IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 December 2023, accepted 21 January 2024, date of publication 26 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3358809

RESEARCH ARTICLE

Echo State Network-Based Robust Tracking Control for Unknown Constrained Nonlinear Systems by Using Integral Reinforcement Learning

CHONG LIU, YALUN LI, ZHONGXING DUAN, ZHOUSHENG CHU, AND ZONGFANG MA
College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, Shaanxi 710055, China

Corresponding author: Zhongxing Duan (zhx_duan@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62103316 and Grant 62276207, and in part by the China Postdoctoral Science Foundation under Grant 2022MD713783.

ABSTRACT It is necessary to consider the robustness in the tracking problem, which can effectively suppress the external disturbance to ensure the tracking performance. Different from previous tracking control methods, considering the robustness, completely unknown nonlinear system dynamics and constrained controller, we propose a data-based echo state network (ESN) approximated algorithm for a class of robust tracking problems. First, the robust tracking control problem (RTCP) is transformed into the optimal control problem of the according nominal system by designing a elaborate value function. To obtain the optimal control policy, we have to solve a Hamilton-Jacobi-Bellman equation (HJBE) about the augmented nominal system. It is well-known that modelling the accurate dynamics for the practical engineering applications is usually difficult, so the model-free integral reinforcement learning (IRL) algorithm is used to learn the optimal control policy and performance function simultaneously by only using systems data. In this IRL algorithm, a reservoir computing based ESN is used to approximate the performance function and control input. Contrast to other neural networks, ESN need not consider the choice of activation function, which can greatly reduce the difficulty and effort of neural network structure design. The output weights of the ESNs are iteratively updated towards the optimal ones by using least square algorithm and the pre-collected off-line system data. Then, using the converged output weights and ESNs, the tracking control input can be derived without knowing any system dynamic information. Finally, we demonstrate that the given system can be controlled to track the desired trajectory well under the proposed method by using two simulation examples.

INDEX TERMS Adaptive dynamic programming (ADP), echo state network (ESN), integral reinforcement learning (IRL), robust control, tracking control.

I. INTRODUCTION

In practical engineering applications, external disturbances are inevitable, to enhance the robustness of the system, then we have to consider suppressing disturbances while designing the controllers [1], [2], [3], [4]. System robustness means that the controlled system can maintain a certain performance characteristics in spite of suffering external disturbances. In recent years, various methods have emerged to cope with the robust problem [5]. For example, robust fault-tolerant control based on $H\infty$ method [6], boundary control with output feedback [7] and Multi-H∞ controls for unknown input-interference [8]. Especially in the tracking problem, the external disturbances may cause the system to fail to track the desired trajectory. The robust tracking control problem (RTCP) has been a focal point [9], whose purpose

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

is to design a control policy so that the given system with disturbance signal can track the target trajectory, in control theory [10], [11], [12], [13], [14], [15] and applications, such as vehicle control [16], aircraft control [17], [18] and so on. In addition, due to safety considerations, there are certain physical constraints on the controller in reality [19], [20]. Therefore, it is interesting to study the RTCP of the constrained nonlinear systems.

In the past decades, various methods have been proposed to solve the RTCP. Now, a system transformation method is usually used to solve the tracking problem by constructing a corresponding augmented system of the tracking error and the desired trajectory [21], [22]. Thus the tracking problem of the nonlinear systems can be transformed into an optimal control problem by introducing a performance index. In [11], a robust approximate optimal tracking control problem was studied. Modares et al. [12] proposes the disturbances by using the $H\infty$ theory and optimal control method. Recently, [13] studied the RTCP for the nonlinear-constrained systems by using RL. The aim of optimal control is to design the control policy, which make the system stable and performance index optimal [23], [24]. The key of solving the nonlinear optimal control problem is how to solve the Hamilton-Jacobi-Bellman equation(HJBE). However, HJBE is a nonlinear partial differential equation, which has no analytical solution and is difficult to solve. In the past decades, many scholars have devoted themselves to seeking solutions for the nonlinear optimal control methods, in which adaptive dynamic programming (ADP) [25], [26] and reinforcement learning (RL) are the prominent ones. ADP was first proposed by Werbos in [27], and was based on the dynamic programming proposed by Bellman in [28].

In practical engineering applications, it is difficult to model the accurate dynamics of a real system, especially when there are disturbances in the system. Among the above methods, RL is a type of data-based model-free methods that maximizes the payoff by learning the experiences and updating the strategies during the interaction with the environment [29], [30], [31]. In [32], RL method was proposed and was applied to the linear optimal control problem. In [29], RL method was applied to the nonlinear systems, which is more universal. In [33], Yang et al. proposed an integral RL (IRL) for the robust control problem, which is a great inspiration to this paper. In [34], Yao et al. proposed a model-based IRL algorithm for electrohydraulic position servo systems. Further, to cope with the tracking control problem, [35] used data-driven policy iteration method to search the optimal controller by only using system data.

To implement the RL, it is necessary to design appropriate neural networks (NNs) to approximate the value function and the control policy. Most of the above references used the polynomial NNs, which were composed of a combination of system states. However, the hidden layer structure design of polynomial NN is always a open problem. Especially for augmented systems in the tracking problems, the dimension of the objective system is relatively large, which makes the design of NN more difficult. In this paper, echo state network (ESN) will be used as an approximation to implement the RL algorithm. ESN is a kind of reservoir network as an improved recurrent NN model, which was proposed by Prof. Jaeger et al., in 2001 [36]. The biggest advantage of ESN is that the reservoir of the hidden layer is randomly generated without elaborate design. The ESN consists of input layer, a reservoir, and output layer. In this paper, we use ESN to approximate the value function in HJBE. In addition, ESN only need to train the output weights, which greatly reduces the computation. Therefore, ESN has been successfully adopted in online ADP methods for the optimal control problems [37], [38].

In short, to solve the RTCP of unknown nonlinear systems with constrained input, this paper presents an ESN-based IRL method to obtain a stable closed-loop controller. The proposed method has the following advantages:

1) The RTCP is transformed into an optimal control problem by constructing an augmented system and introducing an auxiliary term into the performance function.
2) To cope with the unknown system problem, IRL method is used to solve the optimal controller by only the system data.
3) ESN is used to implement the IRL method, which greatly reduces the design difficulty and computational burden.

The structure of this paper has six parts. In Section II, problem formulation and preliminaries of the robust tracking is shown. In Section III, the robust controller is obtained by solving a transformed optimal control problem with a novel performance function. Besides, the availability of the proposed method is proved to be uniformly ultimately bounded (UUB). In Section IV, an IRL algorithm is derived. Then, ESN is designed to implement the approximate iteration only by using the system data. In Section V, two simulations is given to verify the validity of the proposed method. A simulation example is shown in Section VI gives a brief conclusion and prospect.

II. PROBLEM FORMULATION AND PRELIMINARIES

Consider the following disturbed continuous-time (CT) affine nonlinear system described by:

$$\dot{x} = f\left(x\left(t\right)\right) + g\left(x\left(t\right)\right) u\left(t\right) + g\left(x\left(t\right)\right) \omega\left(x\left(t\right)\right) \quad (1)$$

where $x\left(t\right) = \left[x_1\left(t\right), \ldots, x_n\left(t\right)\right]^T \in \Omega \in \mathcal{R}^n$ is the system state vector, $u\left(t\right) = \left[u_1\left(t\right), \ldots, u_m\left(t\right)\right]^T \in \mathcal{U} \in \mathcal{R}^m$ is the system control vector, which is constrained by $\left|u_m\left(t\right)\right| \leqslant \sigma, \sigma > 0$, and σ is the saturating bound. $f\left(x\left(t\right)\right) \in \mathcal{R}^n$ and $g\left(x\left(t\right)\right) \in \mathcal{R}^{n\times m}$ are the unknown system dynamics, and $\omega\left(x\left(t\right)\right) \in \mathcal{R}^n$ is the unknown disturbing function. $x_0 = x\left(0\right)$ is the initial state of the system and $f\left(0\right) = 0$. To complete the following development, based on the literature [25], two assumptions are pre-given.

Assumption 1: The given affine nonlinear CT System (1) is controllable and $f(x(t)) + g(x(t))u(t)$ is Lipschitz continuous on the set Ω, i.e., the solution of the CT nonlinear system exists and unique whatever initial state $x_0 \in \Omega$ and $u \in \mathcal{U}$.

Assumption 2: For arbitrary $x \in \mathcal{R}^n$, the input dynamic satisfies $0 < \|g(x)\| < g_{\mathcal{M}}$ with $g_{\mathcal{M}}$ is a given constant. At the same time, the disturbing functions $\omega(x(t))$ is bounded by a given function $\omega_{\mathcal{M}}(x)$, and $\|\omega(x)\| \leqslant \omega_{\mathcal{M}}(x)$ for arbitrary $x \in \mathcal{R}^n$. Furthermore, $\omega(0) = 0$ and $\omega_{\mathcal{M}}(0) = 0$.

