


# Robust Near-optimal Control for Constrained Nonlinear System via Integral Reinforcement Learning

Yu-Qing Qiu, Yan Li, and Zhong Wang\* 

**Abstract:** This paper proposes a robust near-optimal control algorithm for uncertain nonlinear systems with state constraints and input saturation. By incorporating a barrier function and a non-quadratic term, the robust stabilization problem with constraints and uncertainties is converted into an unconstrained optimal control problem of the nominal system, which requires the solution of the Hamilton-Jacobi-Bellman (HJB) equation. The proposed integral reinforcement learning (IRL)-based method can obtain the approximate solution of the HJB equation without requiring any knowledge of system drift dynamics. An online gain-adjustable update law of the actor-critic architecture is developed to relax the persistence of excitation (PE) condition and ensure the closed-loop system stability throughout learning. The uniform ultimate boundedness of the closed-loop system is verified using Lyapunov's direct method. Simulation results demonstrate the effectiveness and feasibility of the proposed method.

**Keywords:** Constrained nonlinear system, integral reinforcement learning, optimal control, robust control.

## 1. INTRODUCTION

Various uncertainties could significantly degrade the closed-loop system stability and performance [1,2]. Recently, the robust optimal control problems for nonlinear systems with uncertainties have received intensive attentions [3-7], which can take both optimality and robustness into consideration. These robust optimal controllers are obtained by handling optimal control problems, which frequently require solving the Hamilton-Jacobi-Bellman (HJB) equation. However, the analytic solution of the HJB equation is difficult to get due to the nonlinearity.

In order to address the challenge, reinforcement learning (RL) method is employed to solve the optimization problems, and it has gained a lot of interests from investigators [8-10]. In [11-14], the approximate optimal solutions are obtained by the actor-critic algorithms for nonlinear optimal control problems. In [15], a novel RL-based robust adaptive controller is presented for nonlinear systems subject to input constraints, which can remove the condition imposed on the initial control. Whereas, the information about system drift dynamics is required either explicitly or from an identifier in the traditional RL technique.

Integral reinforcement learning (IRL), as a variation of the RL, is proposed in [16,17], which can relax the requirement of system dynamics by introducing the integral Bellman equations. Thus, it has been utilized in the

optimal control problems for uncertain nonlinear systems [18-21] and input-constrained systems [22]. In [23], an approximate optimal critic learning algorithm is developed for tackling input-constrained optimal control problems. A robust control problem is transformed into a constrained optimal control problem in [24], which is solved utilizing the critic and actor neural networks (NNs). The study on designing optimal controllers using the IRL framework for a class of input-constrained nonlinear systems can be tracked to the work [25]. All of the preceding studies employ a non-quadratic penalty function to remove the input saturation. One property of these methods is that they require the persistence of excitation (PE) condition to be used. To relax the PE condition, an IRL algorithm based on the actor-critic structure is derived in [26], which employs the experience replay (ER) technique to learn the online solution of the HJB equation for partially-unknown input-constrained systems. The study [27] presents a critic-only IRL controller coupled with an ER-based identifier to solve the optimal control problem for unknown systems with actuator constraints. Mishra [28] proposes a novel parameter update law for tuning the weights of critic, actor and disturbance NNs based on the variable-gain gradient descent and ER methods.

This paper focuses on the robust optimal control problem for uncertain nonlinear systems under state constraints and input saturation based on the IRL technique. Requiring both an initial admissible control input and the

Manuscript received August 10, 2021; revised March 16, 2022; accepted May 13, 2022. Recommended by Associate Editor Niket Kaisare under the direction of Senior Editor PooGyeon Park.

Yu-Qing Qiu, Yan Li, and Zhong Wang are with the School of Automation, Northwestern Polytechnical University, Xi'an 710129, China (e-mails: qiuyuqing@mail.nwpu.edu.cn, liyan@nwpu.edu.cn, zhong.wang@mail.nwpu.edu.cn).

\* Corresponding author.

PE condition simultaneously is challenging for the traditional IRL formulations like [25]. In this paper, a barrier function and a non-quadratic penalty term are introduced to handle the state and input constraints. In particular, an online update law for the actor-critic architecture is designed to address the aforementioned challenge. As a result, an online IRL-based robust near-optimal control method is proposed for constrained nonlinear systems with uncertainties. It should be noted that the proposed scheme is in a similar spirit as [25], but the novel update laws in this paper can relax the PE condition and ensure the system stability throughout learning without an initial admissible control.

