Proceedings Paper:
Tang, Z., Rossiter, J.A., Dong, Y. et al. (2024) Reinforcement learning-based output stabilization control for nonlinear systems with generalized disturbances. In: 2024 IEEE International Conference on Industrial Technology (ICIT). 2024 IEEE International Conference on Industrial Technology (ICIT), 25-27 Mar 2024, Bristol, United Kingdom. Institute of Electrical and Electronics Engineers, pp. 1-6. ISBN: 979-8-3503-4026-6. ISSN: 2641-0184. EISSN: 2643-2978.

https://doi.org/10.1109/icit58233.2024.10540609

Reinforcement Learning-Based Output Stabilization Control for Nonlinear Systems With Generalized Disturbances

Zezhi Tang*, J Anthony Rossiter, Yi Dong, George Panoutsos

***Abstract*—This paper proposes a new disturbance observer (DO)-based reinforcement learning (RL) control approach for nonlinear systems with unmatched (generalized) disturbances. While a nonlinear disturbance observer (NDO) is utilized to measure the plant uncertainties, disturbances can exist in the plant via distinct channels from those of the control signals; so-called mismatched disturbances are theoretically difficult to attenuate within the channel of the system's states. A generalized disturbance observer-based compensator is implemented to address the uncertainty cancellation problem by removing the influence of uncertainties from the output channels. Concurrently, a composite actor-critic RL scheme is utilized for approximating the optimal control policy as well as the ideal value function pertaining to the compensated system by solving a Hamilton-Jacobi-Bellman (HJB) equation for both online and offline iterations simultaneously. Stability analysis verifies the convergence of the proposed framework. Simulation results are included to illustrate the effectiveness of the proposed scheme.**

I. Introduction

Reinforcement learning (RL), as a goal-directed computational method, has gained significant attention in recent years considering its ability to address intricate decision-making problems across various domains [1], [2]. In the realm of controller design, RL offers a promising alternative approach for updating control policies in complex and dynamic systems, particularly in scenarios where traditional control methods may be unsuitable [3]. Recent studies have revealed a notable surge in the utilization of RL as an effective approach for addressing complex sequential decision-making problems in the field of optimal controller design [4].

Originally, the emphasis on developing RL-based control systems was primarily on linear systems due to their relatively simpler nature. A composite iterative learning controller is applied to optimize the tracking performance of the magnetic bearing in [5]. In addition, recent developments in RL have enabled its extension to nonlinear systems featuring high-dimensional state and action spaces. Robustness is crucial in designing nonlinear systems for consistent performance under varying conditions [6]. The composite RL scheme designed for robust control under worst-case uncertainties is presented in [7]. Integration of H_∞ control and RL schemes has been demonstrated in [8]. In order to ensure an adequate variety in the agent's actions, which covers an extensive range of states, the so-called persistence of excitation (PE) condition is generally employed [9], [10]. However, the realization of the PE condition requires the incorporation of probing signals into the control inputs, consequently causing a compromise in transient performance.

Disturbances are pervasive problems in control systems, as they lead to degraded performance, instability, and even system failure. In real applications, disturbances can be attributed to diverse sources, ranging from external environmental factors to uncertainties arising from unmatched models, nonlinear coupling, and parameter perturbations [11], [12]. Some control approaches utilize feedback control to eliminate disturbances rather than employing feedforward compensation. In such cases, the methodology behind those approaches is to reject the uncertainties via the error between measured output and desired setpoints [13]. In order to mitigate the impact of disturbances, a feedforward compensation scheme based on estimations of the uncertainties can be utilized to counteract their effect in advance, thereby enhancing the overall robustness of the system [11], [14]. In practical implementations, the presence of lumped uncertainties within the plant via distinct channels to input signals is a common occurrence. These uncertainties, often referred to as mismatched or unmatched disturbances, present significant difficulties in direct compensation. Consequently, achieving asymptotic stability in the presence of such mismatched uncertainties poses a considerable challenge in the control design [15].

To address the aforementioned challenge, we demonstrate a nonlinear disturbance observer (NDO)-based reinforcement learning control scheme. The proposed approach employs the actor-critic framework to compute the optimal solution to a Hamilton-Jacobi-Bellman (HJB) equation, which is obtained from the underlying nonlinear system dynamics. In addition, by introducing a disturbance attenuation scheme that serves to counteract mismatched uncertainties in the output channel, we augment the disturbance compensation ability as well as enhance the system's robustness. In this paper, the following three principal contributions have been outlined:

1) An RL-based optimal control strategy is designed, while the actor-critic-observer framework is employed to approximate the optimal solution of the derived HJB equation.
2) The robustness against a broader array of uncertainties is enhanced by asymptotically canceling uncertainty impacts in the output channel, without adjusting the structure of the nominal RL-based controllers.

