

Received September 3, 2021, accepted October 3, 2021, date of publication October 8, 2021, date of current version October 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3118723

SODO Based Reinforcement Learning Anti-Disturbance Fault Tolerant Control for a Class of Nonlinear Uncertain Systems With Matched and Mismatched Disturbances

SHIYI HUANG¹, ZHENG WANG^{2,3,4}, ZHAOHUI YUAN¹, KANG CHEN³, AND TAO LI⁵

¹Software Institute, East China Jiaotong University, Nanchang 330000, China

²Research Center for Unmanned System Strategy Development, Northwestern Polytechnical University, Xi'an 710072, China

³Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, China

⁴National Key Laboratory of Aerospace Flight Dynamics, Northwestern Polytechnical University, Xi'an 710072, China

⁵Beijing Institute of Space Long March Vehicle, Beijing 100076, China

Corresponding author: Zheng Wang (wz_nwpu@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grants no.11772256 and supported by Science and Technology Foundation on Electromechanical Dynamic Control Laboratory of China under Grants no. 6142601190210.


ABSTRACT This paper proposes a reinforcement learning anti-disturbance fault tolerant control structure for a class of nonlinear uncertain systems with time varying matched and mismatched disturbances. To deal with the time varying matched and mismatched disturbances, two second order disturbance observers (SODOs) are designed for the inner and outer loop dynamic equations. For the purpose of enhancing the robustness and adaptivity with respect to the system uncertainties, two long short-term memory (LSTM) networks those possesses perfect fitting ability, have been introduced as the critic and actor networks. Moreover, to overcome the difficulty caused by the unknown perturbations of the control effectiveness, several fault tolerant adaptive laws have been designed. Consequently, a novel reinforcement learning anti-disturbance fault tolerant control structure has been established for the concerned disturbed nonlinear uncertain system. Two numerical examples are provided finally, demonstrating the satisfactory performance of the proposed control structure.

INDEX TERMS Adaptive control, reinforcement learning control, disturbance observer, anti-disturbance control, actuator faults.

I. INTRODUCTION

Nonlinear control system is always one of the focuses and difficulties in the control field [1]–[3]. In order to solve the control problem of nonlinear systems, the researchers have proposed a series of control methods and strategies [4], [5]. Isidori and his colleges first proposed a feedback linearization method based on differential geometry to solve nonlinear system problems [6]. A discontinuous nonlinear synovial membrane control method for nonlinear systems was proposed in [7]. In [8], to deal with the nonlinear control system with disturbances and input uncertainties, a novel disturbance observer has been designed and a stable control law has been proposed. In [9], the authors proposed an optimal H_∞ tracking control method for nonlinear multivariable dynamic

systems. In [10], a gap measurement method is introduced to design a multi-model stable controller for a class of nonlinear systems. In [11], taking the matched and mismatched disturbances into consideration, a trajectory linearization control method has been designed for the disturbed nonlinear systems. In [12], for a class of nonlinear systems with time-varying de-lay and state constraints, a novel quantitative adaptive control strategy has been established. In [13], for the miniature high-precision nonlinear system, a proportional integral-differential control method based on the dynamic hysteresis nonlinear model and inverse model has been proposed. However, in the above-mentioned results, the intelligent methods such as the neural networks, the deep learning methods, the reinforcement learning approaches, have never been utilized to construct the control laws, and the suppression performance for the unknown nonlinearities or system uncertainties may have to be enhanced.

The associate editor coordinating the review of this manuscript and approving it for publication was Haiquan Zhao .

Among the plenty of the intelligent algorithms, the reinforcement learning strategies and methods possess the advantages of autonomous learning ability and ability of handling the complex dynamics [14]. Reinforcement learning control is a deep combination of the control technique and reinforcement learning methods, possessing the excellent ability of handling the complex or uncertain dynamics existing in the control systems, and so as to effectively realize the stabilization or tracking control. Recently, many of the reinforcement learning control methods have been investigated or reported.

In [15], a novel adaptive fault-tolerant attitude control approach has been designed based on the long short-term memory (LSTM) network for the fixed-wing UAV subject to the high dynamic disturbances and actuator faults. In [16], a reinforcement learning state feedback control method has been designed. In [17], a reinforcement learning control structure using the hidden reward function has been constructed. In [18] control method has been synthesized by using the reinforcement learning and behavior-critic strategy. In [19], a deep reinforcement learning control method has been provided, and a novel model-free reinforcement learning fault tolerant control structure has been established. In [20], an improved adaptive reinforcement learning control method has been proposed for the deformation control of the aerospace unmanned systems.

Moreover, because of the excellent ability of handling the complex or uncertain dynamics, the reinforcement learning control methods have been applied to a plenty of the practical engineering systems. In [21] and [22], two reinforcement learning trajectory tracking control methods has been investigated for the underactuated ships and the soft robots. In [23], a reinforcement learning control has been proposed for vehicle robot with variable gravitational center. In [24], by using reinforcement learning algorithm, a novel precise control strategy has been proposed for the nonlinear fast hot machining control system. In [25], a strategy-based reinforcement learning control method has been reported, minimizing switching time and overshoot of the nonlinear floating piston system. In [26], a real-time reinforcement learning control method has been proposed for experiential playback. In [27], a model-free reinforcement learning controller has been designed for the electrically driven cold heat storage system. Besides, the reinforcement learning control methods have also been applied to the air injection systems [28], the humanoid robots [29], the HVAC (Heating Ventilation and Air Conditioning) systems [30].

However, the reinforcement learning control methods has never been designed for the nonlinear system with both matched and mismatched disturbances. Moreover, for the matched and mismatched disturbances with time-varying features, the reinforcement learning anti-disturbance control law has been rarely reported. Furthermore, for the nonlinear system suffering from the actuator faults and the multiple disturbances simultaneously, the reinforcement learning controllers is lacking. Therefore, this paper carries out the research of SODO based reinforcement learning anti-disturbance fault

tolerant control for a class of nonlinear uncertain systems with matched and mismatched disturbances. The main contributions of this paper can be summarized as follows:

- To the best of the authors knowledge, the reinforcement learning fault tolerant anti-disturbance controller has been firstly proposed for the nonlinear uncertain systems with matched and mismatched disturbances.
- By using the LSTM networks as the critic and actor networks, the robustness and adaptivity with respect to the system uncertainties can be enhanced.
- Benefitting from the estimation ability of the SODO, both of the matched and mismatched time-varying disturbances can be handled.

II. PROBLEM FORMULATION

A. THE UNCERTAIN NONLINEAR SYSTEMS WITH MATCHED AND MISMATCHED UNCERTAINTIES

Consider the following nonlinear uncertain system with matched and mismatched uncertain-ties:

$$\begin{aligned}\dot{x}_1(t) &= x_2(t) + d_0(t) \\ \dot{x}_2(t) &= f(x_1(t), x_2(t)) + \Delta f(x_1(t), x_2(t)) \\ &\quad + BN[\Lambda + \Delta\Lambda]u(t) + d_1(t) + d_2(t) \\ y(t) &= x_1(t)\end{aligned}\quad (1)$$

where $x_1 \in \mathbb{R}^n$ and $x_2 \in \mathbb{R}^n$ denote the system states, $u \in \mathbb{R}^m$ is the control input. $B \in \mathbb{R}^{n \times m}$, $N \in \mathbb{R}^{m \times m}$. $f(x_1(t), x_2(t))$ and $\Delta f(x_1(t), x_2(t))$ are the known and unknown nonlinearities existing in the considered nonlinear uncertain system. $\Lambda = \Lambda^T \in \mathbb{R}^{m \times m}$ is a known matrix, representing the control effectiveness. $\Delta\Lambda = [\Delta\Lambda]^T \in \mathbb{R}^{m \times m}$ denotes the unknown perturbations of the control effectiveness. $d_0(t)$ denotes the mismatched time-varying disturbance, while $d_1(t)$ and $d_2(t)$ are the matched disturbances.

