Brief Papers

Reinforcement Learning Output Feedback NN Control Using Deterministic Learning Technique

Bin Xu, Chenguang Yang, *Member, IEEE,* and Zhongke Shi

***Abstract*— In this brief, a novel adaptive-critic-based neural network (NN) controller is investigated for nonlinear pure-feedback systems. The controller design is based on the transformed predictor form, and the actor–critic NN control architecture includes two NNs, whereas the critic NN is used to approximate the strategic utility function, and the action NN is employed to minimize both the strategic utility function and the tracking error. A deterministic learning technique has been employed to guarantee that the partial persistent excitation condition of internal states is satisfied during tracking control to a periodic reference orbit. The uniformly ultimate boundedness of closed-loop signals is shown via Lyapunov stability analysis. Simulation results are presented to demonstrate the effectiveness of the proposed control.**

***Index Terms*— Approximate dynamic programming, discrete-time system, output feedback control, pure-feedback system, radial basis function neural network (RBF NN).**

I. Introduction

Neural networks (NNs) are widely employed for the control of complex nonlinear systems owing to their excellent function approximation ability. NN control has been extensively studied for both continuous-time and discrete-time systems. For continuous-time systems, much research work has been carried out on affine nonlinear systems through back-stepping or dynamic surface design [1]–[3]. In practice, there are many systems falling into this category of nonaffine systems, such as aircraft flight systems [4], biochemical processes [5], etc. There are fewer analysis tools for nonaffine systems compared to affine systems, and control of nonaffine systems is more challenging [6].

While much effort has been made in controlling continuous-time nonaffine system, NN control of discrete-time systems has also drawn much attention [7]. Lyapunov design for nonlinear discrete-time systems is much more intractable than for continuous-time systems because the linearity property of the derivative of a Lyapunov function in continuous time is not present in the difference of Lyapunov function. For the control of pure-feedback systems [8], [9], by investigating the relationship between outputs and states, the pure-feedback system is transformed into an input–output predictor model and, to overcome the difficulty of nonaffine appearance of the control input, the implicit function theorem is used. Only a single NN is used and the singularity is completely avoided. In [10] and [11], elegant controllers are first designed for nonlinear discrete-time systems by estimating the norm of the ideal weights and constructing the new adaptive NN approximation.

However, stability is only a basic necessity for the controller design. A further consideration is the optimality based on a predefined cost function [12], [13]. In [14] and [15], the continuous-time direct adaptive optimal control with infinite horizon cost with state feedback is presented. Output feedback controller schemes are necessary when certain states of the plants become unavailable for measurement. In the discrete-time case, a reinforcement-learning-based online neural controller has been designed for affine nonlinear systems [16], [17], while a class of the strict feedback system with input saturation is studied in [18]. In [19], the nonlinear discrete-time system with disturbances written in nonlinear autoregressive moving average with eXogenous input is controlled by system transformation to an affine-like form.

For adaptive system, persistent excitation (PE) is of great importance [20]. The concept was first introduced in context of system identification to express the idea that the input signal to the plant should be sufficiently rich such that all the modes of the plants are excited and convergence of the model parameters is achieved. In [21], the definition of PE in discrete-time case is presented as (6). The deterministic learning theory recently proposed in [22] uses the localized radial basis function (RBF) NN for identification of nonlinear dynamical systems undergoing periodic or recurrent motions, and it rigorously proved that a partial PE condition, i.e., the PE condition of a certain regression subvector constructed out of the RBFs along the periodic or periodic-like (recurrent) trajectories, is satisfied. The method has been applied for rapid dynamical pattern recognition [23] and ocean surface ship control [24].

In this brief, different from the previous design on affine system [17] or strict feedback system [18], a novel reinforcement-learning-based NN output feedback control scheme is developed for single-input–single-output (SISO) nonlinear systems in the pure-feedback form. Following [8] and [9], the system is first transformed into the input–output predictor. An action NN is employed to generate the near-optimal control signal to track the desired system output and to minimize the long-term cost function while the critic NN

Manuscript received December 27, 2012; revised June 12, 2013; accepted October 20, 2013. Date of publication December 11, 2013; date of current version February 14, 2014. This work was supported in part by the National Science Foundation of China under Grant 61304098 and Grant 61134004, in part by the Foundation of Key Laboratory of Autonomous Systems and Networked Control (Ministry of Education, China) under Grant 2012A04, in part by the EU FP-7 (PIIFR-GA-2010-910078-H2R) project, and in part by the NWPU Basic Research Funding under Grant JC20120236.

B. Xu and Z. Shi are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: smileface.binxu@gmail.com; zkeshi@nwpu.edu.cn).

C. Yang is with the School of Computing and Mathematics, University of Plymouth, Plymouth PL4 8AA, U.K., and with School of Automation, Beijing Institute of Technology, Beijing, 100086, China. (e-mail: cyang@ieee.org).

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/TNNLS.2013.2292704

is used to approximate the new "strategic" utility function. On satisfying the partial PE condition, the closed system achieves uniformly ultimate boundedness (UUB) stability and the tracking error converges to a bounded compact set.

