Optimal Control of Nonlinear System with Uncertainty Based on Reinforcement Learning

Miguel Garcia and Wenjie Dong

Department of Electrical and Computer Engineering, the University of Texas Rio Grande Valley, Edinburg, TX 78539, USA

ARTICLE HISTORY

Compiled February 10, 2025

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 three neural network approximation. Simulation results are presented to show the effectiveness of the proposed algorithms.

KEYWORDS

 $\label{eq:control} \ensuremath{\mathsf{Reinforcement}}\xspace \ensuremath{\mathsf{e}}\xspace \ensuremath{\mathsf{Reinforcement}}\xspace \ensuremath{\mathsf{e}}\xspace \$

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 in Bertsekas (1995); Powell (2007); Werbos (1992).

Reinforcement learning (RL) is one of 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, 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

CONTACT W. Dong. Email: wenjie.dong@utrgv.edu

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, Xue, Zhao, Luo, & Wei, 2021).

RL has been studied for continuous-time systems under the assumption that the system model information is well-known in Doya (2000); Murray, Cox, Saeks, and Lendaris (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, 2012; Lewis, Vrabie, & Vamvoudakis, 2012; D. L. 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 inputoutput data along the trajectory of the system state. In D. Vrabie and Lewis (2009); D. Vrabie, Lewis, and M.Abu-Khalaf (2008); D. Vrabie, Pastravanu, Abu-Khalaf, 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 in Gao, Mynuddin, Wunsch, and Jiang (2022); Jiang and Jiang (2012). For nonlinear systems with unknown dynamics, IRL algorithms are proposed in Jiang and Jiang (2014, 2015). In Gao, Jiang, Jiang, and Chai (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 from 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 identifieractor-critic reinforcement learning and IRL, in our proposed reinforcement learning approach there is neither 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 organization 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 $x_1 \in \mathbb{R}^n$ and $x_2 \in \mathbb{R}^n$ are the states, $x^{\top} = [x_1^{\top}, x_2^{\top}]^{\top} \in \Omega \subset \mathbb{R}^{2n}$, $u \in U \subset \mathbb{R}^n$ is the input, $f(x) \in \mathbb{R}^n$ with f(0) = 0 being an unknown vector function, and $g(x) \in \mathbb{R}^n$ is a known input matrix function. It is assumed that f(x) + g(x)u is Lipschitz continuous and the system (1)-(2) is stabilizable. For the purpose of control, it is assumed that g(x) is nonsingular for any state x. Let

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

(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

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

is minimized, 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

$$V(x(t), u(t)) = \int_{t}^{\infty} (Q(x) + u^{\top} P u) d\tau.$$
(5)

The Hamiltonian function corresponding to the above optimal control problem is

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

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

$$V^*(x(t)) = \inf_u \int_t^\infty (Q + u^\top P u) d\tau.$$
(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

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

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

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

$$= Q(x) + \nabla_x V^{*\top} F - \frac{1}{4} \nabla_x V^{*\top} G P^{-1} G^{\top} \nabla_x V^*$$
$$= 0.$$
(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, 2012). If f(x) is unknown, the VI and PI iteration methods do not work. In order to solve the optimization 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. $||f(x)|| \leq \gamma_1 x^{\top} x + \gamma_2 ||x||$, 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

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

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

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, Stinchcombe, and White (1989), and Stone (1948), there exists a vector S_f such that

$$f(x) = W_f^{\top} S_f(x) + \epsilon_f(x)$$
(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 choose the basis vector ϕ carefully. The gradient of V^* is

$$\nabla_x V^* = \nabla_x V_1^* + \nabla_x V_2^*
= 2\Lambda + \nabla_x \phi^\top W_V + \nabla_x \epsilon
= 2\Lambda + S_V^\top W_V + \epsilon_V$$
(14)

where

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

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

$$u^{*} = -\frac{1}{2}P^{-1}G^{\top}(2\Lambda + S_{V}^{\top}W_{V} + \epsilon_{V})$$

$$= -P^{-1}G^{\top}\Lambda - \frac{1}{2}P^{-1}G^{\top}S_{V}^{\top}W_{V} - \frac{1}{2}P^{-1}G^{\top}\epsilon_{V}.$$
 (15)