The desired trajectories $x^r(t)$ of the affine nonlinear CT system satisfy

$$\dot{x}^r(t) = z\left(x^r(t)\right) \tag{2}$$

where $x^r(t)$ is bounded and $x^r(t) \in \mathcal{R}^n$. The input dynamic $z(x^r(t))$ is Lipschitz continuous function and $z(0) = 0$. Define the system tracking error $e(t)$ as follows:

$$e(t) = x(t) - x^r(t) \tag{3}$$

Taking the derivative of the tracking error (3), according to equations (1) and (2), one has

$$\begin{aligned}\dot{e}(t) =& f(x(t)) + g(x(t))u(t) + g(x(t))\omega(x(t)) \\ &- z\left(x^r(t)\right)\end{aligned} \tag{4}$$

Under Assumptions 1 and 2, the goal of robust tracking control is to minimize the performance function so that original system (1) eventually tracks the desired trajectory (2) and stays stable in the sense of UUB. To address this goal, the RTCP is solved by using a transformed optimal control problem of a nominal system with an appropriate performance function. The so-called nominal system is that

$$\dot{x} = f(x(t)) + g(x(t))u(t) \tag{5}$$

Subsequently, We define the tracking error of the nominal system to be

$$\dot{e}(t) = f(x(t)) + g(x(t))u(t) - z\left(x^r(t)\right) \tag{6}$$

Let $Z(t) = \left[e(t)^T \; x^r(t)^T\right]^T$, the augmented matrix is defined by (6) and (2) as follows

$$\dot{Z} = \mathcal{A}(Z(t)) + \mathcal{B}(Z(t))u(t) \tag{7}$$

where

$$\mathcal{A}(Z(t)) = \begin{bmatrix} f(e(t) + x^r(t)) - z(x^r(t)) \\ z(x^r(t)) \end{bmatrix} \tag{8}$$

$$\mathcal{B}(Z(t)) = \begin{bmatrix} g(e(t) + x^r(t)) \\ \mathbf{O} \end{bmatrix} \tag{9}$$

where $\mathbf{O}$ denotes zero matrix with the corresponding dimension.

We define the tracking performance function of the augmented system (7) as follows:

$$\begin{aligned}V(Z(t)) =& \int_t^\infty e^{-\gamma(\varsigma - t)}\big(\eta\omega_{\mathcal{M}}^2(Z) \\ &+ Z(\varsigma)^T QZ(\varsigma) + \varpi(u(\varsigma))\big)d\varsigma\end{aligned} \tag{10}$$

where γ is a discount factor and $\gamma > 0$, to ensure the boundedness of the performance index function. $Q = \operatorname{diag}\{Q_e, \mathbf{O}_{\mathrm{n}\times\mathrm{n}}\}$, Q_e is a symmetric constant positive definite matrix about augmented system state Z and $Q_e \in \mathcal{R}^{\mathrm{n}\times\mathrm{n}}$, η is a constant parameter and $\eta > 0$, to solve the bounded control problem in the system, we define $\varpi(u(\mathrm{t}))$ as follows:

$$\varpi(u(\mathrm{t})) = 2\sigma\int_0^{\mathrm{u(t)}}\left(\tanh^{-1}(\varsigma/\sigma)\right)^T R d\varsigma \tag{11}$$

where $R = \operatorname{diag}(r_1, \ldots, r_m)$ and $r_i > 0, i = 1, 2, 3\cdots m$. $\varpi(u(\mathrm{t}))$ is a non-quadratic cost index function, which can cope with the input constraint problem. The term w_M^2 is used to suppress the disturbances and ensure the tracking error (4) to be UUB, which can be shown in Theorem 1.

Definition 1 (Admissible Control): For a practical control problem, there exists an admissible control set Ω. A controller $u(Z)$ is called to be admissible on Ω if $u(Z)$ is continuous, $u(0) = 0$, and stabilizes the system (7) with the performance function (10) is finite for $\forall \mathrm{Z} \in \Omega$. Then, it is formulated as $u(Z) \in \pi(\Omega)$.

For ease of presentation, the time variate t will be simplified in the latter part.

III. ROBUST TRACKING CONTROLLER DESIGN

Based on the augmented system (7) and the new tracking performance function (10), differentiate $V(Z)$, one has

$$\begin{aligned}\dot{V}(Z) =& \gamma\int_t^\infty e^{-\gamma(\varsigma - t)}\Big(\eta\omega_{\mathcal{M}}^2(Z) \\ &+Z(\varsigma)^T QZ(\varsigma) + \varpi(u(\varsigma))\Big)d\varsigma \\ &- \eta\omega_{\mathcal{M}}^2(Z) - Z^T QZ - \varpi(u) \\ =& \gamma V(Z) - \eta\omega_{\mathcal{M}}^2(Z) - Z^T QZ - \varpi(u)\end{aligned} \tag{12}$$

We present the Hamiltonian function as follows

$$\begin{aligned}\mathcal{H}(\mathrm{Z}, \mathrm{u}, \nabla\mathrm{V}) \equiv& \eta\omega_{\mathcal{M}}^2(Z) + Z^T QZ + \varpi(u) - \gamma V(Z) \\ &+ \nabla\mathrm{V}^T(Z)(\mathcal{A}(Z(t)) + \mathcal{B}(Z(t))u(\mathrm{Z})) \\ =& 0\end{aligned} \tag{13}$$

where $\nabla V = \partial V(Z)/\partial Z$. Let $V^*(Z)$ denotes the optimal value function as follows:

$$\begin{aligned}V^*(Z) = \min_{u\in\mathcal{U}}& \int_t^\infty e^{-\gamma(\varsigma - t)}\Big(\eta\omega_{\mathcal{M}}^2(Z) \\ &+Z(\varsigma)^T QZ(\varsigma) + \varpi(u(\varsigma))\Big)d\varsigma\end{aligned} \tag{14}$$

Substituting (14) into (13), the optimal robust tracking HJBE is derived as

$$\begin{aligned}&\mathcal{H}\left(\mathrm{Z}, \mathrm{u}^*, \nabla\mathrm{V}^*\right) \\ &\quad\equiv \eta\omega_{\mathcal{M}}^2(Z) + Z^T QZ + \varpi(u) - \gamma V^*(Z) \\ &\quad\quad+ \nabla\mathrm{V}^{*T}(Z)\left(\mathcal{A}(Z) + \mathcal{B}(Z)u^*(Z)\right)\end{aligned} \tag{15}$$

According to the literature [22], the analytical solution for optimal control is as follows:

$$u^* = -\sigma\tanh\left((1/2\sigma)R^{-1}\mathcal{B}^T(Z)\nabla\mathrm{V}^*(\mathrm{Z})\right) \tag{16}$$

According to [35], the optimal control u^* in (16) can minimize value function V^* and guarantee the tracking error (3) converges to zero while $\gamma \to 0$.

Theorem 1: Consider the nominal system (5) of the disturbed system (1) with related HJBE (15). If the following conditions hold:

$$\omega^T(Z) R\omega(Z) \le \eta\omega_{\mathcal{M}}^2(Z) \tag{17}$$

and

$$\gamma \to 0, \tag{18}$$

the optimal control policy u^* given in (16) can ensure the tracking error (4) to be UUB.

Proof: Considering the solution V^* of HJBE (15), we can obtain that $V^*(x) > 0$ for arbitrary $x \neq 0$ and $V^*(x) = 0$ for $x = 0$, $\nabla \mathrm{V}^*$ is bounded, denoting as $\|\nabla \mathrm{V}^*\| \le \mathcal{V}_{\mathcal{M}}$ with $\mathcal{V}_{\mathcal{M}} > 0$. Differentiating $V^*(Z)$ along the trajectory of the augmented system (7), one has

$$\begin{aligned}\frac{dV^*(Z)}{dt} &= \nabla V^{*T}(Z)\left(\mathcal{A}(Z) + \mathcal{B}(Z)\left(u(Z) + \omega(Z)\right)\right)\\ &= \nabla V^{*T}(Z)\left(\mathcal{A}(Z) + \mathcal{B}(Z)u(Z)\right)\\ &\quad + \nabla V^{*T}(Z)\mathcal{B}(Z)\omega(Z)\end{aligned} \tag{19}$$

Consider the control policy given in (16)

$$\begin{aligned}u(Z) &= u^*(Z)\\ &= -\sigma\tanh\left((1/2\sigma)R^{-1}\mathcal{B}^T(Z)\nabla \mathrm{V}^{*T}(Z)\right)\end{aligned} \tag{20}$$

According to the Hamiltonian function (13), we have

$$\begin{aligned}&\nabla V^{*T}(Z)\left(\mathcal{A}(Z) + \mathcal{B}(Z)u(Z)\right)\\ &\quad = \gamma V^*(Z) - \eta\omega_{\mathcal{M}}^2(Z) - Z^TQZ - \varpi\left(u^*\right)\end{aligned} \tag{21}$$

According to the optimal control solution (16), one has

$$\begin{aligned}&\nabla V^{*T}(Z)\mathcal{B}(Z)\omega(Z)\\ &\quad = -2\sigma\left(\tanh^{-1}\left(u^*/\sigma\right)\right)^T R\omega(Z)\end{aligned} \tag{22}$$

Combining (21) and (22), we can derive that

$$\begin{aligned}\frac{dV^*(Z)}{dt} &= \gamma V^*(Z) - \eta\omega_{\mathcal{M}}^2(Z) - Z^TQZ - \varpi\left(u^*\right)\\ &\quad - 2\sigma\left(\tanh^{-1}\left(u^*/\sigma\right)\right)^T R\omega(Z)\end{aligned} \tag{23}$$

$$\begin{aligned}&\frac{dV^*(Z)}{dt} - \gamma V^*(Z)\\ &\quad = -\eta\omega_{\mathcal{M}}^2(Z) - Z^TQZ - \varpi\left(u^*\right)\\ &\qquad - 2\sigma\left(\tanh^{-1}\left(u^*/\sigma\right)\right)^T R\omega(Z)\end{aligned} \tag{24}$$