II. Problem Formulation and Preliminaries

A. Pure-Feedback System

Consider the following SISO discrete-time systems in pure-feedback form:

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

where $\bar{\xi}_j(k) = [\xi_1(k), \xi_2(k), \ldots, \xi_j(k)]^T$, $j = 1, 2, \ldots, n$, $n \geq 1$, are system states, $f_i(\cdot, \cdot)$ and $f_n(\cdot, \cdot, \cdot)$ are unknown nonlinear functions, $u(k) \in \Re$ and $y(k) \in \Re$ are system input and output, respectively, and $d(k)$ denotes the external disturbance, which is bounded by an unknown constant $\bar{d}$ so that $|d(k)| \leq \bar{d}$.

Assumption 1: [9] System functions $f_i(\cdot, \cdot)$ and $f_n(\cdot, \cdot, 0)$ in (1) are continuous with respect to all the arguments and continuously differentiable with respect to the second argument.

Assumption 2: [9] There exist constants $\bar{g}_i > \underline{g}_i > 0$ such that $0 < \underline{g}_i \leq g_i(\cdot) \leq \bar{g}_i$, $i = 1, 2, \ldots, n$, where $g_j(\cdot) = \partial f_j(\bar{\xi}_j(k), \xi_{j+1}(k))/\partial \xi_{j+1}(k)$, $j = 1, 2, \ldots, n-1$, and $g_n(\cdot) = \partial f_n(\bar{\xi}_n(k), u(k), d(k))/\partial u(k)$.

For convenience, let us introduce the notations $g = \Pi_{i=1}^n g_i$, $\underline{g} = \Pi_{i=1}^n \underline{g}_i$, and $\bar{g} = \Pi_{i=1}^n \bar{g}_i$.

The control objective is to synthesize an adaptive NN control $u(k)$ for the system (1), such that all signals in the closed-loop systems are bounded and the output $y(k)$ tracks a bounded periodic reference trajectory $y_d(k)$.

B. Preliminary Results

1) Universal Approximation: The RBF NN can be described in the following form [22]:

$$\phi(W, z) = W^T S(z) = \sum_{i=1}^{N_l} \omega_i s_i(z) \tag{2}$$

where $W \in \Re^{N_l}$ is the weight vector and $S(z) = \left[s_1(z), s_2(z), \ldots, s_{N_l}(z)\right]^T$, with $s_i(\cdot)$ being the radial basis functions.

For a smooth function $\varphi(z)$ over a compact set $\mathbf{\Omega}_z \subset \Re^m$, given a small constant real number $\mu^* > 0$, if the NN nodes number N_l is sufficiently large, there exist a set of ideal bounded weights W^* such that

$$\max |\varphi(z) - \phi(W^*, z)| < \mu(z) \qquad |\mu(z)| < \mu^*. \tag{3}$$

Consider the basis functions of RBF NN (2), with z being the input vector. The following property of RBF will be used for the parameter selection in the simulation:

$$S^T(z) S(z) < N_l. \tag{4}$$

2) Spatially Localized Approximation: For the bounded trajectory $Z_\zeta(t)$ within the compact set $\mathbf{\Omega}_Z$, $\varphi(z)$ can be approximated by using neurons located in a local region along the trajectory [22], [23]

$$\varphi(z) = W_\zeta^{*T} S_\zeta(z) + \mu_\zeta(z) \tag{5}$$

where $S_\zeta(z) \in \Re^{N_\zeta}$ is a subset of $S(z)$, $N_\zeta < N_l$, $W_\zeta^* \in \Re^{N_\zeta}$, and $\mu_\zeta(z)$ is the approximation error.

Definition [21]: An input sequence $x(k)$ is said to be PE if there exist $\lambda > 0$ and $k_1 \geq 1$ such that

$$\lambda_{\min}\left[\sum_{k=k_0}^{k_1} S(x(k)) S^T(x(k))\right] > \lambda \quad \forall k_0 \geq 0 \tag{6}$$

where $\lambda_{\min}(M)$ denotes the smallest eigenvalue of M.

Lemma [22], [25]: Consider any recurrent sequence $Z(k)$. Assume that $Z(k)$ is a discrete map from $[0, \infty)$ into a compact set $\mathbf{\Omega}_Z \subset \Re^q$. For the RBF network with centers placed on a regular lattice (large enough to cover the compact set $\mathbf{\Omega}_Z$), $S_\zeta(z)$ is PE almost always.

III. Output-Feedback Online Reinforcement Learning Controller Design

Considering the system (1), for the first equation we can design the following ideal virtual controller ξ_{2f}:

$$\xi_{2f}(k) = f_1^c(\xi_1(k), y_d(k+1)). \tag{7}$$

Similarly, for the second equation, we can construct another virtual controller ξ_{3f}

$$\xi_{3f}(k) = f_2^c(\bar{\xi}_2(k), \xi_{2f}(k+1)). \tag{8}$$

However, $\xi_{2f}(k+1)$ is the future virtual control input and not available in practice. This causes the noncausal problem.

