UKAEA-RACE-PR(21)02

Jing Na, Jun Zhao, Kaiqiang Zhang, Yongfeng Lv

Adaptive Identifier-Critic-Based Optimal Tracking Control for Nonlinear Systems With Experimental Validation

Adaptive Identifier-Critic-Based Optimal Tracking Control for Nonlinear Systems With Experimental Validation

Jing Na, Jun Zhao, Kaiqiang Zhang, Yongfeng Lv

This is a preprint of a paper submitted for publication in
IEEE Transactions on Systems, Man, and Cybernetics: Systems

Adaptive Identifier-Critic-Based Optimal Tracking Control for Nonlinear Systems With Experimental Validation

Jing Na, *Member, IEEE*, Yongfeng Lv, *Graduate Student Member, IEEE*, Kaiqiang Zhang, *Member, IEEE*, and Jun Zhao, *Graduate Student Member, IEEE*

***Abstract*—This article presents and practically validates an identifier-critic-based approximate dynamic programming (ADP) method to online address the optimal tracking control problem for nonlinear continuous-time unknown systems. The imposed assumption on precisely known system dynamics is obviated via a neural network (NN) identifier. A static control is first adopted to retain the steady-state tracking response, while an optimal control derived via the ADP method is proposed to regulate the tracking error by minimizing a cost function. A critic NN is then trained online to obtain the solution of the associated Hamilton–Jacobi–Bellman (HJB) equation. The learning of the identifier NN and critic NN is performed online simultaneously by tailoring a novel adaptation method, which can guarantee the convergence of the estimated NN weights. Consequently, the critic NN can be used to construct the optimal control policy directly, such that the actor NN used in the previous ADP schemes is avoided. Simulations are performed to verify the suggested control, and experiments on a helicopter plant are carried out to show its feasibility and improved control response.**

***Index Terms*—Adaptive dynamic programming, adaptive control, neural network (NN), optimal control.**

I. Introduction

THE PRIMARY tracking control design objective is to find proper control actions such that the state (or output) of a system can track a given trajectory. Among different control schemes, adaptive control [1] has been developed for systems with uncertain parameters. To minimize a predefined cost function, optimal control [2] has been derived based on the Pontryagin's minimum principle. Although it is practically useful, classical optimal control was solved offline based on fully known system dynamics [3]. Hence, some effort has been recently made to use adaptive techniques in the optimal control synthesis. Specifically, optimal control can also be solved via the reinforcement learning (RL) algorithms [4]. In this line, Werbos [5] introduced an RL-based approximate dynamic programming (ADP) with a critic–actor framework, which employs two neural networks (NNs) to obtain optimal control solutions. This idea has been further developed for solving various optimal control problems [2], [3], [6]–[14].

The ADP approaches have initially been developed for discrete-time (DT) systems [10], [15]–[20], whereas it is a nontrivial challenge to develop ADP control algorithms for continuous time (CT) systems [2]–[4], [6], [21]. Abu-Khalaf and Lewis [22] proposed an offline ADP scheme to design an approximately optimal control. In [3] and [23], an integral RL-based online policy iteration (PI) was developed, where a critic NN and an actor NN are trained sequentially. A synchronous ADP algorithm [24] was further proposed, with both the critic and actor NNs being trained simultaneously. To relax assumptions on accurate system dynamics, an identifier-critic–actor ADP structure was presented using an identifier to estimate the unknown drift dynamics [25]. It is noted that the convergence of identifier NN weights was not addressed, though the identifier output error converges to zero in [25]. The work of [26] further relaxed the assumption on the input dynamics and presented a new adaptation in terms of the experience replay method. Note that an extra actor NN is used in these ADP schemes to derive control actions, thereby, imposing demanding computational costs. In parallel, Jiang and Jiang [27], [28] presented a new PI method, which solves the algebraic Riccati equation (ARE) to find an optimal control of linear CT systems without the need for NNs. Recently, an off-policy RL scheme [29] was also used to solve robust control of uncertain systems using an optimal control method. However, all of the above results address optimal regulation problem only.

Compared with the optimal regulation problem, the tracking control is more complicated due to its noncausal property [2], [30]. The inverse optimal control [31], [32] was studied for specific systems to minimize the cost function without solving the Hamilton–Jacobi–Bellman (HJB) equation. Alternatively, an NN approximator was used in the optimal control design of linear systems [33]. However, ADP-based optimal tracking control of nonlinear CT systems has been less developed in comparison to the optimal regulation. In

Manuscript received February 8, 2019; revised December 24, 2019 and May 5, 2020; accepted June 8, 2020. This work was supported by the National Natural Science Foundation of China under Grant 61922037 and Grant 61873115. This article was recommended by Associate Editor T.-M. Choi. *(Corresponding author: Jing Na.)*

Jing Na and Jun Zhao are with the Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650500, China (e-mail: najing@163.com; junzhao1993@163.com).

Yongfeng Lv is with the School of Automation, Beijing Institute of Technology, Beijing 100081, China (e-mail: lvyilian1989@foxmail.com).

Kaiqiang Zhang is with the Remote Applications in Challenging Environments, U.K. Atomic Energy Authority, Abingdon OX14 3DB, U.K. (e-mail: kaiqiang.zhang@ukaea.uk).

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

Digital Object Identifier 10.1109/TSMC.2020.3003224

fact, the existing ADP solutions for optimal tracking control can be mainly divided in two categories: In [34] and [35], linear quadratic output tracking of linear systems was solved by using a system augmentation approach, where the tracking error dynamics and command generator are merged into one augmented system. However, extension of this idea to fully unknown nonlinear systems is challenging. Another solution for optimal tracking control was given in [36], where a static control is used to transform the tracking control of the original system into the optimal regulation of an induced error system. In this line, the work of [37] combined the system augmentation and steady-state control to present an optimal tracking control for uncertain systems. The subsequent work [38] adopted a concurrent-earning-based identifier for tracking control of nonlinear systems with unknown drift dynamics. In [39], Q-learning was incorporated into the linear virtual reference feedback tuning to achieve a model-free tracking control. An observer was used together with the critic–actor-based ADP control in [40] to remedy the unknown system state and dynamics. However, these optimal tracking controllers all depend on the identifier-critic–actor-based ADP structure. The other issue is these approaches require an extra actor NN to prove the closed-loop stability, because the convergence of the identifier NN weights is not addressed for the adaptive laws designed via the gradient descent algorithm [36]. Hence, this ADP structure with triple approximators has slow convergence speed and heavy computational costs. This problem was tackled in our previous work [41] by developing a dual approximator-based ADP method. However, the input dynamics of the system are assumed to be a known constant in [41].

This article proposes and practically validates a simplified ADP framework based on an identifier NN and a critic NN only, motivated to solve the optimal tracking control problem of nonlinear systems with fully unknown dynamics. Specifically, an error feedback term is used to enhance the convergence in this work, therefore, remedying previously imposed assumptions on the input dynamics [41]. Then a composite control with a steady-state control and an optimal control is suggested. First, an NN identifier is constructed to online estimate the unknown dynamics. The identified dynamics are used to retain the steady-state tracking response, while the optimal control is then proposed to regulate the control error and minimize a cost function. To realize this optimal control, a critic NN is trained online to solve the derived optimal equation. Both the identifier NN and critic NN can be trained simultaneously by using adaptive algorithms with the obtained estimation error [42]. Since this adaptation is designed to ensure the convergence of the NN weights to unknown ideal values, the critic NN can be utilized to directly derive the optimal control action. Therefore, the widely used actor NN is avoided, resulting in an effective ADP framework with dual approximators, reduced computational burden, and fast convergence. Consequently, it can be seen that the adaptive laws presented in this article are different to the Least-Squares [25] or Levenberg–Marquardt algorithms [24] in other ADP schemes. Simulations are given to verify the proposed methods. More specifically, practical experiments are carried out on a Quanser helicopter to exemplify the proposed ADP method and demonstrate its improved control performance.

In brief, this article has the following contributions in comparison to other ADP methods.

1) An identifier-critic-based ADP scheme is used. This scheme has a dual approximation structure, resulting in reduced computational costs. Since the identifier and critic NNs are trained via a novel adaptation rather than the classical gradient schemes [24], [25], [36], the convergence of critic NN weights can be retained. Thus, the actor NN used in the existing identifier-critic–actor-based ADP structure [25], [36] is avoided.
2) The previously imposed assumptions on the accurate model of drift dynamics [3], [24], [37] and input dynamics [36], [38], [41] are obviated. This work suggests an improved identifier over previous work [41].
3) Practical experiments are given to exemplify the presented ADP method, besides numerical simulations. It is found that the steady-state control with an optimal compensation derived via ADP can achieve better tracking performance.

The problem formulation is shown in Section II. The NN identifier is designed in Section III. Section IV gives the adaptive optimal control design. Simulations and experimental results are provided in Sections V and VI to validate the proposed control method, respectively. Section VII draws some conclusions.

II. Problem Formulation and Preliminaries

A class of nonlinear affine multi-input multi-output (MIMO) systems are considered as

$$\dot{x} = f(x) + g(x)u \tag{1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$ are the system state and control vectors; $f(x) \in \mathbb{R}^n$ and $g(x) \in \mathbb{R}^{n\times m}$ are unknown functions.

The aim of this article is to design a controller $u(t)$, which not only guarantees that the state $x(t)$ of system (1) tracks a given reference trajectory $x_d(t)$ but also makes the tracking error converge to zero in an optimal manner, i.e., to minimize a predefined performance index [36] given by

$$V(e(t)) = \int_t^\infty r(e(\tau), u(\tau))d\tau \tag{2}$$

where $e = x - x_d$ denotes the control error, $r(\bullet, \bullet) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ with $r(e(\tau), u(\tau)) \geq 0$ is the utility function; $x_d(t)$ and $\dot{x}_d(t)$ are bounded references to be tracked; $f(x) + g(x)u$ is Lipschitz on a compact set Ω, thus system (1) is stabilizable [24]. The tracking error $e(t)$ is utilized in the cost function (2) as this article studies optimal tracking control problem. In order to address the unknown dynamics in system (1), we first propose an NN identifier, and then use the reconstructed dynamics to design a controller that realize an effective composite control with a new ADP scheme.

Definition 1 [1]: A vector or matrix ϕ is persistently excited (PE) if there exist constants $\tau > 0, \varepsilon > 0$ such that $\int_t^{t+\tau} \phi(r)\phi^T(r)dr \geq \varepsilon I, \ \forall t \geq 0$.

Throughout this article, $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ represent the maximum and minimum eigenvalues of matrices.

III. Adaptive NN-Based Identification

To handle the unknown dynamics in system (1), an adaptive identifier is proposed. Without loss of generality, the unknown dynamics $f(x)$, $g(x)$ are assumed to be smooth functions on a compact set Ω, and thus approximated via NNs [43]–[45] as

$$f(x) = \theta\xi(x) + \varepsilon_f, \quad g(x) = \psi\varsigma(x) + \varepsilon_g \tag{3}$$

with $\theta \in \mathbb{R}^{n\times k_\theta}$, $\psi \in \mathbb{R}^{n\times k_\psi}$ the ideal NN weights, $\xi \in \mathbb{R}^{k_\theta}$, $\varsigma \in \mathbb{R}^{k_\psi \times m}$ the regressors, and ε_f, ε_g the approximation errors.

Substituting the NN approximation (3) into system (1), we can rewrite system (1) in a compact form as

$$\dot{x} = W_1^T\phi_1(x,u) + \varepsilon_T \tag{4}$$

with $W_1 = [\theta, \psi]^T \in \mathbb{R}^{d\times n}$ the augmented weights matrix, $\phi_1(x,u) = [\xi^T(x), u^T\varsigma^T(x)]^T \in \mathbb{R}^d$ the augmented regressor for $d = k_\theta + k_\psi$, and $\varepsilon_T = \varepsilon_f + \varepsilon_g u$ the augmented NN error. Note that the NNs can also be replaced by fuzzy systems [46]–[48]. In practice, the basis function in the regressors ξ, ς can be implemented by sigmoid functions. For specific systems where the unknown nonlinearities can be formulated in the linearly parameterized form as (3), the regressors can be designed based on the plant information as shown in the case studies (see simulations in Section V).

