Data-Based Self-Learning Optimal Control for Continuous-Time Unknown Nonlinear Systems With Disturbance

Qinglai Wei[1], Derong Liu[2], Ruizhuo Song[2], Pengfei Yan[1]

1. The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
E-mail: qinglai.wei@ia.ac.cn; pengfei.yan@ia.ac.cn

2. School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, 100083, China
E-mail: derong@ustb.edu.cn; ruizhuosong@ustb.edu.cn

Abstract: In this paper, a new data-based self-learning control scheme is developed to solve infinite horizon optimal control problems for continuous-time nonlinear systems. The developed optimal control scheme can be implement without knowing the mathematical model of the system. According to the input-output data of the nonlinear systems, a recurrent neural network (RNN) is employed to reconstruct the dynamics of the nonlinear system. According to the RNN model of the system, a new two-person zero-sum adaptive dynamic programming (ADP) algorithm is developed to obtain the optimal control, where the reconstruction error and the system disturbance are considered the control input of the system. Single-layer neural networks are used to construct the critic and action networks, which are presented to approximate the performance index function and the control law, respectively. Finally, simulation results will show the effectiveness of the developed data-based ADP methods.

Key Words: Adaptive critic designs, Adaptive dynamic programming, Approximate dynamic programming, Neuro-dynamic programming, Recurrent neural network, Data-based control, Optimal control

1 INTRODUCTION

Optimal control of nonlinear systems has always been the key focus of the control field in the latest several decades [1–3]. Adaptive dynamic programming (ADP) [4, 5] is a powerful brain-like intelligent optimal control method for nonlinear systems [6–8]. Iterative methods are widely used in ADP to obtain the solution of the Hamilton-Jacobi-Bellman (HJB) equation indirectly [9–12]. In [13], a complex-valued policy iteration algorithm was discussed, where for the first time the optimal control problem of complex-valued nonlinear systems was successfully solved by ADP. In [14], based on neurocognitive psychology, a novel policy iteration algorithm based on multiple actor-critic structures was developed for unknown systems and the proposed controller traded off fast actions based on stored behavior patterns with real-time exploration using current input-output data.

In most previous iterative ADP algorithms, the system model were generally required to update the iterative control law and the iterative value functions. However, accurate mathematical models of nonlinear systems are difficult to obtain. In this situation, recurrent neural network (RNN) is an effective tool to reconstruct the system dynamics [15–17] using the system input-output data. In [18], an affine-type RNN, inspired by [19], was proposed to reconstruct the nonlinear system, which makes the system control law can explicitly be expressed by the optimality principle.

In this paper, inspired by [18, 19], a new data-based ADP method for continuous-time unknown nonlinear systems with disturbance will be developed. For the unknown nonlinear system, a recurrent neural network is employed to reconstruct the system dynamics. According to the RNN model, the optimal control problem with disturbance is effectively transformed into a two-person zero-sum optimal control one. The optimal control law by ADP is achieved under the worst case disturbance. Single-layer neural networks (SLNNs) are introduced to approximate the performance index function, the system control law, and the disturbance control, respectively, for facilitating the implementation of the ADP method. Finally, simulation results will show the effectiveness of the developed data-based ADP methods.

2 Problem Formulation

In this paper, we consider the following general nonlinear continuous-time systems with disturbance

$$\dot{x} = \mathcal{F}(x, u) + \tau_d, \tag{1}$$

where $x \in \Omega_x$ is the n-dimensional state vector and $u \in \Omega_u$ is the m-dimensional control vector. Let Ω_x and Ω_u be the domains of definition for state and control, which are defined as $\Omega_x = \{x \,|\, x \in \mathbb{R}^n \text{ and } \|x\| < \infty\}$ and $\Omega_u = \{u \,|\, u \in \mathbb{R}^m \text{ and } \|u\| < \infty\}$, respectively, where $\|\cdot\|$ denotes the Euclidean norm [1]. Let $\mathcal{F}$ be an unknown smooth nonlinear function and τ_d be a finite n-dimensional measurable system disturbance. Let $\tau_d^B > 0$

This work was supported in part by the National Natural Science Foundation of China under Grants 61374105, 61233001, 61273140, and 61304079.

is a positive constant. Let Ω_w be the domains of definition for disturbance, which is defined as $\Omega_{\tau_d} = \{\tau_d \,|\, \tau_d \in \mathbb{R}^n \text{ and } \|\tau_d\| < \tau_d^B\}$.