The objective of this paper is to design a reinforcement learning anti-disturbance control to realize stably tracking for desired signal $y_d(t)$, in the presence of the unknown nonlinearities, the unknown perturbations of the control effectiveness, and the mismatched and matched time-varying disturbances.

To achieve the design objective, the following assumptions are required:

Assumption 1: The matched and mismatched disturbances are all bounded, i.e., there exists a constant $\bar{d}_0, \bar{d}_1, \bar{d}_2$ such that $\forall t \geq 0, \|d_0\| \leq \bar{d}_0, \|d_1\| \leq \bar{d}_1, \|d_2\| \leq \bar{d}_2$.

Assumption 2: Define $M = [\Lambda + \Delta\Lambda]/\Lambda$. It is assumed that $\lambda_{\min}(M) > 0$.

Assumption 3: The desired signal $y_d(t)$ is assumed to be smooth and twice differential.

B. LONG SHORT-TERM MEMORY NETWORK

To achieve the reinforcement learning anti-disturbance control for the concerned nonlinear uncertain system with matched and mismatched disturbances, the long short-term memory networks are introduced as action network and critic network.

The output of the LSTM can be formulated by

$$z = W^T \Phi(y) \quad (2)$$

where $y \in \mathbb{R}^n$ and $z \in \mathbb{R}$ represent the input and output signals of the LSTM. The LSTM includes the forget gates, input gates, memory states, update gate and output gates, those can be described as follows:

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\ \tilde{c}_t &= \tanh(W_c[h_{t-1}, x_t] + b_c) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \end{aligned} \quad (3)$$

The final output state is $h_t = o_t \circ \tanh(c_t)$.

Lemma 1 ([38], [39]): For any unknown smooth function, the LSTM network can achieve approximation with bounded errors. In details, for any given smooth function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, the following equation holds:

$$f(y) = W^T \Phi(y) + \varepsilon$$

where W is the weight matrix of LSTM, ε is the error of LSTM approximation.

III. MAIN RESULTS

A. THE CONTROL STRUCTURE OF THE REINFORCEMENT LEARNING ANTI-DISTURBANCE CONTROLLER

The control structure of the proposed reinforcement learning anti-disturbance control law is shown in Fig 1. Two SODOs are utilized to handle the matched and mismatched time-varying disturbances. The critic network is utilized to evaluate the anti-disturbance control performance of the closed-loop system, and the actor network is introduced as a component in the anti-disturbance fault tolerant control law.

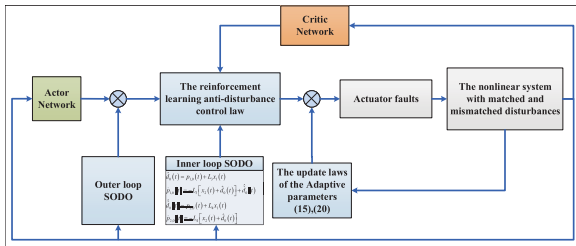


FIGURE 1. The control structure of the proposed reinforcement learning anti-disturbance control law.

B. REINFORCEMENT LEARNING ANTI-DISTURBANCE CONTROL LAW

Define $e_1(t) = x_1(t) - y_d(t)$. Based on the dynamic equation of the nonlinear uncertain system (1), the cost function is selected as follows.

$$J = \int_0^\infty [e_1^T(\tau) Q e_1(\tau) + u^T(\tau) R u(\tau)] d\tau \quad (4)$$

To approximate the cost function, a LSTM network is selected as the critic network, which is

$$J = W_c^T \Phi_c(x) + \varepsilon_c \quad (5)$$

where $W_c \in \mathbb{R}^{p_c}$ is the desired weight of the critic network. ε_c is a bounded error of the critic network. p_c is nodes number of the NN, and $\Phi_c(x) \in \mathbb{R}^{p_c}$ is a vector of the primary functions.

Define \hat{J} and \hat{W}_c as the estimated value of J and W_c , respectively. Hence, we can get that

$$\hat{J} = \hat{W}_c^T \Phi_c(x) \quad (6)$$

Construct the residual mean square error function of the critic network as

$$\begin{aligned} E_c &= \frac{1}{2} e_c^T e_c \\ e_c &= e_1^T Q e_1 + u^T R u + \hat{W}_c^T \nabla \Phi_c \dot{x}_1 \end{aligned} \quad (7)$$

where $\nabla \Phi_c = \partial \Phi_c(x) / \partial x$, $\nabla \Phi_c \in \mathbb{R}^{p_c \times n}$. The updating objective of the weight of the critic network is to minimize E_c . Therefore, according to the gradient descent method, the update law for the weight of the critic network can be designed as

$$\dot{\hat{W}}_c = c_{W_c} [\sigma_{W_c} (\sigma_{W_c}^T \hat{W}_c + Q e_1^2 + R u^2)] - c_{W_c} \sigma_c \hat{W}_c \quad (8)$$

where $\sigma_{W_c} = \nabla \Phi_c \dot{x}_1$. $c_{W_c}, \sigma_{W_c} > 0$ are positive design constants.

In this paper, the actor network is fused into the adaptive fault tolerant controller. Considering that $\Delta f(x_1(t), x_2(t))$ is unknown, a LSTM network is introduced as the actor network, which is

$$\Delta f = W_a^T \Phi_a(x) + \varepsilon_f \quad (9)$$

where $W_a \in \mathbb{R}^{p_a \times n}$, $\Phi_a(x) \in \mathbb{R}^{p_a}$ represents the desired weight and the primary function vector of the actor network. ε_f is the bounded error of the actor network, satisfying that $\|\varepsilon_f\| \leq \bar{\varepsilon}_f$. The estimated value of W_a is defined as \hat{W}_a .

