Data-Based Adaptive Critic Designs for Nonlinear Robust Optimal Control With Uncertain Dynamics

Ding Wang, *Member, IEEE*, Derong Liu, *Fellow, IEEE*, Qichao Zhang, and Dongbin Zhao, *Senior Member, IEEE*

***Abstract*—In this paper, the infinite-horizon robust optimal control problem for a class of continuous-time uncertain nonlinear systems is investigated by using data-based adaptive critic designs. The neural network identification scheme is combined with the traditional adaptive critic technique, in order to design the nonlinear robust optimal control under uncertain environment. First, the robust optimal controller of the original uncertain system with a specified cost function is established by adding a feedback gain to the optimal controller of the nominal system. Then, a neural network identifier is employed to reconstruct the unknown dynamics of the nominal system with stability analysis. Hence, the data-based adaptive critic designs can be developed to solve the Hamilton–Jacobi–Bellman equation corresponding to the transformed optimal control problem. The uniform ultimate boundedness of the closed-loop system is also proved by using the Lyapunov approach. Finally, two simulation examples are presented to illustrate the effectiveness of the developed control strategy.**

***Index Terms*—Adaptive critic designs, adaptive dynamic programming, intelligent control, neural networks, policy iteration, robust optimal control, system identification, uncertain nonlinear systems.**

I. Introduction

MODEL uncertainties arise frequently in practical control systems, such as mechanical systems, transportation systems, and power systems and can severely degrade the closed-loop system performance. Therefore, the problem of designing robust controllers for nonlinear systems with uncertainties has drawn considerable attention in the literature for many years [1]–[6]. Although various direct robust control approaches have been proposed previously, the relationship between robust control and optimal control has been studied recently to derive new robust control methods [4]–[6]. Lin *et al.* [4] showed that the robust control problem could be solved by studying the optimal control of corresponding nominal system, but detailed procedures were not given. Lin and Brandt [5] presented an optimal control approach to achieve robust control of robot manipulators. The nominal part of the controlled system was linear and hence, the optimal controller could be obtained by solving an algebraic Riccati equation. Since many practical systems possess nonlinearity and uncertainty, it is necessary to study the robust control problem when the nominal parts are nonlinear systems. Wang *et al.* [6] developed a novel iterative algorithm for online design of robust control for a class of continuous-time nonlinear systems. This was a meaningful result which used the advanced computational intelligence technique to deal with the traditional nonlinear robust control problem. However, the optimality of the robust controller with respect to a specified cost function was not discussed, not to mention that the dynamics of the nominal system were assumed to be known. This restricts its application to some extent and also motivates our research.

The basic idea of the design strategy in this paper comes from neural-network-based optimal control, or neuro-optimal control. As is known, dealing with the nonlinear optimal control problem always requires solving the Hamilton–Jacobi–Bellman (HJB) equation. Although dynamic programming is a conventional method in solving optimization and optimal control problems, it often suffers from the curse of dimensionality. To avoid the difficulty, based on function approximators, such as neural networks, adaptive or approximate dynamic programming (ADP) was proposed by Werbos [7] as a method to solve optimal control problems forward-in-time. Recently, research on ADP and related fields has gained much attention from various scholars [6], [8]–[50]. The comprehensive research progress and prospects of ADP for optimal control can be found in [8] and [9]. Remarkably, more and more researchers have pointed out that ADP is a biologically inspired and computational method to construct truly brain-like systems in the field of computational intelligence and intelligent control [7], [8], [10], [20], [33], [45].

Reinforcement learning is a class of approaches used in the field of machine learning to derive the optimal action of an agent based on responses from its environment. Lewis and Vrabie [10] stated that the ADP technique was closely related to reinforcement learning and that policy iteration was one of the basic algorithms of reinforcement learning. In addition, the information of system dynamics is necessarily required when the traditional policy iteration algorithm is employed. Vamvoudakis and Lewis [13] discussed

Manuscript received June 11, 2015; revised August 16, 2015; accepted September 26, 2015. Date of publication October 30, 2015; date of current version October 13, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 61233001, Grant 61273136, Grant 61273140, Grant 61304086, Grant 61374105, Grant 61533017, and Grant 61573353, and in part by the Early Career Development Award of the State Key Laboratory of Management and Control for Complex Systems. This paper was recommended by Associate Editor Z. Wang.

D. Wang, Q. Zhang, and D. Zhao are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: ding.wang@ia.ac.cn; zhangqichao2014@ia.ac.cn; dongbin.zhao@ia.ac.cn).

D. Liu is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: derong@ustb.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/TSMC.2015.2492941

an online algorithm based on policy iteration for learning the continuous-time optimal control solution with infinite horizon cost for nonlinear systems with known dynamics. They presented an online adaptive algorithm which involved simultaneous tuning for both actor and critic neural networks. Modares *et al.* [15] proposed an online learning algorithm, based on the policy iteration technique, to find the optimal control solution for continuous-time systems subject to input constraints. However, for many complex systems, it is difficult to acquire accurate models of controlled plants. Then, Modares *et al.* [16] presented an online policy iteration algorithm to learn the continuous-time optimal control solution for unknown constrained-input systems. Unlike existing results which require complete or at least partial knowledge about the system dynamics, the proposed method does not need any knowledge about the system dynamics. Liu *et al.* [27] developed an online synchronous approximate optimal learning algorithm based on policy iteration to solve a multiplayer nonzero-sum game without requiring exact knowledge of dynamic systems. Besides, Luo *et al.* [32] addressed the model-free nonlinear optimal control problem based on data by introducing the reinforcement learning technique. They proposed a data-based approximate policy iteration method by using real system data rather than a system model. However, system uncertainties are not considered in most works. Recently, Jiang and Jiang [35] studied the robust optimal control design for a class of uncertain nonlinear systems from a perspective of robust ADP. It is an important work of integrating tools from nonlinear control with the idea of ADP, which not only stabilizes the original uncertain system, but also achieves optimality in the absence of dynamic uncertainty. Note that the optimization issue related to the original uncertain system is not included. In many situations, it is necessary to define suitable cost functions corresponding to nonlinear systems with uncertainties and discuss the optimality. Note that though the robust optimal control of nonlinear systems has been studied in [37], it is reconsidered in this paper from the following two aspects. On one hand, the dynamics of nominal system is not required by constructing a neural network identifier. On the other hand, the model-free policy iteration algorithm is presented to solve the transformed optimal control problem with stability analysis different from that of [37]. Overall, to the best of our knowledge, there are no results on robust optimal control of uncertain nonlinear systems through data-based adaptive critic designs method. This is the motivation of this paper.

Actually, in this paper, it is the first time that the robust optimal control scheme for a class of uncertain nonlinear systems via data-based adaptive critic learning technique is established. First, the optimal controller of the nominal system is designed. It can be proved that the modification of the optimal control law is in fact the robust controller of the original uncertain system, which also achieves optimality under the definition of a cost function. Then, a data-based ADP technique, which relies on two neural networks, namely, a model network and a critic network, is developed to solve the transformed optimal control problem. The uniform ultimate boundedness of the closed-loop system is also proved via the well-known Lyapunov approach. At last, two simulation examples are given to show the effectiveness of the robust optimal control scheme. It is found that the developed approach not only extends the application scope of ADP to nonlinear optimal control design under uncertain environment, but also provides a novel robust optimal control method for uncertain nonlinear systems. The significance lies in the fact that it employs the idea of computational intelligence to construct and design self-learning and intelligent control systems.

The rest of this paper is organized as follows. In Section II, the robust control design of uncertain nonlinear system is stated with some backgrounds of nonlinear optimal control design. In Section III, the robust optimal control method of uncertain nonlinear system is provided with theoretical proof. In Section IV, the optimal control implementation via data-based adaptive critic learning approach is developed with stability analysis, by using neural network and policy iteration techniques. In Section V, two numerical examples are given to demonstrate the effectiveness of the established approach. In Section VI, concluding remarks and discussion on future work are presented.

II. Problem Statement and Preliminaries

In this paper, we study a class of continuous-time uncertain nonlinear systems described by

$$\dot{x}(t) = f(x(t)) + g(x(t))\big(\bar{u}(t) + \bar{d}(x(t))\big) \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $\bar{u}(t) \in \mathbb{R}^m$ is the control vector, $f(\cdot)$ and $g(\cdot)$ are differentiable in their arguments with $f(0) = 0$, and $\bar{d}(x) \in \mathbb{R}^m$ is the unknown nonlinear perturbation. We let $x(0) = x_0$ be the initial state and assume $\bar{d}(0) = 0$ ensuring that $x = 0$ is an equilibrium of (1). Similar to many other literature, for the corresponding nominal system

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t) \tag{2}$$

we also assume that $f + gu$ is Lipschitz continuous on a set Ω in $\mathbb{R}^n$ containing the origin and that (2) is controllable.

For designing the robust optimal control of (1), we should find a feedback control law $\bar{u}(x)$, such that the closed-loop system is globally asymptotically stable for all uncertainties $\bar{d}(x)$ and the optimality related to a specified cost function is attained. Next, we will show that it can be transformed into solving the optimal control problem for the nominal system with an appropriate cost function.

Let $R \in \mathbb{R}^{m\times m}$ be a symmetric positive definite matrix. We denote $d(x) = R^{1/2}\bar{d}(x)$ with $d(x) \in \mathbb{R}^m$ bounded by a known function $d_M(x)$, i.e., $\|d(x)\| \le d_M(x)$ with $d_M(0) = 0$. For (2), for the purpose of solving the infinite horizon optimal control problem, we should derive the control law $u(x)$ that minimizes the cost function

$$J(x_0) = \int_0^\infty \left\{d_M^2(x(\tau)) + u^{\mathsf{T}}(x(\tau))Ru(x(\tau))\right\}\mathrm{d}\tau. \tag{3}$$

According to the classical optimal control theory, the feedback control must not only stabilize the controlled system

on Ω, but also guarantee that the cost function (3) is finite (i.e., the designed control law must be admissible). The definition of admissible control can be found in [12], [13], and [26].