According to [18, 19], the nonlinear system (1) can be reconstructed by the following RNN model

$$\dot{x} = \mathcal{A}^{\mathsf{T}} x + \mathcal{B}^{\mathsf{T}} f(x) + \mathcal{C}^{\mathsf{T}} u + \mathcal{D}^{\mathsf{T}} + \tau_\varepsilon + \tau_d. \quad (2)$$

Here, $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$ are the unknown ideal weight matrices, which are assumed to satisfy $\|\mathcal{A}\|_F \leq \mathcal{A}^B$, $\|\mathcal{B}\|_F \leq \mathcal{B}^B$, $\|\mathcal{C}\|_F \leq \mathcal{C}^B$, $\|\mathcal{D}\|_F \leq \mathcal{D}^B$, respectively, where is $\mathcal{A}^B$, $\mathcal{B}^B$, $\mathcal{C}^B$, $\mathcal{D}^B$ are all positive constants. Let the activation function $f(x)$ be a Lipschitz continuous function on Ω_x, i.e., $\forall x, y \in \Omega_x$, there exits a positive constant $\chi > 0$ that satisfies the following inequality

$$\|f(x) - f(y)\| \leq \chi \|x - y\|. \quad (3)$$

Let τ_ε be the finite approximate error, which satisfies $\|\tau_\varepsilon\| \leq \tau_\varepsilon^B$, where $\tau_\varepsilon^B > 0$ is a positive constant.

Based on (2), the data-based RNN model can be constructed as

$$\dot{\hat{x}} = \hat{\mathcal{A}}^{\mathsf{T}} \hat{x} + \hat{\mathcal{B}}^{\mathsf{T}} f(\hat{x}) + \hat{\mathcal{C}}^{\mathsf{T}} u + \hat{\mathcal{D}}^{\mathsf{T}} + \mathcal{E} z_m, \quad (4)$$

where $\hat{\mathcal{A}}$, $\hat{\mathcal{B}}$, $\hat{\mathcal{C}}$ and $\hat{\mathcal{D}}$ are the estimated weight matrices of the ideal unknown weight matrices $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$, respectively. Let $\mathcal{E}$ be a square matrix that satisfies

$$\lambda_{\min}\Big(\mathcal{E} - \mathcal{A}^{\mathsf{T}} - \frac{1}{2}\mathcal{B}^{\mathsf{T}}\mathcal{B}\Big) > \frac{1}{2}\chi^2, \quad (5)$$

where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix. Let the state estimation error be

$$z_m = x - \hat{x}. \quad (6)$$

Define the weight estimation error matrices as $\tilde{\mathcal{A}} = \mathcal{A} - \hat{\mathcal{A}}$, $\tilde{B} = B - \hat{\mathcal{B}}$, $\tilde{\mathcal{C}} = \mathcal{C} - \hat{\mathcal{C}}$, $\tilde{\mathcal{D}} = \mathcal{D} - \hat{\mathcal{D}}$, and define $\tilde{f}(z_m) = f(x) - f(\hat{x})$. According to [18], we can prove that the state estimation error z_m and the weight estimation error matrices $\tilde{\mathcal{A}}$, $\tilde{B}$, $\tilde{\mathcal{C}}$, and $\tilde{\mathcal{D}}$ are all UUB.

3 Data-Based Self-Learning Optimal Control for Unknown Nonlinear Systems with Disturbance Using ADP

In this section, adaptive dynamic programming (ADP) is developed to design the optimal controller for the unknown nonlinear system with disturbance. First, the optimal control problem for the nonlinear system with disturbance is transformed into a two-person zero-sum optimal control problem. Next, the detailed ADP implementation by neural networks are developed and the expressions of the optimal controller is obtained.

3.1 Derivation of the Zero-Sum Optimal Control Problem

Using RNN, we can see that as $t \to \infty$, the RNN-based system state $\hat{x}(t)$ will converge to a finite neighborhood of the ideal sate x. The matrices of the RNN, i.e., $\hat{\mathcal{A}}$, $\hat{\mathcal{B}}$, $\hat{\mathcal{C}}$, and $\hat{\mathcal{D}}$ will also converge to finite neighborhoods of the ideal matrices $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$, respectively. Hence, for $t \to \infty$, we can let $\lim_{t\to\infty} \hat{\mathcal{A}} = A$, $\lim_{t\to\infty} \hat{\mathcal{B}} = B$, $\lim_{t\to\infty} \hat{\mathcal{C}} = C$ and $\lim_{t\to\infty} \hat{\mathcal{D}} = D$, respectively, where A, B, C, and D are corresponding steady weight matrices.

