Reinforcement Learning-based Disturbance Rejection Control for Uncertain Nonlinear Systems

Maopeng Ran, Juncheng Li, and Lihua Xie, Fellow, IEEE

Abstract—This paper investigates the reinforcement learning (RL) based disturbance rejection control for uncertain nonlinear systems having non-simple nominal models. An extended state observer (ESO) is first designed to estimate the system state and the total uncertainty, which represents the perturbation to the nominal system dynamics. Based on the output of the observer, the control compensates for the total uncertainty in real time, and simultaneously, online approximates the optimal policy for the compensated system using a simulation of experience based RL technique. Rigorous theoretical analysis is given to show the practical convergence of the system state to the origin and the developed policy to the ideal optimal policy. It is worth mentioning that, the widely-used restrictive persistence of excitation (PE) condition is not required in the established framework. Simulation results are presented to illustrate the effectiveness of the proposed method.

Index Terms—Uncertain nonlinear systems, reinforcement learning (RL), disturbance rejection, optimal control, extended state observer (ESO).

I. INTRODUCTION

Reinforcement learning (RL), inspired by learning behavior in nature, is a goal-oriented learning strategy wherein the agent learns the policy to optimize a pre-defined reward by interacting with the environment [1][2]. From a perspective of control theory, RL is strongly connected with traditional optimal control and adaptive control algorithms [3]. The goal of an RL-based controller is to learn the optimal policy and value function by finding the solution to the Hamilton-Jacobi-Bellman (HJB) equation online. For its uniqueness in data-driven concepts, effectiveness in reaching optimal behavior, and adaptiveness to uncertain environment, RL has undergone rapid progress in control community [3][4].

Early-efforts on RL-based control were mainly devoted to linear systems. For example, a policy iteration technique to solve the continuous time LQR problem without the knowledge of the state matrix was proposed in [5]. The results in [5] were generalized to linear systems with completely unknown dynamics in [6]. In the last decade, the extension of RL-based control to nonlinear systems has gained considerable attention as well. Model-free optimal control problems for affine and nonaffine nonlinear systems were addressed based on RL in [7] and [8], respectively. In [9], RL-based optimal control for polynomial systems was considered. A new policy iteration scheme was developed in [9] which attempts to solve a convex optimization problem rather than a partial differential equation as usual. In [10] and [11], the problem of RL-based optimal tracking control for nonlinear systems in strict-feedback form was investigated. Ref. [10] handled this problem by transforming it into an equivalent optimal regulation problem through a feedforward adaptive input, while [11] addressed this problem by using a novel policy iteration technique to solve semidefinite HJB equations.

In practice, almost all control systems are inherently nonlinear and subject to uncertainties and disturbances [12][13]. Thus it is important to investigate the robustness of RL-based control policies for nonlinear systems. The traditional robust control methods, such as the nonlinear small-gain theorem [14][15], sliding mode control [16], and $H_{\infty}$ control [17][18] were integrated with the RL designs. In [19][20], novel adaptive critic strategies were developed with robustness guarantees for disturbed nonlinear systems. In [21][22], event-triggered mechanisms were introduced into the design of robust RL controllers to mitigate unnecessary communications. Note that to guarantee parameter convergence, the RL-based control laws generally require persistence of excitation (PE) of the state system. In [14][16][19][20], to fulfill the PE condition, probing signals are added into the control inputs, which will inevitably decrease the transient performance. What is more, the approaches in [14][22] were developed only for relatively simple uncertainties or disturbances (e.g., independent from the control signal).

On the other hand, the active disturbance rejection control (ADRC), which was proposed by Han [23], is an efficient methodology to deal with uncertainties and disturbances. The basic idea of ADRC is to first estimate the total uncertainty online via a device called extended state observer (ESO), and then compensate for it in the control loop in real time. Accordingly, an ADRC law contains two components: a disturbance rejection term compensates for the estimated uncertainty, and a nominal controller guarantees the performance of the compensated system. The existing ADRC literature mainly focuses on its engineering applications [24] and theoretical verifications for different classes of uncertain systems, such as MIMO systems [25], nonaffine-in-control systems [26], time-delay systems [27], and so on. On the contrary, there are very few works that concern the design of the nominal controller, since the compensated system is generally simplified into a chain of integrators (see, e.g., [25][28]).

Motivated by the observations stated above, in this paper, we consider the RL-based disturbance rejection control for uncertain nonlinear systems. The systems are assumed to
have non-simple nominal models (i.e., the nominal models may also be complex and nonlinear) and are subject to multiple uncertainties. In order to relax the requirement of the PE condition and avoid using the probing signal, concurrent learning (CL), also known as experience replay, is employed in this paper. The original idea of concurrent learning is to use the recorded data and current data simultaneously for parameter adaptation in the framework of model reference adaptive control (MRAC) [29]. In recent years, this idea was extended to develop RL-based control policies to remove the PE condition [17, 21, 22, 30, 31]. Note that in RL, estimates for the parameters of the value function are updated using the Bellman error (BE) as a performance metric. In this paper, we evaluate the BE along the system trajectory, and simultaneously, extrapolate the BE to any unexplored data points by using the nominal model. This concurrent learning technique is interpreted as simulation of experience [32, 33]. The main contributions of this paper are twofold:

1) An ESO-based reinforcement learning and disturbance rejection framework for uncertain nonlinear systems having non-simple nominal models is established. The ESO provides estimates of the system state and total uncertainty to the RL-algorithm and disturbance rejection term, respectively. The disturbance rejection term compensates for the total uncertainty in real time, while the output feedback simulation of experience based RL explores the optimal policy for the compensated system simultaneously. To the best of the authors’ knowledge, this work is the first attempt that bridges the gap between RL and the philosophy of ADRC.

2) A rigorous theoretical analysis that establishes online simultaneous disturbance rejection and optimal policy approximation without the PE condition is conducted. Due to the strong coupling between the ESO estimation error and the RL approximation error, the theoretical problem faced in this paper is much more challenging than the previous ADRC [25, 28] and RL results [32, 33]. Specifically, in [25, 28] the nominal controller is designed to stabilize a chain of integrators, while here it is a learning controller which is able to online approximate the optimal policy for the non-simple compensated system. Compared with [32, 33], we provide a more practical RL solution for nonlinear systems, since the developed controller is capable of handling multiple uncertainties and it is output feedback based.

The rest of this paper is organized as follows. Section II presents some preliminaries and formulates the problem. Section III states the design and analysis of the developed ESO-based reinforcement learning and disturbance rejection framework. In Section IV, some simulation results are given to demonstrate the effectiveness of the proposed method. Finally, Section V concludes the paper.

Notations: Throughout this paper, we use $C$ to represent the set of all continuously differentiable functions. For any continuously differentiable function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^l$, $f_x : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{l \times n}$ represents its gradient with respect to the first vector argument, i.e., $f_x(\nu_1, \nu_2) = \partial f(\nu_1, \nu_2)/\partial \nu_1$. $\lambda_{\max}(P)$ and $\lambda_{\min}(P)$ denote the maximum and minimum eigenvalues of matrix $P$, respectively. Big $O$-notation in terms of $\nu$ is denoted as $O(\nu)$ and it is assumed that this holds for $\nu$ positive and sufficiently small. sat : $\mathbb{R} \to \mathbb{R}$ denotes the standard unity saturation function defined by $\text{sat}(\nu) = \text{sign}(\nu) \cdot \min\{1, |\nu|\}$. $1_A(\nu)$ is the indicator function defined by $1_A(\nu) = \begin{cases} 1 & \text{if } \nu \in A \setminus \emptyset \\ 0 & \text{if } \nu \notin A. \end{cases}$ Throughout this paper, for the sake of brevity, the time variable $t$ of a signal will be omitted except when the dependence of the signal on $t$ is crucial for clear presentation.