As shown in [25], we know that the NN regressors ξ, ς, and approximation errors ε_f, ε_g are all bounded. Based on Weierstrass theorem [22], [24], it is true that the NN errors ε_f and ε_g will vanish as the number of neurons increases (i.e., $k_\theta, k_\psi \to \infty$).

Remark 1: Some recent works have reported several adaptive identifier designs for system (4), e.g., [25], [36] and references therein. However, the adaptive laws used to update the NN weights are driven by the identifier output error (the error between x and the identifier output $\hat{x}$), so that the convergence of identifier NN weights cannot be retained. This article will address this issue by developing a new adaptive law that guarantees the convergence and simplifies the control design.

We first impose filtering operations on the measured x and ϕ_1 in (4) to obtain the filtered variables x_f and ϕ_{1f} as [49]

$$\begin{cases} k\dot{x}_f + x_f = x \\ k\dot{\phi}_{1f} + \phi_{1f} = \phi_1 \end{cases} \tag{5}$$

where $k > 0$ is a filter coefficient.

Based on (4) and (5), it can be verified

$$\dot{x}_f = \frac{x - x_f}{k} = W_1^T\phi_{1f} + \varepsilon_{Tf} \tag{6}$$

where ε_{Tf} is the filtered bounded error $k\dot{\varepsilon}_{Tf} + \varepsilon_{Tf} = \varepsilon_T$.

The auxiliary matrices $P_1 \in \mathbb{R}^{d\times d}$ and $Q_1 \in \mathbb{R}^{d\times n}$ now can be calculated as

$$\begin{cases} \dot{P}_1 = -\ell_1 P_1 + \phi_{1f}\phi_{1f}^T, & P_1(0) = 0 \\ \dot{Q}_1 = -\ell_1 Q_1 + \phi_{1f}\left[\frac{x - x_f}{k}\right]^T, & Q_1(0) = 0 \end{cases} \tag{7}$$

where $\ell_1 > 0$ is a forgetting factor to guarantee the boundedness of P_1 and Q_1.

The aim to introduce the filters in (5) and variables P_1, Q_1 is to online calculate the matrix $M_1 \in \mathbb{R}^{d\times n}$ as

$$M_1 = P_1\hat{W}_1 - Q_1. \tag{8}$$

With the matrix M_1, the NN weights can be estimated by

$$\dot{\hat{W}}_1 = -\Gamma_1 M_1 \tag{9}$$

with $\Gamma_1 > 0$ is the learning gain. The motivation of introducing the auxiliary matrices P_1 and Q_1 is to extract the estimation error variable M_1, which enables the adaptive law (9) to retain convergence of the estimated NN weighted, using the known system dynamics x, ϕ_1.

The PE condition is necessary for guaranteeing the convergence of the adaptive laws in [22], [24], [25], and [36]. In this work, the relationship between the PE condition and the positive definiteness of matrix P_1 is first investigated as follows.

Lemma 1: Suppose the augmented regressor ϕ_1 is PE, then P_1 is positive definite, i.e., there exists a constant $\sigma_1 > 0$ such that $\lambda_{\min}(P_1) > \sigma_1$.

Proof: The transfer function of (5) is given by $1/(ks + 1)$, which is stable, minimum phase and strictly proper. Consequently, the PE property of the filtered regressor ϕ_{1f} equals to the PE property of ϕ_1 [1].

Suppose ϕ_1 is PE, indicating ϕ_{1f} is PE, then from Definition 1, the PE condition $\int_t^{t+\tau}\phi_{1f}^T(r)\phi_{1f}(r)dr \geq \varepsilon I$ is equivalent to $\int_{t-\tau}^{t}\phi_{1f}^T(r)\phi_{1f}(r)dr \geq \varepsilon I$ for $t > \tau > 0$. Hence, the following inequality is true:

$$e^{-\ell_1\tau}\int_{t-\tau}^{t}\phi_{1f}^T(r)\phi_{1f}(r)dr \geq e^{-\ell_1\tau}\varepsilon I. \tag{10}$$

Within the time interval $r \in [t-\tau, t]$, we know $t - r \leq \tau$, and thus $e^{-\ell_1(t-r)} \geq e^{-\ell_1\tau} > 0$ holds, such that

$$\begin{aligned}\int_{t-\tau}^{t} e^{-\ell_1(t-r)}\phi_{1f}^T(r)\phi_{1f}(r)dr &\geq \int_{t-\tau}^{t} e^{-\ell_1\tau}\phi_{1f}^T(r)\phi_{1f}(r)dr \\ &\geq e^{-\ell_1\tau}\varepsilon I.\end{aligned} \tag{11}$$

Moreover, it can be verified for all $t > \tau > 0$ that

$$\int_0^t e^{-\ell_1(t-r)}\phi_{1f}^T(r)\phi_{1f}(r)dr > \int_{t-\tau}^{t} e^{-\ell_1(t-r)}\phi_{1f}^T(r)\phi_{1f}(r)dr. \tag{12}$$

From (11) and (12), one can conclude that

$$\begin{aligned}P_1 = \int_0^t e^{-\ell_1(t-r)}\phi_{1f}^T(r)\phi_{1f}(r)dr &> e^{-\ell_1\tau}\int_{t-\tau}^{t}\phi_{1f}^T(r)\phi_{1f}(r)dr \\ &\geq e^{-\ell_1\tau}\varepsilon I.\end{aligned} \tag{13}$$

Hence, P_1 is positive definite, therefore, $\lambda_{\min}(P_1) > \sigma_1 > 0$ is true with $\sigma_1 = e^{-\ell_1\tau}\varepsilon$. This finishes the proof. ■

Remark 2: Although the PE condition is well-recognized in the ADP literatures [22], [24], [25], [36], it is difficult to test the PE condition directly [1]. Lemma 1 shows that the PE condition implies the positive definite property of P_1. Therefore, it is possible to online validate the PE property by assessing whether $\lambda_{\min}(P_1) > \sigma_1 > 0$. This online verifiable condition $\lambda_{\min}(P_1) > \sigma_1 > 0$ is used in this article and it also provides a feasible method to test the PE condition.

However, Lemma 1 does not necessarily show the relaxation of the PE condition, though this can also be done by using the concurrent-learning or experience replay (see [26], [38]).

The main conclusion of this section is presented as follows.

Theorem 1: For nonlinear system (4) with the adaptive algorithm (9), if ϕ_1 satisfies the PE condition, the identifier weights error $\tilde{W}_1 = W_1 - \hat{W}_1$ converges to a small set around zero. Moreover, in the absence of NN approximation errors (i.e., $\varepsilon_T = 0$), $\tilde{W}_1$ converges to zero.

Proof: We first derive the solution of (7) as

$$\begin{cases} P_1(t) = \int_0^t e^{-\ell_1(t-r)}\phi_{1f}(r)\phi_{1f}^T(r)dr \\ Q_1(t) = \int_0^t e^{-\ell_1(t-r)}\phi_{1f}(r)\left[\frac{x(r)-x_f(r)}{k}\right]^T dr. \end{cases} \tag{14}$$

Then from (6)–(14), it can be validated that

$$Q_1 = P_1 W_1 - \upsilon_1 \tag{15}$$

with $\upsilon_1 = -\int_0^t e^{-\ell_1(t-r)}\phi_{1f}(r)\varepsilon_{Tf}^T(r)dr$ being a bounded term since the regressor ϕ_{1f} and error ε_T are all bounded, that is $\|\upsilon_1\| \le \varepsilon_{\upsilon 1}$ for $\varepsilon_{\upsilon 1} > 0$.

From (8) and (15), one can verify

$$M_1 = P_1\hat{W}_1 - P_1 W_1 + \upsilon_1 = -P_1\tilde{W}_1 + \upsilon_1. \tag{16}$$

Here, the matrix M_1 contains the information of estimation error $\tilde{W}_1$. In this sense, the proposed adaptive law is an improved gradient algorithm to minimize the estimation error rather than the observer error used in other ADP schemes. Given a Lyapunov function $V_1 = tr(\tilde{W}_1^T\Gamma_1^{-1}\tilde{W}_1)/2$, $\dot{V}_1$ is obtained along (9) and (16) as

$$\begin{aligned} \dot{V}_1 &= tr\left(\tilde{W}_1^T\Gamma_1^{-1}\dot{\tilde{W}}_1\right) = -tr\left(\tilde{W}_1^T P_1\tilde{W}_1\right) + tr\left(\tilde{W}_1^T\upsilon_1\right) \\ &\le -\left\|\tilde{W}_1\right\|\left(\sigma_1\left\|\tilde{W}_1\right\| - \varepsilon_{\upsilon 1}\right). \end{aligned} \tag{17}$$

Based on the Lyapunov Theorem, the error $\tilde{W}_1$ converges to a small set around zero $\Omega_1 : \{\tilde{W}_1 | \|\tilde{W}_1\| \le \varepsilon_{\upsilon 1}/\sigma_1\}$, whose ultimate bound is determined by the NN error ε_T and σ_1.

Moreover, in the ideal case of $\varepsilon_T = 0$ and thus $\upsilon_1 = 0$, (17) is reduced to

$$\dot{V}_1 = -tr\left(\tilde{W}_1^T P_1\tilde{W}_1\right) < -\sigma_1\left\|\tilde{W}_1\right\|^2 \le -\mu_1 V_1 \tag{18}$$

with $\mu_1 = 2\sigma_1/\lambda_{\max}(\Gamma_1^{-1})$ a positive constant. Equation (18) implies that $\tilde{W}_1$ exponentially converges to zero. ■

Remark 3: As shown in the proof of Theorem 1 [e.g., (16)], the variable M_1 defined in (8) includes the estimation error $\tilde{W}_1$ with a bounded residual error υ_1. This residual error υ_1 will vanish by using sufficient neurons in the identifier NN (4) as shown in [22]. Thus, in this article, M_1 is adopted to design the adaptive law (9) to achieve the convergence of $\hat{W}_1$ to W_1.

IV. Optimal Tracking Control Design

The optimal tracking control is designed by incorporating the identified dynamics into the ADP synthesis. Therefore, system (1) can be rewritten as

$$\dot{x} = \hat{\theta}\xi(x) + \hat{\psi}\varsigma(x)u + \varepsilon_N + \varepsilon_T \tag{19}$$

where $\hat{\theta}$ and $\hat{\psi}$ are the estimates of θ and ψ. Here, $\hat{\theta}$ and $\hat{\psi}$ can be derived from $\hat{W}_1$ given by the adaptive law (9).

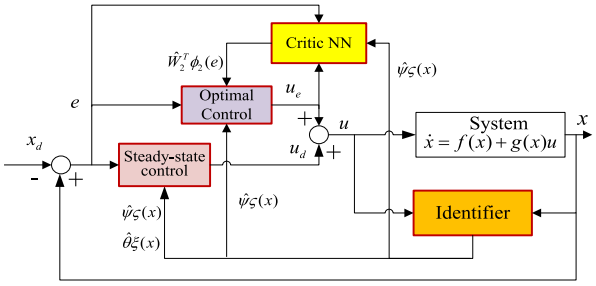

Fig. 1. Designed ADP-based control system structure.

$\varepsilon_N = \tilde{W}_1\phi_1$ denotes the identifier error, which is also bounded as $\tilde{W}_1$ is bounded (proved in Theorem 1) and the NN regressor ϕ_1 is bounded. Thus, we have $\|\varepsilon_N\| \le \eta_N$ for a positive constant η_N.

For the purpose of tracking control design, the tracking error can be defined as $e = x - x_d$, so that

$$\dot{e} = \dot{x} - \dot{x}_d = \hat{\theta}\xi(x) + \hat{\psi}\varsigma(x)u + \varepsilon_N + \varepsilon_T - \dot{x}_d. \tag{20}$$

The objective of tracking control is to design a controller such that $e \to 0$ in an optimal manner. Thus, a composite control action u with two parts [36], [41] can be designed

$$u = u_d + u_e \tag{21}$$

where u_d is used to retain the steady-state performance, which should compensate the undesired system dynamics, and u_e is used to regulate the tracking error by minimizing a given cost function. The schematic of the proposed identifier-critic-based optimal control system is shown in Fig. 1.