Consequently, the nonlinear system (1) can be rewritten as

$$\dot{x} = A^{\mathsf{T}} x + B^{\mathsf{T}} f(x) + C^{\mathsf{T}} u + D^{\mathsf{T}} + w, \quad (7)$$

As $x \in \Omega_x$ and $u \in \Omega_u$ are both finite, then there exist positive constants $w_x^B > 0$, $w_u^B > 0$, $w_d^B > 0$ that satisfy $w_x^B = \sup_{x\in\Omega_x} \|(\mathcal{A} - A)x + (\mathcal{B} - B)f(x)\|$, $w_u^B = \sup_{u\in\Omega_u} \|(\mathcal{C} - C)u\|$, and $w_d^B \geq \|\mathcal{D} - D\|$, respectively. Hence, the parameter w in (7) can be seen as a finite system disturbance, which satisfies

$$\|w\| \leq w_x^B + w_u^B + w_d^B + \tau_\varepsilon^B + \tau_d^B. \quad (8)$$

Considering the disturbance w as a system control input, according to [20,21], the optimal control for system (7) can be transformed into a two-person zero-sum optimal control problem, where the performance index function can be defined as

$$V = \int_t^\infty (Q(x(s)) + u^{\mathsf{T}}(s) R u(s) - \gamma^2 w^{\mathsf{T}}(s) P w(s)) \mathrm{d}s. \quad (9)$$

In (9), we let $Q(x)$ be a positive definite function which satisfies $Q(x) \geq \mathcal{Q} x^{\mathsf{T}} x$ for a certain positive constant $\mathcal{Q}$. Let R and P are both positive definite matrices and let $\gamma > 0$ be a positive constant. According to (9), we can define the Hamilton function as

$$\begin{aligned} H(x, u, w, V) =& V_x^{\mathsf{T}}(A^{\mathsf{T}} x + B^{\mathsf{T}} f(x) + C^{\mathsf{T}} u + D^{\mathsf{T}} + w) \\ &+ Q(x) + u^{\mathsf{T}} R u - \gamma^2 w^{\mathsf{T}} P w, \end{aligned} \quad (10)$$

where $V_x = \dfrac{\mathrm{d}V}{\mathrm{d}x}$. Let $r(x, u, w) = Q(x) + u^{\mathsf{T}} R u - \gamma^2 w^{\mathsf{T}} P w$ be the utility function. To guarantee the existence of the optimal performance index function (saddle point), we assume that the following L_2-gain is less than or equal to γ.

Definition 1 *Let γ be certain prescribed level of disturbance attenuation. The system (7) is said to have L_2-gain less than or equal to γ if the inequality*

$$\int_t^\infty (Q(x(s)) + u^{\mathsf{T}}(s) R u(s)) ds \leq \gamma^2 \int_t^\infty (w^{\mathsf{T}}(s) P w(s)) \mathrm{d}s \quad (11)$$

holds for all w.

According to [20,21], the optimal performance index function can be defined as

$$\begin{aligned} V^* = \min_u \max_w \int_t^\infty & (Q(x(s)) + u^{\mathsf{T}}(s) R u(s) \\ & - \gamma^2 w^{\mathsf{T}}(s) P w(s)) \mathrm{d}s. \end{aligned} \quad (12)$$

Thus, we desire to find a state feedback control law such that the closed-loop system is stable and simultaneously

makes the performance index function (9) reach the optimum. According to [20, 21], for arbitrary γ that satisfies (11), we have the optimal performance index function V^* satisfies

$$\begin{aligned} V^* &= \min_u \max_w \int_t^\infty r(x,u,w)\mathrm{d}s \\ &= \max_w \min_u \int_t^\infty r(x,u,w)\mathrm{d}s. \end{aligned} \tag{13}$$

According to (10) and (13), the optimal control pair (u^*, w^*) satisfies the following Hamilton-Jacobi-Isaacs (HJI) equation

$$\begin{aligned} &H(x,u^*,w^*,V^*) \\ &\quad = V_x^{*T}\big(A^\mathsf{T}x + B^\mathsf{T}f(x) + C^\mathsf{T}u^* + D^\mathsf{T} \\ &\qquad + w^*\big) + Q(x) + u^{*T}Ru^* - \gamma^2 w^{*T}Pw^* \\ &\quad = 0. \end{aligned} \tag{14}$$

According to the principle of optimality, the optimal control pair (u^*, w^*) can be expressed as

$$\begin{aligned} u^* &= -\frac{1}{2}R^{-1}CV_x^*, \\ w^* &= \frac{1}{2\gamma^2}P^{-1}V_x^*. \end{aligned} \tag{15}$$

Generally, the optimal performance index function $V^*(x)$ is a non-analytical nonlinear function. It is nearly impossible to obtain $V^*(x)$ by solving the HJI equation (14). To overcome this problem, a new ADP algorithm for the zero-sum optimal control problem is developed. Three neural networks, which are critic, u-action and w-action networks are established to approximate the performance index function, the system control law u and the disturbance control law w, respectively, to implement the developed ADP algorithm.

