2 is the control gain.

The critic NN which can be used to appraise control performance and make feedback to the actor NN will be introduced in detail. Firstly, we define the integral penalty function of the controlled system as follows

$$J(t) = \int_0^\infty [e_1^T(\tau) Q e_1(\tau) + u^T(\tau) R u(\tau)] d\tau \tag{21}$$

Then, we approximate the penalty function $J(t)$ by designing the critic NN

$$\hat{J}(t) = \hat{W}_c^T \Phi_c(Z_c) \tag{22}$$

where $\hat{W}_c \in \mathbb{R}^{p_2 \times n}$, $\Phi_c(Z_c) \in \mathbb{R}^{p_2 \times 1}$, $\Phi_c(Z_c) = e^{-\frac{(Z_c - \mu)^2}{2\sigma^2}}$, $Z_c = e_1(t)$.

Constructing the residual mean square error function of the critic NN structure, one has

$$e_c(t) = e_1^T(t) Q e_1(t) + u^T(t) R u(t) + \hat{W}_c^T \nabla \Phi_c \dot{x}_1(t) \tag{23}$$

$$E_c(t) = \frac{1}{2} e_c^T(t) e_c(t)$$

where $\nabla \Phi_c = \partial \Phi_c(x_1) / \partial x_1$ and $\nabla \Phi_c \in \mathbb{R}^{p_2 \times n}$. The update goal of the weight of the critic NN is to minimise $E_c(t)$, thus the update rate of the critic network weight is obtained according to the gradient descent method

$$\begin{aligned} \dot{\hat{W}}_c &= -\Gamma_{w_c} \lambda_{w_c} e_c - \Gamma_{w_c} \lambda_{w_c} \hat{W}_c \\ &= -\Gamma_{w_c} \left[\lambda_{w_c} \left(\lambda_{w_c}^T \hat{W}_c + e_1^T(t) Q e_1(t) + u^T(t) R u(t) \right) \right] - \Gamma_{w_c} \lambda_{w_c} \hat{W}_c \end{aligned} \tag{24}$$

where $\lambda_{w_c} = \nabla \Phi_c \dot{x}(t)$, Γ_{w_c} , $\lambda_{w_c} > 0$.

Finally, the adaptive laws of \hat{W}_a , \hat{W}_c , $\hat{\varepsilon}_v$ are listed as follows

$$\begin{aligned} \dot{\hat{W}}_a &= \Gamma_{w_a} \Phi_a(Z_a) \left(e_2^T(t) + \hat{J} \Omega^T \right) - \Gamma_{w_a} \lambda_{w_a} \hat{W}_a \\ \dot{\hat{W}}_c &= -\Gamma_{w_c} \left[\lambda_{w_c} \left(\lambda_{w_c}^T \hat{W}_c + e_1^T Q e_1 + u^T R u \right) \right] - \Gamma_{w_c} \lambda_{w_c} \hat{W}_c \\ \dot{\hat{\varepsilon}}_v &= e_2^T(t) \tanh\left(\frac{e_2(t)}{\alpha}\right) - \lambda_\varepsilon \hat{\varepsilon}_v \end{aligned} \tag{25}$$

For the sake of analysis, we define $\tilde{*} = \hat{*} - *$ to represent the estimation error of the unknown variable $*$.

4.0 Stability analysis

Defining $e_0(t) = \int_0^t e_1(s) ds$ and combining Equations (7), (8) and (20), then the closed relation can be obtained

$$\begin{aligned} \dot{e}_0(t) &= e_1(t) \\ \dot{e}_1(t) &= -k_0 e_0(t) - k_1 e_1(t) + e_2(t) - \tilde{d}_1(t) \\ \dot{e}_2(t) &= -k_2 e_2(t) - e_1(t) + d_2(t) + \chi(x_2(t), \eta(t)) \\ &\quad - \hat{W}_a^T \Phi_a(Z_a) - \hat{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) - \frac{1}{4} e_2(t) - \varphi_\rho(t, e_2(t)) \end{aligned} \tag{26}$$

The purpose of this paper is to construct an efficient controller to ensure the stability of the closed relation described in Equation (26). The stability of the straight air compound missile system with the proposed control scheme can be revealed by the following theorem.

Theorem 1. Consider the straight air compound missile system described in Equation (3). Suppose Assumption 1 and Assumption 2 can be satisfied. If the inner loop control law and the outer loop control law are given by Equations (8) and (20), the STDO is designed as Equation (9), the adaptive laws are designed as Equation (25), then the closed-loop control system in the existence of unmodeled dynamics is stable and all the signals are upper bounded.

Proof. The Lyapunov function V is selected as

$$\begin{aligned} V &= V_1 + V_2 \\ V_1 &= \frac{1}{2} e_0^T(t) e_0(t) + \frac{1}{2} e_1^T(t) e_1(t) \\ V_2 &= \frac{1}{2} e_2^T(t) e_2(t) + \frac{1}{2} \text{Tr}(\tilde{W}_a^T \Gamma_{w_a}^{-1} \tilde{W}_a) + \frac{1}{2} \text{Tr}(\tilde{W}_c^T \Gamma_{w_c}^{-1} \tilde{W}_c) + \frac{1}{2} \tilde{\varepsilon}_v^T \tilde{\varepsilon}_v + \frac{r(t)}{\Gamma_r} + J^*(x_1) \end{aligned} \tag{27}$$

Taking the derivative of both sides of the Equation (27), we can get that

$$\begin{aligned} \dot{V} &= \dot{V}_1 + \dot{V}_2 \\ \dot{V}_1 &= e_0^T(t) \dot{e}_0(t) + e_1^T(t) \dot{e}_1(t) \\ \dot{V}_2 &= e_2^T(t) \dot{e}_2(t) + \text{Tr}(\tilde{W}_a^T \Gamma_{w_a}^{-1} \dot{\tilde{W}}_a) + \text{Tr}(\tilde{W}_c^T \Gamma_{w_c}^{-1} \dot{\tilde{W}}_c) + \tilde{\varepsilon}_v^T \dot{\tilde{\varepsilon}}_v - \frac{\gamma_0}{\Gamma_r} r(t) + \frac{\rho(t)}{\Gamma_r} + J_x^{*T} \dot{x}_1 \end{aligned} \tag{28}$$