Let $\Psi(\Omega)$ be the set of admissible controls on Ω. For any admissible control law $u \in \Psi(\Omega)$, if the associated cost function (3) is continuously differentiable, its infinitesimal version is the nonlinear Lyapunov equation

$$0 = d_M^2(x) + u^{\mathsf{T}}(x)Ru(x) + (\nabla J(x))^{\mathsf{T}}(f(x) + g(x)u(x)) \quad (4)$$

with $J(0) = 0$. In (4), the symbol $\nabla(\cdot) \triangleq \partial(\cdot)/\partial x$ is the notation of the gradient operator, for example, $\nabla J(x) = \partial J(x)/\partial x$.

Define the Hamiltonian of (2) as

$$H(x, u, \nabla J(x)) = d_M^2(x) + u^{\mathsf{T}}(x)Ru(x) + (\nabla J(x))^{\mathsf{T}}(f(x) + g(x)u(x)). \quad (5)$$

The optimal cost function of (2) is formulated as

$$J^*(x_0) = \min_{u\in\Psi(\Omega)} \int_0^\infty \left\{ d_M^2(x(\tau)) + u^{\mathsf{T}}(x(\tau))Ru(x(\tau)) \right\} \mathrm{d}\tau.$$

In view of optimal control theory, the optimal cost function $J^*(x)$ satisfies the HJB equation

$$0 = \min_{u\in\Psi(\Omega)} H\big(x, u, \nabla J^*(x)\big). \quad (6)$$

Assume that the minimum on the right-hand side of (6) exists and is unique. Then, the optimal control law is

$$u^*(x) = -\frac{1}{2}R^{-1}g^{\mathsf{T}}(x)\nabla J^*(x). \quad (7)$$

Based on (5) and (7), the HJB equation (6) of (2) becomes

$$0 = d_M^2(x) + (\nabla J^*(x))^{\mathsf{T}} f(x) - \frac{1}{4}\big(\nabla J^*(x)\big)^{\mathsf{T}} g(x)R^{-1}g^{\mathsf{T}}(x)\nabla J^*(x) \quad (8)$$

with $J^*(0) = 0$. In the following, we will discuss how the optimal control problem of nominal system (2) is linked with the robust optimal control of original uncertain system (1).

III. Robust Optimal Control Methodology of Uncertain Nonlinear Systems

In this section, we first develop a robust control law for the original uncertain system (1) and then show that the robust control law possesses the property of optimality under a specified cost function. Some results of [37] will be used to build the theoretical basis of the nonlinear robust optimal control methodology, which is necessary and helpful to the development of data-based robust optimal control strategy in the next section.

To establish the robust stabilizing control strategy of (1), we modify the optimal control law (7) of (2) by adding a feedback gain π, that is

$$\bar{u}(x) = \pi u^*(x) = -\frac{1}{2}\pi R^{-1}g^{\mathsf{T}}(x)\nabla J^*(x). \quad (9)$$

Considering (8) and (9), the derivative of $L_1(t) = J^*(x(t))$ along the trajectory of the closed-loop system can be formulated as

$$\dot{L}_1(t) = \big(\nabla J^*(x)\big)^{\mathsf{T}}(f(x) + g(x)\bar{u}(x)) = -d_M^2(x) - \frac{1}{2}\left(\pi - \frac{1}{2}\right)\left\|R^{-1/2}g^{\mathsf{T}}(x)\nabla J^*(x)\right\|^2.$$

Clearly, $\dot{L}_1(t) < 0$ whenever $\pi \geq 1/2$ and $x \neq 0$. Hence, for (2), the feedback control given by (9) ensures that the closed-loop system is asymptotically stable for all $\pi \geq 1/2$.

Lemma 1 [37]: For (1), there exists a positive number $\pi_1^* \geq 1$, such that for any $\pi > \pi_1^*$, the feedback control developed by (9) ensures that the closed-loop system is asymptotically stable.

The proof of Lemma 1 is provided in [37]. According to Lemma 1, $\bar{u}(x)$ with $\pi > \pi_1^* \geq 1$ is a robust control law of the original system (1). Next, we show that it also holds the property of optimality. To this end, we have to define a cost function related to the original system (1). Consider

$$\bar{J}(x_0) = \int_0^\infty \left\{ Q(x(\tau)) + \frac{1}{\pi}\bar{u}^{\mathsf{T}}(x(\tau))R\bar{u}(x(\tau)) \right\} \mathrm{d}\tau \quad (10)$$

where

$$Q(x) = d_M^2(x) - \big(\nabla J^*(x)\big)^{\mathsf{T}} g(x)\bar{d}(x) + \frac{1}{4}(\pi - 1)\big(\nabla J^*(x)\big)^{\mathsf{T}} g(x)R^{-1}g^{\mathsf{T}}(x)\nabla J^*(x). \quad (11)$$

By adding and subtracting $(1/(\pi-1))d^{\mathsf{T}}(x)d(x)$ to (11) and noticing the fact that $d^{\mathsf{T}}(x)d(x) \leq d_M^2(x)$, we can easily find that

$$Q(x) \geq d_M^2(x) - \frac{1}{\pi - 1}d^{\mathsf{T}}(x)d(x) \geq \frac{\pi - 2}{\pi - 1}d_M^2(x).$$

Clearly, there exists a positive number $\pi_2^* \geq 2$, such that for all $\pi > \pi_2^*$, $Q(x)$ is a positive definite function. Hence, it is important to derive that the definition of new utility function, i.e., $Q(x)+(1/\pi)\bar{u}^{\mathsf{T}}R\bar{u}$, corresponding to the original uncertain system (1) is reasonable. Then, we obtain the main theorem of this section.

Theorem 1: Consider (1) with infinite-horizon cost function (10). There exists a positive number π^* such that for any $\pi > \pi^*$, the feedback control law obtained by (9) is an asymptotically stabilizing solution of the designed optimal control problem.

Proof: The Hamiltonian of (1) with cost function (10) is

$$\bar{H}\big(x, \bar{u}, \nabla\bar{J}(x)\big) = Q(x) + \frac{1}{\pi}\bar{u}^{\mathsf{T}}(x)R\bar{u}(x) + \big(\nabla\bar{J}(x)\big)^{\mathsf{T}}\big(f(x) + g(x)\big(\bar{u}(x) + \bar{d}(x)\big)\big)$$

where $\pi > \pi_2^* \geq 2$. Using (8), (9), and (11), we can derive that

$$\bar{H}\big(x, \bar{u}, \nabla\bar{J}(x)\big) = \big(\nabla\bar{J}(x) - \nabla J^*(x)\big)^{\mathsf{T}} \times \big(f(x) + g(x)\big(\bar{u}(x) + \bar{d}(x)\big)\big).$$

By replacing $\bar{J}(x)$ with $J^*(x)$, we obtain $\bar{H}(x, \bar{u}, \nabla J^*(x)) = 0$. This shows that $J^*(x)$ is a solution to the HJB equation of (1). Correspondingly, the optimal control law of (1) is πu^*. Then, we say that the control law (9) achieves optimality with cost function (10). Overall, there exists a positive number $\pi^* \triangleq \max\{\pi_1^*, \pi_2^*\}$ such that for any $\pi > \pi^*$, the control law (9) is an asymptotically stabilizing solution to the corresponding optimal control problem. This completes the proof. ■

Remark 1: Based on Theorem 1, there exists a $\pi > \pi^*$ such that the control law (9) can not only stabilize (1), but also achieve optimality with the defined cost function (10). Moreover, we find that the function $\bar{J}(x)$ relies on the choice of feedback gain π. When π varies, the cost function $\bar{J}(x)$ varies, and then the optimal control of (1) also varies. However, the form of the optimal control is fixed, i.e., πu^*.

According to Theorem 1, in order to design the robust optimal control of (1), we should put emphasis upon solving the optimal control problem of nominal system (2). As we observe from the previous parts, the ADP method is effective to solve the nonlinear optimization and optimal control problems. Then, in the following, we will provide a neural-network-based data-driven optimal control approach for (2) and prove the stability of the closed-loop system.

IV. Optimal Control Implementation via Data-Based Adaptive Critic Designs

In this section, we present the optimal control implementation via data-based adaptive critic designs. A neural network identifier is constructed and trained to learn the system dynamics. Then, a model-free policy iteration algorithm for the transformed optimal control problem is developed and implemented by building a critic neural network. Stability analysis of the closed-loop system is provided in detail as well.