Z. Tang, J.A. Rossiter and G. Panoutsos are with the Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, United Kingdom (emails: zezhi.tang@sheffield.ac.uk, j.a.rossiter@sheffield.ac.uk, g.panoutsos@sheffield.ac.uk)

Y. Dong is with the Department of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom (email: yi.dong@soton.ac.uk)

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC, grant no. EP/V051261/1).

*Corresponding author

The remainder of the paper is organized as follows: Section II introduces the formulation of the control problem. Section III presents the actor-critic RL methodology, including the associated update laws. Subsequently, Section IV proposes a composite control architecture, incorporating the proposed NDO and corresponding compensator. A rigorous stability analysis of the integrated control scheme is delivered in Section V. In Section VI, simulation results are illustrated to verify the correctness of the presented scheme. Finally, Section VII provides the conclusion and outlines future work.

II. Closed-loop optimal control problem formulation

Suppose a nonlinear system is given by

$$\begin{cases} \dot{x} = f(x) + g_u(x)u + g_d(x)\omega, \\ y_o = C_o x, \end{cases} \tag{1}$$

where $x \in \mathbb{R}^n, \omega \in \mathbb{R}, u \in \mathbb{R}, y_o \in \mathbb{R}$ refer to the system states, total uncertainties, control input and controlled output, respectively. $f(x) \in \mathbb{R}^n, g_u(x) \in \mathbb{R}^n$ and $g_d(x) \in \mathbb{R}^n$ are nonlinear functions supposed to be continuously differentiable, while $C_o \in \mathbb{R}^{1\times n}$ represents the output matrix.

A standard optimal control approach can be employed to achieve system stability by formulating the cost function as follows [16]:

$$J = \frac{1}{2}\int_0^{+\infty} (Q(x) + u_c^{\mathrm{T}} R u_c) dt, \tag{2}$$

where $u_c \in \mathbb{R}$ denotes the approximated optimal control policy derived from subsequent RL algorithms; Consider $Q(x) = x^{\mathrm{T}} Q_c x \in \mathbb{R}$, where $Q_c \in \mathbb{R}^{n\times n}$ is the positive symmetric weighting matrix, while $R \in \mathbb{R}$ is positive.

Suppose the output has been given in (1). The performance index for the system output can be rewritten as [17]:

$$J_y = \frac{1}{2}\int_0^{+\infty} (y_o^{\mathrm{T}} \overline{Q} y_o + u_c^{\mathrm{T}} R u_c) dt, \tag{3}$$

where $\overline{Q} \in \mathbb{R}$ and $Q_c = C_o^{\mathrm{T}} \overline{Q} C_o$. A universal suggestion for choosing matrices Q_c and R is given in [18], which enables the definition of the optimal output performance index from (2) in a convenient manner.

Based on the (3), according to [19], the Hamiltonian $H \in \mathbb{R}$ of the system (1) can be derived as:

$$\begin{aligned} H(x, u_c, W_q) \triangleq & W_q(x)(f(x) + g_u(x)u_c \\ & + g_d(x)\omega) + Q(x) \\ & + {u_c}^{\mathrm{T}}(x) R u_c(x), \end{aligned} \tag{4}$$

where $W_q = \frac{\partial W}{\partial x}$ refers to the gradient of the value function $W \in \mathbb{R}^n$. An optimal value function with corresponding control input that minimizes the total performance index (2), can be specified as W^* and u_c^* respectively. The HJB function is deduced as:

$$\begin{aligned} H(x, u_c^*, W_q^*) = & W_q^*(x)(f(x) + g_u(x)u_c^* \\ & + g_d(x)\omega) + Q(x) \\ & + {u_c^*}^{\mathrm{T}}(x) R u_c^*(x) = 0. \end{aligned} \tag{5}$$

For the optimal control policy, one can demonstrate a closed-form expression as:

$$u_c^*(x) = -\frac{1}{2} R^{-1} g_u^{\mathrm{T}} {W_q^*}^{\mathrm{T}}(x). \tag{6}$$

By applying an approximation of the optimal control input $\widehat{u}_c(x)$, estimation of system states $\widehat{x}$, and optimal value function $\widehat{W}(x)$ to the (5), the Hamiltonian is expressed as $H(\widehat{x}, \widehat{u}_c, \widehat{W}_q)$.

