International Journal of Control

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tcon20

Reinforcement learning-based optimal control of uncertain nonlinear systems

Miguel Garcia & Wenjie Dong

To cite this article: Miguel Garcia & Wenjie Dong (25 Jan 2024): Reinforcement learning-based optimal control of uncertain nonlinear systems, International Journal of Control, DOI: 10.1080/00207179.2024.2305727

To link to this article: https://doi.org/10.1080/00207179.2024.2305727

INTERNATIONAL JOURNAL OF CONTROL
https://doi.org/10.1080/00207179.2024.2305727

Taylor & Francis
Taylor & Francis Group

Reinforcement learning-based optimal control of uncertain nonlinear systems

Miguel Garcia and Wenjie Dong

Department of Electrical and Computer Engineering, The University of Texas Rio Grande Valley, Edinburg, TX, USA

ABSTRACT

This paper considers the optimal control of a second-order nonlinear system with unknown dynamics. A new reinforcement learning based approach is proposed with the aid of direct adaptive control. By the new approach, actor-critic reinforcement learning algorithms are proposed with neural network approximation. Simulation results are presented to show the effectiveness of the proposed algorithms.

ARTICLE HISTORY
Received 9 January 2023
Accepted 9 January 2024

KEYWORDS
Reinforcement learning; optimal control; nonlinear systems; uncertain systems; stabilisation

1. Introduction

Optimal control has lots of applications in control engineering, especially in space engineering. An optimal control problem can be solved with the aid of the dynamic programming (DP) or the Pontryagin's maximum principle. In the DP, by solving the Hamilton–Jacobi–Bellman (HJB) equation backward in time an optimal controller can be obtained. However, in practice it is extremely hard to solve the HJB equation because the HJB equation is a partial differential equation which contains the information of the dynamics of the system. In order to overcome this difficulty, approximate dynamic programming (ADP) and adaptive dynamic programming (ADP) have been proposed by Werbos (1992), Bertsekas (1995), Powell (2007).

Reinforcement learning (RL) is one of the effective methods to solve the optimal control problems. RL is inspired by natural learning mechanisms, where animals adjust their actions based on rewards and punishment stimuli received from the environment (Mendel & McLaren, 1970). In RL an actor or agent interacts with its environment and modifies its actions based on the stimuli received in response to its actions (Lewis, Vrabie & Vamvoudakis et al., 2012). A RL algorithm is designed based on the idea that a successful control decision should be a decision that increases the reward or decreases the punishment. RF learning algorithms have different forms in dealing with different optimal problems. RL can be applied to solve the optimal problems and the dynamic programming (DP) problems. Adaptive online controllers can be obtained. One type of RL algorithms employs the actor-critic structure. In this structure, the critic evaluates the reward or punishment based on the measured data and the actor finds an improved action and applies the action to the environment. Noting that DP problems can be solved by the approximate/adaptive dynamic programming (ADP) techniques, in literature the terms RL and ADP are used interchangeably (Liu et al., 2021).

RL has been studied for continuous-time systems under the assumption that the system model information is well-known in Doya (2000), Murray et al. (2001). The value iteration method and the policy iteration method have been proposed. However, in practice, the information of the model may not be available. In order to deal with this case, two types of methods have been proposed: the identifier-based RL (Bhasin et al., 2013) and the integral RL (Jiang & Jiang, 2012; Lewis, Vrabie & Syrmos et al., 2012; Lewis, Vrabie & Vamvoudakis et al., 2012; Vrabie & Lewis, 2009). In the identifier-based reinforcement learning (Bhasin et al., 2013) an identifier system is designed for the uncertain system and then reinforcement learning algorithms are proposed with the aid of the identifier system in which there is no uncertainty. In IRL, RL algorithms are proposed with the aid of integrating the value function over a period of time to partially circumvent or circumvent the unknown dynamics of the plant. In this method, there is no explicit identification of the unknown dynamics though there are calculations of some parameters using input-output data along the trajectory of the system state. In Vrabie et al. (2008), Vrabie et al. (2009), Vrabie and Lewis (2009), IRL algorithms are proposed when partial information of the dynamics is known. For linear systems with unknown dynamics an adaptive optimal algorithm is proposed by Jiang and Jiang (2012), Gao et al. (2022). For nonlinear systems with unknown dynamics, IRL algorithms are proposed in Jiang and Jiang (2014, 2015). In Gao et al. (2016), IRL algorithms are proposed for output feedback systems with unknown dynamics.

In this paper, we consider the optimal control of a second-order nonlinear system with partially unknown dynamics. A new reinforcement learning approach is proposed. In this approach, the idea of direct adaptive control is applied and the unknown dynamics is estimated by a neural network during the reinforcement learning controller design. In the proposed RL controller, three neural networks are designed for the actor, the critic, and the unknown dynamics, respectively. Compared with the identifier-actor-critic reinforcement learning and IRL, in our proposed reinforcement learning approach there is neither

CONTACT W. Dong wenjie.dong@utrgv.edu Department of Electrical and Computer Engineering, The University of Texas Rio Grande Valley, Edinburg 78539, TX, USA

explicit identification of the unknown plant nor integrating the value function over a period of time. Furthermore, the proposed approach can be extended to solve the optimal control problem of more general nonlinear systems.

The organisation of the remaining part of this paper is as follows. In Section 2, the problem considered in this paper is defined. In Section 3, the reinforcement learning algorithm is proposed. The simulation is presented in Section 4. The last section concludes this paper.

2. Problem statement

Consider the following second-order nonlinear system

$$\dot{x}_1 = x_2 \tag{1}$$

$$\dot{x}_2 = f(x) + g(x)u \tag{2}$$

where $x_1 \in R^n$ and $x_2 \in R^n$ are the states, $x^\top = [x_1^\top, x_2^\top]^\top \in \Omega \subset R^{2n}$, $u \in U \subset R^n$ is the input, $f(x) \in R^n$ with $f(0) = 0$ being an unknown vector function, and $g(x) \in R^n$ is a known input matrix function. It is assumed that $f(x) + g(x)u$ is Lipschitz continuous and the system (1)-(2) is stabilisable. For the purpose of control, it is assumed that $g(x)$ is nonsingular for any state x. Let

$$F = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ f(x)a \end{bmatrix}, \quad G = \begin{bmatrix} \mathbf{0}_{n\times n} \\ g(x) \end{bmatrix}$$

(1)–(2) can be written in a compact form

$$\dot{x} = F(x) + G(x)u. \tag{3}$$

The control problem considered in this paper is defined as follows.

Optimal Control Problem: Design a control law u for system (1)–(2) such that the state x converges to zero and the cost

$$J = \int_0^\infty (Q(x) + u^\top Pu)\mathrm{d}\tau \tag{4}$$

is minimised, where $Q(x) \in R$ is a positive definite function of x and P is a positive definite matrix.

In order to solve the optimal problem, the value function at time t for an input u and the state $x(t)$ is defined as

$$V(x(t), u(t)) = \int_t^\infty (Q(x) + u^\top Pu)\mathrm{d}\tau. \tag{5}$$

The Hamiltonian function corresponding to the above optimal control problem is

$$H(x, u, V) = \nabla V^\top (F + Gu) + Q(x) + u^\top Pu. \tag{6}$$

Let V^* be the value function, i.e.

$$V^*(x(t)) = \inf_u \int_t^\infty (Q + u^\top Pu)\mathrm{d}\tau. \tag{7}$$

The optimal control u^* for system (1)-(2) with the cost function (4) can be obtained with the aid of the Hamiltonian (6) as

$$u^* = \arg\min_u H(x, u, V) = -\frac{1}{2}P^{-1}G^\top \nabla_x V^*. \tag{8}$$

The value function and the optimal control satisfy the following Hamilton–Jacobi–Bellman (HJB) equation

$$H^*(x, u^*, V^*) = \nabla_x V^{*\top}(F + Gu^*) + Q(x) + u^{*\top}Pu^* \tag{9}$$

$$= Q(x) + \nabla_x V^{*\top} F - \frac{1}{4}\nabla_x V^{*\top} GP^{-1}G^\top \nabla_x V^*$$

$$= 0. \tag{10}$$

In order to apply the optimal control law (8), it is required to solve the nonlinear partial differential Equation (10). Generally, it is impossible to find the analytic solution V^*. To overcome this difficulty, iterative methods have been proposed in the past decades. For example, if $f(x)$ is known the optimal controller can be obtained by solving (10) with the aid of the value iteration (VI) or the policy iteration (PI) algorithms (Lewis, Vrabie & Syrmos et al., 2012). If $f(x)$ is unknown, the VI and PI iteration methods do not work. In order to solve the optimisation problem, the identifier-actor-critic reinforcement learning-based algorithms and the IRL algorithms have been proposed. In this paper, we propose a new actor-critic reinforcement learning-based algorithm with the aid of the idea of direct adaptive control.