Multiplying $e^{-\gamma t}$ to both parts of the equation (24) and considering $\frac{d\left(e^{-\gamma t}V^*(Z)\right)}{dt} = e^{-\gamma t}\left(\frac{dV^*(Z)}{dt} - \gamma V^*(Z)\right)$, one has

$$\begin{aligned}&\frac{d\left(e^{-\gamma t}V^*(Z)\right)}{dt}\\ &\quad = e^{-\gamma t}\Big(- \eta\omega_{\mathcal{M}}^2(Z) - Z^TQZ - \varpi\left(u^*\right)\\ &\qquad - 2\sigma\left(\tanh^{-1}\left(u^*/\sigma\right)\right)^T R\omega(Z)\Big)\end{aligned} \tag{25}$$

According to equation (11), one has

$$\begin{aligned}\varpi\left(u^*\right) &= 2\sigma\int_0^{\mathrm{u}^*}\left(\tanh^{-1}(\varsigma/\sigma)\right)^T R d\varsigma\\ &= 2\sigma\sum_{p=1}^{m}\int_0^{\mathrm{u}^*} r_p\tanh^{-1}(\varsigma/\sigma)\, d\varsigma\end{aligned} \tag{26}$$

Briefly, let $\varrho = \tanh^{-1}(\varsigma/\sigma)$, we have

$$\begin{aligned}\varpi\left(u^*\right) &= 2\sigma^2\sum_{p=1}^{m}\int_0^{\tanh^{-1}(\mathrm{u}^*/\sigma)} r_p\varrho\left(1 - \tanh^2(\varrho)\right)d\varrho\\ &= 2\sigma^2\sum_{p=1}^{m}\int_0^{\tanh^{-1}(\mathrm{u}^*/\sigma)} r_p\varrho d\varrho\\ &\quad - 2\sigma^2\sum_{p=1}^{m}\int_0^{\tanh^{-1}(\mathrm{u}^*/\sigma)} r_p\varrho\tanh^2(\varrho)\, d\varrho\\ &= \sigma^2\sum_{p=1}^{m} r_j\left(\tanh^{-1}\left(\mathrm{u}^*/\sigma\right)\right)^2\\ &\quad - 2\sigma^2\sum_{p=1}^{m}\int_0^{\tanh^{-1}(\mathrm{u}^*/\sigma)} r_p\varrho\tanh^2(\varrho)\, d\varrho\end{aligned} \tag{27}$$

Denoting $\omega(Z) = [\omega_1(Z), \ldots, \omega_m(Z)]^T \in \mathcal{R}^m$ with $\omega_p(Z) \in \mathcal{R}\quad p \in \{1, 2, 3\ldots, m\}$, we have

$$\begin{aligned}&-2\sigma\left(\tanh^{-1}\left(u^*/\sigma\right)\right)^T R\omega(Z)\\ &\qquad = -2\sigma\sum_{p=1}^{m} r_p\tanh^{-1}\left(u^*/\sigma\right)\omega(Z)\end{aligned} \tag{28}$$

$$\sum_{p=1}^{m} r_p\omega^2(Z) = \omega^T(Z) R\omega(Z) \tag{29}$$

Substituting (27), (28) and (29) into (25), let

$$\begin{aligned}\varLambda &= -\eta\omega_{\mathcal{M}}^2(Z) - Z^TQZ\\ &\quad + 2\sigma^2\sum_{p=1}^{m}\int_0^{\tanh^{-1}(\mathrm{u}^*/\sigma)} r_p\varrho\tanh^2(\varrho)\, d\varrho\\ &\quad - \sigma^2\sum_{p=1}^{m} r_p\left(\tanh^{-1}\left(\mathrm{u}^*/\sigma\right)\right)^2 + \sum_{p=1}^{m} r_p\omega^2(Z)\\ &\quad - \sum_{p=1}^{m} r_p\omega^2(Z) - 2\sigma\sum_{p=1}^{m} r_p\tanh^{-1}\left(u^*/\sigma\right)\omega(Z)\\ &= -\eta\omega_{\mathcal{M}}^2(Z) - Z^TQZ + \sum_{p=1}^{m} r_p\omega^2(Z)\\ &\quad + 2\sigma^2\sum_{p=1}^{m}\int_0^{\tanh^{-1}(\mathrm{u}^*/\sigma)} r_p\varrho\tanh^2(\varrho)\, d\varrho\\ &\quad - \sum_{p=1}^{m} r_p\left(\sigma\tanh^{-1}\left(\mathrm{u}^*/\sigma\right) + \omega(Z)\right)^2\end{aligned}$$

$$\leq -\eta\omega_{\mathcal{M}}^{2}(Z) - Z^{T}QZ + \sum_{p=1}^{m} r_{p}\omega^{2}(Z) + 2\sigma^{2}\sum_{p=1}^{m}\int_{0}^{\tanh^{-1}(\mathrm{u}^{*}/\sigma)} r_{p}\varrho\tanh^{2}(\varrho)\,d\varrho \tag{30}$$

What's more, according to the integral mean-value theorem, one has

$$2\sigma^{2}\sum_{p=1}^{m}\int_{0}^{\tanh^{-1}(\mathrm{u}^{*}/\sigma)} r_{p}\varrho\tanh^{2}(\varrho)\,d\varrho = 2\sigma^{2}\sum_{p=1}^{m} r_{p}\tanh^{-1}\left(\mathrm{u}^{*}/\sigma\right)\varepsilon\tanh^{2}(\varepsilon) \tag{31}$$

where $\varepsilon \in (0, \tanh^{-1}(\mathrm{u}^{*}/\sigma))$, $2\sigma^{2}\sum_{p=1}^{m}\int_{0}^{\tanh^{-1}(\mathrm{u}^{*}/\sigma)} r_{p}\varrho\tanh^{2}(\varrho)\,d\varrho > 0$, and $0 < \tanh^{2}(\varepsilon) \leq 1$. Besides,

$$\begin{aligned}
&2\sigma^{2}\sum_{p=1}^{m}\int_{0}^{\tanh^{-1}(\mathrm{u}^{*}/\sigma)} r_{p}\varrho\tanh^{2}(\varrho)\,d\varrho\\
&\leq 2\sigma^{2}\sum_{p=1}^{m} r_{p}\tanh^{-1}\left(\mathrm{u}^{*}/\sigma\right)\varepsilon\\
&\leq 2\sigma^{2}\sum_{p=1}^{m} r_{p}\left(\tanh^{-1}\left(\mathrm{u}^{*}/\sigma\right)\right)^{2}\\
&= 2\sigma^{2}\left(\tanh^{-1}\left(\mathrm{u}^{*}/\sigma\right)\right)^{T} R\left(\tanh^{-1}\left(\mathrm{u}^{*}/\sigma\right)\right)\\
&= \frac{1}{2}\left(\nabla V^{*}\right)^{T}\mathcal{B}(Z)R^{-1}\mathcal{B}^{T}(Z)\nabla V^{*}\\
&\leq \frac{1}{2}\lambda_{\max}\left(R^{-1}\right)\mathcal{B}_{\mathcal{M}}^{2}\mathcal{V}_{\mathcal{M}}^{2}
\end{aligned} \tag{32}$$

where $\lambda_{\max}\left(R^{-1}\right)$ is the largest eigenvalue of R^{-1}, $\mathcal{B}_{\mathcal{M}}$ denotes the upper bound of $\mathcal{B}(Z)$.

Considering the given assumption (17), according to (30) and (32), one has

$$\begin{aligned}
\frac{d\left(e^{-\gamma t}V^{*}(Z)\right)}{dt} &\leq e^{-\gamma t}\Big(-\lambda_{\min}(Q)\|Z\|^{2} + \sum_{p=1}^{m} r_{p}\omega^{2}(Z)\\
&\quad + \frac{1}{2}\lambda_{\max}\left(R^{-1}\right)\mathcal{B}_{\mathcal{M}}^{2}\mathcal{V}_{\mathcal{M}}^{2} - \eta\omega_{\mathcal{M}}^{2}(Z)\Big)\\
&\leq e^{-\gamma t}\Big(-\lambda_{\min}(Q)\|Z\|^{2}\\
&\quad + \frac{1}{2}\lambda_{\max}\left(R^{-1}\right)\mathcal{B}_{\mathcal{M}}^{2}\mathcal{V}_{\mathcal{M}}^{2}\Big)
\end{aligned} \tag{33}$$

According to [22] and [35], when $\gamma \to 0$, we have

$$\frac{d\left(V^{*}(Z)\right)}{dt} \leq -\lambda_{\min}(Q)\|Z\|^{2} + \frac{1}{2}\lambda_{\max}\left(R^{-1}\right)\mathcal{B}_{\mathcal{M}}^{2}\mathcal{V}_{\mathcal{M}}^{2} \tag{34}$$

Obviously, $\frac{d(V^{*}(Z))}{dt} < 0$ when Z is out of the set Ψ_{Z}:

$$\Psi_{Z} = \left\{ Z : \|Z\| \leq \mathcal{B}_{\mathcal{M}}\mathcal{V}_{\mathcal{M}}\sqrt{\frac{\lambda_{\max}\left(R^{-1}\right)}{2\lambda_{\min}(Q)}} \right\} \tag{35}$$

By the Lyapunov extension theorem we can prove that the trajectory of the augmented system (7) is UUB under the optimal control u^{*} (16) with the bound $\mathcal{B}_{\mathcal{M}}\mathcal{V}_{\mathcal{M}}\sqrt{\frac{\lambda_{\max}(R^{-1})}{2\lambda_{\min}(Q)}}$. In other words, the tracking error (3) is UUB.