Using the same transformation procedure in [8]–[11], it is shown that system (1) under Assumptions 1 and 2 is transformable to system (9) with new function $F_n(\cdot)$

$$y(k+n) = F_n(\underline{z}(k), u(k), \underline{d}(k)) \tag{9}$$

where

$$\underline{y}(k) = [y(k), y(k-1), \ldots, y(k-n+1)]^T \tag{10}$$

$$\begin{aligned} \underline{u}(k-1) &= [u(k-1), \ldots, u(k-n+1)]^T \\ \underline{z}(k) &= [\underline{y}^T(k), \underline{u}^T(k-1)]^T \\ \underline{d}(k) &= [d(k), d(k-1), \ldots, d(k-n+1)]^T. \end{aligned} \tag{11}$$

It can be easily shown that function $F_n(\cdot)$ is continuous and continuously differentiable with respect to $u(k)$. With this function, the noncausal problem is avoided. Rewrite system (9) as

$$y(k+n) = \phi_o(\underline{z}(k), u(k)) + d_o(k) \tag{12}$$

where

$$\begin{aligned} \phi_o(\underline{z}(k), u(k)) &= F_n(\underline{z}(k), u(k), \mathbf{0_{[n]}}) \\ d_o(k) &= F_n(\underline{z}(k), u(k), \underline{d}(k)), -F_n(\underline{z}(k), u(k), \mathbf{0_{[n]}}). \end{aligned} \tag{13}$$

Similar as in [8] and [9], there exists a finite constant $\bar{d}_o$ such that $|d_o(k)| \leq \bar{d}_o$.

IV. Adaptive NN Control Design

A. Strategic Utility Function

Inspired by but different from [18], we introduce a utility function $p(k)$ based on the tracking error $e(k) = y(k) - y_d(k)$, as follows:

$$p(k) = \alpha_0 |e(k)| \tag{14}$$

where $p(k) \in \Re$, $\alpha_0 \in \Re$ is the positive design parameter. The utility function $p(k)$ is viewed as the current system-performance index. The long-term system-performance measure or the strategic utility function $Q \in \Re$ is defined using

$$Q(k) = \alpha^N p(k+1) + \alpha^{N-1} p(k+2) + \cdots + \alpha^{k+1} p(N) + \cdots \tag{15}$$

where $\alpha \in \Re$, $0 < \alpha < 1$ and N is the horizon. Equation (15) can also be expressed as $Q(k) = \min_{u(k)} \left[\alpha Q(k-1) - \alpha^{N+1} p(k)\right]$ [18].

B. Critic NN

The critic NN is used to approximate the strategic utility function $Q(k)$. The estimation is presented with the formulation

$$\hat{Q}(k) = \hat{W}_c^T(k) S_c(\underline{z}(k)),\ S_c(\underline{z}(k)) \in \Re^{l_c} \tag{16}$$

where $\hat{W}_c(k) \in \Re^{l_c}$ is the estimation of optimal NN weights W_c^*. For convenience, $S_c(\underline{z}(k))$ is denoted as $S_c(k)$.

Remark 1: In [17], [18], and [26], $\xi_n(k)$ is selected as the critic NN inputs. Since the states are not available, from the input–output transformation result [9, Eq. (50)], we know that there is a mapping between $\xi_n(k)$ and $\underline{z}(k)$ so that in the output feedback design $\underline{z}(k)$ is selected as the critic NN inputs.

We define the prediction error as

$$e_c(k) = \hat{Q}(k) - \alpha \hat{Q}(k-1) + \alpha^{N+1} p(k). \tag{17}$$

The objective function to be minimized by the critic NN is defined as

$$E_c(k) = \frac{1}{2} e_c^2(k). \tag{18}$$

The update rule for the critic NN is a gradient-based adaption, which is given by

$$\hat{W}_c(k+1) = \hat{W}_c(k) + \Delta \hat{W}_c(k) \tag{19}$$

where

$$\Delta \hat{W}_c(k) = -\alpha_c \frac{\partial E_c(k)}{\partial \hat{W}_c(k)} \tag{20}$$

or

$$\hat{W}_c(k+1) = \hat{W}_c(k) - \alpha_c S_c(k) \times \Big(\hat{W}_c^T(k) S_c(k) + \alpha^{N+1} p(k) - \alpha \hat{W}_c^T(k-1) S_c(k-1)\Big) \tag{21}$$

where $\alpha_c \in \Re$ is the NN adaption gain. The critic NN weights are tuned by the reinforcement learning signal and discounted values of critic NN past outputs.

C. Action NN

From the derivation of $F_n(\cdot)$, we see that

$$\frac{\partial \phi_o(\underline{z}(k), u(k))}{\partial u(k)} = \frac{\partial F_n(\cdot)}{\partial u(k)} = g(\cdot) \neq 0.$$

The dynamics of the tracking error $e(k) = y(k) - y_d(k)$ is given by

$$e(k+n) = \phi_o(\underline{z}(k), u(k)) - y_d(k+n) + d_o(k). \tag{22}$$

It is easy to show that

$$\frac{\partial(\phi_o(\underline{z}(k), u(k)) - y_d(k+n))}{\partial u(k)} \neq 0.$$

Therefore, according to [9, Lemma 2], there exists an ideal control input $u^*(\bar{z}(k))$ such that

$$\phi_o(\underline{z}(k), u^*(\bar{z}(k))) - y_d(k+n) = 0,$$
$$\bar{z}(k) = [\underline{z}^T(k), y_d(k+n)]^T. \tag{23}$$

Using the ideal control $u^*(\bar{z}(k))$, we have $e(k) = 0$ after n steps if $d_o(k) = 0$. It implies that the ideal control $u^*(\bar{z}(k))$ is an n-step deadbeat control.