3. Actor-critic reinforcement learning controller design

In order to solve the optimal control problem, the following assumptions are made (Vamvoudakis & Lewis, 2010).

Assumption 3.1: *The solution V^* to (10) is smooth and positive definite.*

Assumption 3.2: *$\|f(x)\| \le \gamma_1 x^\top x + \gamma_2 \|x\|$, where γ_1 and γ_2 are non-negative constants.*

The value function can be written as

$$V^* = V_1^* + V_2^* \tag{11}$$

where

$$V_1^* = 2\int_0^{x_2} [g^{-\top}(x_1, \tau)Pg^{-1}(x_1, \tau)f(x_1, \tau)]^\top \mathrm{d}\tau$$

$$V_2^* = V^* - V_1^*$$

and we use the notations $f(x) = f(x_1, x_2)$ and $g(x) = g(x_1, x_2)$. V_1^* is chosen in this way because we want to make sure there is one term to cancel f in the optimal control. Since $f(x_1, x_2)$ is smooth, with the aid of the universal approximation theorems of functions in Cybenko (1989), Hornik et al. (1989), and Stone (1948), there exists a vector S_f such that

$$f(x) = W_f^\top S_f(x) + \epsilon_f(x) \tag{12}$$

where W_f is the ideal weight vector and ϵ_f is the residue error and can be made as small as possible by choosing the basis matrix S_f carefully.

Since V^* is smooth, V_2^* is also smooth. There exists a vector ϕ such that

$$V_2^* = W_V^\top \phi(x) + \epsilon(x) \tag{13}$$

where W_V is the ideal weight vector and ϵ is the residue error and can be made as small as possible by choosing the basis vector ϕ carefully.

The gradient of V^* is

$$\begin{aligned}\nabla_x V^* &= \nabla_x V_1^* + \nabla_x V_2^* \\ &= 2\Lambda + \nabla_x \phi^\top W_V + \nabla_x \epsilon \\ &= 2\Lambda + S_V^\top W_V + \epsilon_V \end{aligned} \tag{14}$$

where

$$\begin{aligned}\Lambda &= 0.5\nabla_x V_1^* = \begin{bmatrix} 0.5\nabla_{x_1} V_1^* \\ g^{-\top} P g^{-1} f \end{bmatrix} \\ &= \begin{bmatrix} Y^\top W_f + \epsilon_2 \\ g^{-\top} P g^{-1} (S_f^\top W_f + \epsilon_f) \end{bmatrix} \\ S_V &= \nabla_x \phi \\ \epsilon_V &= \nabla_x \epsilon \\ Y &= \int_0^{x_2} \nabla_{x_1} [S_f(x_1, \tau) g^{-\top}(x_1, \tau) P g^{-1}(x_1, \tau)] \mathrm{d}\tau \\ \epsilon_2 &= \nabla_{x_1} \int_0^{x_2} [g^{-\top}(x_1, \tau) P g^{-1}(x_1, \tau) \epsilon_f(x_1, \tau)]^\top \mathrm{d}\tau \end{aligned}$$

and we apply the notations $S_f(x) = S_f(x_1, x_2)$ and $\epsilon_f(x) = \epsilon_f(x_1, x_2)$. The optimal control input is

$$\begin{aligned} u^* &= -\frac{1}{2} P^{-1} G^\top (2\Lambda + S_V^\top W_V + \epsilon_V) \\ &= -P^{-1} G^\top \Lambda - \frac{1}{2} P^{-1} G^\top S_V^\top W_V - \frac{1}{2} P^{-1} G^\top \epsilon_V. \end{aligned} \tag{15}$$

The Hamiltonian in (9) can be written as

$$\begin{aligned} &H^*(x, u^*, \nabla_x V^*) \\ &= Q(x) + (u^*)^\top P u^* + [2\Lambda + S_V^\top W_V + \epsilon_V]^\top \\ &\quad [F - G P^{-1} G^\top \Lambda \\ &\quad - \frac{1}{2} G P^{-1} G^\top S_V^\top W_V - \frac{1}{2} G P^{-1} G^\top \epsilon_V] \end{aligned} \tag{16}$$

$$\begin{aligned} &= Q(x) + (u^*)^\top P u^* + [2\Lambda + S_V^\top W_V \\ &\quad + \epsilon_V]^\top [F - \bar{P}\Lambda - \frac{1}{2} \bar{P} S_V^\top W_V - \frac{1}{2} \bar{P} \epsilon_V] \end{aligned} \tag{17}$$

where

$$\bar{P} = G P^{-1} G^\top = \begin{bmatrix} \mathbf{0}_{n\times n} & \mathbf{0} \\ \mathbf{0} & g P^{-1} g^\top \end{bmatrix}.$$

Since ϵ_V, W_f, and W_V are unknown, it is impossible to implement the optimal control u^* in (15). In order to make the control input implementable, we employ the actor-critic architecture of the reinforcement learning to implement the controller (15).

Let the estimate of the unknown function $f(x)$ be

$$\hat{f}(x) = \hat{W}_f^\top S_f(x) \tag{18}$$

where $\hat{W}_f$ is an estimate of W_f and will be proposed later. Let an estimate of the gradient of V^* be

$$\nabla_x \hat{V}^* = 2\hat{\Lambda} + S_V^\top W_c \tag{19}$$

where W_c is an estimate of W_V for the *critic* (13) and will be proposed later, and

$$\hat{\Lambda} = \begin{bmatrix} \hat{\Lambda}_1 \\ \hat{\Lambda}_2 \end{bmatrix} = \begin{bmatrix} Y^\top \hat{W}_f \\ g^{-\top} P g^{-1} S_f^\top \hat{W}_f \end{bmatrix}.$$

With the aid of the approximations of $f(x)$ and $\nabla_x V^*$, the optimal control input is

$$\begin{aligned} u &= -P^{-1} G^\top \hat{\Lambda} - \frac{1}{2} P^{-1} G^\top S_V^\top W_a \\ &= -P^{-1} g^\top \hat{\Lambda}_2 - \frac{1}{2} P^{-1} G^\top S_V^\top W_a \end{aligned} \tag{20}$$

where W_a is an estimate of W_V for the *actor* (15) and will be proposed later.

With the aid of (18)–(20), the approximation of the Hamiltonian in (9) is

$$\begin{aligned} &H^*(x, u, \nabla_x \hat{V}^*) \\ &= Q + u^\top P u + (2\hat{\Lambda} + S_V^\top W_c)(\hat{F} + G u) \\ &= Q(x) + \hat{\Lambda}^\top \bar{P} \hat{\Lambda} + \hat{\Lambda}^\top \bar{P} S_V^\top W_a + \frac{1}{4} W_a^\top S_V \bar{P} S_V^\top W_a \\ &\quad + 2\hat{\Lambda}^\top (\hat{F} - \bar{P}\hat{\Lambda}) - \hat{\Lambda}^\top \bar{P} S_V^\top W_a + \hat{W}_c^\top \xi \end{aligned}$$

where $\hat{F} = [\ x_2^\top, (S_f^\top \hat{W}_f)^\top\ ]^\top$ and $\xi = S_V(\hat{F} + G u) = S_V[\hat{F} - \bar{P}\hat{\Lambda} - \frac{1}{2}\bar{P} S_V^\top W_a]$.

The Bellman residue error is defined as

$$\begin{aligned} z &= H^*(x, u, \nabla \hat{V}^*) - H^*(x, u^*, \nabla V^*) \\ &= H^*(x, u, \nabla \hat{V}^*). \end{aligned} \tag{21}$$

In order to solve the problem, the following assumption is made.

Assumption 3.3 ((Uniform Approximations)): *The vector functions S_f and S_V, the value function approximation errors ϵ_f, and ϵ_V, and the Hamiltonian residual error z are all uniformly bounded on the set $\Omega \subset R^{2n}$, in the sense that there exist finite positive constants δ_f, δ_V, δ_z, α_V, α_g, and α_f such that $\|S_f\| \le \alpha_f$, $\|g^{-\top} P g^{-1}\| \le \alpha_g$, $\|S_V\| \le \alpha_V$, $\|\epsilon_f\| \le \delta_f$, $\|\epsilon_V\| \le \delta_V$, and $|z| \le \delta_z$.*

In order to find the critic update law W_c, we minimise the residue error z^2 by the gradient descent method. The update law W_c is proposed as

$$\dot{W}_c = -k_c \frac{\partial z^2}{\partial W_c} = -2k_c z \frac{\partial z}{\partial W_c} = -2k_c \xi z \tag{22}$$

where k_c is a positive constant.

