# Optimal Neuro-Control Strategy for Nonlinear Systems With Asymmetric Input Constraints

Xiong Yang, *Member, IEEE* and Bo Zhao, *Member, IEEE*

***Abstract*****—In this paper, we present an optimal neuro-control scheme for continuous-time (CT) nonlinear systems with asymmetric input constraints. Initially, we introduce a discounted cost function for the CT nonlinear systems in order to handle the asymmetric input constraints. Then, we develop a Hamilton-Jacobi-Bellman equation (HJBE), which arises in the discounted cost optimal control problem. To obtain the optimal neurocontroller, we utilize a critic neural network (CNN) to solve the HJBE under the framework of reinforcement learning. The CNN's weight vector is tuned via the gradient descent approach. Based on the Lyapunov method, we prove that uniform ultimate boundedness of the CNN's weight vector and the closed-loop system is guaranteed. Finally, we verify the effectiveness of the present optimal neuro-control strategy through performing simulations of two examples.**

***Index Terms*****—Adaptive critic designs (ACDs), asymmetric input constraint, critic neural network (CNN), nonlinear systems, optimal control, reinforcement learning (RL).**

## I. Introduction

REINFORCEMENT learning (RL), known as a research branch of machine learning, has been an effective tool in solving nonlinear optimization problems [1]. The main idea behind RL is to create an architecture to learn optimal policies without systems' information. A well-known architecture used in RL is the actor-critic structure, which is comprised of two neural networks (NNs), that is, actor and critic NNs. The mechanism of implementing the actor-critic structure is as follows: The actor NN generates a control policy to surroundings or plants, and the critic NN (CNN) estimates the cost stemming from that control policy and gives a positive/negative signal to the actor NN [2]. Owing to this mechanism of actor-critic structure, one is able to not only obtain optimal policies without knowing systems' prior knowledge, but also avoid "the curse of dimensionality" occurring [3]. According to [4], adaptive dynamic programming (ADP) also takes the actor-critic structure as an implementation architecture and shares similar spirits as RL. Thus, researchers often use ADP and RL as two interchangeable names. During the past few years, quite a few ADP and RL approaches emerged, such as goal representation ADP [5], policy/value iteration ADP [6], [7], event-sampled/triggered ADP [8], [9], robust ADP [10], integral RL [11], [12], online RL [13], [14], off-policy RL [15], [16].

Doubtlessly, the actor-critic structure utilized in RL has achieved great success in solving nonlinear optimization problems (see aforementioned literature). However, when tackling optimal control problems of nonlinear systems with available systems' information, researchers found that the actor-critic structure could be reduced to a structure with only the critic, i.e., the critic-only structure [17]. The early research on solving optimization problems via a critic-only structure can be tracked to the work of Widrow *et al.* [18]. Later, Prokhorov and Wunsch [19] named this critic-only structure as a kind of adaptive critic designs (ACDs), which were originated from RL. After that, Padhi *et al.* [20] suggested a single network ACD to learn an optimal control policy for input-affine discrete-time (DT) nonlinear systems. Recently, Wang *et al.* [21] introduced a data-based ACD to acquire the robust optimal control of continuous-time (CT) nonlinear systems. Apart from the identifier NN used to reconstruct system dynamics, Wang *et al.* [21] proposed a unique CNN to implement the data-based ACD. Later, Luo *et al.* [22] reported a critic-only $Q$-learning method to derive an optimal tracking control of input-nonaffine DT nonlinear systems with unknown models. Following the line of [20]–[22], this paper aims at presenting a single CNN to obtain an optimal neuro-control law of CT nonlinear systems with asymmetric input constraints.

System's inputs/actuators suffering from constraints are common phenomena. This is because the design of stabilizing controllers must take safety or the physical restriction of actuators into consideration. In recent years, many scholars have paid their attention to nonlinear-constrained optimization problems. For DT nonlinear systems, Zhang *et al.* [23] presented an iterative ADP to derive an optimal control of nonlinear systems subject to control constraints. To implement the iterative ADP, they employed the model NN, the CNN, and the actor NN. By using a similar architecture as [23], Ha *et al.* [24] suggested an event-triggered ACD to solve nonlinear-constrained optimization problems. The key feature distinguishing [23] and [24] is whether the optimal control was obtained in an event-triggering mechanism. For CT nonlinear systems, Abu-Khalaf and Lewis [25] first proposed an off-line policy iteration algorithm to solve an optimal

Manuscript received August 14, 2019; revised November 27, 2019; accepted January 21, 2020. This work was supported by the National Natural Science Foundation of China (61973228, 61973330). Recommended by Associate Editor Yanjun Liu. *(Corresponding author: Xiong Yang.)*

Citation: X. Yang and B. Zhao, "Optimal neuro-control strategy for nonlinear systems with asymmetric input constraints," *IEEE/CAA J. Autom. Sinica*, vol. 7, no. 2, pp. 575–583, Mar. 2020.

X. Yang is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: xiong.yang@tju.edu.cn).

B. Zhao is with the School of Systems Science, Beijing Normal University, Beijing 100875, China (email: zhaobo@bnu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2020.1003063

control problem of nonlinear systems with input constraints. To implement the policy iteration algorithm, they employed aforementioned actor-critic structure. By using the same structure, Modares *et al.* [26] reported an online policy iteration algorithm together with the experience replay technique to obtain an optimal control of nonlinear constrained-input systems with totally unavailable systems' information. After that, Zhu *et al.* [27] suggested an ADP combined with the concurrent learning technique to design an optimal event-triggered controller for nonlinear systems with input constraints as well as partially available systems' knowledge. Recently, Wang *et al.* [28] reported various ACD methods to obtain the time/event-triggered robust (optimal) control of constrained-input nonlinear systems. Later, Zhang *et al.* [29] proposed an ADP-based robust optimal control method for nonlinear constrained-input systems with unknown systems' prior information. More recently, unlike [28] and [29] studying nonlinear-constrained regulation problems, Cui *et al.* [30] solved the nonlinear-constrained optimal tracking control problem via a single network event-triggered ADP.

Though nonlinear-constrained optimization problems were successfully solved in aforementioned literature, all of them assumed that the system's input/actuator suffered from *symmetric* input constraints. Actually, in engineering industries, there exist many nonlinear plants subject to *asymmetric* input constraints [31]. Thus, one needs to develop adaptive control strategies, especially adaptive optimal neuro-control schemes for such systems. Recently, Kong *et al.* [32] proposed an asymmetric bound adaptive control for uncertain robots by using NNs and the backstepping method together. They tackled asymmetric control constraints via introducing a switching function. In general, it is challengeable to find such a switching function owing to the complexity of nonlinear systems. More recently, Zhou *et al.* [33] presented an ADP-based neuro-optimal tracking controller for continuous stirred tank reactor subject to asymmetric input constraints. They analyzed the convergence of the proposed ADP algorithm. But they did not discuss the stability of the closed-loop system. Moreover, they designed the optimal tracking controller for DT nonlinear systems, not for CT nonlinear systems. To the best of authors' knowledge, there lacks the work on designing optimal neuro-controller for CT nonlinear systems with asymmetric input constraints. This motivates our investigation.

In this study, we develop an optimal neuro-control scheme for CT nonlinear systems subject to asymmetric input constraints. First, we introduce a discounted cost function for the CT nonlinear systems in order to deal with asymmetric input constraints. Then, we present the Hamilton-Jacobi-Bellman equation (HJBE) originating from the discounted-cost optimal control problem. After that, under the framework of RL, we use a unique CNN to solve the HJBE in order to acquire the optimal neuro-controller. The CNN's weight vector is updated through the gradient descent approach. Finally, uniform ultimate boundedness (UUB) of the CNN's weight vector and the closed-loop system is proved via the Lyapunov method.

The novelties of this paper are three aspects.