II. Problem Formulation

A single-input, single-output system with relative degree $n$, under a suitable diffeomorphism, can be written in the following normal form [12]:

\[
\begin{align*}
\dot{z} &= f_x(z, \omega), \\
\dot{x} &= Ax + B[f(x, z, \omega) + g(x, z, \omega)u], \\
y &= Cx,
\end{align*}
\]  

where $x = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$ and $z \in \mathbb{R}^p$ are the states, $\omega \in \mathbb{R}$ is the external disturbance, $u \in \mathbb{R}$ is the control input, $y \in \mathbb{R}$ is the measured output, $f_x : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}^n$ and $f, g : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$ are continuously differentiable functions, and matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times 1}$, and $C \in \mathbb{R}^{1 \times n}$ are given by

\[
A = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 \\ \vdots \\ 0 \end{bmatrix}.
\]

For system [1], the following assumptions are made:

Assumption A1: The external disturbance $\omega$ and its time derivative $\dot{\omega}$ are bounded.

Assumption A2: The zero-dynamics $\dot{z} = f_z(z, \omega)$ with input $(x, \omega)$ is bounded-input-bounded-state stable.

In this paper, the functions $f(\cdot)$ and $g(\cdot)$ are assumed to be uncertain and partially known, that is,

\[
f(x, z, \omega) = f_0(x) + \Delta f(x, z, \omega), \\
g(x, z, \omega) = g_0(x) + \Delta g(x, z, \omega),
\]

where $f_0(\cdot), g_0(\cdot) \in C(\mathbb{R}^n, \mathbb{R})$ are known and globally bounded, $g_0(\cdot) \neq 0$, and $\Delta f(\cdot), \Delta g(\cdot) \in C(\mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}, \mathbb{R})$ are unknown.

Remark 1: Similar to many previous ESO works (see, e.g., [34, 35]), we require the global boundedness of $f_0(\cdot)$ and $g_0(\cdot)$. It should be pointed out that this requirement does not exclude linear functions or any unbounded function because in this paper we achieve semi-global results, and the global boundedness can be always satisfied by smoothly saturating the function outside a compact set of interest.

Remark 2: In system [1], the uncertainties are located in the same channel as the control signal. For a general nonlinear system with smooth unmatched uncertainties, one can conduct a well-defined state transformation to transform the system into the form of system [1] [36].
Note that system (1) is subjected to multiple uncertainties, including the uncertainty in zero dynamics \( \dot{z} = f_z(x, z, \omega) \), drift dynamics modeling error \( \Delta f(\cdot) \), mismatch of control \( \Delta g(\cdot)u \), and external disturbance \( \omega \). Inherited from the basic philosophy of ADRC [23], the total uncertainty is regarded as an extended state of the system, which is denoted by
\[
x_{n+1} = \Delta f(x, z, \omega) + \Delta g(x, z, \omega)u.
\]
(4)

With assumption A2, we concentrate on the control design for the \( x \)-subsystem. Had \( x \) and \( x_{n+1} \) been available for feedback, we could have the following controller:
\[
u = u_0^*(x) - \frac{x_{n+1}}{g_0(x)}.
\]
(5)

Here the second term \(-\frac{x_{n+1}}{g_0(x)}\) is to compensate the total uncertainty, and \( u_0^*(x) \) is the optimal policy for the compensated system
\[
\dot{x} = Ax + B[f_0(x) + g_0(x)u_0],
\]
with the cost functional
\[
J(x, u_0) = \int_0^\infty r(x(\tau), u_0(\tau))d\tau.
\]
(7)
The function \( r : \mathbb{R}^n \times \mathbb{R}^+ \rightarrow \mathbb{R}^+ \) denotes the instantaneous cost given by
\[
r(x, u_0) = Q(x) + u_0^TRu_0,
\]
(8)
where \( Q : \mathbb{R}^n \rightarrow \mathbb{R}^+ \) is positive definite, and \( R > 0 \).

Note that in the control (5), the optimal control \( u_0^*(x) \) is generally analytically infeasible, and the state \( x \) and total uncertainty \( x_{n+1} \) are both unavailable. The objective of this paper is then to develop an ESO-based reinforcement learning and disturbance rejection scheme, which is able to online simultaneously compensate the uncertainty and approximate the optimal policy.

III. ESO-BASED REINFORCEMENT LEARNING AND DISTURBANCE REJECTION

A. Control Design

First of all, an ESO is designed for system (1) to provide the estimates of the state and total uncertainty:
\[
\begin{align*}
\dot{x}_1 &= x_2 + \frac{l_1}{\varepsilon}(x_1 - \hat{x}_1), \\
\vdots \\
\dot{x}_{n-1} &= \hat{x}_n + \frac{l_{n-1}}{\varepsilon^{n-1}}(x_1 - \hat{x}_1), \\
\dot{x}_n &= \hat{x}_{n+1} + \frac{l_n}{\varepsilon^{n}}(x_1 - \hat{x}_1) + f_0(\hat{x}) + g_0(\hat{x})u, \\
\dot{\hat{x}}_{n+1} &= \frac{l_{n+1}}{\varepsilon^{n+1}}(x_1 - \hat{x}_1),
\end{align*}
\]
(9)
where \( \hat{x} = [\hat{x}_1, \ldots, \hat{x}_n]^T \in \mathbb{R}^n, \varepsilon < 1 \) is a small positive constant, \( L = [l_1, l_2, \ldots, l_{n+1}]^T \in \mathbb{R}^{n+1} \) is the observer gain selected such that the following matrix is Hurwitz:
\[
E = \begin{bmatrix}
-l_1 & 1 & 0 & \cdots & 0 \\
-l_2 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
-l_n & 0 & 0 & \cdots & 1 \\
-l_{n+1} & 0 & 0 & \cdots & 0
\end{bmatrix} \in \mathbb{R}^{(n+1) \times (n+1)}.
\]
Since \( \varepsilon \) is a small positive constant, the observer (9) exhibits peaking phenomenon during the initial fast transient [25] [34].

To protect the peaking of the observer from propagating into other variables, we employ the well-known saturation technique \([34]\) to saturate the output of the observer as
\[
\bar{x}_i = M_i s\left( \frac{\hat{x}_i}{M_i} \right), \quad 1 \leq i \leq n + 1,
\]
(10)
where \( M_i, 1 \leq i \leq n + 1, \) are saturation bounds to be selected such that the saturation will not be invoked in the steady-period of the observer, and \( s(\cdot) \) is an odd saturation-like function defined by
\[
s(\nu) = \begin{cases} \nu, & 0 \leq \nu \leq 1 \\
\nu + \frac{\nu - 1}{\varepsilon} - \frac{\nu^2 - 1}{2\varepsilon}, & 1 \leq \nu \leq 1 + \varepsilon \\
\frac{1 + \varepsilon}{\varepsilon}, & \nu > 1 + \varepsilon
\end{cases}
\]
Note that this function is nondecreasing, continuously differentiable with locally Lipschitz derivative. What is more, \( 0 \leq s'(\nu) \leq 1 \) and \( |s'(\nu) - \text{sat}(\nu)| \leq \frac{\varepsilon}{2}, \forall \nu \in \mathbb{R} \), where \( s'(\nu) = \frac{d s(\nu)}{d \nu} \).