In order to find the actor update laws W_a and $\hat{W}_f$, we choose a Lyapunov function

$$V_3 = V^* + \frac{1}{2}\tilde{W}_a^\top \Gamma_a \tilde{W}_a + \frac{1}{2}\tilde{W}_c^\top \Gamma_c \tilde{W}_c + \frac{1}{2}\tilde{W}_f^\top \Gamma_f \tilde{W}_f$$

where $\tilde{W}_a = W_V - W_a$, $\tilde{W}_c = W_V - W_c$, $\tilde{W}_f = W_f - \hat{W}_f$, Γ_a, Γ_f, and Γ_c are positive definite matrices. With the control law (20), we have

$$\begin{aligned}
\dot{V}_3 &= \nabla V^{*\top}(F + Gu^*) + \nabla V^{*\top} G(u - u^*) \\
&\quad + \tilde{W}_a^\top \Gamma_a \dot{\tilde{W}}_a + \tilde{W}_c^\top \Gamma_c \dot{\tilde{W}}_c + \tilde{W}_f^\top \Gamma_f \dot{\tilde{W}}_f \\
&= -Q - (u^*)^\top P u^* + (2\Lambda + S_V^\top W_V + \epsilon_V)^\top \\
&\quad \bar{P}(\tilde{\Lambda} + 0.5 S_V^\top \tilde{W}_a + 0.5\epsilon_V) \\
&\quad + \tilde{W}_a^\top \Gamma_a \dot{\tilde{W}}_a + \tilde{W}_c^\top \Gamma_c \dot{\tilde{W}}_c + \tilde{W}_f^\top \Gamma_f \dot{\tilde{W}}_f \\
&= -Q - (u^*)^\top P u^* + (2\hat{\Lambda} + S_V^\top W_c)^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} S_f^\top \tilde{W}_f \\
&\quad + 0.5(2\hat{\Lambda} + S_V^\top W_c)^\top \bar{P} S_V^\top \tilde{W}_a + 2\tilde{\Lambda}^\top \bar{P} \tilde{\Lambda} \\
&\quad + \tilde{\Lambda}^\top \bar{P} S_V^\top \tilde{W}_a + \tilde{W}_c^\top S_V \bar{P} \tilde{\Lambda} \\
&\quad + 0.5 \tilde{W}_c^\top S_V \bar{P} S_V^\top \tilde{W}_a + \Lambda^\top \bar{P} \epsilon_V + 0.5 W_V^\top S_V \bar{P} \epsilon_V \\
&\quad + \epsilon_V^\top \bar{P} \tilde{\Lambda} + 0.5 \epsilon_V^\top \bar{P} S_V^\top \tilde{W}_a \\
&\quad + 0.5 \epsilon_V^\top \bar{P} \epsilon_V + \tilde{W}_a^\top \Gamma_a \dot{\tilde{W}}_a + \tilde{W}_c^\top \Gamma_c \dot{\tilde{W}}_c + \tilde{W}_f^\top \Gamma_f \dot{\tilde{W}}_f.
\end{aligned}$$

We choose

$$\dot{\hat{W}}_f = \Gamma_f^{-1} S_f [\mathbf{0}, \mathbf{I}] \left(2\hat{\Lambda} + S_V^\top W_c\right) - k_f \Gamma_f^{-1} S_f S_f^\top \hat{W}_f \tag{23}$$

$$\begin{aligned}
\dot{W}_a &= \frac{1}{2}\Gamma_a^{-1} S_V \bar{P} \left(2\hat{\Lambda} + S_V^\top W_c\right) - k_a \Gamma_a^{-1} S_V S_V^\top (W_a - W_c) \\
&\quad - k_e \Gamma_a^{-1} S_V S_V^\top W_a
\end{aligned} \tag{24}$$

and modify the update law for W_c in (22) as follows:

$$\begin{aligned}
\dot{W}_c &= -2k_c \Gamma_c^{-1} \xi z - k_a \Gamma_c^{-1} S_V S_V^\top (W_c - W_a) \\
&\quad - \Gamma_c^{-1} k_d S_V S_V^\top W_c.
\end{aligned} \tag{25}$$

Then

$$\begin{aligned}
\dot{V}_3 &= -Q + \tilde{W}_f^\top S_f [\mathbf{0}, \mathbf{I}] S_V^\top \tilde{W}_a + 2\tilde{W}_f^\top S_f g^{-\top} P g^{-1} S_f^\top \tilde{W}_f \\
&\quad + \tilde{W}_c^\top S_V \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} S_f^\top \tilde{W}_f \\
&\quad - (u^*)^\top P u^* + \frac{1}{2}\tilde{W}_c^\top S_V \bar{P} S_V^\top \tilde{W}_a + f^\top [\mathbf{0}, \mathbf{I}] \epsilon_V \\
&\quad + \frac{1}{2} W_V^\top S_V \bar{P} \epsilon_V + \epsilon_V^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} S_f^\top \tilde{W}_f \\
&\quad + 0.5 \epsilon_V^\top \bar{P} S_V^\top \tilde{W}_a + 0.5 \epsilon_V^\top \bar{P} \epsilon_V \\
&\quad - k_a \tilde{W}_a^\top S_V S_V^\top \tilde{W}_a - k_a \tilde{W}_c^\top S_V S_V^\top \tilde{W}_c \\
&\quad + k_a \tilde{W}_a^\top S_V S_V^\top \tilde{W}_c + k_a \tilde{W}_c^\top S_V S_V^\top \tilde{W}_a \\
&\quad - \frac{k_f}{2} \tilde{W}_f^\top S_f S_f^\top \tilde{W}_f - \frac{k_f}{2} \hat{W}_f^\top S_f S_f^\top \hat{W}_f \\
&\quad + \frac{k_f}{2} W_f^\top S_f S_f^\top W_f - \frac{k_e}{2} \tilde{W}_a^\top S_V S_V^\top \tilde{W}_a \\
&\quad - \frac{k_e}{2} W_a^\top S_V S_V^\top W_a + \frac{k_e}{2} W_V^\top S_V S_V^\top W_V \\
&\quad - \frac{k_d}{2} \tilde{W}_c^\top S_V S_V^\top \tilde{W}_c - \frac{k_d}{2} W_c^\top S_V S_V^\top W_c \\
&\quad + \frac{k_d}{2} W_V^\top S_V S_V^\top W_V + 2k_c z \tilde{W}_c^\top S_V \begin{bmatrix} x_2 \\ \mathbf{0} \end{bmatrix} \\
&\quad + k_c z \tilde{W}_c^\top S_V \bar{P} S_V^\top \tilde{W}_a - k_c z \tilde{W}_c^\top S_V \bar{P} S_V^\top W_V.
\end{aligned}$$

Since $Q(x)$ is positive definite, there exists a positive matrix q such that $Q(x) \geq x^\top \bar{Q} x$ for $x \in \Omega$. Then

$$\begin{aligned}
\dot{V}_3 &\leq -x^\top \bar{Q} x - (u^*)^\top P u^* + \tilde{W}_f^\top S_f [\mathbf{0}, \mathbf{I}] S_V^\top \tilde{W}_a \\
&\quad + 2\tilde{W}_f^\top S_f g^{-\top} P g^{-1} S_f^\top \tilde{W}_f + \tilde{W}_c^\top S_V \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} S_f^\top \tilde{W}_f \\
&\quad + 0.5 \tilde{W}_c^\top S_V \bar{P} S_V^\top \tilde{W}_a + \gamma_1 \delta_V x^\top x + \gamma_2 \|x\| \delta_V \\
&\quad + 0.5 W_V^\top S_V \bar{P} \epsilon_V + \epsilon_V^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} S_f^\top \tilde{W}_f \\
&\quad + 0.5 \epsilon_V^\top \bar{P} S_V^\top \tilde{W}_a + 0.5 \epsilon_V^\top \bar{P} \epsilon_V - k_a \tilde{W}_a^\top S_V S_V^\top \tilde{W}_a \\
&\quad - k_a \tilde{W}_c^\top S_V S_V^\top \tilde{W}_c \\
&\quad + k_a \tilde{W}_a^\top S_V S_V^\top \tilde{W}_c + k_a \tilde{W}_c^\top S_V S_V^\top \tilde{W}_a \\
&\quad - \frac{k_f}{2} \tilde{W}_f^\top S_f S_f^\top \tilde{W}_f - \frac{k_f}{2} \hat{W}_f^\top S_f S_f^\top \hat{W}_f \\
&\quad + \frac{k_f}{2} W_f^\top S_f S_f^\top W_f - \frac{k_e}{2} \tilde{W}_a^\top S_V S_V^\top \tilde{W}_a \\
&\quad - \frac{k_e}{2} W_a^\top S_V S_V^\top W_a + \frac{k_e}{2} W_V^\top S_V S_V^\top W_V \\
&\quad - \frac{k_d}{2} \tilde{W}_c^\top S_V S_V^\top \tilde{W}_c - \frac{k_d}{2} W_c^\top S_V S_V^\top W_c \\
&\quad + \frac{k_d}{2} W_V^\top S_V S_V^\top W_V + 2k_c z \tilde{W}_c^\top S_V \begin{bmatrix} x_2 \\ \mathbf{0} \end{bmatrix} \\
&\quad + k_c z \tilde{W}_c^\top S_V \bar{P} S_V^\top \tilde{W}_a - k_c z \tilde{W}_c^\top S_V \bar{P} S_V^\top W_V.
\end{aligned}$$