A. Steady-State Control

As aforementioned, u_d should be designed to guarantee an ideal tracking performance $e = x - x_d = 0$ for (20). Hence, u_d needs to compensate for the effect of $\hat{\psi}\varsigma(x)$, $\hat{\theta}\xi(x)$, and it can be designed as

$$u_d = \hat{g}^+(x)\left[\dot{x}_d - \hat{\theta}\xi(x) - K_e e\right] \tag{22}$$

with $K_e > 0$ the feedback gain, $\hat{g}^+(x) = ([\hat{\psi}\varsigma(x)]^T\hat{\psi}\varsigma(x))^{-1}[\hat{\psi}\varsigma(x)]^T$ is the Moore–Penrose inverse of matrix $\hat{\psi}\varsigma(x)$ as used in [37] and [38]. To avoid the potential singularity in calculating $\hat{g}^+(x)$, the projection operator in [25] can be applied to the adaptive law (9), such that $\hat{\psi}\varsigma(x) \ne 0$ is retained. Clearly, u_d can be obtained based on e, $\dot{x}_d$ and $\hat{\psi}\varsigma(x)$, $\hat{\theta}\xi(x)$ from identifier (4).

Then by substituting (22) into (20), the dynamics of tracking error e are described as

$$\dot{e} = -K_e e + \hat{\psi}\varsigma(x)u_e + \varepsilon_N + \varepsilon_T. \tag{23}$$

With the help of steady-state control (22), the tracking error in (20) is simplified to (23). Consequently, the optimal tracking control for system (19) is reformulated to the regulation of system (23) by using an optimal control u_e. In comparison to previous work [41], the imposed assumption on the input dynamics $g(x)$ is removed and a feedback term $K_e e$ is added in the steady-state control to enhance convergence.

B. Optimal Control

In this section, a simplified ADP framework with a critic NN only will be introduced to design u_e to stabilize (23). As explained above, the steady-state control u_d can compensate for the undesired dynamics $\hat{\psi}\varsigma(x), \hat{\theta}\xi(x)$ and reduce (20) into (23). Therefore, the cost function (2) can be reformulated as

$$V(e) = \int_t^{\infty} r(e(\tau), u_e(e(\tau)))d\tau \tag{24}$$

with $r(e, u_e) = e^T Qe + u_e^T Ru_e$ the utility function of the tracking error e and optimal control u_e, where Q and R are the weighting matrices.

To solve the optimal control problem of system (23), the Hamiltonian is given as

$$H(e, u_e, V) = V_e^T\Big[-K_e e + \hat{\psi}\varsigma(x)u_e + \varepsilon_N + \varepsilon_T\Big] + e^T Qe + u_e^T Ru_e \tag{25}$$

where $V_e \triangleq \partial V/\partial e$ is the partial derivative of V.

Then based on the optimality principle in [2], an optimal cost function V^* can be denoted by

$$V^*(e) = \min_{u_e \in \Psi(\Omega)} \left(\int_t^{\infty} r(e(\tau), u_e(e(\tau)))d\tau \right) \tag{26}$$

which must satisfy the following HJB equation:

$$0 = \min_{u_e \in \Psi(\Omega)} \big[H\big(e, u_e^*, V^*\big)\big]. \tag{27}$$

The optimal control u_e^* is then derived based on (25)–(27) by using $\partial H(e, u_e^*, V^*)/\partial u_e^* = 0$ as

$$u_e^* = -\frac{1}{2}R^{-1}\Big[\hat{\psi}\varsigma(x)\Big]^T \frac{\partial V^*(e)}{\partial e}. \tag{28}$$

To implement optimal control (28), the HJB (27) should be solved to provide V^*. However, it is generally difficult to solve (27) as it is a nonlinear partial differential equation (PDE). Inspired by the ADP principle [3], [22]–[25], a critic NN is used to estimate V^*. Assuming V^* is a continuous function on the compact set Ω, V^* can be approximated by a critic NN in the following way:

$$V^*(e) = W_2^T \phi_2(e) + \varepsilon_v \tag{29}$$

where its partial derivative $\partial V^*(e)/\partial e$ is given by

$$\frac{\partial V^*(e)}{\partial e} = \nabla\phi_2^T W_2 + \nabla\varepsilon_v. \tag{30}$$

Here, $W_2 \in \mathbb{R}^l$and $\phi_2(e) \in \mathbb{R}^l$ are the critic NN weights and regressor vector. ε_v is the critic NN error. l is the number of NN nodes. $\nabla\phi_2 = \partial\phi_2/\partial e$ and $\nabla\varepsilon_v = \partial\varepsilon_v/\partial e$ are the partial derivatives of ϕ_2 and ε_v.

By using (28) and (29), u_e^* can be given as

$$u_e^* = -\frac{1}{2}R^{-1}\Big[\hat{\psi}\varsigma(x)\Big]^T \big(\nabla\phi_2^T W_2 + \nabla\varepsilon_v\big). \tag{31}$$

Assumption 1 [22]: The NN weights W_2, regressors $\phi_2(\bullet)$ and $\nabla\phi_2(\bullet)$ are bounded by $\|W_2\| \le W_N$, $\|\phi_2\| \le \phi_N$, $\|\nabla\phi_2\| \le \phi_M$. The error $\nabla\varepsilon_v$ is also bounded by $\|\nabla\varepsilon_v\| \le \phi_\varepsilon$.

In the control implementation, the regressor vector ϕ_2 can be selected such that its components $\{\phi_{2i}(e) : i = 1, \ldots, l\}$ provide an independent basis of V^*. Based on the Weierstrass theorem [22], [24], it is true that the approximation error and its partial derivative $\varepsilon_v, \nabla\varepsilon_v \to 0$ in the critic NN (29), (30) for $l \to +\infty$.

Since the critic NN weights W_2 in (29) are unknown, the practical critic NN output $\hat{V}(e)$ to estimate $V^*(e)$ is

$$\hat{V}(e) = \hat{W}_2^T \phi_2(e) \tag{32}$$

where $\hat{W}_2$ is the estimated NN weights to be updated via the online adaptation algorithm to be presented.

Then from (28) and (32), the practical optimal control is described by

$$\begin{aligned}\hat{u}_e &= -\frac{1}{2}R^{-1}[\hat{\psi}\varsigma(x)]^T \frac{\partial\hat{V}(e)}{\partial e} \\ &= -\frac{1}{2}R^{-1}\Big[\hat{\psi}\varsigma(x)\Big]^T \nabla\phi_2^T(e)\hat{W}_2\end{aligned} \tag{33}$$

with $\partial\hat{V}(e)/\partial e = \nabla\phi_2^T\hat{W}_2$ the derivative of the critic NN output.

Consequently, the practical composite controller u is

$$u = u_d + \hat{u}_e \tag{34}$$

where u_d is the steady-state control given in (22) and $\hat{u}_e$ is the optimal control defined in (33).

Remark 4: In most of existing ADP control designs for (23) (e.g., [22]–[25] and references therein), the classical critic–actor structure is utilized, where a critic NN and an actor NN are used to approximate the value function and the control policy, respectively. Hence, they need fairly long transient to achieve convergence and significant computational costs. Moreover, most of these approaches either run offline [22], [23] or assume that the system dynamics are known [24], [25]. In contrary to these approaches, the ADP framework proposed in this article uses the critic NN only to derive the optimal control as (33). This could reduce the computational cost since the actor NN is not needed. Moreover, both the identifier and critic NNs are updated simultaneously in this article, leading to a synchronous implementation rather than the sequential scheme in [3].

C. Updating Critic NN Weights

The final problem to be addressed is to present an online learning scheme to derive the critic NN weights $\hat{W}_2$ for control (33). For facilitating the following developments, the HJB equation (25) with (30) is rewritten as:

$$0 = W_2^T\nabla\phi_2\Big[-K_e e + \hat{\psi}\varsigma(x)u_e\Big] + e^T Qe + u_e^T Ru_e + \varepsilon_{HJB} \tag{35}$$

where $\varepsilon_{HJB} = W_2^T\nabla\phi_2(\varepsilon_N + \varepsilon_T) + \nabla\varepsilon_v(-K_e e + \hat{\psi}\varsigma u_e + \varepsilon_N + \varepsilon_T)$ is the residual HJB equation error, which is bounded and sufficiently small with sufficient amount of NN nodes in the critic NN [22], [24]. This is because $\varepsilon_N, \varepsilon_T \to 0$ holds if we set $k_\theta, k_\psi \to +\infty$, and $\nabla\varepsilon_v \to 0$ is true if we set $l \to +\infty$. It can also be found from (35) that the convergence of $\tilde{W}_1$ is essential for the convergence of $\hat{W}_2$ since the identifier error ε_N is involved in the residual error term ε_{HJB}.

Define $\Xi = \nabla\phi_2[-K_e e + \hat{\psi}\varsigma(x)u_e]$ and $\Theta = e^T Q e + u_e^T R u_e$, so that the HJB (35) is reformulated as

$$\Theta = -W_2^T \Xi - \varepsilon_{HJB}. \tag{36}$$

It can be found from (36) that the weights W_2 to be updated appear in a parameterized formulation. Therefore, the adaptation design proposed in Section III can be further tailored to estimate W_2 rather than to minimize the HJB equation error ε_{HJB} by using the gradient method in [24], [25], and [36]. In this case, the convergence of $\hat{W}_2$ to W_2 can be proved.

The auxiliary matrix $P_2 \in \mathbb{R}^{l\times l}$ and vector $Q_2 \in \mathbb{R}^l$ can be calculated as

$$\begin{cases} \dot{P}_2 = -\ell_2 P_2 + \Xi\Xi^T, & P_2(0) = 0 \\ \dot{Q}_2 = -\ell_2 Q_2 + \Xi\Theta, & Q_2(0) = 0 \end{cases} \tag{37}$$

with $\ell_2 > 0$ the forgetting factor as used in (7).

The adaptive law for updating $\hat{W}_2$ is represented by

$$\dot{\hat{W}}_2 = -\Gamma_2 M_2 \tag{38}$$

with $\Gamma_2 > 0$ the learning gain. The vector $M_2 \in \mathbb{R}^l$ is derived from the variables P_2 and Q_2 in (37) by

$$M_2 = P_2\hat{W}_2 + Q_2. \tag{39}$$

Similar to Lemma 1, the following lemma is true.

Lemma 2: The PE condition of the regressor Ξ in (36) implies the positive definiteness of matrix P_2 in (37), that is $\lambda_{\min}(P_2) > \sigma_2 > 0$ for a constant σ_2.

The convergence of adaptive law (38) is given as

Theorem 2: Considering adaptive algorithm (38) for critic NN (32), if the regressor Ξ in (36) is PE, then the weights error $\tilde{W}_2 = W_2 - \hat{W}_2$ converges to a small set around zero. Moreover, for null NN error $\varepsilon_{HJB} = 0$ thus, $\tilde{W}_2$ converges to zero exponentially.

Proof: The solution of (37) can be deduced as

$$\begin{cases} P_2(t) = \int_0^t e^{-\ell_2(t-r)}\Xi(r)\Xi^T(r)dr \\ Q_2(t) = \int_0^t e^{-\ell_2(t-r)}\Theta\Xi^T(r)dr. \end{cases} \tag{40}$$

From (36) and (40), it holds $Q_2 = -P_2 W_2 + \upsilon_2$, where $\upsilon_2 = -\int_0^t e^{-\ell_2(t-r)}\varepsilon_{HJB}(r)\Xi(r)dr$ is bounded by a positive constant $\varepsilon_{\upsilon 2}$ as $\|\upsilon_2\| \le \varepsilon_{\upsilon 2}$.