For subsequent use, denote \( \pi = [\pi_1, \ldots, \pi_n]^T \), and \( \bar{x}_i = s'(\hat{x}_i/M_i)\hat{x}_i, 1 \leq i \leq n + 1 \).

According to the optimal control theory \([37]\), the optimal control problem of system (6) with cost functional (7) can be converted to solve the following HJB equation:
\[
V^*_x(x) [Ax + B(f_0(x) + g_0(x)u_0^*(x))] + Q(x) + u_0^TRu_0 = 0,
\]
(11)
where \( V^* \in C(\mathbb{R}^n, \mathbb{R}_{\geq 0}), V^*(0) = 0, \) is the optimal value function. The optimal control policy can be determined from the optimal value function as
\[
u_0^*(x) = -\frac{1}{2}R^{-1}g_0^T(x)B^TV_x^*(x).
\]
(12)
Generally, the analytical solution of the HJB equation is not feasible. However, the optimal value function \( V^*(x) \) and the optimal control policy \( u_0^*(x) \) can be approximated by an actor-critic neural network (NN) based approach \([7,8]\). For any given compact set \( \mathcal{X} \subset \mathbb{R}^n \) and positive constant \( \pi \), the optimal value function \( V^*(x) \) can be represented by
\[
V^*(x) = \Theta^T \phi(x) + \kappa(x),
\]
(13)
where \( \phi : \mathcal{X} \rightarrow \mathbb{R}^l \) is a continuously differentiable activation function, \( \Theta \in \mathbb{R}^l \) is the ideal weight vector, \( l \in \mathbb{N} \) is the number of neurons, and \( \kappa : \mathbb{R}^n \rightarrow \mathbb{R} \) is the approximation error function which satisfies \( |\kappa(x)| \leq \pi \) and \( |\kappa_x(x)| \leq \pi, \forall x \in \mathcal{X} \). It follows from the NN representation of the value function that the optimal control policy can be obtained as
\[
u_0^*(x) = -\frac{1}{2}R^{-1}g_0^T(x)B^T(\Theta^T \phi(x) + \kappa^T(x))
\]
(14)
Therefore, bearing in mind that the ESO (9) provides an estimate of the state \( x \), the NN-based approximations of \( V^*(x) \) and \( u_0^*(x) \) are given by
\[
\hat{V}(\bar{x}, \hat{\Theta}_V) = \hat{\Theta}_V^T \phi(\bar{x}),
\]
(15)
\[
\hat{u}_0(\bar{x}, \hat{\Theta}_E) = -\frac{1}{2}R^{-1}g_0^T(\bar{x})B^T \phi^T(\bar{x})\hat{\Theta}_E,
\]
(16)
where $\hat{\Theta}_v, \hat{\Theta}_c \in \mathbb{R}^l$ are estimates of $\Theta_v$ and weights for the critic and actor NNs, respectively. By (15) and (16), the approximated instantaneous BE $\delta_t : \mathbb{R}^n \times \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}$ can be computed as

$$
\delta_t = \delta \left( \bar{x}, \hat{\Theta}_v, \hat{\Theta}_c \right)
\triangleq \bar{U}_x(\bar{x}, \hat{\Theta}_c) \left[ A\bar{x} + B \left( f_0(\bar{x}) + g_0(\bar{x})\hat{u}_0(\bar{x}, \hat{\Theta}_c) \right) \right] 
+ Q(\bar{x}) + \hat{u}_0^T(\bar{x}, \hat{\Theta}_c) R \hat{u}_0(\bar{x}, \hat{\Theta}_c).
$$

(17)

In this paper, motivated by [32] [33], we implement an ESO based learning strategy via simulation of experience, i.e., use the estimated state $\bar{x}$ and the knowledge of $f_0(\cdot)$ and $g_0(\cdot)$ to extrapolate the BE to a predefined set of points $\{x^i \in \mathbb{R}^n | i = 1, \ldots, N\}$. The approximated BE extrapolated to the point $x^i$ is given by

$$
\delta_i = \delta \left( x^i, \hat{\Theta}_v, \hat{\Theta}_c \right).
$$

(18)

Then the actor-critic NN uses the BEs $\delta_i$ and $\delta_t$ to update the estimates $\hat{\Theta}_v$ and $\hat{\Theta}_c$. A least-square update law for the critic NN is given by

$$
\dot{\hat{\Theta}}_v = -\lambda_{v1} \Gamma \frac{\hat{\Theta}}{\rho} - \lambda_{v2} \sum_{i=1}^{N} \mu_i \delta_i,
$$

(19)

$$
\hat{\Gamma} = \left( \beta \Gamma - \lambda_{v1} \frac{\mu_i \Gamma \mu_i^T \Gamma}{\rho^2} \right) \Gamma \leq s_1,
$$

(20)

where $\Gamma : \mathbb{R}_{>0} \to \mathbb{R}^{l \times l}$ is a time-varying least-square gain matrix, $s_1 > 0$ is a saturation constant, $\lambda_{v1}, \lambda_{v2} > 0$ are constant adaptation gains, $\beta > 0$ is a constant forgetting factor, and

$$
\mu = \phi_\delta(x) \left[ A \bar{x} + B \left( f_0(\bar{x}) + g_0(\bar{x})\hat{u}_0(\bar{x}, \hat{\Theta}_c) \right) \right],
$$

$$
\mu_i = \phi_\delta(x^i) \left[ A x^i + B \left( f_0(x^i) + g_0(x^i)\hat{u}_0(x^i, \hat{\Theta}_c) \right) \right],
$$

$$
\rho = 1 + \gamma \mu \Gamma \mu,
$$

$$
\rho_i = 1 + \gamma \mu_i \Gamma \mu_i,
$$

with $\gamma > 0$ a constant normalization gain. The update law (20) indicates that the time-varying gain matrix $\Gamma$ is bounded in the sense that

$$
\frac{1}{s_0} I \leq \Gamma(t) \leq s_1 I, \quad t \geq 0,
$$

(21)

where $s_0 \in \mathbb{R}^+$ and $I$ is the identity matrix. According to the subsequent stability analysis, the update law for the actor NN is given by

$$
\dot{\hat{\Theta}}_c = -\lambda_{c1} \left( \hat{\Theta}_c - \hat{\Theta}_v \right) - \lambda_{c2} \hat{\Theta}_v
$$

$$
+ \lambda_{c1} \frac{G_i \hat{\Theta}_c \mu_i \Gamma \mu_i^T}{4 \rho} \hat{\Theta}_v 
+ \sum_{i=1}^{N} \lambda_{c2} G_i \hat{\Theta}_c \mu_i \Gamma \mu_i^T \hat{\Theta}_v,
$$

(22)

where $\lambda_{c1}, \lambda_{c2} > 0$ are constant adaptation gains, and

$$
G_i \triangleq \phi_\delta(x) B g_0(\bar{x}) R^{-1} \hat{g}_0(\bar{x}) \bar{x}^T \phi_\delta(x),
$$

$$
G_i \triangleq \phi_\delta(x^i) B g_0(x^i) R^{-1} \hat{g}_0(x^i) x^i \bar{x}^T \phi_\delta(x^i).
$$

Finally, based on the output of the ESO [9] and (16), the control that simultaneously compensates the total uncertainty and approximates the optimal policy is given by

$$
u = \hat{u}_0 \left( \bar{x}, \hat{\Theta}_c \right) - \frac{\pi_{n+1}}{g_0(\bar{x})}.
$$

(23)

The block diagram of the developed ESO-based reinforcement learning and disturbance rejection is depicted in Fig. 1.