1) In comparison with [25]–[30], this paper presents an optimal neuro-control strategy for CT nonlinear systems with asymmetric input constraints rather than symmetric input constraints. Thus, the present optimal control scheme is suitable for a wider range of dynamical systems, in particular, those nonlinear systems subject to asymmetric input constraints.

2) Unlike [32] handling asymmetric input constraints via proposing a switching function, this paper introduces a modified hyperbolic tangent function into the cost function to tackle such constraints (Note: here "the modified hyperbolic tangent function" means that the equilibrium point of the hyperbolic tangent function is nonzero). Thus, the present optimal control scheme can obviate the challenge arising in constructing the switching function.

3) Though both this paper and [31], [33] study optimal control problems of nonlinear systems with asymmetric input constraints, an important difference between this paper and [31], [33] is that, this paper develops an optimal neruo-control strategy for CT nonlinear systems rather than DT nonlinear systems. In general, control methods developed for DT nonlinear systems are not applicable to those CT nonlinear systems. Furthermore, in comparison with [31] and [33], this papers provided stability analyses of the closed-loop system, which guarantee the validity of the obtained optimal neuro-control policy.

*Notations:* $\mathbb{R}$, $\mathbb{R}^m$, and $\mathbb{R}^{n\times m}$ denote the set of real numbers, the Euclidean space of real *m*-vectors, and the space of $n\times m$ real matrices, respectively. $\Omega$ is a compact subset of $\mathbb{R}^n$. $I_m$ represents the $m\times m$ identity matrix. $C^1$ means the function with continuous derivative. $\|x\|$ and $\|A\|$ denote the norms of the vector $x\in\mathbb{R}^m$ and the matrix $A$, respectively. $\mathcal{A}(\Omega)$ denotes the set of admissible control on $\Omega$. $V^*(x) = \frac{\partial V^*(x)}{\partial x}$.

## II. Problem Formulation

We consider the following CT nonlinear systems

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t),\ \ x_0 = x(0) \tag{1}$$

where $x\in\Omega\subset\mathbb{R}^n$ is the state variable, $u\in\mathcal{U}\subset\mathbb{R}^m$ is the control input, $\mathcal{U} = \{(u_1,u_2,\ldots,u_m)\in\mathbb{R}^m : u_{\min}\le u_i\le u_{\max}, |u_{\min}|\ne |u_{\max}|,\ i=1,2,\ldots,m\}$ with $u_{\min}$ and $u_{\max}$ being the minimum and maximum bound of $u_i$, respectively, $f(x)\in\mathbb{R}^n$ and $g(x)\in\mathbb{R}^{n\times m}$ are known continuous functions on $\Omega$, and $x_0\in\mathbb{R}^n$ is the initial state.

*Remark 1:* Generally speaking, the knowledge of system dynamics is not necessary to be known when one applies RL to design neuro-controllers for nonlinear systems, such as [34] and [35]. Here we need the prior information of system (1) (i.e., $f(x)$ and $g(x)$). This is because the neuro-controller will be designed only using a unique critic NN rather than the typical actor-critic dual NNs.

*Assumption 1:* $f(0) = 0$, i.e., $x = 0$ is the equilibrium point of system (1) if letting $u = 0$ or $g(0) = 0$. In addition, $f(x) + g(x)u$ satisfies the Lipschitz condition guaranteeing that $x = 0$ is the unique equilibrium point on $\Omega$ (Note: $0\in\Omega$).

*Assumption 2:* For every $x\in\Omega$, $\|g(x)\|\le g_M$ with $g_M > 0$ the known constant. Moreover, $g(0) = 0$.

Considering that system (1) suffers from asymmetric input constraints, we propose a discounted cost function as follows

$$V^u(x(t)) = \int_t^{+\infty} e^{-\alpha(\tau - t)} \mathcal{S}(x(\tau), u(\tau)) d\tau \tag{2}$$

where $\alpha > 0$ is the discount factor, $\mathcal{S}(x,u) = x^T Qx + \mathcal{R}(u)$, $Q \in \mathbb{R}^{n\times n}$ is a positive-definite constant matrix, and $\mathcal{R}(u) \in \mathbb{R}$ is defined as

$$\mathcal{R}(u) = 2\beta \sum_{i=1}^{m} \int_b^{u_i} \psi^{-1}\Big(\beta^{-1}(s_i - b)\Big) ds_i \tag{3}$$

where

$$\beta = \frac{u_{\max} - u_{\min}}{2}, \quad b = \frac{u_{\max} + u_{\min}}{2} \tag{4}$$

and $\psi^{-1}(\cdot) \in C^1(\Omega)$ is an odd monotonic function satisfying $\psi^{-1}(0) = 0$. Observing the characteristic of $\psi^{-1}(\cdot)$, we choose $\psi^{-1}(\cdot) = \tanh^{-1}(\cdot)$, where $\tanh(\cdot)$ is the hyperbolic tangent function.

*Remark 2:* Two notes are provided to make (2) and (3) better for understanding, i.e.,

a) Even if $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ is a symmetric function, $\mathcal{R}(u)$ in (3) still gives rise to asymmetric constraints on the input (see $u^*(x)$ in later (8)). This is because $b \neq 0$ in (4) (Note: $|u_{\min}| \neq |u_{\max}|$ implies that $b \neq 0$). This feature of $\mathcal{R}(u)$ differs from [25]–[30], which studied symmetric input constraints.

b) Owing to the asymmetric input constraints handled via (3), the optimal control will not converge to zero when the steady states are obtained (Note: According to $u^*(x)$ in later (8), we can find $u^*(0) \neq 0$. Moreover, simulation results also verify this conclusion). Therefore, if letting $\alpha = 0$ (i.e., no the decay term $e^{-\alpha(\tau-t)}$), then $V^u(x(t))$ might be unbounded. That is why we introduce the discounted cost function (2).

The optimum of $V^u(x)$, denoted by $V^*(x)$, is defined as

$$V^*(x) = \min_{u \in \mathcal{A}(\Omega)} V^u(x). \tag{5}$$

As pointed out by [4], $V^*(x)$ is the solution of the HJBE

$$\min_{u \in \mathcal{A}(\Omega)} H(x, \nabla V^*(x), u) = 0 \tag{6}$$

with $V^*(0) = 0$ and $H(x, \nabla V^*(x), u)$ given as

$$\begin{aligned} H(x, \nabla V^*(x), u) &= (\nabla V^*(x))^T (f(x) + g(x)u) \\ &\quad - \alpha V^*(x) + x^T Qx + \mathcal{R}(u) \end{aligned} \tag{7}$$

where $H(x, \nabla V^*(x), u)$ in (7) is called the Hamiltonian [4].

Applying the stationary condition [36, Theorem 5.8] to (7) (that is, $\frac{\partial H(x, \nabla V^*(x), u^*)}{\partial u^*} = 0$), we have the optimal control formulated as

$$\begin{aligned} u^*(x) &= \arg \min_{u \in \mathcal{A}(\Omega)} H(x, \nabla V^*(x), u) \\ &= -\beta \tanh\left(\frac{1}{2\beta} g^T(x) \nabla V^*(x)\right) + \ell_b \end{aligned} \tag{8}$$

where

$$\ell_b = [b, b, \ldots, b]^T \in \mathbb{R}^m$$

with $b$ being defined as (4).

Inserting (8) into (6), we are able to rewrite the HJBE as

$$\begin{aligned} &(\nabla V^*(x))^T f(x) - \alpha V^*(x) + x^T Qx + (\nabla V^*(x))^T g(x)\ell_b \\ &+ \mathcal{R}\left(-\beta \tanh\Big(\frac{1}{2\beta} g^T(x) \nabla V^*(x)\Big) + \ell_b\right) \\ &- \beta(\nabla V^*(x))^T g(x) \tanh\left(\frac{1}{2\beta} g^T(x) \nabla V^*(x)\right) = 0 \end{aligned} \tag{9}$$

with $V^*(0) = 0$. The expression (9) indicates that it is in essence a nonlinear equation with respect to $V^*(x)$. As emphasized by [1] and [4], there often does not exist an analytical method to solve such a nonlinear equation like (9). In this paper, we are devoted to presenting a CNN to approximately solve (9) under the framework of RL.