The major contributions of this paper are as follows:

- 1) A new gain-adjustable update law of the actor-critic structure is proposed, in which the weights are learned by employing both historical and current data. The benefit is that the PE condition is relaxed and the system stability is ensured throughout learning.
- 2) The system drift dynamics are not required in the proposed method. The stability and convergence analysis of the closed-loop system is given, which includes the online update laws of the critic and actor NNs.

The rest of this paper is arranged as follows: The problem formulation and preliminaries are given in Section 2. An online IRL-based controller is designed in Section 3, and the stability and convergence analysis of the closed-loop system is presented in this section too. Section 4 gives simulation results to verify the proposed control algorithm, followed by the conclusion in Section 5.

**Notation:** Throughout this paper,  $|\cdot|$  denotes the magnitude of a scalar.  $\|\cdot\|$  is the Euclidean norm of a vector and the induced norm of a matrix.  $\lambda_{\max}(\cdot)$  and  $\lambda_{\min}(\cdot)$  are the maximum eigenvalue and minimum eigenvalue of a matrix respectively. The superscripts  $-1$  and  $T$  represent the inverse and transpose of a matrix respectively.

## 2. PROBLEM FORMULATION AND PRELIMINARIES

### 2.1. Problem formulation

Consider a nonlinear system described as

$$\dot{x} = f(x) + g(x)(u + d(x)), \quad (1)$$

where  $x \in \mathbb{R}^n$  denotes the state vector and  $u \in \mathbb{R}^m$  is the control vector.  $f(x) \in \mathbb{R}^n$  and  $g(x) \in \mathbb{R}^{n \times m}$  are continuous functions. The uncertainty term  $d(x) \in \mathbb{R}^m$  represents both the model parameters uncertainties and unmodeled dynamics. The state constraint and input saturation are considered as follows:

$$\begin{aligned} \mathcal{D}_x &= \{x : |x_j| \leq \beta, j = 1, \dots, n\}, \\ \mathcal{D}_u &= \{u : |u_i| \leq \eta, i = 1, \dots, m\}, \end{aligned} \quad (2)$$

where  $\mathcal{D}_x$  and  $\mathcal{D}_u$  denote the state and control constraint sets, while the bounds are represented by  $\beta$  and  $\eta$ , respectively.

For convenience, the following assumptions and definition are given below.

**Assumption 1:** The functions  $f(x)$  and  $g(x)$  are Lipschitz continuous on a compact set  $\Omega$  containing the origin, such that the system (1) with (2) is controllable.

**Assumption 2:** It is assumed that there exist three known positive constants  $b_f$ ,  $b_g$  and  $\rho$  such that  $\|f(x)\| \leq b_f \|x\|$  with  $f(\mathbf{0}) = \mathbf{0}$ ,  $\|g(x)\| \leq b_g$ , and the uncertainties  $\|d(x)\| \leq \rho \|x\|$  with  $d(\mathbf{0}) = \mathbf{0}$ .

**Definition 1** (Uniformly ultimately bounded (UUB) stability) [29]: Consider the nonlinear system  $\dot{x} = f(x, t)$ . The solution of system is UUB stable if for  $\forall x \subset \Omega$ , there exists a bound  $B$  and time  $T(B, x_0)$  such that  $\|x(t) - x_0\| \leq B$  for all  $t \geq t_0 + T$ .

For the nonlinear system (1), the robust optimal stabilization problem investigated in this paper is to design a robust controller  $u(x) \in \mathcal{D}_u$  based on the optimal control, such that the closed-loop system is stable in the sense of UUB under constraints (2) and uncertainties  $d(x)$ .

### 2.2. Preliminaries

The nominal system is defined as

$$\dot{x} = f(x) + g(x)u(x), \quad (3)$$

and the corresponding performance index is

$$J(x(0), u(x)) = \int_0^\infty (\rho^2 \|x\|^2 + Q(x) + U(u)) d\tau, \quad (4)$$

where  $x(0)$  shows the initial state,  $Q(x)$  and  $U(u)$  will be given in the following. As a result of [24], the robust control for the system (1) with (2) can be obtained by solving a constrained optimal control problem denoted by (3), (4) and (2). However, the problem is difficult to solve because the constraints are involved.