Then, (39) is further derived as

$$M_2 = P_2\hat{W}_2 + Q_2 = -P_2\tilde{W}_2 + \upsilon_2. \tag{41}$$

Hence, M_2 contains the information of estimation error, and is used to derive the adaptive law with guaranteed convergence. Select a Lyapunov function as $V_2 = \tilde{W}_2^T\Gamma_2^{-1}\tilde{W}_2/2$, then $\dot{V}_2$ is calculated from (38) and (41) as

$$\begin{aligned} \dot{V}_2 &= \tilde{W}_2^T\Gamma_2^{-1}\dot{\tilde{W}}_2 = -\tilde{W}_2^T P_2\tilde{W}_2 + \tilde{W}_2^T\upsilon_2 \\ &\le -\sigma_2\|\tilde{W}_2\|^2 + \tilde{W}_2^T\upsilon_2 \le -\|\tilde{W}_2\|\big(\sigma_2\|\tilde{W}_2\| - \varepsilon_{\upsilon 2}\big). \end{aligned} \tag{42}$$

Based on the Lyapunov Theorem, it follows that $\tilde{W}_2$ will converge to a small set $\Omega_2 : \{\tilde{W}_2 | \|\tilde{W}_2\| \le \varepsilon_{\upsilon 2}/\sigma_2\}$, whose size is determined by the NN error ε_υ and the excitation level σ_2.

Moreover, in the ideal case of $\varepsilon_{HJB} = 0$ (and thus $\upsilon_2 = 0$), (42) is reduced to

$$\dot{V}_2 = -\tilde{W}_2^T P_2\tilde{W}_2 < -\sigma_2\|\tilde{W}_2\|^2 \le -\mu_2 V_2 \tag{43}$$

where $\mu_2 = 2\sigma_1/\lambda_{\max}(\Gamma_2^{-1})$ is a positive constant. Thus, the NN error $\tilde{W}_2$ will converge to zero *exponentially*. ■

As shown in Theorem 2, the estimated weights $\hat{W}_2$ converge to a set around W_2. Thus, $\hat{W}_2$ can be used to directly calculate optimal control (33), so that the actor NN can be avoided. This suggests a simplified ADP method with dual approximators.

Remark 5: In the proposed ADP scheme, the filter constant ℓ_i in (7) and (37) is a forgetting factor to retain the boundedness of P_i and Q_i. This constant introduces a d.c. gain $1/\ell_i$ for the filter $1/(s + \ell_i)$; hence it cannot be set too small to retain the convergence response. The constant k in (5) determines the "bandwidth" of the filter $(\bullet)_f = (\bullet)/(ks + 1)$, and thus it needs to be small for enhancing the convergence speed. Moreover, as shown in the proof of Theorems 1 and 2, the learning gain $\Gamma_i, i = 1, 2$ determines the convergence rate of the estimation error $\tilde{W}_i$. In practice, it can be set small initially and then increased gradually to seek for better convergence via a trial-and-error process.

D. Stability and Convergence Analysis

The controlled system stability and the convergence of optimal control (33) can be summarized as follows.

Theorem 3: For nonlinear system (1), design the composite control (34) with the steady-state control (22) and the optimal control (33), where the adaptive laws (9) and (38) are used. If the regressors ϕ_1 and Ξ are PE, then:

1) the tracking error e and the NN weights errors $\tilde{W}_1$, $\tilde{W}_2$ are uniformly ultimately bounded (UUB); the practical control $\hat{u}_e$ in (33) will converge to a set around the optimal solution u_e^* in (31), i.e., $\|\hat{u}_e - u_e^*\| \le \varepsilon_u$ holds for a positive constant ε_u;
2) if the approximation errors are zero, the tracking error e and the NN weights errors $\tilde{W}_1$, $\tilde{W}_2$ converge to zero; the practical control $\hat{u}_e$ converges to u_e^*.

Proof: The detailed proof is shown in the Appendix. ■

V. Simulations

In this section, two simulation examples are given to validate the theoretical studies.

Example 1: The following nonlinear system is studied as [50]:

$$\begin{cases} \dot{x}_1 = -x_1 + x_2 \\ \dot{x}_2 = -0.5x_1 - 0.5x_2\big(1 - (\cos(2x_1) + 2)^2\big) \\ \quad +(\cos(2x_1) + 2)u. \end{cases} \tag{44}$$

The purpose of tracking control is to make the system states x_1, x_2 track the given demands $x_{1d} = \sin(t)$ and $x_{2d} = \cos(t) + \sin(t)$. Since the unknown nonlinearities $f(x), g(x)$ can be formulated in a linearly parameterized form (3), the regressor of identifier is set as $\phi_1 = \begin{bmatrix} x_1 & x_2 & 0 & 0 & 0 \\ x_1 & 0 & x_2(1 - x_2(\cos(2x_1) + 2)^2) & u\cos(2x_1) & u \end{bmatrix}^T$, and the unknown weights $W_1 = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ -0.5 & 0 & -0.5 & 1 & 2 \end{bmatrix}^T$ are estimated using the adaptive law (9) to validate its convergence.

To achieve the tracking control target, the steady-state control (22) is accomplished with the optimal control (33), aiming

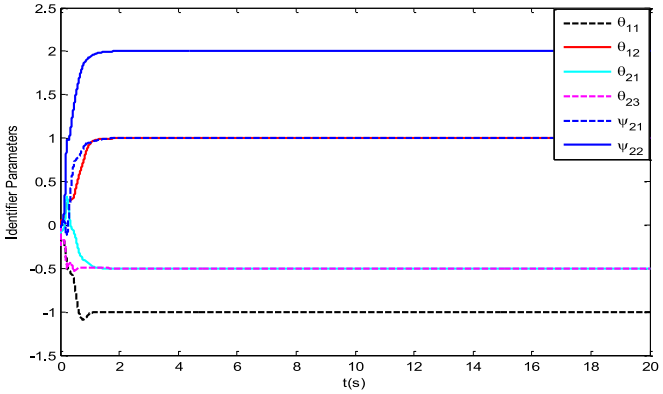

Fig. 2. Estimated identifier weights $\hat{W}_1$.

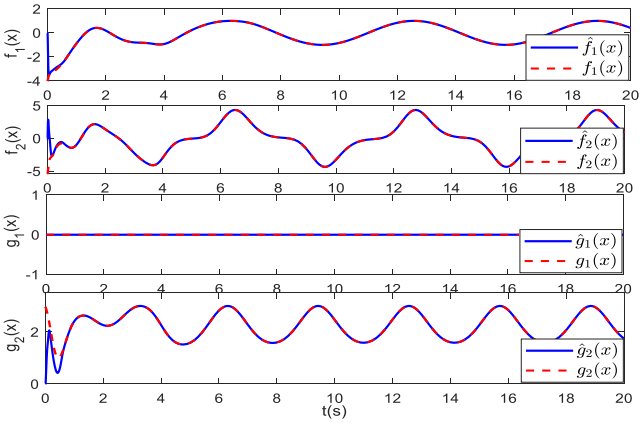

Fig. 3. Estimation performance of $f(x)$ and $g(x)$.

to minimize the following optimal cost function [50]:

$$V^*(e) = \frac{1}{2}e_1^2 + e_2^2 \tag{45}$$

with Q and R in (24) being identity matrices [50]. Hence, the ideal optimal control is derived by

$$u_e^* = -\frac{1}{2}R^{-1}[g(x)]^T \frac{\partial V^*(e)}{\partial e} = -(\cos(2e_1) + 2)e_2. \tag{46}$$

To approximate the optimal value function (45), the regressor of critic NN is chosen as $\phi_2(e) = [e_1^2, e_1e_2, e_2^2]^T$, such that the associated weights are $W_2 = [0.5, 0, 1]^T$. The simulation parameters are set as $k = 0.001$, $\ell_1 = \ell_2 = 1$, $\Gamma_1 = \Gamma_2 = 150$, $K_e = 1.65$. The initial NN weights are $\hat{W}_1(0) = \hat{W}_2(0) = 0$ and the initial conditions are $x_1(0) = 3$, $x_2(0) = -1$. Since the tracking control is considered, the system is forced to track a sinusoidal command in the simulations. Based on Lemma 1, one can check the condition $\lambda_{\min}(P_i) > \sigma_i > 0$ and find that the regressors ϕ_2 and Ξ fulfill the required excitation condition to retain the convergence of the adaptive laws (9) and (38). Hence, different to the ADP-based regulation (e.g., [24] and references therein), probing noise is not needed in this case study.

Fig. 2 shows the profiles of the no-null elements of the identifier weights $\hat{W}_1$, which converge to their ideal values in around 1 s, showing the merit of the proposed adaptive law (9) with the estimation error. Fig. 3 gives the estimation performances of unknown nonlinearities $f(x)$ and $g(x)$. It shows that

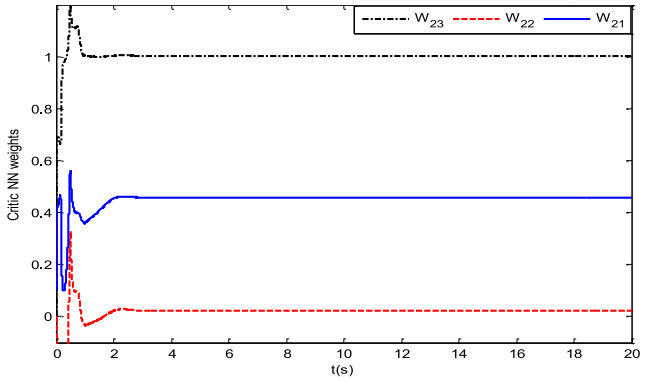

Fig. 4. Estimated critic NN weights $\hat{W}_2$.

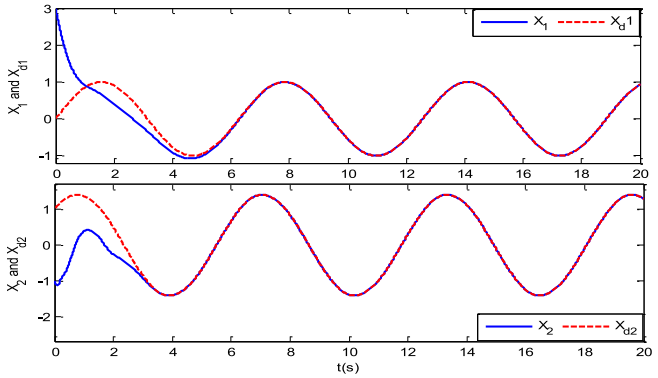

Fig. 5. State tracking performance (Example 1).

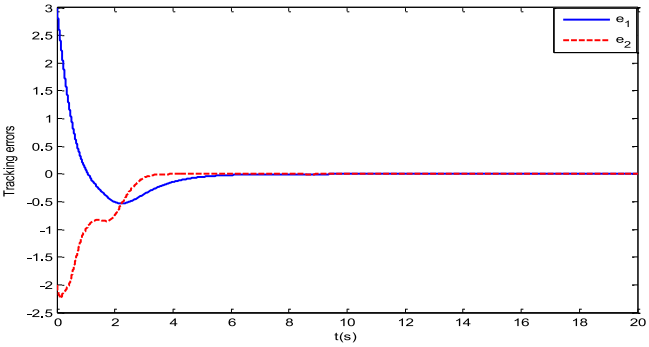

Fig. 6. Tracking errors.

the proposed adaptive identifier can reconstruct the unknown system dynamics. The estimated critic NN weights $\hat{W}_2$ are depicted in Fig. 4, which also converge closely to the ideal values $W_2 = [0.5, 0, 1]^T$ in about 2 s. This implies that the proposed optimal control (33) approaches to the ideal solution in (46). Moreover, the system states and the command to be tracked are depicted in Fig. 5, and the tracking errors are shown in Fig. 6, which indicate accurate tracking response. Finally, Fig. 7 illustrates the error between the estimated cost function $\hat{V}(e) = \hat{W}_2^T\phi_2(e)$ and the optimal cost function (45), showing that good approximation response is achieved.

For comparison, an ADP control with a triple approximation structure (consisting of an identifier NN, a critic NN, and an actor NN) in [36] is also tested for the system (44) in the same

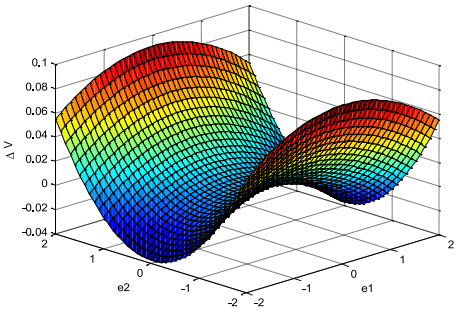

Fig. 7. Steady-state cost function error $\Delta V = \hat{V} - V^*$.