Compared with the HJB function (5), the instantaneous Bellman error $\sigma_{hjb} \in \mathbb{R}$ can be written as:

$$\begin{aligned} \sigma_{hjb} \triangleq & H(\widehat{x}, \widehat{u}_c, \widehat{W}_q) - H(x, u_c^*, W_q^*) \\ = & H(\widehat{x}, \widehat{u}_c, \widehat{W}_q) \\ = & \widehat{W}_q(x)(f(\widehat{x}) + g_u(\widehat{x})\widehat{u}_c + g_d(x)\omega) \\ & + Q(x) + {\widehat{u}_c}^{\mathrm{T}}(x) R \widehat{u}_c(x). \end{aligned} \tag{7}$$

III. Actor-critic reinforcement learning design

For the aforementioned HJB equation, it is generally difficult to derive its solution due to its nonlinear nature [20]. Therefore, we employ a novel actor-critic-based RL algorithm to approximate an optimized solution by updating weights in the constructed NNs.

A. Approximated Bellman Error (BE) Derivation

Using the closed-loop expression (6), we can explicitly demonstrate the control policies and their corresponding value functions as follows:

$$W^*(x) = K^{\mathrm{T}}\theta(x) + \tau(x), \tag{8}$$

$$u_c^*(x) = -\frac{1}{2} R^{-1} g_u^{\mathrm{T}} \left(\theta_x^{\mathrm{T}}(x) K + \tau_x^{\mathrm{T}}(x)\right), \tag{9}$$

The optimized NN weight is denoted by $K \in \mathbb{R}^m$, where $m \in \mathbb{N}$ refers to the number of neurons. $\theta \in \mathbb{R}^m$ refers to the activation function basis, and $\theta_x = \frac{\partial \theta}{\partial x}$ refers to its gradient. The $\tau \in \mathbb{R}$ refers to the reconstruction error in the updating process and $\tau x = \frac{\partial \tau}{\partial x}$ refers to its gradient.

We can derive the NN-based representation for control policy $\widehat{u}_c\left(\widehat{x}, \widehat{K}_c\right) = -\frac{1}{2} R^{-1}(\widehat{x}) g_u^{\mathrm{T}}(x) \theta_x^{\mathrm{T}}(\widehat{x}) \widehat{K}_c$, and corresponding approximated value function $\widehat{W}\left(\widehat{x}, \widehat{K}_v\right) = \widehat{K}_v^{\mathrm{T}} \theta(\widehat{x})$, where $\widehat{K}_v \in \mathbb{R}^m$ and $\widehat{K}_c \in \mathbb{R}^m$ represent the optimal critic and actor weights estimations, respectively.

It is proved in [21] that the approximation error $\tau \to 0$ when the number of neurons $m \to \infty$, therefore a reformulation of the approximated instantaneous BE can be presented (7) as follows:

$$\begin{aligned} \sigma_{hjb} \triangleq & \widehat{W}_q\left(\widehat{x}, \widehat{K}_v\right)\left[f(\widehat{x}) + g_u(\widehat{x})\widehat{u}_c\left(\widehat{x}, \widehat{K}_c\right)\right] \\ & + Q(\widehat{x}) + \widehat{u}_c^{\mathrm{T}}\left(\widehat{x}, \widehat{K}_c\right) R \widehat{u}_c\left(\widehat{x}, \widehat{K}_c\right). \end{aligned} \tag{10}$$

Meanwhile, the approximation of the BE $\sigma_i \in \mathbb{R}$ at setpoint $x^i \in \mathbb{R}^n$ can be calculated utilizing the information acquired from the proposed observer.

B. Critic NN Update Law

Based on the proposed actor-critic architecture, we employ a least square (LS) law to update the weights of critic NN. The fundamental estimation method of weights can be found in [22]. A recursive formulation of the LS approach is employed by utilizing derivatives of representation for the LS critic weight estimate, which is presented in [23] as:

$$\begin{aligned}\dot{\widehat{K}}_v =& -\rho_{v1}\Theta\frac{\xi}{1+\lambda\xi^{\mathrm{T}}\Theta\xi}\sigma_{hjb}\\ &-\frac{\rho_{v2}}{N}\Theta\sum_{i=1}^{N}\frac{\xi_i}{1+\lambda\xi_i^{\mathrm{T}}\Theta\xi_i}\sigma_i,\end{aligned} \tag{11}$$

where λ, ρ_{v1} and ρ_{v2} are the constant gains. $\Theta \in \mathbb{R}^{m\times m}$, refers to the gain matrix and is given by:

$$\dot{\Theta} = \phi\Theta - \rho_{v1}\frac{\Theta\xi\xi^{\mathrm{T}}\Theta}{(1+\lambda\xi^{\mathrm{T}}\Theta\xi)^2}, \|\Theta(0)\| \le \eta_1, \tag{12}$$

of which $\xi = \theta_x(\widehat{x})\left[f(\widehat{x}) + g_u\widehat{u}_c\left(\widehat{x},\widehat{K}_c\right)\right]$ and $\xi_i = \theta_x\left(x^i\right)\left[f(x^i) + g_u\widehat{u}_c\left(x^i,\widehat{K}_c\right)\right]$ and ϕ is the positive forgetting parameter.