To copy with this difficulty, a non-quadratic penalty term and a barrier function are introduced for the performance index (4). For the input saturation, the non-quadratic penalty term  $U(u)$  is defined as [26]

$$U(u) = 2 \int_0^u (\eta \tanh^{-1}(\xi/\eta))^T R d\xi, \quad (5)$$

where  $\tanh^{-1}(\xi/\eta) = [\tanh^{-1}(\xi_1/\eta), \dots, \tanh^{-1}(\xi_m/\eta)]^T$ ,  $\tanh^{-1}(\cdot) = \text{artanh}(\cdot)$  and  $R = \text{diag}(r_1, \dots, r_m)$  with  $r_i > 0$ . Note that  $U(u)$  is positive definite because  $\tanh^{-1}(\cdot)$  is a strictly monotonic odd function and  $R$  is positive definite.

To deal with the state constraints, the barrier function  $Q_s(x)$  is designed as follows:

$$Q_s(x) = \alpha_s \sum_{i=1}^n |x_i|^2 \ln \left( \frac{\beta^2}{\beta^2 - x_i^2} \right), \quad (6)$$

where  $\alpha_s > 0$  is a small constant. It is obvious that the cost rate  $Q(x) = x^T S x + Q_s(x) \geq 0$  for all  $|x_i| < \beta$  and the equation holds if and only if  $x = 0$ , where  $S$  is a symmetric positive definite matrix. Moreover,  $Q(x) \rightarrow +\infty$  when  $x_i \rightarrow \beta$ .

**Remark 1:** Compared with [24–27], the state and input constraints are handled simultaneously in this paper. The constraints have been involved in the cost functions  $U(u)$  and  $Q(x)$ . The state and control will be confined as long as the performance index  $J(x(0), u(x))$  is finite. Given these cost functions, the constrained optimal control problem will be converted into an unconstrained one.

**Definition 2** (Admissible control) [11]: A control policy  $\mu(x)$  is said to be admissible on a compact set  $\Omega$  for (3) with respect to (4), denoted by  $\mu \in \Pi(\Omega)$ , if  $\mu(x)$  is continuous with  $\mu(\mathbf{0}) = \mathbf{0}$ ,  $u(x) = \mu(x)$  stabilizes (3), and  $V(x_0)$  is finite for all  $x_0$ .

Considering an admissible control policy  $u(x)$ , the value function is written as

$$V(x) = \int_t^\infty (\rho^2 \|x\|^2 + Q(x) + U(u)) d\tau, \quad (7)$$

with  $V(x) \in \mathcal{C}^1$ , where  $\mathcal{C}^n$  denotes the set of  $n$ -times continuously differentiable functions. For the nominal system (3), define the Hamiltonian function

$$H(x, u(x), \nabla_x V^*(x)) = \nabla_x V^{*T}(x)(f(x) + g(x)u(x)) + \rho^2 \|x\|^2 + Q(x) + U(u), \quad (8)$$

where  $\nabla_x V^{*T}(x) = (\partial V^*(x)/\partial x)^T$ .  $V^*(x)$  denotes the optimal value function, which is obtained from the following HJB equation

$$\min_{u(x)} H(x, u(x), \nabla_x V^*(x)) = 0, \quad (9)$$

with  $V^*(\mathbf{0}) = 0$ , and the optimal control is derived as

$$u^*(x) = \eta \tanh \left( -\frac{1}{2\eta} R^{-1} g^T(x) \nabla_x V^*(x) \right). \quad (10)$$

Then, the HJB equation (9) can be rewritten as

$$\begin{aligned} & \nabla_x V^{*T}(x) f(x) + \rho^2 \|x\|^2 + Q(x) - 2\eta^2 \mathcal{A}^T R \tanh(\mathcal{A}) \\ & + 2 \int_0^{\eta \tanh(-\mathcal{A})} (\eta \tanh^{-1}(\xi/\eta))^T R d\xi = 0, \end{aligned} \quad (11)$$

where  $\mathcal{A} = 1/(2\eta)R^{-1}g^T(x)\nabla_x V^*(x)$ .