IV. ESN-BASED IRL ALGORITHM FOR THE SOLUTION

In this part, an ESN-based IRL algorithm is proposed to search the solution of the HJBE (15). In the first part, the off-policy algorithm is shown. In second part, a date-based IRL algorithm is introduced. Finally, ESN is designed to complete the IRL algorithm, which reduces the design complexity and the computation burden.

A. BASIC OFF-POLICY ITERATION ALGORITHM

On-policy iteration and off-policy iteration both are reinforcement learning methods. In on-policy iteration algorithm, the next policy relies on the currently improved policy and the performance function is evaluated by using the system data generated by that improved policy. However, the data is inaccurate, which results in an increasing approximation error. Different from the on-policy iteration, in off-policy iteration algorithm, the next policy do not rely on the currently improved policy and the performance function is evaluated by using the system data generated by an arbitrary control [39], [40], [41]. This paper uses the off-policy iteration algorithm, which includes two parts: policy evaluation and policy improvement.

Algorithm 1. (Steps of the off-policy iteration)

- Step 1: Initialization. Initialize the admissible policy $u^{(0)}(Z)$ and the calculation accuracy $\iota > 0$ at step $\ell = 0$.
- Step 2: Policy evaluation by compute the performance function $V^{(\ell)}(Z)$.

$$\eta\omega_{\mathcal{M}}^{2}(Z) + Z^{T}QZ + \varpi\left(u^{(\ell)}\right) - \gamma V^{(\ell)}(Z) + \left(\nabla \mathrm{V}^{(\ell)}(Z)\right)^{T}\left(\mathcal{A}(Z) + \mathcal{B}(Z)u^{(\ell)}(Z)\right) = 0 \tag{36}$$

where $\varpi\left(u^{(\ell)}\right) = 2\sigma\int_{0}^{u^{(\ell)}}\left(\tanh^{-1}(\varsigma/\sigma)\right)^{T}Rd\varsigma$.

- Step 3: Policy improvement via

$$u^{(\ell+1)}(Z) = -\sigma\tanh\left((1/2\sigma)R^{-1}\mathcal{B}^{T}(Z)\nabla \mathrm{V}^{(\ell)}(Z)\right) \tag{37}$$

- Step 4: Termination Conditions. Stop the iteration when $\left\|V^{(\ell+1)}(Z) - V^{(\ell)}(Z)\right\| \leq \iota$ for each $Z \in \Upsilon$. The final corresponding policy is approximately optimal. If not, let $\ell = \ell + 1$ and return to the Step 2.

The convergence of the policy iterations has been shown in [23].

B. DATA-BASED IRL ALGORITHM

Algorithm 1 indicates that system dynamics $\mathcal{A}$ and $\mathcal{B}$ are required. When the system dynamic is unknown, we will present a data-based IRL algorithm.

Adding an auxiliary variate $u^{(\ell)}(Z)$ to the augmented system (7), we can obtain

$$\begin{aligned}\dot{Z} &= \mathcal{A}(Z) + \mathcal{B}(Z)\, u^{(\ell)}(Z)\\ &\quad + \mathcal{B}(Z)\left(u(Z) - u^{(\ell)}(Z)\right)\end{aligned} \tag{38}$$

Taking the derivative of $V^{(\ell+1)}(Z)$ along the augmented system (38), one has

$$\begin{aligned}\frac{dV^{(\ell)}(Z)}{dt} &= \left(\nabla V^{(\ell)}\right)^T \Big(\mathcal{A}(Z) + \mathcal{B}(Z(t))\, u^{(\ell)}(Z)\\ &\qquad + \mathcal{B}(Z)\left(u(Z) - u^{(\ell)}(Z)\right)\Big)\\ &= \left(\nabla V^{(\ell)}\right)^T \left(\mathcal{A}(Z) + \mathcal{B}(Z)\, u^{(\ell)}(Z)\right)\\ &\quad + \left(\nabla V^{(\ell)}\right)^T \left(\mathcal{B}(Z)\left(u(Z) - u^{(\ell)}(Z)\right)\right)\end{aligned} \tag{39}$$

According to equations (36) and (37) in Algorithm 1, we can obtain

$$\begin{aligned}&\left(\nabla V^{(\ell)}(Z)\right)^T \left(\mathcal{A}(Z) + \mathcal{B}(Z)\, u^{(\ell)}(Z)\right)\\ &\quad = -\eta\omega_{\mathcal{M}}^2(Z) - Z^T Q Z - \varpi\left(u^{(\ell)}\right) + \gamma V^{(\ell)}(Z)\\ &\left(\nabla V^{(\ell)}(Z)\right)^T \mathcal{B}(Z)\\ &\quad = -2\sigma\left(\tanh^{-1}\left(u^{(\ell+1)}/\sigma\right)\right)^T R\end{aligned} \tag{40}$$

Substituting the above two equations into (39), we can obtain

$$\begin{aligned}\frac{dV^{(\ell)}(Z)}{dt} &= -\eta\omega_{\mathcal{M}}^2(Z) - Z^T Q Z - \varpi\left(u^{(\ell)}\right)\\ &\quad + \gamma V^{(\ell)}(Z) - 2\sigma\left(\tanh^{-1}\left(u^{(\ell+1)}/\sigma\right)\right)^T R\end{aligned} \tag{41}$$

Integrating both sides of (41) over the time interval $[t, t+\Delta t]$, one has

$$\begin{aligned}&V^{(\ell)}(Z(t+\Delta t)) - V^{(\ell)}(Z(t))\\ &= -\int_t^{t+\Delta t}\Big(\eta\omega_{\mathcal{M}}^2(Z(\tau))\\ &\quad + Z^T(\tau) Q Z(\tau) + \varpi\left(u^{(\ell)}(\tau)\right) - \gamma V^{(\ell)}(Z(\tau))\Big)d\tau\\ &\quad - \int_t^{t+\Delta t}\Big(2\sigma\left(\tanh^{-1}\left(u^{(\ell+1)}(\tau)/\sigma\right)\right)^T\\ &\quad \times R\left(u(\tau) - u^{(\ell)}(\tau)\right)\Big)d\tau\end{aligned} \tag{42}$$

Algorithm 2. (Steps of the data-based IRL)

- Step 1: Collecting data. Given a calculation accuracy $\iota > 0$, let $\ell = 0$, select the initial admissible control $u^{(0)}(Z)$.
- Step 2: Let $\ell > 0$, according to control policy $u^{(\ell)}(Z)$, simultaneously solving for $V^{(\ell)}(Z)$ and $u^{(\ell+1)}(Z)$ from equation (42).
- Step 3: If $\left\|V^{(\ell+1)} - V^{(\ell)}\right\| \le \iota$ for each $Z \in \Upsilon$. The final corresponding policy is approximately optimal. If not, let $\ell = \ell + 1$ and return to the Step 2.

Differ from the off-policy algorithm, the data-based IRL algorithm can iterate and compute the performance function $V^{(\ell)}(Z)$ and the control policy $u^{(\ell+1)}(Z)$, simultaneously. We can find that the proposed data-based IRL algorithm does not require any system dynamics information.

C. IMPLEMENTATION OF THE IRL ALGORITHM BY USING ESN

In this subsection, an ESN-based actor-critic architecture is used to implement the IRL algorithm. First, performance function $V^{(\ell)}(Z)$ is approximated by ESN as follows

$$\dot{\mathcal{Z}}_1(t) = \frac{1}{\mathfrak{B}_1}\left(-\kappa_1 \mathcal{Z}_1(t) + \Phi_1\left(W_{in1} Z(t) + W_1 \mathcal{Z}_1(t)\right)\right) \tag{43}$$

$$\hat{V}^{(\ell)}(Z) = \left(W_{out1}^{(\ell)}\right)^T [Z(t); \mathcal{Z}_1(t)] \tag{44}$$

where augmented system states $Z(t)$ is used as the input of ESN. $\mathcal{Z}_1(t)$ is the reservoir states. Parameters $\kappa_1 > 0$, $\mathfrak{B}_1 > 0$ and $\Phi_1(\cdot)$ denote the leaky rate, time constant and active function, respectively. Weight matrixes $W_{in1} \in \mathcal{R}^{p_1 \times 2n}$, $W_1 \in \mathcal{R}^{p_1 \times p_1}$ and $W_{out1} \in \mathcal{R}^{1 \times (2n+p_1)}$ link the input vector, reservoir states and critic ESN output vector, respectively. $[\cdot;\cdot]$ denotes the concatenation operation between two vectors. Among these weight matrixes, only the output weight $W_{out1} \in \mathcal{R}^{1 \times (2n+p_1)}$ needs to be trained, and the other matrices are randomly generated according to the given sparsity. Unlike polynomial NN, ESN do not need to elaborately select the hidden layers and only need to train the output weights, which greatly reduces the design difficulty and computational burden.