The setting of Θ is carefully considered to ensure that it is positive definite and its value is not excessively small in certain directions, as discussed in [23]. It can be concluded that the inequality $\eta_0 I \le \Theta \le \eta_1 I, t \ge 0$ holds since $\dot{\Theta} \le 0$, and η_0 and η_1 are positive constants.

C. Actor NN Update Law

The minimization of BE in the actor's NN update process bears similarity to that of a critic's NN. However, due to the nonlinearity of the actor's NN, the LS method is unsuitable for use here. To address this issue, a gradient update method for minimizing the squared BE has been introduced in [21], and the NN update law is presented as:

$$\begin{aligned}\dot{\widehat{K}}_c =&\rho_{c2}\widehat{K}_v + \frac{\rho_{v1}P_t^{\mathrm{T}}\widehat{K}_c\xi^{\mathrm{T}}}{4(1+\lambda\xi^{\mathrm{T}}\Theta\xi)}\widehat{K}_v - \widehat{K}_c\left(\rho_{c1}+\rho_{c2}\right)\\ &+\sum_{i=1}^{N}\frac{\rho_{v2}P_i^{\mathrm{T}}\widehat{K}_c\xi_i^{\mathrm{T}}}{4N(1+\lambda\xi_i^{\mathrm{T}}\Theta\xi_i)}\widehat{K}_v,\end{aligned} \tag{13}$$

of which $P_t = \theta_x(\widehat{x})g_u(\widehat{x})R^{-1}g_u^{\mathrm{T}}(\widehat{x})\theta_x^{\mathrm{T}}(\widehat{x})$ and $P_i = \theta_x\left(x^i\right)g_u(x^i)R^{-1}g_u^{\mathrm{T}}(x^i)\theta_x^{\mathrm{T}}\left(x^i\right)$. Similar to the critic NN, the variables ρ_{c1} and ρ_{c2} in (13) are positive constants.

IV. Universal nonlinear disturbance observer based reinforcement learning control

A. Nonlinear disturbance observer design

From the aforementioned analysis, the estimation of lumped uncertainties is needed for constructing the actor-critic RL framework.

Suppose an NDO is proposed for estimating total uncertainties with the following model [24]:

$$\begin{cases}\widehat{\omega} = z + r(x),\\ \dot{z} = -l(x)g_d(x)z - l(x)\left[g_d(x)r(x) + f(x) + g_u(x)u\right],\end{cases} \tag{14}$$

of which $\widehat{\omega}$ and z represent the estimation of disturbances proposed in (1), and respective intermediate states for the proposed NDO. $r(x)$ refers to the nonlinear observer function to be developed. The observer gain $l(x)$ can be designed as $l(x) = \frac{\partial r(x)}{\partial x}$. The internal states can be easily estimated using a nonlinear observer according to [11].

Assumption 1: The disturbance ω and its derivative $\dot{\omega}$ are bounded, i.e., $||\omega||_2 \le H_1$,$||\dot{\omega}||_2 \le H_2$, where H_1, H_2 are positive constants.

Assumption 2: Lumped disturbances ω are slowly varying relative to the proposed observer dynamics.

Given that Assumption 2 is satisfied, $\widehat{\omega}$ converges to ω asymptotically when $l(x)$ is chosen to guarantee the stability of the following estimation error system:

$$\dot{e}(t) = -l(x)g_d(x)e(t), \tag{15}$$

applicable for any $x \in R^n$. Here $e = \omega - \widehat{\omega}$ denotes the estimation error. Consequently, the developed function $l(x)$, which ensures the asymptotic stability of (15) can be easily selected. [11].

B. Nonlinear disturbance observer-based control architecture

A composite control law of the optimal output stabilization control is derived as:

$$u = u_c + u_d\widehat{\omega}, \tag{16}$$

of which u_d is the compensation input to be designed. It should be noted that the compensator component solely improves the disturbance rejection capability of the composite approach. Additionally, in the absence of lumped uncertainties, the system's global asymptotic stability is maintained with the RL input.

Assuming the existence of a nonlinear function $\bar{f}(x)$ satisfying $f(x) = \bar{f}(x)x$ and $\overline{u}_c x = u_c$, to obtain the appropriate compensator gain, we assume there exists a nonlinear function $\bar{f}(x)$ satisfy $f(x) = \bar{f}(x)x$ and $\overline{u}_c x = u_c$. Therefore the the system given by equation (1) can be represented as:

$$\begin{cases}\dot{x} = \bar{f}(x)x + g_u(x)u + g_d(x)\omega,\\ y_o = C_o x,\end{cases} \tag{17}$$

therefore the composite controller can be demonstrated as:

$$u = \overline{u}_c x + u_d\widehat{\omega}, \tag{18}$$

where

$$\begin{aligned}u_d =& -\left[C_o(\bar{f}(x) + g_u(x)\overline{u}_c)^{-1}g_u(x)\right]^{-1} \times C_o(\bar{f}(x)\\ &+ g_u(x)\overline{u}_c)^{-1}g_d(x).\end{aligned} \tag{19}$$

Based on the compensation gain proposed in (19), a generalized procedure for designing the nonlinear disturbance observer-based control (NDOBC) can be shown as follows [11]:

1) Confirm the closed-loop stability for the system in the absence of uncertainties, with the baseline controller.
2) Design the disturbance observer with respective lumped uncertainties.