As mentioned in Section II-B, there exists an ideal constant weights vector $W^* \in \Re^{l_a}$, such that

$$\begin{aligned} u_{nn}^*(\bar{z}(k)) &= W_a^{*T} S(\bar{z}(k)) \qquad S_a(\bar{z}(k)) \in \Re^{l_a} \\ u^*(\bar{z}(k)) &= u_{nn}^*(\bar{z}(k)) + \mu(\bar{z}(k)) \qquad \forall \bar{z} \in \mathbf{\Omega}_{\bar{z}} \end{aligned} \tag{24}$$

where $\mu(\bar{z}(k))$ is the NN approximation error and $\mathbf{\Omega}_{\bar{z}}$ is a sufficiently large compact set.

Using the RBF NN as an approximator of $u^*(\bar{z}(k))$, we propose the controller with the formulation as

$$u(k) = \hat{W}_a^T(k) S_a(\bar{z}(k)). \tag{25}$$

Adding and subtracting $\phi_o(\bar{z}(k), u^*(\bar{z}(k)))$ on the right-hand side of (22) leads to

$$\begin{aligned} e(k+n) &= \phi_o(\underline{z}(k), u(k)) - \phi_o(\underline{z}(k), u^*(\bar{z}(k))) + d_o(k) \\ &= g(\underline{z}(k), u^c(k))(u(k) - u^*(\bar{z}(k))) + d_o(k) \end{aligned} \tag{26}$$

where

$$g(\underline{z}(k), u^c(k)) = \frac{\partial \phi_o(\underline{z}(k), u^c(k))}{\partial u^c(k)}$$

with $u^c(k) \in [\min\{u^*(\bar{z}(k)), u(k)\}, \max\{u^*(\bar{z}(k)), u(k)\}]$. For convenience, let us introduce the following notations:

$$g(k) = g(\underline{z}(k), u^c(k)),\ S_a(k) = S_a(\bar{z}(k)),\ \mu(k) = \mu(\bar{z}(k)) \tag{27}$$

and it is obvious that $\underline{g} \leq g(k) \leq \bar{g}$.

Substituting (24) into (26) and noting that $\tilde{W}_a(k) = \hat{W}_a(k) - W_a^*$, we obtain

$$e(k+n) = g(k) \tilde{W}_a^T(k) S_a(k) + d^*(k) \tag{28}$$

where $\tilde{W}_a = \hat{W}_a - W_a^*$, $d^*(k) = -g(k)\mu(k) + d_o(k)$ and it is trivial to show that $|d^*(k)| \leq \bar{g}\mu^* + \bar{d}_0 := d_0^*$.

The action NN adaption law is tuned by using the system tracking error and the error between the desired strategic utility function $Q_d(k)$ and the critic signal $\hat{Q}(k)$. Our desired value for the utility function $Q_d(k)$ is "0" at each step, and then the nonlinear system can track the reference signal well. Define $k_1 = k - n$ and

$$e_{a1}(k) = g(k_1)^{-\frac{1}{2}} e(k) \tag{29}$$

$$e_{a2}(k) = g(k_1)^{-\frac{1}{2}} \hat{Q}(k). \tag{30}$$

The objective function to be minimized by the action NN is given by

$$E_a(k) = \frac{1}{2} e_{a1}^2(k) + \frac{1}{2} e_{a2}^2(k). \tag{31}$$

The update rule for the action NN is also a gradient-based adaption, which is given by

$$\hat{W}_a(k+1) = \hat{W}_a(k_1) + \Delta \hat{W}_a(k_1) \tag{32}$$

$$\begin{aligned}\Delta \hat{W}_a(k_1) &= -\alpha_a \frac{\partial E_a(k)}{\partial \hat{W}_a(k_1)}\\ &= -\alpha_a \frac{\partial E_a(k)}{\partial e_{a1}(k)} \frac{\partial e_{a1}(k)}{\partial e(k)} \frac{\partial e(k)}{\partial \hat{W}_a(k_1)}\\ &\quad -\alpha_a \frac{\partial E_a(k)}{\partial e_{a2}(k)} \frac{\partial e_{a2}(k)}{\partial e(k)} \frac{\partial e(k)}{\partial \hat{W}_a(k_1)}\\ &= -\alpha_a S_a(k_1) e(k)\\ &\quad +\alpha_a S_a(k_1) \hat{Q}(k) \alpha_0 \mathrm{sign}(e(k)) \alpha^{N+1}\\ &= -\alpha_a S_a(k_1)[e(k) - \hat{Q}_e(k)]\end{aligned}$$

or

$$\hat{W}_a(k+1) = \hat{W}_a(k_1) - \alpha_a S_a(k_1)[e(k) - \hat{Q}_e(k)] \tag{33}$$

where $\alpha_a \in \Re$ is the NN adaption gain, $\hat{Q}_e(k) = \hat{Q}(k)\alpha_0$ $\mathrm{sign}(e(k))\alpha^{N+1}$.

D. Stability Analysis

In this section, a theorem is presented to show how the controller parameters and adaption gains can be selected to ensure the performance of the closed-loop system and the UUB of all the internal signals.

Theorem 1: Consider the adaptive closed-loop system consisting of system (1) under Assumptions 1 and 2 with controller (25) and the NN weights adaptation law (21) and (33). All the signals in the closed-loop system are UUB with the bounds specifically given by (48) and (49) provided the controller design parameters are selected as follows:

1) $\alpha_c \|S_c(k)\|^2 < 1$;
2) $\alpha_a \|S_a(k)\|^2 < \frac{1}{g}$;
3) $0 < \alpha < \frac{\sqrt{3}}{3}$.