Taking the derivatives of both sides of $e_1(t)$ yields that

$$\dot{e}_1(t) = x_2(t) + d_0(t) - \dot{y}_d(t) \quad (10)$$

To force the inner loop of the nonlinear uncertain system to be stable, the virtual control signal is designed as

$$x_{2c}(t) = -K_1 e_1(t) - \hat{d}_0(t) + \dot{y}_d(t) \quad (11)$$

where $K_1 > 0$ is the control gain. $\hat{d}_0(t)$ is the estimated value of the mismatched time varying disturbance $d_0(t)$, obtained from the following second order disturbance observer:

$$\begin{aligned} \hat{d}_0(t) &= p_{1,0}(t) + L_3 x_1(t) \\ \dot{p}_{1,0}(t) &= -L_3 [x_2(t) + \hat{d}_0(t)] + \dot{\hat{d}}_0(t) \\ \hat{d}_0(t) &= p_{2,0}(t) + L_4 x_1(t) \\ \dot{p}_{2,0}(t) &= -L_4 [x_2(t) + \hat{d}_0(t)] \end{aligned} \quad (12)$$

Define $e_2(t) = x_2(t) - x_{2c}(t)$. We can get that

$$\begin{aligned} \dot{e}_2(t) &= f(x_1(t), x_2(t)) + \Delta f(x_1(t), x_2(t)) \\ &\quad + B N [\Lambda + \Delta \Lambda] u + d_1(t) + d_2(t) - \dot{x}_{2c}(t) \end{aligned} \quad (13)$$

By combining the actor network, the baseline control signal is designed as

$$u_c = \frac{[BN\Lambda]^T}{BN\Lambda[BN\Lambda]^T} \begin{pmatrix} -K_2 e_2(t) - e_1(t) \\ -f(x_1(t), x_2(t)) \\ -\hat{W}_a^T \Phi_a(x) - \hat{d}_1(t) \\ -\hat{d}_2(t) + \dot{x}_{2c}(t) \end{pmatrix} \quad (14)$$

where $K_2 > 0$ denotes the control gain matrix of the outer loop. $\hat{d}_1(t)$ is an adaptive parameter, generated from the following equation:

$$\dot{\hat{d}}_1 = c_d [e_2 - \sigma_d \hat{d}_1] \quad (15)$$

$\hat{d}_2(t)$ is the estimated value of mismatched time varying disturbance $d_2(t)$, generated from the following second order disturbance observer:

$$\begin{aligned} \hat{d}_2(t) &= p_{1,2}(t) + L_1 x_2(t) \\ \dot{p}_{1,2}(t) &= -L_1 [f(x_1(t), x_2(t)) + \hat{W}_a^T \Phi_a(x)] \\ &\quad - L_1 [BN\Lambda u(t) + \hat{d}_1(t) + \hat{d}_2(t)] + \hat{d}_2(t) \\ \hat{d}_2(t) &= p_{2,2}(t) + L_2 x_2(t) \\ \dot{p}_{2,2}(t) &= -L_2 [f(x_1(t), x_2(t)) + \hat{W}_a^T \Phi_a(x)] \\ &\quad - L_2 [BN\Lambda u(t) + \hat{d}_1(t) + \hat{d}_2(t)] \end{aligned} \quad (16)$$

The update law of \hat{W}_a is designed as follows:

$$\dot{\hat{W}}_a = c_{W_a} \Phi_a [e_2^T + \hat{J} \Omega^T] - c_{W_a} \sigma_a \hat{W}_a \quad (17)$$

The final control signal is designed as

$$u(t) = u_c(t) + u_a(t) \quad (18)$$

where u_a is utilized to deal with the unknown perturbations of the control effectiveness, de-signed by:

$$u_a(t) = -\hat{M}(t) u_c(t) \quad (19)$$

where \hat{M} is the estimated value of M . The adaptive law of \hat{M} is designed as

$$\dot{\hat{M}} = c_M [N^T B^T e_2 u_c^T - \sigma_M \hat{M}] \quad (20)$$

C. STABILITY ANALYSIS

Theorem 1: Consider the nonlinear uncertain system (1) with matched and mismatched uncertainties. Suppose Assumption 1, 2 and 3 are satisfied. If the critic network and the actor network are selected as (5) and (9) respectively, the reinforcement learning anti-disturbance control law is designed as (18), (15) (20), the update law for the network weights are designed as (8) and (17), then all the signals of the closed-loop control system will be bounded and the tracking error can be forced to converge into a compact neighborhood of zero.

Proof: Define $\tilde{d}_0 = \hat{d}_0 - d_0$, $\tilde{d}_1 = \hat{d}_1 - d_1$, $\tilde{d}_2 = \hat{d}_2 - d_2$, $\tilde{d}_2 = \hat{d}_2 - d_2$. By using equation (12) and (16), we and take the derivative to get:

$$\dot{\tilde{d}}_0 = -L_3 \tilde{d}_0 + \tilde{d}_0$$

$$\begin{aligned} \dot{\tilde{d}}_0 &= -L_4 \tilde{d}_0 \\ \dot{\tilde{d}}_2 &= -L_1 \tilde{d}_2 - L_1 \tilde{d}_1 + \tilde{d}_2 - L_1 \tilde{W}_a^T \Phi_a + L_1 BN \Delta \Lambda u \\ \dot{\tilde{d}}_2 &= -L_2 \tilde{d}_2 - L_2 \tilde{W}_a^T \Phi_a + L_2 BN \Delta \Lambda u \end{aligned} \quad (21)$$

Define $\tilde{W}_a = \hat{W}_a - W_a$, $\tilde{W}_c = \hat{W}_c - W_c$, $\tilde{M} = \hat{M} - M$, $\tilde{d}_1 = \hat{d}_1 - d_1$, the following closed loop equations can be obtained:

$$\begin{aligned} \dot{e}_2 &= -K_2 e_2 - e_1 - \tilde{W}_a^T \Phi_a - \tilde{d}_1 - \tilde{d}_2 - BN (\Delta \Lambda + \Lambda) \tilde{M} u_c \\ \dot{\tilde{W}}_a &= c_{W_a} \Phi_a [e_2^T + \hat{J} \Omega^T] - c_{W_a} \sigma_a \hat{W}_a \\ \dot{\tilde{W}}_c &= -c_{W_c} [\sigma_{W_c} (\sigma_{W_c}^T \hat{W}_c + \varepsilon_c)] - c_{W_c} \sigma_c \hat{W}_c \end{aligned} \quad (22)$$

The Lyapunov function $L(t)$ is selected as follows:

$$\begin{aligned} L(t) &= \frac{1}{2} e_1^T e_1 + \frac{1}{2} e_2^T e_2 + \frac{1}{2} \tilde{d}_0^T \tilde{d}_0 + \frac{1}{2} \tilde{d}_1^T \tilde{d}_1 \\ &\quad + \frac{1}{2c_d} \tilde{d}_1^T \tilde{d}_1 + \frac{1}{2} \tilde{d}_2^T \tilde{d}_2 + \frac{1}{2} \tilde{d}_2^T \tilde{d}_2 \\ &\quad + \frac{1}{2c_M} Tr(\tilde{M}^T (\Delta \Lambda + \Lambda) \tilde{M}) + \frac{1}{2c_{W_a}} Tr(\tilde{W}_a^T \tilde{W}_a) \\ &\quad + \frac{1}{2c_{W_c}} Tr(\tilde{W}_c^T \tilde{W}_c) + J^*(x_1) \end{aligned} \quad (23)$$

Take the derivative of (23) as follows

$$\begin{aligned} \dot{L}(t) &= \frac{1}{2} e_1^T e_1 + \frac{1}{2} e_2^T e_2 + \frac{1}{2} \tilde{d}_0^T \tilde{d}_0 + \frac{1}{2} \tilde{d}_1^T \tilde{d}_1 \\ &\quad + \frac{1}{2c_d} \tilde{d}_1^T \tilde{d}_1 + \frac{1}{2} \tilde{d}_2^T \tilde{d}_2 + \frac{1}{2} \tilde{d}_2^T \tilde{d}_2 \\ &\quad + \frac{1}{2c_M} Tr(\tilde{M}^T (\Delta \Lambda + \Lambda) \tilde{M}) + \frac{1}{2c_{W_a}} Tr(\tilde{W}_a^T \tilde{W}_a) \\ &\quad + \frac{1}{2c_{W_c}} Tr(\tilde{W}_c^T \tilde{W}_c) + J^*(x_1) \end{aligned} \quad (24)$$