**Theorem 1:** Let Assumptions 1 and 2 hold. Consider the nominal system (3) with the value function (7) and the HJB equation (11). If  $d^T(x)Rd(x) \leq \rho^2 \|x\|^2$ ,

- 1) the optimal control  $u^*(x)$  in (10) ensures the state of system (1) to be UUB, and
- 2) the relationship (2) is satisfied.

**Proof:** For 1), since the additional cost rate  $Q_s(x)$  has no impact on the proof, the details can be found in [24]. Based on 1), we can find that the optimal value function  $V^*(x)$  is finite. Thus, the state and input constraints are guaranteed. This completes the proof.  $\square$

**Remark 2:** The resulting robust control  $u^*(x)$  of system (1) is optimal with respect to the predefined cost function. Therefore,  $u^*(x)$  is also called the robust optimal controller in this paper.

Obviously, the robust optimal control  $u^*(x)$  can be obtained by solving the optimal value function  $V^*(x)$ . However, it is intractable to solve the HJB equation analytically for general nonlinear systems. In order to conquer this challenge, an online IRL-based algorithm is proposed to find the approximate solution of the HJB equation in the subsequent sections.

### 3. NEAR-OPTIMAL CONTROLLER DESIGN

In this section, an online partially model-free method is presented to obtain the near-optimal controller. An actor-critic structure is constructed to approximate the optimal value function and optimal control firstly. Next, an online update law with adjustable gain for the actor-critic scheme is presented. Finally, the stability and convergence analysis for the closed-loop system is provided.

#### 3.1. Actor-critic structure

According to the Bellman's principle of optimality, for any time  $t > T$  and time interval  $T > 0$ , when control  $u(x)$  is admissible, the value function  $V(x)$  satisfies the following integral Bellman equation

$$\begin{aligned} V(x(t)) &= V(x(t+T)) \\ &+ \int_t^{t+T} (\rho^2 \|x\|^2 + Q(x) + U(u)) d\tau. \end{aligned} \quad (12)$$

Note that the system dynamics  $f(x)$  and  $g(x)$  are not included in (12). The corresponding Lyapunov equation that is used to design the learning algorithm is as follows:

$$\begin{aligned} & V(x(t+T)) - V(x(t)) \\ & + \int_t^{t+T} (\rho^2 \|x\|^2 + Q(x) + U(u)) d\tau = 0. \end{aligned} \quad (13)$$

To approximate the optimal value function and optimal control policy, an actor-critic structure is employed in this paper. Based on the Weierstrass high-order approximation theorem [30], there exists a single-layer NN such that the optimal value function can be expressed as

$$V^*(x) = w_c^{*T} \Theta_c(x) + \varepsilon_c(x), \quad (14)$$

where  $\Theta_c(x) \in \mathbb{R}^{N_c}$  is a vector of polynomial basis functions,  $N_c$  is the number of neurons,  $w_c^* \in \mathbb{R}^{N_c}$  is the ideal

weights and  $\varepsilon_c(x)$  is the approximation error. Since the optimal critic weights  $w_c^*$  are unknown in general, the current weights estimation  $\hat{w}_c$  will be used. As a result, the critic NN can be obtained as follows:

$$\hat{V}(x) = \hat{w}_c^T \Theta_c(x), \quad (15)$$

where  $\hat{V}(x)$  is the estimation of the optimal value function  $V^*(x)$ .

Similarly, the optimal control is approximated by

$$u^*(x) = \eta \tanh(\Theta_a^T(x) w_a^*/\eta) + \varepsilon_a(x), \quad (16)$$

where the notions of  $\Theta_a(x) \in \mathbb{R}^{N_a \times m}$ ,  $N_a$ ,  $w_a^* \in \mathbb{R}^{N_a}$  and  $\varepsilon_a(x) \in \mathbb{R}^m$  are same as critic NN. Then, the actor NN is denoted as

$$\hat{u}(x) = \eta \tanh(\Theta_a^T(x) \hat{w}_a/\eta), \quad (17)$$

where  $\hat{w}_a$  and  $\hat{u}(x)$  show the estimations of the ideal actor weights  $w_a^*$  and the optimal control  $u^*(x)$ , respectively.

