Neurocomputing 402 (2020) 50–65

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Reinforcement Learning-Based Control for Nonlinear Discrete-Time Systems with Unknown Control Directions and Control Constraints

Miao Huang[a,*], Cong Liu[b], Xiaoqi He[c], Longhua Ma[d], Zheming Lu[e], Hongye Su[e]

[a] College of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai 201209, PR China
[b] Department of Internet of Things Engineering, Shanghai Business School, Shanghai 200235, PR China
[c] Ningbo Industrial Internet Institute, Ningbo 315000, PR China
[d] Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, PR China
[e] Zhejiang University, Hangzhou 310058, PR China

ARTICLE INFO

Article history:
Received 13 December 2018
Revised 3 February 2020
Accepted 21 March 2020
Available online 8 April 2020

Communicated by Prof. Huaguang Zhang

Keywords:
Neural networks
Reinforcement learning
Non-affine nonlinear systems
Output feedback
Unknown control directions

ABSTRACT

In this work, output-feedback control problems for a class of discrete-time non-affine nonlinear systems with unknown control directions and input constraints are considered by using reinforcement learning (RL) method. Two neural networks (NNs) implement the control: 1) a critic NN that estimates a non-quadratic strategic utility function (SUF) and 2) an action NN that generates optimized control input and minimizes the SUF. The implicit function theorem is applied to obtain the optimal control law since the control is appeared in a non-affine form. For the first time, the discrete Nussbaum gain is introduced to overcome the difficulty that the control directions are unknown and a non-quadratic SUF is used to deal with the control constraints in the RL-based control. The theoretical derivation of the uniformly ultimately boundedness of the NN weights and the closed-loop output tracking error is given. And two numerical examples have been supplied to valid the proposed method.

1. Introduction

Adaptive dynamic programming (ADP) is an optimal control method which is implemented to ensure the closed-loop system stable and minimize the predefined cost functions ([1,2]). However, most of ADP algorithms are designed by an off-line iterative process which requires the dynamics of nonlinear systems are known a prior. Since the dynamics of nonlinear systems are rarely obtained, it makes the implementation of these algorithms hard in the practical control processes. To overcome this drawback, reinforcement learning (RL) is adopted to solve optimal control problems.

Reinforcement learning is an interaction between an actor (or agent) with its environments. The actor (or agent) improves its actions or control strategies according to the results in answer to its actions. In discrete time, a direct RL adaptive controller was provided to guarantee a requested tracking performance for nonlinear systems in the present of unknown bounded disturbances ([3]). A minimal-learning-parameter mechanism was introduced in the RL controller to cope with the pure feedback discrete-time systems in [4]. Radac et al. provided a nonlinear state-feedback Q-learning controller with a batch fitted Q iteration algorithm and a linear virtual reference feedback tuning technique to solve the model reference control problem of discrete-time nonlinear systems ([5]). An extra NN was added as an estimator of the unavailable system states in RL based output feedback adaptive controllers for affine nonlinear systems and strict feedback nonlinear systems in [6] and [7], respectively. Novel critic-based adaptive NN controllers have been implemented for a class of nonlinear pure-feedback systems, in which a deterministic learning technique was employed to make internal states satisfy the partial persistent excitation condition in a periodic reference orbit tracking problem ([8]). Shih et al. introduced one action NN to obtain the virtual control input and another action NN to provide the actual input into the RL based output feedback controller for the non-strict feedback nonlinear systems in [9].

In continuous time, Zhu et al. used an integral online RL algorithm to find a suboptimal controller in an output-feedback form for linear time invariant systems in [10]. A RL algorithm based on a model-free off-policy learning technique was utilized to approximate the optimal solution of the output-feedback control problem of linear systems in [11]. For the nonlinear systems with constrained-input, the infinite-horizon optimal control problem was solved by the RL-based adaptive control method proposed by [12]. In this method, an action-critic structure was de-

* Corresponding author.
E-mail address: huangmiao@sspu.edu.cn (M. Huang).

https://doi.org/10.1016/j.neucom.2020.03.061
0925-2312/

veloped, in which the optimal control input and the minimized cost were approximated by two feed-forward NNs called actor and critic, respectively. Liu et al. implemented an integral RL based decentralized state-feedback controller to a constrained-input system which has large-scale interconnection subsystems so that the optimal state tracking performance can be accomplished in [13]. For marine surface vessel systems, by using a policy iteration technique, RL algorithm was applied in [14] to find the solution of the optimal control problem when the system dynamic was known a prior. Most of the works mentioned above were assumed that the system control direction, i.e. the sign of the control gain, was known. However, in practical applications, the control gains of systems may be unknown. This case bought a difficult control problem that the control operate direction was hardly decided.

In last decades, an increasing number of attention was received by the system control problem with unknown control directions. Nussbaum gains were introduced to output feedback adaptive controllers to overcome the drawback that the control direction was unknown in general nonlinear systems in [15,16] and [17] provided the adaptive control methods by using back-stepping design idea to solve control problem of the non-strict-feedback nonlinear systems in the present of unknown backlash-like hysteresis nonlinearity and full state constraints, respectively. For strict-feedback nonlinear systems, Liu et al. investigated a generalized fuzzy hyperbolic model-based adaptive predefined performance control method with Nussbaum gains ([18]). Nussbaum gains were also utilized in state feedback control design in [19] and the effects of un-modeled dynamics and unknown dead zones were eliminated. However, most of the past literature dealt with the affine strict or non-strict-feedback systems. The optimal control problems of non-affine nonlinear systems with unknown control directions were not considered.

In this paper, an optimal output-feedback control problem of the discrete-time non-affine nonlinear systems with unknown control directions and input constraints was considered. An on-line real-time RL method with Actor-Critic structure and Nussbaum gains is chosen to implement the control such that the optimal control problem can be solved using only fewer data measured along system trajectories. Temporal difference (TD) and value function approximation (VFA) are the two key machine learning technologies applied to this structure. VFA implies a strategic utility function (SUF), can be approximated by a estimator with fewer parameters, i.e. critic NN. TD error means a prediction error between the values of predicted SUF and observed SUF in response to an action, i.e. the output of the actor NN, applied to the systems. Then, the optimal control problem can be converted to find the control law to make the TD error zero.

The main contributions of this paper are as follows.

(1) An optimal control law is developed by using Implicit function theorem and Bellman's principle of optimality since the non-affine appearance of the control input.
(2) Discrete-time Nussbaum gains are firstly introduced in the optimal control problem of non-affine nonlinear systems to cope with the unknown control directions.
(3) A non-quadratic strategic utility function (SUF) is firstly used to overcome the control constraints in the RL-based output-feedback control design.

The rest of this paper is organized as follows. Section 2 presents the problem formulation and preliminaries. An actor-critic structure is provided in Section 3 to estimate the optimal control law and the minimized SUF. The uniformly ultimately bounded (UUB) performance of the tracking error and the NN weights is illustrated and two numerical examples are presented to show the effectiveness of the proposed method in Section 4 and 5, respectively. Finally, the conclusions are drawn in Section 6.

2. Problem Formulation and Preliminaries

Consider the SISO discrete-time system in a non-affine pure-feedback form as follows:

$$\begin{cases} \xi_i(k+1) = f_i\big(\bar{\xi}_i(k), \xi_{i+1}(k)\big), i = 1, 2, \ldots, n-1 \\ \xi_n(k+1) = f_n\big(\bar{\xi}_n(k), u(k), d(k)\big) \\ \quad y(k) = \xi_1(k) \end{cases} \tag{1}$$

where $\bar{\xi}_j(k) = \big[\xi_1(k), \xi_2(k), \ldots, \xi_j(k)\big]^T, j = 1, 2, \ldots, n$ are the state variables of the system, $n \geq 1$ is the system order, $f_i(\cdot\ ,\ \cdot)$ and $f_n(\cdot\ ,\ \cdot\ ,\ \cdot)$ are the unknown nonlinear functions, $u(k) \in R$ and $y(k) \in R$ are the input and output of the system, respectively, $|u(k)| \leq \upsilon$, where υ is the saturating bound, and $d(k)$ denotes the external disturbance, which has the unknown constant bound $\bar{d}_s$, i.e. $|d(k)| \leq \bar{d}_s$.

Assumption 1. The system functions $f_i(\cdot\ ,\ \cdot)$ and $f_n(\cdot, \cdot, 0), i = 1, \ldots, n-1$ in (1) are continuous with respect to all the arguments and continuously differentiable with respect to the second argument.

Assumption 2. There exist constants $\bar{g}_i > \underline{g}_i > 0$ so that $0 \leq \underline{g}_i \leq |g_i(\cdot)| \leq \bar{g}_i, i = 1, 2, \ldots, n$, where $g_j(\cdot) = \big(\partial f_j\big(\bar{\xi}_j(k), \xi_{j+1}(k)\big)/\partial \xi_{j+1}(k)\big), j = 1, 2, \ldots, n-1$ and $g_n(\cdot) = \big(\partial f_n\big(\bar{\xi}_n(k), u(k), d(k)/\partial u(k)\big)\big)$ are the system control gains.

Assumption 3. The system functions $f_i(\cdot\ ,\ 0)$ and $f_n(\cdot\ ,\ 0,\ \cdot)$ are Lipschitz continuous on Ω_i and $\Omega_n \times \Omega_d$, where $\Omega_i \in R^i, i = 1, 2, \ldots, n-1$, $\Omega_n \in R^n$ and $\Omega_d \in R$ are some known compact sets.

Remark 1. Assumption 1 is a standard assumption in nonlinear control systems with the non-affine form, which can be found in lots of existing related works such as [20] and [21]. The system functions $f_i(\cdot\ ,\ \cdot)$ and $f_n(\cdot\ ,\ \cdot\ ,\ \cdot)$ in (1) are continuously differentiable with respect to the second argument, which ensures the existent of $g_j(\cdot)$ and $g_n(\cdot)$ of Assumption 2.

Remark 2. Assumption 2 implies that System (1) has either strictly positive or negative control gains $g_j(\cdot)$ and $g_n(\cdot)$. However, the signs of $g_j(\cdot)$ and $g_n(\cdot)$, i.e. the control directions, are unknown. Noting that the constraints $\bar{g}_i$ and $\underline{g}_i$ are unknown, since they are used for analysis instead of control design.

In order to simplify the controller design, system (1) is transformed into an input-output form without the future states according to the derivative process given in [20].

$$y(k+n) = \phi(\underline{z}(k), u(k)) + d_t(k) \tag{2}$$

where $\underline{z}(k) = [y(k), \ldots, y(k-n+1), u(k-1), \ldots, u(k-n+1)]$, $\phi(\cdot\ ,\ \cdot)$: $R^{2n} \to R$ is an unknown nonlinear function. There exists a finite constant $\bar{d}$ such that $|d_t(k)| \leq \bar{d}$.

Remark 3. The transformation from System (1) to System (2) is guaranteed by Assumption 1 and 3. In addition, $\phi(\cdot\ ,\ \cdot)$ is also Lipschitz function, since it is obtained by iterative substitution of system functions $f_i(\cdot\ ,\ \cdot)$ and $f_n(\cdot\ ,\ \cdot\ ,\ \cdot)$.

The transformation procedure of a second-order system from Eq. (1) to Eq. (2) will be given as an example, which can be divided into two steps.