## III. Optimal Neuro-control Strategy

The approximation characteristic of NNs indicated in [37] guarantees that $V^*(x)$ in (5) can be restated on $\Omega$ in the form

$$V^*(x) = \omega_c^T \sigma_c(x) + \varepsilon_c(x) \tag{10}$$

where $\omega_c \in \mathbb{R}^{\tilde{n}_c}$ is the ideal weight vector often unavailable, $\tilde{n}_c$ denotes the number of neurons used in the NN, $\sigma_c(x) \in \mathbb{R}^{\tilde{n}_c}$ is the vector activation function comprised of $\tilde{n}_c$ linearly independent elements $\sigma_{c1}(x), \sigma_{c2}(x), \ldots, \sigma_{c\tilde{n}_c}(x)$ (Note: $\sigma_{ci}(x) \in \mathbb{R}$ and $\sigma_{ci}(0) = 0$ , $i = 1, 2, \ldots, \tilde{n}_c$), and $\varepsilon_c(x)$ is the error originating from reconstructing $V^*(x)$.

Then, we obtain from (10) that

$$\nabla V^*(x) = \nabla \sigma_c^T(x) \omega_c + \nabla \varepsilon_c(x) \tag{11}$$

where $\nabla C(x) = \frac{\partial C(x)}{\partial x}$ with $\nabla C(\cdot) = \nabla \sigma_c(\cdot)$ or $\nabla \varepsilon_c(\cdot)$.

Inserting (11) into (8), it follows:

$$u^*(x) = -\beta \tanh(\mathcal{D}_1(x)) + \varepsilon_{u^*}(x) + \ell_b \tag{12}$$

where

$$\mathcal{D}_1(x) = \frac{1}{2\beta} g^T(x) \nabla \sigma_c^T(x) \omega_c$$

$$\varepsilon_{u^*}(x) = -\frac{1}{2}(I_m - C(\nu(x))) g^T(x) \nabla \varepsilon_c(x)$$

with $C(\nu(x)) = \operatorname{diag}\{\tanh^2(\nu_i(x))\}$, $i = 1, 2, \ldots, m$, $\nu(x) = [\nu_1(x), \nu_2(x), \ldots, \nu_m(x)]^T \in \mathbb{R}^m$, and $\nu(x) \in \mathbb{R}^m$ be selected between $(1/(2\beta)) g^T(x) \nabla V^*(x)$ and $\mathcal{D}_1(x)$.

*Remark 3:* To make (12) easy for understanding, we present the detailed procedure as follows. Let

$$\mathcal{T}(\mathcal{D}_k(x)) = -\beta \tanh(\mathcal{D}_k(x)), \ k = 0, 1$$

where

$$\mathcal{D}_0(x) = \frac{1}{2\beta} g^T(x) \nabla V^*(x)$$

and $\mathcal{D}_1(x)$ is given in (12).

Then, using the mean value theorem [36, Theorem 5.10], we find

$$\begin{aligned} \mathcal{T}(\mathcal{D}_0(x)) - \mathcal{T}(\mathcal{D}_1(x)) &= -\beta\big(\tanh(\mathcal{D}_0(x)) - \tanh(\mathcal{D}_1(x))\big) \\ &= -\frac{1}{2}(I_m - C(\nu(x))) g^T(x) \nabla \varepsilon_c(x) \end{aligned}$$

where $C(\nu(x)) = \operatorname{diag}\{\tanh^2(\nu_i(x))\}$, $i = 1, 2, \ldots, m$, and $\nu(x) \in \mathbb{R}^m$ is chosen between $\mathcal{D}_0(x)$ and $\mathcal{D}_1(x)$. Thus, we have

$$\begin{aligned}u^*(x) &= -\beta\tanh(\mathcal{D}_0(x))+\ell_b\\ &= \mathcal{T}(\mathcal{D}_1(x))+(\mathcal{T}(\mathcal{D}_0(x))-\mathcal{T}(\mathcal{D}_1(x)))+\ell_b\\ &= -\beta\tanh(\mathcal{D}_1(x))\\ &\quad -\frac{1}{2}(I_m-C(\nu(x)))g^T(x)\nabla\varepsilon_c(x)+\ell_b.\end{aligned}$$

This verifies that (12) holds.

As previously stated, $\omega_c$ in (10) is generally unavailable. Thus, $u^*(x)$ in (12) cannot be implemented in control process. To tackle this issue, we replace $\omega_c$ with its estimated value $\hat{\omega}_c$. Then, the estimated value of $V^*(x)$, denoted by $\hat{V}(x)$, can be expressed as an output of CNN, that is

$$\hat{V}(x)=\hat{\omega}_c^T\sigma_c(x). \tag{13}$$

So, the estimated control policy of $u^*(x)$ can be expressed as

$$\hat{u}(x)=-\beta\tanh(\mathcal{D}_2(x))+\ell_b \tag{14}$$

where

$$\mathcal{D}_2(x)=\frac{1}{2\beta}g^T(x)\nabla\sigma_c^T(x)\hat{\omega}_c.$$

Using $\hat{V}(x)$ in (13) and $\hat{u}(x)$ in (14) to replace $\nabla V^*(x)$ and $u$ in (7), we have the approximate Hamiltonian formulated as

$$\hat{H}(x,\nabla\hat{V}(x),\hat{u}(x))=\hat{\omega}_c^T\phi+x^TQx+\mathcal{R}(\hat{u}(x))$$

where

$$\phi=\nabla\sigma_c(x)(f(x)+g(x)\hat{u}(x))-\alpha\sigma_c(x).$$

Then, we can describe the error between $H(x,\nabla V^*(x),u^*)$ and $\hat{H}(x,\nabla\hat{V}(x),\hat{u}(x))$ as (Note: $H(x,\nabla V^*(x),u^*)=0$)

$$\begin{aligned}e_c &= \hat{H}(x,\nabla\hat{V}(x),\hat{u}(x))-H(x,\nabla V^*(x),u^*(x))\\ &= \hat{\omega}_c^T\phi+x^TQx+\mathcal{R}(\hat{u}(x)).\end{aligned} \tag{15}$$

To make $\hat{\omega}_c\rightarrow\omega_c$, we resort to forcing $e_c\rightarrow 0$. To achieve this goal, we choose the target function to be $E=(1/2)e_c^Te_c/(1+\phi^T\phi)^2$ and let the gradient descent method be applied to $E$. Then, we have the tuning rule for the CNN's weight vector formulated as

$$\dot{\hat{\omega}}_c=-\frac{\gamma}{(1+\phi^T\phi)^2}\frac{\partial E}{\partial\hat{\omega}_c}=-\frac{\gamma\phi}{(1+\phi^T\phi)^2}e_c \tag{16}$$

with $e_c$ being defined as (15) and $\gamma>0$ being the adjustable parameter. Letting $\tilde{\omega}_c=\omega_c-\hat{\omega}_c$ be the CNN's weight estimation error, we get the error dynamics of $\tilde{\omega}_c$ as

$$\dot{\tilde{\omega}}_c=-\gamma\varphi\varphi^T\tilde{\omega}_c+\frac{\gamma\varphi}{1+\phi^T\phi}\varepsilon_H \tag{17}$$

with $\varphi=\phi/(1+\phi^T\phi)$ and $\varepsilon_H=-\nabla\varepsilon_c^T(x)(f(x)+g(x)\hat{u}(x))+\alpha\varepsilon_c(x)$ (Note: Since $\varepsilon_H$ can be obtained via a similar procedure as shown in [38], we omit the process here).