B. Stability Analysis

To establish the stability of the closed-loop system formed by (1), (9), (19), (20), (22), and (23), we need two additional assumptions. By Assumption A1, let $\omega(t) \in \mathcal{W}$, $\forall t \in [0, \infty)$, for some compact set $\mathcal{W} \subseteq \mathbb{R}$. Assume that the initial condition of system (1), $x(0)$, belongs to a compact set $\mathcal{X}_0$ which is slightly smaller than $\mathcal{X}$ (i.e., $\mathcal{X}_0 \subseteq \mathcal{X}$ and their boundaries are disjoint). By Assumption A2, there exists a constant $c_2$ such that the compact set $\mathcal{Z} = \{ z \in \mathbb{R}^p; \| z \| \leq c_2 \}$ is a positive invariant set of $\dot{z} = f(z, x, \omega)$, for all $x \in \mathcal{X}$ and $\omega \in \mathcal{W}$. Denote the compact set $\Xi$ by $\Xi = \mathcal{X} \times \mathcal{Z} \times \mathcal{W}$.

Assumption A3: The known control coefficient $g_0(x)$ satisfies

$$
k_g \triangleq \max_{(x, z, \omega) \in \Xi} \left( g(x, z, \omega) - g_0(x) \right) / g_0(x) < 1.
$$

(24)

Assumption A4: There exists a finite set of fixed points $\{x^i \in \mathbb{R}^n | i = 1, \ldots, N\}$ such that,

$$
e = \frac{1}{N} \inf_{t \geq 0} \left( \lambda_{\min} \left( \sum_{i=1}^{N} \frac{\mu_i \mu_i^T}{\rho_i} \right) \right) > 0.
$$

(25)

Remark 3: Assumption A3 is a standard assumption for the ESO-based controller [25] [34]. The rational behind this assumption is that the total uncertainty $x_{n+1}$ to be compensated in the control law contains the mismatch of control (i.e., $\Delta g(x, z, \omega) u$). To guarantee the “compensation capability” of the ESO-based controller, the nominal control coefficient $g_0(x)$ should not be too far away from the actual control coefficient $g(x, z, \omega)$. Condition (24) is not difficult to be satisfied in practice [34]. Specifically, for any uncertain control gain $g(x, z, \omega)$ satisfying $g < g(x, z, \omega)(x, z, \omega) \in \Xi \leq g$ (or $-g < g(x, z, \omega)(x, z, \omega) \in \Xi \leq -g$), where $g$ and $-g$ are positive constants. By selecting $g_0(x)|_{x \in \mathcal{X}} > g/2$ (or $g_0(x)|_{x \in \mathcal{X}} < -g/2$), the condition (24) is satisfied. ■

Remark 4: Condition (25) depends on the estimate $\hat{\Theta}_c$, and hence it cannot be guaranteed a prior. However, different
from the well-known PE condition [33], the condition (22) can be monitored online [32]. Moreover, this condition can be heuristically met by selecting more points than the number of the neurons, i.e., choosing $N \gg l$ (see the numerical example in Section IV).

Now, we are in a position to state our main result.

**Theorem 1:** Consider the closed-loop system formed by plant (1), ESO (9), control (23), and RL-based update laws (19), (20), and (22). Suppose Assumptions A1 to A4 are satisfied, and the initial condition of the plant $x(0)$ is an interior point of $X_0$. Then there exists $\varepsilon^* > 0$ such that for any $\varepsilon \in (0, \varepsilon^*)$:

- for any $T > 0$, $\lim_{\varepsilon \to 0^+} |x_i - \hat{x}_i| \to 0$, $1 \leq i \leq n + 1$, uniformly in $t \in [T, \infty)$;
- the state $x$ and the weight estimation errors $\hat{\Theta}_v = \hat{\Theta}_v - \Theta$ and $\check{\Theta}_c = \check{\Theta}_c - \Theta$ are uniformly bounded.

**Proof:** First of all, we give some notations and definitions for subsequent use. Let $Z(t) = [x^T(t), \hat{\Theta}_v^T(t), \check{\Theta}_c^T(t)]^T, V_v = \frac{1}{2} \hat{\Theta}_v^T \Gamma \hat{\Theta}_v, V_c = \frac{1}{2} \check{\Theta}_c^T \check{\Theta}_c$. Denote $\tau_v = \frac{1}{2} \varepsilon_1 \| \hat{\Theta}_v(0) \| + 1$, $\tau_c = \max \{ V_v(\hat{\Theta}_v(0)), V_c(\check{\Theta}_c(0)) \}_{t \geq t_0} + 1$, where $t_0$ and $\tau_\varepsilon$ are given by (45). Define several compact sets:

\[
\begin{aligned}
\Omega^0_v &= \{ \hat{\Theta}_v \in \mathbb{R}^l; V_v(\hat{\Theta}_v) \leq \tau_v \}, \\
\Omega^1_v &= \{ \hat{\Theta}_v \in \mathbb{R}^l; V_v(\hat{\Theta}_v) \leq \tau_v + 1 \}, \\
\Omega^\varepsilon_v &= \{ \hat{\Theta}_v \in \mathbb{R}^l; V_v(\hat{\Theta}_v) \leq \tau_\varepsilon \}, \\
\Omega^\varepsilon_c &= \{ \check{\Theta}_c \in \mathbb{R}^l; V_c(\check{\Theta}_c) \leq \tau_c \}.
\end{aligned}
\]

Based on the above definitions, denote $\Omega^0 = X_0 \times \Omega^0_v \times \Omega^\varepsilon_c$ and $\Omega^\varepsilon = \mathcal{X} \times \Omega^1_v \times \Omega^\varepsilon_c$. It can be clearly observed that $Z(0)$ is an internal point of $\Omega^0$ and $\Omega^\varepsilon \subseteq \Omega^0 \cap \partial \Omega^1$.

The following proof will be divided into two steps. In the first step, we show that, for sufficiently small $\varepsilon$ and appropriately selected learning gains, $Z(t) \in \Omega^1_v, \forall t \in [0, \infty)$. Then, in the second step, the convergence of the ESO and the uniformly boundedness of $x, \hat{\Theta}_v$, and $\check{\Theta}_c$ are proved.

**Step 1:** There exists $\varepsilon^1 > 0$ such that for any $\varepsilon \in (0, \varepsilon^1)$, $Z(t) \in \Omega^1_v, \forall t \in [t_0, \infty)$.