Step 1. a second-order system with Form (1) to the NARMAX form.

$$\begin{cases} \xi_1(k+1) = f_1\big(\bar{\xi}_1(k), \xi_2(k)\big) \\ \xi_2(k+1) = f_2\big(\bar{\xi}_2(k), u(k), d(k)\big) \\ \quad y(k) = \xi_1(k) \end{cases}$$

where $\bar{\xi}_1(k) = \xi_1(k), \bar{\xi}_2(k) = [\xi_1(k), \xi_2(k)]^T$. Let $\phi_{1,1}\big(\bar{\xi}_2(k)\big) = f_1\big(\bar{\xi}_1(k), \xi_2(k)\big)$ and $\phi_{1,2}(\cdot,\cdot,\cdot) = f_2(\cdot,\cdot,\cdot)$. Then, we have

$$\begin{cases} \xi_1(k+1) = \phi_{1,1}\big(\bar{\xi}_2(k)\big) \\ \xi_2(k+1) = \phi_{1,2}\big(\bar{\xi}_2(k), u(k), d(k)\big) \\ y(k+2) = \xi_1(k+2) \end{cases}$$

In $k+2$ step, $\xi_1(k+2) = \phi_{1,1}\big(\phi_{1,1}\big(\bar{\xi}_2(k)\big), \xi_2(k+1)\big)$. Then, we obtain that

$$y(k+2) = \phi_1\big(\bar{\xi}_2(k), u(k), d(k)\big),$$

where $\phi_1(\bar{\xi}_2(k), u(k), d(k)) = \phi_{1,1}(\phi_{1,1}(\bar{\xi}_2(k)), \phi_{1,2}(\bar{\xi}_2(k), u(k), d(k)))$. Rewrite the first equation of (1) as

$$\xi_1(k+1) - f_1(\xi_1(k), \xi_2(k)) = 0.$$

According to Assumption 2, the derivative of the left-hand side of the above equation with respect to $\xi_2(k)$ is not zero. From Implicit Function Theorem given by [22] there exists an implicit function $p_2'(\cdot)$ such that $\xi_2(k)$ can be seen as a function of $\xi_1(k+1)$ and $\xi_1(k)$ as follows:

$$\begin{aligned} \xi_2(k) &= p_2'(\xi_1(k+1), \xi_1(k)) \\ &:= p_2(y(k+1), y(k)) \end{aligned}$$

Substituting $\xi_2(k)$ in $y(k+2)$ with the above equation, the NARMAX form of System (1) with second-order can be obtained.

$$\begin{aligned} y(k+2) &= \phi_1(y(k), p_2(y(k+1), y(k)), u(k), d(k)) \\ &= \phi_s\big(\underline{y}(k+1), u(k), d(k)\big), \end{aligned}$$

where $\underline{y}(k+1) = [y(k+1), y(k)]^T$.

Step 2. The NARMAX system to System (2).

To overcome the difficulty in controlling NARMAX system lies in the existence of future outputs $y(k+1)$, which are not available at the current step, the output prediction approach is considered.

Moving back 1 step of $y(k+2)$, we have

$$y(k+1) = \phi_s\big(\underline{y}(k), u(k-1), d(k-1)\big).$$

Substituting $y(k+1)$ in $y(k+2)$, it follows that

$$\begin{aligned} y(k+2) &= \phi_s(\phi_s(\underline{y}(k), u(k-1), d(k-1)), u(k), d(k)) \\ &= \phi_p(\underline{y}(k), u(k), u(k-1), d(k), d(k-1)) \\ &= \phi(\underline{z}(k), u(k)) + d_t(k) \end{aligned}$$

where $\underline{z}(k) = [y(k), y(k-1), u(k-1)]^T$, $\phi(\underline{z}(k), u(k)) = \phi_p(\underline{y}(k), u(k), u(k-1), 0, 0))$ and $d_t(k) = \phi_p(\underline{y}(k), u(k), u(k-1), \bar{d}(k), d(k-1)) - \phi_p(\underline{y}(k), u(k), u(k-1), 0, 0))$. Since ϕ_p is obtained by iterative substitution of the system functions $f_i, i = 1, 2$, which satisfies Lipschitz condition in Assumption 3. Then, there exists a finite constant L_d such that $|d_t(k)| \le L_d|d(k)| + L_d|d(k-1)| \le \bar{d}$.

The general optimal control objective is to obtain the admissible control $u(k)$ which can guarantee the system stability and minimize the non-quadratic SUF defined in [23].By introducing an infinite vector $\bar{z}(k) = [y(k), u(k), y(k+1), u(k+1), \ldots]^T$, the SUF is defined as

$$J(\bar{z}(k)) = \sum_{i=0}^{\infty} \{W(u(k+i)) + q(y(k+i))\} \tag{3}$$

where $W(u(k)) = 2\int_0^{u(k)} \varphi^{-1}\big(\upsilon^{-1}s\big)\upsilon r ds$, r is a positive constant, $\varphi(\cdot)$ is a bounded one-to-one function satisfying $|\varphi(\cdot)| \le 1$ and belonging to $L_2(\Omega_n)$. Moreover, $\varphi(\cdot)$ is an odd function and increases monotonically. The gradient of $\varphi(\cdot)$ is bounded by a constant M. Such a function is easy to find, and one example is the hyperbolic tangent function, i.e. $\varphi(\cdot) = \tanh(\cdot)$. $q(y(k)) = (y(k) - y_r(k))^2$, $y_r(k)$ is referred as the desired trajectory which is a known smooth bounded function over the compact subset of R. It should be noticed that by the definition above, $W(u(\cdot))$ is assured to be positive because $\varphi^{-1}(\cdot)$ is a monotonic odd function and r is positive. In addition, the symbol $\bar{z}(k)$ is only used to analysis instead of control design.

By rewriting (3) as

$$J(\bar{z}(k)) = W(u(k)) + q(y(k)) + \sum_{i=1}^{\infty} \{W(u(k+i)) + q(y(k+i))\},$$

it can be equivalent to a difference equation given by

$$J(\bar{z}(k)) = W(u(k)) + q(y(k)) + J(\bar{z}(k+1)), J(\mathbf{0}) = 0.$$

It means that the value of a current policy $u(k)$ can be obtained by solve the above difference equation. This equation refers as the *Bellman equation*. The Bellman equation is applied to evaluate the value of a current policy $u(k)$ and is solved online in real-time using observed data from the system trajectories.

Define a discrete-time Hamiltonian function as

$$H(\bar{z}(k)) = W(u(k)) + q(y(k)) + J(\bar{z}(k+1)) - J(\bar{z}(k)),$$

The Bellman equation requires the Hamiltonian associated with the specified strategy to be zero.

The optimal value can be written using the Bellman equation as

$$J^*(\bar{z}(k)) = \min_{u(\cdot)} \{W(u(k)) + q(y(k)) + J(\bar{z}(k+1))\}.$$

Bellman's principle states that "An optimal policy has the property that no matter what the previous decisions (i.e. controls) have been, the remaining decisions must constitute an optimal policy with regard to the state resulting from those previous decisions". For this $J^*(z(k))$, it means that

$$J^*(\bar{z}(k)) = \min_{u(\cdot)} \{q(y(k)) + W(u(k)) + J^*(\bar{z}(k+1))\}. \tag{4}$$

This is known as the Bellman optimality equation, or the discrete-time Hamilton-Jacobi-Bellman (HJB) equation.

As shown by Implicit Function Theorem, there exists an optimal policy $u^*(k)$ defined as

$$u^*(k) = \arg\min_{u(k)} \{q(y(k)) + W(u(k)) + J^*(\bar{z}(k+1))\}. \tag{5}$$

Assuming that the value function $J^*(\cdot)$ is smooth, the minimum of the right-hand side of (4) can exactly be solved by letting the gradient of $q(y(k)) + W(u(k)) + J^*(\bar{z}(k+1))$ with respect to $u(k)$ equal to zero. From System (2), it follows that $y(k+n), \ldots, y(k+2n-1)$ depend on $u(k)$ such that

$$\begin{aligned} \frac{\partial J^*(\bar{z}(k))}{\partial u(k)} &= \frac{\partial (q(y(k)) + W(u(k)))}{\partial u(k)} \\ &+ \sum_{i=0}^{n-1} \frac{\partial J^*(\bar{z}(k+1))}{\partial y(k+n+i)} \cdot \frac{\partial y(k+n+i)}{\partial u(k)} = 0. \end{aligned} \tag{6}$$

Therefore, the corresponding optimal control law $u^*(k)$ can be obtained by solving the above equation, i.e.

$$u^*(k) = \upsilon\varphi\left(-\frac{1}{2}(\upsilon r)^{-1} \sum_{i=0}^{n-1} h_i(k) \frac{\partial J^*(\bar{z}(k+1))}{\partial y(k+n+i)}\right) \tag{7}$$

where $h_i(k) = \frac{\partial y(k+n+i)}{\partial u(k)}, i = 0, \ldots, n-1$.

To facilitate the control design, the definition of Nussbaum gain is first reviewed.

Definition 1 ([20]). Consider a discrete nonlinear function $N(x(k))$ defined on a sequence $x(k)$ with $x_s(k) = \sup_{i \le k} \{x(i)\}$. $N(x(k))$ is a discrete Nussbaum gain if and only if it satisfies the following two properties:

(i) If $x_s(k)$ increases without bound, then for any given constant δ_0

$$\sup_{x_s(k)\geq\delta_0}\frac{S_N(x(k))}{x_s(k)}=+\infty,\ \inf_{x_s(k)\geq\delta_0}\frac{S_N(x(k))}{x_s(k)}=-\infty.$$

(ii) If $x_s\leq\delta_1$, then $|S_N(x(k))|\leq\delta_2$ with some positive constants δ_1 and δ_2, where $S_N(x(k))$ is defined with $\triangle x(k)=x(k+1)-x(k)$ as follows:

$$S_N(x(k))=\sum_{i=0}^{k}N(x(i))\triangle x(i).$$

Let $\{x(k)\}$ be a discrete sequence with

$$x(0)=0, x(k)\geq 0, \forall k$$

and

$$|\triangle x(k)|=|x(k+1)-x(k)|\leq c_1$$

where c_1 is a constant.

In this paper, the discrete Nussbaum gain $N(x(k))$ proposed in [24] will be used, which is defined as:

$$N(x(k))=x_s(k)s(x(k))$$

where $x_s(k)=\sup_{\sigma\leq k}\{x(\sigma)\}$ and $s(x(k))$ is defined in the following manner:

$$s(x(0))=+1.$$

At $k=k_1$, if $s(x(k_1))=+1$, then if

$$\sum_{\sigma=0}^{k_1}N(x(\sigma))\triangle x(\sigma)>x_s^{3/2}(k_1)$$

set $s(x(k_1+1))=-1$ otherwise set $s(x(k_1+1))=+1$.

But if $s(x(k_1))=-1$ then if

$$\sum_{\sigma=0}^{k_1}N(x(\sigma))\triangle x(\sigma)<-x_s^{3/2}(k_1)$$

set $s(x(k_1+1))=+1$, otherwise set $s(x(k_1+1))=-1$.

The next lemma gives a fact of the discrete Nussbaum gain $N(x(k))$.

Lemma 1 ([25]). *Let $V(k)$ be a positive definite function defined for $\forall k$, $N(x(k))$ be a discrete Nussbaum gain, and ϑ be a nonzero constant. If the following inequality holds, for $\forall k$*

$$V(k)\leq\left[\sum_{i=k_1}^{k}(\rho_1+\vartheta N(x(i)))\triangle x(i)\right]+\rho_2x(k)+\rho_3$$

where ρ_1, ρ_2 and ρ_3 are some constants, k_1 is a positive integer, then $V(k)$, $x(k)$ and $\sum_{i=k_1}^{k}(\rho_1+\vartheta N(x(i)))\triangle x(i)+\rho_2x(i)+\rho_3$ must be bounded for $\forall k$.

3. Output Feedback Controller Design

In this section, an output feedback controller which has an actor-critic architecture is developed by using reinforcement learning methods. The design processes of the critic NN and the action NN are introduced, respectively.

3.1. Critic NN and weight update law

In this section, a critic NN is used to approximate the SUF $J(k)$. Since $J(k)$ is unavailable at the kth time instant, the critic NN is tuned online to ensure its output converges close to $J(k)$.

Define a prediction error of the critic NN, i.e. the TD error, as:

$$e_c(k)=r_c\hat{J}(k)-\hat{J}(k-1)+D^{-2}(k)(q(k)+W(k))a(k) \tag{8}$$

where $\hat{J}(k)=\hat{W}_c^T(k)\phi_c\left(V_c^Tz(k)\right)$ represents the output of the critic NN, $z(k)=[\underline{z}(k),u(k)]^T$, and $0<r_c<1$ is the temporal difference coefficient. $a(k)$ and $D^{-2}(k)$ are defined later in Section 3.2. The critic NN has a two-layer structure, while $\hat{W}_c(k)\in R^{n_c\times 1}$ and $V_c\in R^{n_z\times n_c}$ indicate its actual weight vector of the output and the weight matrix of hidden layers, respectively. The term n_c denotes the number of the neurons in the hidden layer and $n_z=2n+1$. The regression $z(k)\in R^{n_z}$, which is composed of the past value of the input and output measurements, are chosen as the input of the critic NN. The activation function vector of the hidden layer $\phi_c\left(V_c^Tz(k)\right)\in R^{n_z}$ can be written as $\phi_c(z(k))$ or short. If there are enough number of the neurons in the hidden layer, the critic network can approximate the optimal SUF J^* with arbitrarily small estimation error $\varepsilon_c(k)$ as

$$\begin{aligned}J^*(\bar{z}(k))&=W_c^T\phi_c\left(V_c^Tz(k)\right)+\varepsilon_c(z(k))\\&=W_c^T\phi_c(z(k))+\varepsilon_c(z(k))\end{aligned} \tag{9}$$

where W_c denotes the desired weight matrix of the optimal SUF, $\varepsilon_c(z(k))$ denotes the bounded error.