The Hamiltonian in (9) can be written as

$$H^{*}(x, u^{*}, \nabla_{x}V^{*}) = Q(x) + (u^{*})^{\top}Pu^{*} + [2\Lambda + S_{V}^{\top}W_{V} + \epsilon_{V}]^{\top}[F - GP^{-1}G^{\top}\Lambda - \frac{1}{2}GP^{-1}G^{\top}S_{V}^{\top}W_{V} - \frac{1}{2}GP^{-1}G^{\top}\epsilon_{V}]$$
(16)
$$= Q(x) + (u^{*})^{\top}Pu^{*} + [2\Lambda + S_{V}^{\top}W_{V} + \epsilon_{V}]^{\top}[F - \bar{P}\Lambda - \frac{1}{2}\bar{P}S_{V}^{\top}W_{V} - \frac{1}{2}\bar{P}\epsilon_{V}]$$
(17)

where

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

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

$$u = -P^{-1}G^{\top}\hat{\Lambda} - \frac{1}{2}P^{-1}G^{\top}S_{V}^{\top}W_{a}$$

= $-P^{-1}g^{\top}\hat{\Lambda}_{2} - \frac{1}{2}P^{-1}G^{\top}S_{V}^{\top}W_{a}$ (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

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

where $\hat{F} = \begin{bmatrix} x_2^\top, (S_f^\top \hat{W}_f)^\top \end{bmatrix}^\top$ and $\xi = S_V(\hat{F} + Gu) = S_V[\hat{F} - \bar{P}\hat{\Lambda} - \frac{1}{2}\bar{P}S_V^\top W_a].$ The Bellman residue error is defined as

$$z = H^{*}(x, u, \nabla \hat{V}^{*}) - H^{*}(x, u^{*}, \nabla V^{*})$$

= $H^{*}(x, u, \nabla \hat{V}^{*}).$ (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 $\Omega \subset \mathbb{R}^{2n}$, in the sense that there exist finite positive constants δ_f , δ_V , δ_z , α_V , α_g , and α_f such that $\|S_f\| \leq \alpha_f$, $\|g^{-\top}Pg^{-1}\| \leq \alpha_g$, $\|S_V\| \leq \alpha_V$, $\|\epsilon_f\| \leq \delta_f$, $\|\epsilon_V\| \leq \delta_V$, and $|z| \leq \delta_z$.

In order to find the critic update law W_c , we minimize 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

$$\begin{split} \dot{V}_{3} &= \nabla V^{*^{\top}} (F + Gu^{*}) + \nabla V^{*^{\top}} G(u - u^{*}) + \tilde{W}_{a}^{\top} \Gamma_{a} \dot{\tilde{W}}_{a} + \tilde{W}_{c}^{\top} \Gamma_{c} \dot{\tilde{W}}_{c} + \tilde{W}_{f}^{\top} \Gamma_{f} \dot{\tilde{W}}_{f} \\ &= -Q - (u^{*})^{\top} Pu^{*} + (2\Lambda + S_{V}^{\top} W_{V} + \epsilon_{V})^{\top} \bar{P} (\tilde{\Lambda} + 0.5 S_{V}^{\top} \tilde{W}_{a} + 0.5 \epsilon_{V}) \\ &+ \tilde{W}_{a}^{\top} \Gamma_{a} \dot{\tilde{W}}_{a} + \tilde{W}_{c}^{\top} \Gamma_{c} \dot{\tilde{W}}_{c} + \tilde{W}_{f}^{\top} \Gamma_{f} \dot{\tilde{W}}_{f} \\ &= -Q - (u^{*})^{\top} Pu^{*} + (2\hat{\Lambda} + S_{V}^{\top} W_{c})^{\top} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} S_{f}^{\top} \tilde{W}_{f} \\ &+ 0.5(2\hat{\Lambda} + S_{V}^{\top} W_{c})^{\top} \bar{P} S_{V}^{\top} \tilde{W}_{a} + 2\tilde{\Lambda}^{\top} \bar{P} \tilde{\Lambda} + \tilde{\Lambda}^{\top} \bar{P} S_{V}^{\top} \tilde{W}_{a} + \tilde{W}_{c}^{\top} S_{V} \bar{P} \tilde{\Lambda} \\ &+ 0.5 (2\hat{\Lambda} + S_{V}^{\top} W_{c})^{\top} \bar{P} S_{V}^{\top} \tilde{W}_{a} + 2\tilde{\Lambda}^{\top} \bar{P} \tilde{\Lambda} + \tilde{\Lambda}^{\top} \bar{P} S_{V}^{\top} \tilde{W}_{a} + \tilde{W}_{c}^{\top} S_{V} \bar{P} \tilde{\Lambda} \\ &+ 0.5 (\hat{W}_{c}^{\top} S_{V} \bar{P} S_{V}^{\top} \tilde{W}_{a} + \Lambda^{\top} \bar{P} \epsilon_{V} + 0.5 W_{V}^{\top} S_{V} \bar{P} \epsilon_{V} + \epsilon_{V}^{\top} \bar{P} \tilde{\Lambda} + 0.5 \epsilon_{V}^{\top} \bar{P} S_{V}^{\top} \tilde{W}_{a} \\ &+ 0.5 \epsilon_{V}^{\top} \bar{P} \epsilon_{V} + \tilde{W}_{a}^{\top} \Gamma_{a} \dot{\tilde{W}}_{a} + \tilde{W}_{c}^{\top} \Gamma_{c} \dot{\tilde{W}}_{c} + \tilde{W}_{f}^{\top} \Gamma_{f} \dot{\tilde{W}}_{f}. \end{split}$$