To summarize aforementioned descriptions of the proposed optimal control scheme, we present a block diagram in Fig. 1.

## IV. Stability Analysis

Before proceeding further, we give two indispensable assumptions, which were employed in [38] and [39].

*Assumption 3:* For all $x\in\Omega$, there are $\|\nabla\varepsilon_c(x)\|\le b_{\varepsilon_c}$, $\|\nabla\sigma_c(x)\|\le b_{\sigma_c}$, and $\|\varepsilon_H\|\le b_{\varepsilon_H}$, where $b_{\varepsilon_c}$, $b_{\sigma_c}$, and $b_{\varepsilon_H}$ are positive constants.

*Assumption 4:* $\varphi$ in (17) satisfies the persistence of excitation (PE) condition. Specifically, we have constants $0<\rho_1<\rho_2$ and $T_0>0$ such that, for arbitrary time $t$

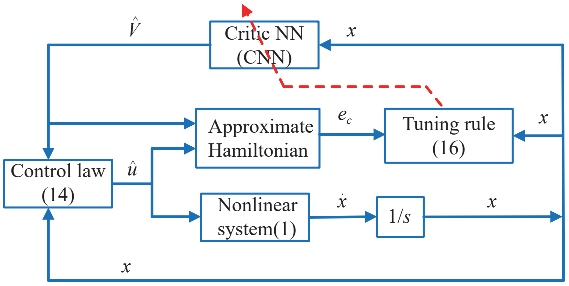

Fig. 1. Block diagram of the present optimal control scheme.

$$\rho_1 I_{\tilde{n}_c}\le\int_t^{t+T_0}\varphi(s)\varphi^T(s)ds\le\rho_2 I_{\tilde{n}_c}. \tag{18}$$

*Theorem 1:* Consider system (1) with the related control (14). Given that Assumptions 1–4 hold and the update rule for CNN's weight vector is described as (16). Meanwhile, let the initial control for system (1) be admissible. Then, UUB of all signals in the closed-loop system is guaranteed.

*Proof:* Let the Lyapunov function candidate be

$$\mathcal{L}(t)=V^*(x(t))+\frac{1}{2}\tilde{\omega}_c^T\tilde{\omega}_c. \tag{19}$$

Considering the derivative of $V^*(x(t))$ (i.e., $\frac{dV^*(x(t))}{dt}$) along the solution of $\dot{x}=f(x)+g(x)\hat{u}$, we have

$$\begin{aligned}\dot{V}^*(x(t)) &= (\nabla V^*(x))^T(f(x)+g(x)\hat{\mu}(x))\\ &= (\nabla V^*(x))^T(f(x)+g(x)u^*(x))\\ &\quad +(\nabla V^*(x))^Tg(x)(\hat{\mu}(x)-u^*(x)).\end{aligned} \tag{20}$$

According to (6)–(8), there holds

$$\begin{cases}(\nabla V^*(x))^T(f(x)+g(x)u^*(x))\\ \quad =-x^TQx-\mathcal{R}(u^*(x))+\alpha V^*(x)\\ (\nabla V^*(x))^Tg(x)=-2\beta\left(\tanh^{-1}\left(\dfrac{u^*(x)-\ell_b}{\beta}\right)\right)^T.\end{cases}$$

Then, after performing calculations, (20) can be restated as

$$\dot{V}^*(x)=-x^TQx-\mathcal{R}(u^*(x))+\alpha V^*(x)+\pi_1 \tag{21}$$

with

$$\pi_1=-2\beta\left(\tanh^{-1}\left(\frac{u^*(x)-\ell_b}{\beta}\right)\right)^T(\hat{u}(x)-u^*(x)).$$

Note that, for vectors $c$ and $d$ with suitable dimensions

$$2c^Td\le c^Tc+d^Td \quad\text{or}\quad 2c^Td\le\|c\|^2+\|d\|^2$$

holds. Then, using (12) and (14), we have $\pi_1$ in (21) satisfied

$$\begin{aligned}\pi_1 &\le\beta^2\left\|\tanh^{-1}\left(\frac{u^*(x)-\ell_b}{\beta}\right)\right\|^2+\left\|\hat{u}(x)-u^*(x)\right\|^2\\ &\le\beta^2\sum_{i=1}^m\left(\tanh^{-1}\left(\frac{u_i^*(x)-b}{\beta}\right)\right)^2\\ &\quad+\underbrace{\left\|\beta(\tanh(\mathcal{D}_1(x))-\tanh(\mathcal{D}_2(x)))-\varepsilon_{u^*}(x)\right\|^2}_{\pi_2}.\end{aligned} \tag{22}$$

Likewise, $\|c+d\|^2\le 2\|c\|^2+2\|d\|^2$ holds for vectors $c$ and $d$

with proper dimensions (Note: actually, it is a kind of Young's inequalities [40]). Thus, using the facts that $\|\tanh(\mathcal{D}_k(x))\| \le \sqrt{m}$ $(k=1,2)$ and $\|I_m - \operatorname{diag}\{\tanh^2(\nu_i(x))\}_{i=1}^{m}\| \le 2$ [41, Lemma 1] as well as Assumptions 2 and 3, we have $\pi_2$ in (22) satisfied

$$\begin{aligned}\pi_2 &\le 2\beta^2\|\tanh(\mathcal{D}_1(x)) - \tanh(\mathcal{D}_2(x))\|^2 + 2\|\varepsilon_{u^*}(x)\|^2\\ &\le 4\beta^2\left(\|\tanh(\mathcal{D}_1(x))\|^2 + \|\tanh(\mathcal{D}_2(x))\|^2\right)\\ &\quad + 2\left\|\frac{1}{2}(I_m - C(\nu(x)))g^T(x)\nabla\varepsilon_c(x)\right\|^2\\ &\le 8\beta^2 m + 2b_{\varepsilon_c}^2 g_M^2. \end{aligned} \tag{23}$$

On the other hand, letting $\xi_i = s_i - b$ in (3) and using $u^*$ in (12) to replace $u$, we can write $\mathcal{R}(u^*(x))$ as (Note: for brevity, $\mathcal{R}(u^*(x))$ is denoted by $\mathcal{R}(u^*)$)

$$\mathcal{R}(u^*) = 2\beta\sum_{i=1}^{m}\int_0^{u_i^*(x)-b}\tanh^{-1}\left(\frac{\xi_i}{\beta}\right)d\xi_i. \tag{24}$$

Similar to the proof of [42, Theorem 1], after performing some calculations, we can restate $\mathcal{R}(u^*)$ in (24) as

$$\begin{aligned}\mathcal{R}(u^*) = &-2\beta^2\sum_{i=1}^{m}\int_0^{\tanh^{-1}\left(\frac{u_i^*(x)-b}{\beta}\right)}\theta_i\tanh^2(\theta_i)d\theta_i\\ &+\beta^2\sum_{i=1}^{m}\left(\tanh^{-1}\left(\frac{u_i^*(x)-b}{\beta}\right)\right)^2. \end{aligned}\tag{25}$$

Thus, combining (22), (23), and (25), we find that (21) yields

$$\begin{aligned}\dot{V}^*(x) \le &-\lambda_{\min}(Q)\|x\|^2\\ &+\underbrace{2\beta^2\sum_{i=1}^{m}\int_0^{\tanh^{-1}\left(\frac{u_i^*(x)-b}{\beta}\right)}\theta_i\tanh^2(\theta_i)d\theta_i}_{\pounds(x)}\\ &+\alpha V^*(x) + 8\beta^2 m + 2b_{\varepsilon_c}^2 g_M^2\end{aligned}\tag{26}$$

with $\lambda_{\min}(Q)$ being the minimum eigenvalue of $Q$ in (2). From the process of demonstrating [42, Theorem 1], we can have the conclusion that $\pounds(x)$ in (26) is bounded. So, we write $\|\pounds(x)\| \le \delta_M$ with $\delta_M > 0$ the known constant. Note that $V^*(x)$ is associated with the admissible control $u^*(x)$ in (8). Thus, according to the definition of admissible control [25, Definition 1], $V^*(x)$ is bounded. We denote $\left\|V^*(x)\right\| \le b_{V^*}$ with $b_{V^*} > 0$ the known constant. Then, from (26), we have

$$\dot{V}^*(x) \le -\lambda_{\min}(Q)\|x\|^2 + c_0 \tag{27}$$

where

$$c_0 = 8\beta^2 m + 2b_{\varepsilon_c}^2 g_M^2 + \alpha b_{V^*} + \delta_M. \tag{28}$$