Since $Z(0)$ is an interior point of $\Omega^0$, $\Omega^0 \subseteq \Omega^1 \cap \partial \Omega^1$, and the output of the ESO (9) is saturated, there exists an $\varepsilon$-independent $t_0 > 0$ such that $Z(t) \in \Omega^1_v, \forall t \in [t_0, \infty)$. We prove the conclusion in this step by contradiction. To this end, assume that there exist $t_2 > t_1 > t_0$ such that

\[
\begin{aligned}
Z(t_1) &\in \partial \Omega^0_v, \\
Z(t) &\in \Omega^1_v, \quad t \in [t_1, t_2], \\
Z(t_2) &\in \partial \Omega^1_v.
\end{aligned}
\]

Consider the scaled ESO estimation error $\eta = [\eta_1, \ldots, \eta_{n+1}]^T$ with

\[
\eta_i = \frac{x_i - \hat{x}_i}{\varepsilon^{n+1-i}}, \quad 1 \leq i \leq n + 1.
\]

By (1), (9), and (23), the dynamics of $\eta$ can be formulated as

\[
\begin{aligned}
\varepsilon \eta_1 &= -l_1 \eta_1 + \eta_{n+1}, \quad 1 \leq i \leq n - 1, \\
\varepsilon \eta_n &= -\eta_{n-1} + \eta_{n+1} + \eta_1, \\
\varepsilon \eta_{n+1} &= -l_{n+1} \eta_1 - \varepsilon \eta_2 + \varepsilon \eta_3, \\
\theta_1(x, \hat{x}, \check{x}_{n+1}, \check{\Theta}_c) &= f_0(x) - f_0(\hat{x}) + (g_0(x) - g_0(\hat{x})) \left[ \bar{u}_0(\bar{\tau}, \check{\Theta}_c) - \frac{\tau_{n+1}}{\tau_{n+1}} \right], \\
\theta_2(x, z, \omega) &= \frac{g_0(x, z, \omega) - g_0(\hat{x})}{\tau_{n+1}} s'(\check{x}_{n+1}/\check{\tau}_{n+1}), \\
\theta_3(x, z, \omega, \hat{x}, \check{x}_{n+1}, \check{\Theta}_c) &= \check{f}(x, z, \omega) - f_0(x) + (\check{g}(x, z, \omega) - \check{g}(\hat{x})) \left[ \frac{\tau_{n+1}}{\tau_{n+1}} \right], \\
\theta_4(x, z, \omega) &= \frac{g_0(x, z, \omega) - g_0(\hat{x})}{\tau_{n+1}} \bar{u}_0(\bar{\tau}, \check{\Theta}_c).
\end{aligned}
\]

The equations of $\eta$ can be rewritten into the following compact form:

\[
\dot{\eta} = \frac{1}{\varepsilon} E \eta - F_1 \partial_2(\cdot) \gamma_1 + \left[ F_2 \frac{\partial_1(\cdot)}{\varepsilon} + F_1 \partial_3(\cdot) \right], \tag{29}
\]

for $F_1 = [0 \ B^T]^T, F_2 = [B^T \ 0]^T$. The matrix $E$ is Hurwitz by design. To show the convergence of $\eta$, we first consider the following system which is simplified from (29):

\[
\dot{\eta} = \frac{1}{\varepsilon} E \eta - F_1 \partial_2(\cdot) \gamma_1. \tag{30}
\]

The system above can be regarded as a negative feedback connection of the time-varying gain $\partial_2(\cdot)$ and the transfer function

\[
G(\varepsilon s) = \frac{l_{n+1}}{(\varepsilon s)^{n+1} + l_1 (\varepsilon s)^n + \cdots + l_{n+1}}. \tag{31}
\]

Note that $|\partial_2(\cdot)| \leq k_{\beta} s'|(\check{x}_{n+1}/\check{\tau}_{n+1})| \leq k_{\beta}$. By Assumption A3 and (31), one has $k_{\beta} \| G \|_{\infty} < 1$. It then follows from the circle criterion [13] that the origin of (30) is globally exponentially stable. Therefore, applying a loop transformation to (30) and using the Kalman-Yakubovich-Popov lemma [13], there exists a quadratic Lyapunov function $W(\eta) = \eta^T P \eta$ for system (30) satisfying $W(\eta) \leq -\frac{1}{2} P \eta(\eta)$, where $P$ is a positive definite matrix and $\alpha$ is an $\varepsilon$-independent positive constant. Utilizing $W(\eta)$ as a Lyapunov function candidate for system (29), and calculating its derivative yields

\[
W(\eta) \leq -\alpha \varepsilon W(\eta) + 2 \xi \| \| \left[ F_2 \frac{\partial_1(\cdot)}{\varepsilon} + F_1 \partial_3(\cdot) \right], \tag{32}
\]

where $\xi = \lambda_{\max}(P)$. Note that 1) the functions $f_0(\cdot), g_0(\cdot)$, and $s(\cdot)$ are continuously differentiable with locally Lipschitz derivatives and globally bounded; 2) by (26), $x, \hat{x}_i, \check{x}_i$, and $\check{\Theta}_c$ are all bounded in the time interval $[0, t_2]$; 3) by (27), $|\check{x}_i| \leq |x_i| + \varepsilon^{n+1-i} |\eta_i|, 1 \leq i \leq n + 1$. Based on these observations, we can conclude that $\partial_1(\cdot)$ and $\partial_3(\cdot)$ are locally Lipschitz, and bounded from above by affine in $\| \eta \|$ functions, uniformly in $\varepsilon$. Here, we take the term $f_0(x) - f_0(\hat{x})$ as an example to further clarify this point. According to the Hadamard’s lemma [41], one has