We choose

$$\dot{\hat{W}}_{f} = \Gamma_{f}^{-1}S_{f}[\mathbf{0},\mathbf{I}]\left(2\hat{\Lambda}+S_{V}^{\top}W_{c}\right)-k_{f}\Gamma_{f}^{-1}S_{f}S_{f}^{\top}\hat{W}_{f}$$

$$\dot{W}_{a} = \frac{1}{2}\Gamma_{a}^{-1}S_{V}\bar{P}\left(2\hat{\Lambda}+S_{V}^{\top}W_{c}\right)-k_{a}\Gamma_{a}^{-1}S_{V}S_{V}^{\top}(W_{a}-W_{c})$$

$$-k_{e}\Gamma_{a}^{-1}S_{V}S_{V}^{\top}W_{a}$$
(23)
$$(23)$$

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

$$\dot{W}_{c} = -2k_{c}\Gamma_{c}^{-1}\xi z - k_{a}\Gamma_{c}^{-1}S_{V}S_{V}^{\top}(W_{c} - W_{a}) - \Gamma_{c}^{-1}k_{d}S_{V}S_{V}^{\top}W_{c}.$$
(25)

Then

$$\begin{split} \dot{V}_{3} &= -Q + \tilde{W}_{f}^{\top} S_{f}[\mathbf{0},\mathbf{I}] S_{V}^{\top} \tilde{W}_{a} + 2 \tilde{W}_{f}^{\top} S_{f} g^{-\top} P g^{-1} S_{f}^{\top} \tilde{W}_{f} + \tilde{W}_{c}^{\top} S_{V} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} S_{f}^{\top} \tilde{W}_{f} \\ &- (u^{*})^{\top} P u^{*} + \frac{1}{2} \tilde{W}_{c}^{\top} S_{V} \bar{P} S_{V}^{\top} \tilde{W}_{a} + f^{\top}[\mathbf{0},\mathbf{I}] \epsilon_{V} + \frac{1}{2} W_{V}^{\top} S_{V} \bar{P} \epsilon_{V} + \epsilon_{V}^{\top} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} S_{f}^{\top} \tilde{W}_{f} \\ &+ 0.5 \epsilon_{V}^{\top} \bar{P} S_{V}^{\top} \tilde{W}_{a} + 0.5 \epsilon_{V}^{\top} \bar{P} \epsilon_{V} - k_{a} \tilde{W}_{a}^{\top} S_{V} S_{V}^{\top} \tilde{W}_{a} - k_{a} \tilde{W}_{c}^{\top} S_{V} S_{V}^{\top} \tilde{W}_{c} \\ &+ k_{a} \tilde{W}_{a}^{\top} S_{V} S_{V}^{\top} \tilde{W}_{c} + k_{a} \tilde{W}_{c}^{\top} S_{V} S_{V}^{\top} \tilde{W}_{a} - \frac{k_{f}}{2} \tilde{W}_{f}^{\top} S_{f} S_{f}^{\top} \tilde{W}_{f} - \frac{k_{f}}{2} \tilde{W}_{f}^{\top} S_{f} S_{f}^{\top} \tilde{W}_{f} \\ &+ \frac{k_{f}}{2} W_{f}^{\top} S_{f} S_{f}^{\top} W_{f} - \frac{k_{e}}{2} \tilde{W}_{a}^{\top} S_{V} S_{V}^{\top} \tilde{W}_{a} - \frac{k_{e}}{2} W_{a}^{\top} S_{V} S_{V}^{\top} W_{a} + \frac{k_{e}}{2} W_{V}^{\top} S_{V} S_{V}^{\top} W_{V} \\ &- \frac{k_{d}}{2} \tilde{W}_{c}^{\top} S_{V} S_{V}^{\top} \tilde{W}_{c} - \frac{k_{d}}{2} W_{c}^{\top} S_{V} S_{V}^{\top} W_{c} + \frac{k_{d}}{2} W_{V}^{\top} S_{V} S_{V}^{\top} W_{V} + 2k_{c} z \tilde{W}_{c}^{\top} S_{V} \begin{bmatrix} x_{2} \\ \mathbf{0} \end{bmatrix} \\ &+ k_{c} z \tilde{W}_{c}^{\top} S_{V} \bar{P} S_{V}^{\top} \tilde{W}_{a} - k_{c} z \tilde{W}_{c}^{\top} S_{V} \bar{P} S_{V}^{\top} W_{V}. \end{split}$$