3) Develop the appropriate disturbance compensation gain and construct the composite NDOBC architecture by integrating the DO-based compensator with the baseline control scheme.

C. System Analysis

The system analysis will be shown in two parts in this section. The rigorous stability proof is not shown in this paper due to the page limit. The main proof idea is to first demonstrate the effectiveness of the RL system without considering the disturbance observer-based compensator, and a Lyapunov function exists which provides the stability criterion if the number of neurons is enough to satisfy the proposed assumptions. Then the output compensator analysis will be presented to demonstrate the disturbance rejection ability of the compensation scheme in the output channel, with closed-loop stability guaranteed.

In the first part of the analysis, we will focus on the analysis of the disturbance-free system. Assumptions below are proposed in [21] to relax the PE condition.

Assumption 3: A group of offline points $\left\{x^i \in \mathbb{R}^n \mid i = 1, \ldots, N\right\}$ exist to assure the following holds

$$j \triangleq \frac{1}{N} \inf_{t \geq 0} \left(\sigma_{eig} \left\{ \sum_{i=1}^{N} \frac{\xi_i \xi_i^{\mathrm{T}}}{1 + \lambda \xi_i^{\mathrm{T}} \Theta \xi_i} \right\} \right) > 0, \tag{20}$$

of which the σ_{eig} is a minimum eigenvalue of the corresponding matrix [21].

The weight estimation error for the actor and critic NN weights are presented as:

$$\widetilde{e}_c = \widehat{K}_c - K_c, \widetilde{e}_v = \widehat{K}_v - K_v, \tag{21}$$

The variable K_c represents the estimated weight of K for the actor NN, while K_v represents the estimated weight of K for the critic NN. The convergence of the two values to a shared value has been demonstrated in [21].

Theorem 1: Consider the closed-loop system consists of the nonlinear system (1), controller (16), update law (11) and (13). The estimation error of the weight is uniformly ultimately bounded (UUB) if: i) Assumption 3 holds and ii) the effect of the uncertainties is attenuated in the output channel.

Now we are going to analyze the system stability as well as the disturbance compensation ability of the complete architecture. First of all, an analysis of the system's input-to-state stability (ISS) is provided with the proof below.

Theorem 2: An NDOBC architecture composed of the system plant (1), NDO (14) and composite controller (16) is input-to-state stable if: i) system plant (1) with baseline controller u_c is globally asymptotically stable without considering the influence of uncertainties; ii) the observer function is selected appropriately to ensure the error is globally asymptotically stable and iii) the compensation function is chosen appropriately to ensure that $P_d(x) = g_d(x) + g_u(x)u_d$ is continuously differentiable.

A subsequent closed-loop system is developed by combining the system equation, control signals and estimation error representation, a subsequent representation of the system in the presence of control signals is demonstrated as:

$$\begin{cases} \dot{x} = [f(x) + g_u(x)u_c] - g_u(x)u_d e \\ \qquad + [g_d(x) + g_u(x)u_d]\,\omega, \\ \dot{e} = -\dfrac{\partial r(x)}{\partial x} g_d(x) e. \end{cases} \tag{22}$$

with disturbances ω as the input, lumped x and the observer states e as new states for the reconstructed system. Suppose a augmented state $\overline{x}_d = \left[\ x, e\ \right]^{\mathrm{T}}$, therefore

$$F(\overline{x}_d) = \begin{bmatrix} f(x) + g_u(x)u_c - g_u(x)u_d e \\ -l(x)g_d(x)e \end{bmatrix}. \tag{23}$$

Combining (22) with (23), a reformulated system is represented as:

$$\dot{\overline{x}}_d = F(\overline{x}_d) + \begin{bmatrix} P_d(x) \\ 0 \end{bmatrix} \omega. \tag{24}$$

The model $\dot{\overline{x}}_d = F(\overline{x}_d)$ demonstrates asymptotically stability as the first two conditions given in Theorem 2 are satisfied. The disturbance-free system stability in the presence of the baseline control input is demonstrated in Theorem 1.

Now we will give the analysis of the uncertainty attenuation ability of the composite disturbance compensation system.