Let $\Theta(Z) = [Z(t); \mathcal{Z}_1(t)]$, we have

$$\hat{V}^{(\ell)}(Z) = \left(W_{out1}^{(\ell)}\right)^T \Theta(Z) \tag{45}$$

To solve the constraint input problem, inspired by the literature [33], an intermediate variable is defined as

$$\mu^{(\ell)}(Z) = \tanh^{-1}\left(u^{(\ell)}(Z)/\sigma\right) \tag{46}$$

Applying the ESN to approximate this intermediate variable (46), one has

$$\dot{\mathcal{Z}}_{2p}(t) = \frac{1}{\mathfrak{B}_{2p}}\left(-\kappa_{2p}\mathcal{Z}_{2p}(t) + \Phi_{2p}\left(W_{in2p} Z(t) + W_{2p}\mathcal{Z}_{2p}(t)\right)\right) \tag{47}$$

$$\hat{\mu}_p^{(\ell)}(Z) = \left(W_{out2p}^{(\ell)}\right)^T \left[Z(t); \mathcal{Z}_{2p}(t)\right] \tag{48}$$

The settings of ESN parameters are similar to the performance function in (43) and (44), which will not be described here. It should be noted here that $W_{out2p}^{(\ell)}$ is the actor ESN output weight vector. Let $\psi_p^{(\ell)}(Z) = \left[Z(t); \mathcal{Z}_{2p}(t)\right]$, then $\hat{\mu}_p^{(\ell)}(Z)$ is deduced to be

$$\hat{\mu}_p^{(\ell)}(Z) = \left(W_{out2p}^{(\ell)}\right)^T \psi_p^{(\ell)}(Z) \tag{49}$$

where $\hat{\mu}_p^{(\ell)}(Z)$ means the p-th intermediate variable of the control input, $p \in \{1, 2, 3\ldots, m\}$. So the control policy is formulated as

$$\begin{aligned}\hat{u}^{(\ell)}(Z) &= \sigma \tanh\left(\hat{\mu}^{(\ell)}(Z)\right)\\ &= \left[\sigma \tanh\left(\hat{\mu}_1^{(\ell)}(Z)\right), \ldots, \sigma \tanh\left(\hat{\mu}_m^{(\ell)}(Z)\right)\right]^T\\ &= \left[\sigma \tanh\left(\left(W_{out21}^{(\ell)}\right)^T \psi_1^{(\ell)}(Z)\right), \ldots\right.\\ &\quad \left.\sigma \tanh\left(\left(W_{out2m}^{(\ell)}\right)^T \psi_m^{(\ell)}(Z)\right)\right]^T \end{aligned} \tag{50}$$

Substituting (50) into (11), we have

$$\begin{aligned}\varpi\left(u^{(\ell)}\right) &= 2\sigma \int_0^{u^{(\ell)}} \left(\tanh^{-1}(\varsigma/\sigma)\right)^T R d\varsigma\\ &= 2\sigma \int_0^{\sigma \tanh\left(\mu^{(\ell)}(Z)\right)} \left(\tanh^{-1}(\varsigma/\sigma)\right)^T R d\varsigma \end{aligned} \tag{51}$$

Then the final formulation of IRL algorithm (42) is

$$\begin{aligned}&V^{(\ell)}(Z(t+\Delta t)) - V^{(\ell)}(Z(t))\\ &= -\int_t^{t+\Delta t}\left(\eta\omega_{\mathcal{M}}^2(Z(\tau)) + Z^T(\tau)QZ(\tau) - \gamma V^{(\ell)}(Z(\tau))\right.\\ &\quad \left.+ 2\sigma \int_0^{\sigma \tanh\left(\mu^{(\ell)}(Z)\right)} \left(\tanh^{-1}(\varsigma/\sigma)\right)^T R d\varsigma\right)d\tau\\ &\quad - \int_t^{t+\Delta t}\left(2\sigma\left(u^{(\ell+1)}(\tau)\right)^T\right.\\ &\quad \left.\times R\left(u(\tau) - \sigma \tanh\left(\mu^{(\ell)}(Z(\tau))\right)\right)\right)d\tau \end{aligned} \tag{52}$$

According to the integral operation of the inverse hyperbolic tangent, we have

$$\begin{aligned}&2\sigma \int_0^{\sigma \tanh\left(\mu^{(\ell)}(Z)\right)} \left(\tanh^{-1}(\varsigma/\sigma)\right)^T R d\varsigma\\ &\quad = 2\sigma^2\left(\mu^{(\ell)}(Z)\right)^T R \tanh\left(\mu^{(\ell)}(Z)\right)\\ &\quad\quad + \sigma^2 \sum_{p=1}^m r_p \ln\left(1 - \tanh^2\left(\mu_p{}^{(\ell)}(Z)\right)\right) \end{aligned} \tag{53}$$

So (52) can be derived as

$$\begin{aligned}&V^{(\ell)}(Z(t+\Delta t)) - V^{(\ell)}(Z(t))\\ &= -\int_t^{t+\Delta t} \eta\omega_{\mathcal{M}}^2(Z(\tau)) + Z^T(\tau)QZ(\tau) - \gamma V^{(\ell)}(Z(\tau))\, d\tau\\ &\quad - \int_t^{t+\Delta t} 2\sigma^2\left(\mu^{(\ell)}(Z(\tau))\right)^T R \tanh\left(\mu^{(\ell)}(Z(\tau))\right)d\tau\\ &\quad - \int_t^{t+\Delta t} \sigma^2 \sum_{p=1}^m r_p \ln\left(1 - \tanh^2\left(\mu_p^{(\ell)}(Z(\tau))\right)\right)d\tau\\ &\quad - \int_t^{t+\Delta t}\left(2\sigma\left(\mu^{(\ell)}(Z(\tau))\right)^T\right.\\ &\quad \left.\times R\left(u(\tau) - \sigma \tanh\left(\mu^{(\ell)}(Z(\tau))\right)\right)\right)d\tau \end{aligned} \tag{54}$$

When the performance function and the control policy are not optimal, we define the approximated values as $\hat{V}$ and $\hat{\mu}$. We substitute the approximated values into the data-based IRL algorithm (54), there is the residual error ζ as follows:

$$\begin{aligned}&\zeta(Z)\\ &= \left(W_{out1}^{(\ell)}\right)^T \Theta(Z(t)) - \left(W_{out1}^{(\ell)}\right)^T \Theta(Z(t+\Delta t))\\ &\quad - \int_t^{t+\Delta t}\left(\eta\omega_{\mathcal{M}}^2(Z(\tau)) + Z^T(\tau)QZ(\tau)\right.\\ &\quad \left.- \gamma\left(W_{out1}^{(\ell)}\right)^T \Theta(Z(\tau))\right)d\tau\\ &\quad - \int_t^{t+\Delta t}\left(2\sigma^2\left(\left(W_{out2}^{(\ell)}\right)^T \psi^{(\ell)}(Z(\tau))\right)^T\right.\\ &\quad \left.\times R \tanh\left(\left(W_{out2}^{(\ell)}\right)^T \psi^{(\ell)}(Z(\tau))\right)\right)d\tau\\ &\quad - \int_t^{t+\Delta t} \sigma^2 \sum_{p=1}^m r_p \ln\left(1 - \tanh^2\left(\left(W_{out2p}^{(\ell)}\right)^T \psi_p^{(\ell)} Z)\right)\right)d\tau\\ &\quad - 2\sigma \sum_{p=1}^m r_p \int_t^{t+\Delta t}\left(\left(W_{out2p}^{(\ell)}\right)^T \psi_p^{(\ell)}(Z(\tau))\right)^T\\ &\quad \times\left(u_p(\tau) - \sigma \tanh\left(\left(W_{out2p}^{(\ell)}\right)^T \psi_p^{(\ell)}(Z(\tau))\right)\right)d\tau \end{aligned} \tag{55}$$

Let

$$\begin{aligned}&\mathrm{J}(Z)\\ &= (\Theta(Z(t)) - \Theta(Z(t+\Delta t)))^T\\ &\quad + \int_t^{t+\Delta t} \gamma\Theta(Z(t(\tau)))^T d\tau\\ &\mathrm{K}(Z)\\ &= 2\sigma r_p \int_t^{t+\Delta t}\left(u(\tau)\right.\\ &\quad \left.- \sigma \tanh\left(\left(W_{out2p}^{(\ell)}\right)^T \psi_p^{(\ell)}(Z(\tau))\right)\right)\left(\psi_p^{(\ell)}(Z(\tau))\right)d\tau\\ &\mathrm{L}\left(Z, \hat{u}^{(\ell)}\right)\\ &= \int_t^{t+\Delta t} \eta\omega_{\mathcal{M}}^2(Z(\tau)) + Z^T(\tau)QZ(\tau)d\tau\\ &\quad + \int_t^{t+\Delta t} 2\sigma^2\left(\left(W_{out2}^{(\ell)}\right)^T \psi^{(\ell)}(Z(\tau))\right)^T\\ &\quad \times R \tanh\left(\left(W_{out2}^{(\ell)}\right)^T \psi^{(\ell)}(Z(\tau))\right)d\tau\\ &\quad + \int_t^{t+\Delta t} \sigma^2 \sum_{p=1}^m r_p \ln\left(1 - \tanh^2\left(\left(W_{out2p}^{(\ell)}\right)^T \psi_p^{(\ell)} Z)\right)\right)d\tau\\ &\Xi(Z)\\ &= \left[\mathrm{J}(Z)^T, \mathrm{K}_1(Z)^T, \ldots, \mathrm{K}_m(Z)^T\right]^T \in \mathcal{R}^{1\times(N+mM)} \end{aligned}$$

$$\begin{aligned} &W_{out}^{(\ell)} \\ &= \left[\left(W_{out1}^{(\ell)}\right)^T, \left(W_{out2}^{(\ell)}\right)^T, \ldots, \left(W_{outm}^{(\ell)}\right)^T\right]^T \\ &\zeta(Z) \\ &= \varXi(Z) W_{out}^{(\ell)} - \mathrm{L}\left(Z, \hat{u}^{(\ell)}\right) \end{aligned} \tag{56}$$