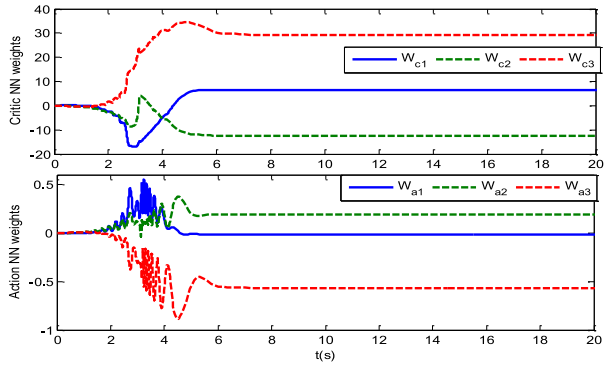

Fig. 8. Comparative simulation results of [36].

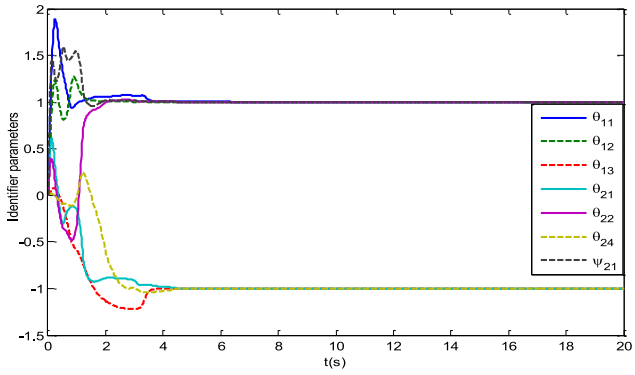

Fig. 9. Estimated identifier weights $\hat{W}_1$.

simulation conditions. Fig. 8 provides the profiles of the critic NN weights and actor NN weights. Since the adaptive laws for the identifier, critic and actor NNs in [36] are obtained based on the gradient method, the NN weights shown in Fig. 8 do not converge to their ideal values though the tracking response can be retained owing to the steady-state control. In contrast, the adaptive laws (9) and (38) proposed in this article can retain a fast convergence performance (Figs. 2 and 4), such that the actor NN can be avoided. Moreover, differing to our previous result [41], the input dynamics are unknown here.

Example 2: The following nonlinear system is studied:

$$\begin{cases} \dot{x}_1 = p_1 x_1 + p_2 x_2 + p_3 x_1 \left(x_1^2 + x_1^2\right) \\ \dot{x}_2 = p_4 x_1 + p_5 x_2 + p_6 x_2 \left(x_1^2 + x_1^2\right) + p_7 u \end{cases} \tag{47}$$

where $p = [1, 1, -1, -1, 1, -1, 1]$ are the unknown parameters to be estimated. The desired trajectory is set as $x_{1d} = \sin(t)$ and $x_{2d} = 2\sin^3(t) + \cos(t) - \sin(t)$. Simulation parameters of the identifier are $k = 0.001$, $\ell_1 = 1$, $\Gamma_1 = 1000$, $\hat{W}_1(0) = \hat{W}_2(0) = 0$, $x_1(0) = -0.15, x_2(0) = -0.1$. The unknown dynamics can be formulated as (3) with the identifier regressor and weights

$$\phi_1 = \begin{bmatrix} x_1 & x_2 & x_1(x_1^2 + x_1^2) & 0 & 0 \\ x_1 & x_2 & 0 & x_2(x_1^2 + x_1^2) & u \end{bmatrix}^T$$

$$W_1 = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & -1 & 1 \end{bmatrix}^T.$$

The optimal value function is set as $V^*(e) = 0.5e_1^2 + e_2^2$. In this case, the regressor of the critic NN is $\phi_2(e) = [e_1^2, e_1 e_2, e_2^2]^T$. The adaptive law (38) is adopted to update the critic NN weights. The simulation parameters of control are $\ell_2 = 6$, $\Gamma_2 = 700$ and $K_e = 1.65$. Fig. 9 shows the profiles of identifier weights $\hat{W}_1$, which converge to the ideal values. Hence, the unknown system dynamics can be precisely reconstructed. The critic NN weights $\hat{W}_2$ are depicted in Fig. 10, which shows the convergence to their ideal values $W_2 = [0.5, 0, 1]^T$. The tracking performance is given in Fig. 11, and the derived tracking error is indicated in Fig. 12. To prove the necessity of using the optimal control u_e to improve the tracking performance, four indices of error e_1 with the composite control u and the steady-state control u_d are considered: root mean square error (RMSE $= \sqrt{\int_0^{t_0} e_1^2 dt / t_0}$), integral absolute error (IAE $= \int |e_1| dt$), and integrated square error (ISDE $= \int (e_1 - e_{1e})^2 dt$) with the mean error e_{1e}. The values of these indices are given in Table I, which indicates that

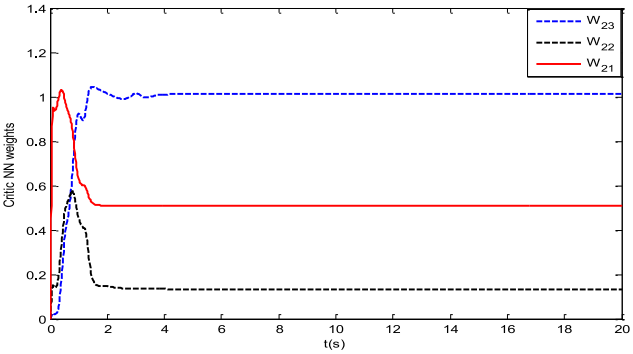

Fig. 10. Estimated critic NN weights $\hat{W}_2$.

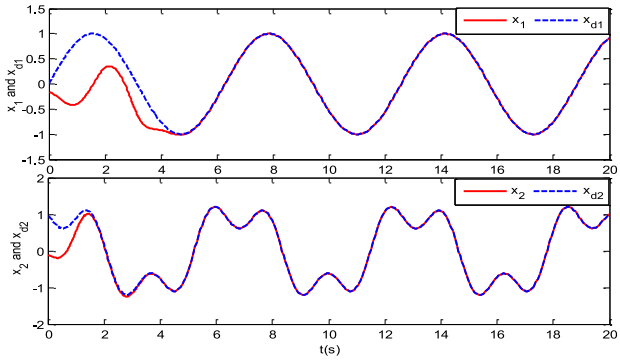

Fig. 11. State tracking performance (Example 2).

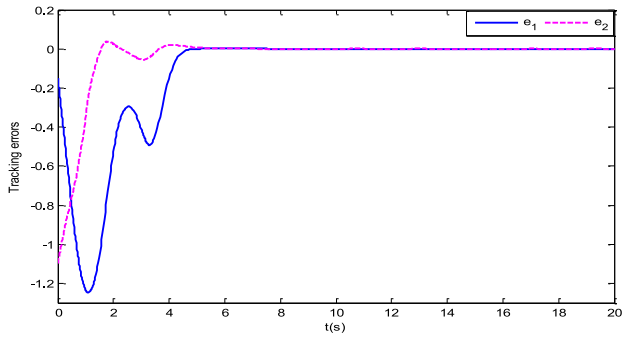

Fig. 12. Tracking errors.

TABLE I
COMPARATIVE ERROR PERFORMANCES OF u AND u_d

Performance	Optimal control u	Steady-state control u_d
RMSE	0.4458	0.5815
IAE	36.725	51.693
ISDE	12.854	17.552

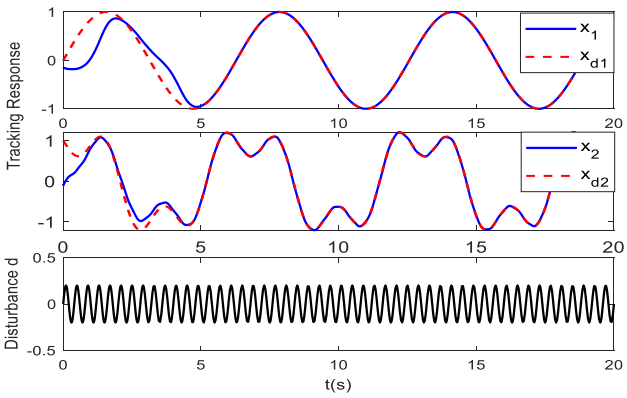

Fig. 13. Control performances with disturbance.

the proposed control u has smaller index values, and a better control performance than the steady-state control u_d.

Finally, to verify the robustness of the proposed optimal tracking control, a disturbance $d = 0.2\sin(5\pi t)$ is added to the system state x_2. The corresponding tracking performance is presented in Fig. 13, which indicates that this control can effectively eliminate the effect of disturbances, because this disturbance can be taken as a part of unknown dynamics, which are identified and compensated for.

In these simulations, it is observed that the proposed identifier estimates the unknown system dynamics accurately, and the critic NN can estimate the optimal value function well. Hence, a satisfactory tracking control response is achieved by using the suggested composite optimal control.

VI. EXPERIMENTAL VALIDATION

To assess the applicability of the suggested optimal control, practical experiments were carried out on a helicopter test-rig (Quanser Company Ltd.) as shown in Fig. 14. The target is to control the elevation angle x_1 and the elevation velocity x_2 to track given references. The two propellers are identical,

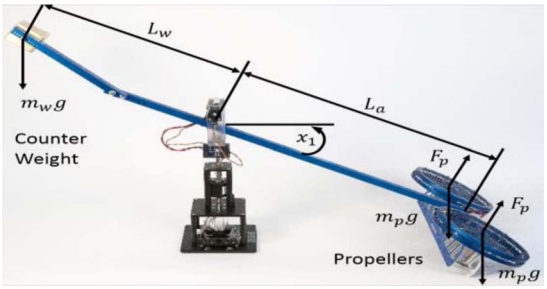

Fig. 14. Experimental setup.

and the output thrust-forces of two propellers are linear with respect to the control voltages. Therefore, the dynamics of two propellers are taken as the same presentation containing the gravity $m_p g$ and the thrust force $F_p = K_f u$, where K_f= 0.1188 N/V is the propeller force-thrust constant.

Hence, the elevation dynamics model is derived as

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}0\\ \frac{2L_aK_fK}{2m_pL_a^2+m_wL_w^2}\end{bmatrix} \times \left[u - \frac{(2m_pL_a - m_wL_w)g}{L_aK_f}\cos(x_1+\theta_0)\right] \quad (48)$$

where $m_p = 0.575$ kg is the mass of propeller, $m_w = 1.87$ kg defines the mass of the counter weight, and $g = 9.8$ N/kg denotes the gravity factor, $L_a = 0.6604$ m is the distance between the travel axis and the helicopter body, $L_w = 0.4699$ m represents the distance from the elevation axis to the counter weight gravity center, $\theta_0 = -23$ deg indicates the initial offset of the elevation (which is determined by the physical configuration), and $K = 10$ N/V is the lumped gain (including the amplifiers gain and motor drive board gain) from the control output to the propeller motors. Moreover, the cost function is set as $V(e(t)) = \int_t^\infty (e^2+u_e^2)d\tau$, which is used to derive the optimal control action u_e.