\[
\begin{aligned}
\frac{1}{\varepsilon} [f_0(x) - f_0(\hat{x})] &= \frac{1}{\varepsilon} (x - \hat{x}) \int_0^1 \frac{\partial f}{\partial x}(\bar{x} + \lambda (x - \hat{x})) d\lambda \\
&= \left[ \varepsilon^{n+1-1} \eta_1, \ldots, \varepsilon \eta_{n+1} \right]^T \int_0^1 \frac{\partial f}{\partial x}(\bar{x} + \lambda (x - \hat{x})) d\lambda.
\end{aligned}
\]
Therefore, from (32), one has that there exists an \( \varepsilon_1 > 0 \) such that for any \( \varepsilon \in (0, \varepsilon_1) \) and \( \bar{t}_1 \in (t_0 / 2, t_0) \),
\[
\| \eta \| = O(\varepsilon), \quad \forall t \in [\bar{t}_1, t_2].
\]
By the convergence of the ESO, and the definition of \( x_{n+1} \), the bounds to saturate the output of the observer are selected to satisfy
\[
M_i > \sup_{x \in \mathcal{X}} |x_i|, \quad 1 \leq i \leq n,
\]
\[
M_{n+1} > \sup_{(x, z, \omega) \in \bar{Z} \times \bar{\Theta}_c \times \Omega_1} \left| \frac{\Delta f(x, z, \omega)}{g(x, z, \omega)} \right| + \frac{\Delta g(x, z, \omega) \hat{u}_0(x, \hat{\Theta}_c)}{g(x, z, \omega)}.
\]
It follows that the saturation elements \( M_i s(\tilde{x}_i / M_i) \), \( 1 \leq i \leq n + 1 \), all work in the linear zone in the time interval \( [t_1, t_2] \), i.e., \( \pi_i = \tilde{x}_i \).
In the following, we consider the dynamics of \( Z \). To facilitate the subsequent analysis, the NN-based approximations \( \tilde{V}_z(\pi, \hat{\Theta}_c) \) and \( \tilde{u}_0(\pi, \hat{\Theta}_c) \) are expressed in terms of \( \Theta_v, \Theta_c, \kappa(\tilde{x}) \), and \( \kappa_c(\tilde{x}) \) as
\[
\tilde{V}_z(\pi, \hat{\Theta}_c) = \Theta^T \phi_x(\tilde{x}) + (\Theta_v - \Theta)^T \phi_x(\tilde{x})
\]
\[
= \tilde{V}_z^{*}(\pi) + \Theta_v^T \phi_x(\tilde{x}) - \kappa_x(\tilde{x}),
\]
\[
\tilde{u}_0(\pi, \hat{\Theta}_c) = \frac{1}{2} \left[ \frac{\Delta g(x, z, \omega) \hat{u}_0(x, \hat{\Theta}_c)}{g(x, z, \omega)} \right].
\]
By (11) and (13), and inserting (34) and (37) into (17), the instantaneous BE \( \delta_t \) can be written as
\[
\delta_t = \mu^T \Theta_v - \kappa_x(\tilde{x}) \left[ A\tilde{x} + B \left( f_0(x) + g_0(\tilde{x}) \hat{u}_0(x, \hat{\Theta}_c) \right) \right]
\]
\[
+ 2u_0^T \tilde{r} \left[ \Theta_v^T \phi_x(\tilde{x}) \Theta_c \right]
\]
\[
+ \frac{1}{2} R^{-1} \left[ \Theta_v^T \phi_x(\tilde{x}) \Theta_c \right] + \frac{1}{2} R^{-1} \left[ \Theta_v^T \phi_x(\tilde{x}) \Theta_c \right]
\]
\[
= \mu^T \Theta_v + \frac{1}{4} \Theta_v^T \Theta_c \Theta_v + \Delta,
\]
where
\[
\Delta = 1 - \kappa_x(\tilde{x}) \left[ A\tilde{x} + B f_0(\tilde{x}) \right]
\]
\[
+ \frac{1}{4} \Theta_v^T \Theta_c \Theta_v + \Delta,
\]
Similarly, the extrapolated BE \( \delta_i \) can be expressed as
\[
\delta_i = \mu_1^T \Theta_v + \frac{1}{4} \Theta_v^T \Theta_c \Theta_v + \Delta_i,
\]
where
\[
\Delta_i = 1 - \kappa_c(x^i) \left[ A x^i + B f_0(x^i) \right]
\]
\[
+ \frac{1}{4} \Theta_v^T \Theta_c \Theta_v + \Delta.
\]
By (37), the derivative of \( V^*(x) \) under the control (23) is given by
\[
\dot{V}(x) = V^*_z(x) \left[ A x + B (f_0(x) + g(x, z, \omega) u) \right]
\]
\[
= V^*_z(x) \left[ A x + B (f_0(x) + g_0(x) u_0(x)) \right]
\]
\[
+ V^*_z(x) B \left[ x_{n+1} - \frac{g_0(x)}{g_0(\tilde{x})} \right]
\]
\[
+ g_0(\tilde{x}) \left( \hat{u}_0(x, \hat{\Theta}_c) - u_0^*(\tilde{x}) \right)
\]
\[
+ g_0(\tilde{x}) \left( u_0^*(\tilde{x}) - u_0^*(x) \right).
\]
\[- \lambda_{c1} \epsilon \| \Theta_c \|^2 - (\lambda_{c1} + \lambda_{c2}) \| \Theta_c \|^2 \\
+ \frac{\lambda_{c1}}{2} (\| \Theta_c \|^2 + \| \Theta_c \|^2) + \lambda_{c2} \| \Theta_c \| \| \Theta_c \|^2 \\
+ \chi_1 \| \Theta_c \| (\| \Theta_c \|^2 + \| \Theta_c \|^2) \\
- \frac{1}{2} \beta \| \Theta_c \|^2 + \lambda_{c1} \| \Theta_c \|^2 \\
+ \eta_2 \| \eta \| + \frac{1}{2} \beta \| \Theta_c \|^2 \\
+ \frac{\lambda_{c2}}{2} \| \Theta_c \|^2 \sum_{i=1}^{N} | \Delta_i | \\
\] 

where

\[ \eta_2 \triangleq \frac{- \lambda_{c2} \epsilon - \beta}{\epsilon_1} + \lambda_{c1} \| \Theta \| + \frac{\lambda_{c1}}{2}, \]
\[ \eta_6 \triangleq \frac{- \lambda_{c1}}{2} - \lambda_{c2} + \frac{3}{2} \chi_1 \| \Theta \|, \]
\[ \eta_7 \triangleq \lambda_{c2} \| \Theta \| + \chi_1 \| \Theta \|^2, \]
\[ \eta_8 \triangleq \eta_2 \| \eta \| + \epsilon_3 \| \Theta \| + \lambda_{c1} \| \Theta \|^2 \\
+ \frac{\lambda_{c2}}{2} \| \Theta \|^2 \sum_{i=1}^{N} | \Delta_i |. \]
\[ \chi_1 \triangleq \frac{\lambda_{c1}}{8} \sup_{x \in \chi} \| G_i \| + \frac{\lambda_{c2}}{8} \| \Theta \|^2 \frac{\max_{i \leq N} \| G_i \|}{| \Delta_i |}. \]

Sufficient conditions for the gains \( \lambda_{c1}, \lambda_{c1}, \lambda_{c2}, \) and \( \beta \) are then given by

\[ \beta > \chi_1 \| \Theta \| + \lambda_{c2} \| \Theta \|, \]
\[ \lambda_{c1} > 3 \chi_1 \| \Theta \|. \]  

Provided the gains are selected satisfying (46), the ESO is convergent in the sense of (33), the upper bound of the approximation error \( \kappa_x \) can be made sufficiently small by increasing the number of NN neurons, and (26), the derivative of \( V \) satisfies

\[ \dot{V}(Z(t)) \leq 0, \quad t \in [t_1, t_2]. \]  

This contradicts (26). Thus the statement in Step 1) holds.

Step 2): There exists \( \epsilon^* \in (0, \epsilon] \) such that for any \( \epsilon \in (0, \epsilon^*], \) \( \| \eta(t) \| = O(\epsilon), \) \( \forall t \in [T, \infty), \) \( T > 0, \) and \( Z(t) \) is uniformly ultimately bounded.

Since \( Z(t) \in \Omega^1, \forall t \in [0, \infty), \) by (32), one has that for any \( \epsilon \in (0, \epsilon^*], \) \( T > 0, \) and \( \| \eta(t) \| = O(\epsilon), \) \( \forall t \in [T, \infty). \) What is more, (45) holds for \( t \in [T, \infty). \) Consequently, selecting the parameters according to (46), and by (45), one can conclude the uniformly ultimately boundedness of \( Z. \) Also note that the system state \( x \) converges to the neighbourhood of origin, and the estimate weight \( \hat{\Theta}_c \) approximates the ideal weight \( \Theta \) if \( \eta_7/\| \Theta \| \) (i.e., \( \lambda_{c2}/\lambda_{c1} \)) is sufficiently small. This completes the proof of Theorem 1. 

Remark 5: In this paper, similar to (7), (8), (32), we use two weights \( \hat{\Theta}_c \) and \( \hat{\Theta}_c \) to estimate the same ideal weight \( \Theta. \) The use of two weights rather than one weight is mainly motivated by the stability analysis. Note that in (38) and (39), benefited from the use of two weights, the BEs are linear with respect to the weight estimation error \( \Theta_c, \) which enables us to develop a least-square update law for \( \hat{\Theta}_c \) based on the BEs. The update law for \( \hat{\Theta}_c \) guarantees that \( \hat{\Theta}_c \rightarrow \Theta_c. \)

Remark 6: We should point out that the update law in this paper is different from the standard CL update law in existing works such as (17, 21, 22, 29–31). As depicted in Fig. 2, for a standard CL update law, the recorded data can be only selected along the system trajectory. In this case, to guarantee that the system state visits sufficient number of points in the domain of operation, a probing signal is generally required to excite the system (31). In this paper, with the known nominal model of the system, the BE can be extrapolated to any desired data points, and the system trajectory doesn’t need to really reach to these data points (see Fig. 2). Benefited from this, our developed approach doesn’t need a probing signal, and hence it is able to achieve better transient performance.