The weight estimation error of the critic network NN is defined as

$$\tilde{W}_c(k)=\hat{W}_c(k)-W_c \tag{10}$$

and the approximation error is described as

$$\zeta_c(k)=\tilde{W}_c^T(k)\phi_c(k). \tag{11}$$

Thus, the prediction error can be

$$\begin{aligned}e_c(k)&=r_c\hat{J}(k)-\hat{J}(k-1)+a(k)(q(k)+W(k))D^{-2}(k)\\&=r_c\zeta_c(k)+r_cJ^*(\bar{z}(k))+\zeta_c(k-1)-J^*(\bar{z}(k-1))\\&\quad+\varepsilon_c(k)-\varepsilon_c(k-1)+a(k)(q(k)+W(k))D^{-2}(k).\end{aligned} \tag{12}$$

Define a quadratic function of the prediction errors as the minimization object of the critic NN:

$$E_c(k)=\frac{1}{2}e_c^2(k). \tag{13}$$

The weight update rule for the critic NN is a gradient-based adaptation which is given by

$$\hat{W}_c(k)=\hat{W}_c(k-n)+\triangle\hat{W}_c(k) \tag{14}$$

where

$$\triangle\hat{W}_c(k)=\alpha_c\left[-\frac{\partial E_c(k)}{\partial\hat{W}_c(k)}\right] \tag{15}$$

where $\alpha_c\in R$ is the adaptation gain. Then, the following Lemma is given to obtain the specific weight updating law.

Lemma 2 ([7]). *Given the matrices $A\in R^{m\times m}$, $X\in R^{n\times m}$ and vectors $b\in R^n$ and $q\in R^m$, the derivative of the following quadratic term with respect to the matrix X is given by*

$$\frac{d\left(\left(AX^Tb+q\right)^T\left(AX^Tb+q\right)\right)}{dX}=2b\left(A^T\left(AX^Tb+q\right)\right)^T \tag{16}$$

where the matrix A, vectors b and q are independent of the matrix X.

Combining (11),(12),(13) with (15), the weight updating law of the critic NN can be derived as

$$\begin{aligned}\hat{W}_c(k+n)=\hat{W}_c(k)-\alpha_c\phi_c(z(k))\cdot\big(&r_c\hat{J}(k)+a(k)D^{-2}(k)(q(k)\\&+W(k)-\hat{J}(k-1)\big).\end{aligned} \tag{17}$$

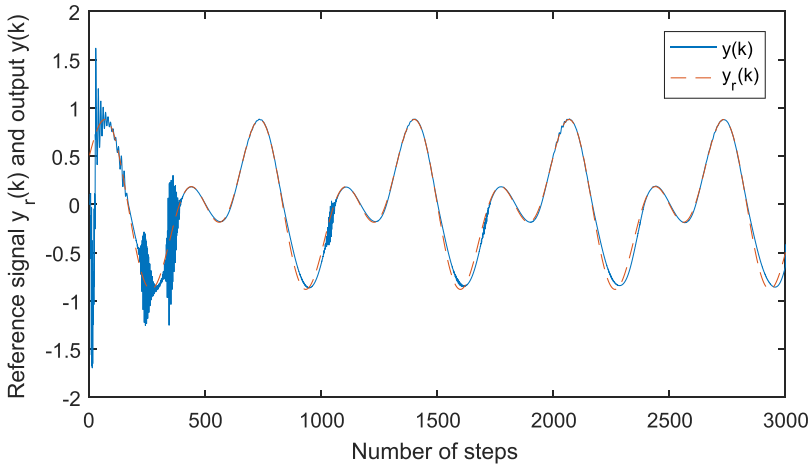

Fig. 1. Reference signal and system output.

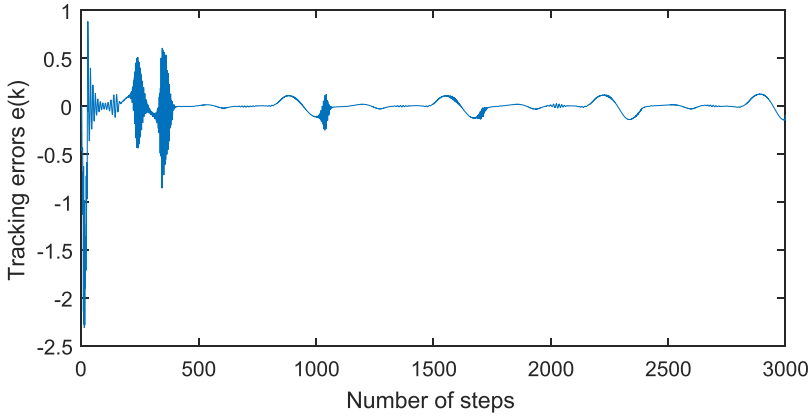

Fig. 2. Tracking error between reference signal $y_r(k)$ and output $y(k)$.

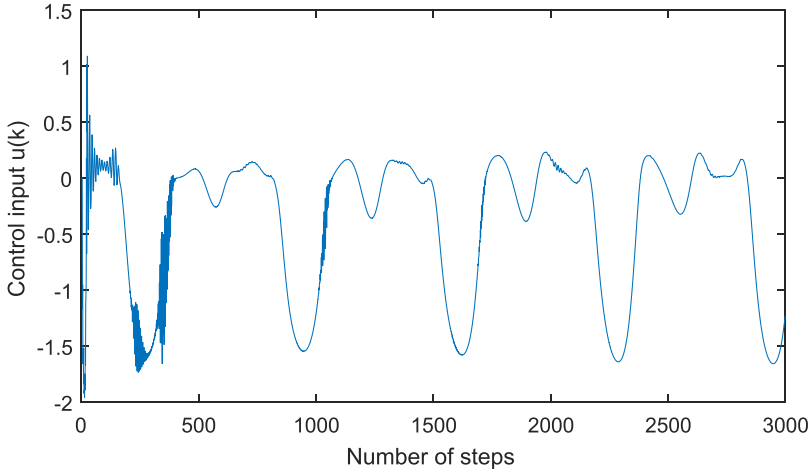

Fig. 3. Control signal $u(k)$.

3.2. Action NN and weight update law

In this section, an action NN is used to generate the input signal to approximate the desired control input (5). The desired control signal can be approximated as

$$u(k) = \upsilon\varphi\left(-\frac{1}{2}(\upsilon r)^{-1}U(k)\right) \tag{18}$$

where $U(k) = \hat{W}_a(k)S(V_a z(k))$ denotes the output of action NN. $\hat{W}_a(k)$ and V_a denote the weighted vector of output layer and the weighted matrix of hidden layer, respectively. $S(z(k))$ is shorted for $S(V_a z(k))$ which denotes the activation function vector of the hidden layer, $\hat{W}_a \in R^{n_a\times 1}$ and $V_a \in R^{n_a\times n_z}$, n_a is the number of the neurons in the hidden layer.

Define an auxiliary variable:

$$U^*(k) = W_a^T S(z(k)) \tag{19}$$

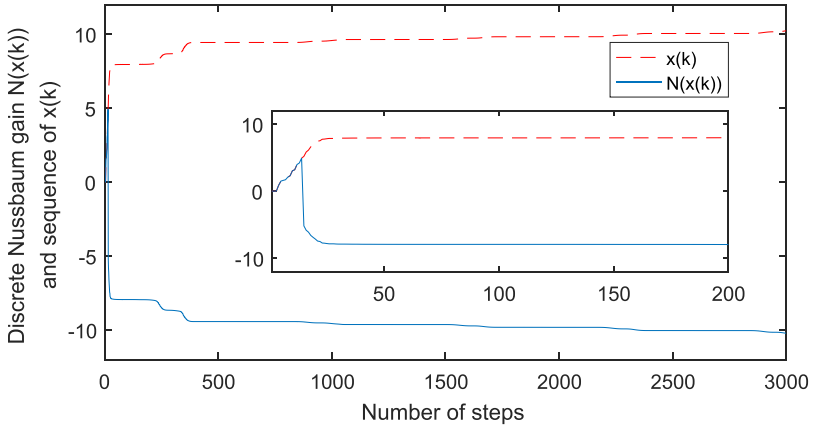

Fig. 4. Discrete Nussbaum gain $N(x(k))$ and its argument $x(k)$.

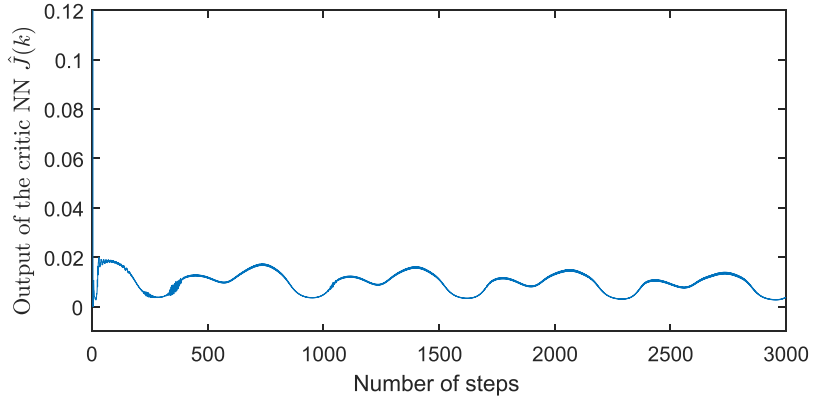

Fig. 5. Output of the critic NN $\hat{J}(k)$.

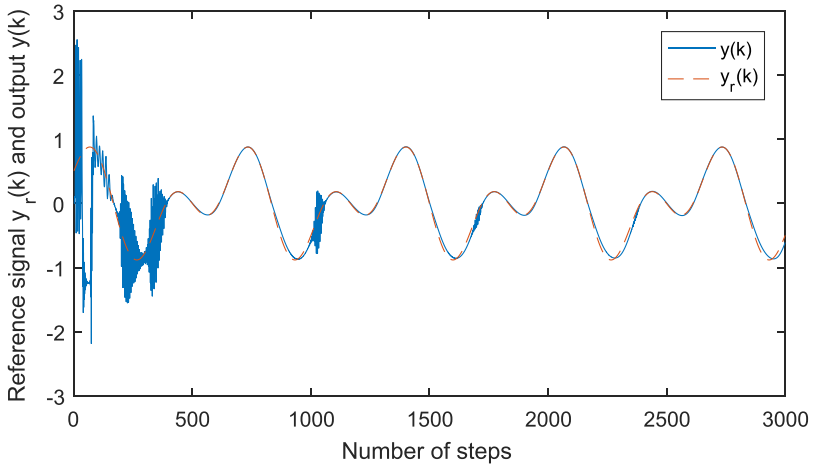

Fig. 6. Reference signal and system output.

where W_a denotes the desired weighted vector so that $u^*(k) = \upsilon\varphi\left(-\frac{1}{2}(\upsilon r)^{-1}(U^*(k)+d^*(k))\right)$, $d^*(k)$ denotes a bounded error.