Substituting Equation (26) into \dot{V}_1 term of Equation (28), one has

$$\dot{V}_1 = e_0^T(t) e_1(t) - k_0 e_0^T(t) e_0(t) - k_1 e_1^T(t) e_1(t) + e_1^T(t) e_2(t) - e_1^T(t) \tilde{d}_1(t) \tag{29}$$

Defining $\bar{e}_1 = [e_0^T(t), e_1^T(t)]^T$ and utilising the following inequality

$$\begin{aligned} e_1^T(t) \tilde{d}_1(t) &\leq \frac{1}{2} e_1^T(t) e_1(t) + \frac{1}{2} \tilde{d}_1^T(t) \tilde{d}_1(t) \\ &\leq \frac{1}{2} e_1^T(t) e_1(t) + \frac{1}{2} \varepsilon_d^2 \end{aligned} \tag{30}$$

Equation (29) can be rewritten as

$$\begin{aligned} \dot{V}_1 &\leq -\bar{e}_1^T(t) A \bar{e}_1(t) + e_1^T(t) e_2(t) + \frac{1}{2} \varepsilon_d^2 \\ A &= \begin{bmatrix} 0 & -1 \\ k_0 & -\frac{1}{2} + k_1 \end{bmatrix} \end{aligned} \tag{31}$$

where we assume that $\|\tilde{d}_1(t)\| \leq \varepsilon_d$ holds and ε_d is a positive constant.

Thanks to Equation (16) and $\dot{e}_2(t)$ term of Equation (26), we can get the following inequality

$$\begin{aligned}
 e_2^T(t) \dot{e}_2(t) &\leq -k_2 e_2^T(t) e_2(t) - e_2^T(t) e_1(t) - e_2^T(t) \varphi_\rho(t, e_2(t)) - e_2^T(t) \hat{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) \\
 &\quad - e_2^T(t) \tilde{W}_a^T \Phi_a(Z_a) + e_2^T(t) (d_2(t) + \varepsilon_{w_a}) + \sum_{i=1}^3 \varepsilon_i
 \end{aligned} \tag{32}$$

where

$$\begin{aligned}
 &e_2^T(t) (d_2(t) + \varepsilon_{w_a}) - e_2^T(t) \hat{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) \\
 &= e_2^T(t) (D(t)) - e_2^T(t) \hat{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) \\
 &\leq e_2^T(t) \varepsilon_v \tanh\left(\frac{e_2(t)}{\alpha}\right) + \kappa \alpha \varepsilon_v - e_2^T(t) \hat{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) \\
 &= -e_2^T(t) \tilde{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) + \kappa \alpha \varepsilon_v
 \end{aligned} \tag{33}$$

Thus, the following inequality can be readily obtained

$$\begin{aligned}
 e_2^T(t) \dot{e}_2(t) &\leq -k_2 e_2^T(t) e_2(t) - e_2^T(t) e_1(t) - e_2^T(t) \varphi_\rho(t, e_2(t)) \\
 &\quad - e_2^T(t) \tilde{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) - e_2^T(t) \tilde{W}_a^T \Phi_a(Z_a) + \sum_{i=1}^3 \varepsilon_i + \kappa \alpha \varepsilon_v
 \end{aligned} \tag{34}$$

Substituting Equation (34) into \dot{V}_2 term of Equation (28), one has

$$\begin{aligned}
 \dot{V}_2 &\leq -k_2 e_2^T(t) e_2(t) - e_2^T(t) e_1(t) - e_2^T(t) \varphi_\rho(t, e_2(t)) - e_2^T(t) \tilde{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) - e_2^T(t) \tilde{W}_a^T \Phi_a(Z_a) \\
 &\quad + \text{Tr}\left(\tilde{W}_a^T \Gamma_{w_a}^{-1} \dot{\tilde{W}}_a\right) + \text{Tr}\left(\tilde{W}_c^T \Gamma_{w_c}^{-1} \dot{\tilde{W}}_c\right) + \tilde{\varepsilon}_v^T \dot{\tilde{\varepsilon}}_v - \frac{\gamma_0}{\Gamma_r} r(t) + \frac{\rho(t)}{\Gamma_r} + J_x^{*T} \dot{x}_1 + \sum_{i=1}^3 \varepsilon_i + \kappa \alpha \varepsilon_v
 \end{aligned} \tag{35}$$

For any vector $\xi \in \mathbb{R}^n$, we define

$$\text{Tanh}(\xi(t)) = [\tanh \xi_1(t), \tanh \xi_2(t), \dots, \tanh \xi_n(t)]^T \tag{36}$$

Therefore, the following formula holds

$$\begin{aligned}
 \frac{\rho(t)}{\Gamma_r} &= \frac{\rho(t)}{\Gamma_r} \left(1 - 16 \text{Tanh}^T\left(\frac{e_2(t)}{\varepsilon_\rho}\right) \text{Tanh}\left(\frac{e_2(t)}{\varepsilon_\rho}\right)\right) + e_2^T(t) \varphi_\rho(t, e_2(t)) \\
 \varphi_\rho(t, e_2(t)) &= \frac{16 e_2(t) \rho(t)}{\Gamma_r e_2^T(t) e_2(t)} \text{Tanh}^T\left(\frac{e_2(t)}{\varepsilon_\rho}\right) \text{Tanh}\left(\frac{e_2(t)}{\varepsilon_\rho}\right)
 \end{aligned} \tag{37}$$