By using equation (17), it can be known that

$$\begin{aligned} c_{W_a}^{-1} Tr(\tilde{W}_a^T \tilde{W}_a) &= e_2^T \tilde{W}_a^T \Phi_a + \tilde{W}_c^T \Phi_c \Omega^T \tilde{W}_a^T \Phi_a \\ &\quad + W_c^T \Phi_c \Omega^T \tilde{W}_a^T \Phi_a - \sigma_a Tr(\tilde{W}_a^T \hat{W}_a) \\ &\leq \rho_8 e_2^T e_2 + \frac{\bar{\Phi}_a^2}{4\rho_8} Tr[\tilde{W}_a^T \tilde{W}_a] + \rho_9 \tilde{W}_c^T \tilde{W}_c \\ &\quad + \frac{\lambda_{\max}(\Phi_c \Omega^T \Omega \Phi_c^T) \bar{\Phi}_a^2}{4\rho_9} Tr[\tilde{W}_a^T \tilde{W}_a] \\ &\quad + \rho_{10} W_c^T W_c + \frac{\lambda_{\max}(\Phi_c \Omega^T \Omega \Phi_c^T) \bar{\Phi}_a^2}{4\rho_{10}} Tr[\tilde{W}_a^T \tilde{W}_a] \\ &\quad - \frac{\sigma_a}{2} \tilde{W}_a^T \tilde{W}_a + \frac{\sigma_a}{2} W_a^T W_a \end{aligned} \quad (25)$$

Similarly, the following inequality can be get:

$$\begin{aligned} c_{W_c}^{-1} Tr(\tilde{W}_c^T \tilde{W}_c) &= -\tilde{W}_c^T \sigma_{W_c} \sigma_{W_c}^T \tilde{W}_c \\ &\quad - \tilde{W}_c^T \sigma_{W_c} \varepsilon_c - \tilde{W}_c^T \sigma_c \hat{W}_c \\ &\leq \left(\rho_6 + \frac{\lambda_{\max}(\bar{\sigma}_{W_c})}{4\rho_6} \right) \tilde{W}_c^T \tilde{W}_c \end{aligned}$$

$$\begin{aligned}
 & + \rho_7 \lambda_{\max} \left(\sigma_{W_c} \sigma_{W_c}^T \right) \tilde{W}_c^T \tilde{W}_c \\
 & + \frac{1}{4\rho_7} \varepsilon_c^2 - \frac{\sigma_c}{2} \tilde{W}_c^T \tilde{W}_c + \frac{\sigma_c}{2} W_c^T W_c
 \end{aligned} \tag{26}$$

where $\bar{\sigma}_{W_c} = \sigma_{W_c} \sigma_{W_c}^T \sigma_{W_c} \sigma_{W_c}^T$. Accordingly, it can be known that $J_x^{*T} \dot{x}_1$ satisfies:

$$J_x^{*T} \dot{x}_1 \leq -\lambda_{\min}\{Q\} \|e_1\|^2 - \lambda_{\min}\{R\} \|u\|^2 \tag{27}$$

Substituting (21) into (24) yields:

$$\tilde{d}_0^T \dot{\tilde{d}}_0 + \tilde{d}_0^T \ddot{\tilde{d}}_0 = -\tilde{d}_0^T L_3 \tilde{d}_0 + \tilde{d}_0^T \tilde{d}_0 - \tilde{d}_0^T L_4 \tilde{d}_0 \tag{28}$$

It follows that

$$\tilde{d}_0^T \tilde{d}_0 \leq \rho_{11} \tilde{d}_0^T \tilde{d}_0 + \frac{1}{4\rho_{11}} \tilde{d}_0^T \tilde{d}_0 \tag{29}$$

Meanwhile, we know that

$$\begin{aligned}
 \tilde{d}_2^T \dot{\tilde{d}}_2 + \tilde{d}_2^T \ddot{\tilde{d}}_2 & = -\tilde{d}_2^T L_1 \tilde{d}_2 - \tilde{d}_2^T L_1 \tilde{d}_1 \\
 & + \tilde{d}_2^T \tilde{d}_2 - \tilde{d}_2^T L_1 \tilde{W}_a^T \Phi_a + \tilde{d}_2^T L_1 B N \Delta \Lambda u \\
 & - \tilde{d}_2^T L_2 \tilde{d}_2 - \tilde{d}_2^T L_2 \tilde{W}_a^T \Phi_a + \tilde{d}_2^T L_2 B N \Delta \Lambda u
 \end{aligned} \tag{30}$$

By using the Young's inequities, we can get that:

$$\begin{aligned}
 -\tilde{d}_2^T L_1 \tilde{d}_1 & \leq \rho_0 \tilde{d}_2^T \tilde{d}_2 + \frac{\lambda_{\max} [L_1^T L_1]}{4\rho_0} \tilde{d}_1^T \tilde{d}_1 \\
 \tilde{d}_2^T \tilde{d}_2 & \leq \rho_1 \tilde{d}_2^T \tilde{d}_2 + \frac{1}{4\rho_1} \tilde{d}_2^T \tilde{d}_2 \\
 -\tilde{d}_2^T L_1 \tilde{W}_a^T \Phi_a & \leq \rho_2 \tilde{d}_2^T \tilde{d}_2 + \frac{\lambda_{\max} [L_1^T L_1]}{4\rho_2} \bar{\Phi}_a^2 Tr [\tilde{W}_a^T \tilde{W}_a] \\
 \tilde{d}_2^T L_1 B N \Delta \Lambda u & \leq \rho_3 \lambda_{\max} (\bar{L}_1) \tilde{d}_2^T \tilde{d}_2 + \frac{1}{4\rho_3} \varepsilon_\delta^2 \\
 -\tilde{d}_2^T L_2 \tilde{W}_a^T \Phi_a & \leq \rho_4 \tilde{d}_2^T \tilde{d}_2 + \frac{\lambda_{\max} [L_2^T L_2]}{4\rho_4} \bar{\Phi}_a^2 Tr [\tilde{W}_a^T \tilde{W}_a] \\
 \tilde{d}_2^T L_2 B N \Delta \Lambda u & \leq \rho_5 \lambda_{\max} (\bar{L}_2) \tilde{d}_2^T \tilde{d}_2 + \frac{1}{4\rho_5} \varepsilon_\delta^2
 \end{aligned} \tag{31}$$

where $\bar{\Phi}_a$ is the upper bound on the demonstration number $\|\Phi_a\|$. By combining equations (27), (28) and (29), we know that:

$$\begin{aligned}
 & \tilde{d}_2^T \dot{\tilde{d}}_2 + \tilde{d}_2^T \ddot{\tilde{d}}_2 \\
 & \leq - \left[L_1 - \sum_{i=0}^2 \rho_i - \rho_3 \lambda_{\max} (\bar{L}_1) \right] \tilde{d}_2^T \tilde{d}_2 \\
 & - \left[L_2 - \frac{1}{4\rho_1} - \rho_4 - \rho_5 \lambda_{\max} (\bar{L}_2) \right] \tilde{d}_2^T \tilde{d}_2 \\
 & + \frac{\lambda_{\max} [L_1^T L_1]}{4\rho_0} \tilde{d}_1^T \tilde{d}_1 + \left[\frac{1}{4\rho_3} + \frac{1}{4\rho_5} \right] \varepsilon_\delta^2 \\
 & + \left[\frac{\lambda_{\max} [L_1^T L_1]}{4\rho_2} + \frac{\lambda_{\max} [L_2^T L_2]}{4\rho_4} \right] \bar{\Phi}_a^2 Tr [\tilde{W}_a^T \tilde{W}_a]
 \end{aligned} \tag{32}$$