Since Q(x) is positive definite, there exists a positive matrix q such that $Q(x) \ge 1$

 $x^{\top} \bar{Q} x$ for $x \in \Omega$. Then

$$\begin{split} \dot{V}_{3} &\leq -x^{\top}\bar{Q}x - (u^{*})^{\top}Pu^{*} + \tilde{W}_{f}^{\top}S_{f}[\mathbf{0},\mathbf{I}]S_{V}^{\top}\tilde{W}_{a} + 2\tilde{W}_{f}^{\top}S_{f}g^{-\top}Pg^{-1}S_{f}^{\top}\tilde{W}_{f} + \tilde{W}_{c}^{\top}S_{V}\begin{bmatrix}\mathbf{0}\\\mathbf{I}\end{bmatrix}S_{f}^{\top}\tilde{W}_{f} \\ &+ 0.5\tilde{W}_{c}^{\top}S_{V}\bar{P}S_{V}^{\top}\tilde{W}_{a} + \gamma_{1}\delta_{V}x^{\top}x + \gamma_{2}||x||\delta_{V} + 0.5W_{V}^{\top}S_{V}\bar{P}\epsilon_{V} + \epsilon_{V}^{\top}\begin{bmatrix}\mathbf{0}\\\mathbf{I}\end{bmatrix}S_{f}^{\top}\tilde{W}_{f} \\ &+ 0.5\epsilon_{V}^{\top}\bar{P}S_{V}^{\top}\tilde{W}_{a} + 0.5\epsilon_{V}^{\top}\bar{P}\epsilon_{V} - k_{a}\tilde{W}_{a}^{\top}S_{V}S_{V}^{\top}\tilde{W}_{a} - k_{a}\tilde{W}_{c}^{\top}S_{V}S_{V}^{\top}\tilde{W}_{c} \\ &+ k_{a}\tilde{W}_{a}^{\top}S_{V}S_{V}^{\top}\tilde{W}_{c} + k_{a}\tilde{W}_{c}^{\top}S_{V}S_{V}^{\top}\tilde{W}_{a} - \frac{k_{f}}{2}\tilde{W}_{f}^{\top}S_{f}S_{f}^{\top}\tilde{W}_{f} - \frac{k_{f}}{2}\hat{W}_{f}^{\top}S_{f}S_{f}^{\top}\tilde{W}_{f} \\ &+ \frac{k_{f}}{2}W_{f}^{\top}S_{f}S_{f}^{\top}W_{f} - \frac{k_{e}}{2}\tilde{W}_{a}^{\top}S_{V}S_{V}^{\top}\tilde{W}_{a} - \frac{k_{e}}{2}W_{a}^{\top}S_{V}S_{V}^{\top}W_{a} + \frac{k_{e}}{2}W_{V}^{\top}S_{V}S_{V}^{\top}W_{V} \\ &- \frac{k_{d}}{2}\tilde{W}_{c}^{\top}S_{V}S_{V}^{\top}\tilde{W}_{c} - \frac{k_{d}}{2}W_{c}^{\top}S_{V}S_{V}^{\top}W_{c} + \frac{k_{d}}{2}W_{V}^{\top}S_{V}S_{V}^{\top}W_{V} + 2k_{c}z\tilde{W}_{c}^{\top}S_{V}\left[\begin{array}{c}x_{2}\\\mathbf{0}\end{array}\right] \\ &+ k_{c}z\tilde{W}_{c}^{\top}S_{V}\bar{P}S_{V}^{\top}\tilde{W}_{a} - k_{c}z\tilde{W}_{c}^{\top}S_{V}\bar{P}S_{V}^{\top}W_{V}. \end{split}$$