Remark 7: According to Theorem 1, \( \hat{x} \rightarrow x \) and the actor NN weight \( \hat{\Theta}_c \) approximates the ideal weight \( \Theta. \) Therefore, the control policy given by \( \hat{u}_0(x, \hat{\Theta}_c) \) approximates the ideal optimal control policy \( u^*_0(x). \) In this sense and following also (24), \( \hat{u}_0(x, \hat{\Theta}_c) \) is referred to as an approximate optimal control. Also note that, similar to (32), (33), (39), the convergence results in this paper are practical. Thus the cost defined by (7), when evaluated along the approximate optimal trajectory, will be infinite. This, however, will not cause any theoretical concern. Note that the partial derivative \( \hat{V}_x \) leveraged in the BEs (17)–(18) to update the NN weights is bounded for all \( t \in [0, \infty). \) Actually, the design and analysis of the RL-based controller doesn’t need the infinite horizon cost along the approximate optimal trajectory to be finite, it only needs the cost, when evaluated along the ideal optimal trajectory to be finite. To avoid infinite cost, one could use a discounted cost formulation with some additional conditions to guarantee stability (40).

Remark 8: There are several groups of parameters that need to be selected for the developed ESO-based reinforcement learning and disturbance rejection scheme.

- For the ESO, the parameters to be selected include the observer gains \( L \) and \( \epsilon, \) the saturation bounds for the observer output \( (M_i, 1 \leq i \leq n + 1), \) and the saturation
bounds for the functions \( f_0(\cdot) \) and \( g_0(\cdot) \) (denoted by \( M_f \) and \( M_g \), respectively). The observer gain \( L \) can be selected by a pole placement method such that the matrix \( E \) is Hurwitz. Theoretically, \( \varepsilon \) can be selected arbitrarily small to achieve more accurate estimation. However, in practice, the lower-bound of \( \varepsilon \) is limited due to noise and sampling constraints \([26, 34]\). The bounds \( M_i, 1 \leq i \leq n + 1 \), are selected according to \([34-45]\). Note that the calculation of \( M_{n+1} \) is generally not straightforward, and one might end up with a conservative bound. From a practice viewpoint, the values of \( \varepsilon \) and \( M_{n+1} \) can be decided by a simple trial and error procedure, based on the obtained performance. Our numerical experience and many previous ESO results (see, e.g., \([26, 34]\)) indicate that it is generally very easy to select a group of satisfactory \( \varepsilon \) and \( M_{n+1} \). Finally, the bounds \( M_f \) and \( M_g \) are selected such that \( M_f \geq \sup_{x \in X} |f_0(x)| \) and \( M_g \geq \sup_{x \in X} |g_0(x)| \). respectively.

- For the simulation of experience based RL, the parameters to be selected include the positive adaptation gains \( \lambda_1, \lambda_2, \lambda_3, \lambda_4, \) and \( \beta \). The sufficient conditions for the adaptation gains based on the stability analysis are given by \([46]\). Note that the inequalities in \([46]\) depend on unknown parameters \( \Theta, \zeta_0, \zeta_1, \) and \( c \). Therefore, the selection of these adaptation gains also needs a simple trial and error procedure. What is more, \( \lambda_2/\lambda_1 \) is required to be sufficiently small.

- For the data set \( \{x^i \in \mathbb{R}^n | i = 1, \ldots, N\} \) to extrapolate the BE, it is required to satisfy Assumption A4. We mention that, similar to the probing signals used in previous RL algorithms \([14,16,19,20]\), the data set is also selected offline and then the RL algorithm is implemented online. It is difficult to provide a theoretical guarantee to make sure that the offline selected data points (and probing signals) achieve satisfactory online performance. In practice, to fulfill Assumption A4, one can select \( x^i, 1 \leq i \leq N \), on an \( a \times a \times \cdots \times a \) data grid which covers the interested compact set \( \mathcal{X} \subseteq \mathbb{R}^n \), where \( a \) is an appropriately large positive integer.

- For the basis function \( \phi(x) \), and the initial NN weights \( \hat{\Theta}_v(0) \) and \( \hat{\Theta}_z(0) \), their selections are important for RL-based controllers \([2,4]\). The basic principle of selecting the basis function is to let the real basis be contained in the selected basis to make sure that the NN representation error \( \kappa(x) \) is small. In this paper, since the system nominal model (i.e., \( f_0(x) \) and \( g_0(x) \)) is known to the designer, the selection of the basis function will be relatively easier than that for an unknown system. However, the selection of a good basis for very general nonlinear systems is challenging and is still largely open in machine learning \([39]\). On the other hand, similar to \([32,33]\), benefited from the mechanism that the BE can be extrapolated to any desired data points with a known system nominal model, the initial NN weights are not required to be admissible. However, in practice, better initial NN weights will lead to better closed-loop transient performance.

Remark 9: Compared with the existing RL results for uncertain nonlinear systems \([14,22]\), the advantages of the proposed approach are threefold: 1) The uncertainties considered in this paper are more general and complex, since it involves the system states \( x \) and \( z \), the control input \( u \), and the external disturbance \( \omega \), and hence the analysis is much more challenging. 2) This paper inherits the idea of ADRC and handles the uncertainty in an “observation+compensation” scheme, which is fundamentally different from the ideas in \([14,22]\). 3) This paper provides a more practical solution to the RL design problem for uncertain nonlinear systems since our developed approach is output feedback based and does not require the probing signal. The approaches in \([14,22]\) are full state-feedback based and in \([14,16,19,20]\) require the probing signal.

IV. Examples

This section presents two simulation examples to illustrate the effectiveness of our developed approach.
A. Example 1

Consider the following uncertain nonlinear system

\[
\begin{align*}
\dot{z} &= -(x_2^2 + \omega^2)z, \\
\dot{x}_1 &= x_2, \\
\dot{x}_2 &= x_1 - 2.5x_2 + \omega + z^2 + 0.5(x_1 + x_2)(\cos(2x_1) + 2)^2 f(x,z,\omega) \\
&+ (\cos(2x_1) + 2 + \sin(x_1)\omega) u, \\
y &= x_1,
\end{align*}
\]

where the external disturbance is numerically set as \(\omega = 0.5\sin(t)\). The known nominal models of \(f(x,z,\omega)\) and \(g(x,z,\omega)\) are taken as

\[
\begin{align*}
f_0(x) &= -x_1 - 1.5x_2 + 0.5(x_1 + x_2)(\cos(2x_1) + 2)^2, \\
g_0(x) &= \cos(2x_1) + 2.
\end{align*}
\]

The cost functional for the nominal system is given by (7) with \(Q = x^TQx\) and \(R = 1\), where \(Q = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}\) is positive definite. The nominal models are selected above since the corresponding optimal control problem has an analytical solution, which is helpful for the simulation to verify the correctness of the developed approach. Specifically, according to the procedure in [42], the optimal value function is

\[
V^*(x) = 1.5x_1^2 + 2x_1x_2 + x_2^2, \quad \text{and the optimal control policy is } \quad u^*_0(x) = -(\cos(2x_1) + 2)(x_1 + x_2). \quad \text{The total uncertainty for system (48) is denoted by}
\]

\[
x_3 = -x_2 + \omega + z^2 + \sin(x_1)\omega u. \quad (49)
\]

It can be verified that Assumptions A1 to A3 are satisfied.