Then, combining Equations (31), (35)–(37), we have

$$\begin{aligned}
 \dot{V} &= \dot{V}_1 + \dot{V}_2 \\
 &\leq -\bar{e}_1^T(t) A \bar{e}_1(t) - k_2 e_2^T(t) e_2(t) - e_2^T(t) \tilde{\varepsilon}_v \tanh\left(\frac{e_2(t)}{\alpha}\right) - e_2^T \tilde{W}_a^T \Phi_a(Z_a)
 \end{aligned}$$

$$\begin{aligned}
 & + \text{Tr}\left(\tilde{W}_a^T \Gamma_{W_a}^{-1} \dot{\tilde{W}}_a\right) + \text{Tr}\left(\tilde{W}_c^T \Gamma_{W_c}^{-1} \dot{\tilde{W}}_c\right) + \tilde{\varepsilon}_v^T \dot{\tilde{\varepsilon}}_v - \frac{\gamma_0}{\Gamma_r} r(t) + J_x^{*T} \dot{x}_1 \\
 & + \rho(t) \left(1 - 16 \text{Tanh}^T\left(\frac{e_2(t)}{\varepsilon_\rho}\right) \text{Tanh}\left(\frac{e_2(t)}{\varepsilon_\rho}\right)\right) / \Gamma_r + \frac{1}{2} \varepsilon_d^2 + \sum_{i=1}^3 \varepsilon_i + \kappa \alpha \varepsilon_v
 \end{aligned} \tag{38}$$

By using the adaptive laws in Equation (25) and considering the following inequalities

$$-\tilde{\varepsilon}_v^T \dot{\tilde{\varepsilon}}_v \leq -\frac{1}{2} \dot{\tilde{\varepsilon}}_v^2 + \frac{1}{2} \varepsilon_v^2 \tag{39}$$

then, it can be concluded that

$$\begin{aligned}
 \dot{V} & \leq -\bar{e}_1^T(t) A \bar{e}_1(t) - k_2 e_2^T(t) e_2(t) - \frac{\lambda_\varepsilon}{2} \tilde{\varepsilon}_v^2 + \text{Tr}\left(\tilde{W}_a^T \left(\Phi_a \hat{J} \Omega^T - \lambda_{W_a} \hat{W}_a\right)\right) \\
 & + \text{Tr}\left(\tilde{W}_c^T \left(-\lambda_{W_c} \left(\lambda_{W_c}^T \hat{W}_c + \varepsilon_c\right) - \lambda_{W_c} \hat{W}_c\right)\right) + J_x^{*T} \dot{x}_1 - \frac{\gamma_0}{\Gamma_r} r(t) + \frac{1}{2} \varepsilon_d^2 \\
 & + \sum_{i=1}^3 \varepsilon_i + \frac{\lambda_\varepsilon}{2} \varepsilon_v^2 + \rho(t) \left(1 - 16 \text{Tanh}^T\left(\frac{e_2(t)}{\varepsilon_\rho}\right) \text{Tanh}\left(\frac{e_2(t)}{\varepsilon_\rho}\right)\right) / \Gamma_r
 \end{aligned} \tag{40}$$

Considering $\text{Tr}\left(\tilde{W}_a^T \left(\Phi_a \hat{J} \Omega^T - \lambda_{W_a} \hat{W}_a\right)\right)$ term of Equation (40), we can obtain that

$$\begin{aligned}
 & \text{Tr}\left(\tilde{W}_a^T \left(\Phi_a \hat{J} \Omega^T - \lambda_{W_a} \hat{W}_a\right)\right) \\
 & = \text{Tr}\left(\tilde{W}_a^T \Phi_a \tilde{W}_c^T \Phi_c \Omega^T\right) + \text{Tr}\left(\tilde{W}_a^T \Phi_a W_c^T \Phi_c \Omega^T\right) - \text{Tr}\left(\tilde{W}_a^T \lambda_{W_a} \hat{W}_a\right) \\
 & = \tilde{W}_c^T \Phi_c \Omega^T \tilde{W}_a^T \Phi_a + W_c^T \Phi_c \Omega^T \tilde{W}_a^T \Phi_a - \lambda_{W_a} \text{Tr}\left(\tilde{W}_a^T \hat{W}_a\right) \\
 & \leq \rho_1 \tilde{W}_c^T \tilde{W}_c + \frac{\lambda_{\max}\left(\Phi_c \Omega^T \Omega \Phi_c^T\right) \bar{\Phi}_a^2}{4 \rho_1} \text{Tr}\left(\tilde{W}_a^T \tilde{W}_a\right) \\
 & + \rho_2 W_c^T W_c + \frac{\lambda_{\max}\left(\Phi_c \Omega^T \Omega \Phi_c^T\right) \bar{\Phi}_a^2}{4 \rho_2} \text{Tr}\left(\tilde{W}_a^T \tilde{W}_a\right) - \frac{\lambda_{W_a}}{2} \tilde{W}_a^T \tilde{W}_a + \frac{\lambda_{W_a}}{2} W_a^T W_a
 \end{aligned} \tag{41}$$

Then, considering $\text{Tr}\left(\tilde{W}_c^T \left(-\lambda_{W_c} \left(\lambda_{W_c}^T \hat{W}_c + \varepsilon_c\right) - \lambda_{W_c} \hat{W}_c\right)\right)$ term of Equations (40), the following inequality holds

$$\begin{aligned}
 & \text{Tr}\left(\tilde{W}_c^T \left(-\lambda_{W_c} \left(\lambda_{W_c}^T \hat{W}_c + \varepsilon_c\right) - \lambda_{W_c} \hat{W}_c\right)\right) \\
 & = \text{Tr}\left(-\tilde{W}_c^T \lambda_{W_c} \lambda_{W_c}^T \hat{W}_c - \tilde{W}_c^T \lambda_{W_c} \varepsilon_c - \tilde{W}_c^T \lambda_{W_c} \hat{W}_c\right) \\
 & = -\tilde{W}_c^T \lambda_{W_c} \lambda_{W_c}^T \tilde{W}_c - \tilde{W}_c^T \lambda_{W_c} \varepsilon_c - \tilde{W}_c^T \lambda_{W_c} \hat{W}_c \\
 & \leq \left(\rho_3 + \frac{\lambda_{\max}\left(\bar{\lambda}_{W_c}\right)}{4 \rho_3}\right) \tilde{W}_c^T \tilde{W}_c + \rho_4 \lambda_{\max}\left(\lambda_{W_c} \lambda_{W_c}^T\right) \tilde{W}_c^T \tilde{W}_c + \frac{1}{4 \rho_4} \varepsilon_c^2 - \frac{\lambda_c}{2} \tilde{W}_c^T \tilde{W}_c + \frac{\lambda_c}{2} W_c^T W_c
 \end{aligned} \tag{42}$$