Let

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} x \\ S_f^\top \tilde{W}_f \\ S_V^\top \tilde{W}_c \\ S_V^\top \tilde{W}_a \end{bmatrix}$$

then

$$\begin{aligned}
\dot{V}_3 &\leq -y_1^\top \bar{Q} y_1 - (u^*)^\top P u^* + 2\alpha_g y_2^\top y_2 \\
&\quad + y_2^\top [\mathbf{0}, \mathbf{I}] y_4 + y_3^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} y_2 \\
&\quad + 0.5 y_3^\top \bar{P} y_4 + \gamma_1 \delta_V y_1^\top y_1 + \gamma_2 \delta_V \|y_1\| - k_a y_4^\top y_4 \\
&\quad - k_a y_3^\top y_3 + 2k_a y_4^\top y_3 \\
&\quad - \frac{k_f}{2} y_2^\top y_2 - \frac{k_e}{2} y_4^\top y_4 - \frac{k_d}{2} y_3^\top y_3 + 2k_c z y_3^\top \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} y_1 \\
&\quad + k_c z y_3^\top \bar{P} y_4
\end{aligned}$$

$$
\begin{aligned}
&- k_c z y_3^\top \bar{P} S_V^\top W_V + 0.5 W_V^\top S_V \bar{P} \epsilon_V + \epsilon_V^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} y_2 \\
&+ 0.5 \epsilon_V^\top \bar{P} y_4 + 0.5 \epsilon_V^\top \bar{P} \epsilon_V \\
&- \frac{k_f}{2} \hat{W}_f^\top S_f S_f^\top \hat{W}_f + \frac{k_f}{2} W_f^\top S_f S_f^\top W_f \\
&- \frac{k_e}{2} W_a^\top S_V S_V^\top W_a + \frac{k_e}{2} W_V^\top S_V S_V^\top W_V \\
&- \frac{k_d}{2} W_c^\top S_V S_V^\top W_c + \frac{k_d}{2} W_V^\top S_V S_V^\top W_V \\
\le\ & - y_1^\top \bar{Q} y_1 - (u^*)^\top P u^* + 2\alpha_g y_2^\top y_2 \\
&+ y_2^\top [\mathbf{0}, \mathbf{I}] y_4 + y_3^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} y_2 \\
&+ 0.5 y_3^\top \bar{P} y_4 + \gamma_1 \delta_V y_1^\top y_1 + \gamma_2 \delta_V \|y_1\| \\
&- k_a y_4^\top y_4 - k_a y_3^\top y_3 + 2 k_a y_4^\top y_3 \\
&- \frac{k_f}{2} y_2^\top y_2 - \frac{k_e}{2} y_4^\top y_4 - \frac{k_d}{2} y_3^\top y_3 \\
&+ k_c \delta_z y_3^\top y_3 + k_c \delta_z y_1^\top y_1 + 0.5 k_c \delta_z y_3^\top y_3 + 0.5 k_c \delta_z y_4^\top y_4 \\
&- k_c z y_3^\top \bar{P} S_V^\top W_V + 0.5 W_V^\top S_V \bar{P} \epsilon_V \\
&+ \epsilon_V^\top \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} y_2 + 0.5 \epsilon_V^\top \bar{P} y_4 + 0.5 \epsilon_V^\top \bar{P} \epsilon_V \\
&- \frac{k_f}{2} \hat{W}_f^\top S_f S_f^\top \hat{W}_f + \frac{k_f}{2} W_f^\top S_f S_f^\top W_f \\
&- \frac{k_e}{2} W_a^\top S_V S_V^\top W_a + \frac{k_e}{2} W_V^\top S_V S_V^\top W_V \\
&- \frac{k_d}{2} W_c^\top S_V S_V^\top W_c + \frac{k_d}{2} W_V^\top S_V S_V^\top W_V \\
\le\ & - y^\top H y + y^\top D + \gamma_2 \delta_V \|y_1\| + \epsilon_1 \qquad (26)
\end{aligned}
$$

where

$$
H = \begin{bmatrix} H_{11} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{k_f - 4\alpha_g}{2} I & H_{23} & H_{24} \\ \mathbf{0} & H_{23}^\top & H_{33} & H_{34} \\ \mathbf{0} & H_{24}^\top & H_{34}^\top & H_{44} \end{bmatrix}
$$

$$
\begin{aligned}
H_{11} &= \bar{Q} - \gamma_1 \delta_V I - k_c \delta_z, \quad H_{23} = -[\mathbf{0}, 0.5\mathbf{I}] \\
H_{24} &= -[\mathbf{0}, 0.5\mathbf{I}] \\
H_{34} &= -\frac{0.5\bar{P} + 2k_a I}{2} \\
H_{33} &= (k_a + 0.5 k_d - 1.5 k_c \delta_z) I \\
H_{44} &= (k_a + 0.5 k_e - 0.5 k_c \delta_z) I
\end{aligned}
$$

$$
D = \begin{bmatrix} \mathbf{0} \\ [\mathbf{0}, \mathbf{I}]\, \epsilon_V \\ -k_c z \bar{P} S_V^\top W_V \\ 0.5 \bar{P} \epsilon_V \end{bmatrix}
$$

$$
\epsilon_1 = 0.5 \epsilon_V^\top \bar{P} \epsilon_V + \frac{k_f}{2} W_f^\top S_f S_f^\top W_f
$$

$$
+ \frac{k_e}{2} W_V^\top S_V S_V^\top W_V + \frac{k_d}{2} W_V^\top S_V S_V^\top W_V + 0.5 W_V S_V \bar{P} \epsilon_V
$$

Based on the above procedure, we have the following results.

Theorem 3.4: *The controller in (20) with the critic-actor weight update laws in (24)–(25) and the learning update law (23) of uncertainty $f(x)$ in the model ensure that the state x and the weight errors $\tilde{W}_c$, $\tilde{W}_a$, and $\tilde{W}_f$ uniformly ultimately bounded (UUB) if the control parameters are chosen such that H is positive definite. Furthermore, the state x can be made as small as possible by choosing large control parameters.*

Proof: Based on (26), we have

$$
\begin{aligned}
\dot{V}_3 &\le -\|y\|^2 \sigma_m + \|D\| \|y\| + \gamma_2 \delta_V \|y_1\| + \|\epsilon_1\| \\
&\le -\|y\|^2 \sigma_m + (\|D\| + \gamma_2 \delta_V) \|y\| + \|\epsilon_1\|.
\end{aligned}
$$

where σ_m is the smallest eigenvalue of H. If

$$
\|y\| > \frac{\|D\| + \delta_V \gamma_2}{2\sigma_m} + \sqrt{\frac{(\|D\| + \delta_V \gamma_2)^2}{4\sigma_m^2} + \frac{\|\epsilon_1]}{\sigma_m}},
$$

$\dot{V}_3$ is negative. Therefore, the state and the estimate errors of the weights W_f, W_a, and W_c are UUB. Furthermore, the state x can be made as small as possible by choosing large control parameters.

The block diagram of the proposed controller is shown in Figure 1. Different from the Identifier-Actor-Crtic RL and the IRL, in Theorem 1 the unknown dynamics $f(x)$ is estimated online with the aid of direct adaptive theory. An identifier system is not required.

In (12) and (13) the vectors S_f and ϕ should be chosen carefully such that ϵ_f and ϵ are small. One can choose each element of S_f and ϕ to be a sigmoid function with appropriate weights (see Cybenko, 1989; Hornik et al., 1989) or high-order polynomials of x (see Stone, 1948).

In this paper, we considered the second-order nonlinear system. The proposed method can be applied to the optimal control of the high-order nonlinear system. ■

4. Simulation

In order to show the effectiveness of the proposed results, two examples are considered.