Second, we consider the time derivative of $(1/2)\tilde{\omega}_c^T\tilde{\omega}_c$. Using (17), we can see that $\frac{d(1/2)\tilde{\omega}_c^T\tilde{\omega}_c}{dt}$ becomes

$$\frac{d\left(\frac{1}{2}\tilde{\omega}_c^T\tilde{\omega}_c\right)}{dt} = -\gamma\tilde{\omega}_c^T\varphi\varphi^T\tilde{\omega}_c + \frac{\gamma}{1+\phi^T\phi}\tilde{\omega}_c^T\varphi\varepsilon_H. \tag{29}$$

Based on aforementioned inequality $2c^Td \le c^Tc + d^Td$ and the fact that $1/(1+\phi^T\phi) \le 1$, we get

$$\frac{\gamma}{1+\phi^T\phi}\tilde{\omega}_c^T\varphi\varepsilon_H \le \frac{\gamma}{2}\tilde{\omega}_c^T\varphi\varphi^T\tilde{\omega}_c + \frac{\gamma}{2}\varepsilon_H^T\varepsilon_H. \tag{30}$$

Then, using Assumptions 3 and 4 as well as (30), we can further write (29) as

$$\begin{aligned}\frac{d\left(\frac{1}{2}\tilde{\omega}_c^T\tilde{\omega}_c\right)}{dt} &\le -\frac{\gamma}{2}\tilde{\omega}_c^T\varphi\varphi^T\tilde{\omega}_c + \frac{\gamma}{2}\varepsilon_H^T\varepsilon_H\\ &\le -\frac{\gamma}{2}\lambda_{\min}\left(\varphi\varphi^T\right)\|\tilde{\omega}_c\|^2 + \frac{\gamma}{2}b_{\varepsilon_H}^2\end{aligned}\tag{31}$$

with $\lambda_{\min}(\varphi\varphi^T)$ being the minimum eigenvalue of $\varphi\varphi^T$.

Combining (27) and (31), it can be observed that $\mathcal{L}(t)$ in (19) satisfies

$$\dot{\mathcal{L}}(t) \le -\lambda_{\min}(Q)\|x\|^2 - \frac{\gamma}{2}\lambda_{\min}\left(\varphi\varphi^T\right)\|\tilde{\omega}_c\|^2 + \frac{\gamma}{2}b_{\varepsilon_H}^2 + c_0$$

with $c_0$ given as (28). Thus, $\dot{\mathcal{L}}(t) < 0$ holds only if we can guarantee either $x \notin \Omega_x$ or $\tilde{\omega}_c \notin \Omega_{\tilde{\omega}_c}$ with $\Omega_x$ and $\Omega_{\tilde{\omega}_c}$ given as follows:

$$\Omega_x = \left\{x\colon \|x\| \le \sqrt{\frac{\gamma b_{\varepsilon_H}^2 + 2c_0}{2\lambda_{\min}(Q)}} = r_1\right\} \tag{32}$$

$$\Omega_{\tilde{\omega}_c} = \left\{\tilde{\omega}_c\colon \|\tilde{\omega}_c\| \le \sqrt{\frac{\gamma b_{\varepsilon_H}^2 + 2c_0}{\gamma\lambda_{\min}(\varphi\varphi^T)}} = r_2\right\}. \tag{33}$$

This demonstrates UUB of $x$ and $\tilde{\omega}_c$. Their ultimate bounds are given as $r_1$ in (32) and $r_2$ in (33), respectively. Furthermore, noticing that the ideal weight vector $\omega_c$ is typically bounded [1] and $\tilde{\omega}_c = \omega_c - \hat{\omega}_c$, we thus conclude that $\hat{\omega}_c$ is stable in the sense of UUB. ■

*Remark 4:* The key to making the inequality (31) valid lies in that there is $\lambda_{\min}(\varphi\varphi^T) > 0$. Obviously, (18) guarantees that $\lambda_{\min}(\varphi\varphi^T) > 0$ holds. That is why we need $\varphi$ to satisfy the PE condition in Assumption 4.

*Theorem 2:* With the same condition as Theorem 1, the estimated control policy $\hat{u}(x)$ in (14) can converge to $u^*(x)$ in (12) within an adjustable bound.

*Proof:* According to (12) and (14) and using the mean value theorem [36, Theorem 5.10], it follows

$$\begin{aligned}\hat{u}(x) - u^*(x) &= \beta(\tanh(\mathcal{D}_1(x)) - \tanh(\mathcal{D}_2(x))) - \varepsilon_{u^*}(x)\\ &= \frac{1}{2}(I_m - C(\vartheta(x)))g^T(x)\nabla\sigma_c^T(x)\tilde{\omega}_c - \varepsilon_{u^*}(x)\end{aligned}\tag{34}$$

where

$$C(\vartheta(x)) = \operatorname{diag}\{\tanh^2(\vartheta_i(x))\}, \quad i = 1,2,\ldots,m$$

with $\vartheta(x) = [\vartheta_1(x), \vartheta_2(x), \ldots, \vartheta_m(x)]^T \in \mathbb{R}^m$ being selected between $\mathcal{D}_1(x)$ and $\mathcal{D}_2(x)$.

Noticing that $\|I_m - C(\vartheta(x))\| \le 2$ and using Assumptions 2 and 3, we can derive from (34) that

$$\left\|\hat{u}(x) - u^*(x)\right\| \le (b_{\sigma_c}\|\tilde{\omega}_c\| + b_{\varepsilon_c})g_M. \tag{35}$$

According to Theorem 1, the ultimate bound of $\tilde{\omega}_c$ is $r_2$ in (33). Hence, from (35), we have

$$\left\|\hat{u}(x) - u^*(x)\right\| \le (b_{\sigma_c}r_2 + b_{\varepsilon_c})g_M.$$

This proves that $\hat{u}(x)$ converges to $u^*(x)$ within the bound $(b_{\sigma_c} r_2 + b_{\varepsilon_c}) g_M$. Here, $r_2$ and $b_{\varepsilon_c}$ are in essence the bounds connected with $\nabla\varepsilon_c(x)$. As stated in [37] and [38], one has $\varepsilon_c(x) \to 0$ and $\nabla\varepsilon_c(x) \to 0$ when letting $\tilde{n}_c \to \infty$ in (10). Therefore, both $r_2$ and $b_{\varepsilon_c}$ are adjustable. Or rather, the bound $(b_{\sigma_c} r_2 + b_{\varepsilon_c}) g_M$ is adjustable and made small. ■

## V. Simulation Results

To test the effectiveness of established theoretical results, we perform simulations of two examples in this section.