3.2 Designs of Critic and Action Networks

In this subsection, based on the well-trained RNN (7), three single-layer neural networks (SINNs), which are critic, u-action and w-action networks, respectively, are introduced to implement the developed ADP algorithm.

3.2.1 Critic Network

The goal of the critic network is to approximate the performance index function. Based on the trained RNN model (7), the ideal function critic network can be expressed by

$$V = W_c^\mathsf{T}\psi_c + \tau_c, \tag{16}$$

where V is the performance index function and W_c is the ideal weight matrix of the critic network. Let $\psi_c\colon \mathbb{R}^n \to \mathbb{R}^{N_c}$, be the activation function, where N_c is the number of neurons in the hidden layer. Let τ_c be the finite approximation error of the critic network, which satisfies $\|\tau_c\| \le \tau_c^B$ for a certain positive constant τ_c^B.

Let $\hat{W}_c$ be the estimation weight matrix of W_c. Then, the actual output of the critic network can be expressed as

$$\hat{V} = \hat{W}_c^\mathsf{T}\psi_c, \tag{17}$$

where $\hat{V}$ is the estimation value of V. Define the weight estimation error of the critic network as

$$\tilde{W}_c = W_c - \hat{W}_c. \tag{18}$$

Define the approximate Hamilton function as

$$\begin{aligned} H(x,u,w,\hat{W}_c) &= \hat{W}_c^\mathsf{T}\nabla\psi_c\big(A^\mathsf{T}x + B^\mathsf{T}f(x) + C^\mathsf{T}u + D^\mathsf{T} \\ &\qquad + w\big) + Q(x) + u^\mathsf{T}Ru - \gamma^2 w^\mathsf{T}Pw \\ &= z_c. \end{aligned} \tag{19}$$

If we let

$$\tau_H = W_c^\mathsf{T}\nabla\psi_c\dot{x} + r, \tag{20}$$

then z_c in (19) can be rewritten as

$$z_c = -\tilde{W}_c^\mathsf{T}\nabla\psi_c\dot{x} + \tau_H, \tag{21}$$

The objective of the critic network is to select $\hat{W}_c$ which minimizes the following squared residual error

$$E_c = \frac{1}{2}z_c^2. \tag{22}$$

Based on the gradient descent rule, letting $\vartheta_1 = \nabla\psi_c\dot{x}$, the update law of the critic weight matrix can be expressed as

$$\dot{\hat{W}}_c = -l_c\frac{\partial E_c}{\partial \hat{W}_c} = -l_c\frac{\vartheta_1(\vartheta_1^\mathsf{T}\hat{W}_c + r)}{(\vartheta_1^\mathsf{T}\vartheta_1 + 1)^2}, \tag{23}$$

where $l_c > 0$ is the learning rate of the critic network. Define $\vartheta_2 = \dfrac{\vartheta_1}{\vartheta_3}$, where $\vartheta_3 = \vartheta_1^\mathsf{T}\vartheta_1 + 1$. Then we can obtain

$$\dot{\tilde{W}}_c = l_c\frac{\vartheta_1(\vartheta_1^\mathsf{T}\hat{W}_c + r)}{\vartheta_3^2} = -l_c\vartheta_2\vartheta_2^\mathsf{T}\tilde{W}_c + l_c\vartheta_2\frac{\tau_H}{\vartheta_3}. \tag{24}$$

3.2.2 u-Action Network

The goal of the u-action network is to compute the optimal feedback control law of system (7) with respect to the performance index function (9). The ideal function of the u-action network can be expressed by

$$u = W_a^\mathsf{T}\psi_a + \tau_a, \tag{25}$$

where W_a is the weight matrix of u-action network. Let $\psi_a\colon \mathbb{R}^n \to \mathbb{R}^{N_a}$ be the activation function vector, where N_a is the number of neurons in the hidden layer of u-action network. Let τ_a be the finite approximation error of u-action network, which satisfies $\|\tau_a\| \le \tau_a^B$ for a certain constant τ_a^B. Let $\hat{W}_a$ be the estimation weight matrix of W_a. Then, the actual output of the u-action network can be expressed as

$$\hat{u} = \hat{W}_a^\mathsf{T}\psi_a, \tag{26}$$

where $\hat{u}$ is the estimation vector of u. According to the Hamilton function (19), from $\dfrac{\partial H(x,u,w,\hat{W}_c)}{\partial u} = 0$, we can obtain the desired feedback optimal control law as

$$u = -\frac{1}{2}R^{-1}C\nabla\psi_c^\mathsf{T}\hat{W}_c. \tag{27}$$

Then, the approximation error of the u-action network can be defined as

$$z_a = \hat{W}_a^{\mathsf{T}}\psi_a + \frac{1}{2}R^{-1}C\nabla\psi_c^{\mathsf{T}}\hat{W}_c. \tag{28}$$