Proof: Choose a positive-definite function $V(k)$ as

$$V(k) = V_1(k) + V_2(k) + V_3(k) \tag{34}$$

where

$$\begin{aligned}V_1(k) &= \frac{1}{\alpha_c} \mathrm{tr}\left[\tilde{W}_c^T(k) \tilde{W}_c(k)\right]\\ V_2(k) &= \frac{1}{\gamma_c} \|\zeta_c(k-1)\|^2\\ V_3(k) &= \frac{1}{\gamma_a \alpha_a} \sum_{j=0}^{n} \mathrm{tr}\left[\tilde{W}_a^T(k-n+j) \tilde{W}_a(k-n+j)\right]\end{aligned}$$

where $\tilde{W}_c = \hat{W}_c - W_c^*$, $\zeta_c(k-1) = \tilde{W}_c^T(k-1)S_c(k-1)$, and $\gamma_c > 0$, $\gamma_a > 0$. The first difference of the Lyapunov function is calculated as

$$\Delta V(k) = \Delta V_1(k) + \Delta V_2(k) + \Delta V_3(k). \tag{35}$$

Define $A = \tilde{W}_c^T(k) S_c(k)$, $B = W_c^{*T} S_c(k) - \alpha W_c^{*T} S_c(k-1)$, $C = \alpha \tilde{W}_c^T(k-1) S_c(k-1)$, $D = g(k_1) \tilde{W}_a^T(k_1) S_a(k_1)$, $E = \alpha_0 \alpha^{N+1}$, $F = E|e(k)|$, and $G = d^*(k_1) - E W_c^{*T}$ $S_c(k) \mathrm{sign}(e(k))$.

Take the first term in the first difference of (35) and rewrite it as

$$\Delta V_1(k) = \frac{1}{\alpha_c} \mathrm{tr}\left[\tilde{W}_c^T(k+1) \tilde{W}_c(k+1) - \tilde{W}_c^T(k) \tilde{W}_c(k)\right]. \tag{36}$$

From (21), it is easy to see

$$\begin{aligned}\tilde{W}_c(k+1) &= \tilde{W}_c(k) - \alpha_c S_c(k)\\ &\quad \times \left(\hat{W}_c^T(k) S_c(k) + \alpha^{N+1} p(k) - \alpha \hat{W}_c^T(k-1) S_c(k-1)\right).\end{aligned}$$

Then

$$\tilde{W}_c(k+1) = \tilde{W}_c(k) - \alpha_c S_c(k)(A + B + F - C). \tag{37}$$

Substituting (37) into (36), we get

$$\begin{aligned}\Delta V_1(k) &= \alpha_c \|S_c(k)\|^2 (A + B + F - C)^2\\ &\quad -2\tilde{W}_c^T(k) S_c(k)(A + B + F - C)\\ &= \alpha_c \|S_c(k)\|^2 (A + B + F - C)^2\\ &\quad -2A(A + B + F - C)\\ &\le -(1 - \alpha_c \|S_c(k)\|^2)(A + B + F - C)^2\\ &\quad +3B^2 + 3C^2 + 3F^2 - A^2\end{aligned} \tag{38}$$

and from (28), we know

$$F^2 \le 2E^2 D^2 + 2E^2 d_0^{*2}. \tag{39}$$

Now, taking the second term in (35), we get

$$\begin{aligned}\Delta V_2(k) &= \frac{1}{\gamma_c}\left(\|\zeta_c(k)\|^2 - \|\zeta_c(k-1)\|^2\right)\\ &= \frac{1}{\gamma_c}\left(A^2 - \frac{1}{\alpha^2} C^2\right).\end{aligned} \tag{40}$$

The third term in (35) is expanded as

$$\Delta V_3(k) = \frac{1}{\gamma_a \alpha_a} \mathrm{tr}\left[\tilde{W}_a^T(k+1) \tilde{W}_a(k+1) - \tilde{W}_a^T(k_1) \tilde{W}_a(k_1)\right]. \tag{41}$$

From (33), we know

$$\tilde{W}_a(k+1) = \tilde{W}_a(k_1) - \alpha_a S_a(k_1)[D + G - E\mathrm{sign}(e(k))A].$$

Substituting the weight updates and simplifying it, we get

$$\begin{aligned}\Delta V_3(k) &= \frac{1}{\gamma_a} \alpha_a \|S_a(k)\|^2 [D + G - E\mathrm{sign}(e(k))A]^2\\ &\quad -\frac{2}{\gamma_a g(k_1)} D[D + G - E\mathrm{sign}(e(k))A]\\ &\le -\frac{1}{\gamma_a}\left(\frac{1}{g(k_1)} - \alpha_a \|S_a(k)\|^2\right)\\ &\quad \times [D + G - E\mathrm{sign}(e(k))A]^2\\ &\quad -\frac{1}{\gamma_a g(k_1)} D^2 + \frac{2}{\gamma_a g(k_1)}(E^2 A^2 + G^2).\end{aligned} \tag{42}$$