To obtain the optimal ESN output weights, it is necessary to collect enough system data. Consider the q-th sampled data, we have

$$\zeta^{[q]}(Z) = \varXi^{[q]}(Z) W_{out}^{(\ell)} - \mathrm{L}^{[q]}\left(Z, \hat{u}^{(\ell)}\right) \tag{57}$$

According to [33], the output weights can be iterated by using the least squares method as follows

$$\begin{aligned} W_{out}^{(\ell)} = &\left(\sum_{q=1}^{S}\left(\left(\zeta^{[q]}(Z)\right)^T \zeta^{[q]}(Z)\right)\right)^{-1} \\ &\times \sum_{q=1}^{S}\left(\zeta^{[q]}(Z)\right)^T \mathrm{L}^{[q]}\left(Z, \hat{u}^{(\ell)}\right) \end{aligned} \tag{58}$$

By using the (58), we can obtain the optimal weights for critic ESN and actor ESN until $\left\|V^{(\ell+1)} - V^{(\ell)}\right\| \leq \iota$ and stop the iteration. Then we can acquire the optimal control policy by using equations (48) and (50).

Remark 1: To implement (58), we need to collect enough data from the system (38), which is generally chosen to be at least $N + mM$. To guarantee the approximate accuracy, according to [42], the size of the reservoir of the ESN should be at least equal to the dimension of the system states. In order to improve the approximation accuracy, it is suggested that the size of the reservoir should be more than twice of the system states dimension.

V. SIMULATIONS

The two simulations, one is about linear system and another is about nonlinear system, are given in the following.

A. LINEAR SYSTEM SIMULATION

Consider the disturbed well-known spring, mass and damper system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{k}{m}x_1 - \frac{c}{m}x_2 + \frac{1}{m}(u + w) \end{aligned} \tag{59}$$

where x_1 denotes the position and x_2 denotes the velocity, m denotes the mass, k denotes the stiffness coefficient and c denotes the damping. These parameters are always be selected as $m = 1kg$, $c = 0.5N \cdot s/m$ and $k = 5N/m$. Besides, the disturbance is defined as $w(t) = x_1 sin^2(x_2) cos(0.5x_1)$.

The desired trajectories are written as

$$\dot{x}^r = \begin{bmatrix} 0 & 1 \\ -5 & 0 \end{bmatrix} x^r \tag{60}$$

Consider the nominal system of the system (59), the corresponding augmented system is

$$\dot{Z} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -5 & -0.5 & 0 & -0.5 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -5 & 0 \end{bmatrix} Z + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} u \tag{61}$$

The tracking performance function is selected as

$$\begin{aligned} V(Z(t)) = \int_0^\infty e^{-0.01(\varsigma - t)} \Bigg(&3Z^2 + Z(\varsigma)^T QZ(\varsigma) \\ &+ 2\sigma \int_0^{\mathrm{u(t)}} \left(\tanh^{-1}(\upsilon/\sigma)\right)^T R d\upsilon\,(u(\varsigma))\Bigg) d\varsigma \end{aligned} \tag{62}$$

where Q and R are set as identity matrices. Besides, the system dynamics are assumed to be totally unknown. The two ESNs are selected as $\mathfrak{B}_1 = \mathfrak{B}_{2p} = 100$, $\kappa_1 = \kappa_{2p} = 1$, $p_1 = p_2 = 10$. The reservoir function $\varPhi_1$ and $\varPhi_2$ are selected as identity function. Then, the output weights of ESNs are vector with 14 dimensions and are initially set as zero.

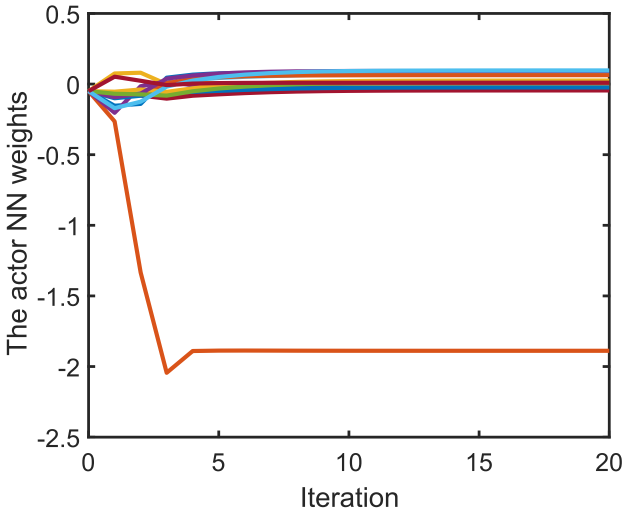

FIGURE 1. The actor NN weights of the linear tracking system.

The simulation results are shown in Figs. 1–3. Fig. 1 shows the iteration of the weights of the actor ESN during the training process of the linear system tracking problem. Fig. 2 shows the iteration of the weights of the actor ESN during the training process. Substitute the actor weights into the equation (49) to compute the intermediate variable. Then, substitute the intermediate variable into the equation (50) to compute the optimal control policy. To verify the effectiveness, the obtained controller is used to control this linear system (59) to track the predefined desired (60), the tracking trajectory is shown in Fig. 3. We can see that this linear system tracks the desired system in 6s and maintains the tracking state well under the disturbance. Fig. 4 shows the tracking trajectory under the polynomial network for the linear system, which is similar to our simulation result.

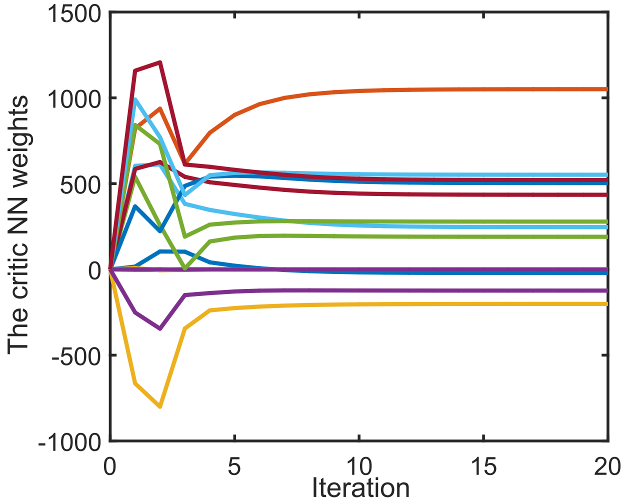

FIGURE 2. **The critic NN weights of the linear tracking system.**

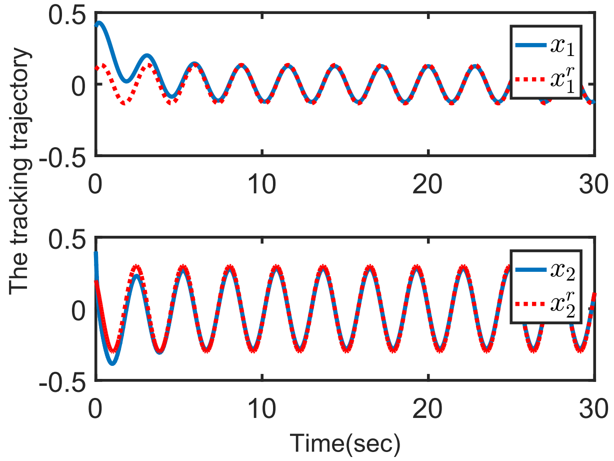

FIGURE 3. **The tracking trajectory of the linear system.**

B. NONLINEAR SYSTEM SIMULATION

The nonlinear system is

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ -0.5(x_1 + x_2) \end{bmatrix} + \begin{bmatrix} 0 \\ 0.5x_2(2 + \cos^2(2x_1))^2 + (2 + \cos^2(2x_1)) \end{bmatrix} \times (u + w) \tag{63}$$

where x denotes the system states. The disturbance is defined as $w(t) = x_1 sin^2(x_1)cos(0.5x_1x_2)$.