The objective of the u-action network is to select $\hat{W}_a$ which minimizes the following squared residual error

$$E_a = \frac{1}{2}z_a^2. \tag{29}$$

Based on the gradient descent algorithm, the update rule for the u-action network weight is given by

$$\dot{\hat{W}}_a = -l_a\psi_a(\hat{W}_a^{\mathsf{T}}\psi_a + \frac{1}{2}R^{-1}C\nabla\psi_c^{\mathsf{T}}\hat{W}_c)^{\mathsf{T}}, \tag{30}$$

where $l_a > 0$ is the learning rate of u-action network. If we define the weight estimation error of the u-action network as

$$\tilde{W}_a = W_a - \hat{W}_a, \tag{31}$$

then we can obtain

$$\begin{aligned}\dot{\tilde{W}}_a =& l_a\psi_a\big((W_a - \tilde{W}_a)^{\mathsf{T}}\psi_a + \frac{1}{2}R^{-1}C\nabla\psi_c^{\mathsf{T}}(W_c - \tilde{W}_c)\big)^{\mathsf{T}}\\ =& l_a\psi_a\big(-\tilde{W}_a^{\mathsf{T}}\psi_a - \frac{1}{2}R^{-1}C\nabla\psi_c^{\mathsf{T}}\tilde{W}_c + W_a^{\mathsf{T}}\psi_a\\ &+ \frac{1}{2}R^{-1}C\nabla\psi_c^{\mathsf{T}}W_c\big)^{\mathsf{T}}.\end{aligned} \tag{32}$$

3.2.3 w-Action Network

The goal of the w-action network is to approximate the optimal disturbance control law of system (7). The ideal function of the w-action network can be expressed by

$$w = W_d^{\mathsf{T}}\psi_d + \tau_w, \tag{33}$$

where W_d is the weight matrix of w-action network. Let $\psi_d\colon \mathbb{R}^n \to \mathbb{R}^{N_d}$ be the activation function vector, where N_d is the number of neurons in the hidden layer of w-action network. Let τ_w be the finite approximation error of w-action network, which satisfies $\|\tau_w\| \le \tau_w^B$, for a certain positive constant τ_w^B. Let $\hat{W}_d$ be the estimation weight matrix of W_d. Then, the actual output of the w-action network can be expressed as

$$\hat{w} = \hat{W}_d^{\mathsf{T}}\psi_d, \tag{34}$$

where $\hat{w}$ is the estimation vector of w. According to the Hamilton function (19), from $\dfrac{\partial H(x,u,w,\hat{W}_c)}{\partial w} = 0$, we can obtain the desired feedback optimal control law as

$$w = \frac{1}{2\gamma^2}P^{-1}\nabla\psi_c^{\mathsf{T}}\hat{W}_c. \tag{35}$$

Then, the approximation error of the w-action network can be defined as

$$z_d = \hat{W}_d^{\mathsf{T}}\psi_d - \frac{1}{2\gamma^2}P^{-1}\nabla\psi_c^{\mathsf{T}}\hat{W}_c. \tag{36}$$

The objective of the w-action network is to select $\hat{W}_d$ which minimizes the following squared residual error

$$E_d = \frac{1}{2}z_d^2. \tag{37}$$

Based on the gradient descent algorithm, the update rule for the w-action network weight is given by

$$\dot{\hat{W}}_d = -l_d\psi_d(\hat{W}_d^{\mathsf{T}}\psi_d - \frac{1}{2\gamma^2}P^{-1}\nabla\psi_c^{\mathsf{T}}\hat{W}_c)^{\mathsf{T}}, \tag{38}$$

where $l_d > 0$ is the learning rate of w-action network. If we define the weight estimation error of the w-action network as

$$\tilde{W}_d = W_d - \hat{W}_d, \tag{39}$$

then we can obtain

$$\begin{aligned}\dot{\tilde{W}}_d =& l_d\psi_d((W_d - \tilde{W}_d)^{\mathsf{T}}\psi_d - \frac{1}{2\gamma^2}P^{-1}\nabla\psi_c^{\mathsf{T}}(W_c - \tilde{W}_c))^{\mathsf{T}}\\ =& l_d\psi_d(-\tilde{W}_d^{\mathsf{T}}\psi_d + \frac{1}{2}P^{-1}\nabla\psi_c^{\mathsf{T}}\tilde{W}_c + W_d^{\mathsf{T}}\psi_d\\ &- \frac{1}{2\gamma^2}P^{-1}\nabla\psi_c^{\mathsf{T}}W_c)^{\mathsf{T}}.\end{aligned} \tag{40}$$