Theorem 3: For the system (1) under NDO (14) and control framework which is composed of the RL-based controller and compensation design (16), the effect of unmatched uncertainties is removed from the output, if uncertainty compensation function u_d is appropriately designed to ensure that the system (22) is ISS, while (19) holds.

Consider the system (22), together with (18) and (17), we can represent the states as:

$$\begin{aligned} x = [\bar{f}(x) + g_u(x)\overline{u}_c]^{-1} \{ & \dot{x} - g_u(x)u_d e \\ & - [g_u(x)u_d + g_d(x)]\,\omega \}. \end{aligned} \tag{25}$$

Combining (25) with compensation gain (19), and rewriting the system (17) gives:

$$\begin{aligned} y = & C_o [\bar{f}(x) + g_u(x)\overline{u}_c]^{-1} \dot{x} \\ & + C_o \left[\bar{f}(x) + g_u(x)\overline{u}_c\right]^{-1} g_d(x) e. \end{aligned} \tag{26}$$

With the convergence of the system, it can be demonstrated that the effect of uncertainties is removed from the output, as the estimation error and derivatives of the system state exhibit convergence.

V. Simulation results

To substantiate the proposed NDOBC-RL scheme, simulation examples are demonstrated based on a nonlinear system under the presence of matched and unmatched lumped uncertainties.

A nonlinear system subject to a constant disturbance is proposed in (27). Both the disturbance and control signals are transmitted through the same channel. The corresponding optimal control solution for this system has been presented

in [21] to verify the validity of the RL-based approximation scheme.

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 - 1.5x_2 + 0.5\left(x_1 + x_2\right)\left(\cos\left(2x_1\right)\right. \\ \qquad \left. + 2\right)^2 + \left(\cos\left(2x_1\right) + 2\right)u + \omega. \end{cases} \tag{27}$$

The control output can be represented as follows:

$$y_o = 10x_1 + x_2, \tag{28}$$

Based on the analysis in (3), matrices are chosen as $\overline{Q} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$, according to [21], to be positive definite. Subsequently, the optimal output performance index can be derived using (2). These parameters were selected to simplify the difficulty of the verification process since optimal solutions under a specific basis are provided by analytical methods.

Initial parameter settings in this section are chosen based on prior research [21]. The complete offline set is defined as $\mathcal{X} = [-3 \quad 3] \times [-3 \quad 3]$, while the DO gain is chosen as $l = [5 \quad 5]^{\mathrm{T}}$. The basis function is selected as $\theta(x) = [x_1^2 \quad x_1x_2 \quad x_2^2]^{\mathrm{T}}$, with initial values set to 0.5. The RL controller gains are selected as $\rho_{v1} = 1, \rho_{v2} = 5, \rho_{c1} = 100, \rho_{c2} = 0.1$, $\lambda = 0.5$, and $\phi = 100$, according to [21].

Remark 1: In this paper, the optimized basis function is selected to be consistent with the analytical results from [25], allowing for neuron weight comparisons. Nevertheless, determining a suitable basis function for a general nonlinear system remains a considerable challenge and remains an ongoing area of research.

The nonlinear system proposed in (27) is now extended to include a mismatched sinusoidal disturbance, as presented in (29).

$$\begin{cases} \dot{x}_1 = x_2 + \omega, \\ \dot{x}_2 = -x_1 - 1.5x_2 + 0.5\left(x_1 + x_2\right) \\ \qquad \cdot \left(\cos\left(2x_1\right) + 2\right)^2 + \left(\cos\left(2x_1\right) + 2\right)u, \\ y_m = y_o = 10x_1 + x_2, \end{cases} \tag{29}$$

Illustrative results for both matched and mismatched cases are presented in Figs. 1–5. Specifically, Figs. 1 and 3 demonstrate the effectiveness of the proposed architecture under the matched case described in (27). Fig. 1 shows the trajectories of internal states, lumped disturbances, and system outputs. And it proves the effectiveness of the proposed architecture by stabilizing the system while attenuating the disturbances. Fig. 2 demonstrates the convergence of the weight parameters to their analytical results of 2, 1.5, and 1, respectively. Then for the compensated plant, the optimal control signals u_c derived from the proposed RL algorithm are shown in Fig. 3. Meanwhile, Figs. 4 and 5 illustrate the more generalized results under the mismatched sinusoidal disturbances described in (29).

Mismatched sinusoidal uncertainties that exist in different channels than the control signal are unable to be eliminated theoretically. However, the idea of the proposed algorithm is to attenuate uncertainties from the output channel by designing proper compensation functions. Fig. 4 indicates that the disturbance observer has the ability to track lumped uncertainties effectively. Meanwhile, Fig. 5 shows the trajectories of the composite controlled output; one can see that the convergence of the output is free from the influence of the perturbation.