A. Neural Network Identification

In this paper, we assume that the internal and drift dynamics of (2) are unknown. A three-layer neural network identifier is used to reconstruct the unknown dynamics by using input–output data. Let the number of hidden layer neurons be denoted by l_m. The corresponding nominal system (2) based on neural network can be represented as

$$\dot{x} = Ax + \omega_m^{\mathsf{T}}\sigma_m\left(\nu_m^{\mathsf{T}} z\right) + \varepsilon_m. \tag{12}$$

Let $\bar{z} = \nu_m^{\mathsf{T}} z$, where $\bar{z} \in \mathbb{R}^{l_m}$. In (12), A is a designed stable matrix, $z = [x^{\mathsf{T}}, u^{\mathsf{T}}]^{\mathsf{T}} \in \mathbb{R}^{n+m}$ is the neural network input vector, $\nu_m \in \mathbb{R}^{(n+m)\times l_m}$ is the ideal weight matrix between input layer and hidden layer, $\omega_m \in \mathbb{R}^{l_m \times n}$ is the ideal weight matrix between hidden layer and output layer, $\varepsilon_m \in \mathbb{R}^n$ is the functional approximation error, and $\sigma_m(\cdot) \in \mathbb{R}^{l_m}$ is the activation function selected as a monotonically increasing one, such as $\sigma_m(\cdot) = \tanh(\cdot)$. Similar to [27], [49], and [50], for any $y_1, y_2 \in \mathbb{R}$ $(y_1 \geq y_2)$, there exists a constant λ_0 $(\lambda_0 > 0)$, such that

$$\sigma_m(y_1) - \sigma_m(y_2) \leq \lambda_0(y_1 - y_2). \tag{13}$$

During system identification, let the weight matrix between input layer and hidden layer be constant while only tuning the weight matrix between hidden layer and output layer. Hence, the output of neural network identifier can be presented as

$$\dot{\hat{x}} = A\hat{x} + \hat{\omega}_m^{\mathsf{T}}(t)\sigma_m\big(\hat{z}\big)$$

where $\hat{\omega}_m(t)$ is the current estimated matrix of the ideal weight matrix ω_m at time t, $\hat{x}$ is the estimated system state, and $\hat{z} = \nu_m^{\mathsf{T}}[\hat{x}^{\mathsf{T}}, u^{\mathsf{T}}]^{\mathsf{T}}$. Then, the dynamics of the identification error can be obtained by

$$\dot{\tilde{x}} = A\tilde{x} + \tilde{\omega}_m^{\mathsf{T}}(t)\sigma_m\big(\hat{z}\big) + \omega_m^{\mathsf{T}}\big(\sigma_m(\bar{z}) - \sigma_m\big(\hat{z}\big)\big) + \varepsilon_m \tag{14}$$

where $\tilde{\omega}_m = \omega_m - \hat{\omega}_m$ is the weight estimation error of the identifier and $\tilde{x} = x - \hat{x}$ is the system identification error.

Here, we provide the following two assumptions, which are commonly used in papers such as [27], [49], and [50].

Assumption 1: The ideal neural network weight matrices are bounded by two positive constants, i.e., $\|\omega_m\| \leq \lambda_{\omega_m}$ and $\|\nu_m\| \leq \lambda_{\nu_m}$.

Assumption 2: The functional approximation error ε_m is upper bounded by a function of identification error, such that $\varepsilon_m^{\mathsf{T}}\varepsilon_m \leq \lambda_{\varepsilon_m}\tilde{x}^{\mathsf{T}}\tilde{x}$, where λ_{ε_m} is a positive constant.

The stability of identification error dynamics is proved in the following theorem.

Theorem 2: Suppose that Assumptions 1 and 2 are satisfied. The identification error $\tilde{x}$ is asymptotically stable, if the weight matrix of the neural network identifier is updated by

$$\dot{\hat{\omega}}_m = \Gamma_m\sigma_m\big(\hat{z}\big)\tilde{x}^{\mathsf{T}} \tag{15}$$

where $\Gamma_m \in \mathbb{R}^{l_m \times l_m}$ is a symmetric positive definite matrix of learning rates.

Proof: Choose the following Lyapunov function:

$$L_2(t) = \frac{1}{2}\tilde{x}^{\mathsf{T}}\tilde{x} + \frac{1}{2}\mathrm{tr}\Big(\tilde{\omega}_m^{\mathsf{T}}\Gamma_m^{-1}\tilde{\omega}_m\Big).$$

We take the derivative of $L_2(t)$ along the trajectory generated by the identification error (14) as

$$\dot{L}_2(t) = \tilde{x}^{\mathsf{T}}\dot{\tilde{x}} + \mathrm{tr}\Big(\tilde{\omega}_m^{\mathsf{T}}\Gamma_m^{-1}\dot{\tilde{\omega}}_m\Big).$$

Based on (14) and (15), we have

$$\dot{L}_2(t) = \tilde{x}^{\mathsf{T}}A\tilde{x} + \tilde{x}^{\mathsf{T}}\omega_m^{\mathsf{T}}\big(\sigma_m(\bar{z}) - \sigma_m\big(\hat{z}\big)\big) + \tilde{x}^{\mathsf{T}}\varepsilon_m. \tag{16}$$

According to (13) and Assumption 1, we can obtain

$$\begin{aligned}\tilde{x}^{\mathsf{T}}\omega_m^{\mathsf{T}}\big(\sigma_m(\bar{z}) - \sigma_m\big(\hat{z}\big)\big) &\leq \frac{1}{2}\tilde{x}^{\mathsf{T}}\omega_m^{\mathsf{T}}\omega_m\tilde{x} + \frac{1}{2}\big(\sigma_m(\bar{z}) - \sigma_m\big(\hat{z}\big)\big)^{\mathsf{T}} \\ &\quad \times \big(\sigma_m(\bar{z}) - \sigma_m\big(\hat{z}\big)\big) \\ &\leq \frac{1}{2}\tilde{x}^{\mathsf{T}}\omega_m^{\mathsf{T}}\omega_m\tilde{x} + \frac{1}{2}\lambda_0^2\|\bar{z} - \hat{z}\|^2 \\ &\leq \frac{1}{2}\tilde{x}^{\mathsf{T}}\omega_m^{\mathsf{T}}\omega_m\tilde{x} + \frac{1}{2}\lambda_0^2\lambda_{\nu_m}^2\tilde{x}^{\mathsf{T}}\tilde{x}. \end{aligned} \tag{17}$$

Based on (16) and (17) and considering Assumption 2, we have

$$\begin{aligned}\dot{L}_2(t) &\le \tilde{x}^{\mathsf{T}}A\tilde{x} + \frac{1}{2}\tilde{x}^{\mathsf{T}}\omega_m^{\mathsf{T}}\omega_m\tilde{x} + \frac{1}{2}\lambda_0^2\lambda_{\nu_m}^2\tilde{x}^{\mathsf{T}}\tilde{x} \\ &\quad + \frac{1}{2}\tilde{x}^{\mathsf{T}}\tilde{x} + \frac{1}{2}\lambda_{\varepsilon_m}\tilde{x}^{\mathsf{T}}\tilde{x} \\ &= \tilde{x}^{\mathsf{T}}\left(A + \frac{1}{2}\omega_m^{\mathsf{T}}\omega_m + \frac{1}{2}\left(1 + \lambda_{\varepsilon_m} + \lambda_0^2\lambda_{\nu_m}^2\right)I_n\right)\tilde{x} \\ &\triangleq -\tilde{x}^{\mathsf{T}}\Xi\tilde{x} \end{aligned} \tag{18}$$

where

$$\Xi = -A - \frac{1}{2}\omega_m^{\mathsf{T}}\omega_m - \frac{1}{2}\left(1 + \lambda_{\varepsilon_m} + \lambda_0^2\lambda_{\nu_m}^2\right)I_n$$

and I_n stands for the identity matrix with dimension n. If A is selected to make $\Xi > 0$, the Lyapunov derivative is negative such that $\dot{L}_2(t) \le 0$. Hence, it can be concluded that the identification error approaches zero, i.e., $\tilde{x}(t) \to 0$ as $t \to \infty$. This completes the proof. ■

From Theorem 2, we know the model neural network is an asymptotically stable identifier. Consequently, after a sufficient learning session, we can obtain the following neural network identifier:

$$\dot{x} = f(x) + g(x)u = Ax + \omega_m^{\mathsf{T}}\sigma_m(\bar{z}). \tag{19}$$

In addition, by taking the partial derivative of both sides of (19) with respect to u, we can obtain

$$\begin{aligned} g(x) &= \frac{\partial\left(Ax + \omega_m^{\mathsf{T}}\sigma(\bar{z})\right)}{\partial u} \\ &= \omega_m^{\mathsf{T}}\frac{\partial\sigma(\bar{z})}{\partial\bar{z}}\nu_m^{\mathsf{T}}\left[\begin{array}{c} 0_{n\times m} \\ \hline I_m \end{array}\right]. \end{aligned} \tag{20}$$

Remark 2: By virtue of neural network identification, the unknown system dynamics and control matrix of (2) can be approached by (19) and (20), respectively. Actually, as approximated values, the state derivative $\dot{x}$ in (19) and control matrix $g(x)$ in (20) should be denoted by $\hat{\dot{x}}$ and $\hat{g}(x)$, respectively. However, we still use $\dot{x}$ and $g(x)$ in the following for convenience of analysis.

Remark 3: The expressions of $\dot{x}$ and $g(x)$ in (19) and (20) are related with the converged weight vectors of the neural network identifier. In this sense, it is feasible to develop a data-based optimal control method under the framework of ADP, which is helpful to achieve the robust optimal control of uncertain nonlinear system.

B. Model-Free Policy Iteration Algorithm

In this section, a model-free policy iteration algorithm working together with neural network identifier for nominal system (2) is presented. Via system identification, we can acquire the weight matrices ω_m and ν_m. Then, based on (19) and (20), we can develop the model-free policy iteration algorithm for the transformed optimal control problem as shown in Algorithm 1.

Remark 4: Note that the above algorithm can converge to the optimal cost function and optimal control law, i.e., $J^{(i)}(x) \to J^*(x)$ and $u^{(i)}(x) \to u^*(x)$ as $i \to \infty$. The convergence proof of policy iteration algorithm has been given in [12] and related references therein and is therefore omitted here.

Algorithm 1 Model-Free Policy Iteration Algorithm for the Transformed Optimal Control Problem

1: Initialization
Let the initial iteration index be $i = 0$ and $J^{(0)}(\cdot) = 0$.
Give a small positive real number ϵ.
Start with an initial admissible control law $u^{(0)}$.