Combining (38), (40), and (42) to get the first difference of Lyapunov (35), we get

$$\begin{aligned}\Delta V(k) \le & -\left(1-\frac{1}{\gamma_c}-\frac{2E^2}{\gamma_a g(k_1)}\right)A^2\\ & -\left(\frac{1}{\gamma_c\alpha^2}-3\right)C^2-\left(\frac{1}{\gamma_a g(k_1)}-6E^2\right)D^2\\ & +3B^2+\frac{2}{\gamma_a g(k_1)}G^2+6E^2 d_0^{*2}\\ & -\left(1-\alpha_c\|S_c(k)\|^2\right)(A+B+F-C)^2\\ & -\frac{1}{\gamma_a}\left(\frac{1}{g(k_1)}-\alpha_a\|S_a(k)\|^2\right)\\ & \times [D+G-E\mathrm{sign}(e(k))A]^2.\end{aligned}$$

Define

$$H^2 = 3B^2+\frac{2}{\gamma_a g(k_1)}G^2+6E^2 d_0^{*2}. \tag{43}$$

The upper bound H_m for H is

$$H^2 \le H_m^2 = \left(6+6\alpha^2+\frac{4E^2}{\underline{g}\gamma_a}\right)\|W_c^*\|^2 l_c+\frac{4d_o^{*2}}{\underline{g}\gamma_a}+6E^2 d_0^{*2}. \tag{44}$$

Choose

$$\gamma_c \le \frac{\sqrt{3}}{3\alpha} \tag{45}$$

$$\frac{2E^2}{\underline{g}\left(1-\frac{1}{\gamma_c}\right)} < \gamma_a < \frac{1}{6\bar{g}E^2}. \tag{46}$$

With the assumption and (44), we get

$$\begin{aligned}\Delta V(k) \le & -\left(1-\frac{1}{\gamma_c}-\frac{2E^2}{\gamma_a g(k_1)}\right)A^2-\left(\frac{1}{\gamma_c\alpha^2}-3\right)C^2\\ & -\left(\frac{1}{\gamma_a g(k_1)}-6E^2\right)D^2+H_m^2.\end{aligned} \tag{47}$$

This further implies that $\Delta V(k) \le 0$ as long as

$$A > \sqrt{\frac{1}{1-\frac{1}{\gamma_c}-\frac{2E^2}{\gamma_a \underline{g}}}}H_m \tag{48}$$

or

$$D > \sqrt{\frac{1}{\frac{1}{\gamma_a \bar{g}}-6E^2}}H_m \tag{49}$$

which implies that there exist a finite K for all $k > K$, $A < 1/(1-1/\gamma_c-2E^2/\gamma_a\underline{g})^{(1/2)}H_m$ and $D < 1/(1/\gamma_a\bar{g}-6E^2)^{(1/2)}H_m$. From the definition of A and D, the boundedness of $\tilde{W}_c^T(k)S_c(k)$, $\tilde{W}_a^T(k)S_a(k)$ can be deduced. It is known that $W_c^{*T}S_c(k)$ and $W_a^{*T}S_a(k)$ are bounded so that $\hat{W}_c^T(k)S_c(k)$ and $\hat{W}_a^T(k)S_a(k)$ are bounded.

Since the NN input follows a periodic (or periodic-like) orbit, based on the Lemma 1, the PE of $S_\zeta(k)$ is satisfied. Then we see that the boundedness of $\hat{W}_c^T(k)S_c(k)$ and $\hat{W}_a^T(k)S_a(k)$ further implies that $\hat{W}_c(k)$ and $\hat{W}_a(k)$ are bounded.

With the boundedness of $g(k)$ and $d^*(k)$, we know the tracking error $e(k)$ is bounded as

$$|e(k)| \le \sqrt{\frac{1}{\frac{1}{\gamma_a\bar{g}}-6E^2}}H_m+d_0^*. \tag{50}$$

The proof is complete.

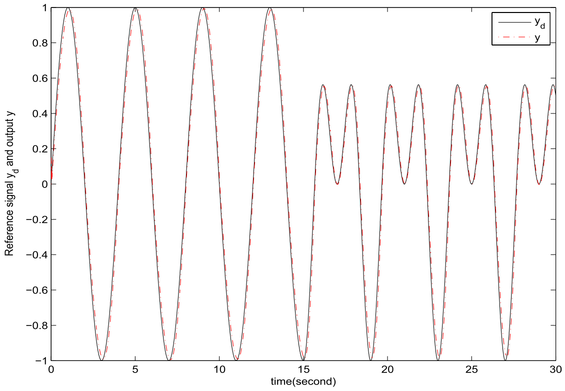

Fig. 1. Reference signal and system output.

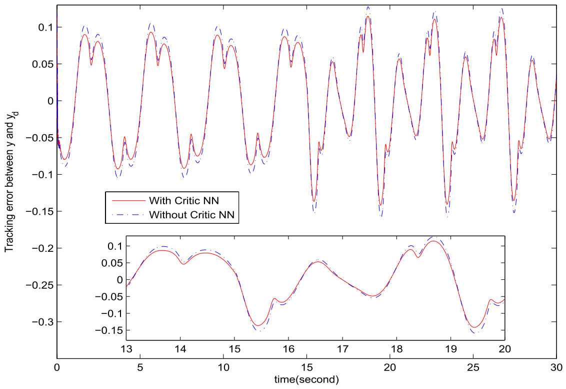

Fig. 2. Tracking error.