Consider the scenario that the compact set of interest is \(\mathcal{X} = [-2,2] \times [-2,2]\). The ESO is designed with \(L = [3 \ 3 \ 1]^T\) and \(\varepsilon = 0.02\). The saturation bounds for the nominal functions \(f_0(x)\) and \(g_0(x)\) are selected as \(M_f = 6.5 > \sup_{x \in \mathcal{X}} |f_0(x)|\) and \(M_g = 3 > \sup_{x \in \mathcal{X}} |g_0(x)|\), respectively. The saturation bounds for the output of the observer \(\hat{x}_1, \hat{x}_2, \text{ and } \hat{x}_3\), are selected as \(M_1 = M_2 = M_3 = 2\).

**Simulation with known basis function:** First, we simulate the simple case, i.e., the basis function is known. In this case, the basis function \(\phi: \mathbb{R}^2 \to \mathbb{R}^3\) is selected as \(\phi(x) = [x_1^2 \ x_1x_2 \ x_2^2]^T\). According to the analytical solution of \(V^*(x)\), the ideal weight \(\Theta = [1.5 \ 2 \ 1]^T\). The data points to extrapolate the BE are selected on a \(5 \times 5\) data grid cover the domain \(\mathcal{X}\) (see Fig. 3). The gains for the RL are selected as \(\lambda_{c1} = 1, \lambda_{c2} = 5, \lambda_{c3} = 100, \lambda_{c4} = 0.1, \gamma = 0.5, \text{ and } \beta = 100\). The upper bound of \(\Gamma\) is set as \(\varsigma_1 = 2000\). Simulation is done with initial conditions \(z(0) = 1\),
ideal value of the critical and actor NN weights is \( \tilde{\Theta} = \left[ \begin{array}{c} \Theta_c(0) \\ \Theta_a(0) \end{array} \right] = \left[ \begin{array}{c} 0.5 \\ 0.5 \end{array} \right] \), and \( \Gamma(0) = \text{diag}\{100, 100, 100\} \). All other settings are the same with the simulation with known basis function. Trajectories of the system state \( x \) and the actor NN weight \( \hat{\Theta}_a \) are illustrated in Figs. 9 and 10 respectively. It can be observed that the system state converges to the origin while the actor NN weight converges to the ideal value. However, as expected, the transient period is longer than the known basis function case.

### B. Example 2

In this example, we consider a class of third-order kinematic models described by

\[
\begin{align*}
\dot{p} &= v, \\
\dot{v} &= f_a - bv, \\
\dot{f}_a &= u - f_a + \omega,
\end{align*}
\]

(50)

where \( p, v, \) and \( f_a \) denote the position, velocity, and actuator force, respectively; \( u \) is the control input; \( m, b, \) and \( \tau \) represent the mass, velocity constant, and time constant, respectively; and \( \omega \) is the external disturbance. The above system represent a large class of physical systems, including human motor system, autonomous vehicle, power system, etc [43]. By letting
An ESO-based reinforcement learning and disturbance rejection framework is established for uncertain nonlinear systems having non-simple nominal models. The developed approach compensates for the total uncertainty and approximates the optimal policy for the compensated system simultaneously. Simulation of experience based RL is employed to utilize the nominal model and to relax the requirement of the PE condition. The obtained results provide a novel learning-based solution for the disturbance rejection of uncertain nonlinear systems, especially those having non-simple nominal models, which are quite common in practice. Future research works will be directed at the extension of the results to systems with constraints [44, 45], and the application of the developed approach to robotic systems [46].

REFERENCES

[1] E. O. Neftci and B. B. Averbeck, “Reinforcement learning in artificial and biological systems,” Nature Machine Intelligence, vol. 1, no. 3, pp. 133-143, 2019.
[2] F. L. Lewis and D. Vrabie, “Reinforcement learning and adaptive dynamic programming for feedback control,” IEEE Circuits and Systems Magazine, vol. 09, no. 3, pp. 32-50, 2009.
[3] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, “Optimal and autonomous control using reinforcement learning: A survey,” IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 6, pp. 2042-2062, 2018.
[4] D. Wang, H. He, and D. Liu, “Adaptive critic nonlinear robust control: A survey,” IEEE Transactions on Cybernetics, vol. 47, no. 10, pp. 3429-3451, 2017.

V. CONCLUSION

\[
x_1 = p, \ x_2 = v, \text{ and } x_3 = \frac{f_a - bv}{m}, \text{ the above system can be written as}
\]

\[
\begin{cases}
\dot{x}_1 = x_2, \\
\dot{x}_2 = x_3, \\
\dot{x}_3 = -\frac{b}{\tau m} x_2 - \left(\frac{1}{\tau} + \frac{b}{m}\right) x_3 + \frac{1}{\tau m} u + \frac{1}{\tau m} \omega.
\end{cases}
\]

In the simulation, the values of the model parameters are given by \( m = 1\text{kg}, b = 1\text{N \cdot s/m}, \text{ and } \tau = 0.1s \). Let \( f_0(x) = -\frac{b}{\tau m} x_2 - \left(\frac{1}{\tau} + \frac{b}{m}\right) x_3, \ g_0(x) = \frac{1}{\tau m}, \text{ and } \omega = 0.2\sin(t) \). The cost functional for the nominal system is selected as \( r(x, u_0) = \|x\|^2 + u_0^2 \), and consequently the optimal value function and optimal control policy are given by \( V^*(x) = 2.2669x_1^2 + 2.3580x_2^2 + 0.0470x_3^2 + 3.1390x_1 x_2 + 0.2x_1 x_3 + 0.2534x_2 x_3 \) and \( u^*_0(x) = x_1 + 1.2669 x_2 + 0.4695 x_3 \), respectively. The compact set of interest is set as \( \chi = [-1 \cdot 1] \times [-1 \cdot 1] \times [-5 \cdot 5] \), and the data points to extrapolate the BE are selected on a 5 × 5 × 5 data grid covering the domain \( \chi \). The ESO is designed with \( L = [4 \ 6 \ 4 \ 1]^T \) and \( \varepsilon = 0.01 \). The basis function is selected as \( \phi(x) = [x_1^2 \ x_2^2 \ x_3^2 \ x_1 x_2 \ x_1 x_3 \ x_2 x_3]^T \). The adaptation gains of the RL adaptive laws are the same as those in Example 1.

Simulation is done with initial conditions \( x(0) = [0.5 \ 0.5 \ 4]^T, \ \hat{x}_1(0) \ \hat{x}_2(0) \ \hat{x}_3(0) \)^T = [0 0 0]^T, \ \Gamma(0) = \text{ diag\{100, 100, 100, 100, 100, 100\}, \text{ and the elements of } \Theta_{\phi}(0) \text{ and } \Theta_{\chi}(0) \text{ randomly selected on } [0 \ 2] \). Figs. 11-13 show the simulation results, from which one can see that the developed approach achieves satisfactory performance for the third-order kinematic model.

Fig. 11. Trajectories of the system state and ESO output (Example 2).

Fig. 12. Trajectory of the control input \( u \) (Example 2).

Fig. 13. Trajectories of the actor NN weight (Example 2).