The desired trajectories are written as

$$\dot{x}^r = \begin{bmatrix} -1 & 1 \\ -2 & 1 \end{bmatrix} x^r \tag{64}$$

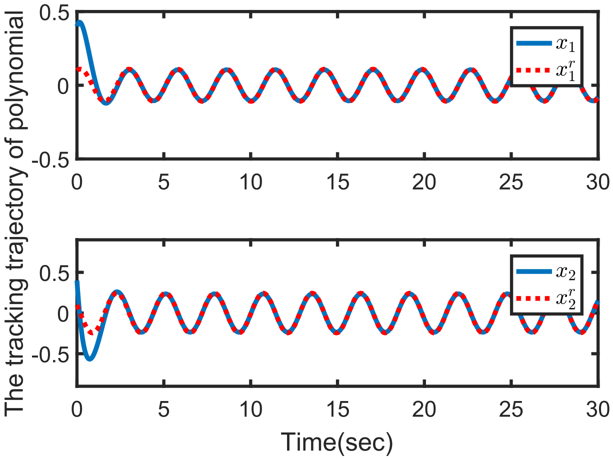

FIGURE 4. **The tracking trajectory of polynomial network for linear systems.**

By using the nominal system (63), the corresponding augmented system is

$$\dot{Z} = \begin{bmatrix} -Z_1 + Z_2 \\ -0.5(Z_1 + Z_2 + Z_3 + Z_4) + 2Z_3 - Z_4 \\ -Z_3 + Z_4 \\ -2Z_3 + Z_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0.5(Z_2 + Z_4)(2 + \cos(2(Z_1 + Z_3))^2)^2 \\ +2 + \cos^2(2(Z_1 + Z_3)) \\ 0 \\ 0 \end{bmatrix} u \tag{65}$$

The tracking performance function is selected as

$$V(Z(t)) = \int_0^{\infty} e^{-0.01(\varsigma - t)} \Big(2Z^2 + Z(\varsigma)^T QZ(\varsigma) + 2\sigma \int_0^{\mathrm{u(t)}} \left(\tanh^{-1}(\upsilon/\sigma)\right)^T Rd\upsilon\,(u(\varsigma)) \Big) d\varsigma \tag{66}$$

where Q and R are selected as identity matrices. Besides, the system dynamics are assumed to be totally unknown. The two ESNs are set as $\mathfrak{B}_1 = \mathfrak{B}_{2p} = 100$, $\kappa_1 = \kappa_{2p} = 1$, $p_1 = p_2 = 10$. The reservoir function Φ_1 and Φ_2 are selected as $tanh(\cdot)$. Then, the output weights of ESNs are also vector with 14 dimensions and are initially set as zero.

The simulation results are shown in Figs. 5–7. Fig. 5 shows the iteration steps of the actor ESN weights the during the training process of the nonlinear system tracking problem. Fig. 6 shows the iteration steps of the actor ESN weights during the training process. Substitute the actor weights into the equation (49) to compute the intermediate variable. Then, also substitute the intermediate variable into the equation (50) to compute the optimal control policy. To verify the effectiveness, the obtained controller is used to control this nonlinear system (63) to track the predefined desired (64), the tracking trajectory is shown in Fig. 7. We can see that this nonlinear system tracks the desired system in 4s

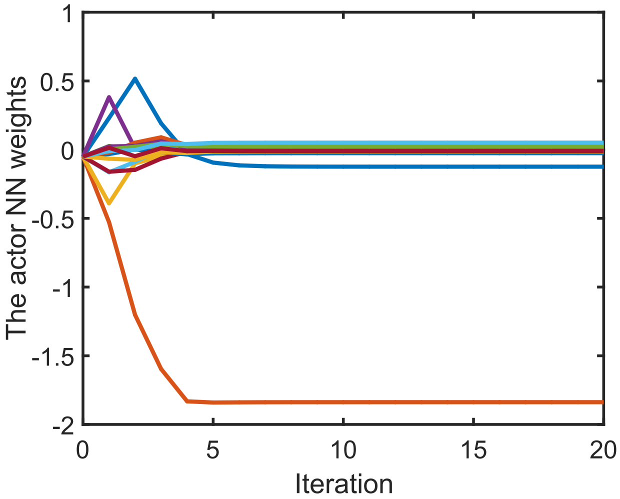

FIGURE 5. The actor NN weights of the nonlinear tracking system.

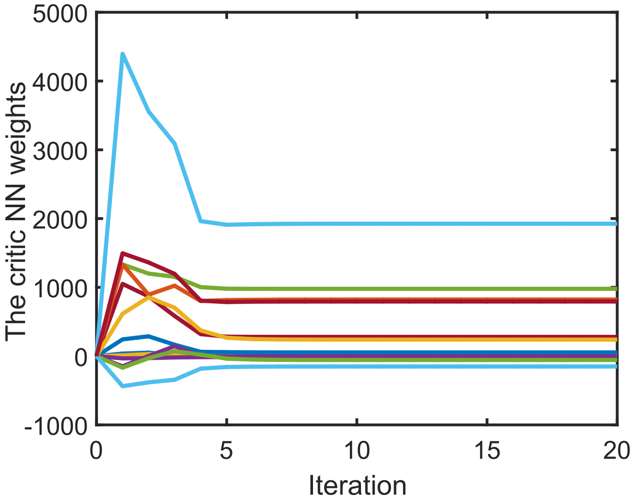

FIGURE 6. The critic NN weights of the nonlinear tracking system.

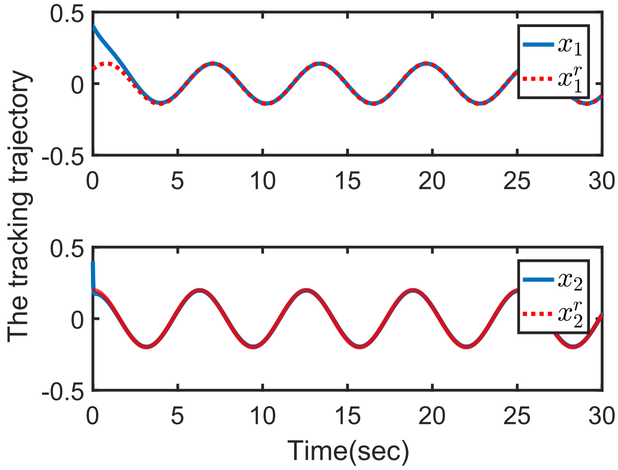

FIGURE 7. The tracking trajectory of the nonlinear system.

and maintains the tracking state well under the disturbance. Fig. 8 shows the tracking trajectory of polynomial network for nonlinear system, which is similar to our simulation result.

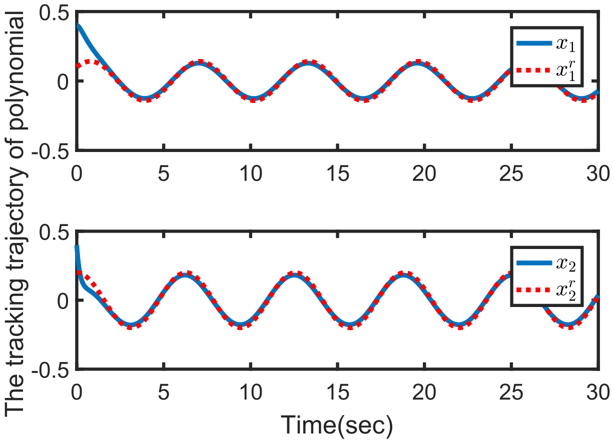

FIGURE 8. The tracking trajectory of polynomial network for nonlinear systems.

According to these two comparisons, we can see that the algorithm used in this paper not only guarantees tracking effectiveness, but also does not need to consider the choice of activation function.

VI. CONCLUSION

This paper utilizes a data-based IRL algorithm to solve the RTCP for a class of constrained CT nonlinear systems. It extends the IRL technique in robust tracking problem. The tracking problem is solved by using an optimal control method by defining an augmented system and a corresponding discounted performance function. Different from the existing methods, ESN is utilized as the approximate structure in the IRL algorithm, thus reducing the difficulty of neural network design and calculation burden of the weights training. The optimal performance function and control policy can be solved by using only the system data generated off-line. Two simulations present good results that the controlled system can effectively suppress the external disturbance and ensure the system to track the given target. Because of the random characteristics of ESN, the stability of the proposed algorithm needs to be further improved. In the future, we will continue to study how to design the parameters of ESN to improve the IRL algorithm, and concern with the problem of asymmetric input constraints.

REFERENCES

[1] X. Wang, Y. Zhou, Z. Zhao, W. Wei, and W. Li, "Time-delay system control based on an integration of active disturbance rejection and modified twice optimal control," *IEEE Access*, vol. 7, pp. 130734–130744, 2019.

[2] Z. Wang, C. Hu, Y. Zhu, S. He, K. Yang, and M. Zhang, "Neural network learning adaptive robust control of an industrial linear motor-driven stage with disturbance rejection ability," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2172–2183, Oct. 2017.

[3] X. Yang, D. Liu, H. Ma, and Y. Xu, "Online approximate solution of HJI equation for unknown constrained-input nonlinear continuous-time systems," *Inf. Sci.*, vol. 328, pp. 435–454, Jan. 2016.

[4] Y. Luo, S. Zhao, D. Yang, and H. Zhang, "A new robust adaptive neural network backstepping control for single machine infinite power system with TCSC," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 1, pp. 48–56, Jan. 2020.

[5] C. Qin, X. Qiao, J. Wang, D. Zhang, Y. Hou, and S. Hu, ''Barrier-critic adaptive robust control of nonzero-sum differential games for uncertain nonlinear systems with state constraints,'' *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 54, no. 1, pp. 50–63, Jan. 2024, doi: 10.1109/TSMC.2023.3302656.

[6] X. Zhang, W. Huang, and Q.-G. Wang, ''Robust h8 adaptive sliding mode fault tolerant control for T-S fuzzy fractional order systems with mismatched disturbances,'' *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 68, no. 3, pp. 1297–1307, Mar. 2021.

[7] J.-P. Humaloja, M. Kurula, and L. Paunonen, ''Approximate robust output regulation of boundary control systems,'' *IEEE Trans. Autom. Control*, vol. 64, no. 6, pp. 2210–2223, Jun. 2019.