V. Simulation

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

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

where the system functions are

$$\begin{aligned}f_1(\xi_1(k),\xi_2(k)) &= 0.2\frac{\xi_1^2(k)\xi_2(k)}{1+\xi_1^2(k)}+0.5\xi_2(k)\\ f_2(\bar{\xi}_2(k),u(k)) &= \frac{\xi_1(k)}{1+\xi_1^2(k)+\xi_2^2(k)}+u(k)+0.2\sin(u(k))\end{aligned}$$

where the disturbance is $d(k) = 0.05\cos(0.01k)\cos(\xi_1(k))$. The control objective is to make the output $y(k)$ track the desired reference trajectory

$$y_d(k) = \begin{cases}\sin(\pi kT/2) & \text{if } k < 3000\\ 0.5\cos(\pi kT)+0.5\sin(\pi kT/2) & \text{if } k \ge 3000\end{cases}$$

where $T = 0.005$, and guarantee the boundedness of all the closed-loop signals. The system initial states are $\bar{\xi}_2(0) = [0,\ 0.3]^T$. The action RBF NN is constructed with $l_a = 81$ neurons with centers evenly spaced in $[-1;1]\times[-1;1]\times[-1;1]\times[-1;1]$ while the action RBF NN is constructed with $l_c = 27$ neurons with centers evenly spaced

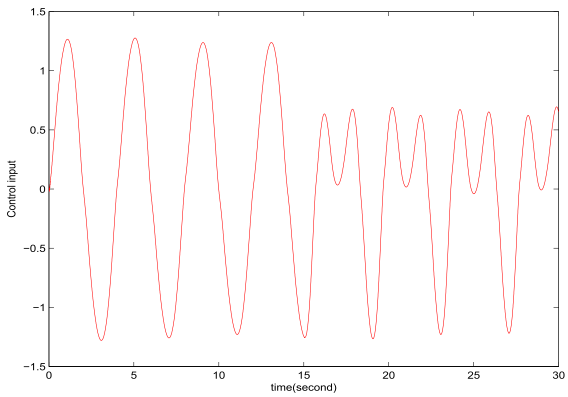

Fig. 3. Control input.

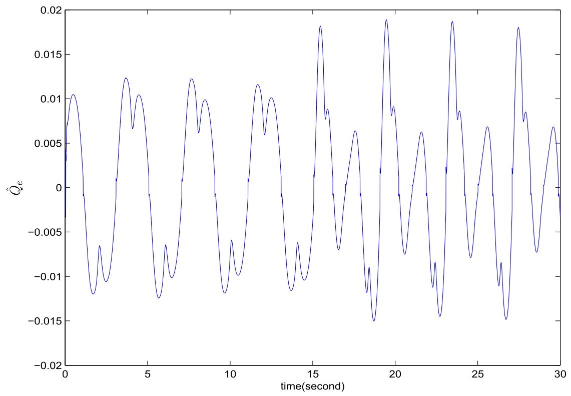

Fig. 4. Critic signal $\hat{Q}_e$.

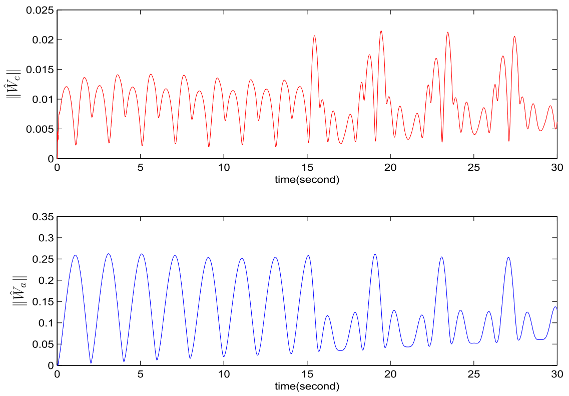

Fig. 5. Critic and action NN weights.

in $[-1; 1] \times [-1; 1] \times [-1; 1]$. The width for two NNs is selected as 1.

It can be checked that $g_{1,1}(\cdot) = \partial f_1(\cdot)/\partial \xi_2(k) = 0.5 + 0.2\xi_1^2(k)/1 + \xi_1^2(k) \in [0.5, 0.7]$ and $g_{1,2}(\cdot) = \partial f_2(\cdot)/\partial u(k) = 1 + 0.2\cos(u(k)) \in [0.8, 1.2]$ such that $g(\cdot) = \partial f(\cdot)/\partial u(k) = g_{1,1}(\cdot)g_{1,2}(\cdot) \in [0.4, 0.84)$ and Assumption 2 is satisfied. According to the condition of the theorem, the parameters are selected as $\alpha_c = 1/27$, $\alpha_a = 0.01$, $\alpha = 0.515$, $\alpha_0 = 7$, and $N = 4$.

The simulation results are presented in Figs. 1–5. Fig. 1 shows that the controller can track the reference signal very well. To show the difference from the controller without critic signal, the tracking error is illustrated in Fig. 2, and it can be observed obviously that the proposed method with critic signal $\hat{Q}_e$ depicted in Fig. 4 achieves better tracking performance. Figs. 3 and 5 illustrate the boundedness of the control input $u(k)$ and the NN weights vector estimate $\hat{W}_a(k)$,$\hat{W}_c(k)$.

VI. Conclusion

In this brief, a novel adaptive-critic-based NN controller has been investigated for the nonlinear pure-feedback systems. To overcome the noncausal problem, the system was transformed to a predictor for output feedback control design. The critic NN was used to approximate the "strategic" utility function, whereas an action NN was employed to minimize both the strategic utility function and the tracking error. Results demonstrated that the Lyapunov-based adaptive critic design renders satisfactory performance while ensuring closed-loop stability.