In practice, the original reference signal is a square wave signal with a step size of 10° with a period of 20 s. The offset is set as 10° to avoid contact between the helicopter body and the test ground. The elevation references x_{d1}, x_{d2} are calculated by injecting the original reference signal into a second order model with natural frequency $\omega = 1$ rad/s and a damping ratio $\zeta = 0.707$. The proposed controller was then implemented by using Simulink module built in dSpace. Based on the model (48), it is clear that the unknown dynamics can be formulated as the linearly parameterized form (3), such that the regressor of the identifier can be set as $\phi_1(x,u) = [u, \cos(x_1+\theta_0)]^T$, where the term x_2 will not be involved. The parameters in the identifier are set as $k = 0.01, \ell_1 = 5$ and $\Gamma_1 = \text{diag}([0.8, 18])$. As shown in Fig. 15, all the estimated weights rapidly converge to the ideal values of $W_1 = [1.72 - 25.59]$. Note that the convergence of one parameter (with true value -25.59) is slightly slower than the other one since the amplitude of the related regressor $\cos(x_1+\theta_0)$ is smaller. Moreover, due to the unavoidable uncertainties and measurement noise in practice, the estimated identifier weights oscillate slightly around the expected values after a transient period. The regressor for the critic NN

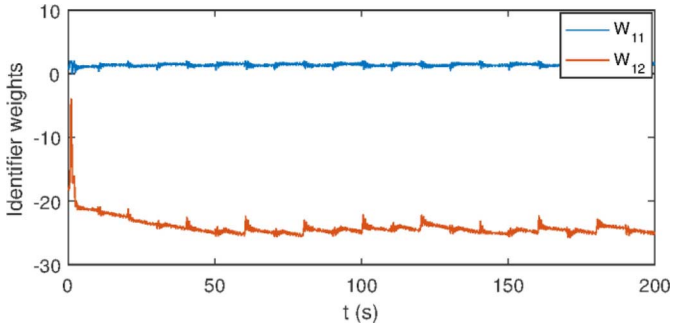

Fig. 15. Convergence of identifier weights.

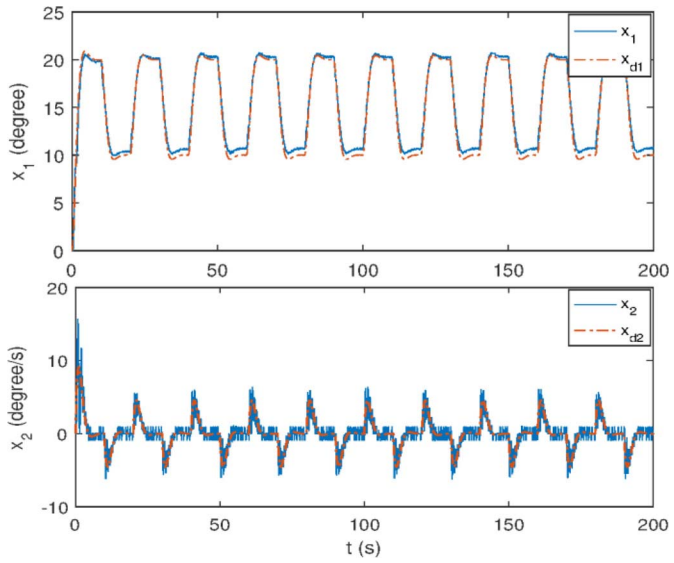

Fig. 16. Tracking performance of the proposed controller.

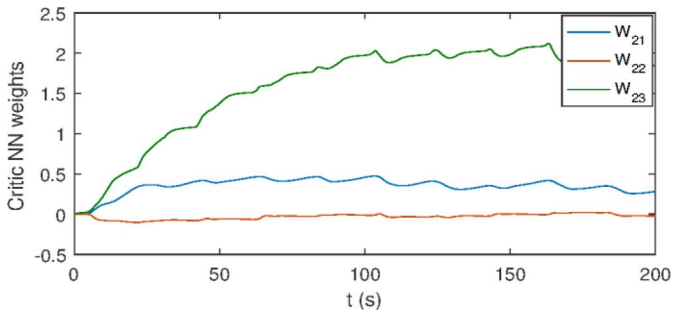

Fig. 17. Critic NN weights.

is $\phi_2(e) = [e_1^2, e_1e_2, e_2^2]^T$, while the learning parameters are $k = 1, \ell_2 = 0.5$, and $\Gamma_2 = \text{diag}([0.8, 2, 3])$. The tracking response using the proposed composite control is illustrated in Fig. 16. One may find from Fig. 16 that fairly satisfactory tracking control responses can be obtained. The oscillations in the velocity come from numerical differentiator used to calculate the velocity via the measured position signal. Moreover, the online updated critic NN weights are shown in Fig. 17, which illustrates the convergences in 100 s.

To show the advantages of the suggested optimal control action, a nominal feedback controller $u = K_x x + K_r d$ with $K_x = [-0.97, -1.1]$ and $K_r = 0.97$ is also applied to the

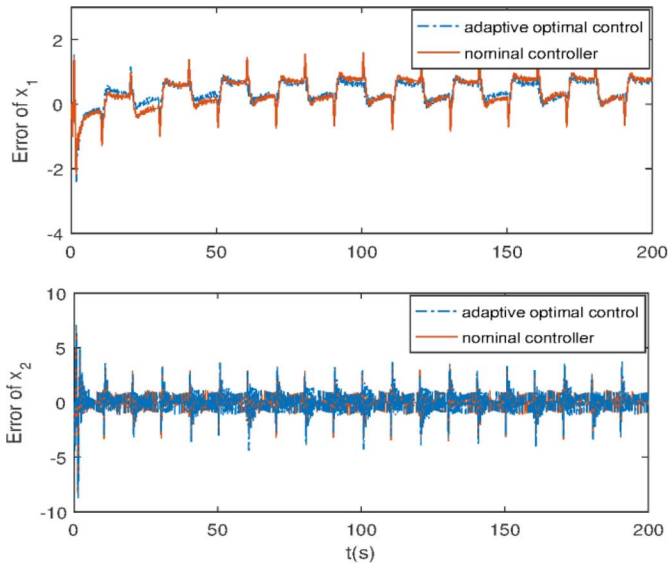

Fig. 18. Comparative tracking errors.

TABLE II
COMPARATIVE ERROR PERFORMANCES

Performance	Nominal control	Optimal control
IAE	93.658	90.939
ISDE	39.590	30.011
IAU	3427.9	3428.4

experimental setup for comparisons. The tracking errors of this nominal controller and the proposed composite controller are shown in Fig. 18, where it is shown that the tracking errors using the proposed controller are smaller than that of the nominal controller. This implies that the use of optimal control leads to a better tracking response. Specifically, the suggested optimal control achieves a faster convergence rate than the nominal controller under large initial conditions. In the steady-state, the optimal controller has less oscillations in the velocity error compared to the nominal controller. To quantify the control performance, three performance indices: 1) the integrated absolute error: IAE $= \int |e(t)|dt$; 2) the integrated square error: ISDE $= \int (e(t) - e_0)^2 dt$ with the mean error e_0; and 3) the integrated absolute control; IAU $= \int |u(t)|dt$ are calculated. Table II provides comparative results of e_1 for both controllers within the time interval $t = 0 \sim 100$ s. From Table II, one may conclude that the proposed optimal control obtains smaller IAE and ISDE, and thus better performance. However, this is partially at the expense of a larger control effort (i.e., IAU).

VII. CONCLUSION

This article proposed and practically validated an optimal tracking control of nonlinear systems with unknown dynamics. It introduces a new identifier-critic-based ADP framework, where the unknown dynamics are reconstructed by using the identifier. Then, an optimal control that minimizes a given cost function is accomplished with a steady-state control to achieve

optimal tracking. A critic NN is used to approximate the solution of the derived HJB equation. In this ADP approach, only two NNs (i.e., an identifier NN and a critic NN) are needed, while the widely used actor NN in other ADP schemes is avoided. Consequently, the computational costs can be reduced and the convergence speed can be improved. Another salient feature is that the novel adaptive laws are proposed to update the weights of the NNs simultaneously to retain the convergence to the ideal values. Simulations and experiments all illustrate the efficacy of the proposed control scheme. Future work will focus on the optimal tracking control for nonaffine systems, and relax the required PE conditions.

APPENDIX
PROOF OF THEOREM 3

Proof: By substituting (33) into (23), if follows:

$$\begin{aligned}\dot{e} &= -K_e e + \hat{\psi}\varsigma u_e + \varepsilon_N + \varepsilon_T \\ &= -K_e e + \hat{\psi}\varsigma\Big\{-\frac{1}{2}R^{-1}\Big(\hat{\psi}\varsigma\Big)^T\nabla\phi_2^T\hat{W}_2 \\ &\qquad + \frac{1}{2}R^{-1}\Big(\hat{\psi}\varsigma\Big)^T\big(\nabla\phi_2^T W_2 + \nabla\varepsilon_v\big)\Big\} \\ &\quad + \hat{\psi}\varsigma u_e^* + \varepsilon_N + \varepsilon_T \\ &= -K_e e + \frac{1}{2}\hat{\psi}\varsigma R^{-1}\Big(\hat{\psi}\varsigma\Big)^T\nabla\phi_2^T\tilde{W}_2 + \frac{1}{2}\hat{\psi}\varsigma R^{-1}\Big(\hat{\psi}\varsigma\Big)^T\nabla\varepsilon_v \\ &\quad + \hat{\psi}\varsigma u_e^* + \varepsilon_N + \varepsilon_T. \end{aligned} \tag{49}$$

Choose the following Lyapunov function:

$$\begin{aligned}V &= V_1 + V_2 + V_3 + V_4 + V_5 \\ &= \frac{1}{2}tr\Big(\tilde{W}_1^T\Gamma_1^{-1}\tilde{W}_1\Big) + \frac{1}{2}\tilde{W}_2^T\Gamma_2^{-2}\tilde{W}_2 + \Gamma e^T e + KV^* \\ &\quad + \Upsilon_1\upsilon_1^T\upsilon_1 + \Upsilon_2\upsilon_2^T\upsilon_2 \end{aligned} \tag{50}$$

where V^* is the value function given in (26); $K > 0$, $\Gamma > 0$, $\Upsilon_1 > 0$, $\Upsilon_2 > 0$ are all positive constants. Within a compact set $\tilde{\Omega} \in \mathbb{R}^{d\times n}\times\mathbb{R}^l\times\mathbb{R}^n\times\mathbb{R}^{d\times n}\times\mathbb{R}^l\times\mathbb{R}^n\times\mathbb{R}^n$of the coordinates $(\tilde{W}_1, \tilde{W}_2, e, \upsilon_1, \upsilon_2, x_d, \dot{x}_d)$, which contains the origin in its interior, i.e., $(\tilde{W}_1, \tilde{W}_2, e, \upsilon_1, \upsilon_2, x_d, \dot{x}_d) \in \tilde{\Omega}$ implies $(e + x_d, u_d(e + x_d, \dot{x}_d) + u_e(e)) \in \Omega$, one can choose x_{1d} and $\dot{x}_{1d}$ within the compact set $\tilde{\Omega}$. Then, for any initial condition, the system state x and control action u will be within $\tilde{\Omega}$ in finite time, $t \in [0, T_1]$, guaranteeing that υ_1,υ_2 are bounded in all $t \in [0, T_1]$.