**Remark 3:** This paper considers the continuous optimal value function  $V^*(x)$  and the continuous optimal control  $u^*(x)$ . Therefore, the last term  $\varepsilon_c(x)$  and  $\varepsilon_a(x)$  in (14) and (16) will converge uniformly to zero as  $N_c \rightarrow \infty$  and  $N_a \rightarrow \infty$ . In addition, for fixed  $N_c$  and  $N_a$ , the approximation errors  $\varepsilon_c(x)$  and  $\varepsilon_a(x)$  are bounded by constants on a compact set.

The following assumption is supplied to analyze the system stability conveniently, which is necessary in many studies such as [15,25,26].

**Assumption 3:** The assumptions are given as follows:

- 1) The basis functions and their gradients are bounded for all  $x \in \Omega$ , i.e.,  $\|\Theta_c(x)\| \leq b_c$ ,  $\|\nabla_x \Theta_c(x)\| \leq b_{cx}$ ,  $\|\Theta_a(x)\| \leq b_a$  and  $\|\nabla_x \Theta_a(x)\| \leq b_{ax}$ .
- 2) The approximation errors and their gradients are bounded for all  $x \in \Omega$ , i.e.,  $\|\varepsilon_c(x)\| \leq b_{\varepsilon c}$ ,  $\|\varepsilon_a(x)\| \leq b_{\varepsilon a}$ ,  $\|\nabla_x \varepsilon_c(x)\| \leq b_{\varepsilon cx}$  and  $\|\nabla_x \varepsilon_a(x)\| \leq b_{\varepsilon ax}$ .
- 3) The ideal weights and their estimations are bounded, i.e.,  $\|w_c^*\| \leq b_{wc}^*$ ,  $\|w_a^*\| \leq b_{wa}^*$ ,  $\|\hat{w}_c\| \leq b_{wc}$  and  $\|\hat{w}_a\| \leq b_{wa}$ .

### 3.2. Update laws design

In order to make the weight estimations  $\hat{w}_c$  and  $\hat{w}_a$  converge to ideal values  $w_c^*$  and  $w_a^*$ , an online update law is developed for actor-critic scheme to tune the critic and actor weights simultaneously.

Substituting (15) into (13), the Bellman error  $e_c$  is given by

$$e_c = \hat{w}_c^T \sigma(x) + \int_t^{t+T} (\rho^2 \|x\|^2 + Q(x) + U(u)) d\tau, \quad (18)$$

where  $\sigma(x) = \Theta_c(x(t+T)) - \Theta_c(x(t))$  with  $\|\sigma(x)\| \leq b_\sigma$ , and define  $\sigma \triangleq \sigma(x)$ . The error  $e_a \in \mathbb{R}^m$  between the estimated control input and the optimal control policy is calculated by

$$e_a = \hat{u}(x) - u^*(x). \quad (19)$$

In [25], the critic and actor weights can be adjusted using the gradient descent method to minimize the goal functions  $E_c = \frac{1}{2} e_c^T e_c$  and  $E_a = \frac{1}{2} e_a^T e_a$ . However, this learning algorithm requires the PE condition and an initial admissible control policy throughout learning, both of which are difficult to achieve for general nonlinear systems. To remedy these, the following update laws are proposed.

The update law for the critic NN, employing the gradient descent method and ER technology, is designed as follows:

$$\begin{aligned} \dot{\hat{w}}_c = & -\alpha_c \frac{\sigma}{(\sigma^T \sigma + 1)^2} e_c^T \\ & -\alpha_c \sum_{j=1}^{q_0} \frac{\sigma_j}{(\sigma_j^T \sigma_j + 1)^2} (\sigma_j^T \hat{w}_c + K_j), \end{aligned} \quad (20)$$

where  $\alpha_c > 0$  is the adjustable gain,  $\sigma_j = \Theta_c(x(t_j+T)) - \Theta_c(x(t_j))$ ,  $K_j = \int_{t_j}^{t_j+T} (\rho^2 \|x\|^2 + Q(x) + U(\hat{u})) d\tau$  and  $j$  ( $j = 1, \dots, q_0$ ) denotes the  $j$ th sample data stored in the history stack  $L$ . It is shown that the critic weights are updated using the historical and current data to ensure the convergence of  $\hat{w}_c$ .