Moreover, the $J_x^{*T} \dot{x}_1$ term of Equations (28) satisfies

$$J_x^{*T} \dot{x}_1 \leq -\lambda_{\min}\{Q\} \|e_1\|^2 - \lambda_{\min}\{R\} \|u\|^2 \tag{43}$$

Substituting Equations (41)–(43) into Equation (40), we can get that

$$\begin{aligned} \dot{V} \leq & -\tilde{e}_1^T(t) A \tilde{e}_1(t) - k_2 e_2^T(t) e_2(t) - \frac{\lambda_\varepsilon \tilde{\varepsilon}_v^2}{2} - \frac{\gamma_0}{\Gamma_r} r(t) + \rho(t) \left(1 - 16 \text{Tanh}^T \left(\frac{e_2(t)}{\varepsilon_\rho} \right) \text{Tanh} \left(\frac{e_2(t)}{\varepsilon_\rho} \right) \right) / \Gamma_r \\ & - \left(\frac{\lambda_{W_a}}{2} - \frac{\lambda_{\max}(\Phi_c \Omega^T \Omega \Phi_c^T) \bar{\Phi}_a^2}{4\rho_1} - \frac{\lambda_{\max}(\Phi_c \Omega^T \Omega \Phi_c^T) \bar{\Phi}_a^2}{4\rho_2} \right) \text{Tr}(\tilde{W}_a^T \tilde{W}_a) \\ & - \left(\frac{\lambda_{W_c}}{2} - \rho_1 - \rho_3 - \frac{\lambda_{\max}(\tilde{\lambda}_{W_c})}{4\rho_3} - \rho_4 \lambda_{\max}(\lambda_{W_c} \lambda_{W_c}^T) \right) \tilde{W}_c^T \tilde{W}_c - \lambda_{\min}\{Q\} \|e_1\|^2 - \lambda_{\min}\{R\} \|u\|^2 \\ & + \frac{\lambda_\varepsilon}{2} \varepsilon_v^2 + \frac{1}{2} \varepsilon_d^2 + \sum_{i=1}^3 \varepsilon_i + \frac{\lambda_{W_a}}{2} W_a^T W_a + \frac{2\rho_2 + \lambda_{W_c}}{2} W_c^T W_c + \frac{1}{4\rho_4} \varepsilon_c^2 \end{aligned} \tag{44}$$

Defining

$$\begin{aligned} \gamma = \min & \left\{ \begin{array}{l} 2\lambda_{\min}(A), 2k_2, 2\lambda_{\min}(Q), 2\lambda_{\min}(R), \lambda_\varepsilon, \\ \frac{\lambda_{W_a}}{2} - \frac{\lambda_{\max}(\Phi_c \Omega^T \Omega \Phi_c^T) \bar{\Phi}_a^2}{4\rho_1} - \frac{\lambda_{\max}(\Phi_c \Omega^T \Omega \Phi_c^T) \bar{\Phi}_a^2}{4\rho_2}, \\ \frac{\lambda_c}{2} - \rho_1 - \rho_3 - \frac{\lambda_{\max}(\tilde{\lambda}_{W_c})}{4\rho_3} - \rho_4 \lambda_{\max}(\lambda_{W_c} \lambda_{W_c}^T), \end{array} \right\} \\ \varepsilon_f = & -\frac{\gamma_0}{\Gamma_r} r(t) + \frac{\lambda_\varepsilon}{2} \varepsilon_v^2 + \frac{1}{2} \varepsilon_d^2 + \sum_{i=1}^3 \varepsilon_i + \frac{\lambda_{W_a}}{2} W_a^T W_a + \frac{2\rho_2 + \lambda_{W_c}}{2} W_c^T W_c + \frac{1}{4\rho_4} \varepsilon_c^2 \end{aligned} \tag{45}$$

Thanks to Equation (45), we can obtain that

$$\dot{V} \leq -\gamma V + \varepsilon_f + \rho(t) \left(1 - 16 \text{Tanh}^T \left(\frac{e_2(t)}{\varepsilon_\rho} \right) \text{Tanh} \left(\frac{e_2(t)}{\varepsilon_\rho} \right) \right) / \Gamma_r \tag{46}$$

According to Equation (46), it can be seen that signals $[e_0(t), e_1(t), e_2(t), \tilde{\varepsilon}_v(t), \tilde{W}_a(t), \tilde{W}_c(t)]$ are all stable and bounded. Therefore, the stability of the closed-loop system and the boundedness of all signals can be verified. The proof is complete.

5.0 Simulation study

In this section, some numerical simulations are performed to demonstrate the effectiveness and performance of the proposed STDO-based adaptive reinforcement learning (ARL) control method. To show the advantages of the proposed STDO-ARL method, the STDO-ARL without STDO and the STDO-ARL without reinforcement learning (RL) are also considered for comparison, as shown in Figs. 2 and 3. On the other hand, the robustness of the proposed STDO-ARL method is reflected by several external disturbances $d_1(t)$ and uncertainties $\chi(t)$ of different degrees imposed on the system as listed in Table 1, and the results are shown in Figs. 7 and 8.

The initial values of the system for simulation are listed as follows: $x_1 = [0.0675 \quad -0.5738]^T$, $x_2 = [0 \quad 0]^T$; the weights of the actor network and the critic network are respectively set as: $\hat{W}_a = \text{zeros}(22, 1)$, $\hat{W}_c = \text{zeros}(11, 1)$; the mismatched disturbance of the system is $\hat{d}_1 = [0 \quad 0]^T$; what's more, $P_2 = [0 \quad 0]^T$ and $\hat{x}_2 = [0 \quad 0]^T$.