Moreover, the following inequities can be obtained

$$\begin{aligned}
 -e_1^T \tilde{d}_0 & \leq \rho_{z1} e_1^T e_1 + \frac{1}{4\rho_{z1}} \tilde{d}_0^T \tilde{d}_0 \\
 -e_2^T \tilde{d}_2 & \leq \rho_{z2} e_2^T e_2 + \frac{1}{4\rho_{z2}} \tilde{d}_2^T \tilde{d}_2
 \end{aligned} \tag{33}$$

By combining equations (24), (32) and (33), we know that

$$\begin{aligned}
 \dot{L}(t) & \leq -e_1^T (K_1 - \rho_{z1}) e_1 - e_2^T (K_2 - \rho_{z2}) e_2 \\
 & + \frac{1}{c_{W_a}} Tr (\tilde{W}_a^T \dot{\tilde{W}}_a) - e_2^T \tilde{W}_a^T \Phi_a + \frac{1}{c_d} Tr (\tilde{d}_1^T \dot{\tilde{d}}_1) - e_2^T \tilde{d}_1 \\
 & + \frac{1}{c_M} Tr (\tilde{M}^T (\Delta \Lambda + \Lambda) \dot{\tilde{M}}) - e_2^T B N (\Delta \Lambda + \Lambda) \tilde{M} u_c \\
 & - \left[L_3 - \rho_{11} - \frac{1}{4\rho_{z1}} \right] \tilde{d}_0^T \tilde{d}_0 - \left[L_4 - \frac{1}{4\rho_{11}} \right] \tilde{d}_0^T \tilde{d}_0 \\
 & - \left[L_1 - \sum_{i=0}^2 \rho_i - \rho_3 \lambda_{\max} (\bar{L}_1) - \frac{1}{4\rho_{z2}} \right] \tilde{d}_2^T \tilde{d}_2 \\
 & - \left[L_2 - \frac{1}{4\rho_1} - \rho_4 - \rho_5 \lambda_{\max} (\bar{L}_2) \right] \tilde{d}_2^T \tilde{d}_2 \\
 & + \frac{\lambda_{\max} [L_1^T L_1]}{4\rho_0} \tilde{d}_1^T \tilde{d}_1 \\
 & + \left[\frac{\lambda_{\max} [L_1^T L_1]}{4\rho_2} + \frac{\lambda_{\max} [L_2^T L_2]}{4\rho_4} \right] \bar{\Phi}_a^2 Tr [\tilde{W}_a^T \tilde{W}_a] \\
 & + \left[\frac{1}{4\rho_3} + \frac{1}{4\rho_5} \right] \varepsilon_\delta^2 + \frac{1}{c_{W_c}} Tr (\tilde{W}_c^T \dot{W}_c) + J_x^{*T} \dot{x}_1
 \end{aligned} \tag{34}$$

By using equation (17), (20) and (25) (27), the following equation can be obtained:

$$\begin{aligned}
 \dot{L}(t) & \leq -e_1^T (K_1 - \rho_{z1} I) e_1 - e_2^T (K_2 - \rho_{z2} I - \rho_8 I) e_2 \\
 & - \left[\frac{\sigma_a}{2} - \frac{\lambda_{\max} [L_1^T L_1] \bar{\Phi}_a^2}{4\rho_2} \right. \\
 & \left. - \frac{\lambda_{\max} [L_2^T L_2] \bar{\Phi}_a^2}{4\rho_4} - \frac{\bar{\Phi}_a^2}{4\rho_8} \right] Tr [\tilde{W}_a^T \tilde{W}_a] \\
 & - \left[\frac{4\rho_9}{\lambda_{\max} (\Phi_c \Omega^T \Omega \Phi_c^T) \bar{\Phi}_a^2} \right. \\
 & \left. - \frac{4\rho_9}{\lambda_{\max} (\Phi_c \Omega^T \Omega \Phi_c^T) \bar{\Phi}_a^2} \right] \\
 & - \left[\frac{\sigma_c}{2} - \rho_6 - \rho_9 - \frac{\lambda_{\max} (\bar{\sigma}_{W_c})}{4\rho_6} \right] \tilde{W}_c^T \tilde{W}_c \\
 & - \left[-\rho_7 \lambda_{\max} (\sigma_{W_c} \sigma_{W_c}^T) \right] \\
 & - \frac{\sigma_M}{2} Tr (\tilde{M}^T (\Delta \Lambda + \Lambda) \tilde{M}) - \left[\frac{\sigma_d}{2} - \frac{\lambda_{\max} [L_1^T L_1]}{4\rho_0} \right] \tilde{d}_1^T \tilde{d}_1 \\
 & - \left[L_3 - \rho_{11} - \frac{1}{4\rho_{z1}} \right] \tilde{d}_0^T \tilde{d}_0 - \left[L_4 - \frac{1}{4\rho_{11}} \right] \tilde{d}_0^T \tilde{d}_0 \\
 & - \left[L_1 - \sum_{i=0}^2 \rho_i - \rho_3 \lambda_{\max} (\bar{L}_1) - \frac{1}{4\rho_{z2}} \right] \tilde{d}_2^T \tilde{d}_2
 \end{aligned}$$

$$\begin{aligned}
 & - \left[L_2 - \frac{1}{4\rho_1} - \rho_4 - \rho_5 \lambda_{\max}(\bar{L}_2) \right] \tilde{d}^T \tilde{d}_2 \\
 & - \lambda_{\min}\{Q\} \|e_1\|^2 - \lambda_{\min}\{R\} \|u\|^2 \\
 & + \left[\frac{1}{4\rho_3} + \frac{1}{4\rho_5} \right] \varepsilon_\delta^2 + \frac{\sigma_M}{2} Tr \left(M^T (\Delta\Lambda + \Lambda) M \right) \\
 & + \frac{\sigma_d}{2} d_1^T d_1 + \frac{\sigma_a}{2} Tr \left[W_a^T W_a \right] \\
 & + \left(\rho_{10} + \frac{\sigma_c}{2} \right) W_c^T W_c + \frac{1}{4\rho_7} \varepsilon_c^2
 \end{aligned} \tag{35}$$