Example 4.1: Consider a second-order system in (1)–(2) where

$$
f(x) = -(x_1 + x_2)\left(\frac{9}{4} - \frac{\cos 2(x_1 + x_2)}{2}\right), \quad g(x) = 1.
$$

In the optimal control problem, we choose $P = 1$ and

$$
Q(x) = x_1^2 + (x_1 + x_2)^2 + (x_1 + x_2)^2 \sin^2(x_1 + x_2).
$$

If $f(x)$ is known, it can be verified that

$$
V^* = \frac{1}{2} x_1^2 + \frac{1}{2}(x_1 + x_2)^2.
$$

and

$$
V_1^* = -\frac{9}{4}(x_1 + x_2)^2 + \frac{9}{4} x_1^2
$$

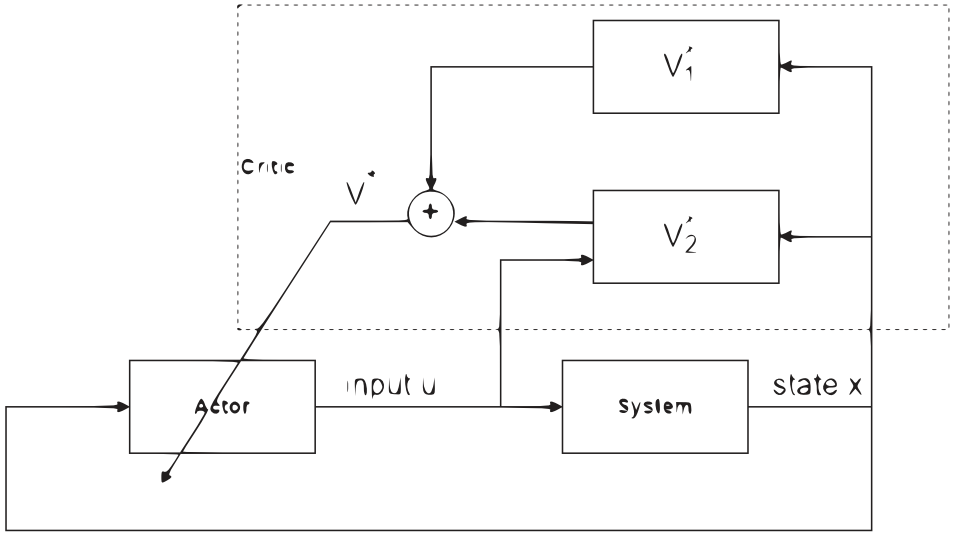

Figure 1. The block diagram of the proposed controller.

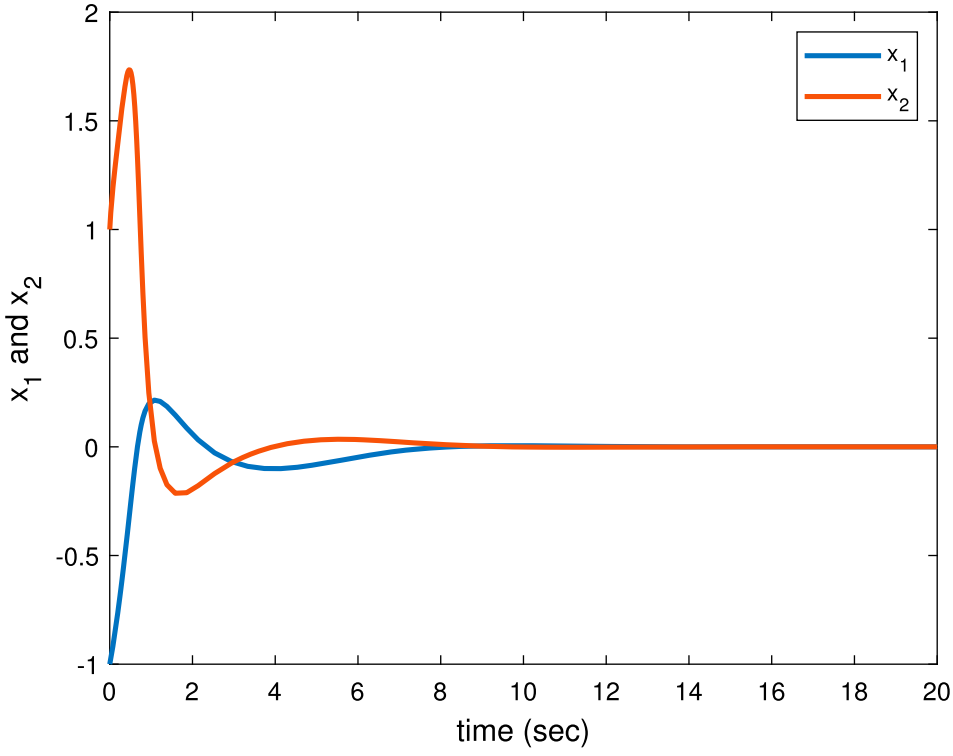

Figure 2. Time response of *x*.

$$+\frac{(x_1+x_2)\sin 2(x_1+x_2)}{2}-\frac{x_1\sin 2x_1}{2}$$
$$+\frac{\cos 2(x_1+x_2)}{4}-\frac{\cos 2x_1}{4}$$

Since $f(x)$ is unknown, V^* and V_1^* are unknown. We apply the proposed method in the last section to design a reinforcement learning-based controller. We choose S_f and ϕ as $S_f = [x_1, x_2, x_1^2, x_2^2, x_1x_2]^\top$ and $\phi = [x_1^2, x_2^2, x_1x_2, x_1^3, x_2^3, x_1^2x_2, x_1x_2^2]^\top$. By Theorem 3.4, the control input is (20) and the update laws for $\hat{W}_f$, W_c, and W_a are (23), (25), and (24), respectively. Simulation was done with the aid of MATLAB Simulink. Figure 2 shows the response of the state. It shows that the state converges to a small neighbourhood of the origin. Figure 3 shows the control input u. Figures 4– 6 show the response of $\hat{W}_f$, W_c, and W_a, respectively. These figures show that $\hat{W}_f$, W_c, and W_a are bounded. The simulation results confirm the statement in Theorem 3.4.

Example 4.2: Consider the system

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2 a\end{bmatrix}=\begin{bmatrix}x_2\\ -0.5x_1-0.5(x_1+x_2)(1-(\cos(2x_1)+2)^2)\end{bmatrix}+\begin{bmatrix}0\\ \cos(2x_1)+2\end{bmatrix}u \tag{27}$$

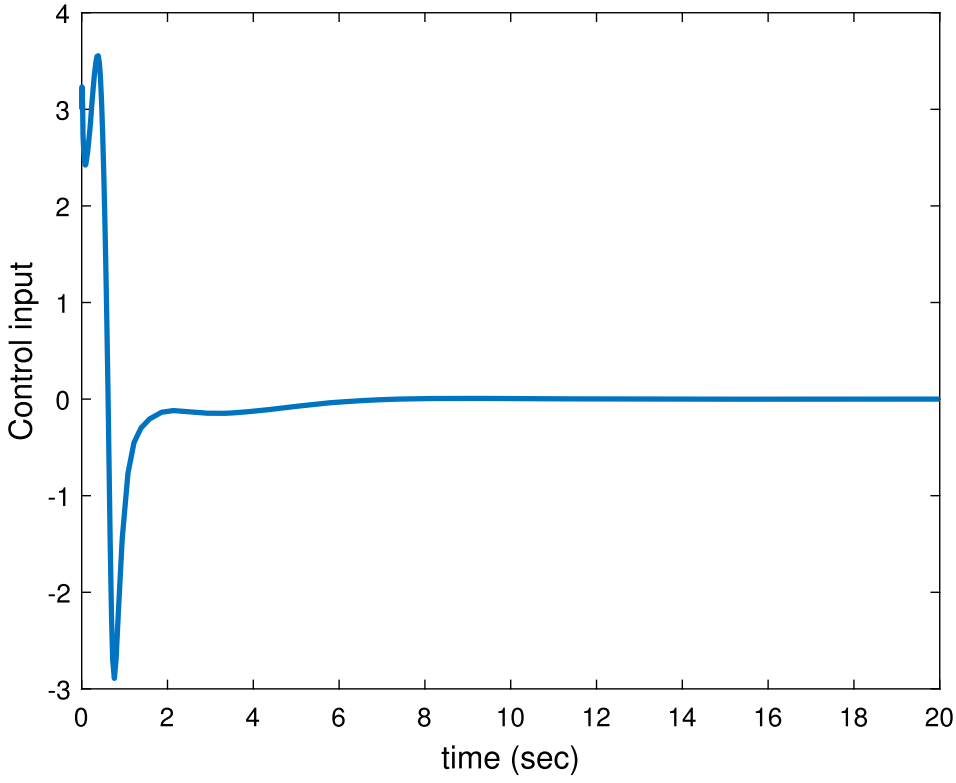

Figure 3. Time response of the input u.