The tracking error at instant k is defined as

$$\begin{aligned} e(k+n) &= y(k+n) - y_r(k+n) \\ &= \phi(\underline{z}(k), u(k)) - \phi(\underline{z}(k), u^*(k)) + d_\phi(k) \\ &= \Phi(\underline{z}(k), U(k)) - \Phi(\underline{z}(k), U^*(k)) + \bar{\eta}(k) \end{aligned} \tag{20}$$

where $\bar{\eta}(k) = d_\phi(k) - \Phi(\underline{z}(k), U^*(k)+d^*(k)) + \Phi(\underline{z}(k), U^*(k))$ and $d_\phi(k) = d_t(k) - y_r(k+n) + \phi(\underline{z}(k), u^*(k))$. From Assumption 3, we know that $\Phi(\cdot,\cdot)$ is a Lipschitz function and the boundedness of $\bar{\eta}(k)$, i.e.$|\bar{\eta}(k)| < \bar{\eta}^*$, where $\bar{\eta}^*$ is a positive constant. Then, (20) becomes

$$e(k+n) = \delta(k)(U(k) - U^*(k)) + \bar{\eta}(k) \tag{21}$$

where $\delta(\underline{z}(k), U^c(k)) = \frac{\partial\Phi(\underline{z}(k), U^c(k))}{\partial U^c(k)}$ is denoted as $\delta(k)$ for simplicity, $U^c(k) \in [\min\{U^*(\underline{z}(k)), U(k)\}, \max\{U^*(\underline{z}(k)), U(k)\}]$ and $\underline{\delta} \le |\delta(k)| \le \bar{\delta}, \bar{\delta} > \underline{\delta} > 0$.

Thus, the dynamic of closed-loop tracking error is expressed as

$$e(k+n) = \delta(k)\zeta_a(k) + \bar{\eta}(k) \tag{22}$$

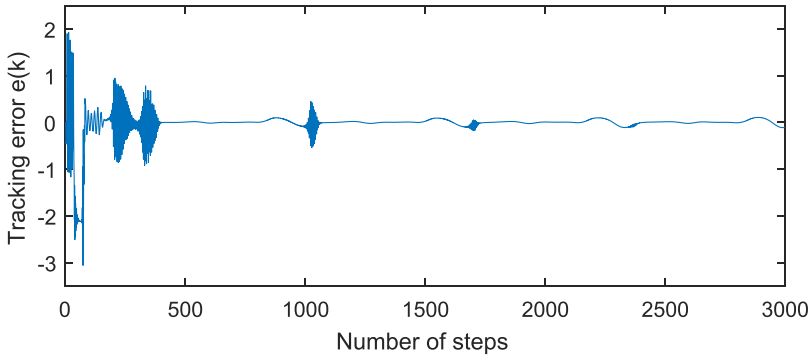

Fig. 7. Tracking error between reference signal $y_r(k)$ and output $y(k)$.

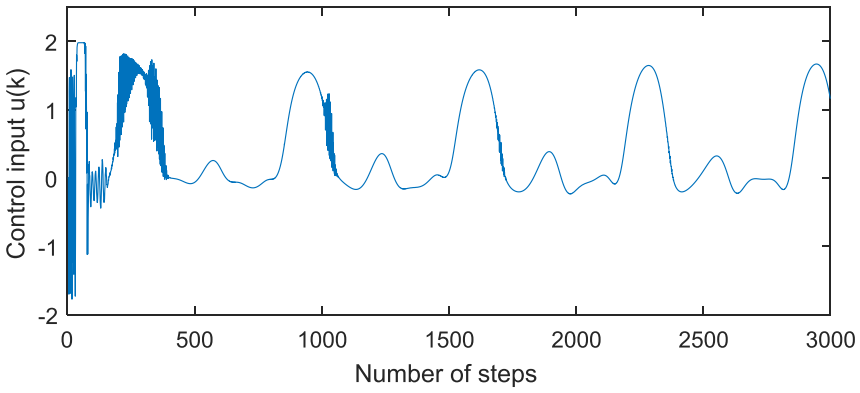

Fig. 8. Control signal $u(k)$.

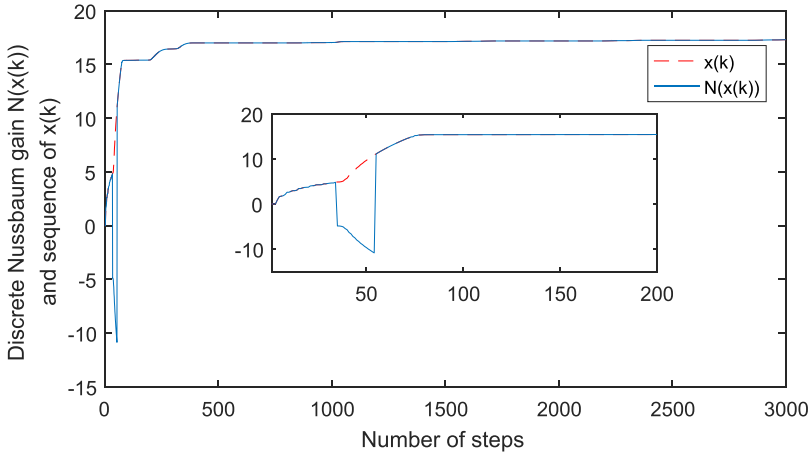

Fig. 9. Discrete Nussbaum gain $N(x(k))$ and its argument $x(k)$.

where $\tilde{W}_a(k) = \hat{W}_a(k) - W_a$ and

$$\zeta_a(k) = \tilde{W}_a(k)S(z(k)). \tag{23}$$

The design principle of the adaption law of the action NN is to minimize the SUF function and track the desired trajectory. The prediction error of action NN is defined as

$$e_a(k) = \left(N(x(k))a(k)\varepsilon(k) + a(k)\hat{f}(k-n)\right)D^{-1}(k) \tag{24}$$

where

$$\varepsilon(k) = \frac{e(k)}{G(k)} \tag{25}$$

$$\triangle x(k) = x(k+1) - x(k) = \frac{a(k)G(k)\varepsilon^2(k)}{D(k)} \tag{26}$$

$$G(k) = 1 + |N(x(k))| \tag{27}$$

$$D(k) = \left(1 + |N(x(k))| + \left|\hat{f}(k-n)\right|\right)\left(1 + \varepsilon^2(k) + \|S(k-n)\|^2\right) \tag{28}$$

$$a(k) = \begin{cases} 1, if|\varepsilon(k)| > \lambda \\ 0, otherwise. \end{cases} \tag{29}$$

where $x(0) = 0$ is the initial value of $x(k)$, λ is a threshold value and $\lambda > 0$.

Remark 4. Noting that $x(k)$ is the discrete sequence, which is the key factor to obtain an appropriate Nussbaum gain $N(x(k))$. The

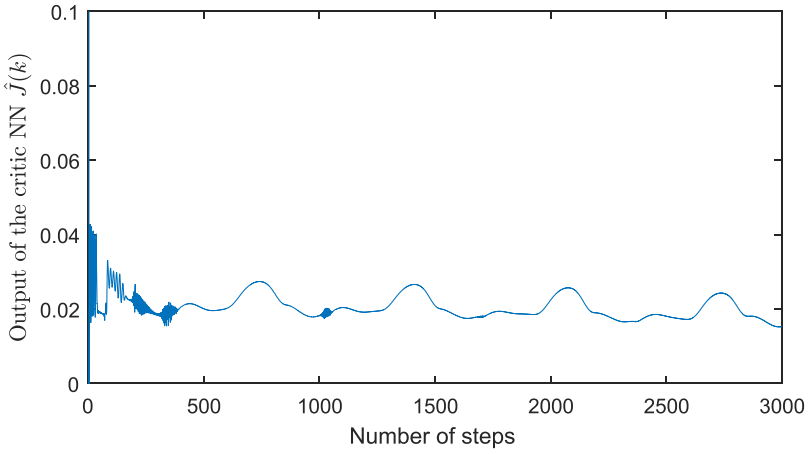

Fig. 10. Output of the critic NN $\hat{J}(k)$.

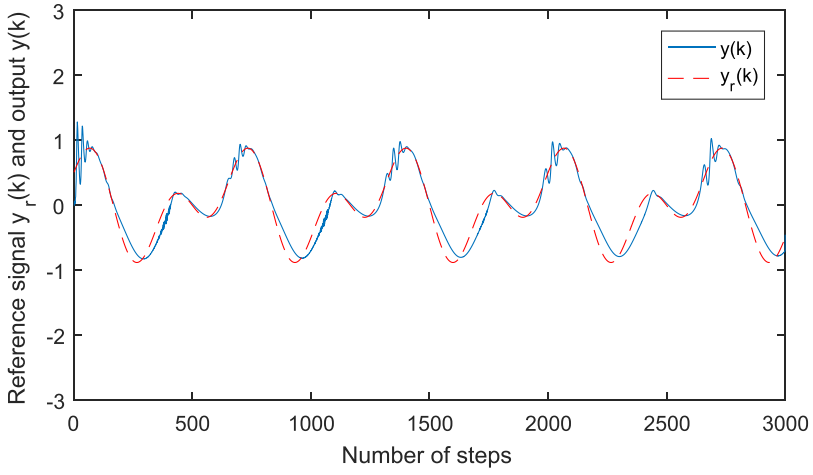

Fig. 11. Reference signal and system output of the output feedback NN control system proposed by [20] with $g = 1$.

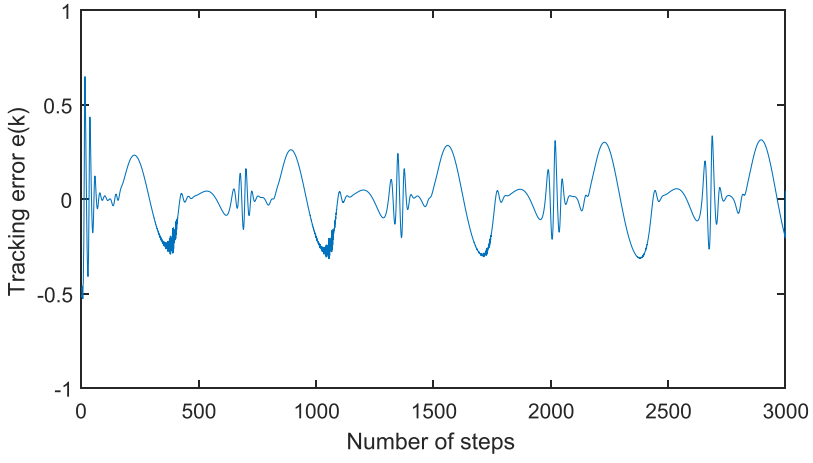

Fig. 12. Tracking error between reference signal $y_r(k)$ and output $y(k)$ of the output feedback NN control system proposed by [20] with g=1.

dynamic of $x(k)$ given by (25)–(29) is different from the existing works, which is one innovation of this paper.

Tune the weight of the action NN $\hat{W}_a(k)$ to minimize the error

$$E_a(k) = \frac{1}{2}e_a^2(k). \tag{30}$$

Combining (23),(24) and (30) with Lemma 2, we have

$$\triangle\hat{W}_a(k) = -\frac{S(k-n)}{D(k)}\Big(N(x(k))\varepsilon(k) + \hat{J}(k-n)\Big)\alpha_a a(k) \tag{31}$$

where $\alpha_a \in R^+$ is the adaptation gain of the action NN. So, the weight updating algorithm for the action NN is obtained as

$$\hat{W}_a(k) = \hat{W}_a(k-n) - \frac{S(k-n)}{D(k)}\Big(N(x(k))\varepsilon(k) + \hat{J}(k-n)\Big)\alpha_a a(k). \tag{32}$$

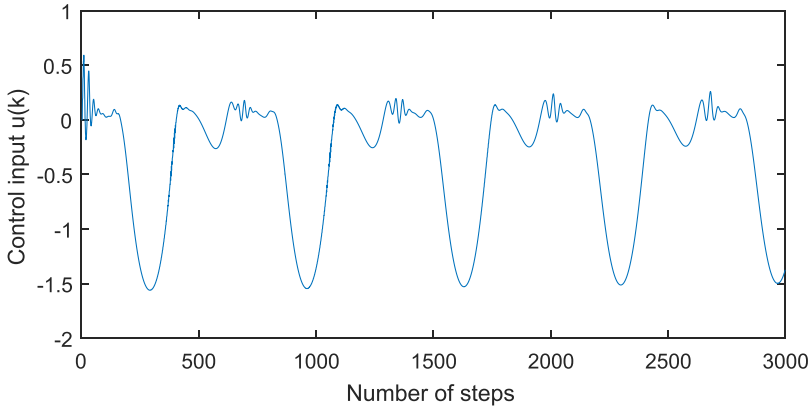

Fig. 13. Control signal $u(k)$ of the output feedback NN control system proposed by [20] with $g = 1$.