According to the two feedback control laws $\hat{u}$ and $\hat{w}$ in (25) and (34) obtained by ADP method, the feedback function of RNN model (7) can be expressed as

$$\begin{aligned}\dot{x} =& A^{\mathsf{T}}x + B^{\mathsf{T}}f(x) + C^{\mathsf{T}}\hat{u} + D^{\mathsf{T}} + \hat{w}\\ =& A^{\mathsf{T}}x + B^{\mathsf{T}}f(x) + C^{\mathsf{T}}(W_a - \tilde{W}_a)\psi_a + D^{\mathsf{T}}\\ &+ (W_d - \tilde{W}_d)\psi_d\\ =& A^{\mathsf{T}}x + B^{\mathsf{T}}f(x) + C^{\mathsf{T}}u + D^{\mathsf{T}} + w - C^{\mathsf{T}}\tilde{W}_a\psi_a\\ &- C^{\mathsf{T}}\tau_a - \tilde{W}_d\psi_d - \tau_w.\end{aligned} \tag{41}$$

According to (23), (30), and (38), the update laws of the critic, u-action, and w-action networks are established, and the ADP algorithm can be implemented by updating the performance index function, and the approximate control law pair $(\hat{u}, \hat{w})$. From (41), the feedback system under the control law pair $(\hat{u}, \hat{w})$ is also constructed. In next subsection, the convergence properties will be developed to show the effectiveness of the developed ADP method.

4 Simulation Study

In this section, we examine the performance of the developed method in a continuously stirred tank reactor system with an exothermic reaction [22]. The nonlinear system is given by

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}\frac{13}{6} & \frac{5}{12}\\ -\frac{50}{3} & -\frac{8}{3}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}-x_1\\ 0\end{bmatrix}u + w, \tag{42}$$

where $x = [x_1, x_2]^{\mathsf{T}}$. Define $\Omega_x = \{x \,|\, x \in \mathbb{R}^2, -1 \le x_1 \le 1, -1 \le x_2 \le 1\}$ and $\Omega_u = \{u \,|\, u \in \mathbb{R}, -3 < u < 3\}$. Let $w = [w_1, 0]$ be the system disturbance, where $|w_1| \le 1$. Let the utility function be expressed as $r(x,u,w) = x^{\mathsf{T}}Qx + u^{\mathsf{T}}Ru - \gamma^2 w^{\mathsf{T}}Pw$, where $Q = I$,

$R = 5I$, $P = I$, $\gamma^2 = 3.5$, and I is the identity matrix with suitable dimensions. First, a data-based model is established to estimate the nonlinear system dynamic. Let us select the RNN as (7) with the activation function $f(x)$ selected as hyperbolic tangent function $\tanh(x)$. Let the initial elements of matrices $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, $\mathcal{D}$ be randomly selected in $[-0.5, 0.5]$. The trajectories of the state estimation errors by RNN are shown in Fig. 1, where we can see that the state estimation errors are UUB around the equilibrium. Hence, we can see that the nonlinear system can be well approximated by the RNN.

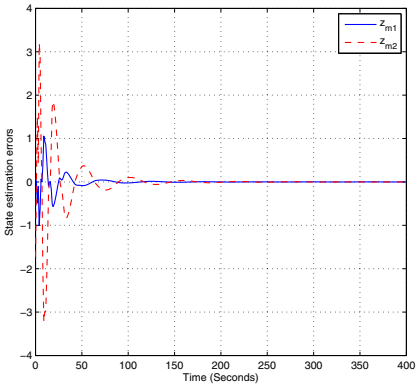

Figure 1: Trajectories of the state estimation errors by RNN

Next, ADP method is implemented to design the optimal control law of the system by SIANNs. Assume that the number of hidden layer neurons is denoted by l. The weight matrix between the input layer and hidden layer is denoted by Y. The weight matrix between the hidden layer and output layer is denoted by W. Let b denote the threshold vector of the neural network. Then, the output of SIANN is expressed by $\hat{F}(X, Y, W, b) = W^{\mathsf{T}}\sigma(Y^{\mathsf{T}}X + b)$, where $\sigma(Y^{\mathsf{T}}X + b) \in R^l$, and $\sigma(\cdot) = \tanh(\cdot)$ is the activation function. During the training procedure of SINN network, only the output weights W are updated during the training, while the hidden weights are arbitrarily chosen and then are kept fixed. Hence the output function of the SINN can be written as