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{split} \dot{V}_{3} &\leq -y_{1}^{\top} \bar{Q}y_{1} - (u^{*})^{\top} Pu^{*} + 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} \\ &+ 0.5y_{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} + 2k_{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} + 2k_{c} zy_{3}^{\top} \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} y_{1} + k_{c} zy_{3}^{\top} \bar{P}y_{4} \\ &- k_{c} zy_{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} \\ \leq -y_{1}^{\top} \bar{Q}y_{1} - (u^{*})^{\top} Pu^{*} + 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} + 2k_{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} zy_{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{1} \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_{V} \\ &- \frac{k_{d}}}{2} W_{V}^{\top} S_{V} S_{V}^{\top} W_{c} + \frac{k_{d}}}{2} W_{V}^{\top} S_{V} S_{V}^{\top} W_{V} \\ &\leq -y^{\top} H y + y^{\top} D + \gamma_{2} \delta_{V} ||y_{1}|| + \epsilon_{1} \end{split}$$

where

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

$$\begin{split} H_{11} &= \bar{Q} - \gamma_1 \delta_V I - k_c \delta_z, \quad H_{23} = -[\mathbf{0}, 0.5\mathbf{I}] \\ H_{24} &= -[\mathbf{0}, 0.5\mathbf{I}] \\ H_{34} &= -\frac{0.5\bar{P} + 2k_a I}{2} \\ H_{33} &= (k_a + 0.5k_d - 1.5k_c \delta_z) I \\ H_{44} &= (k_a + 0.5k_e - 0.5k_c \delta_z) I \\ D &= \begin{bmatrix} \mathbf{0} \\ [\mathbf{0}, \mathbf{I}] \epsilon_V \\ -k_c z \bar{P} S_V^\top W_V \\ 0.5 \bar{P} \epsilon_V \end{bmatrix} \\ \epsilon_1 &= 0.5 \epsilon_V^\top \bar{P} \epsilon_V + \frac{k_f}{2} W_f^\top S_f S_f^\top W_f + \frac{k_e}{2} W_V^\top S_V S_V^\top W_V + \frac{k_d}{2} W_V^\top S_V S_V^\top W_V \\ + 0.5 W_V S_V \bar{P} \epsilon_V \end{split}$$

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

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

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 Fig. 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 *epsilon* are small. One can choose each element of S_f and ϕ to be sigmoidals with appropriate weights (see Cybenko (1989) and 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.



Figure 1. The block diagram of the proposed controller

4. Simulation

Two simulation examples are considered.

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

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

In the optimal problem, 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

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

In controller design, we choose the elements of the vectors S_f and ϕ are high-order polynomials of x_1 and x_2 . In the simulation we choose $S_f = [x_1, x_2, x_1^2, x_2^2, x_1 x_2]^{\top}$ and $\phi = [x_1^2, x_2^2, x_1 x_2, x_1^3, x_2^3, x_1^2 x_2, x_1 x_2^2]^{\top}$. The control input is (20). The update laws for \hat{W}_f , W_c , and W_a are in (23), (25), and (24), respectively.

The simulation results are obtained for a group of control parameters. Fig. 2 shows the response of the state, which converges to a small neighborhood of the origin. Fig.



Figure 3. Time response of the input u.