2: Neural Network Identification
Through system identification, compute the approximate values of $\dot{x}$ and $g(x)$ according to (19) and (20), respectively. Keep the converged weight matrices unchanged.

3: Policy Evaluation
Using the information of $\dot{x}$, solve the following nonlinear Lyapunov equation

$$0 = d_M^2(x) + \left(u^{(i)}(x)\right)^{\mathsf{T}}Ru^{(i)}(x) + \left(\nabla J^{(i+1)}(x)\right)^{\mathsf{T}}\dot{x}$$

with $J^{(i+1)}(0) = 0$.

4: Policy Improvement
Using the information of $g(x)$, update the control law via

$$u^{(i+1)}(x) = -\frac{1}{2}R^{-1}g^{\mathsf{T}}(x)\nabla J^{(i+1)}(x).$$

5: Stopping Criterion
If $\| J^{(i+1)}(x) - J^{(i)}(x) \| \le \epsilon$, stop and obtain the approximate optimal control law $u^{(i+1)}(x)$; else, set $i = i + 1$ and go to step 3.

C. Implementation Process via Critic Network

Considering the universal approximation property of neural networks, $J^*(x)$ can be reconstructed by a single-layer neural network on a compact set Ω as

$$J^*(x) = \omega_c^{\mathsf{T}}\sigma_c(x) + \varepsilon_c(x)$$

where $\omega_c \in \mathbb{R}^{l_c}$ is the ideal weight, $\sigma_c(x) \in \mathbb{R}^{l_c}$ is the activation function, l_c is the number of neurons in the hidden layer, and $\varepsilon_c(x)$ is the approximation error. Then, we have

$$\nabla J^*(x) = (\nabla\sigma_c(x))^{\mathsf{T}}\omega_c + \nabla\varepsilon_c(x). \tag{21}$$

Based on (19) and (21), the Lyapunov equation (4) becomes

$$\begin{aligned} 0 &= d_M^2(x) + u^{\mathsf{T}}(x)Ru(x) \\ &\quad + \left(\omega_c^{\mathsf{T}}\nabla\sigma_c(x) + (\nabla\varepsilon_c(x))^{\mathsf{T}}\right)\left(Ax + \omega_m^{\mathsf{T}}\sigma_m(\bar{z})\right). \end{aligned}$$

As the work of [13], [16], and [36], we assume that ω_c, $\nabla\sigma_c(x)$, and $\varepsilon_c(x)$ and its derivative $\nabla\varepsilon_c(x)$ are all bounded on a compact set Ω.

Since the ideal weights are unknown, a critic neural network is built in terms of the estimated weights as

$$\hat{J}(x) = \hat{\omega}_c^{\mathsf{T}}\sigma_c(x)$$

for the purpose of approximating the optimal cost function, where $\hat{\omega}_c$ represents the estimated weight matrix. Then, we obtain

$$\nabla\hat{J}(x) = (\nabla\sigma_c(x))^{\mathsf{T}}\hat{\omega}_c. \tag{22}$$

Noticing (7), (20), and (21), we find that

$$u^*(x) = -\frac{1}{2}R^{-1}g^{\mathsf{T}}(x)\Big((\nabla\sigma_c(x))^{\mathsf{T}}\omega_c + \nabla\varepsilon_c(x)\Big). \quad (23)$$

Accordingly, considering (7), (20), and (22), the approximate control function is expressed as

$$\hat{u}(x) = -\frac{1}{2}R^{-1}g^{\mathsf{T}}(x)(\nabla\sigma_c(x))^{\mathsf{T}}\hat{\omega}_c. \quad (24)$$

Applying (24) to neural network identifier (19), the closed-loop system dynamics can be rewritten as

$$\dot{x} = f(x) - \frac{1}{2}g(x)R^{-1}g^{\mathsf{T}}(x)(\nabla\sigma_c(x))^{\mathsf{T}}\hat{\omega}_c.$$

Denoting $M = g(x)R^{-1}g^{\mathsf{T}}(x)$ and using the neural network expression (21), the Hamiltonian becomes

$$\begin{aligned} H(x,\omega_c) &= d_M^2(x) + \omega_c^{\mathsf{T}}\nabla\sigma_c(x)f(x) - e_{cH} \\ &\quad - \frac{1}{4}\omega_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\omega_c \\ &= 0 \end{aligned} \quad (25)$$

where

$$e_{cH} = -(\nabla\varepsilon_c(x))^{\mathsf{T}}\Big(Ax + \omega_m^{\mathsf{T}}\sigma_m(\bar{z}^*)\Big) - \frac{1}{4}(\nabla\varepsilon_c(x))^{\mathsf{T}}M\nabla\varepsilon_c(x)$$

denotes the residual error with $\bar{z}^* = \nu_m^{\mathsf{T}}[x^{\mathsf{T}}, u^{*\mathsf{T}}]^{\mathsf{T}}$. Assume that there exists a positive bound $\lambda_{e_{cH}}$ such that $\|e_{cH}\| \le \lambda_{e_{cH}}$. Based on the estimated weight vector, the approximate Hamiltonian can be derived as

$$\begin{aligned} \hat{H}(x,\hat{\omega}_c) &= d_M^2(x) + \frac{1}{4}\hat{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\hat{\omega}_c \\ &\quad + \hat{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)\Big(Ax + \omega_m^{\mathsf{T}}\sigma_m(\bar{z})\Big) \\ &= d_M^2(x) + \hat{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)f(x) \\ &\quad - \frac{1}{4}\hat{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\hat{\omega}_c \\ &\triangleq e_c \end{aligned} \quad (26)$$

where $\bar{z} = \nu_m^{\mathsf{T}}[x^{\mathsf{T}}, \hat{u}^{\mathsf{T}}]^{\mathsf{T}}$. Let the weight estimation error of the critic network be $\tilde{\omega}_c = \omega_c - \hat{\omega}_c$. Combining (23), (25), and (26), we can get

$$\begin{aligned} e_c = -\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)&\Big(f(x) - \frac{1}{2}M(\nabla\sigma_c(x))^{\mathsf{T}}\omega_c\Big) \\ &- \frac{1}{4}\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\tilde{\omega}_c + e_{cH}. \end{aligned} \quad (27)$$

In this paper, in order to train the critic network, we aim at designing $\hat{\omega}_c$ to minimize the objective function

$$E_c = \frac{1}{2}e_c^{\mathsf{T}}e_c.$$

The weights of the critic network are tuned based on the standard steepest descent algorithm, that is

$$\dot{\hat{\omega}}_c = -\alpha_c\left(\frac{\partial E_c}{\partial\hat{\omega}_c}\right) \quad (28)$$

where $\alpha_c > 0$ is the learning rate of the critic network.

In the following, we derive the dynamics of the weight estimation error $\tilde{\omega}_c$. According to (26), we find that

$$\begin{aligned} \frac{\partial e_c}{\partial\hat{\omega}_c} &= \nabla\sigma_c(x)f(x) - \frac{1}{2}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\hat{\omega}_c \\ &= \nabla\sigma_c(x)(f(x) + g(x)\hat{u}(x)) \\ &= \nabla\sigma_c(x)\Big(Ax + \omega_m^{\mathsf{T}}\sigma_m(\bar{z})\Big). \end{aligned} \quad (29)$$

Denoting $\theta = \nabla\sigma_c(x)(Ax + \omega_m^{\mathsf{T}}\sigma_m(\bar{z}))$ and combining (27) and (29), the dynamics of the weight estimation error is written as

$$\begin{aligned} \dot{\tilde{\omega}}_c &= \alpha_c e_c\left(\frac{\partial e_c}{\partial\hat{\omega}_c}\right) \\ &= -\alpha_c\theta\Big\{\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)\Big(f(x) - \frac{1}{2}M(\nabla\sigma_c(x))^{\mathsf{T}}\omega_c\Big) \\ &\qquad + \frac{1}{4}\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\tilde{\omega}_c - e_{cH}\Big\} \end{aligned} \quad (30)$$

which is useful to prove the stability of the weight estimation error of the critic network.

D. Stability Analysis

In this section, the stability analysis of the closed-loop system is established based on the Lyapunov approach.

Definition 1 [16], [29]: For nonlinear system $\dot{x} = f(x(t))$, its solution is said to be uniformly ultimately bounded (UUB), if there exists a compact set $\Omega \subset \mathbb{R}^n$ such that for all $x_0 \in \Omega$, there exist a bound Λ and a time $T(\Lambda, x_0)$ such that $\|x(t) - x_e\| \le \Lambda$ for all $t \ge t_0 + T$, where x_e is an equilibrium point.

Remark 5: The UUB stability emphasizes that after a transition period T, the system state remains within the ball of radius Λ around x_e.

Now, we derive the following theorem.

Theorem 3: Consider the system described by (19). Let the weight tuning law of the critic network be updated by (28) and the control law be computed by (24). Then, the closed-loop system state x, the system identification error $\tilde{x}(t)$, and the weight estimation error $\tilde{\omega}_c$ of the critic network are all UUB.

Proof: Choose the following Lyapunov function:

$$L(t) = L_1(t) + L_2(t) + L_3(t)$$

where $L_1(t) = J^*(x(t))$, $L_2(t)$ is defined as in Theorem 2, and

$$L_3(t) = \frac{1}{2\alpha_c}\tilde{\omega}_c^{\mathsf{T}}\tilde{\omega}_c.$$

The derivative of the Lyapunov function $L(t)$ along the trajectory of (19) is computed as

$$\dot{L}(t) \le \dot{L}_1(t) - \tilde{x}^{\mathsf{T}}\Xi\tilde{x} + \dot{L}_3(t) \quad (31)$$

where Ξ is defined as in (18).