### A. Example 1

We study the plant described by

$$\dot{x} = \begin{bmatrix} -0.5x_1 + x_2 \\ -2x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ -x_1 \end{bmatrix} u \tag{36}$$

where $x = [x_1, x_2]^T$ with $x_0 = [1, -0.5]^T$, and $u \in \mathcal{U} = \{u \in \mathbb{R} : -1 \le u \le 3\}$ (i.e., $u_{\min} = -1$ and $u_{\max} = 3$). Letting $Q = I_2$ and $\alpha = 0.5$ in (2), we have the discounted cost function for system (36) formulated as

$$V^u(x(t)) = \int_t^{+\infty} e^{-0.5(\tau - t)} \left( \|x(\tau)\|^2 + \mathcal{R}(u(\tau)) \right) d\tau \tag{37}$$

where (Note: according to (4), $\beta = 2$ and $b = 1$)

$$\begin{aligned} \mathcal{R}(u(x)) &= 2\int_b^{u(x)} \beta \tanh^{-1}\left(\frac{\tau - b}{\beta}\right) d\tau \\ &= 2\beta(u(x) - 1)\tanh^{-1}\left(\frac{u(x) - 1}{\beta}\right) \\ &\quad + \beta^2 \ln\left(1 - \frac{(u(x) - 1)^2}{\beta^2}\right). \end{aligned} \tag{38}$$

*Remark 5:* In this example, we determine the value of the discount factor $\alpha$ via experiment studies. In fact, there is no general method to determine the accurate range of $\alpha$. We find that selecting $\alpha = 0.5$ in this example can lead to satisfactory results.

To approximate (37), we use the CNN described as (13) with $\tilde{n}_c = 3$. Meanwhile, we choose the vector activation function as $\sigma_c(x) = [x_1^2, x_1 x_2, x_2^2]^T$, and denote its associated weight vector as $\hat{\omega}_c = [\hat{\omega}_{c1}, \hat{\omega}_{c2}, \hat{\omega}_{c3}]^T$. The initial weight vector is set as $\hat{\omega}_c^{\text{initial}} = [1.5040, 0.5259, 1.012]^T$ in order to guarantee the initial control policy for system (36) to be admissible (Note: according to (14), the initial control is associated with the initial weight vector. Thus, we can choose an appropriate initial weight vector to make the initial control admissible). The parameter used in (16) is $\gamma = 0.9$. Meanwhile, an exponential decay signal is added to system's input to guarantee $\varphi$ in (17) to be persistently exciting.

By performing simulations via the MATLAB (2017a) software package, we obtain Figs. 2–4. As displayed in Fig. 2, $\hat{\omega}_c = [\hat{\omega}_{c1}, \hat{\omega}_{c2}, \hat{\omega}_{c3}]^T$ is convergent after the first 6 s. Here we denote the converged value of $\hat{\omega}_c$ as $\hat{\omega}_c^{\text{final}}$. Then, from Fig. 2, we find $\hat{\omega}_c^{\text{final}} = [1.0963, 0.6887, 0.3783]^T$. The evolution of system states $x_1(t)$ and $x_2(t)$ is shown in Fig. 3. Meanwhile, the control policy $\hat{u}(x)$ is illustrated in Fig. 4. It can be observed from Figs. 3 and 4 that the system states converge to the equilibrium point (i.e., $x = 0$) while the control policy converges to a nonzero value (i.e., $\hat{u}^{\text{final}} = 1$). This feature is in accordance with the analyses provided in Remark 2-b). In addition, Fig. 4 indicates that the asymmetric control constraints are overcome.

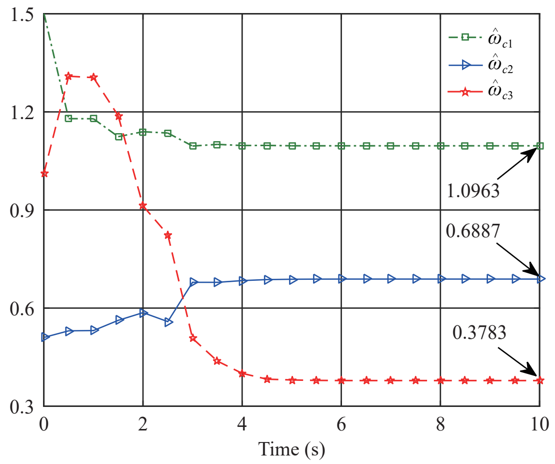

Fig. 2. Performance of $\hat{\omega}_c = [\hat{\omega}_{c1}, \hat{\omega}_{c2}, \hat{\omega}_{c3}]^T$.

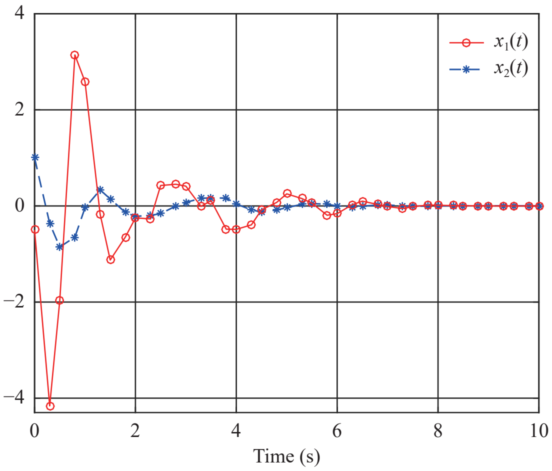

Fig. 3. System states $x_1(t)$ and $x_2(t)$ in Example 1.

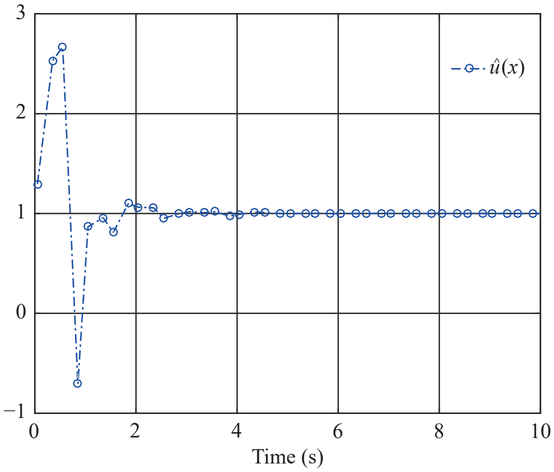

Fig. 4. Control $\hat{u}(x)$ in Example 1.

### B. Example 2

We investigate the nonlinear system given as

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ -0.5(x_1 + x_2) + 0.5x_2 \sin^2(x_1) \end{bmatrix} + \begin{bmatrix} 0 \\ \sin(x_1) \end{bmatrix} u \qquad (39)$$

where $x = [x_1, x_2]^T$ with $x_0 = [1, -1]^T$, and $u \in \mathcal{U} = \{u \in \mathbb{R} : -3 \le u \le 4\}$ (i.e., $u_{\min} = -3$ and $u_{\max} = 4$). The discounted cost function for system (39) is similar to (37). A slight difference is that, according to (4), in this example we have $\beta = 3.5$ and $b = 0.5$.

The CNN described as (13) (Note: $\tilde{n}_c = 8$) is applied to approximate (37). The vector activation function used in (13) is $\sigma_c(x) = [x_1^2, x_2^2, x_1x_2, x_1^4, x_2^4, x_1^3x_2, x_1^2x_2^2, x_1x_2^3]^T$, and its associated weight vector is written as $\hat{\omega}_c = [\hat{\omega}_{c1}, \hat{\omega}_{c2}, \ldots, \hat{\omega}_{c8}]^T$. Similar to Example 1, we choose the initial weight vector for the CNN as $\hat{\omega}_c^{\text{initial}} = [0.075, 1.3466, 0.859, 0.9035, 1.2197, 0.1188, 0.6316, 1.545]^T$, which guarantees the initial control policy for system (39) to be admissible. In addition, we set $\gamma = 0.6$ in (16) and add an exponential decay signal to system's input to ensure $\varphi$ in (17) to meet the PE condition.