We choose the unmodeled dynamics as $\eta = 1$ and the dynamic auxiliary signal as $r = 2$. The mismatched disturbance $d_1(t)$ in the simulation are set as two different trapezoidal waves: $d_1(t) = \begin{bmatrix} D(t, 5, 3) \\ D(t, 5, 1) \end{bmatrix}$. The uncertainties that are affected by the unmodeled dynamics are supposed to

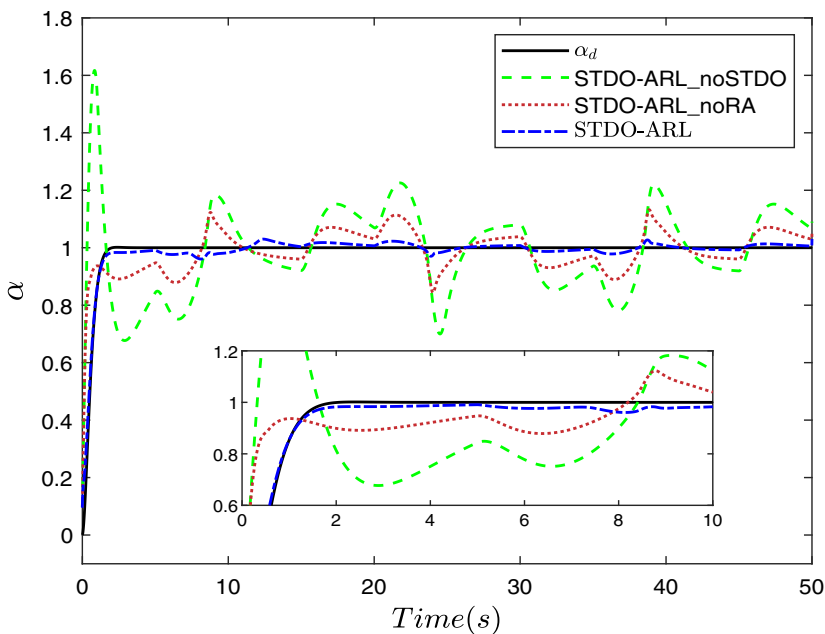


Figure 2. Comparison chart of the tracking performance of α under different methods.

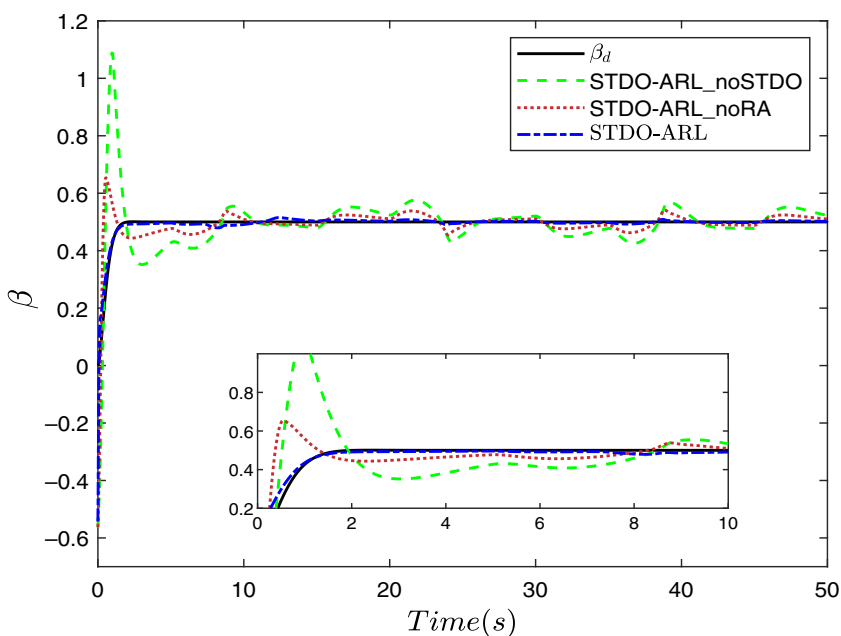


Figure 3. Comparison chart of the tracking performance of β under different methods.

be $\chi(t) = 0.5x_1(t) \sin(t) + \eta x_2(t)$. In this control method, the system matrix is selected as $B = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$ and other control constants are set as $\Gamma_r = 120$, $\varepsilon_\rho = 0.1$. The control gains are designed as $k_0 = 3$, $k_1 = k_2 = 6$. The control parameters of STDO disturbance observer are designed as $K_{d1} = 5$, $K_{P1} = 2$, $K_{P2} = 0.1$. And the adaptive control parameters of reinforcement learning

Table 1. Model parameter values in different cases

	$d_1(t)$	$\chi(t)$
Case1_method	$d_1(t) = \begin{bmatrix} D(t, 5, 3) \\ D(t, 5, 1) \end{bmatrix}$	$0.5x_1\sin(t) + \eta x_2(t)$
Case2_other_disturbance1	$d_1(t) = \begin{bmatrix} \sin(0.5t + 2) \\ \sin(t - 8) + \cos(2t) \end{bmatrix}$	$0.5x_1\sin(t) + \eta x_2(t)$
Case3_other_disturbance2	$d_1(t) = \begin{bmatrix} 2\text{square}(0.2t, 50) \\ 3\text{square}(0.4t, 50) \end{bmatrix}$	$0.5x_1\sin(t) + \eta x_2(t)$
Case4_other_uncertainty	$d_1(t) = \begin{bmatrix} D(t, 5, 3) \\ D(t, 5, 1) \end{bmatrix}$	$0.5x_1\sin(t) + \eta x_2(t) - 0.5x_2(t)\sin(2t)$