Combining (21) and (24), the first term in (31) can be presented as

$$\begin{aligned} \dot{L}_1(t) &= \omega_c^{\mathsf{T}}\nabla\sigma_c(x)f(x) - \frac{1}{2}\omega_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\hat{\omega}_c + \varepsilon_1 \\ &= \omega_c^{\mathsf{T}}\nabla\sigma_c(x)f(x) - \frac{1}{2}\omega_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\omega_c \\ &\quad + \frac{1}{2}\omega_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\tilde{\omega}_c + \varepsilon_1 \end{aligned}$$

where

$$\varepsilon_1 = (\nabla \varepsilon_c(x))^{\mathsf{T}}\Big(Ax + \omega_m^{\mathsf{T}}\sigma_m(\tilde{z})\Big).$$

Based on (25), we can get

$$\begin{aligned}\dot{L}_1(t) &= -d_M^2(x) - \frac{1}{4}\omega_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\omega_c \\ &\quad + \frac{1}{2}\omega_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\tilde{\omega}_c + \varepsilon_1 + e_{cH} \\ &\le -d_M^2(x) + \frac{1}{4}\|\tilde{\omega}_c\|^2 + \frac{1}{4}\Big(\lambda_{1m} + \lambda_{1M}^2\Big)\lambda_{\omega_c}^2 \\ &\quad + \varepsilon_1 + e_{cH} \end{aligned} \tag{32}$$

where $\lambda_{1m} > 0$ and $\lambda_{1M} > 0$ denote the lower and upper bounds of the norm of matrix $\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}$, respectively, and $\lambda_{\omega_c} > 0$ represents the upper bound of $\|\omega_c\|$.

Combining (24) and the definition of θ, we have

$$\theta = \nabla\sigma_c(x)D + \frac{1}{2}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\tilde{\omega}_c \tag{33}$$

where

$$D = f(x) - \frac{1}{2}M(\nabla\sigma_c(x))^{\mathsf{T}}\omega_c.$$

Combining (30) and (33), the last term in (31) can be rewritten as

$$\begin{aligned}\dot{L}_3(t) &= \frac{1}{\alpha_c}\tilde{\omega}_c^{\mathsf{T}}\dot{\tilde{\omega}}_c \\ &= -\Big(\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)D + \frac{1}{2}\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\tilde{\omega}_c\Big) \\ &\quad \times \Big(\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)D \\ &\qquad + \frac{1}{4}\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\tilde{\omega}_c - e_{cH}\Big).\end{aligned}$$

By using the inequalities

$$\begin{aligned} ab &= \frac{1}{2}\left(-\left(\phi_+ a - \frac{b}{\phi_+}\right)^2 + \phi_+^2 a^2 + \frac{b^2}{\phi_+^2}\right) \\ &\le \frac{1}{2}\left(\phi_+^2 a^2 + \frac{b^2}{\phi_+^2}\right) \\ -ab &= -\frac{1}{2}\left(\left(\phi_- a + \frac{b}{\phi_-}\right)^2 - \phi_-^2 a^2 - \frac{b^2}{\phi_-^2}\right) \\ &\le \frac{1}{2}\left(\phi_-^2 a^2 + \frac{b^2}{\phi_-^2}\right)\end{aligned}$$

where ϕ_+ and ϕ_- are nonzero constants, we can find that $\dot{L}_3(t)$ can be rewritten as

$$\begin{aligned}\dot{L}_3(t) &\le -\frac{1}{16}\Big(\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)M(\nabla\sigma_c(x))^{\mathsf{T}}\tilde{\omega}_c\Big)^2 \\ &\quad + 4\Big(\tilde{\omega}_c^{\mathsf{T}}\nabla\sigma_c(x)D\Big)^2 + \frac{33}{8}e_{cH}^2 \\ &\le -\frac{1}{16}\lambda_{1m}^2\|\tilde{\omega}_c\|^4 + 4\lambda_{\nabla\sigma_c}^2\lambda_D^2\|\tilde{\omega}_c\|^2 + \frac{33}{8}e_{cH}^2 \end{aligned} \tag{34}$$

where $\lambda_{\nabla\sigma_c} > 0$ denotes the upper bound of $\|\nabla\sigma_c(x)\|$ and $\lambda_D > 0$ represents the upper bound of $\|D\|$.

Assume that we can determine a quadratic bound of $d(x)$, i.e., $d_M(x) = \rho_0\|x\|$ with a positive constant ρ_0. Then, based on (31), (32), and (34), we obtain

$$\begin{aligned}\dot{L}(t) &\le -d_M^2(x) - \tilde{x}^{\mathsf{T}}\Xi\tilde{x} - \frac{1}{16}\lambda_{1m}^2\|\tilde{\omega}_c\|^4 \\ &\quad + \frac{1 + 16\lambda_{\nabla\sigma_c}^2\lambda_D^2}{4}\|\tilde{\omega}_c\|^2 + \frac{33}{8}e_{cH}^2 \\ &\quad + \frac{1}{4}\Big(\lambda_{1m} + \lambda_{1M}^2\Big)\lambda_{\omega_c}^2 + \varepsilon_1 + e_{cH} \\ &\le -\rho_0^2\|x\|^2 - \lambda_{\min}(\Xi)\|\tilde{x}\|^2 \\ &\quad - \frac{\lambda_{1m}^2}{16}\left(\|\tilde{\omega}_c\|^2 - \frac{2\Big(1 + 16\lambda_{\nabla\sigma_c}^2\lambda_D^2\Big)}{\lambda_{1m}^2}\right)^2 + \zeta\end{aligned}$$

where

$$\begin{aligned}\zeta &= \frac{1}{4}\Big(\lambda_{1m} + \lambda_{1M}^2\Big)\lambda_{\omega_c}^2 + \lambda_{\varepsilon_1} + \lambda_{e_{cH}} \\ &\quad + \frac{33}{8}\lambda_{e_{cH}}^2 + \frac{\Big(1 + 16\lambda_{\nabla\sigma_c}^2\lambda_D^2\Big)^2}{4\lambda_{1m}^2}\end{aligned}$$

$\lambda_{\varepsilon_1} > 0$ denotes the upper bound of $\|\varepsilon_1\|$, and $\lambda_{\min}(\Xi) > 0$ stands for the minimum eigenvalue of the positive definite matrix Ξ.

Given the following inequality:

$$\|x\| > \sqrt{\frac{\zeta}{\rho_0^2}}$$

or

$$\|\tilde{x}\| > \sqrt{\frac{\zeta}{\lambda_{\min}(\Xi)}}$$

or

$$\|\tilde{\omega}_c\| > \sqrt{\sqrt{\frac{16\zeta}{\lambda_{1m}^2}} + \frac{2\Big(1 + 16\lambda_{\nabla\sigma_c}^2\lambda_D^2\Big)}{\lambda_{1m}^2}}$$

holds, then $\dot{L}(t) < 0$. Therefore, using the standard Lyapunov extension theorem, we can derive that the closed-loop system state x, the system identification error $\tilde{x}(t)$, and the weight estimation error $\tilde{\omega}_c$ of the critic network are all UUB. This completes the proof. ■

Remark 6: Currently, the selection of activation function of the critic network is often a natural choice guided by engineering experience and intuition (i.e., it is more of an art than science) [12], [36], [49]. In addition, the initial admissible control law is necessary to perform the model-free policy iteration algorithm, keeping the same characteristic as model-based policy iteration algorithm used in [6], [12], and [15]. Though it is difficult to acquire in some cases, it can be chosen in light of experience and intuition. These ways of choosing the initial parameters are reasonable under the framework of ADP method.

Remark 7: During this section, we can develop an approximate optimal control law of nominal system with unknown dynamics under the definition of a new cost function. According to Theorem 1, by choosing a suitable feedback

gain π, we can establish the robust optimal control strategy of the original nonlinear system with unknown dynamics and uncertainties.

V. Simulation

Two examples are provided in this section to demonstrate the effectiveness of the robust optimal control strategy.

Example 1: Consider the continuous-time nonlinear system given as follows:

$$\dot{x} = \begin{bmatrix} -0.5x_1 + x_2\left(1 + 0.5x_2^2\right) \\ -0.8(x_1 + x_2) + 0.5x_2\left(1 - 0.3x_2^2\right) \end{bmatrix} + \begin{bmatrix} 0 \\ -0.5 \end{bmatrix}\left(\bar{u} + \bar{d}(x)\right) \quad (35)$$

where $x = [x_1, x_2]^{\mathsf{T}} \in \mathbb{R}^2$ and $\bar{u} \in \mathbb{R}$ are the state and control variables, respectively. Note that $\bar{d}(x) = \delta_1 x_2 \cos(\delta_2 x_1 + \delta_3 x_2)$ denotes the dynamics uncertainty of the controlled plant, where δ_1, δ_2, and δ_3 are unknown parameters with $\delta_1 \in [-1, 1]$, $\delta_2 \in [-5, 5]$, and $\delta_3 \in [-3, 3]$. We set $R = I$ and choose $d_M(x) = \|x\|$ as the bound of the term $d(x)$.