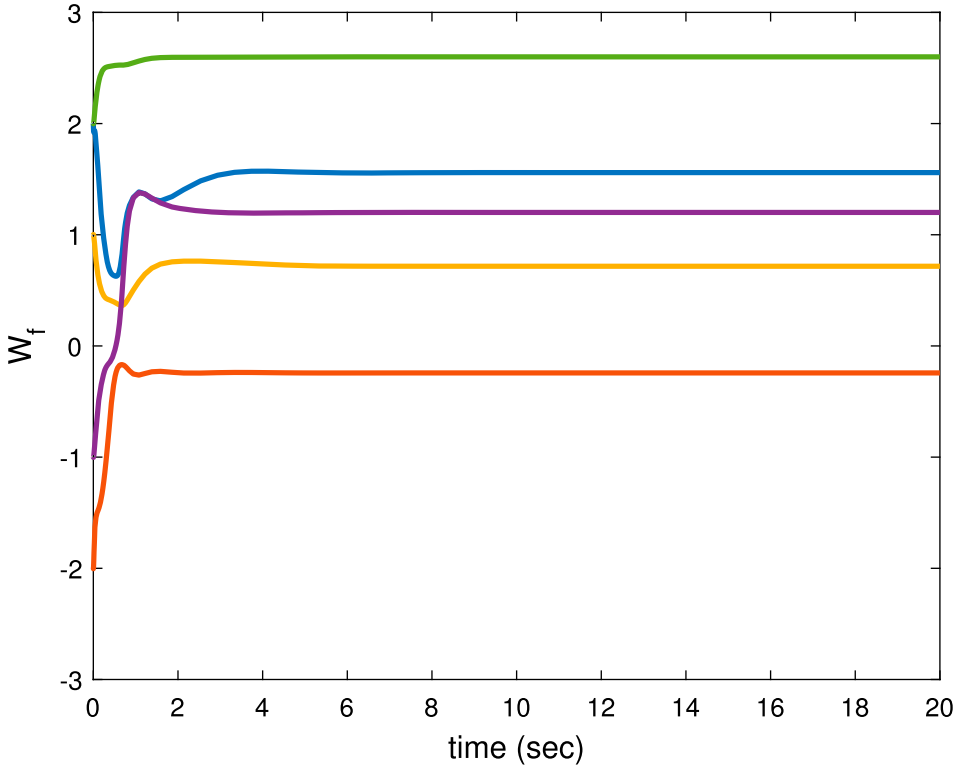

Figure 4. Time response of $\hat{W}_f$.

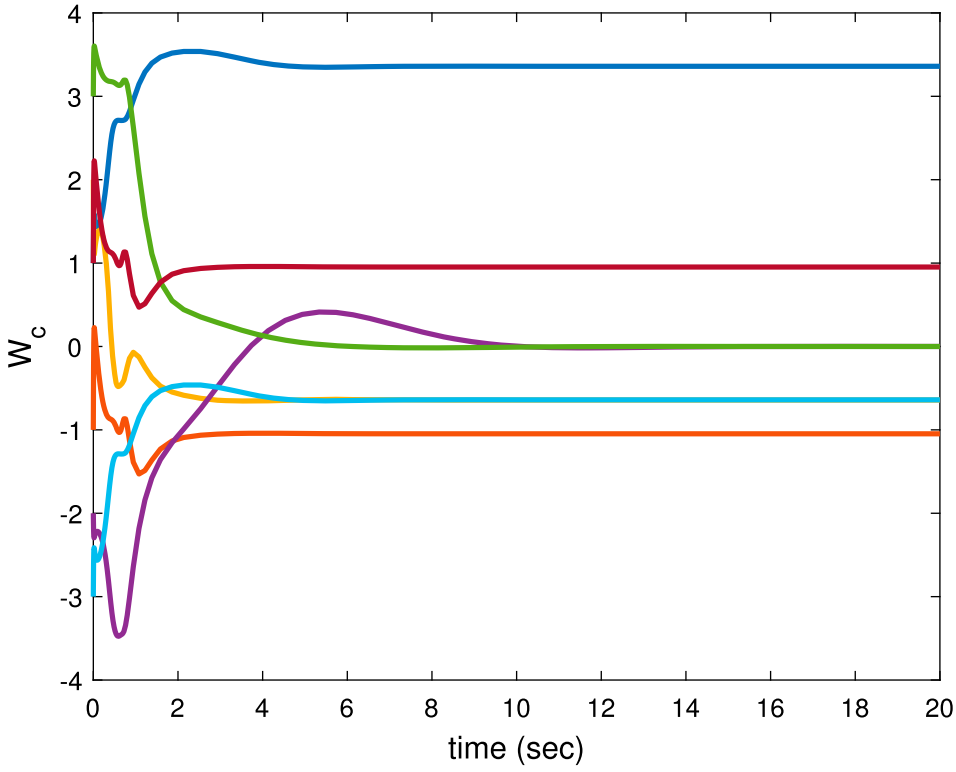

Figure 5. Time response of W_c.

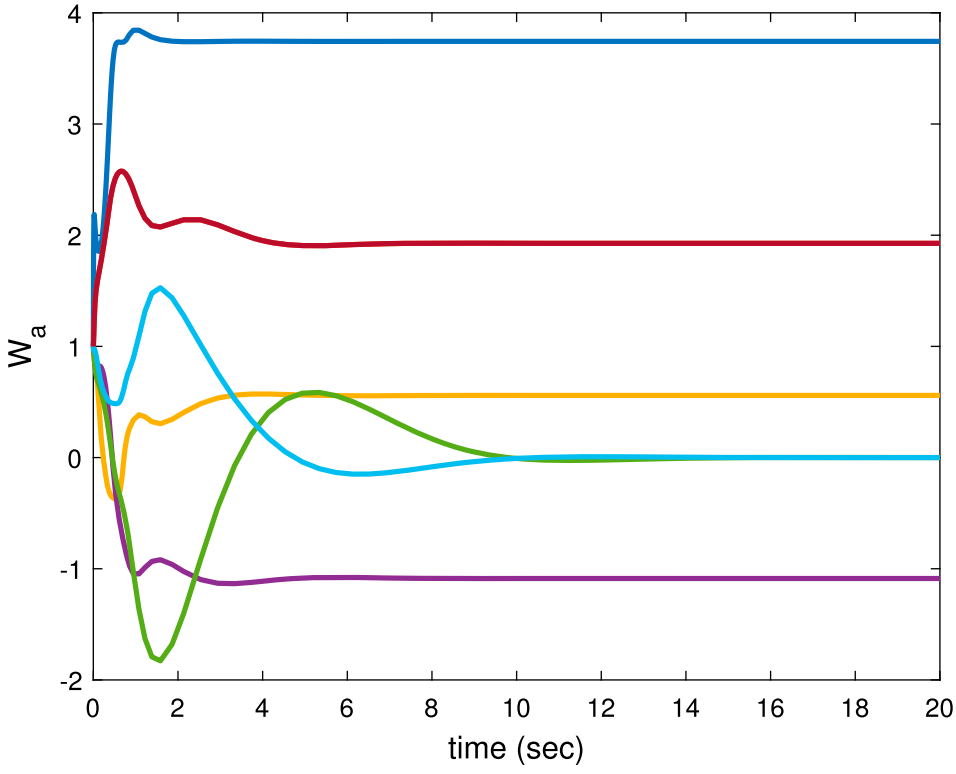

Figure 6. Time response of W_a.