Define:

$$c = \min \left\{ \begin{array}{l} \lambda_{\min}(K_1 - \rho_{z1}I), \lambda_{\min}(K_2 - \rho_{z2}I - \rho_8I), \\ \lambda_{\min}\{Q\}, \lambda_{\min}\{R\}, \\ \frac{c_M \sigma_M}{2}, \frac{c_d \sigma_d}{2} - \frac{c_d \lambda_{\max} [L_1^T L_1]}{4\rho_0}, \\ \frac{c_{W_a} \sigma_a}{2} - \frac{c_{W_a} \lambda_{\max} [L_1^T L_1] \bar{\Phi}_a^2}{4\rho_2}, \\ \frac{c_{W_a} \lambda_{\max} [L_2^T L_2] \bar{\Phi}_a^2}{4\rho_4} - \frac{c_{W_a} \bar{\Phi}_a^2}{4\rho_8}, \\ \frac{c_{W_a} \lambda_{\max} (\Phi_c \Omega^T \Omega \Phi_c^T) \bar{\Phi}_a^2}{4\rho_9}, \\ \frac{c_{W_c} \sigma_c}{2} - c_{W_c} \rho_6 - c_{W_c} \rho_9 - \\ \frac{c_{W_c} \lambda_{\max} (\bar{\sigma}_{W_c})}{4\rho_6} - c_{W_c} \rho_7 \lambda_{\max} (\sigma_{W_c} \sigma_{W_c}^T), \\ L_3 - \rho_{11} - \frac{1}{4\rho_{z1}}, L_4 - \frac{1}{4\rho_{11}}, \\ L_1 - \sum_{i=0}^2 \rho_i - \rho_3 \lambda_{\max}(\bar{L}_1) - \frac{1}{4\rho_{z2}}, \\ L_2 - \frac{1}{4\rho_1} - \rho_4 - \rho_5 \lambda_{\max}(\bar{L}_2) \end{array} \right.$$

$$\begin{aligned}
 \varepsilon_L &= \left[\frac{1}{4\rho_3} + \frac{1}{4\rho_5} \right] \varepsilon_\delta^2 + \frac{\sigma_M}{2} Tr \left(M^T (\Delta\Lambda + \Lambda) M \right) \\
 &+ \frac{\sigma_d}{2} d_1^T d_1 + \frac{\sigma_a}{2} Tr \left[W_a^T W_a \right] \\
 &+ \left(\rho_{10} + \frac{\sigma_c}{2} \right) W_c^T W_c + \frac{1}{4\rho_7} \varepsilon_c^2
 \end{aligned} \tag{36}$$

Then according to (35) and (36), we can get:

$$\dot{L}(t) \leq -cL(t) + \varepsilon_L \tag{37}$$

and

$$L(t) \leq L(0)e^{-ct} + \varepsilon_L \tag{38}$$

Therefore, according to (38), it can be known that the system state, the disturbance estimation error $\tilde{d}_0, \hat{d}_0, \tilde{d}_2, \hat{d}_2$ of SODO and the adaptive estimation error $\tilde{d}_1, \hat{W}_a, \hat{W}_c, \hat{M}$ are all bounded. In addition, the boundedness of $\dot{x}_1, \dot{x}_2, \hat{d}_1, \hat{d}_0, \hat{d}_0, \hat{d}_2, \hat{d}_2, \hat{M}, \hat{W}_a, \hat{W}_c$ can be verified. Moreover, it is obvious that the tracking error can be forced to converge into a compact neighborhood of zero, which completes the proof.

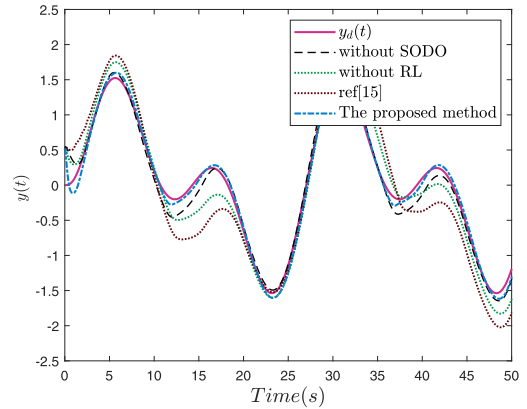


FIGURE 2. The tracking performance for the desired signal under Case 1.

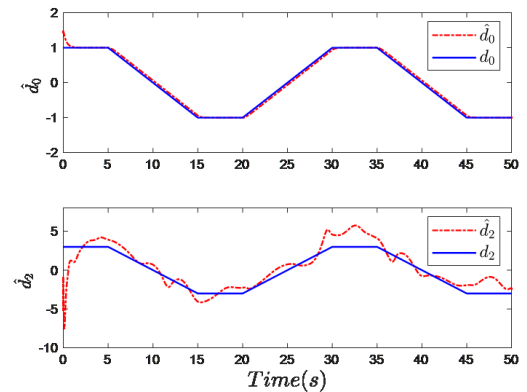


FIGURE 3. The estimation performance for the matched and mismatched disturbances \hat{d}_0 and \hat{d}_2 under Case 1.

IV. SIMULATION STUDY

In order to evaluate the effectiveness and performance of the proposed reinforcement learning-anti-disturbance fault-tolerant control law, a numerical example is provided in this section.

In the simulation, we select $\Delta f = 2 \sin(x_1 + 5) + \cos(6x_2 - 4)$, $B = 1, N = 1$. The desired signal y_d is generated by $\ddot{y}_d + 2\dot{y}_d + y_d = \sin(0.25t) + \sin(0.5t)$, $y_d(0) = 0, \dot{y}_d(0) = 0$. The initial value of the system is set as: $x_1 = 0.5, x_2 = 0, \hat{d}_1 = 0, p_{1,0} = 0, p_{2,0} = 0, p_{1,2} = 0, p_{2,2} = 0, \hat{M} = 0$. The initial weight parameters of the actor network and the critic network are set as:

$$\hat{W}_a = \begin{bmatrix} 0.2 \\ 0.6 \\ -0.3 \\ 0.1 \\ -0.5 \\ 0.8 \\ 0.15 \\ -0.23 \\ 0.35 \end{bmatrix}, \quad \hat{W}_c = \begin{bmatrix} 0.3 \\ 0.1 \\ -0.7 \\ 0.5 \\ -0.64 \\ 0.28 \\ 0.85 \\ -0.23 \\ 0.35 \\ -0.11 \\ -0.92 \end{bmatrix}^T$$

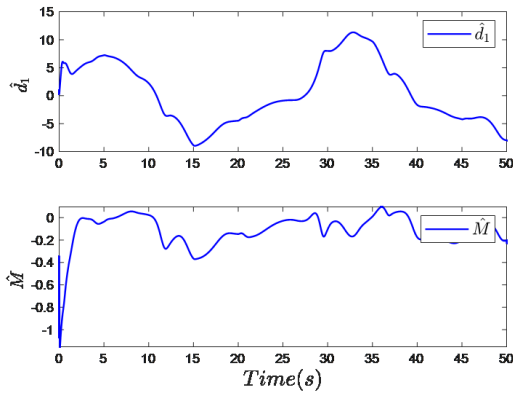


FIGURE 4. The adaptive parameters under Case 1.

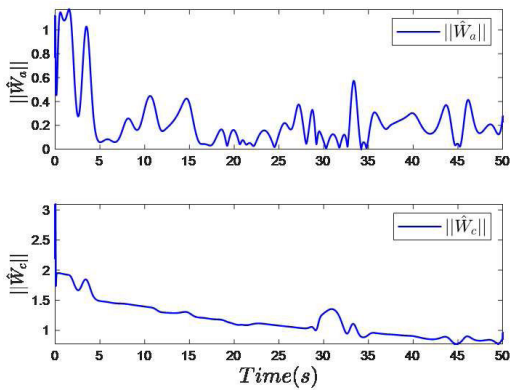


FIGURE 5. The weights of the actor network and the critic network under Case 1.