According to the aforementioned results, for the purpose of obtaining the model-free optimal control of nominal system

$$\dot{x} = \begin{bmatrix} -0.5x_1 + x_2\left(1 + 0.5x_2^2\right) \\ -0.8(x_1 + x_2) + 0.5x_2\left(1 - 0.3x_2^2\right) \end{bmatrix} + \begin{bmatrix} 0 \\ -0.5 \end{bmatrix} u$$

with cost function defined as

$$J(x_0) = \int_0^\infty \left\{ \|x(\tau)\|^2 + u^{\mathsf{T}}(x(\tau)) R u(x(\tau)) \right\} \mathrm{d}\tau$$

we should construct a neural network based on the idea of ADP. Here, the critic network is built in the form of

$$\hat{J}(x) = \hat{\omega}_{c1} x_1^2 + \hat{\omega}_{c2} x_1 x_2 + \hat{\omega}_{c3} x_2^2 + \hat{\omega}_{c4} x_1^4 + \hat{\omega}_{c5} x_1^3 x_2 + \hat{\omega}_{c6} x_1^2 x_2^2 + \hat{\omega}_{c7} x_1 x_2^3 + \hat{\omega}_{c8} x_2^4.$$

We first choose a three-layer feedforward neural network as an identifier with structure 3–8–2. During the system identification process, the constant weight ν_m between input layer and hidden layer is chosen randomly within $[-0.5, 0.5]$, and the initial weight ω_m is initialized to zero. We train the neural network identifier by using the update law (15) for 100 s with the learning matrix $\Gamma_m = 0.01I$. Via simulation, we find that the neural network identifier can learn the unknown nonlinear system successfully. Notice that the identification errors are shown in Fig. 1. Then, we finish the training process of the neural network identifier and fix its weights.

Then, the weights of critic network are initialized in $[0, 1]$ to make the initial control law of policy iteration algorithm admissible. A probing noise is also brought in to satisfy the persistency of excitation condition. Let the learning rate of critic network be $\alpha_c = 0.8$ and the initial state of controlled plant be $x_0 = [0.5, -0.5]^{\mathsf{T}}$. After the simulation process, we can observe that the convergence of the weights occurs after 2500 s. Then, the probing signal is turned off. From simulation

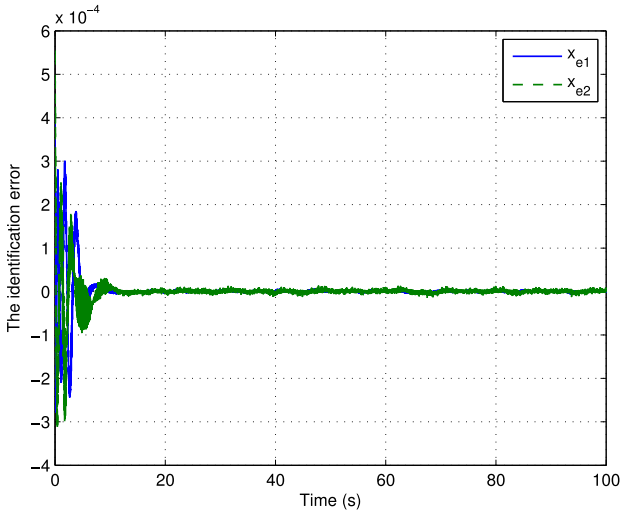


Fig. 1. Identification error (x_{e1} and x_{e2} represent $\tilde{x}_1$ and $\tilde{x}_2$, respectively).

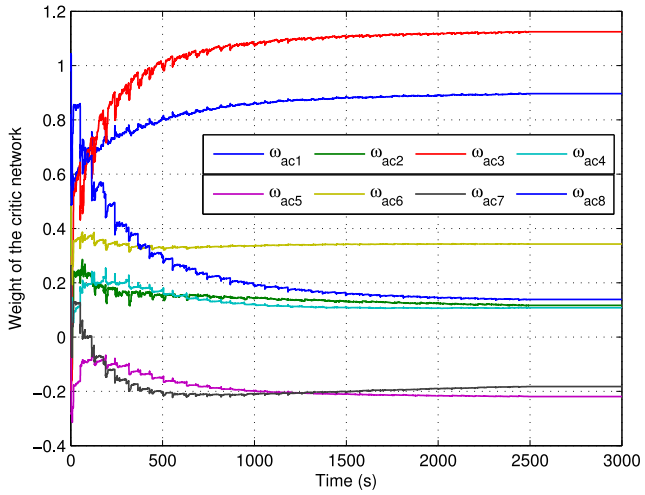


Fig. 2. Convergence of the weight vector of critic network (ω_{aci} represents $\hat{\omega}_{ci}$, $i = 1, 2, \ldots, 8$).

results, we can observe that the weights of critic network converge to

$$[0.8963, 0.1167, 1.1244, 0.1078, -0.2189, 0.3428, -0.1820, 0.1386]^{\mathsf{T}}$$

which is displayed in Fig. 2.

Next, scalar parameters in three different cases are chosen to evaluate the robust control performance. Under the action of the robust control strategy, the state trajectories of uncertain system (35) during the first 20 s in three cases are shown in Fig. 3. In light of Theorem 1, the robust control strategy also achieves optimality with a cost function defined in (10). These simulation results verify the effectiveness of the developed control approach.

Example 2: Consider the following continuous-time nonlinear system:

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ 0.1x_1 - x_2 - x_1 x_3 \\ x_1 x_2 - x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}\left(\bar{u} + \bar{d}(x)\right) \quad (36)$$

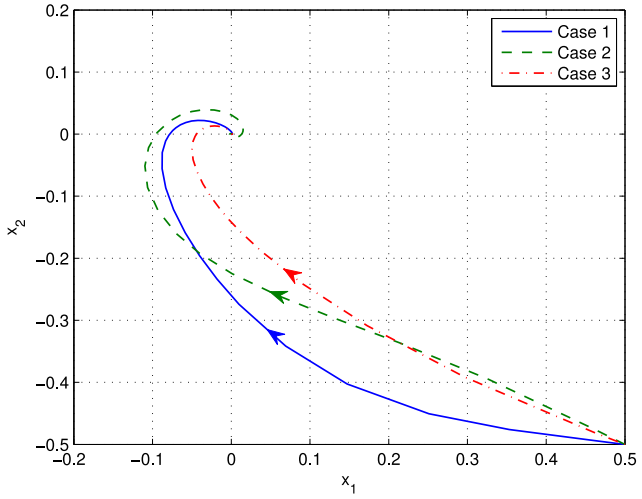


Fig. 3. State trajectories under different system uncertainties. Case 1: $\delta_1 = 0.8$, $\delta_2 = -5$, and $\delta_3 = 3$. Case 2: $\delta_1 = -1$, $\delta_2 = 4$, and $\delta_3 = -2$. Case 3: $\delta_1 = 0.5$, $\delta_2 = 0$, and $\delta_3 = 0$.

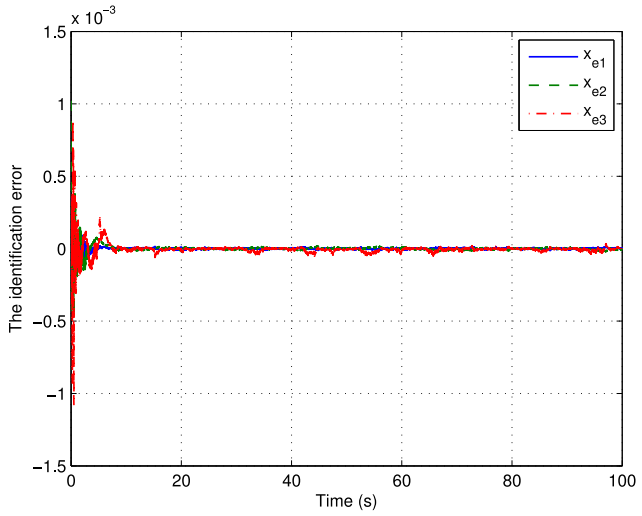


Fig. 4. Identification error (x_{e1}, x_{e2}, and x_{e3} represent $\tilde{x}_1$, $\tilde{x}_2$, and $\tilde{x}_3$, respectively).

where $x = [x_1, x_2, x_3]^{\mathsf{T}}$, $\bar{u} \in \mathbb{R}$, and $\bar{d}(x) = \delta_1 x_1 \sin(\delta_2 x_2 + \delta_3 x_3^3 + \delta_4)$ with $\delta_1 \in [-1, 1]$, $\delta_2 \in [-3, 3]$, $\delta_3 \in [-1, 1]$, and $\delta_4 \in [-5, 5]$.

For nominal system

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ 0.1x_1 - x_2 - x_1 x_3 \\ x_1 x_2 - x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u$$

with a cost function defined the same as the one of Example 1, we employ the data-based adaptive critic learning approach developed in this paper to attain the approximate optimal control law. In this example, the critic network is constructed in the following form:

$$\begin{aligned} \hat{J}(x) = {} & \hat{\omega}_{c1} x_1^2 + \hat{\omega}_{c2} x_2^2 + \hat{\omega}_{c3} x_3^2 + \hat{\omega}_{c4} x_1 x_2 + \hat{\omega}_{c5} x_1 x_3 \\ & + \hat{\omega}_{c6} x_2 x_3 + \hat{\omega}_{c7} x_1^4 + \hat{\omega}_{c8} x_2^4 + \hat{\omega}_{c9} x_3^4 \\ & + \hat{\omega}_{c10} x_1^2 x_2^2 + \hat{\omega}_{c11} x_1^2 x_3^2 + \hat{\omega}_{c12} x_2^2 x_3^2 \\ & + \hat{\omega}_{c13} x_1^2 x_2 x_3 + \hat{\omega}_{c14} x_1 x_2^2 x_3 + \hat{\omega}_{c15} x_1 x_2 x_3^2 \\ & + \hat{\omega}_{c16} x_1^3 x_2 + \hat{\omega}_{c17} x_1^3 x_3 + \hat{\omega}_{c18} x_1 x_2^3 \\ & + \hat{\omega}_{c19} x_2^3 x_3 + \hat{\omega}_{c20} x_1 x_3^3 + \hat{\omega}_{c21} x_2 x_3^3. \end{aligned}$$

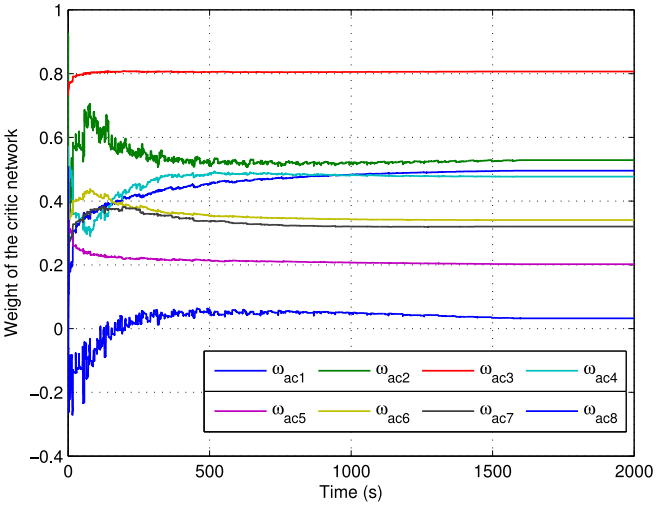


Fig. 5. Convergence of the weight vector of critic network: part 1 (ω_{aci} represents $\hat{\omega}_{ci}$, $i = 1, 2, \ldots, 8$).