We perform simulations via the MATLAB (2017a) software package and then obtain Figs. 5–7. Fig. 5 shows that the CNN's weight vector $\hat{\omega}_c = [\hat{\omega}_{c1}, \hat{\omega}_{c2}, \ldots, \hat{\omega}_{c8}]^T$ converges to $\hat{\omega}_c^{\text{final}} = [0.3, 0.4432, 0.111, 0.582, 0.3317, -0.3627, 0.2622, 0.6552]^T$ after the first 24 s. Figs. 6 and 7 present the evolution of system states $x_1(t)$ and $x_2(t)$ and the control policy $\hat{u}(x)$, respectively. We can see from Figs. 6 and 7 that the system states converge to the equilibrium point (i.e., $x = 0$) while the control policy converges to a nonzero value (i.e., $\hat{u}^{\text{final}} = 0.5$). This verifies the analyses provided in Remark 2-b). Moreover, Fig. 7 indicates that the asymmetric control constraints are conquered.

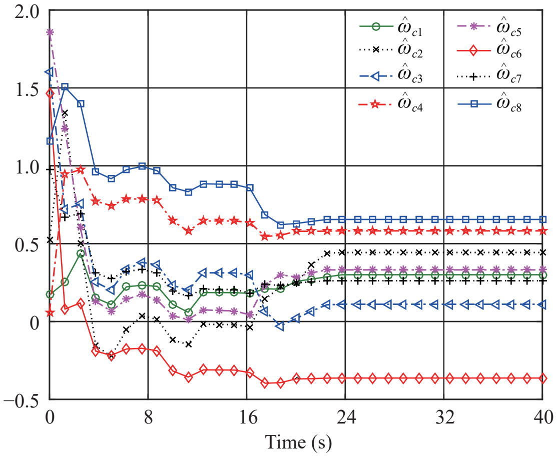

Fig. 5. Performance of $\hat{\omega}_c = [\hat{\omega}_{c1}, \hat{\omega}_{c2}, \ldots, \hat{\omega}_{c8}]^T$.

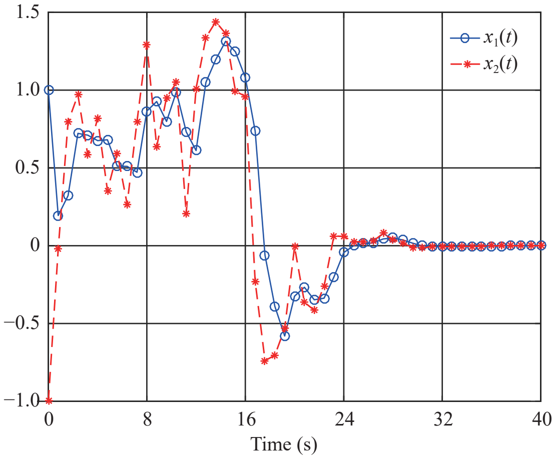

Fig. 6. System states $x_1(t)$ and $x_2(t)$ in Example 2.

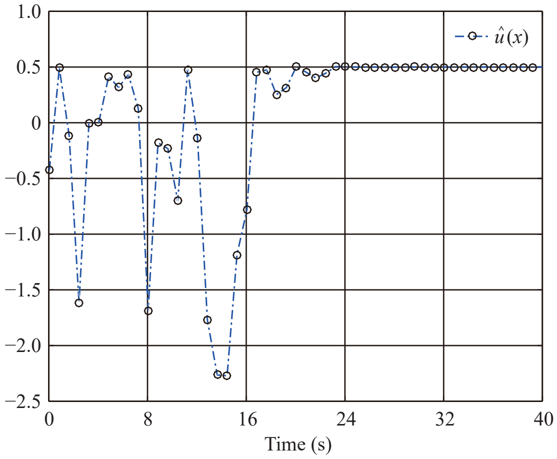

Fig. 7. Control $\hat{u}(x)$ in Example 2.

## VI. Conclusion

An optimal neuro-control scheme has been proposed for CT nonlinear systems with asymmetric input bounds. To implement such a neuro-control strategy, only a CNN is employed, which enjoys a simpler implementation structure compared with the actor-critic structure. However, the PE condition is needed to implement the present neuro-optimal control scheme. Indeed, the PE condition is a strict limitation because of it difficult to verify. Recently, the experience replay technique was introduced to relax the PE condition [43], [44]. In our consecutive work, we shall work on combining RL with the experience replay technique to obtain optimal control policies for nonlinear systems.

On the other hand, it is worth emphasizing here that the steady states generally do not stay at zero, when the optimal control policy does not converge to zero. That is why we need the control matrix in system (1) to satisfy $g(0) = 0$ (see Assumption 2). Thus, this assumption excludes those nonlinear systems with the control matrix $g(0) \neq 0$. To remove this restriction, a promising way is to allow the equilibrium point to be nonzero. Accordingly, our future work also aims at developing optimal nuero-control laws for nonlinear systems with nonzero equilibrium points. More recently, ACDs have been introduced to derive the optimal tracking control policy and the optimal fault-tolerant control policy for DT nonlinear systems, respectively [45], [46]. Therefore, whether the present optimal neuro-control strategy can be extended to solve the nonlinear optimal tracking control problems or the nonlinear optimal fault-tolerant control problems is another issue to be addressed in our consecutive study.

## REFERENCES

[1] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. London: IET, 2013.

[2] X. Yang and H. B. He, "Self-learning robust optimal control for continuoustime nonlinear systems with mismatched disturbances," *Neural Networks*, vol. 99, pp. 19–30, 2018.

[3] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2007.

[4] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive Dynamic Programming with Applications in Optimal Control*. Cham, Switzerland: Springer, 2017.

[5] X. N. Zhong, Z. Ni, and H. B. He, "Gr-GDHP: a new architecture for globalized dual heuristic dynamic programming," *IEEE Trans. Cybernetics*, vol. 47, no. 10, pp. 3318–3330, Oct. 2017.

[6] D. Wang and X. N. Zhong, "Advanced policy learning near-optimal regulation," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 743–749, May 2019.

[7] Q. L. Wei, D. R. Liu, Y. Liu, and R. Z. Song, "Optimal constrained self-learning battery sequential management in microgrid via adaptive dynamic programming," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 2, pp. 168–176, Apr. 2017.

[8] L. Dong, X. N. Zhong, C. Y. Sun, and H. B. He, "Event-triggered adaptive dynamic programming for continuous-time systems with control constraints," *IEEE Trans. Neural Networks and Learning Systems*, vol. 28, no. 8, pp. 1941–1952, Aug. 2017.

[9] B. Zhao and D. R. Liu, "Event-triggered decentralized tracking control of modular reconfigurable robots through adaptive dynamic programming," *IEEE Trans. Industrial Electronics*, vol. 67, no. 4, pp. 3054–3064, Apr. 2020.

[10] Y. Jiang and Z.-P. Jiang, *Robust Adaptive Dynamic Programming*. Hoboken, New Jersey: John Wiley & Sons, 2017.

[11] R. Z. Song, F. L. Lewis, and Q. L. Wei, "Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzerosum games," *IEEE Trans. Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 704–713, Mar. 2017.

[12] H. G. Zhang, K. Zhang, Y. L. Cai, and J. Han, "Adaptive fuzzy fault-tolerant tracking control for partially unknown systems with actuator faults via integral reinforcement learning method," *IEEE Trans. Fuzzy Systems*, vol. 27, no. 10, pp. 1986–1998, Oct. 2019.

[13] L. Liu, Z. S. Wang, and H. G. Zhang, "Adaptive fault-tolerant tracking control for MIMO discrete-time systems via reinforcement learning algorithm with less learning parameters," *IEEE Trans. Automation Science and Engineering*, vol. 14, no. 1, pp. 299–313, Jan. 2017.

[14] Y.-J. Liu, S. Li, S. C. Tong, and C. L. P. Chen, "Adaptive reinforcement learning control based on neural approximation for nonlinear discretetime systems with unknown nonaffine dead-zone input," *IEEE Trans. Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 295–305, Jan. 2019.

[15] J. N. Li, H. Modares, T. Y. Chai, F. L. Lewis, and L. H. Xie, "Off-policy reinforcement learning for synchronization in multiagent graphical games," *IEEE Trans. Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2434–2445, Oct. 2017.