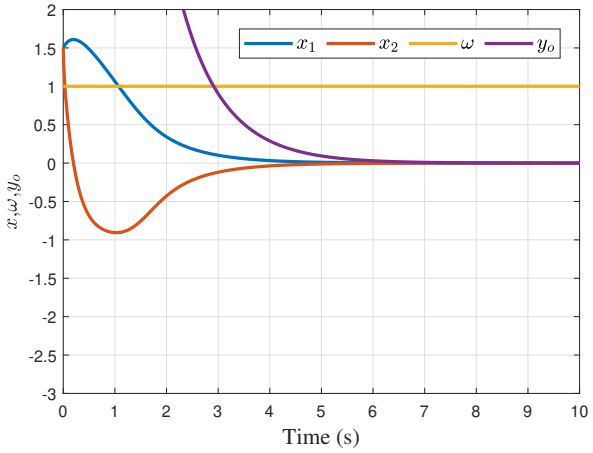

Fig. 1. Trajectories of system states, matched disturbances and controlled output

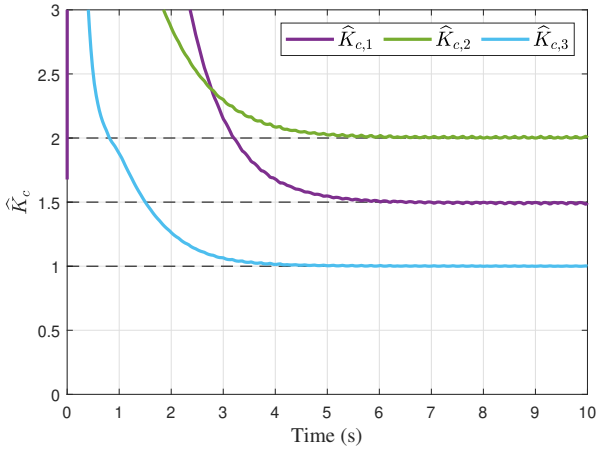

Fig. 2. Convergence of weight parameters

VI. Conclusion

A composite RL-based optimal output stabilization control architecture is demonstrated for generalized nonlinear systems with mismatched disturbances. The significance of the proposed approach lies in the innovative connection between actor-critic methods and uncertain nonlinear systems without requiring matched disturbances. Furthermore, the PE condition, which is commonly assumed in ADP systems, is relaxed by introducing offline data points instead of using probing signals. The simulation results illustrate the effectiveness of the presented method.