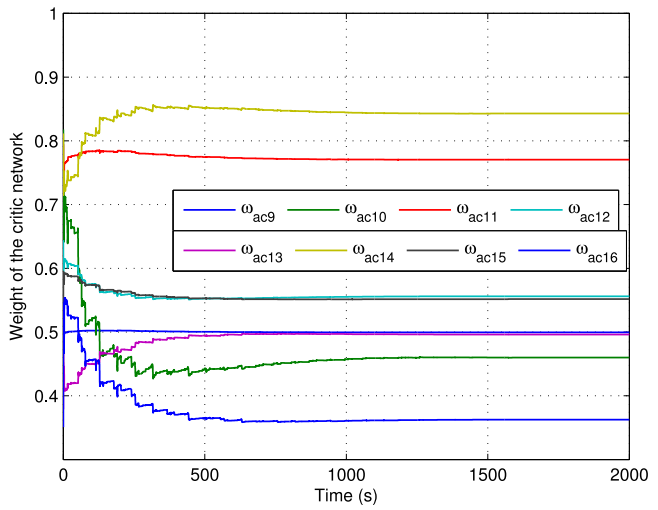


Fig. 6. Convergence of the weight vector of critic network: part 2 (ω_{aci} represents $\hat{\omega}_{ci}$, $i = 9, 10, \ldots, 16$).

We also choose a three-layer feedforward neural network identifier with structure 4–8–3. Other parameters are chosen the same as Example 1. Via simulation, we find that the neural network identifier can learn the unknown nonlinear system successfully. The identification error is shown in Fig. 4. Then, we finish training the neural network and keep the weight vectors unchanged.

Here, let the initial state vector of the controlled system be $x_0 = [1, -1, 0.5]^{\mathsf{T}}$. During the training process of critic network, let the learning rate of the critic network be $\alpha_c = 1.2$. Similar to above, an exploration noise is added to satisfy the persistency of excitation condition. After a sufficient learning session, the weights of the critic network converge to

$$\begin{aligned} & [0.4956, 0.5286, 0.8069, 0.4772, 0.2022, 0.3405, 0.3203, \\ & \quad 0.0324, 0.4995, 0.4599, 0.7706, 0.5561, 0.4960, 0.8429, \\ & \quad 0.5517, 0.3627, 0.7859, 0.4700, 0.7239, 0.5832, 0.6233]^{\mathsf{T}} \end{aligned}$$

as depicted in Figs. 5–7.

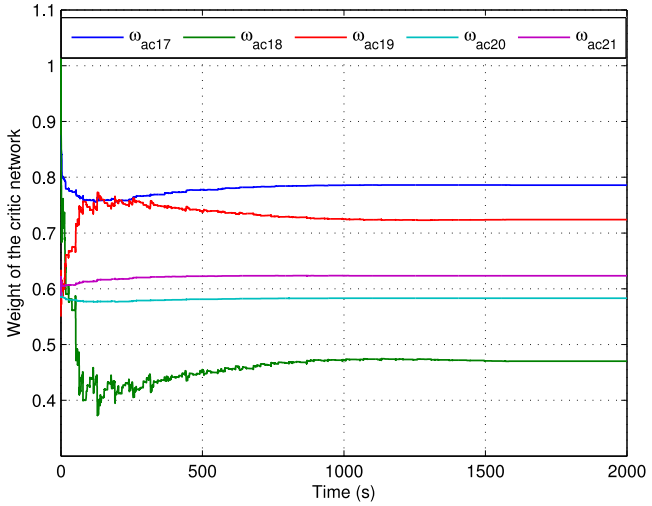


Fig. 7. Convergence of the weight vector of critic network: part 3 (ω_{aci} represents $\hat{\omega}_{ci}$, $i = 17, 18, \ldots, 21$).

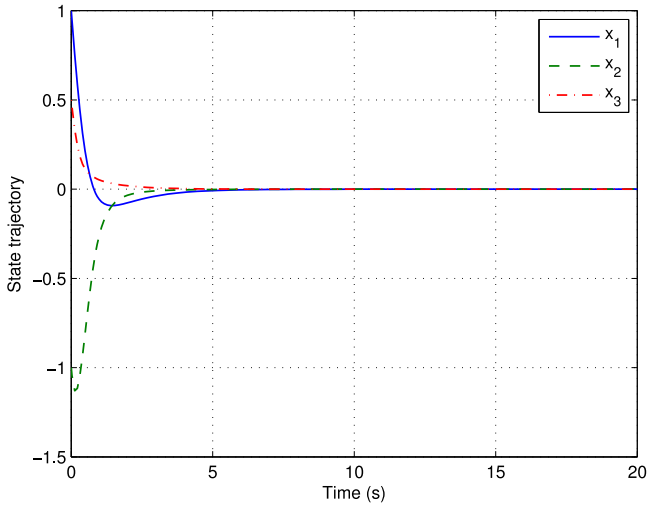


Fig. 8. System state.

At last, the scalar parameters $\delta_1 = -1$, $\delta_2 = 3$, $\delta_3 = -1$, and $\delta_4 = 5$ are chosen for evaluating the robust optimal control performance. The system trajectory is depicted in Fig. 8 when the obtained control law is applied to the uncertain system (36) for 20 s. These simulation results authenticate the validity of the robust optimal control scheme developed in this paper.

VI. Conclusion

A novel adaptive critic learning approach for robust optimal control of a class of uncertain nonlinear systems is developed in this paper, under the framework of data-based ADP. It is proved that the robust controller of the original uncertain system achieves optimality under a specified cost function. Then, the robust optimal control problem is transformed into an optimal control problem. The optimal controller of the nominal system is established without using the system dynamics. The simulation study verifies the good control performance.

As indicated in Remark 6, the developed algorithm of this paper relies on an initial admissible control, which provides a direction for improvement. Although value iteration and policy iteration are two basic algorithms of reinforcement learning, Nodland *et al.* [29] designed an optimal adaptive controller for tracking a trajectory of an unmanned underactuated helicopter forward-in-time without using them. Hence, how to reduce the requirement of the initial admissible control without using value and policy iterations is of great importance. This should be considered in the future research when applying the ADP approach to the framework of nonlinear robust optimal control under uncertain environment. In addition, since the developed approach is only suitable for a class of affine nonlinear systems with matched uncertainties, our future work also contains extending the obtained results to robust optimal control of nonaffine nonlinear systems with unmatched uncertainties.

References

[1] S. Hussain, S. Q. Xie, and P. K. Jamwal, "Robust nonlinear control of an intrinsically compliant robotic gait tranning orthosis," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 43, no. 3, pp. 655–665, May 2013.

[2] H. Gao, X. Meng, and T. Chen, "A new design of robust H_2 filters for uncertain systems," *Syst. Control Lett.*, vol. 57, no. 7, pp. 585–593, Jul. 2008.

[3] Z. Wang and F. T. S. Chan, "A robust replenishment and production control policy for a single-stage production/inventory system with inventory inaccuracy," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 45, no. 2, pp. 326–337, Feb. 2015.

[4] F. Lin, R. D. Brandt, and J. Sun, "Robust control of nonlinear systems: Compensating for uncertainty," *Int. J. Control*, vol. 56, no. 6, pp. 1453–1459, 1992.

[5] F. Lin and R. D. Brandt, "An optimal control approach to robust control of robot manipulators," *IEEE Trans. Robot. Autom.*, vol. 14, no. 1, pp. 69–77, Feb. 1998.

[6] D. Wang, D. Liu, and H. Li, "Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 2, pp. 627–632, Apr. 2014.

[7] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, D. A. White and D. A. Sofge, Eds. New York, NY, USA: Van Nostrand Reinhold, 1992, ch. 13.

[8] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ, USA: Wiley, 2013.

[9] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control: Algorithms and Stability*. London, U.K.: Springer, 2013.

[10] F. L. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, Sep. 2009.

[11] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 943–949, Aug. 2008.

[12] 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.

[13] 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.

[14] S. Bhasin *et al.*, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82–92, Jan. 2013.

[15] H. Modares, M.-B. Naghibi-Sistani, and F. L. Lewis, "A policy iteration approach to online optimal control of continuous-time constrained-input systems," *ISA Trans.*, vol. 52, no. 5, pp. 611–621, Sep. 2013.

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

[17] D. P. Bertsekas, M. L. Homer, D. A. Logan, S. D. Patek, and N. R. Sandell, "Missile defense and interceptor allocation by neuro-dynamic programming," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 30, no. 1, pp. 42–51, Jan. 2000.

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

[19] J. Fu, H. He, and X. Zhou, "Adaptive learning and control for MIMO system based on adaptive dynamic programming," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1133–1148, Jul. 2011.