References

[1] W. Chen and J. Li, "Decentralized output-feedback neural control for systems with unknown interconnections," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 38, no. 1, pp. 258–266, Feb. 2008.

[2] W. Chen and L. Jiao, "Adaptive tracking for periodically time-varying and nonlinearly parameterized systems using multilayer neural networks," *IEEE Trans. Neural Netw.*, vol. 21, no. 2, pp. 345–351, Feb. 2010.

[3] M. Chen, S. Ge, and B. How, "Robust adaptive neural network control for a class of uncertain MIMO nonlinear systems with input nonlinearities," *IEEE Trans. Neural Netw.*, vol. 21, no. 5, pp. 796–812, May 2010.

[4] L. Hunt and G. Meyer, "Stable inversion for nonlinear systems," *Automatica*, vol. 33, no. 8, pp. 1549–1554, Aug. 1997.

[5] M. Krstic, I. Kanellakopoulos, and P. V. Kokotovic, *Nonlinear and Adaptive Control Design*. New York, NY, USA: Wiley, 1995.

[6] C. Wang, D. Hill, S. Ge, and G. Chen, "An ISS-modular approach for adaptive neural control of pure-feedback systems," *Automatica*, vol. 42, no. 5, pp. 723–731, May 2006.

[7] J. Sarangapani, *Neural Network Control of Nonlinear Discrete-Time Systems*. Boca Raton, FL, USA: CRC Press, 2006.

[8] S. Ge, C. Yang, and T. Lee, "Adaptive predictive control using neural network for a class of pure-feedback systems in discrete time," *IEEE Trans. Neural Netw.*, vol. 19, no. 9, pp. 1599–1614, Sep. 2008.

[9] C. Yang, S. Ge, C. Xiang, T. Chai, and T. Lee, "Output feedback NN control for two classes of discrete-time systems with unknown control directions in a unified approach," *IEEE Trans. Neural Netw.*, vol. 19, no. 11, pp. 1873–1886, Nov. 2008.

[10] Y. Liu, G. Wen, and S. Tong, "Direct adaptive NN control for a class of discrete-time nonlinear strict-feedback systems," *Neurocomputing*, vol. 73, nos. 13–15, pp. 2498–2505, 2010.

[11] Y. Liu, C. Chen, G. Wen, and S. Tong, "Adaptive neural output feedback tracking control for a class of uncertain discrete-time nonlinear systems," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1162–1167, Jul. 2011.

[12] D. Liu, Y. Zhang, and H. Zhang, "A self-learning call admission control scheme for cdma cellular networks," *IEEE Trans. Neural Netw.*, vol. 16, no. 5, pp. 1219–1228, Sep. 2005.

[13] D. Liu and Q. Wei, "Finite-approximation-error-based optimal control approach for discrete-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 779–789, Apr. 2013.

[14] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

[15] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Netw.*, vol. 22, no. 3, pp. 237–246, 2009.

[16] D. Liu, D. Wang, D. Zhao, Q. Wei, and N. Jin, "Neural-network-based optimal control for a class of unknown discrete-time nonlinear systems using globalized dual heuristic programming," *IEEE Trans. Autom. Sci. Eng.*, vol. 9, no. 3, pp. 628–634, Jul. 2012.

[17] Q. Yang and S. Jagannathan, "Reinforcement learning controller design for affine nonlinear discrete-time systems using online approximators," *IEEE Trans. Syst., Man, Cybern., B: Cybern.*, vol. 42, no. 2, pp. 377–390, Apr. 2012.

[18] P. He and S. Jagannathan, "Reinforcement learning neural-network-based controller for nonlinear discrete-time systems with input constraints," *IEEE Trans. Syst., Man, Cybern., B: Cybern.*, vol. 37, no. 2, pp. 425–436, Apr. 2007.

[19] Q. Yang, J. Vance, and S. Jagannathan, "Control of nonaffine nonlinear discrete-time systems using reinforcement-learning-based linearly parameterized neural networks," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 38, no. 4, pp. 994–1001, Aug. 2008.

[20] C. Wang and D. Hill, *Deterministic Learning Theory for Identification, Recognition, and Control*, vol. 32. Boca Raton, FL, USA: CRC Press, 2009.

[21] N. Sadegh, "A perceptron network for functional identification and control of nonlinear systems," *IEEE Trans. Neural Netw.*, vol. 4, no. 6, pp. 982–988, Nov. 1993.

[22] C. Wang and D. Hill, "Learning from neural control," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 130–146, Jan. 2006.

[23] C. Wang and D. Hill, "Deterministic learning and rapid dynamical pattern recognition," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 617–630, May 2007.

[24] S. Dai, C. Wang, and F. Luo, "Identification and learning control of ocean surface ship using neural networks," *IEEE Trans. Ind. Informat.*, vol. 8, no. 4, pp. 801–810, Nov. 2012.

[25] T. Chen and C. Wang, "Learning from neural control for a class of discrete-time nonlinear systems," in *Proc. 48th IEEE CDC/CCC*, Dec. 2009, pp. 6732–6737.

[26] L. Yang, J. Si, K. S. Tsakalis, and A. A. Rodriguez, "Direct heuristic dynamic programming for nonlinear tracking control with filtered tracking error," *IEEE Trans. Syst., Man, Cybern., B, Cybern.*, vol. 39, no. 6, pp. 1617–1622, Dec. 2009.