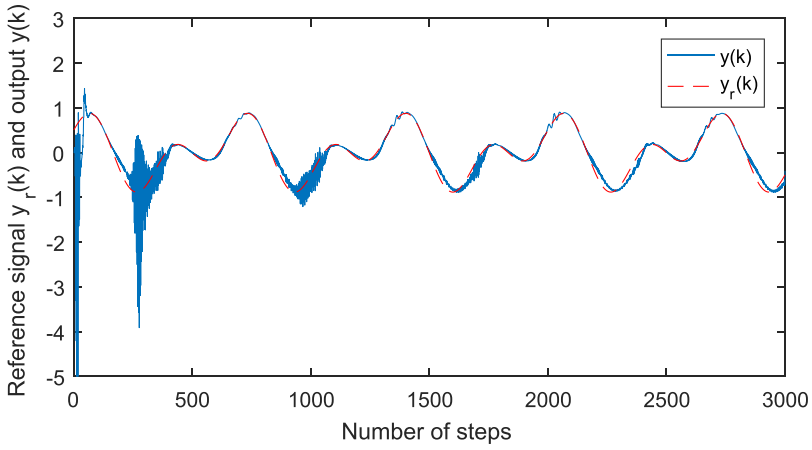

Fig. 14. Reference signal and system output of the output feedback NN control system proposed by [20] with $g = -1$.

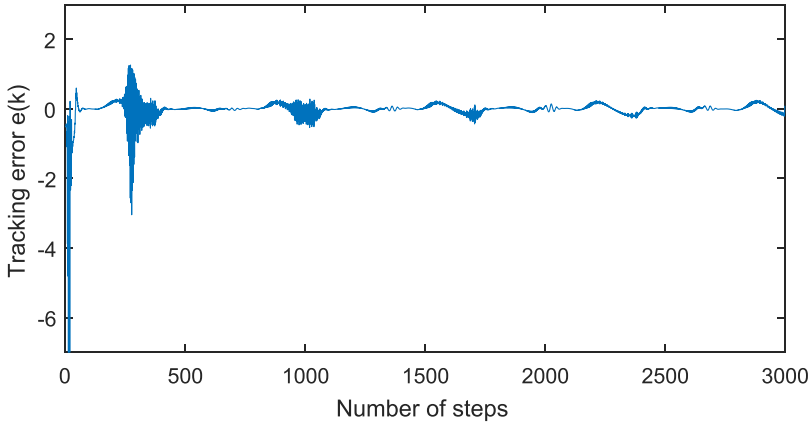

Fig. 15. Tracking error between reference signal $y_r(k)$ and output $y(k)$ of the output feedback NN control system proposed by [20] with g=-1.

Remark 5. Consider the MIMO discrete-time system with n subsystems Σ_j, $j = 1, \ldots, n$ in the non-affine pure-feedback form proposed in [21]. It can be transformed into an input-output model with the following form:

$$y_j(k+n_j) = F_j\big(\underline{y}_1(k+n_1-1), \ldots, \underline{y}_n(k+n_n-1), \bar{u}_j(k), d_j(k)\big),$$

where $n_j$is the order of subsystem Σ_j, $\underline{y}_j(k) = [y_j(k), y_j(k-1), \ldots, y_j(k-n_j+1)]^T$, $\bar{u}_j(k) = [u_1(k), u_2(k), \ldots, u_j(k)]^T$ and $\frac{\partial F_j(\cdot)}{\partial u_j(k)} = \prod_{i_j=1}^{n_j} g_{j,i_j}(\cdot) := g_j(\cdot), \underline{g}_j \le |g_j(\cdot)| \le \bar{g}_j$.

Then, using future output prediction procedure, the input-output model is further transformed into:

$$y_j(k+n_j) = \phi_j\big(\underline{z}_j(k), \bar{u}_j(k)\big) + d'_j(k),$$

where $\underline{z}_j(k) = [Y_1(k), \ldots, Y_{\bar{n}}(k), \underline{U}_1(k-1), \ldots, \underline{U}_{\bar{n}}(k-1)]^T$, $\bar{n} = \max_{j=1}^{n}(n_j)$, $Y_i(k) = \underline{y}_i(k)$, $\underline{U}_i(k-1) = [\bar{u}_i(k-1), \ldots, \bar{u}_i(k-n_i+1)]^T, i = 1, 2, \ldots, \bar{n}$, $d'_j(k) = F_{j,n_j}(\underline{z}_j(k), \bar{u}_j(k), \underline{d}_j(k)) - F_{j,n_j}(\underline{z}_j(k), \bar{u}_j(k), \mathbf{0}_{n_j})$, $\underline{d}_j(k) = [\underline{D}_1(k-1), \ldots, D_{\bar{n}}(k-1), d_j(k)]$ and $\underline{D}_i(k-1) = [d_i(k-1), \ldots, d_i(k-n_i+1)]^T$. $|d'_j(k)| \le \bar{d}_j$ is the bounded disturbance.

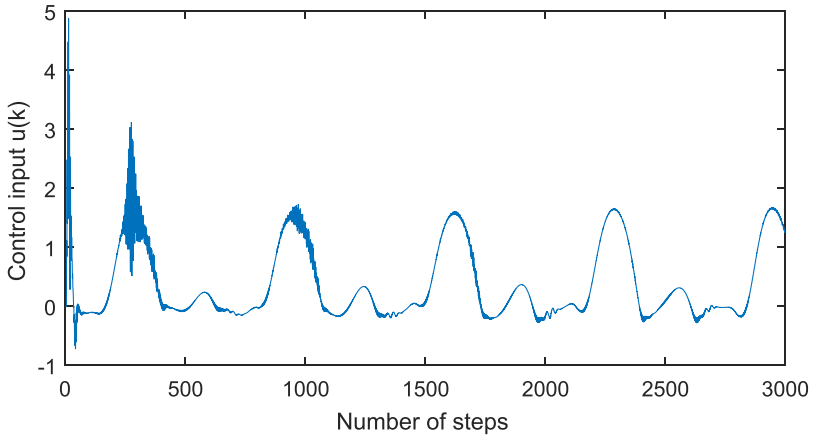

Fig. 16. Control signal $u(k)$ of the output feedback NN control system proposed by [20] with g=-1.

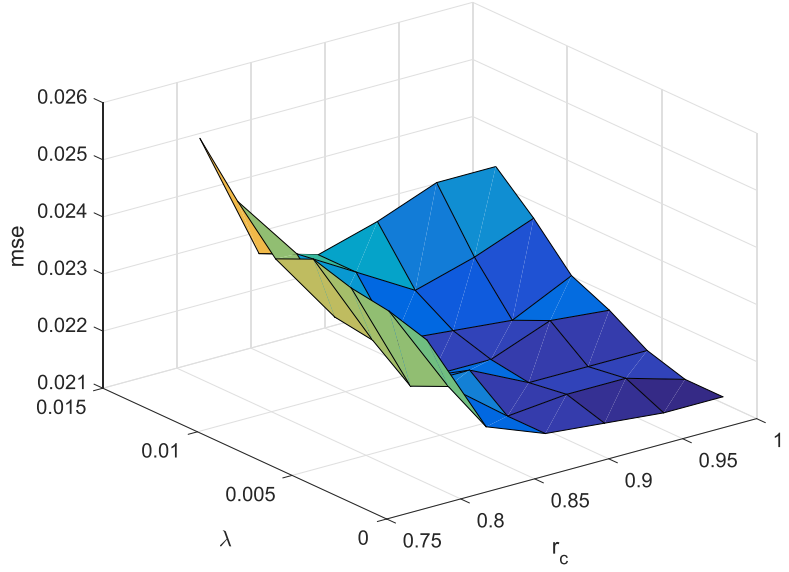

Fig. 17. The MSE trend along with the proposed systems with the values of r_c and λ.

Table 1
The MSE of the proposed systems with different values of r_c and λ

	r_c						
λ	0.76	0.79	0.83	0.87	0.91	0.95	0.99
0.001	0.03	0.0237	0.0219	0.0215	0.0214	0.0213	0.0213
0.003	0.03	0.0239	0.0223	0.0215	0.0216	0.0216	0.0213
0.005	0.0303	0.0240	0.022	0.022	0.0215	0.0215	0.0215
0.007	0.0306	0.0242	0.0225	0.0216	0.0216	0.022	0.0219
0.009	0.0304	0.0239	0.0226	0.0219	0.0218	0.0217	0.0222
0.011	0.0308	0.0246	0.0234	0.0228	0.0222	0.0225	0.0229
0.013	0.0309	0.0254	0.0231	0.0228	0.0231	0.0235	0.0235

Then, for each input-output model $y_j(k+n_j)$, we can design RL based control to obtain $u_j(k)$ using the SISO method proposed in this paper.

4. Theoretic Result

Assumption 4. Let ideal output layer weights W_a and W_c be bounded over the compact set Ω by known positive constants W_{aM} and W_{cM}, respectively. That is,

$\|W_a\| \le W_{aM}, \|W_c\| \le W_{cM}$.

Theorem 1. *Consider the nonlinear discrete-time systems given by (1). Let the Assumptions 1–4 hold and the disturbance bound with an unknown constant $\bar{d}_s$.Let the control input be provided by an output feedback control law (18) with an action NN and a critic NN. Let the weights of the action NN and the critic NN tune along with (32) and (17), respectively. Then the estimated error of NN weight $\tilde{W}_c(k)$ is UUB by positive constants*

$$D_c = \frac{1}{\sigma_{cM}}\sqrt{\frac{D_M^2}{r_2 r_c^2 - 2r_1 - 2r_1\alpha_a - \frac{2r_1\bar{\eta}^{*2}}{\underline{\delta}^2} - r_3}}$$

and the tracking error $e(k)$ and the estimated error of NN weight $\tilde{W}_a(k)$ are UUB provided that the following conditions hold:

(a) $0 < \alpha_c\|\phi_c(k-n)\|^2 < \frac{1}{r_c^2}$

(b) $\alpha_a > 0$

where α_c, α_a are NN adaptation gains and r_1, r_2 and r_3 are positive constants, which satisfy the following in-equations:

$$r_3 - \frac{r_2}{4} > 0, \tag{33}$$

$$r_2 r_c^2 - 2r_1 - 2r_1\alpha_a - \frac{2r_1\bar{\eta}^{*2}}{\underline{\delta}^2} - r_3 > 0, \tag{34}$$

$$r_4 - \frac{r_2}{4} - \frac{r_2}{2\underline{\delta}^2} > 0. \tag{35}$$

where r_4 is a positive constant.

Proof. See the Appendix A. □

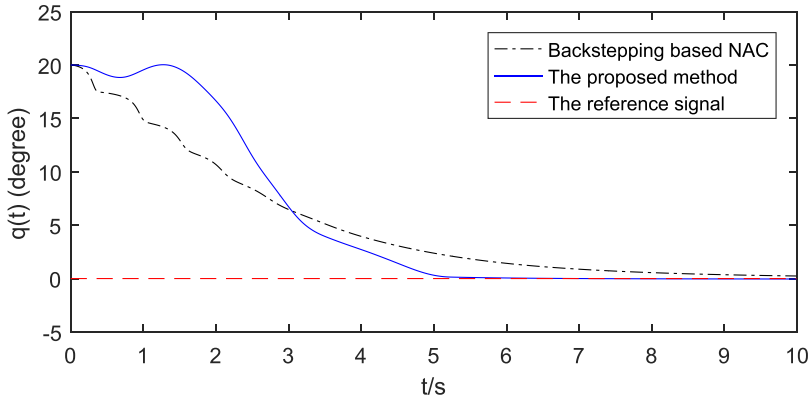

Fig. 18. Output signals of the proposed method and the NAC method proposed by [26].