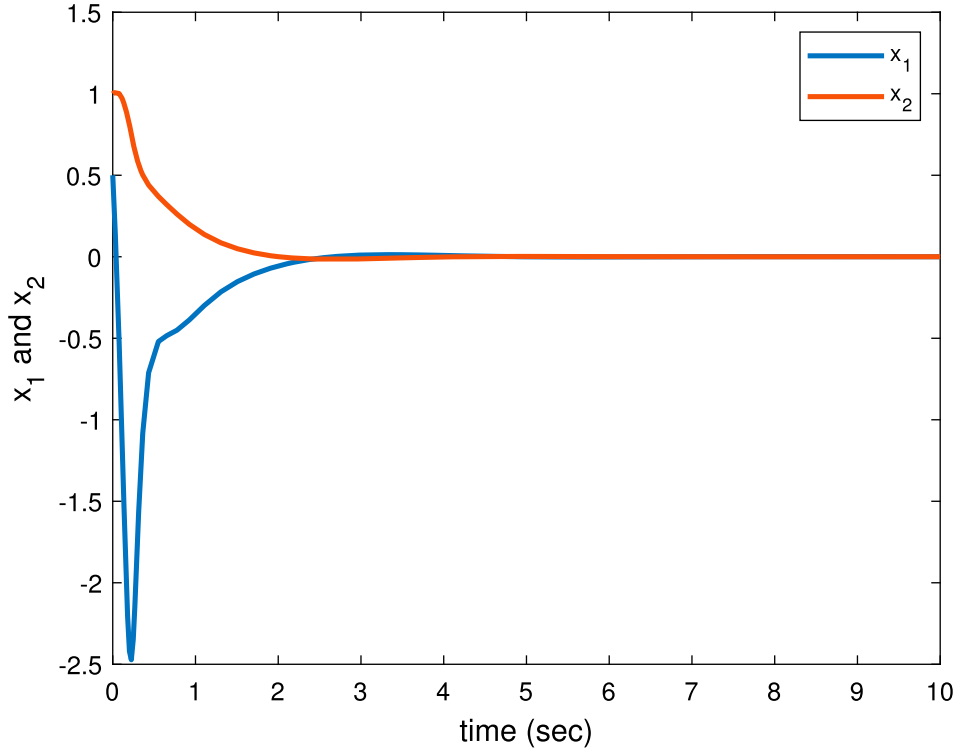

Figure 7. Time response of *x*.

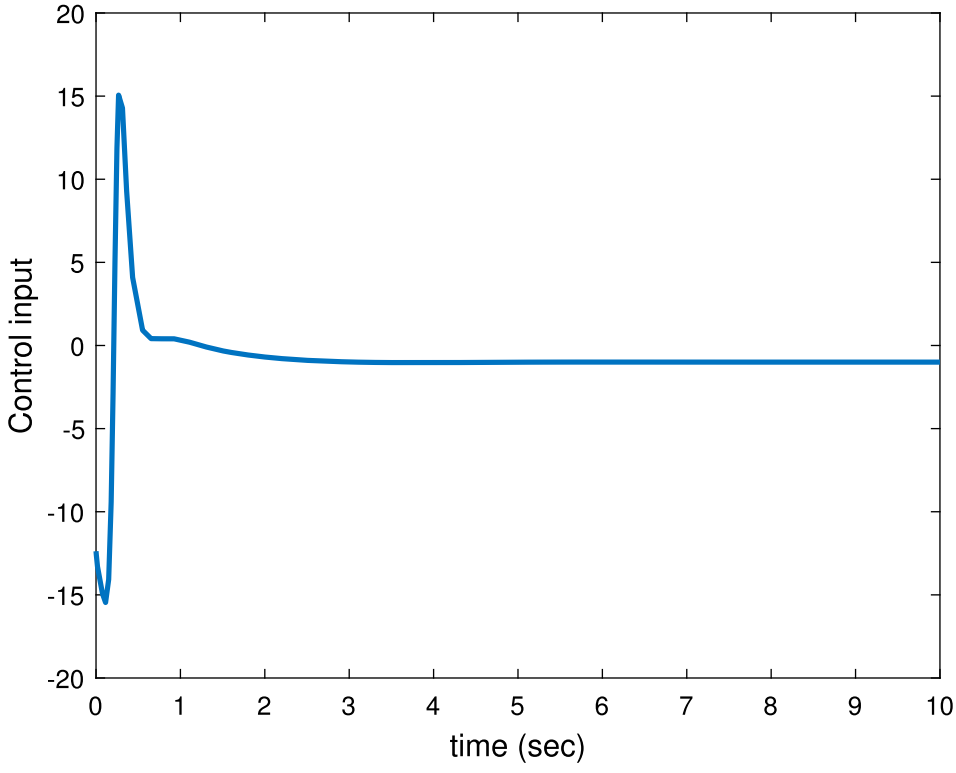

Figure 8. Time response of the input *u*.

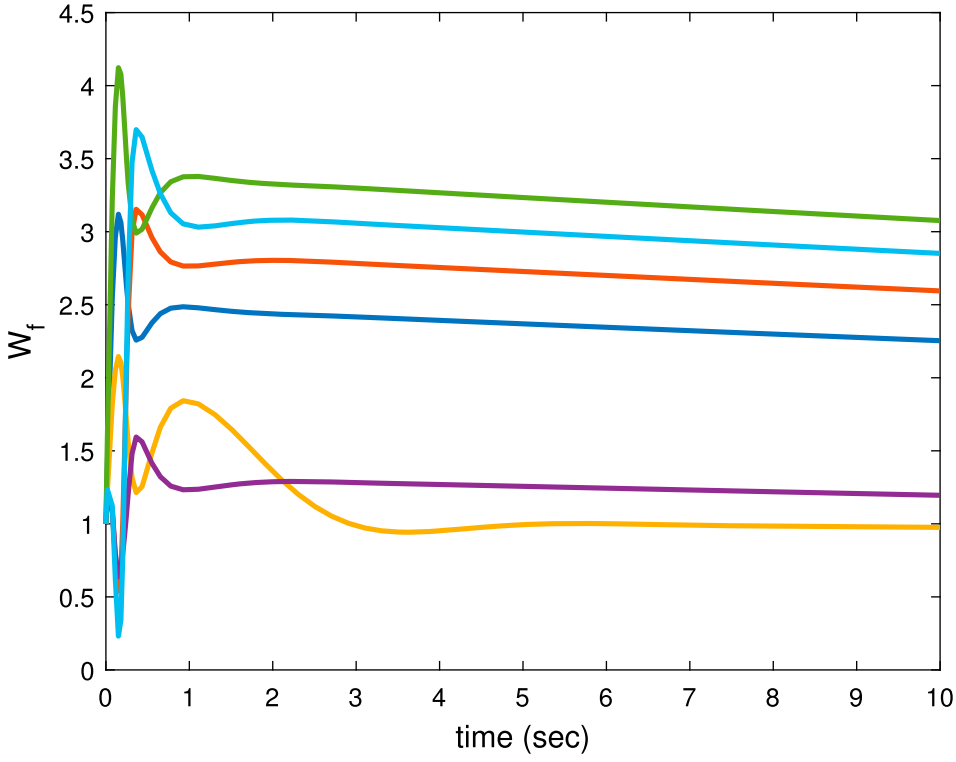

Figure 9. Time response of $\hat{W}_f$.

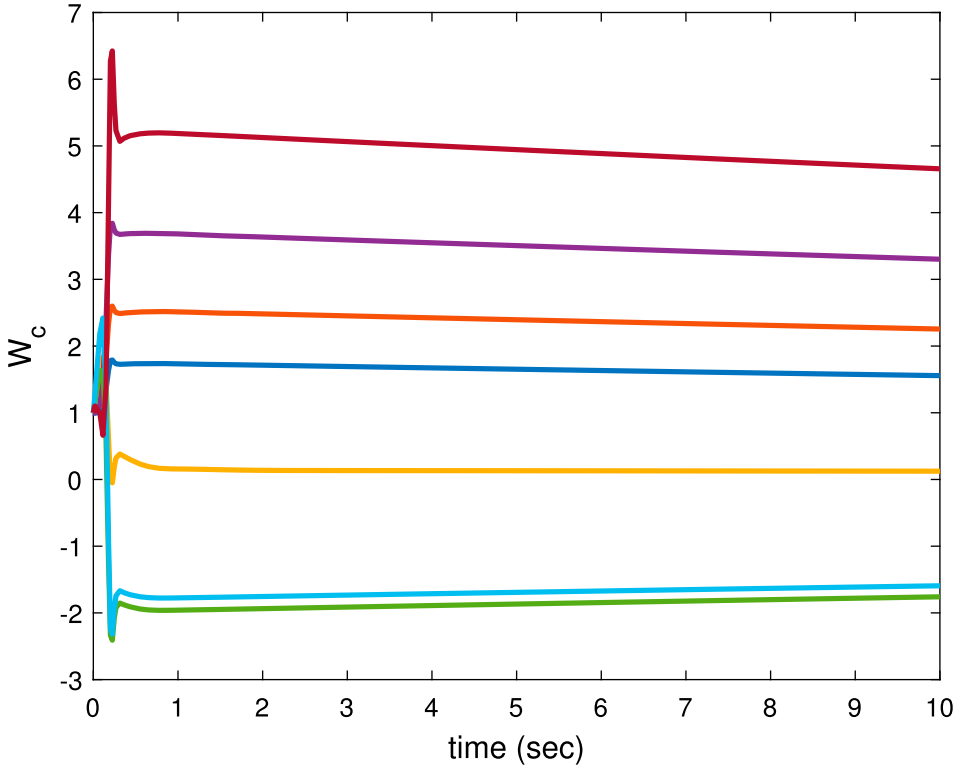

Figure 10. Time response of W_c.