From Young's inequality $ab \le a^2\eta/2 + b^2/2\eta$ for $\eta > 0$, it follows:

$$\begin{aligned}\dot{V}_1 &= -tr\big(\tilde{W}_1^T P_1\tilde{W}_1\big) + tr\big(\tilde{W}_1^T\upsilon_1\big) \le -\sigma_1\big\|\tilde{W}_1\big\|^2 + \big\|\tilde{W}_1^T\upsilon_1\big\| \\ &\le -\Big(\sigma_1 - \frac{1}{2\eta\Upsilon_1}\Big)\big\|\tilde{W}_1\big\|^2 + \frac{\eta\Upsilon_1\|\upsilon_1\|^2}{2} \end{aligned} \tag{51}$$

and

$$\begin{aligned}\dot{V}_2 &= -\tilde{W}_2^T P_2\tilde{W}_2 + \tilde{W}_2^T\upsilon_2 \le -\sigma_2\big\|\tilde{W}_2\big\|^2 + \big\|\tilde{W}_2^T\upsilon_2\big\| \\ &\le -\Big(\sigma_2 - \frac{1}{2\eta\Upsilon_2}\Big)\big\|\tilde{W}_2\big\|^2 + \frac{\eta\Upsilon_2\|\upsilon_2\|^2}{2}. \end{aligned} \tag{52}$$

On the other hand, $\dot{V}_3$ from (26) and (49) is calculated as

$$\begin{aligned}\dot{V}_3 &= 2\Gamma e^T\dot{e} + K\big(-e^TQe - u_e^{*T}Ru_e^*\big) \\ &= 2\Gamma e^T\Big(-K_e e + \frac{1}{2}BR^{-1}B^T\nabla\phi_2^T\tilde{W}_2 + Bu_e^* \\ &\qquad + \frac{1}{2}BR^{-1}B^T\nabla\varepsilon_v + \varepsilon_N + \varepsilon_T\Big) \\ &\quad + K\big(-e^TQe - u_e^{*T}Ru_e^*\big) \\ &\le -\Big[2\lambda_{\min}(K_e)\Gamma + K\lambda_{\min}(Q) \\ &\qquad - \Big(\Big\|B^TR^{-1}B\nabla\phi_2\Big\| + \Big\|B^TR^{-1}B\Big\| + 3\Big)\Gamma\Big]\|e\|^2 \\ &\quad + \frac{1}{4}\Gamma\Big\|B^TR^{-1}B\nabla\phi_2\Big\|\big\|\tilde{W}_2\big\|^2 + \frac{1}{4}\Gamma\Big\|B^TR^{-1}B\Big\|\nabla\varepsilon_v^T\nabla\varepsilon_v \\ &\quad + \Gamma\varepsilon_N^T\varepsilon_N + \Gamma\varepsilon_T^T\varepsilon_T \\ &\quad - \Big(K\lambda_{\min}(R) - \Gamma\|B\|^2\Big)\big\|u_e^*\big\|^2 \end{aligned} \tag{53}$$

with $B = \hat{\psi}\varsigma(x)$ a bounded variable.

Based on (15), one can know $\dot{\upsilon}_1 = -\ell_1\upsilon_1 + \phi_{1f}\varepsilon_{Tf}^T$, so that

$$\begin{aligned}\dot{V}_4 &= 2\Upsilon_1\upsilon_1^T\dot{\upsilon}_1 = 2\Upsilon_1\upsilon_1^T\Big(-\ell_1\upsilon_1 + \phi_{1f}\varepsilon_{Tf}^T\Big) \\ &\le -\Upsilon_1(2\ell_1 - \eta)\|\upsilon_1\|^2 + \Big\|\phi_{1f}\varepsilon_{Tf}^T\Big\|^2/\eta. \end{aligned} \tag{54}$$

From (40), it holds $\dot{\upsilon}_2 = -\ell_2\upsilon_2 + \Xi\varepsilon_{HJB}$, so that

$$\begin{aligned}\dot{V}_5 &= 2\Upsilon_2\upsilon_2^T\dot{\upsilon}_2 \\ &= 2\Upsilon_2\upsilon_2^T\Big\{-\ell_2\|\upsilon_2\| \\ &\qquad + \Xi\Big[\big(W_2^T\nabla\phi_2 + \nabla\varepsilon_v\big)(\varepsilon_N + \varepsilon_T) \\ &\qquad + \nabla\varepsilon_v B\Big(-R^{-1}B^T\nabla\phi_2^T\hat{W}_2/2\Big) - \nabla\varepsilon_v K_e e\Big]\Big\} \\ &\le -\Upsilon_2(2\ell_2 - 4\eta)\|\upsilon_2\|^2 \\ &\quad + \frac{1}{\eta}\Upsilon_2\big\|\Xi\big(W_2^T\nabla\phi_2 + \nabla\varepsilon_v\big)\big\|^2\|\varepsilon_N\|^2 \\ &\quad + \frac{1}{\eta}\Upsilon_2\big\|\Xi\big(W_2^T\nabla\phi_2 + \nabla\varepsilon_v\big)\big\|^2\|\varepsilon_T\|^2 \\ &\quad + \frac{1}{4\eta}\Upsilon_2\Big\|\Xi\nabla\varepsilon_v BR^{-1}B^T\nabla\phi_2^T\hat{W}_2\Big\|^2 \\ &\quad + \frac{\Upsilon_2}{\eta}\|\Xi\nabla\varepsilon_V K_e\|^2\|e\|^2. \end{aligned} \tag{55}$$

Consequently, substitute $\varepsilon_N = \tilde{W}_1\phi_1$ into (55) and have

$$\begin{aligned}\dot{V} &= \dot{V}_1 + \dot{V}_2 + \dot{V}_3 + \dot{V}_4 + \dot{V}_5 \\ &\le -\Big[\sigma_1 - \frac{1}{2\eta\Upsilon_1} \\ &\qquad - \Big(\Gamma + \frac{1}{\eta}\Upsilon_2\big\|\Xi\big(W_2^T\nabla\phi_2 + \nabla\varepsilon_v\big)\big\|^2\Big)\|\phi_1\|^2\Big]\big\|\tilde{W}_1\big\|^2 \\ &\quad - \Big(\sigma_2 - \frac{1}{2\eta\Upsilon_2} - \frac{1}{4}\Gamma\Big\|B^TR^{-1}B\nabla\phi_2\Big\|\Big)\big\|\tilde{W}_2\big\|^2 \\ &\quad - \Big[2\lambda_{\min}(K_e)\Gamma + K\lambda_{\min}(Q) \\ &\qquad - \Big(\Big\|B^TR^{-1}B\nabla\phi_2\Big\| + \Big\|B^TR^{-1}B\Big\| + 3\Big)\Gamma \\ &\qquad - \frac{\Upsilon_2}{\eta}\|\Xi\nabla\varepsilon_V K_e\|^2\Big]\|e\|^2 \end{aligned}$$

$$
\begin{aligned}
&- \left(K\lambda_{\min}(R) - \Gamma\|B\|^2\right)\left\|u_e^*\right\|^2 \\
&- \Upsilon_1(2\ell_1 - 1.5\eta)\|\upsilon_1\|^2 - \Upsilon_2(2\ell_2 - 4.5\eta)\|\upsilon_2\|^2 \\
&+ \left(\Gamma + \frac{1}{\eta}\Upsilon_2\left\|\Xi\left(W_2^T\nabla\phi_2 + \nabla\varepsilon_v\right)\right\|^2\right)\|\varepsilon_T\|^2 \\
&+ \frac{1}{4}\Gamma\left\|B^TR^{-1}B\right\|\nabla\varepsilon_v^T\nabla\varepsilon_v \\
&+ \frac{1}{\eta}\left\|\phi_{1f}\varepsilon_{Tf}^T\right\|^2 + \frac{1}{4\eta}\Upsilon_2\left\|\Xi\nabla\varepsilon_v BR^{-1}B^T\nabla\phi_2^T\hat{W}_2\right\|^2. \quad (56)
\end{aligned}
$$

Hence, it allows to represent (56) as

$$
\begin{aligned}
\dot{V} \le &-a_1\left\|\tilde{W}_1\right\|^2 - a_2\left\|\tilde{W}_2\right\|^2 - a_3\|e\|^2 - a_4\|\upsilon_1\|^2 \\
&- a_5\|\upsilon_2\|^2 + \gamma \quad (57)
\end{aligned}
$$

with

$$
\begin{aligned}
a_1 &= \sigma_1 - \frac{1}{2\eta\Upsilon_1} - \left(\Gamma + \Upsilon_2\left\|\Xi\left(W_2^T\nabla\phi_2 + \nabla\varepsilon_v\right)\right\|^2/\eta\right)\|\phi_1\|^2 \\
a_2 &= \sigma_2 - \frac{1}{2\eta\Upsilon_2} - \frac{1}{4}\Gamma\left\|B^TR^{-1}B\nabla\phi_2\right\| \\
a_3 &= 2\lambda_{\min}(K_e)\Gamma + K\lambda_{\min}(Q) - \frac{1}{\eta}\Upsilon_2\|\Xi\nabla\varepsilon_V K_e\|^2 \\
&\quad - \left(\left\|B^TR^{-1}B\nabla\phi_2\right\| + \left\|B^TR^{-1}B\right\| + 3\right)\Gamma \\
a_4 &= \Upsilon_1(2\ell_1 - 1.5\eta) \\
a_5 &= \Upsilon_2(2\ell_2 - 4.5\eta) \\
\gamma &= \left(\Gamma + \frac{1}{\eta}\Upsilon_2\left\|\Xi\left(W_2^T\nabla\phi_2 + \nabla\varepsilon_v\right)\right\|^2\right)\|\varepsilon_T\|^2 \\
&\quad + \frac{1}{4}\Gamma\left\|B^TR^{-1}B\right\|\nabla\varepsilon_v^T\nabla\varepsilon_v + \frac{1}{\eta}\left\|\phi_{1f}\varepsilon_{Tf}^T\right\|^2 \\
&\quad + \frac{1}{4\eta}\Upsilon_2\left\|\Xi\nabla\varepsilon_v BR^{-1}B^T\nabla\phi_2^T\hat{W}_2\right\|^2.
\end{aligned}
$$

It is clear that γ is a positive constant, which represents the effects of the identifier and critic NN errors ε_T,$\nabla\varepsilon_v$. Moreover, to guarantee the stability of (57), a_1, a_2, a_3, a_4, a_5 should be positive. For this purpose, the design parameters K, Γ, $\Upsilon_1, \Upsilon_2, \eta, \ell_i, i = 1, 2$, and K_e need to be configured properly. In detail, they are selected to fulfill the following condition:

$$
\begin{aligned}
&\Gamma < \min\left\{\frac{\sigma_1}{\|\phi_1\|^2}, \frac{4\sigma_2}{\left\|B^TR^{-1}B\nabla\phi_2\right\|}\right\} \\
&\eta > \max\left(\frac{1/2\Upsilon_1 + \Upsilon_2\left\|\Xi\left(W_2^T\nabla\phi_2 + \nabla\varepsilon_v\right)\right\|^2}{\sigma_1 - \Gamma\|\phi_1\|^2}, \right. \\
&\qquad \left. \frac{1}{\Upsilon_2\left(2\sigma_2 - -\frac{1}{2}\Gamma\left\|B^TR^{-1}B\nabla\phi_2\right\|\right)}, \frac{\Upsilon_2\|\Xi\nabla\varepsilon_V K_e\|^2}{2\lambda_{\min}(K_e)\Gamma}\right) \\
&K > \max\left(\frac{2\lambda_{\min}(K_e)\Gamma + -\frac{1}{\eta}\Upsilon_2\|\Xi\nabla\varepsilon_V K_e\|^2 - \left(\left\|B^TR^{-1}B\nabla\phi_2\right\| + \left\|B^TR^{-1}B\right\| + 3\right)\Gamma}{\lambda_{\min}(Q)}, \right. \\
&\qquad \left. \frac{\Gamma\|B\|^2}{\lambda_{\min}(R)}\right) \\
&\lambda_{\min}(K_e) > \frac{\left\|B^TR^{-1}B\nabla\phi_2\right\| + \left\|B^TR^{-1}B\right\| + 3}{2} \\
&\ell_i > 4.5\eta/2, \Upsilon_1 > 0, \Upsilon_2 > 0.
\end{aligned}
$$

Thus, the constants a_1, a_2, a_3, a_4, and a_5 are all positive.

1) When both the identifier and critic NN errors are nonzero, we know $\gamma \neq 0$. Then, for any
$$
\begin{aligned}
&\left\|\tilde{W}_1\right\| > \sqrt{\gamma/a_1}, \left\|\tilde{W}_2\right\| > \sqrt{\gamma/a_2}, \|e\| > \sqrt{\gamma/a_3}, \\
&\|\upsilon_1\| > \sqrt{\gamma/a_4}\|\upsilon_2\| > \sqrt{\gamma/a_5} \quad (58)
\end{aligned}
$$
it can be verified from (57) that $\dot{V}$ is negative. This together with the Lyapunov theorem illustrates that the control error e, the NN weights errors $\tilde{W}_1$ and $\tilde{W}_2$ are UUB.
To prove $\|\hat{u}_e - u_e^*\| \le \varepsilon_u$, we recall the definition of u_e^* in (31) and $\hat{u}_e$ in (33), and then have
$$
\hat{u}_e - u_e^* = \frac{1}{2}R^{-1}B^T \triangle \phi_2^T\tilde{W}_2 + \frac{1}{2}R^{-1}B^T\nabla\varepsilon_v. \quad (59)
$$
Then for $t \to \infty$, the upper bound of (59) fulfills
$$
\lim_{t\to+\infty}\left\|\hat{u}_e - u_e^*\right\| \le \frac{1}{2}\left\|R^{-1}B^T\right\|\left(\phi_M\left\|\tilde{W}_2\right\| + \phi_\varepsilon\right) \le \varepsilon_u \quad (60)
$$
with $\varepsilon_u > 0$ a constant.