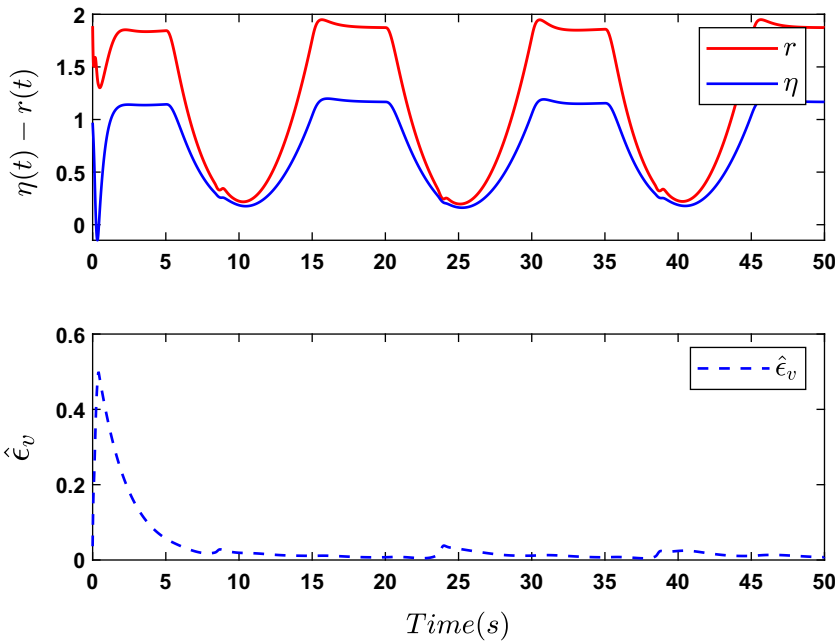


Figure 4. The trajectories of the adaptive parameters of the proposed STDO-ARL scheme.

actor network and critic network are set as $\Gamma_{w_a} = 0.2$, $\lambda_{w_a} = 2.5$, $\Omega = [2 \ 1]^T$, $\Gamma_{w_c} = 0.2$, $\lambda_{w_c} = 2.5$, $Q = \text{diag}([1, 1])$, $R = \text{diag}([2, 1, 0, 1])$. According to the above simulation parameters, the following simulations are carried out: the comparison simulation of different methods and different cases.

5.1 Simulation comparison under different methods

This part is a comparative simulation under different methods. According to Figs. 2 and 3, it is obvious that for the time-varying desired signal, the proposed STDO-ARL control scheme can achieve satisfactory results for the tracking control problems of the straight air compound missile with external disturbances and unmodeled dynamics. While the tracking performance of the proposed method without

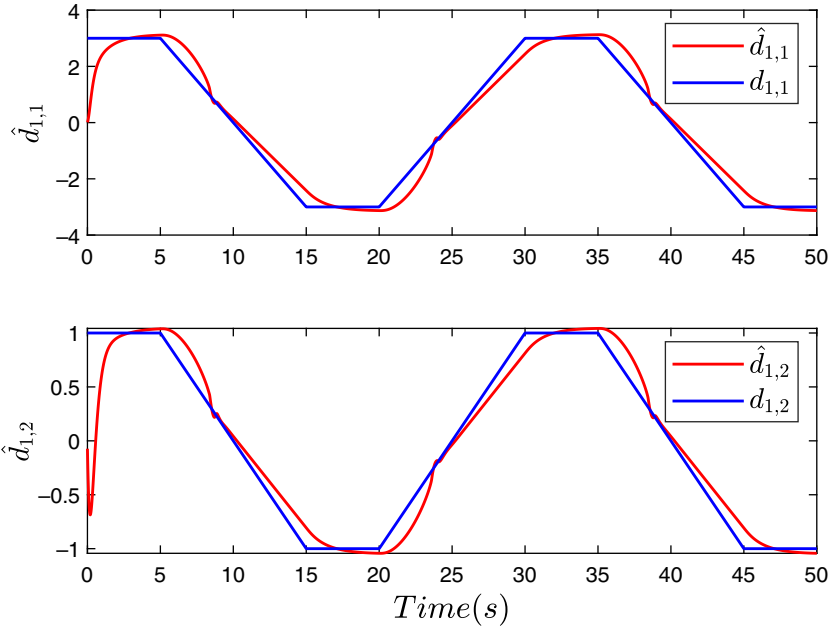


Figure 5. Disturbance estimation effect of $\hat{d}_1(t)$ based on STDO.

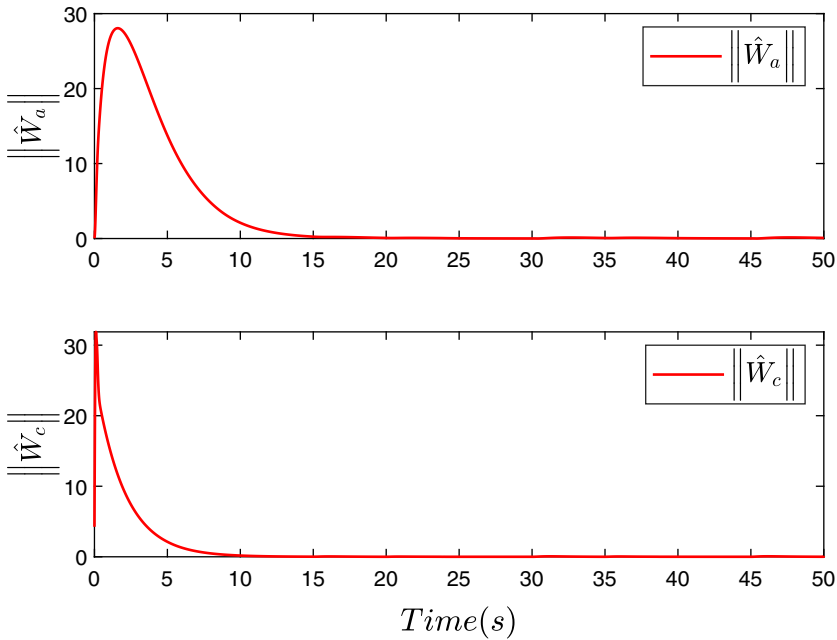


Figure 6. Variation diagram of the weight norm of the reinforcement learning actor NN $\|\hat{W}_a\|$ and the critic NN $\|\hat{W}_c\|$.