[20] Z. Ni, H. He, and J. Wen, "Adaptive learning in tracking control based on the dual critic network design," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 6, pp. 913–928, Jun. 2013.

[21] X. Xu, Z. Hou, C. Lian, and H. He, "Online learning control using adaptive critic designs with sparse kernel machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 762–775, May 2013.

[22] Y. Luo, Q. Sun, H. Zhang, and L. Cui, "Adaptive critic design-based robust neural network control for nonlinear distributed parameter systems with unknown dynamics," *Neurocomputing*, vol. 148, pp. 200–208, Jan. 2015.

[23] D. Wang and D. Liu, "Neuro-optimal control for a class of unknown nonlinear dynamic systems using SN-DHP technique," *Neurocomputing*, vol. 121, pp. 218–225, Dec. 2013.

[24] M. Palanisamy, H. Modares, F. L. Lewis, and M. Aurangzeb, "Continuous-time Q-learning for infinite-horizon discounted cost linear quadratic regulator problems," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 165–176, Feb. 2015.

[25] D. Wang, D. Liu, Q. Wei, D. Zhao, and N. Jin, "Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming," *Automatica*, vol. 48, no. 8, pp. 1825–1832, Aug. 2012.

[26] D. Liu, D. Wang, and H. Li, "Decentralized stabilization for a class of continuous-time nonlinear interconnected systems using online learning optimal control approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 418–428, Feb. 2014.

[27] D. Liu, H. Li, and D. Wang, "Online synchronous approximate optimal learning algorithm for multi-player non-zero-sum games with unknown dynamics," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 44, no. 8, pp. 1015–1027, Aug. 2014.

[28] T. Dierks and S. Jagannathan, "Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1118–1129, Jul. 2012.

[29] D. Nodland, H. Zargarzadeh, and S. Jagannathan, "Neural network-based optimal adaptive output feedback control of a helicopter UAV," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1061–1073, Jul. 2013.

[30] A. Heydari and S. N. Balakrishnan, "Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 145–157, Jan. 2013.

[31] J. Liang, G. K. Venayagamoorthy, and R. G. Harley, "Wide-area measurement based dynamic stochastic optimal power flow control for smart grids with high variability and uncertainty," *IEEE Trans. Smart Grid*, vol. 3, no. 1, pp. 59–69, Mar. 2012.

[32] B. Luo, H. N. Wu, T. Huang, and D. Liu, "Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design," *Automatica*, vol. 50, no. 12, pp. 3281–3290, Dec. 2014.

[33] Z.-P. Jiang and Y. Jiang, "Robust adaptive dynamic programming for linear and nonlinear systems: An overview," *Eur. J. Control*, vol. 19, no. 5, pp. 417–425, Sep. 2013.

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

[35] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 882–893, May 2014.

[36] D. Liu, D. Wang, F.-Y. Wang, H. Li, and X. Yang, "Neural-network-based online HJB solution for optimal robust guaranteed cost control of continuous-time uncertain nonlinear systems," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2834–2847, Dec. 2014.

[37] D. Wang, D. Liu, H. Li, and H. Ma, "Neural-network-based robust optimal control design for a class of uncertain nonlinear systems via adaptive dynamic programming," *Inf. Sci.*, vol. 282, pp. 167–179, Oct. 2014.

[38] J. Na and G. Herrmann, "Online adaptive approximate optimal tracking control with simplified dual approximation structure for continuoustime unknown nonlinear systems," *IEEE/CAA J. Autom. Sinica*, vol. 1, no. 4, pp. 412–422, Oct. 2014.

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

[40] D. Liu, H. Li, and D. Wang, "Error bounds of adaptive dynamic programming algorithms for solving undiscounted optimal control problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1323–1334, Jun. 2015.

[41] X. Yang, D. Liu, and Y. Huang, "Neural-network-based online optimal control for uncertain nonlinear continuous-time systems with control constraints," *IET Control Theory Appl.*, vol. 7, no. 17, pp. 2037–2047, Nov. 2013.

[42] B. Luo, H.-N. Wu, and H.-X. Li, "Adaptive optimal control of highly dissipative nonlinear spatially distributed processes with neuro-dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 684–696, Apr. 2015.

[43] D. Wang, D. Liu, D. Zhao, Y. Huang, and D. Zhang, "A neural-network-based iterative GDHP approach for solving a class of nonlinear optimal control problems with control constraints," *Neural Comput. Appl.*, vol. 22, no. 2, pp. 219–227, Feb. 2013.

[44] R. Song, W. Xiao, and C. Sun, "A new self-learning optimal control laws for a class of discrete-time nonlinear systems based on ESN architecture," *Sci. China Inf. Sci.*, vol. 57, no. 6, pp. 1–10, Jun. 2014.

[45] X. Zhong, H. He, H. Zhang, and Z. Wang, "Optimal control for unknown discrete-time nonlinear Markov jump systems using adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2141–2155, Dec. 2014.

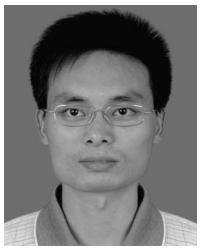
[46] D. Wang, D. Liu, H. Li, H. Ma, and C. Li, "A neural-network-based online optimal control approach for nonlinear robust decentralized stabilization," *Soft Comput.*, vol. 20, no. 2, pp. 707–716, Feb. 2016.

[47] Z. Ni, H. He, X. Zhong, and D. V. Prokhorov, "Model-free dual heuristic dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1834–1839, Aug. 2015.

[48] 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.

[49] X. Yang, D. Liu, and D. Wang, "Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints," *Int. J. Control*, vol. 87, no. 3, pp. 553–566, 2014.

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



Ding Wang (M'15) received the B.S. degree in mathematics from the Zhengzhou University of Light Industry, Zhengzhou, China, in 2007, the M.S. degree in operations research and cybernetics from Northeastern University, Shenyang, China, in 2009, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012.

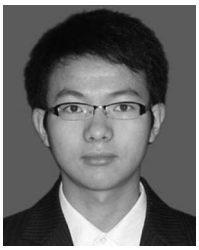
He is currently an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He has published over 50 journal and conference papers, and co-authored two monographs. His current research interests include neural networks, adaptive and learning systems, and complex systems and intelligent control.

Dr. Wang was a recipient of the Excellent Doctoral Dissertation Award of the Chinese Academy of Sciences in 2013, and the nomination of the Excellent Doctoral Dissertation Award of the Chinese Association of Automation (CAA) in 2014. He was the Secretariat of the 2014 IEEE World Congress on Computational Intelligence and the Registration Chair of the 5th International Conference on Information Science and Technology and the 4th International Conference on Intelligent Control and Information Processing. He served as the Program Committee Member of several international conferences. He is the Finance Chair of the 12th World Congress on Intelligent Control and Automation. He has been an Associate Editor of *Neurocomputing* since 2015. He is a member of Asia-Pacific Neural Network Society and CAA.

Derong Liu (S'91–M'94–SM'96–F'05) received the Ph.D. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 1994.

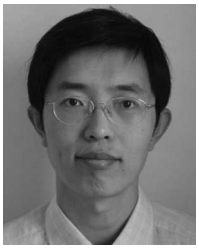
He was a Staff Fellow with General Motors Research and Development Center, Detroit, MI, USA, from 1993 to 1995. He was an Assistant Professor with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA, from 1995 to 1999. He joined the University of Illinois at Chicago, Chicago, IL, USA, in 1999, and became a Full Professor of Electrical and Computer Engineering and Computer Science in 2006. He served as the Associate Director of the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Beijing, China, from 2010 to 2015. He is currently a Full Professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing. He has published 15 books (six research monographs and nine edited volumes).

Dr. Liu was a recipient of the Faculty Early Career Development Award from the National Science Foundation in 1999, the University Scholar Award from the University of Illinois from 2006 to 2009, the Overseas Outstanding Young Scholar Award from the National Natural Science Foundation of China in 2008, and the Outstanding Achievement Award from Asia Pacific Neural Network Assembly in 2014. He was selected for the "100 Talents Program" by the Chinese Academy of Sciences in 2008. He is an elected AdCom Member of the IEEE Computational Intelligence Society and he is the Editor-in-Chief of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. He was the General Chair of the 2014 IEEE World Congress on Computational Intelligence and is the General Chair of the 2016 World Congress on Intelligent Control and Automation. He is a Fellow of the International Neural Network Society.



Qichao Zhang received the B.S. degree in automation from Northeastern Electric Power University, Jilin, China, in 2012, and the M.S. degree in control theory and control engineering from Northeast University, Shenyang, China, in 2014. He is currently pursuing the Ph.D. degree in control theory and control engineering with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His current research interests include reinforcement learning, game theory, and multiagent systems.



Dongbin Zhao (M'06–SM'10) received the Ph.D. degree in materials processing engineering from the Harbin Institute of Technology, Harbin, China, in 2000.

He was a Post-Doctoral Fellow with Tsinghua University, Beijing, China, from 2000 to 2002. He has been an Associate Professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, since 2002, where he has been a Professor with the State Key Laboratory of Management and Control for Complex Systems since 2012. He has published four books, and published over 50 international journal papers. His current research interests include computational intelligence, adaptive dynamic programming, robotics, intelligent transportation systems, and smart grids.

Dr. Zhao has been an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS since 2012 and the IEEE COMPUTATION INTELLIGENCE MAGAZINE since 2014. Since 2015, he has been the Chair of Adaptive Dynamic Programming Technical Committee, Multimedia Subcommittee, and Travel Grant Subcommittee of the IEEE Computational Intelligence Society. He was a Guest Editor of several international journals. He is involved in organizing several international conferences.