$$\hat{F}(X, W) = W^{\mathsf{T}}\bar{\sigma}(X), \tag{43}$$

where $\bar{\sigma}(X) = \sigma(Y^{\mathsf{T}}X + b)$. We can see that a SINN can be effectively constructed by a two-layer networks. Let critic, u-action and w-action networks to approximate the performance index function, optimal control law and optimal disturbance control, respectively. Choose structures of critic, u-action and w-action networks as 2-8-1, 2-8-1 and 2-8-1, respectively. Let the weight matrices Y and b for critic, u-action and w-action networks are arbitrarily chosen and fixed. Letting the initial W weight matrices of the critic, u-action and w-action networks be zero, we implement the ADP method for $t_f = 400$ seconds. The convergence trajectories of the critic weight matrix are shown in Fig. 2. The convergence trajectories of the u-action and w-action weight matrices are shown in Figs. 3 and 4, respectively.

After 400 seconds, from Figs. 2–4, we can see that the critic, u-action and w-action networks are all converged. Choose the disturbance signal $w(t) = 0.5r_d(t)e^{-0.2t}\cos(t)$, where for $\forall\, t = 0, 1, \ldots$, $r_1(t)$ is a random number in $[-1, 1]$. Then, the optimal control trajectory is shown in Fig. 5 and the corresponding system states are shown in Fig. 6.

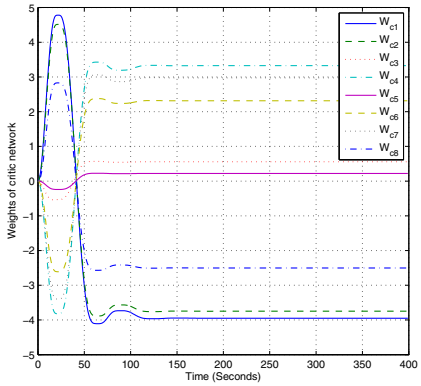

Figure 2: Weight convergence of the critic network

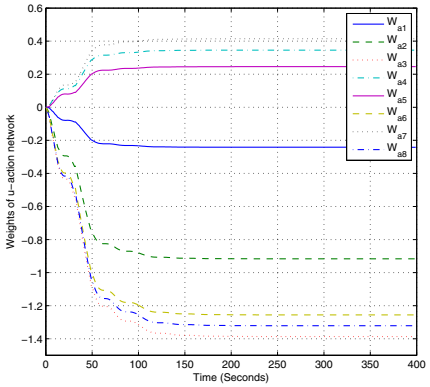

Figure 3: Weight convergence of the u-action network

5 Conclusions

In this paper, a data-based ADP method is developed to solve the optimal control problem for continuous-time unknown nonlinear systems with disturbance. An effective RNN model is used to establish the system dynamics. ADP method is developed to obtain the optimal control law of

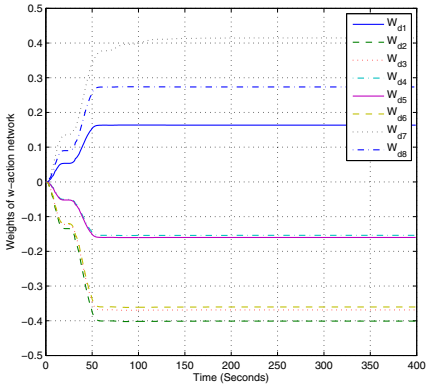

Figure 4: Weight convergence of the w-action network

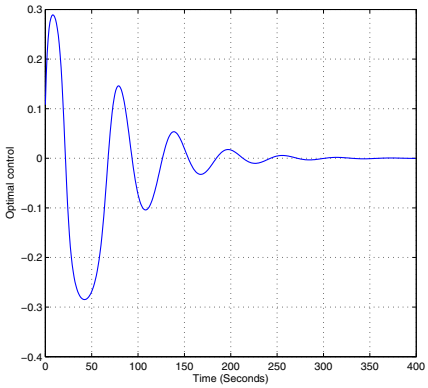

Figure 5: Trajectory of optimal control

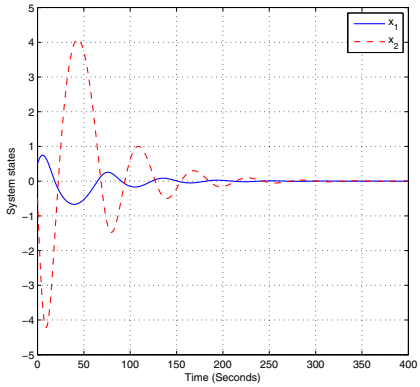

Figure 6: Trajectories of system states

the system based on the RNN model. Finally, numerical results are presented to demonstrate the effectiveness of the developed optimal control scheme.