[8] Y. Lv, J. Na, X. Zhao, Y. Huang, and X. Ren, ''Multi-H_∞ controls for unknown input-interference nonlinear system with reinforcement learning,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 5601–5613, Sep. 2023.

[9] C. Qin, X. Qiao, J. Wang, and D. Zhang, ''Robust trajectory tracking control for continuous-time nonlinear systems with state constraints and uncertain disturbances,'' *Entropy*, vol. 24, no. 6, p. 816, Jun. 2022.

[10] C. Qin, J. Wang, X. Qiao, H. Zhu, D. Zhang, and Y. Yan, ''Integral reinforcement learning for tracking in a class of partially unknown linear systems with output constraints and external disturbances,'' *IEEE Access*, vol. 10, pp. 55270–55278, 2022.

[11] H. Zhang, L. Cui, X. Zhang, and Y. Luo, ''Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method,'' *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 2226–2236, Dec. 2011.

[12] H. Modares, F. L. Lewis, and Z. P. Jiang, ''H∞ tracking control of completely unknown continuous-time systems via off-policy reinforcement learning,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2550–2562, Oct. 2015.

[13] Y. Tang and X. Yang, ''Robust tracking control with reinforcement learning for nonlinear-constrained systems,'' *Int. J. Robust Nonlinear Control*, vol. 32, no. 18, pp. 9902–9919, Dec. 2022.

[14] D. Wang, L. Cheng, and J. Yan, ''Self-learning robust control synthesis and trajectory tracking of uncertain dynamics,'' *IEEE Trans. Cybern.*, vol. 52, no. 1, pp. 278–286, Jan. 2022.

[15] B. Niu, J. Liu, D. Wang, X. Zhao, and H. Wang, ''Adaptive decentralized asymptotic tracking control for large-scale nonlinear systems with unknown strong interconnections,'' *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 1, pp. 173–186, Jan. 2022.

[16] K. Zhang, H. Zhang, W. Xue, and R. Zhang, ''A robust control scheme for autonomous vehicles path tracking under unreliable communication,'' in *Proc. IEEE 11th Data Driven Control Learn. Syst. Conf. (DDCLS)*, Aug. 2022, pp. 1413–1418.

[17] H. Rios, R. Falcon, O. A. Gonzalez, and A. Dzul, ''Continuous sliding-mode control strategies for quadrotor robust tracking: Real-time application,'' *IEEE Trans. Ind. Electron.*, vol. 66, no. 2, pp. 1264–1272, Feb. 2019.

[18] W. Lei, C. Li, and M. Z. Q. Chen, ''Robust adaptive tracking control for quadrotors by combining PI and self-tuning regulator,'' *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 6, pp. 2663–2671, Nov. 2019.

[19] H. Zhang, K. Zhang, G. Xiao, and H. Jiang, ''Robust optimal control scheme for unknown constrained-input nonlinear systems via a plug-n-play event-sampled critic-only algorithm,'' *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 9, pp. 3169–3180, Sep. 2020.

[20] A. Sabanovic, M. Elitas, and K. Ohnishi, ''Sliding modes in constrained systems control,'' *IEEE Trans. Ind. Electron.*, vol. 55, no. 9, pp. 3332–3339, Sep. 2008.

[21] C. Qin, Z. Zhang, Z. Shang, J. Zhang, and D. Zhang, ''Adaptive optimal safety tracking control for multiplayer mixed zero-sum games of continuous-time systems,'' *Int. J. Speech Technol.*, vol. 53, no. 14, pp. 17460–17475, Jan. 2023.

[22] C. Liu, H. Zhang, G. Xiao, and S. Sun, ''Integral reinforcement learning based decentralized optimal tracking control of unknown nonlinear large-scale interconnected systems with constrained-input,'' *Neurocomputing*, vol. 323, pp. 1–11, Jan. 2019.

[23] D. Vrabie and F. L. Lewis, ''Adaptive optimal control algorithm for continuous-time nonlinear systems based on policy iteration,'' in *Proc. 47th IEEE Conf. Decis. Control*, Cancún, Mexico, 2008, pp. 73–79.

[24] H. Zhang, Y. Luo, and D. Liu, ''Neural-Network-Based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints,'' *IEEE Trans. Neural Netw.*, vol. 20, no. 9, pp. 1490–1503, Sep. 2009.

[25] M. Abu-Khalaf and F. L. Lewis, ''Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach,'' *Automatica*, vol. 41, no. 5, pp. 779–791, May 2005.

[26] K. G. Vamvoudakis and F. L. Lewis, ''Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem,'' *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[27] P. J. Werbos, ''Beyond regression: New tools for prediction and analysis in the behavioral sciences,'' Ph.D. dissertation, Dept. Appl. Math., Harvard Univ., Cambridge, MA, USA, 1974.

[28] R. E. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

[29] B. Luo, H.-N. Wu, T. Huang, and D. Liu, ''Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design,'' *Automatica*, vol. 50, no. 12, pp. 3281–3290, Dec. 2014.

[30] Y. Liu and Z. Wang, ''Reinforcement learning-based tracking control for a class of discrete-time systems with actuator fault,'' *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 6, pp. 2827–2831, Jun. 2022.

[31] Z. Zhang, D. Wang, and J. Gao, ''Learning automata-based multiagent reinforcement learning for optimization of cooperative tasks,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4639–4652, Oct. 2021.

[32] Y. Jiang and Z.-P. Jiang, ''Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics,'' *Automatica*, vol. 48, no. 10, pp. 2699–2704, Oct. 2012.

[33] X. Yang, D. Liu, B. Luo, and C. Li, ''Data-based robust adaptive control for a class of unknown nonlinear constrained-input systems via integral reinforcement learning,'' *Inf. Sci.*, vol. 369, pp. 731–747, Nov. 2016.

[34] Z. Yao, X. Liang, G.-P. Jiang, and J. Yao, ''Model-based reinforcement learning control of electrohydraulic position servo systems,'' *IEEE/ASME Trans. Mechatronics*, vol. 28, no. 3, pp. 1446–1455, Jun. 2023.

[35] G. Xiao, H. Zhang, Y. Luo, and H. Jiang, ''Data-driven optimal tracking control for a class of affine non-linear continuous-time systems with completely unknown dynamics,'' *IET Control Theory Appl.*, vol. 10, no. 6, pp. 700–710, Apr. 2016.

[36] H. Jaeger, ''The 'echo state' approach to analysing and training recurrent neural networks,'' German Nat. Res. Center Inform. Technol., St. Augustin, Germany, Tech. Rep., GMD 148, 2001.

[37] C. Liu, H. Zhang, Y. Luo, and H. Su, ''Dual heuristic programming for optimal control of continuous-time nonlinear systems using single echo state network,'' *IEEE Trans. Cybern.*, vol. 52, no. 3, pp. 1701–1712, Mar. 2022.

[38] H. Zhang, C. Liu, H. Su, and K. Zhang, ''Echo state network-based decentralized control of continuous-time nonlinear large-scale interconnected systems,'' *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 10, pp. 6293–6303, Oct. 2021.

[39] R. Song, F. L. Lewis, Q. Wei, and H. Zhang, ''Off-policy actor-critic structure for optimal control of unknown systems with disturbances,'' *IEEE Trans. Cybern.*, vol. 46, no. 5, pp. 1041–1050, May 2016.

[40] R. Song, F. L. Lewis, and Q. Wei, ''Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 704–713, Mar. 2017.

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

[42] M. Lukoševičius, ''A practical guide to applying echo state networks,'' in *Neural Networks: Tricks of the Trade*. Heidelberg, Germany: Springer, 2012, pp. 659–686.

CHONG LIU received the B.S. degree in electronic and information engineering from Inner Mongolia Normal University, Inner Mongolia, China, in 2011, the M.S. degree in electronic science and technology from the Changchun University of Science and Technology, Changchun, China, in 2015, and the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2020. He is currently a Lecturer with the Xi'an University of Architecture and Technology. His research interests include adaptive dynamic programming, neural networks, optimal control, reinforcement learning, and power system control.

YALUN LI received the B.S. degree in mechanical and electrical engineering from Zhoukou Normal University, Zhoukou, China, in 2022. He is currently pursuing the M.S. degree in control science and engineering with the College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, China. His current research interests include adaptive dynamic programming, robust control, and reinforcement learning.

ZHONGXING DUAN received the B.S. degree in industrial electric automation and the M.S. degree in computer application from the Xi'an University of Architecture and Technology, in 1992 and 1999, respectively, and the Ph.D. degree in computer architecture from Xi'an Jiaotong University, in 2006. He is currently a Professor with the College of Information and Control Engineering, Xi'an University of Architecture and Technology. His current research interests include intelligent detection and intelligent control and building environment optimization control.

ZHOUSHENG CHU received the B.S. degree in engineering from the Liren College, Yanshan University, Qinhuangdao, China, in 2022. He is currently pursuing the M.S. degree in control science and engineering with the College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, China. His current research interests include adaptive dynamic programming, neural networks, and optimal control.

ZONGFANG MA was born in Anhui, China, in 1980. He received the bachelor's and master's degrees from the Xi'an University of Architecture and Technology, Xi'an, China, in 2002 and 2006, respectively, and the Ph.D. degree from Northwestern Polytechnical University (NPU), Xi'an, in 2011. He is currently a Professor with the College of Information and Control Engineering, Xi'an University of Architecture and Technology. His current research interests include machine vision and pattern recognition.