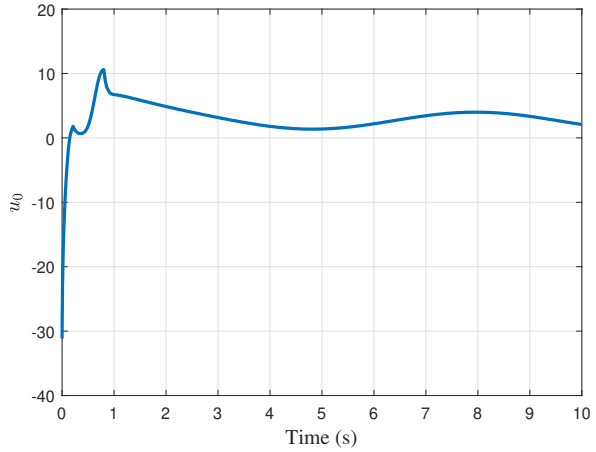

Fig. 3. Trajectories of optimal control policy

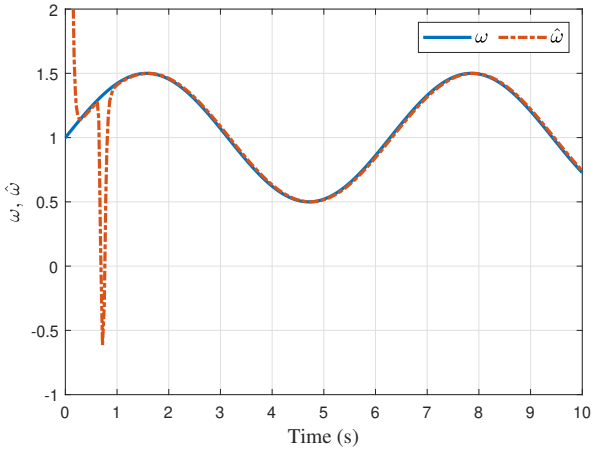

Fig. 4. Lumped disturbances and estimation value from disturbance observer

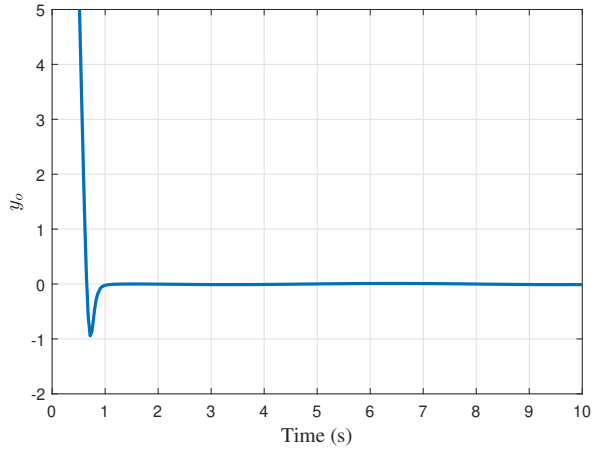

Fig. 5. Convergence of composite controlled output under mismatched disturbances

REFERENCES

[1] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, 2009.

[2] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[3] W. He, H. Gao, C. Zhou, C. Yang, and Z. Li, "Reinforcement learning control of a flexible two-link manipulator: An experimental investigation," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 12, pp. 7326–7336, 2020.

[4] A. Gosavi, "Reinforcement learning: A tutorial survey and recent advances," *INFORMS Journal on Computing*, vol. 21, no. 2, pp. 178–192, 2009.

[5] Z. Tang, Y. Yu, Z. Li, and Z. Ding, "Disturbance rejection via iterative learning control with a disturbance observer for active magnetic bearing systems," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, pp. 131–140, 2019.

[6] H. Zhao, Z. Tang, Z. Li, Y. Dong, Y. Si, M. Lu, and G. Panoutsos, "Real-time object detection and robotic manipulation for agriculture using a yolo-based learning approach," in *2024 IEEE 25th International Conference on Industrial Technology (ICIT)*. IEEE, 2024.

[7] A. Perrusquía and W. Yu, "Continuous-time reinforcement learning for robust control under worst-case uncertainty," *International Journal of Systems Science*, vol. 52, no. 4, pp. 770–784, 2021.

[8] B. Luo, H.-N. Wu, and T. Huang, "Off-policy reinforcement learning for h infinity control design," *IEEE Transactions on Cybernetics*, vol. 45, no. 1, pp. 65–76, 2014.

[9] B. Lian, V. S. Donge, F. L. Lewis, T. Chai, and A. Davoudi, "Data-driven inverse reinforcement learning control for linear multiplayer games," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[10] B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang, "H infinity control of linear discrete-time systems: Off-policy reinforcement learning," *Automatica*, vol. 78, pp. 144–152, 2017.

[11] S. Li, J. Yang, W. Chen, and X. Chen, *Disturbance observer-based control: methods and applications*. Boca Raton: CRC press, 2014.

[12] Z. Tang, C. Wang, and Z. Ding, "Unmatched disturbance rejection for amb systems via dobc approach," in *2016 35th Chinese Control Conference*. IEEE, 2016, pp. 5931–5935.

[13] W. Chen, J. Yang, L. Guo, and S. Li, "Disturbance-observer-based control and related methods—an overview," *IEEE Transactions on Industrial Electronics*, vol. 63, pp. 1083–1095, 2015.

[14] Z. Tang, *Control design for the active magnetic bearing system*. The University of Manchester (United Kingdom), 2019.

[15] Z. Luo, P. Zhang, X. Ding, Z. Tang, C. Wang, and J. Wang, "Adaptive affine formation maneuver control of second-order multi-agent systems with disturbances," in *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2020, pp. 1071–1076.

[16] Z. Tang, P. Chris, J. A. Rossiter, E. Stephen, D. Gary, and G. Panoutsos, "Disturbance observer-based optimal tracking control for slot coating process with mismatched input disturbances," in *2024 UKACC 14th International Conference on Control (CONTROL)*. IEEE, 2024.

[17] Z. Tang, J. A. Rossiter, and G. Panoutsos, "A reinforcement learning-based approach for optimal output tracking in uncertain nonlinear systems with mismatched disturbances," in *2024 UKACC 14th International Conference on Control (CONTROL)*. IEEE, 2024.

[18] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*. London: Routledge, 2018.

[19] D. E. Kirk, *Optimal Control Theory: An Introduction*. Courier Corporation, 2004.

[20] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, pp. 878–888, 2010.

[21] M. Ran, J. Li, and L. Xie, "Reinforcement-learning-based disturbance rejection control for uncertain nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 52, pp. 9621–9633, 2022.

[22] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence and Robustness*. Mineola: Dover Publications, 2011.

[23] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, pp. 82–92, 2013.

[24] W. Chen, "Nonlinear disturbance observer-enhanced dynamic inversion control of missiles," *Journal of Guidance, Control, and Dynamics*, vol. 26, pp. 161–166, 2003.

[25] V. Nevistić and J. A. Primbs, "Constrained nonlinear optimal control: a converse hjb approach," 1996.