[16] J. H. Qin, M. Li, Y. Shi, Q. C. Ma, and W. X. Zheng, "Optimal synchronization control of multiagent systems with input saturation via off-policy reinforcement learning," *IEEE Trans. Neural Networks and Learning Systems*, vol. 30, no. 1, pp. 85–96, Jan. 2019.

[17] X. Yang and H. B. He, "Event-triggered robust stabilization of nonlinear input-constrained systems using single network adaptive critic designs," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, doi: 10.1109/TSMC.2018.2853089, Jul. 2018.

[18] B. Widrow, N. K. Gupta, and S. Maitra, "Punish/reward: learning with a critic in adaptive threshold systems," *IEEE Trans. Systems*, *Man*, *and Cybernetics*, vol. 3, no. 5, pp. 455–465, Sept. 1973.

[19] D. V. Prokhorov and D. C. Wunsch, "Adaptive critic designs," *IEEE Trans. Neural Networks*, vol. 8, no. 5, pp. 997–1007, Sept. 1997.

[20] R. Padhi, N. Unnikrishnan, X. H. Wang, and S. N. Balakrishnan, "A single network adaptive critic (SNAC) architecture for optimal control synthesis for a class of nonlinear systems," *Neural Networks*, vol. 19, no. 10, pp. 1648–1660, 2006.

[21] D. Wang, D. R. Liu, Q. C. Zhang, and D. B. Zhao, "Data-based adaptive critic designs for nonlinear robust optimal control with uncertain dynamics," *IEEE Trans. Systems*, *Man*, *and Cybernetics*: *Systems*, vol. 46, no. 11, pp. 1544–1555, Nov. 2016.

[22] B. Luo, D. R. Liu, T. W. Huang, and D. Wang, "Model-free optimal tracking control via critic-only *Q*-learning," *IEEE Trans. Neural Networks and Learning Systems*, vol. 27, no. 10, pp. 2134–2144, Oct. 2016.

[23] H. G. Zhang, Y. H. Luo, and D. R. Liu, "Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints," *IEEE Trans. Neural Networks*, vol. 20, no. 9, pp. 1490–1503, Sept. 2009.

[24] M. M. Ha, D. Wang, and D. R. Liu, "Event-triggered adaptive critic control design for discrete-time constrained nonlinear systems," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, doi: 10.1109/TSMC.2018.2868510. Sept. 2018.

[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] H. Modares, F. L. Lewis, and M. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Trans. Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1513–1525, Oct. 2013.

[27] Y. H. Zhu, D. B. Zhao, H. B. He, and J. H. Ji, "Event-triggered optimal control for partially-unknown constrained-input systems via adaptive dynamic programming," *IEEE Trans. Industrial Electronics*, vol. 64, no. 5, pp. 4101–4109, May 2017.

[28] D. Wang, H. B. He, and D. R. Liu, "Adaptive critic nonlinear robust control: a survey," *IEEE Trans. Cybernetics*, vol. 47, no. 10, pp. 3429–3451, Oct. 2017.

[29] H. G. Zhang, K. Zhang, G. Y. 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. Systems, Man, and Cybernetics: Systems*, doi: 10.1109/TSMC.2018.2889377, Feb. 2019.

[30] L. L. Cui, X. P. Xie, X. W. Wang, Y. H. Luo, and J. B. Liu, "Event-triggered singlenetwork ADP method for constrained optimal tracking control of continuous-time nonlinear systems," *Applied Mathematics and Computation*, vol. 352, pp. 220–234, Jul. 2019.

[31] Y. Jiang, J. L. Fan, T. Y. Chai, and F. L. Lewis, "Dual-rate operational optimal control for flotation industrial process with unknown operational model," *IEEE Trans. Industrial Electronics*, vol. 66, no. 6, pp. 4587–4599, Jun. 2019.

[32] L. H. Kong, W. He, Y. T. Dong, L. Cheng, C. G. Yang, and Z. J. Li, "Asymmetric bounded neural control for an uncertain robot by state feedback and output feedback," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, doi: 10.1109/TSMC.2019.2901277, Apr. 2019.

[33] W. Zhou, H. C. Liu, H. B. He, J. Yi, and T. F. Li, "Neuro-optimal tracking control for continuous stirred tank reactor with input constraints," *IEEE Trans. Industrial Informatics*, vol. 15, no. 8, pp. 4516–4524, Aug. 2019.

[34] X. Yang, D. R. Liu, D. Wang, and Q. L. Wei, "Discrete-time online learning control for a class of unknown nonaffine nonlinear systems using reinforcement learning," *Neural Networks*, vol. 55, pp. 30–41, 2014.

[35] Y. H. Zhu, D. B. Zhao, X. Yang, and Q. C. Zhang, "Policy iteration for $H_\infty$ optimal control of polynomial nonlinear systems via sum of squares

programming," *IEEE Trans. Cybernetics*, vol. 48, no. 2, pp. 500–509, Feb. 2018.

[36] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill Publishing Co., 1976.

[37] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks*, vol. 3, no. 5, pp. 551–560, 1990.

[38] 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, May 2010.

[39] Z. J. Fu, W. F. Xie, S. Rakheja, and J. Na, "Observer-based adaptive optimal control for unknown singularly perturbed nonlinear systems with input constraints," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 1, pp. 48–57, Jan. 2017.

[40] D. S. Mitrinovic and P. M. Vasic, *Analytic Inequalities*. Berlin: Springer, 1970.

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

[42] D. R. Liu, X. Yang, D. Wang, and Q. L. Wei, "Reinforecement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints," *IEEE Trans. Cybernetics*, vol. 45, no. 7, pp. 1372–1385, Jul. 2015.

[43] X. Yang and H. B. He, "Adaptive critic learning and experience replay for decentralized event-triggered control of nonlinear interconnected systems," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, doi: 10.1109/TSMC.2019.2898370, Mar. 2019.

[44] Z. Ni, N. Malla, and X. N. Zhong, "Prioritizing useful experience replay for heuristic dynamic programming-based learning systems," *IEEE Trans. Cybernetics*, vol. 49, no. 11, pp. 3911–3922, Nov. 2019.

[45] L. Liu, Z. S. Wang, and H. G. Zhang, "Neural-network-based robust optimal tracking control for MIMO discrete-time systems with unknown uncertainty using adaptive critic design," *IEEE Trans. Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 1239–1251, Apr. 2018.

[46] Z. S. Wang, L. Liu, Y. M. Wu, and H. G. Zhang, "Optimal fault-tolerant control for discrete-time nonlinear strict-feedback systems based on adaptive critic design," *IEEE Trans. Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2179–2191, Jun. 2018.

**Xiong Yang** (M'19) received the B.S. degree in mathematics and applied mathematics from Central China Normal University, in 2008, the M.S. degree in pure mathematics from Shandong University, in 2011, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, in 2014. From 2014 to 2016, he was an Assistant Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. From 2016 to 2018, he was a Post-Doctoral Fellow with the Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA. He is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. His research interests include intelligent control, reinforcement learning, deep neural networks, event-triggered control, and their applications. Dr. Yang was a recipient of the Excellent Award of Presidential Scholarship of the Chinese Academy of Sciences in 2014 and the Outstanding Paper Award of IEEE Transactions on Neural Networks and Learning Systems in 2018.

**Bo Zhao** (M'16) received the B.S. degree in automation, and Ph.D. degree in control science and engineering, all from Jilin University, in 2009 and 2014, respectively. From 2014 to 2017, he was a Post-Doctoral Fellow with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. From 2017 to 2018, he joined the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He is currently an Associate Professor with the School of Systems Science, Beijing Normal University. He has authored or coauthored over 80 journal and conference papers. His research interests include adaptive dynamic programming, robot control, fault diagnosis and tolerant control, optimal control, and artificial intelligence-based control.