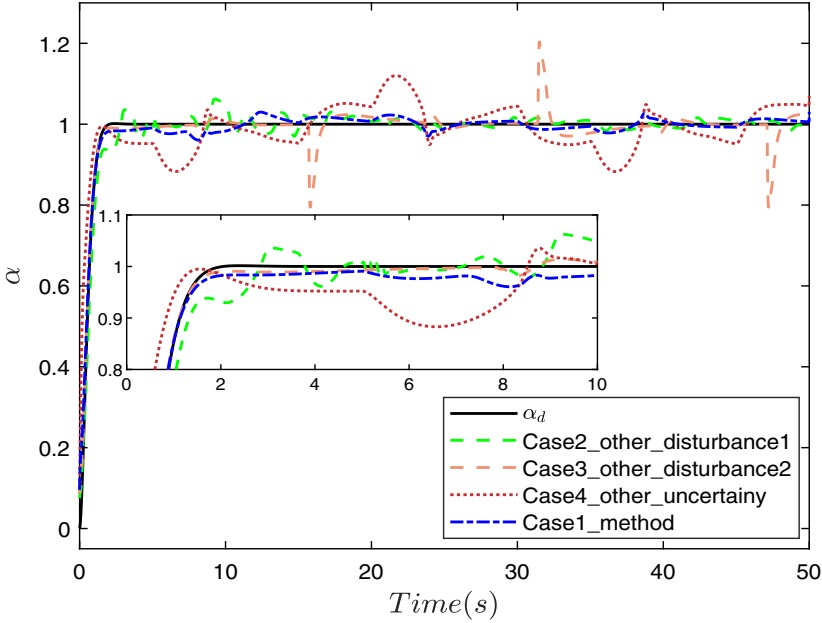


Figure 7. Comparison chart of the tracking performance of α under different cases.

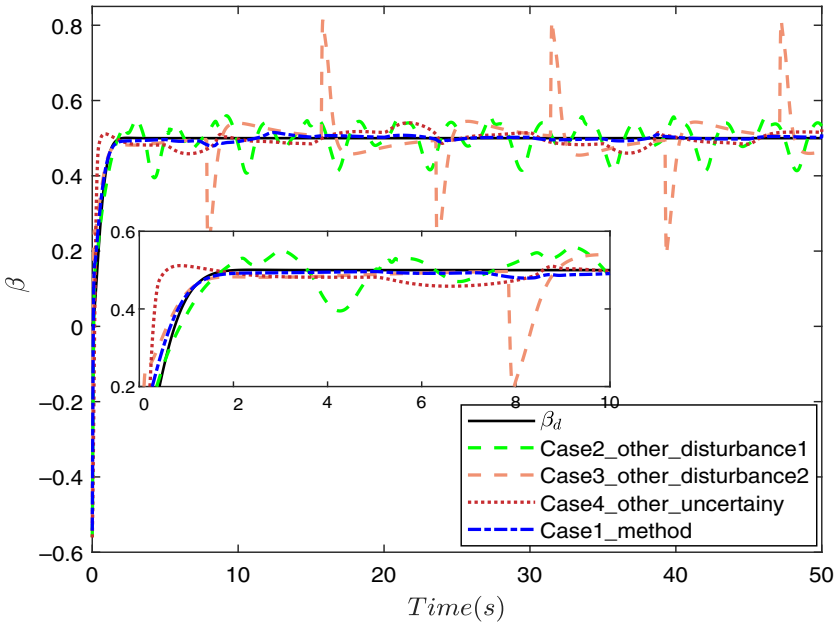


Figure 8. Comparison chart of the tracking performance of β under different cases.

STDO and the proposed method without RL is not ideal, it may produce undesired tracking errors and cannot ensure the tracking accuracy.

Moreover, according to Figs. 4, 5 and 6, all signals in the closed-loop control system are bounded during the whole control process by using the proposed STDO-ARL method. In summary, the proposed

STDO-ARL control method under unmodeled dynamics and disturbances can achieve satisfactory control performance.

5.2 Simulation comparison under different cases

This part is a comparative simulation for the proposed STDO-ARL control scheme under different cases, as shown in Figs. 7 and 8. From the analysis of the simulation results, the proposed STDO-ARL method has the characteristics of high anti-disturbance in dealing with trapezoidal signals and sine-cosine combined signals. However, there will be slight fluctuations when dealing with square-wave signal disturbance. Moreover, for various complex uncertainty conditions, the reinforcement learning structure can be used to fit them effectively. To sum up, the proposed STDO-ARL method has strong robustness and anti-disturbance ability under different cases.

6.0 Conclusion

In this paper, an STDO-based adaptive reinforcement learning control scheme is proposed for the straight air compound missile system with unknown aerodynamic uncertainties and unmodeled dynamics. To deal with the tracking problems for the straight gas compound system, adaptive control with actor-critic design has been investigated in this paper. Considering that the negative impacts of the control signal fluctuation caused by the mismatched disturbance of the straight gas compound system, the STDO disturbance observer has been used to solve the problem well. To improve the control performance, reinforcement learning and neural networks have been adopted in the actor-critic design. The simulation results show that the proposed STDO-ARL controller can guarantee the stability of the straight air compound missile system with unknown aerodynamic uncertainties and unmodeled dynamics. What's more, the effectiveness and robustness of the proposed approach have been illustrated by simulation results. In the future, we will continue to follow up on this problem and consider the reinforcement learning-based anti-coupling control for the straight gas compound system.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grants No.11772256, Science and Technology on Electromechanical Dynamic Control Laboratory, China, No.6142601190210, the Foundation of National Key Laboratory of Science and Technology on Test Physics & Numerical Mathematics, China, and supported by Research Projects KT-KTYWGL-22-22228.