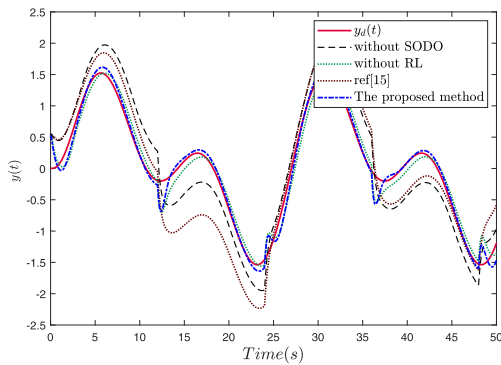


FIGURE 6. The tracking performance for the desired signal under Case 2.

For the proposed control method, the control gain is $K_1 = 3$, $K_2 = 10$ and the adaptive parameters are $c_{W_a} = 5$, $c_{W_c} = 5$, $c_d = 0.5$, $c_M = 0.5$ and $\sigma_a = 3$, $\sigma_c = 0.003$, $\sigma_d = 0.1$, $\sigma_M = 2$.

In Case 1, the matched and mismatched disturbances are set as trapezoidal disturbances those varying with time. The simulation results are shown in Fig 2 - Fig 5. The simulation results show that the proposed control method can achieve satisfactory results under the condition of actuator failure and constant or changing external disturbance. All signals in the

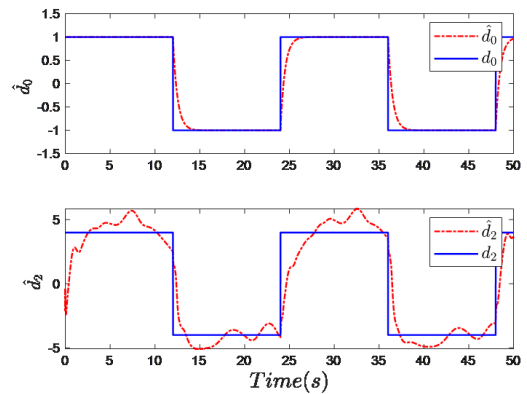


FIGURE 7. The estimation performance for the matched and mismatched disturbances \hat{d}_0 and \hat{d}_2 under Case 2.

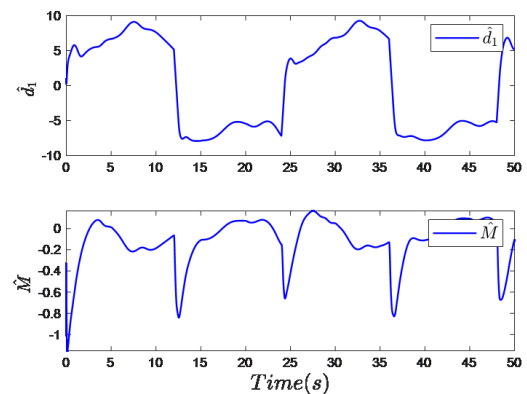


FIGURE 8. The adaptive parameters under Case 2.

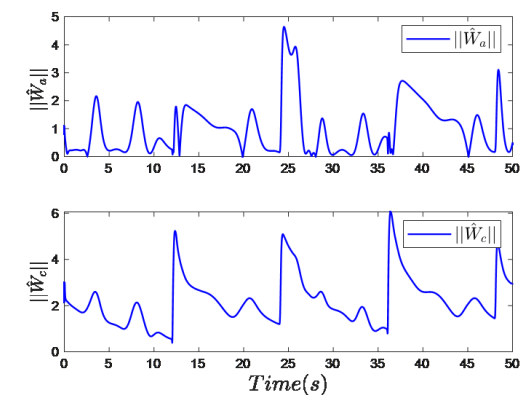


FIGURE 9. The weights of the actor network and the critic network under Case 2.

closed-loop control system are bounded during the whole control process.

In Case 2, the matched and mismatched disturbances are set as square waves those varying with time. The simulation results are shown in Fig 6 - Fig 9. It is obvious that the proposed reinforcement learning- anti-disturbance fault-tolerant control method can still guarantee stable tracking, while the

control methods without SODO or reinforcement learning may produce un-desired tracking errors and time delay. From the simulation results of the two cases, the effectiveness and the advantages of the proposed reinforcement learning- anti-disturbance fault-tolerant control method can be verified.

V. CONCLUSION

This paper addressed the reinforcement learning control problem for the nonlinear uncertain systems with matched and mismatched time-varying disturbances, as well as the unknown perturbations of the control effectiveness. Two SODOs have been designed for the concerned non-linear uncertain system, estimating and compensating the time varying matched and mismatched disturbances. Two LSTM networks those possesses perfect fitting ability have been utilized as the critic and actor networks, improving the adaptivity with respect to the system un-certainties. Then by designing several fault tolerant adaptive laws, the reinforcement learning anti-disturbance fault tolerant control structure which can handle the matched and mismatched time-varying disturbances, has been established. Two cases of simulation have been performed in this paper, and the advantages of the proposed control structure can be known.