2) In the ideal case where both the identifier and critic NNs errors are null, i.e., $\varepsilon_T = \nabla\varepsilon_v = 0$, we can verify that $\gamma = 0$, such that (57) can be rewritten as
$$
\dot{V} = -a_1\left\|\tilde{W}_1\right\|^2 - a_2\left\|\tilde{W}_2\right\|^2 - a_3\|e\|^2 - a_4\|\upsilon_2\|^2 \le 0. \quad (61)
$$
Then, there exists a set $\hat{\Omega} \subset \tilde{\Omega}$ in $(\tilde{W}_1, \tilde{W}_2, e, \upsilon_1, \upsilon_2)$ with $(0, 0, 0, 0, 0)$ in its interior. From the Lyapunov Theorem, we know $V \to 0$ within $\hat{\Omega}$ as $t \to +\infty$ and thus $\tilde{W}_1$, $\tilde{W}_2$, and e all converge to zero. Finally, to show the convergence of the proposed control under $\nabla\varepsilon_v$=0, we can obtain
$$
\begin{aligned}
\hat{u}_e - u_e^* &= -\frac{1}{2}R^{-1}B^T\nabla\phi_2^T\hat{W}_2 + \frac{1}{2}R^{-1}B^T\nabla\phi_2^TW_2 \\
&= \frac{1}{2}R^{-1}B^T\nabla\phi_2^T\tilde{W}_2 \quad (62)
\end{aligned}
$$
so that
$$
\lim_{t\to+\infty}\left\|\hat{u}_e - u_e^*\right\| \le \frac{1}{2}\phi_M\left\|R^{-1}B^T\right\|\left\|\tilde{W}_2\right\| = 0. \quad (63)
$$

The proof is finished. ■

REFERENCES

[1] S. Sastry and M. Bodson, *Adaptive Control: Stability, Convergence and Robustness*. Mineola, NY, USA: Dover Publ. Inc., 1989.

[2] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. New York, NY, USA: Wiley, 2012.

[3] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Netw.*, vol. 22, pp. 237–246, Apr. 2009.

[4] K. Doya, "Reinforcement learning in continuous time and space," *Neural Comput.*, vol. 12, pp. 219–245, Jan. 2000.

[5] P. J. Werbos, "A menu of designs for reinforcement learning over time," in *Neural Networks for Control*. Cambridge, MA, USA: MIT Press, 1990, pp. 67–95.

[6] 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, 3rd Quart., 2009.

[7] H. Zhang, X. Zhang, Y. Luo, and J. Yang, "An overview of research on adaptive dynamic programming," *Acta Automatica Sinica*, vol. 39, pp. 303–311, Apr. 2013.

[8] H. Zhang, J. Zhang, G.-H. Yang, and Y. Luo, "Leader-based optimal coordination control for the consensus problem of multiagent differential games via fuzzy adaptive dynamic programming," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 1, pp. 152–163, Feb. 2015.

[9] Y. Lv and X. Ren, "Approximate Nash solutions for multiplayer mixed-zero-sum game with reinforcement learning," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 12, pp. 2739–2750, Dec. 2019.

[10] Q. Wei, F. L. Lewis, D. Liu, R. Song, and H. Lin, "Discrete-time local value iteration adaptive dynamic programming: Convergence analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 6, pp. 875–891, Jun. 2018.

[11] R. Song, W. Xiao, H. Zhang, and C. Sun, "Adaptive dynamic programming for a class of complex-valued nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1733–1739, Sep. 2014.

[12] Q. Wei, G. Shi, R. Song, and Y. Liu, "Adaptive dynamic programming-based optimal control scheme for energy storage systems with solar renewable energy," *IEEE Trans. Ind. Electron.*, vol. 64, no. 7, pp. 5468–5478, Jul. 2017.

[13] D. Wang, D. Liu, H. Li, B. Luo, and H. Ma, "An approximate optimal control approach for robust stabilization of a class of discrete-time nonlinear systems with uncertainties," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 5, pp. 713–717, May 2016.

[14] J. Na, B. Wang, G. Li, S. Zhan, and W. He, "Nonlinear constrained optimal control of wave energy converters with adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, vol. 66, no. 10, pp. 7904–7915, Oct. 2019.

[15] 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, pp. 1825–1832, Aug. 2012.

[16] 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.

[17] H. Zhang, R. Song, Q. Wei, and T. Zhang, "Optimal tracking control for a class of nonlinear discrete-time systems with time delays based on heuristic dynamic programming," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 1851–1862, Dec. 2011.

[18] 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.

[19] 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.

[20] 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.

[21] T. Hanselmann, L. Noakes, and A. Zaknich, "Continuous-time adaptive critics," *IEEE Trans. Neural Netw.*, vol. 18, no. 3, pp. 631–647, May 2007.

[22] 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, pp. 779–791, May 2005.

[23] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, pp. 477–484, Feb. 2009.

[24] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, pp. 878–888, May 2010.

[25] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, pp. 82–92, Jan. 2013.

[26] 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.

[27] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, pp. 2699–2704, Oct. 2012.

[28] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming with an application to power systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1150–1156, Jul. 2013.

[29] Y. Yang, Z. Guo, H. Xiong, D.-W. Ding, Y. Yin, and D. C. Wunsch, "Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3735–3747, Dec. 2019.

[30] Z. Wang, R. Lu, F. Gao, and D. Liu, "An indirect data-driven method for trajectory tracking control of a class of nonlinear discrete-time systems," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4121–4129, May 2017.

[31] M. Krstic and Z.-H. Li, "Optimal design of adaptive tracking controllers for non-linear systems," *Automatica*, vol. 33, pp. 1459–1473, Aug. 1997.

[32] W. Luo, Y.-C. Chu, and K.-V. Ling, "Inverse optimal adaptive control for attitude tracking of spacecraft," *IEEE Trans. Autom. Control*, vol. 50, no. 11, pp. 1639–1654, Nov. 2005.

[33] A. Mannava, S. N. Balakrishnan, L. Tang, and R. G. Landers, "Optimal tracking control of motion systems," *IEEE Trans. Control Syst. Technol.*, vol. 20, no. 6, pp. 1548–1558, Nov. 2012.

[34] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Trans. Autom. Control*, vol. 59, no. 11, pp. 3051–3056, Nov. 2014.

[35] C. Qin, H. Zhang, and Y. Luo, "Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming," *Int. J. Control*, vol. 87, no. 5, pp. 1000–1009, 2014.

[36] 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.

[37] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, vol. 51, pp. 40–48, Jan. 2015.

[38] R. Kamalapurkar, L. Andrews, P. Walters, and W. E. Dixon, "Model-based reinforcement learning for infinite-horizon approximate optimal tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 753–758, Mar. 2017.

[39] M.-B. Radac, R.-E. Precup, and R.-C. Roman, "Model-free control performance improvement using virtual reference feedback tuning and reinforcement Q-learning," *Int. J. Syst. Sci.*, vol. 48, no. 5, pp. 1071–1083, 2017.

[40] L. Hager, K. R. Uren, G. Van Schoor, and A. J. van Rensburg, "Adaptive Neural network control of a helicopter system with optimal observer and actor-critic design," *Neurocomputing*, vol. 302, pp. 75–90, Aug. 2018.

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

[42] J. Na, M. N. Mahyuddin, G. Herrmann, X. Ren, and P. Barber, "Robust adaptive finite time parameter estimation and control for robotic systems," *Int. J. Robust Nonlinear Control*, vol. 25, pp. 3045–3071, Nov. 2015.

[43] W. He, Y. Dong, and C. Sun, "Adaptive neural impedance control of a robotic manipulator with input saturation," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 46, no. 3, pp. 334–344, Mar. 2016.

[44] Y.-J. Liu, L. Tang, S. Tong, C. P. Chen, and D.-J. Li, "Reinforcement learning design-based adaptive tracking control with less learning parameters for nonlinear discrete-time MIMO systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 165–176, Jan. 2015.

[45] S. Wang, J. Na, and X. Ren, "RISE-based asymptotic prescribed performance tracking control of nonlinear servo mechanisms," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2359–2370, Dec. 2018.

[46] H. Li, Y. Gao, P. Shi, and H.-K. Lam, "Observer-based fault detection for nonlinear systems with sensor fault and limited communication capacity," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2745–2751, Sep. 2016.

[47] C.-J. Kim and D. Chwa, "Obstacle avoidance method for wheeled mobile robots using interval type-2 fuzzy neural network," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 3, pp. 677–687, Jun. 2015.

[48] Y. Gao, F. Xiao, J. Liu, and R. Wang, "Distributed soft fault detection for interval type-2 fuzzy-model-based stochastic systems with wireless sensor networks," *IEEE Trans. Ind. Informat.*, vol. 15, no. 1, pp. 334–347, Jan. 2019.

[49] M. Krstic, I. Kanellakopoulos, and P. V. Kokotovic, *Nonlinear and Adaptive Control Design*, vol. 222, New York, NY, USA: Wiley, 1995.

[50] V. Nevistic and J. A. Primbs, "Constrained nonlinear optimal control: A converse HJB approach," Dept. Control Dyn. Syst., California Inst. Technol., Pasadena, CA, USA, Rep. CIT-CDS 96-021, 1996.

Jing Na (Member, IEEE) received the B.Eng. and Ph.D. degrees in automation and control from the School of Automation, Beijing Institute of Technology, Beijing, China, in 2004 and 2010, respectively.

From 2011 to 2013, he was a Monaco/ITER Postdoctoral Fellow with ITER Organization, Saint-Paul-lez-Durance, France. From 2015 to 2017, he was a Marie Curie Fellow with the Department of Mechanical Engineering, University of Bristol, Bristol, U.K. Since 2010, he has been with the Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming, China, where he became a Professor in 2013. He has coauthored one monograph and more than 100 international journal and conference papers. His current research interests include intelligent control, adaptive parameter estimation, nonlinear control and applications for robotics, vehicle systems, and wave energy convertor.

Dr. Na has been awarded the Best Application Paper Award of IFAC ICONS 2013 and the Hsue-Shen Tsien Paper Award in 2017. He is currently an Associate Editor of the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, *Neurocomputing*, and has served as the Organization Committee Chair of DDCLS 2019 and International Program Committee Chair of ICMIC 2017.

Yongfeng Lv (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in mechatronic engineering from the Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming, China, in 2012 and 2016, respectively. He is currently pursuing the Ph.D. degree in control science and engineering with the School of Automation, Beijing Institute of Technology, Beijing, China.

His current research interests include adaptive dynamic programming, optimal control, game theory, and multi-input system.

Kaiqiang Zhang (Member, IEEE) received the B.Eng. degree in automation and control from the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2012, and the M.Sc. degree (Distinction) in robotics and the Ph.D. degree in control from the University of Bristol, Bristol, U.K., in 2013 and 2019, respectively.

He was a Control System Engineer with the Institute of Microelectronics, Chinese Academy of Sciences, Beijing, China, in 2014. He became a Research Associate in September 2018 and has been a Visiting Senior Research Associate with the University of Bristol since 2020. In January 2020, he has been a Control Systems Engineer with the U.K. Atomic Energy Authority, Abingdon, U.K. His research interests include adaptive control, high precision control, system integration, and robotics.

Jun Zhao (Graduate Student Member, IEEE) received the B.Sc. degree in mechanical design, manufacturing and automation from the Qingdao University of Technology, Qingdao, China, in 2016. He is currently pursuing the Ph.D. degree in mechanical engineering with the Kunming University of Science and Technology, Kunming, China.

His current research interests include robust control output-feedback control, adaptive, and learning systems.