REFERENCES

[1] F. L. Lewis, S. Jagannathan, and A. Yesildirek, *Neural network control of robot manipulators and nonlinear systems*. London, UK: Taylor and Francis, 1999.

[2] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780–1792, Jul. 2014.

[3] D. Liu, Y. Zhang, and H. Zhang, "A self-learning call admission control scheme for CDMA cellular networks," *IEEE Transactions on Neural Networks*, vol. 16, no. 5, pp. 1219–1228, Sep. 2005.

[4] P. J. Werbos, "Advanced forecasting methods for global crisis warning and models of intelligence," *General Systems Yearbook*, vol. 22, pp. 25–38, 1977.

[5] P. J. Werbos, "A menu of designs for reinforcement learning over time," in *Neural Networks for Control*, W.T. Miller, R.S. Sutton and P. J. Werbos, Ed., Cambridge: MIT Press, 1991, pp. 67–95.

[6] M. Fairbank, E. Alonso, and D. Prokhorov, "An equivalence between adaptive dynamic programming with a critic and backpropagation through time," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 12, pp. 2088–2100, Dec. 2013.

[7] Q. Wei, H. Zhang, and J. Dai, "Model-free multiobjective approximate dynamic programming for discrete-time nonlinear systems with general performance index functions," *Neurocomputing*, vol. 72, no. 7–9, pp. 1839–1848, Mar. 2009.

[8] H. Zhang, Q. Wei, and D. Liu, "An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games," *Automatica*, vol. 47, no. 1, pp. 207–214, Jan. 2011.

[9] M. Aurangzeb and F. L. Lewis, "Internal structure of coalitions in competitive and altruistic graphical coalitional games," *Automatica*, vol. 50, no. 2, pp. 335–348, Feb. 2014.

[10] B. Xu, C. Yang, and Z. Shi, "Reinforcement learning output feedback NN control using deterministic learning technique," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 635–641, Mar. 2014.

[11] D. Liu and Q. Wei, "Finite-approximation-error-based optimal control approach for discrete-time nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 779–789, Apri. 2013.

[12] Q. Wei and D. Liu, "Data-driven neuro-optimal temperature control of water gas shift reaction using stable iterative adaptive dynamic programming," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 11, pp. 6399–6408, Nov. 2014.

[13] R. Song, W. Xiao, H. Zhang, and C. Sun, "Adaptive dynamic programming for a class of complex-valued nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 9, pp. 1733–1739, Sep. 2014.

[14] R. Song, F. L. Lewis, Q. Wei, H. Zhang, Z. P. Jiang, and D. Levine, "Multiple actor-critic structures for continuous-time optimal control using input-output data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 4, pp. 851–865, Apr. 2015.

[15] Z. Yan and J. Wang, "Robust model predictive control of nonlinear systems with unmodeled dynamics and bounded uncertainties based on neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 457–469, Mar. 2014.

[16] M. Liu, S. Zhang, Z. Fan, S. Zheng, and We. Sheng, "Exponential H_∞ synchronization and state estimation for chaotic systems via a unified model," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 7, pp. 1114–1126, Jul. 2013.

[17] B. Zhang, D. J. Miller, and Y. Wang, "Nonlinear system modeling with random matrices: echo state networks revisited," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 1, pp. 175–182, Jan. 2012.

[18] 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 Transactions on Neural Networks*, vol. 22, no. 12, pp. 2226–2236, Dec. 2011.

[19] J. D. J. Rubio and W. Yu, "Stability analysis of nonlinear system identification via delayed neural networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 54, no. 2, pp. 161–165, Feb. 2007.

[20] T. Basar and P. Bernhard, *H∞ optimal control and related minimax design problems*. Boston, USA: Birkhäuser Press, 1995.

[21] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory (Second Edition)*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1999.

[22] R. Beard, *Improving the Closed-Loop Performance of Nonlinear Systems*. Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, NY, 1995.