REFERENCES

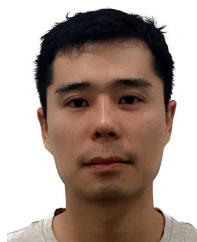
- [1] Y. Li, Z. Ma, and S. Tong, "Adaptive fuzzy fault-tolerant control of nontriangular structure nonlinear systems with error constraint," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 4, pp. 2062–2074, Aug. 2018.
- [2] Y.-J. Liu, S. Lu, S. Tong, X. Chen, C. L. P. Chen, and D.-J. Li, "Adaptive control-based barrier Lyapunov functions for a class of stochastic nonlinear systems with full state constraints," *Automatica*, vol. 87, pp. 83–93, Jan. 2018.
- [3] D. Yu, C. L. P. Chen, and H. Xu, "Fuzzy swarm control based on sliding-mode strategy with self-organized omnidirectional mobile robots system," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, Jan. 20, 2021, doi: 10.1109/TSMC.2020.3048733.
- [4] Y.-M. Li, X. Min, and S. Tong, "Adaptive fuzzy inverse optimal control for uncertain strict-feedback nonlinear systems," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 10, pp. 2363–2374, Oct. 2020.
- [5] Y. Li, T. Yang, and S. Tong, "Adaptive neural networks finite-time optimal control for a class of nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 11, pp. 4451–4460, Nov. 2020.
- [6] A. Isidori, *Nonlinear Control Systems*. New York, NY, USA: Springer, 1995.
- [7] I. V. Utkin, *Sliding Modes in Control and Optimization || Eigenvalue Allocation*. Berlin, Germany: Springer, 1992, ch. 7, pp. 91–110, doi: 10.1007/978-3-642-84379-2.
- [8] K. Umemoto, T. Endo, F. Matsuno, and T. Egami, "Stability analysis of a control system with nonlinear input uncertainty based on disturbance observer," *Int. J. Robust Nonlinear Control*, vol. 30, no. 11, pp. 4433–4448, Jul. 2020.
- [9] G. Rigatos, N. Zervos, M. Abbaszadeh, J. Pomares, and P. Wira, "Non-linear optimal control for multi-DOF electro-hydraulic robotic manipulators," *IET Cyber-Syst. Robot.*, vol. 2, no. 2, pp. 96–106, 2020.
- [10] M. Ahmadi and M. Haeri, "An integrated best-worst decomposition approach of nonlinear systems using gap metric and stability margin," *Proc. Inst. Mech. Eng. I, J. Syst. Control Eng.*, vol. 235, no. 4, pp. 486–502, Apr. 2021.
- [11] R. G. Guerra, J. E. V. Velázquez, L. Fridman, and R. Iriarte, "Robust output trajectory linearisation control for a class of linear time-varying systems," *IET Control Theory Appl.*, vol. 15, no. 6, pp. 877–889, Apr. 2021.
- [12] X. Xia, G. Kang, T. Zhang, and Y. Fang, "Adaptive quantised control of uncertain non-linear systems with state constraints and time-varying delays," *IET Control Theory Appl.*, vol. 14, no. 10, pp. 1308–1320, 2020.
- [13] X. Shan, H. Song, H. Cao, L. Zhang, X. Zhao, and J. Fan, "A dynamic hysteresis model and nonlinear control system for a structure-integrated piezoelectric sensor-actuator," *Sensors*, vol. 21, no. 1, p. 269, Jan. 2021.
- [14] D. B. Zhao, K. Shao, Y. H. Zhu, D. Li, and C. H. Wang, "Review of deep reinforcement learning and discussions on the development of computer Go," *Control Theory Appl.*, vol. 33, no. 6, pp. 701–717, 2016.
- [15] X. Chang, L. Rong, K. Chen, and W. Fu, "LSTM-based output-constrained adaptive fault-tolerant control for fixed-wing UAV with high dynamic disturbances and actuator faults," *Math. Problems Eng.*, vol. 2021, pp. 1–18, Mar. 2021.
- [16] M. B. Radac and R. E. Precup, "Data-driven model-free tracking reinforcement learning control with VRFT-based adaptive actor-critic," *Appl. Sci.*, vol. 9, no. 9, p. 1807, Apr. 2019.
- [17] S. Choi, S. Kim, and H. J. Kim, "Inverse reinforcement learning control for trajectory tracking of a multirotor UAV," *Int. J. Control, Autom. Syst.*, vol. 15, no. 4, pp. 1826–1834, Aug. 2017.
- [18] S. P. Nagesh Rao, G. A. D. Lopes, D. Jeltsema, and R. Babuška, "Passivity-based reinforcement learning control of a 2-DOF manipulator arm," *Mechatronics*, vol. 24, no. 8, pp. 1001–1007, Dec. 2014.
- [19] E. Anderlini, S. Husain, G. G. Parker, M. Abusara, and G. Thomas, "Towards real-time reinforcement learning control of a wave energy converter," *J. Mar. Sci. Eng.*, vol. 8, no. 11, p. 845, Oct. 2020.
- [20] J. Valasek, J. Doebbler, M. D. Tandale, and A. J. Meade, "Improved adaptive-reinforcement learning control for morphing unmanned air vehicles," *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 38, no. 4, pp. 1014–1020, Jun. 2008.
- [21] Z. Zheng, L. Ruan, M. Zhu, and X. Guo, "Reinforcement learning control for underactuated surface vessel with output error constraints and uncertainties," *Neurocomputing*, vol. 399, pp. 479–490, Jul. 2020.
- [22] Q. Wu, Y. Gu, Y. Li, B. Zhang, S. A. Chepinskiy, J. Wang, A. A. Zhilenkov, A. Y. Krasnov, and S. Chernyi, "Position control of cable-driven robotic soft arm based on deep reinforcement learning," *Information*, vol. 11, no. 6, p. 310, Jun. 2020.
- [23] H. Takemi and M. Yokoyama, "1414 reinforcement learning control of a vehicle robot with variable center of gravity," in *Proc. Conf. Hokuriku-Shinetsu Branch*, vol. 49, 2012, pp. 141401–141402.
- [24] D. J. Pradeep and M. M. Noel, "A finite horizon Markov decision process based reinforcement learning control of a rapid thermal processing system," *J. Process Control*, vol. 68, pp. 218–225, Aug. 2018.
- [25] T. Becsi, A. Szabo, B. Kovari, S. Aradi, and P. Gaspar, "Reinforcement learning based control design for a floating piston pneumatic gearbox actuator," *IEEE Access*, vol. 8, pp. 147295–147312, 2020.
- [26] S. Adam, L. Busoniu, and R. Babuska, "Experience replay for real-time reinforcement learning control," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 2, pp. 201–212, Mar. 2012.
- [27] S. Liu and G. P. Henze, "Evaluation of reinforcement learning for optimal control of building active and passive thermal storage inventory," *J. Sol. Energy Eng.*, vol. 129, no. 2, pp. 215–225, May 2007.
- [28] L. Matignon, G. J. Laurent, N. Le Fort-Piat, and Y.-A. Chapuis, "Designing decentralized controllers for distributed-air-jet MEMS-based micromanipulators by reinforcement learning," *J. Intell. Robotic Syst.*, vol. 59, no. 2, pp. 145–166, Aug. 2010.
- [29] D. M. Katić and A. D. Rodić, "Policy gradient fuzzy reinforcement learning control of humanoid walking," *IFAC Proc. Volumes*, vol. 42, no. 19, pp. 98–103, 2009.
- [30] S. H. Cho, "Application study of reinforcement learning control for building HVAC system," *Int. J. Air-Conditioning Refrig.*, vol. 14, no. 4, pp. 138–146, 2006.



SHIYI HUANG received the bachelor's degree in major of software engineering from Nanchang Hangkong University, Nanchang, China, in 2019. He is currently pursuing the master's degree with East China Jiaotong University. His current research interests include intelligent control of unmanned aerial vehicle systems, reinforcement learning control and design, and research of complex neural networks.



ZHENG WANG received the B.E., M.E., and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2013, 2016, and 2020, respectively. He is currently an Associate Professor and a Master Supervisor with the Research Center for Unmanned System Strategy Development, Northwestern Polytechnical University. His current research interests include flight dynamics and control, intelligent decision, and autonomous control of the unmanned systems.



KANG CHEN received the B.E., M.E., and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2001, 2004, and 2010, respectively. He is currently an Associate Professor with the Unmanned System Research Institute, Northwestern Polytechnical University. His current research interests include flight control system design and simulation.



ZHAOHUI YUAN received the B.S. degree in computer science from Huazhong Normal University, China, in 2004, and the Ph.D. degree in computer science from Wuhan University, China, in 2009. He is currently an Associate Professor with East China Jiaotong University, China. His research interests include signal processing, wireless sensor networks, and mobile computing.



TAO LI received the M.E. degree in flight vehicle design from the China Academy of Launch Vehicle Technology, Beijing, China, in 2009. He is currently a Senior Engineer with the Design and Simulation of Spacecraft Control System Group, Beijing Institute of Space Long March Vehicle. His current research interests include flight dynamics, guidance, navigation, and control and hardware-in-the-loop simulation of spacecraft.

...