3 shows the control input u. Figs 4, 5, and 6 show the response of \hat{W}_f , W_c , and W_a , respectively. It is shown that they are all bounded. The simulation results show that the state converge to a small neighborhood of the origin.

Example 2: Consider the system

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

the optimal control problem is to find an optimal control u such that the cost J is minimized where $Q(x) = x_1^2 + (x_1 + x_2)^2$ and P = 1. This optimal control problem has been studied in D. 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 in this article.

If the dynamics is well-known, the optimal value function is

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



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



Figure 5. Time response of W_c



Figure 6. Time response of W_a

and

$$V_1^* = -(2 - (\cos(2x_1) + 2)^2)x_1x_2 - \frac{x_2^2}{2}(1 - (\cos(2x_1) + 2)^2)x_1x_2 - \frac{x_2^2}{2}(1 - (\cos(2x_1) + 2)^2)x_1x_2 + \frac{x_2^2}{2}(1 - (\cos(2x_1) + 2)^2)x_1x_2 + \frac{x_2^2}{2}(1 - (\cos(2x_1) + 2)^2)$$

and the optimal control law is

$$u^*(x) = -(\cos(2x_1) + 2)(x_1 + x_2)$$

Since f(x) is unknown, V^* , V_1^* , and V_2^* are unknown. In the simulation, we choose S_f and ϕ are high-order polynomials of x_1 and x_2 as follows $S_f = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]^{\top}$ and $\phi = [x_1^2, x_2^2, x_1 x_2, x_1^3, x_2^3, x_1^2 x_2, x_1 x_2^2]^{\top}$. The control input is (20). The update laws for \hat{W}_f , W_c , and W_a are in (23), (25), and (24), respectively.

The simulation results are obtained for a group of control parameters. Fig. ?? shows the response of the state, which converges to a small neighborhood of the origin. Fig. ?? shows the control input u. Figs ??, ??, and ?? show the response of \hat{W}_f , W_c , and W_a , respectively. It is shown that they are all bounded. The simulation results show that the state converge to a small neighborhood of the origin.

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. The future research is on how to extend the results in this paper to more general nonlinear system with uncertainty.

Acknowledgements

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 no. ECCS-2037649. 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.

References

- Bertsekas, D. P. (1995). Dynamic programming and optimal control, vol. i. Belmont, MA, USA: 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.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics* of Control, Signals and Systems(2), 303-314.

- Doya, K. (2000). Reinforcement learning in continuous time and space. Neural Computation, 12(1), 219-245.
- 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, 37-45.
- Gao, W., Mynuddin, M., Wunsch, D. C., & Jiang, Z.-P. (2022). Reinforcement learningbased cooperative optimal output regulation via distributed adaptive internal model. *IEEE Transactions on Neural Networks and Learning Systems*, 33(10), 5229-5240.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359-366.
- Jiang, Y., & Jiang, Z.-P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. Automatica, 48(10), 2699-2704.
- Jiang, Y., & Jiang, Z.-P. (2014). Robust adaptive dynamic programming and feedback stabilization of nonlinear systems. *IEEE Transactions on Neural Networks and Learning Sys*tems, 25(5), 882-893.
- Jiang, Y., & Jiang, Z.-P. (2015). Global adaptive dynamic programming for continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, 60(11), 2917-2929.
- Lewis, F. L., Vrabie, D., & Syrmos, V. L. (2012). Optimal control. Hoboken, NJ, USA: 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.
- 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.
- 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, p. 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, p. 2901-2906 vol.4).
- Powell, W. B. (2007). Approximate dynamic programming: Solving the curses of dimensionality, vol. 703. Hoboken, NJ, USA: Wiley.
- Stone, M. H. (1948). The generalized weierstrass approximation theorem. Mathematics Magazine, 21(4), 167–184.
- 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.
- 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. (Goal-Directed Neural Systems)
- Vrabie, D., Lewis, F., & M.Abu-Khalaf. (2008). Biologically inspired scheme for continuoustime approximate dynamic programming. Transactions of the Institute Measurement and Control., 30(3/4), 207-223.
- 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.
- 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, 237-46.
- Werbos, P. J. (1992). Approximate dynamic programming for real-time control and neural modeling. in Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches, D. A. White and D. A. Sofge, Eds. New York, NY, USA: Van Nostrand Reinhold, 1-30.