References

- [1] Antonios, T. and Brian, A. Modern missile flight control design: an overview, *IFAC Proc. Vol.*, 2001, **34**, (15), pp 425–430.
- [2] Song, C., Kim, S.J., Kim, S.H. and Nam, H.S. Robust control of the missile attitude based on quaternion feedback, *Control Eng. Pract.*, July 2006, **14**, (7), pp 811–818.
- [3] Yang, J., Chen, W.H. and Li, S. Non-linear disturbance observer-based robust control for systems with mismatched disturbances/uncertainties, *IET Control Theory Appl.*, December 2011, **5**, (18), pp 2053–2062.
- [4] Lee, Y., Kim, Y. and Moon, G. Sliding-mode-based missile-integrated attitude control schemes considering velocity change, *J. Guid. Control Dynam.*, 2016, **39**, (3), pp 423–436.
- [5] Zhou, J. and Yang, J. Smooth sliding mode control for missile interception with finite-time convergence, *J. Guid. Control Dynam.*, 2015, **38**, (7), pp 1311–1318.
- [6] Shao, X. and Wang, H. Back-stepping active disturbance rejection control design for integrated missile guidance and control system via reduced-order ESO, *ISA Trans.*, 2015, **57**, pp 10–22.
- [7] Guo, P., Yang, S. and Zhao, L. Second order sliding mode control with back stepping approach for moving mass spinning missiles, *J. Beijing Inst. Technol.*, 2016, **1**, pp 17–22.
- [8] Wang, L., Zhang, W., Wang, D., Peng, K. and Yang, H. Command filtered back-stepping missile integrated guidance and autopilot based on extended state observer, *Adv. Mech. Eng.*, 2017, **9**, (11), pp 1–13.
- [9] Fan, Y., Li, X., Yang, J. and Zhang, Y. Design of autopilot for aerodynamic/reaction-jet multiple control missile using variable structure control, *2008 27th Chinese Control Conference*, 2008, pp 642–645.
- [10] Shao, L., Zhang, J. and Cao, Y. Blended robust control method with lateral thrust and aerodynamic force based on robust trail tracking, *Aero Weapon*, 2016, **291**, (1), pp 35–39.

- [11] Liu, X., Li, A., Guo, Y., Wang, S. and Wang, C. Fixed-time convergence blended control for air-to-air missile with lateral thrusters and aerodynamic force, *J. Harbin Inst. Technol.*, 2019, **51**, (09), pp 29–34+42.
- [12] Zhao, Y., Liao, Z., Duan, C. and Zhang, G. Design of blended lateral thrust and aerodynamic control system based on terminal sliding mode, *Navig. Position. Timing*, 2015, **2**, (03), pp 49–54.
- [13] Xu, B. and Zhou, D. Backstepping and control allocation for dual aero/propulsive missile control, *Syst. Eng. Electron.*, 2014, **36**, (03), pp 527–531.
- [14] Zhang, X. Design of compound control system with direct lateral thrust and aerodynamics adopting backstepping method, *Modern Defence Technol.*, 2009, **37**, (04), pp 43–46.
- [15] Shi, Z., Ma, W., Zhang, Y. and Lin, Q. Fuzzy control algorithm and realization of compound control missile, *Harbin Gongcheng Daxue Xuebao/J. Harbin Eng. Univ.*, 2014, **35**, (02), pp 195–201.
- [16] Luo, X. and Zhang, T. The application of fuzzy control in combined-guidance, *J. Project. Rockets, Miss. Guid.*, 2001, **02**, pp 1–4.
- [17] Liu, S., Qu, X. and Liu, Y. Design of missile autopilot based on fuzzy control, *2016 IEEE International Conference on Information and Automation (ICIA)*, 2016, pp 1339–1343.
- [18] Fan, Y. and Yang, J. The design of aerodynamic/reaction-jet compound controller of missile actuator using neural network model reference control, *Fire Control Command Control*, 2008, **163**, (10), pp 85–87.
- [19] Zhou, X., Peng, M. and Li, Y. Autopilot design for dual aero/propulsive missile using genetic algorithm LQR control, *Comput. Meas. Control*, 2014, **22**, (04), pp 1157–1159+1162.
- [20] Dong, Z., Chen, J., Song, C. and Cao, H. Design of longitudinal control system for target missiles based on fuzzy adaptive PID control, *2017 29th Chinese Control and Decision Conference (CCDC)*, 2017, pp 398–402.
- [21] Chwa, D. Fuzzy adaptive disturbance observer-based robust adaptive control for skid-to-turn missiles, *IEEE Trans. Aerosp. Electron. Syst.*, 2015, **51**, (01), pp 468–478.
- [22] Yang, P., Fang, Y., Chai, D. and Wu, Y. Fuzzy control strategy for hypersonic missile autopilot with blended aero-fin and lateral thrust, *Proc. Inst. Mech. Eng. I: J. Syst. Control Eng.*, 2016, **230**, (01), pp 72–81.
- [23] Cai, J., Xing, L., Zhang, M. and Shen, L. Adaptive neural network control for missile systems with unknown hysteresis input, *IEEE Access*, 2017, **05**, pp 15839–15847.
- [24] Chen, K. Full state constrained stochastic adaptive integrated guidance and control for STT missiles with non-affine aerodynamic characteristics, *Inform. Sci.*, 2020, **529**, pp 42–58.
- [25] Zhang, H., Chen, Z. and Hao, L. Optimization of missile allocation based on adaptive genetic algorithm, *Tactical Miss. Technol.*, 2007, **124**, (04), pp 28–30+36.
- [26] Shi, Z., Ma, W. and Wang, F. Intelligent control algorithm for missile with lateral jets and aerodynamic surfaces, *J. Nanjing Univ. Sci. Technol.*, 2014, **38**, (04), pp 481–489.
- [27] Polycarpou, M.M. and Ioannou, P.A. A robust adaptive nonlinear control design, *1993 American Control Conference*, 1993, pp 1365–1369.