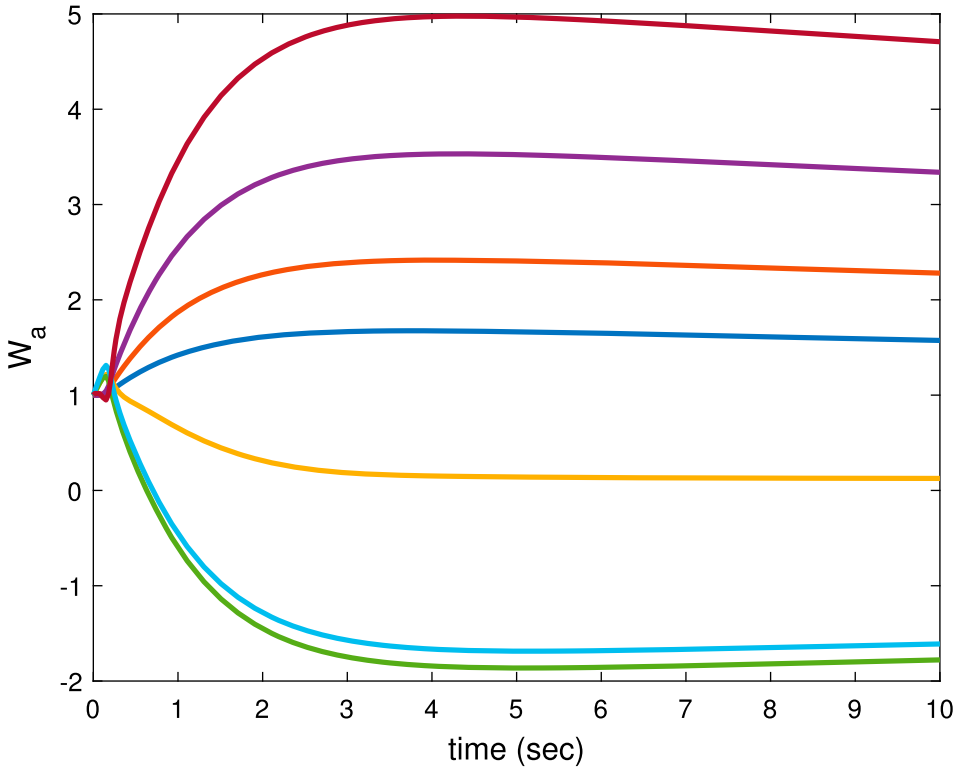

Figure 11. Time response of W_a.

the optimal control problem is to find an optimal control u such that the cost J is minimised where $Q(x) = x_1^2 + (x_1 + x_2)^2$ and $P = 1$. This optimal control problem has been studied in Vrabie and Lewis (2009) with a state transformation $y_1 = x_1$ and $y_2 = x_1 + x_2$. Here we solve this problem using the results proposed in this paper.

Since $f(x)$ is unknown, V^*, V_1^*, and V_2^* are unknown. In the simulation, we choose S_f and ϕ as $S_f = [1, x_1, x_2, x_1^2, x_2^2, x_1x_2]^\top$ and $\phi = [x_1^2, x_2^2, x_1x_2, x_1^3, x_2^3, x_1^2x_2, x_1x_2^2]^\top$. The control input is (20) and the update laws for $\hat{W}_f$, W_c, and W_a are in (23), (25), and (24), respectively. The simulation was done by Matlab Simulink for a group of control parameters. Figure 7 shows the response of the state which converges to a small neighbourhood of the origin. Figure 8 shows the control input u. Figures 9– 11 show the response of $\hat{W}_f$, W_c, and W_a, respectively. From the simulation results, it is shown that $\hat{W}_f$, W_c, and W_a are all bounded. The simulation results verify the statement in the last section.

The proposed results can be applied to solve optimal attitude control of unmanned aerial vehicles, optimal cruise control of autonomous vehicles, etc. The applications of the proposed results will be presented in the future.

5. Conclusion

This paper considered the optimal control of a second-order nonlinear system with unknown dynamics. A new reinforcement learning algorithm was proposed with the aid of direct adaptive control. Future research is on how to extend the results in this paper to more general nonlinear system with uncertainty.

Acknowledgments

The opinions expressed in this paper (or thesis or report or dissertation) are solely those of the author(s), and do not necessarily represent those of the NSF.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) would like to acknowledge funding provided by the National Science Foundation CREST Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (NSF Award No. 2112650) and the NSF [grant number ECCS-2037649].

References

Bertsekas, D. P. (1995). *Dynamic programming and optimal control*. (Vol. i). Athena Scientific.

Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K., Lewis, F., & Dixon, W. (2013). A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems. *Automatica*, *49*(1), 82–92. https://doi.org/10.1016/j.automatica.2012.09.019

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, *12*(1), 219–245. https://doi.org/10.1162/089976600300015961

Gao, W., Jiang, Y., Jiang, Z. P., & Chai, T. (2016). Output-feedback adaptive optimal control of interconnected systems based on robust adaptive dynamic programming. *Automatica*, *72*(1), 37–45. https://doi.org/10.1016/j.automatica.2016.05.008

Gao, W., Mynuddin, M., D. C. Wunsch, & Jiang, Z. P. (2022). Reinforcement learning-Based cooperative optimal output regulation via distributed adaptive internal model. *IEEE Transactions on Neural Networks and*

Learning Systems, *33*(10), 5229–5240. https://doi.org/10.1109/TNNLS.2021.3069728

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Jiang, Y., & Jiang, Z. P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, *48*(10), 2699–2704. https://doi.org/10.1016/j.automatica.2012.06.096

Jiang, Y., & Jiang, Z. P. (2014). Robust adaptive dynamic programming and feedback stabilization of nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 882–893. https://doi.org/10.1109/TNNLS.2013.2294968

Jiang, Y., & Jiang, Z. P. (2015). Global adaptive dynamic programming for continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, *60*(11), 2917–2929. https://doi.org/10.1109/TAC.2015.2414811

Lewis, F. L., Vrabie, D., & Syrmos, V. L. (2012). *Optimal control*. Wiley.

Lewis, F. L., Vrabie, D., & Vamvoudakis, K. G. (2012). Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems Magazine*, *32*(6), 76–105. https://doi.org/10.1109/MCS.2012.2214134

Liu, D., Xue, S., Zhao, B., Luo, B., & Wei, Q. (2021). Adaptive dynamic programming for control: A survey and recent advances. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(1), 142–160. doi:10.1109/TSMC.2020.3042876

Mendel, J., & McLaren, R. (1970). Reinforcement-Learning control and pattern recognition systems. In J. Mendel K. Fu (Eds.), *Adaptive, learning and pattern recognition systems* (Vol. 66, pp. 287–318). Elsevier.

Murray, J., Cox, C., Saeks, R., & Lendaris, G. (2001). Globally convergent approximate dynamic programming applied to an autolander. In *Proceedings of the 2001 American Control Conference* (Vol. 4, pp. 2901–2906).

Powell, W. B. (2007). *Approximate dynamic programming: Solving the curses of dimensionality*. Vol. 703. Wiley.

Stone, M. H. (1948). The generalized weierstrass approximation theorem. *Mathematics Magazine*, *21*(4), 167–184. https://doi.org/10.2307/3029750

Vamvoudakis, K. G., & Lewis, F. L. (2010). Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, *46*(5), 878–888. https://doi.org/10.1016/j.automatica.2010.02.018

Vrabie, D., & Lewis, F. (2009). Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, *22*(3), 237–246. https://doi.org/10.1016/j.neunet.2009.03.008 Goal-Directed Neural Systems.

Vrabie, D., Lewis, F., & Abu-Khalaf, M. (2008). Biologically inspired scheme for continuous-time approximate dynamic programming. *Transactions of the Institute Measurement and Control*, *30*(3-4), 207–223. https://doi.org/10.1177/0142331207088188

Vrabie, D., Pastravanu, O., Abu-Khalaf, M., & Lewis, F. (2009). Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, *45*(2), 477–484. https://doi.org/10.1016/j.automatica.2008.08.017

Vrabie, D. L., & Lewis, F. L. (2009). Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks: The Official Journal of the International Neural Network Society*, *22*(3), 237–246. https://doi.org/10.1016/j.neunet.2009.03.008

Werbos, P. J. (1992). Approximate dynamic programming for real-time control and neural modeling. In D. A. White and D. A. Sofge (Eds.), *Handbook of intelligent control: Neural, fuzzy, and adaptive approaches* (pp. 1–30). Van Nostrand Reinhold.