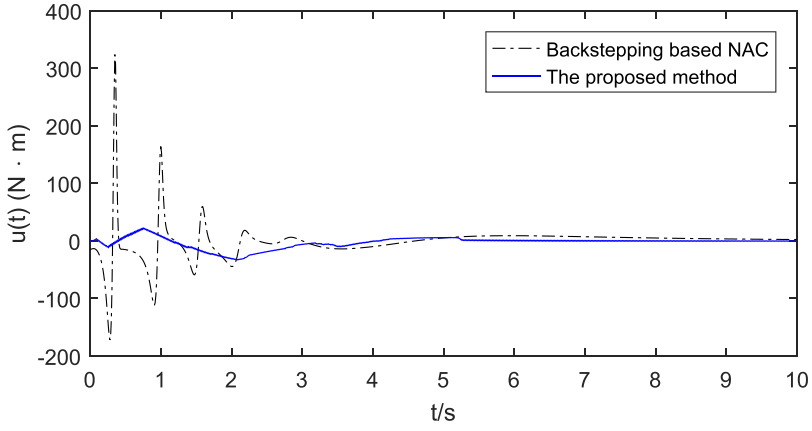

Fig. 19. Input signals of the proposed method and the NAC method proposed by [26].

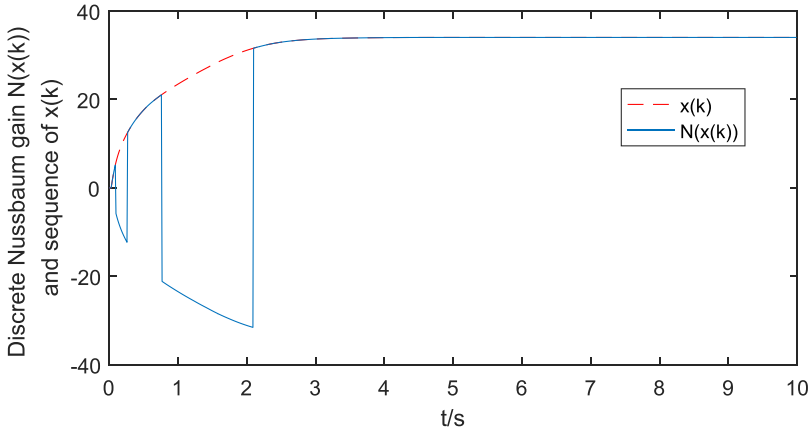

Fig. 20. Discrete Nussbaum gain $N(x(k))$ and its argument $x(k)$.

5. Numerical examples

5.1. Example 1

In this section, the following second-order nonlinear pure-feedback plant is considered for simulation studies:

$$\begin{aligned}\xi_1(k+1) &= f_1(\xi_1(k),\xi_2(k))\\ \xi_2(k+1) &= f_2(\xi_1(k),\xi_2(k),u(k))+d(k)\end{aligned} \tag{36}$$

where

$$f_1(\xi_1(k),\xi_2(k)) = 1.4\frac{\xi_1^2(k)}{1+\xi_1^2(k)} + 0.1\xi_2^3(k) + 0.5\xi_2(k)$$

$$f_2\big(\bar{\xi}_2(k),u(k)\big) = \frac{\xi_1(k)}{1+\xi_1^2(k)+\xi_2^2(k)} + gu(k)$$

where $g=\pm 1$ and the disturbance is $d(k) = 0.1cos(0.05k)cos(\xi_1(k))$. The control objective is to make the output $y(k)$ track the desired reference trajectory $y_r(k) = (1/2)sin((\pi/5)kT) + (1/2)cos((\pi/10)kT)$, where $T=0.05$, and guarantee the boundedness of all the closed-loop signals.

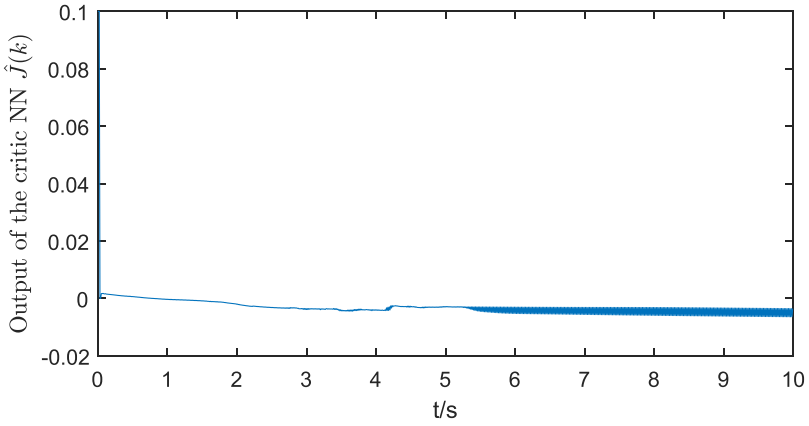

Fig. 21. Output of the critic NN $\hat{J}(k)$.

The system initial states are $\bar{\xi}_2(0) = [0.1, 0.1]^T$. The controller is constructed in the same manner as in Section 3. The tuning rates of the critic NN and the action NN are $\alpha_a = 5.6$ and $\alpha_c = 0.01$, respectively. The threshold value and the temporal difference coefficient and are chosen as $\lambda = 0.01$ and $r_c = 0.99$, respectively. The saturating bound of $u(k)$ is $\upsilon = 2$.

The system responses are presented in Figs. 1–5 when $g = 1$ is chosen. Fig. 1 shows the results of reference signal $y_r(k)$ and system output $y(k)$. Fig. 2 and 3 illustrate the boundedness of the tracking error $e(k)$ and the control input $u(k)$. Fig. 4 shows the boundedness of $x(k)$ and $N(x(k))$. Fig. 5 gives the output of the critic NN $\hat{J}(k)$.

Then, let $g = -1$. Employing the same control law and NN weights adaption laws, the system responses are shown in Figs. 6–10.

To illustrate the effectiveness of the proposed control method, the simulation results of the control method proposed by [20] are also shown in Figs. 11–16 as a comparison. The proposed method outperforms that proposed by [20] either the control gain g of simulation system (36) is assumed positive or negative, respectively.

Remark 6. The choice of the design parameters r_c and λ affects the convergence property of the proposed systems. How to tune them to obtain better performance is a topic worth to be discussed. Take System (36) with the positive control gain i.e. $g = 1$, for example. Fig. 17 shows the MSE trend along with the proposed systems with the values of r_c and λ changing from 0.76 to 0.99 and 0.001 to 0.013 within certain steps. The MSE values are listed in Table 1, when some values of r_c and λ chosen in this domain are applied. It can be seen that the MSE value decrease along the increase of r_c and the decrease of λ both from Fig. 17 and Table 1.

5.2. Example 2

In this section, a networked-control based robotic manipulator system is considered. Its dynamic is described by a Lagrangian equation:

$$J\ddot{q}(t) + B\dot{q} + MgLsin(q(t)) = u(t), \tag{37}$$

where q and $\dot{q}$ are the angle and angular velocity of the rigid link, respectively. J denotes the rotation inertia of the servo motor, B is the damping coefficient, L is the length from the axis of joint to the mass center, M is the mass of the link, and g is the gravitational acceleration. For this example, $J = 1$, $MgL = 10$ and $B = 2$ are assumed to be unknown parameters with respect to the controller design. The controller design task is to make the angle q move back to zero. The output feedback controller proposed in Section 3 is applied to System (37). The saturating bound of $u(k)$ is $\upsilon = 50$. The tuning rates of the critic NN and the action NN are $\alpha_c = 4.7$ and $\alpha_a = 0.01$. The threshold value and the temporal difference coefficient are chosen as $\lambda = 0.001$ and $r_c = 0.9$, respectively.

The simulation results are compared with a backstepping based nonlinear adaptive control (NAC) proposed by [26]. Fig. 18 and Fig. 19 show the output and input responses of the proposed method and the method proposed by [26], respectively. From the figures, it can be seen that the proposed method has faster responses and smaller amplitude of the control signals than the NAC method. Fig. 20 shows the boundedness of $x(k)$ and $N(x(k))$. Fig. 21 gives the output of the critic NN $\hat{J}(k)$.

6. Conclusions

In this paper, we develop an output feedback control algorithm based on RL approaches for a class of non-affine discrete-time nonlinear systems with unknown control directions and control constraints. Firstly, Implicit Function Theorem and Bellman's principle of optimality are employed to obtain the optimal control law. The controller is implemented by two NNs which are used as the actor and the critic to estimate the optimal control and the SUF approximately. In this architecture, the weights of the actor and the critic are simultaneous updated by Nussbaum gain-based adaption laws. Moreover, the UUB performance of the NN estimation weights and systems tracking errors is proved to be guaranteed. Finally, two numerical simulations are provided to demonstrate the effectiveness of the proposed control schemes.

Declaration of Competing Interest

None

CRediT authorship contribution statement

Miao Huang: Conceptualization, Methodology, Software, Writing - original draft. **Cong Liu:** Validation, Formal analysis. **Xiaoqi He:** Investigation, Data curation. **Longhua Ma:** Resources, Project administration. **Zheming Lu:** Writing - review & editing, Visualization. **Hongye Su:** Supervision, Funding acquisition.

Acknowledgments

This work is supported by the National Nature Science Foundation of China under Grant number 61633019, 61272020 and 61673268; Science Fund for Innovative Research Groups of the National Natural Science Foundation of China under Grant Number 61621002; Zhejiang Provincial Natural Science Foundation of China

under Grant number LQ19F030005; Natural Science Foundation of Ningbo City under Grants 2018A610165; Research Programs of Educational Commission Foundation of Zhejiang Province of China under Grant number Y201636903.

Appendix A

Proof of Theorem 1: Define the quadratic functions as

$$V_1(k)=\frac{r_1}{\alpha_a}\sum_{t=1}^{n}\tilde{W}_a^T(k-n+t)\tilde{W}_a(k-n+t) \tag{38}$$

$$V_2(k)=\frac{r_2}{\alpha_c}\sum_{t=1}^{n}\tilde{W}_c^T(k-n+t)\tilde{W}_c(k-n+t) \tag{39}$$

$$V_3(k)=r_3\zeta_c^2(k-n) \tag{40}$$

$$V_4=r_4\sum_{t=1}^{n}\triangle x(k-n+t) \tag{41}$$

where r_1, r_2, r_3 and r_4 are positive constants.

The difference of V_1 is given as

$$\triangle V_1(k)=\frac{r_1}{\alpha_a}\left(\tilde{W}_a^T(k)\tilde{W}_a(k)-\tilde{W}_a^T(k-n)\tilde{W}_a(k-n)\right). \tag{42}$$

Combining (22),(23), (24) and (32), we have

$$\begin{aligned}\triangle V_1(k)&=\frac{r_1}{\alpha_a}\Big(\big(\tilde{W}_a(k-n)+\triangle\hat{W}_a(k)\big)^T\big(\tilde{W}_a(k-n)+\triangle\hat{W}_a(k)\big)\\&\quad-\tilde{W}_a^T(k-n)\tilde{W}_a(k-n)\Big)\\&=\frac{r_1}{\alpha_a}\big(2\tilde{W}_a(k-n)\triangle W_a(k)-\triangle W_a^2(k)\big)\\&=\frac{r_1}{\alpha_a}\left(-\frac{2\alpha_a S(k-n)}{D(k)}\big(N(x(k))a(k)\varepsilon(k)+a(k)\hat{f}(k-n)\big)\right)\tilde{W}_a(k-n)\\&\quad+\frac{S^2(k-n)}{D^2(k)}\alpha_a^2\big(N(x(k))a(k)\varepsilon(k)+a(k)\hat{f}(k-n)\big)^2\end{aligned}$$

$$\begin{aligned}&=-\frac{2r_1S(k-n)}{D(k)}N(x(k))a(k)\varepsilon(k)\tilde{W}(k-n)\\&\quad+\frac{2S^2(k-n)r_1\alpha_a}{D^2(k)}(N(x(k))a(k)\varepsilon(k))^2\\&\quad+\frac{2S^2(k-n)r_1\alpha_a\hat{f}^2(k-n)a^2(k)}{D^2(k)}\\&\quad-\frac{2r_1S(k-n)}{D(k)}\hat{f}(k-n)\tilde{W}_a(k-n)a(k)\\&\le\frac{2r_1\alpha_a a(k)G(k)\varepsilon^2(k)}{D(k)}+\left|\frac{2\bar{\eta}}{\underline{\delta}\lambda}\right|\frac{r_1a(k)G(k)\varepsilon^2(k)}{D(k)}\\&\quad-\frac{2}{\delta(k-n)}N(x(k))\frac{r_1a(k)G(k)\varepsilon^2(k)}{D(k)}+\triangle V_{11}(k)\end{aligned} \tag{43}$$

where

$$\begin{aligned}\triangle V_{11}(k)&=\frac{2S^2(k-n)r_1\alpha_a\hat{f}^2(k-n)a^2(k)}{D^2(k)}\\&\quad-\frac{2r_1S(k-n)}{D(k)}\hat{f}(k-n)\tilde{W}_a(k-n)a(k)\\&\le\frac{2S^2(k-n)r_1\alpha_a\hat{f}^2(k-n)a^2(k)}{D^2(k)}\\&\quad-\frac{2r_1\hat{f}(k-n)a(k)}{D(k)\delta(k-n)}(\varepsilon(k)G(k)-\bar{\eta}(k))\\&\le\frac{2r_1\alpha_a\hat{f}(k-n)a^2(k)}{D(k)}-\frac{2r_1\hat{f}(k-n)a(k)\varepsilon(k)G(k)}{\delta(k-n)D(k)}\\&\quad+\frac{2r_1\hat{f}(k-n)a(k)\bar{\eta}(k)}{D(k)\delta(k-n)}\\&\le\frac{2r_1\alpha_a\hat{f}(k-n)a^2(k)}{D(k)}+\frac{r_1G(k)\varepsilon^2(k)a(k)}{\delta^2(k-n)D(k)}\\&\quad+\frac{r_1G(k)a(k)\hat{f}^2(k-n)}{D(k)}\\&\quad+\frac{2r_1\hat{f}(k-n)a(k)\bar{\eta}(k)}{D(k)\delta(k-n)}.\end{aligned} \tag{44}$$

Then, $\triangle V_1(k)$ becomes

$$\begin{aligned}\triangle V_1(k)&\le\frac{2r_1\alpha_a a(k)G(k)\varepsilon^2(k)}{D(k)}-\frac{2}{\delta(k-n)}N(x(k))\frac{r_1a(k)G(k)\varepsilon^2(k)}{D(k)}\\&\quad+\frac{r_1G(k)\varepsilon^2(k)a(k)}{\underline{\delta}^2D(k)}+r_1a(k)\big(J^*(\bar{z}(k-n))+\zeta_c(k-n)\big)^2\\&\quad+\frac{r_1\alpha_a a^2(k)\hat{f}^2(k-n)}{D(k)}\\&\quad+\frac{r_1a(k)\hat{f}^2(k-n)\bar{\eta}^{*2}}{\underline{\delta}^2D(k)}+\frac{r_1a(k)}{D(k)}+\left|\frac{2\bar{\eta}}{\underline{\delta}\lambda}\right|\frac{r_1a(k)G(k)\varepsilon^2(k)}{D(k)}\\&\quad+\frac{r_1\alpha_a a^2(k)}{D(k)}\\&\le\frac{2r_1\alpha_a a(k)G(k)\varepsilon^2(k)}{D(k)}+\left|\frac{2\bar{\eta}^*}{\underline{\delta}\lambda}\right|\frac{r_1a(k)G(k)\varepsilon^2(k)}{D(k)}\\&\quad-\frac{2}{\delta(k-n)}N(x(k))\frac{r_1a(k)G(k)\varepsilon^2(k)}{D(k)}\\&\quad+\frac{r_1G(k)\varepsilon^2(k)a(k)}{\underline{\delta}^2D(k)}+2r_1a(k)J^{*2}(\bar{z}(k-n))\\&\quad+2r_1a\zeta_c^2(k-n)+2r_1\alpha_a a^2(k)J^{*2}(\bar{z}(k-n))\\&\quad+2r_1\alpha_a a^2(k)\zeta_c^2(k-n)+r_1\alpha_a a^2(k)+\frac{2r_1a(k)\bar{\eta}^{*2}}{\underline{\delta}^2}J^{*2}(\bar{z}(k-n))\\&\quad+\frac{2r_1a(k)\bar{\eta}^{*2}}{\underline{\delta}^2}\zeta_c^2(k-n)+r_1a(k)\end{aligned} \tag{45}$$

The difference of $V_2(k)$ is shown as:

$$\begin{aligned}\triangle V_2(k)=V_2(k)-V_2(k-1)&=\frac{r_2}{a_{cj}}\big(\tilde{W}_{cj}^T(k)\tilde{W}_{cj}(k)\\&\quad-\tilde{W}_{cj}^T(k-n)\tilde{W}_{cj}(k-n)\big)\end{aligned}$$

Using (11), (12), (14) and (17), it can be following that

$$\begin{aligned}\triangle V_2(k)&=\frac{r_2}{\alpha_c}\big(-2\tilde{W}_c^T(k-n)\alpha_c r_c\phi_c(k-n)e_c(k-n)\\&\quad+\triangle W_c^T(k-n)\triangle W_c(k-n)\big)\\&=-2r_2r_c\zeta_c(k-n)e_c(k-n)+r_2\alpha_c r_c^2e_c^2(k-n)\parallel\phi_c(k-n)\parallel^2\\&\le-r_2\big(1-\alpha_c r_c^2\parallel\phi_c(k-n)\parallel^2\big)e_c^2(k-n)-r_cr_c^2\zeta_c^2(k-n)\\&\quad+\frac{r_2}{4}\zeta_c^2(k-n-1)\\&\quad+\frac{r_2}{4}(r_cJ^*(\bar{z}(k-n))-J^*(\bar{z}(k-n-1)))^2\\&\quad+\frac{r_2}{4D^2(k-n)}(q(k-n)+W(k-n))\\&\quad+\frac{r_2}{4}(\varepsilon_c(k-n)-\varepsilon_c(k-n-1))^2\\&\le-r_2\big(1-\alpha_c r_c^2\parallel\phi_c(k-n)\parallel^2\big)e_c^2(k-n)-r_2r_c^2\zeta_c^2(k-n)\\&\quad+\frac{r_2}{4}\zeta_c^2(k-n-1)\\&\quad+\frac{r_2}{4}(r_cJ^*(\bar{z}(k-n))-J^*(\bar{z}(k-n-1)))^2\\&\quad+\frac{a(k-n)r_c\parallel e(k-n)\parallel^2}{4D^2(k-n)}\\&\quad+\frac{r_2}{4}\frac{a(k-n)}{D^2(k-n)}\int_0^{u(k-n)}\varphi^{-1}\big(\bar{u}^{-1}s\big)\bar{u}rds\end{aligned}$$

$$
\begin{aligned}
&+\frac{r_2}{4}(\varepsilon_c(k-n)-\varepsilon_c(k-n-1))^2\\
&\le -r_2\big(1-\alpha_c r_c^2 \parallel \phi_c(k-n)\parallel^2\big)e_c^2(k-n)-r_2r_c^2\zeta_c^2(k-n)\\
&\quad+\frac{r_2}{4}\zeta_c^2(k-n-1)\\
&\quad+\frac{r_2}{4}(r_cJ^*(\bar z(k-n))-J^*(\bar z(k-n-1)))^2\\
&\quad+\frac{a(k-n)r_2\parallel e(k-n)\parallel^2}{4D^2(k-n)}\\
&\quad+\frac{r_2}{4}\frac{a(k-n)}{D^2(k-n)}\int_0^{\upsilon\varphi\left(\frac{1}{2}(\upsilon r)^{-1}(\|\zeta_a(k-n)\|+\|W_aS(z(k-n))\|)\right)}\varphi^{-1}\big(\upsilon^{-1}s\big)\upsilon r ds\\
&\quad+\frac{r_2}{4}(\varepsilon_c(k-n)-\varepsilon_c(k-n-1))^2\\
&\le -r_2\big(1-\alpha_c r_c^2 \parallel \phi_c(k-n)\parallel^2\big)e_c^2(k-n)-r_2r_c^2\zeta_c^2(k-n)\\
&\quad+\frac{r_2}{4}\zeta_c^2(k-n-1)\\
&\quad+\frac{r_2}{4}(r_cJ^*(\bar z(k-n))-J^*(\bar z(k-n-1)))^2\\
&\quad+\frac{a(k-n)r_2\parallel e(k-n)\parallel^2}{4D^2(k-n)}\\
&\quad+\frac{r_2}{4}\frac{a(k-n)}{D^2(k-n)}(\parallel\zeta_a(k-n)\parallel+\parallel W_aS(z(k-n))\parallel)^2\\
&\quad+\frac{r_2}{4}\varepsilon_{cm}^2\\
&\le -r_2\big(1-\alpha_c r_c^2 \parallel \phi_c(k-n)\parallel^2\big)e_c^2(k-n)-r_2r_c^2\zeta_c^2(k-n)\\
&\quad+\frac{r_2}{4}\zeta_c^2(k-n-1)\\
&\quad+\frac{r_2}{4}(r_cJ^*(\bar z(k-n))-J^*(\bar z(k-n-1)))^2\\
&\quad+\frac{a(k-n)r_2\varepsilon^2(k-n)G(k-n)}{4D(k-n)}\\
&\quad+\frac{r_2}{2}\frac{a(k-n)\varepsilon^2(k-n)G(k-n)}{D(k-n)\delta^2(k-n-1)}\\
&\quad+\frac{r_2a(k-n)}{2D^2(k-n)}\parallel W_aS(z(k-n))\parallel^2+\frac{r_2}{4}\varepsilon_{cm}^2. \qquad (46)
\end{aligned}
$$

Then, from the definition of $\triangle x$, let $c_1=2r_1\alpha_a+\left|\frac{2\bar\eta^*}{\underline\delta\lambda}\right|+\frac{r_1}{\underline\delta^2}$ and obtain a positive-definite $V(k)$ as

$$V(k)=V_1(k)+V_2(k)+V_3(k)+V_4(k).$$

The first difference of $V(k)$ is given as

$$\triangle V(k)=\triangle V_1(k)+\triangle V_2(k)+\triangle V_3(k)+\triangle V_4(k). \qquad (47)$$

Combining (45), (46) and (47), we derive

$$
\begin{aligned}
\triangle V(k)=&-\frac{2}{\delta(k-n)}N(x(k))\frac{r_1a(k)G(k)\varepsilon^2(k)}{D(k)}+c_1\triangle x(k)\\
&-\left(r_2r_c^2-2r_1a(k)-2r_1\alpha_aa^2(k)-\frac{2r_1a(k)\bar\eta^{*2}}{\underline\delta^2}\right)\zeta_c^2(k-n)\\
&-r_2\big(1-\alpha_cr_c^2\|\phi_c(k-n)\|^2\big)e_c^2(k-n)+\frac{r_2}{4}\zeta_c^2(k-n-1)\\
&+\frac{r_2}{4}\triangle x(k-n)+\frac{r_2}{2\underline\delta^2}\triangle x(k-n)+r_3\zeta_c^2(k-n)-r_3\zeta_c^2(k-n-1)\\
&+r_4\triangle x(k)-r_4\triangle x(k-n)+D_M^2 \qquad (48)
\end{aligned}
$$

where

$$
\begin{aligned}
D_M^2=&\frac{r_2}{4}(r_cJ^*(\bar z(k-n))-J^*(\bar z(k-n-1)))^2\\
&+\frac{r_2a(k-n)}{2D^2(k-n)}\|W_a\|^2S_m^2\\
&+\frac{r_2}{4}\varepsilon_{cm}^2+2r_1aJ^*(\bar z(k-n))+2r_1\alpha_aa^2(k)J^{*2}(\bar z(k-n))\\
&+r_1\alpha_aa^2(k)+\frac{2r_1a(k)\eta^{*2}}{\underline\delta^2}J^*(\bar z(k-n))+r_1a(k)
\end{aligned}
$$

from the fact that the activation functions are bounded by known positive values, i.e. $\|S(k)\|\le\sigma_{aM}$ and the boundedness of $J^*(\bar z(k))$.

Suppose that the design parameter r_1, r_2, r_3 and r_4 satisfy Inequations (33)-(35). Then, the following condition holds:

$$\|\zeta_c(k-n)\|>\frac{D_M}{\sqrt{r_2r_c^2-2r_1-2r_1\alpha_a-\frac{2r_1\bar\eta^{*2}}{\underline\delta^2}-r_3}}. \qquad (49)$$

Denote $N'(x(k))=\frac{1}{\delta(k-n)}N(x(k))r_1$ and then, noting $1/\bar\delta\le 1/\delta(k-n)\le 1/\underline\delta$ and according to Lemma 1, it can be seen that $N'(x(k))$ is also a discrete Nussbaum gain. Deriving the summation of the right side of (47) and noting $0\le\triangle x(k)\le 1$, we have

$$\triangle V(k)\le(c_1+r_4)\triangle x(k)-2N'(x(k))\triangle x(k)$$

and

$$V(k)\le -2\sum_{k'=0}^{k}N'\big(x(k')\big)\triangle x(k')+c_1x(k)+c_1+r_4. \qquad (50)$$

Applying Lemma 1 to (50), we have the boundedness of $V(k)$ and $x(k)$. Noting that the definition of $V(k)$, the boundedness of $\|\tilde W_a(k)\|$, $\|\tilde W_c(k)\|$ is obtained. Since $|N(x(k))|=\left|\sup_{k'\le k}\{x(k')\}\right|$, it can be implied that $N(x(k))$ and $G(k)=1+|N(x(k))|$ are bounded. In addition, from $\triangle x(k)\ge 0$, we can find $x(k)$ is a non-decreasing sequence. Thus, we have

$$\lim_{k\to 0}\triangle x(k)=0.$$

Note $\|\zeta_c(k)\|\le\sigma_{cM}\|\tilde W_c(k)\|$. Then, by using (49), it can be derived that

$$\|\tilde W_c(k)\|>\frac{1}{\sigma_{cM}}\sqrt{\frac{D_M^2}{r_2r_c^2-2r_1-2r_1\alpha_a-\frac{2r_1\bar\eta^{*2}}{\underline\delta^2}-r_3}},$$

where $\sigma_{cM}\ge\|\phi_c(k)\|$ from the fact that the activation function for the critic NN is bounded by known positive constants over the compact set Ω.

Let us define a time interval as $Z=\{k|a(k)=1\}$ and suppose that Z is an infinite set. Then, we have

$$\lim_{k\to 0,k\in Z}\varepsilon(k)=\lim_{k\to 0,k\in Z}a(k)\varepsilon(k)=0$$

which conflicts with $a(k)=1, k\in Z$, because $|\varepsilon(k)|\ge\lambda$, when $a(k)=1$. Therefore, Z is a finite set, and then it follows

$$\lim_{k\to\infty}a(k)=0,\ \lim_{k\to\infty}\sup\{\varepsilon(k)\}\le\lambda,$$

which indicates that $N(x(k))$ converges to a constant ultimately. From the definition of $\varepsilon(k)$, the tracking errors satisfy

$$\lim_{k\to\infty}\sup\{|e(k)|\}\le\frac{C\lambda}{\alpha_a}$$

where C denotes the limit of $G(k)$.

Then, the conclusion that the tracking error vector $e(k)$, the weights of estimation error for the action NN $\tilde W_a(k)$, and the weights of estimation error for the critic NN $\tilde W_c(k)$ are all UUB can be drawn. The proof is completed.

References

[1] P. J. Werbos, Using ADP to understand and replicate brain intelligence: the next level design, proceedings of the IEEE international symposium on approximate dynamic programming and reinforcement learning, 209-216, honolulu, HI, USA, 1-5 april, 2007.

[2] P.J. Werbos, ADP: the key direction for future research in intelligent control and understanding brain intelligence, IEEE Transactions on Systems, Man, Cybernetics, Part B, Cybernetics 38 (4) (2008) 898–900.

[3] X. Yang, D. Liu, D. Wang, Q. Wei, Discrete-time online learning control for a class of unknown nonaffine nonlinear systems using reinforcement learning, Neural Networks 55 (7) (2014) 30–41.

[4] Y. Zhang, S. Wang, MLP technique based reinforcement learning control of discrete pure-feedback systems, Neurocomputing 168 (11) (2015) 401–407.
[5] M.B. Radac, R.E. Precup, R.C. Roman, Model-free control performance improvement using virtual reference feedback tuning and reinforcement q-learning, International Journal of Systems Science 48 (5) (2007) 1071–1083.
[6] Q. Yang, S. Jagannathan, Reinforcement learning controller design for affine nonlinear discrete-time systems using online approximators, IEEE Transactions on Systems, Man and Cybernetics, Part B, Cybernetics 42 (2) (2012) 377–390.
[7] P. He, S. Jagannathan, Reinforcement learning-based output feedback control of nonlinear systems with input constraints, IEEE Transactions on Systems, Man and Cybernetics Part B Cybernetics 35 (1) (2005) 150–154.
[8] B. Xu, C. Yang, Z. Shi, Reinforcement learning output feedback NN control using deterministic learning technique, IEEE Transactions on Neural Networks and Learning Systems 25 (3) (2014) 635–641.
[9] P. Shih, B.C. Kaul, S. Jagannathan, J.A. Drallmeier, Reinforcement-learning-based output-feedback control of nonstrict nonlinear discrete-time systems with application to engine emission control, IEEE Transactions on Systems, Man and Cybernetics Part B Cybernetics 39 (5) (2009) 1162–1179.
[10] L.M. Zhu, H. Modares, G.O. Peen, F.L. Lewis, B. Ye, Adaptive suboptimal output-feedback control for linear systems using integral reinforcement learning, IEEE Transactions on Control Systems Technology 23 (1) (2014) 264–273.
[11] H. Modares, F.L. Lewis, Z.P. Jiang, Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning, IEEE Transactions on Cybernetics 46 (11) (2016) 2401–2410.
[12] X. Yang, D. Liu, D. Wang, Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints, International Journal of Control 87 (3) (2014) 553–566.
[13] C. Liu, H. Zhang, G. Xiao, et al., Integral reinforcement learning based decentralized optimal tracking control of unknown nonlinear large-scale interconnected systems with constrained-input, Neurocomputing (2018) 1–11.
[14] Z. Yin, W. He, C. Yang, C. Sun, Control design of a marine vessel system using reinforcement learning, Neurocompution 311 (2018) 353–362.
[15] C. Yang, S.S. Ge, T.H. Lee, Output feedback adaptive control of a class of nonlinear discrete-time systems with unknown control directions, Automatica 45 (1) (2009) 270–276.
[16] X.J. Wang, X.H. Yin, Q.H. Wu, F.Q. Meng, Adaptive neural tracking control for nonstrict-feedback nonlinear systems with unknown backlash-like hysteresis and unknown control directions, International Journal of Robust and Nonlinear Control 28 (16) (2018) 5140–5157.
[17] Y. Liu, S. Tong, Barrier lyapunov functions for nussbaum gain adaptive control of full state constrained nonlinear systems, Automatica 76 (2017) 143–152.
[18] L. Liu, Z. Wang, Z. Huang, H. Zhang, Adaptive predefined performance control for MIMO systems with unknown direction via generalized fuzzy hyperbolic model, IEEE Transactions on Fuzzy Systems 25 (3) (2017) 527–542.
[19] H. Ma, H.J. Liang, H.J. Ma, Q. Zhou, Nussbaum gain adaptive backstepping control of nonlinear strict-feedback systems with unmodeled dynamics and unknown dead zone, International Journal of Robust and Nonlinear Control 28 (17) (2018) 5326–5343.
[20] C. Yang, S.S. Ge, C. Xiang, T. Chai, T.H. Lee, Output feedback NN control for two classes of discrete-time systems with unknown control directions in a unified approach, IEEE Transactions on Neural Networks 1 (11) (2008) 1873–1886.
[21] Y. Li, C. Yang, S.S. Ge, T.H. Lee, Adaptive output feedback NN control of a class of discrete-time MIMO nonlinear systems with unknown control directions, IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics 41 (2) (2011) 507–517.
[22] J.R. Mundres, Analysis on Manifolds. Reading, MA: Addison-Wesley, 1991.
[23] H. Zhang, Y. Luo, D. Liu, Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints, IEEE Transactions on Neural Networks 20 (9) (2009) 1490–1503.
[24] T.H. Lee, K.S. Narendra, Stable discrete adaptive control with unknown high-frequency gain, IEEE Transactions on Automatic Control 31 (5) (1986) 477–479.
[25] S.S. Ge, C. Yang, T.H. Lee, Adaptive robust control of a class of nonlinear strict-feedback discrete-time systems with unknown control directions, Systems & Control Letters 57 (11) (2008) 888–895.
[26] L. Xing, C. Wen, Z. Liu, H. Su, J. Cai, Event-triggered adaptive control for a class of uncertain nonlinear systems, IEEE Transactions on Automatic Control 62(4) 2071–2076.

Miao Huang received the Ph.D. degree in control theory and engineering from East China University of Science and Technology, Shanghai, China, in 2015. From 2017 to 2019, she was a postdoctoral fellow with college of control science and engineering, Zhejiang University, Hangzhou, China. Currently, she is with College of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai, China. She has published 15 international journal and conference papers and 5 patents for invention. She has taken charge of a Zhejiang Provincial Natural Science Foundation of China. Her research interests include adaptive control and dynamic wireless power transfer.

Cong Liu received the Ph.D. degree in control theory and engineering from Tongji University, Shanghai, China, in 2016. Currently, he is with Department of Internet of Things Engineering, Shanghai Business School, Shanghai, China. He has published over 10 international journal and conference papers. He has taken charge of a Natural Science Foundation of Ningbo City and Research Programs of Educational Commission Foundation of Zhejiang Province of China. His research interests include medical image processing and computer vision.

Xiaoqi He received the Ph.D. degree in control theory and control engineering from Zhejiang University, Hangzhou, China, in 2003. From 2003 to 2006, he engaged in developing network security products in Hangzhou Guangyang technology co.,LTD, Hangzhou, China. From 2006 to 2019, he was an associate professor with Ningbo Institute of Technology, Zhejiang University, Ningbo, China. Currently, he is with Ningbo Industrial Internet Institute, Ningbo, China. His research interests include industrial internet and network security.

Longhua Ma received the B.S. degree in industrial electrical automation from Lanzhou Jiaotong University, Lanzhou, China, in 1986, the M.S. degree and Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1993 and 2002, respectively. He was an associate research fellow with National engineering research center for industrial automation, Zhejiang University, Hangzhou, China from 1993 to 2008. From 2008 to 2012, he was an associate professor with School of aeronautics and astronautics, Zhejiang University, Hangzhou, China. Currently, he is a professor with Ningbo Industrial Internet Institute, Ningbo, China. He has (co)author four books and published over 70 international journal and conference papers. His currently research interests include network security, new energy and electric vehicle energy management and control and inertial navigation theory and application.

Zheming Lu received the B.S. degree in electromagnetic measurement, the M.S. degree in electric engineering and the Ph.D. degree in test and measure technique and instrument from Harbin Institute of Technology, Harbin, China, respectively. From 2001 to 2003, he was a doctoral researcher with department of electrical engineering, Harbin Institute of Technology. From 2000 to 2006, he was a professor with department of automation test and control, Harbin Institute of Technology, Harbin, China. Currently, he is a professor and head of aerospace department, School of aeronautics and astronautics, Zhejiang University, Hangzhou, China. He was an editorial board member of International Journal of Innovation Computing and Information Control from 2005 to 2008. He was the executive editor of International Journal of Computer Sciences and Engineering Systems from 2007 to 2012. He served as the editor-in-chief of Information Analysis and Processing from 2008 to 2013. Since 2013, he has served as the editorial board member of KSII Transactions on Internet and Information Systems. His currently research interests include multimedia signal processing, information hiding and complex network.

Hongye Su received the Ph.D. degree, majoring in industrial automation from Zhejiang University, Hangzhou, China, in 1995. Then he was promoted to full Professor at Zhejiang University, Hangzhou, China, in 2000. Since October 1999, he has held the position of Vice Director of the Institute of Cyber-Systems and Control (formerly the Institute of Advanced Control) at Zhejiang University. He had also been the Dean of the Department of Control Science and Engineering at Zhejiang University between 2011 and 2013. He was promoted as Chair Professor of Chang Jiang Scholars Program by Ministry of Education of China in 2012. He has taken charge of a large number of significant research projects including National Basic Research Program of China (973 Program) and National Hightechnology Research and Development Program of China (863 Program). He has published more than 100 international journal papers and coauthored four professional books. His main academic activities are in the areas of robust control, nonlinear adaptive control and chemical industrial process control. Dr. Su has been executive members of board of governors at various academic institutions including Chinese Association of Automation (CAA), China Instrument and Control Society (CIS). He received the National Science Fund for Distinguished Young Scholars from National Natural Science Foundation of China, in 2000.