**Remark 4:** Let  $L = [\bar{\sigma}_1, \dots, \bar{\sigma}_{q_0}]$  be the history stack, where  $\bar{\sigma} = \sigma/(\sigma^T \sigma + 1)$  and  $q_0$  is the number of samples. If  $\text{rank}(L) = N_c$ , the history stack  $L$  in the recorded data contains as many linearly independent elements as the number of neurons in (15) [26]. As presented in [31,32], the summation term in (20) can guarantee that the critic weight estimations  $\hat{w}_c$  converge to the ideal value within a residual set provided that  $\text{rank}(L) = N_c$ .

**Remark 5:** In order to satisfy the condition  $\text{rank}(L) = N_c$ , the number of stored samples  $q_0$  must be a fixed value with  $q_0 > N_c$ , and the data in  $L$  is updated periodically in simulations. Note that  $N_c$  is commonly carried out by computer simulations. Thus, the condition  $\text{rank}(L) = N_c$  is equivalent to a PE-like condition, guaranteeing that the Lyapunov derivative is negative definite.

The following tuning law is developed for the actor NN, with the first term minimizing  $E_a$  and the second term (i.e., the stabilization term) ensuring the system stability throughout learning.

$$\begin{aligned} \dot{\hat{w}}_a = & -\alpha_a \Theta_a(x) \tanh'(\hat{A}) e_a^T \\ & -\Upsilon \alpha_a \Theta_a(x) \tanh'(\hat{A}) g^T(x) x, \end{aligned} \quad (21)$$

$$\Upsilon = \begin{cases} 0, & \text{if } x^T(t+T)x(t+T) - x^T(t)x(t) \leq 0, \\ 1, & \text{otherwise,} \end{cases} \quad (22)$$

where  $\alpha_a > 0$  is the adjustable gain and  $\hat{A} = \Theta_a^T(x)\hat{w}_a/\eta$ .

Furthermore, unlike the update laws with constant learning rate in [20,25,26], the adjustable gains  $\alpha_c$  and  $\alpha_a$  are updated in the following way:

$$\dot{\alpha}_c = \begin{cases} 2\alpha_c k_c (|e_c| - e_{\zeta c}), & |e_c| \geq e_{\zeta c}, \\ 0, & |e_c| < e_{\zeta c}, \end{cases} \quad (23)$$

$$\dot{\alpha}_a = \begin{cases} 2\alpha_a k_a (\|e_a\| - e_{\zeta a}), & \|e_a\| \geq e_{\zeta a}, \\ 0, & \|e_a\| < e_{\zeta a}, \end{cases} \quad (24)$$

where  $k_c > 0$  and  $k_a > 0$  are constants and  $e_{\zeta c}$  and  $e_{\zeta a}$  are predefined values.

**Remark 6:** Compared with constant gains, the advantage of the update laws in (20)-(24) is that the critic and actor NNs can modify their learning rate to increase convergence speed based on  $e_c$  and  $e_a$ . It can also guarantee that the approximation errors will converge to within the prescribed values as fast as possible.

A pseudocode (with inline comments to provide guidance after the symbol  $\triangleright$ ) that describes the proposed robust near-optimal control method has the form shown in Algorithm 1.

**Remark 7:** Since the optimal control  $u^*(x)$  is unknown,  $u^*(x)$  in (19) can be computed using (10) and (15) in practice. The system drift dynamics  $f(x)$  are not required in the proposed method.

**Lemma 1:** Suppose Assumption 3 holds. Consider the critic NN (15) and its update law (20) with (23) for an admissible control policy  $u(x)$ . If the history stack  $L$  satisfies the condition  $\text{rank}(L) = N_c$ , the weight estimation error  $\tilde{w}_c = w_c^* - \hat{w}_c$  is UUB.

**Proof:** The Bellman error  $\epsilon_B$  is

$$\epsilon_B = w_c^{*T} \sigma + \int_t^{t+T} (\rho^2 \|x\|^2 + Q(x) + U(u)) d\tau. \quad (25)$$

There exists a bound  $b_{\epsilon B}$  under Assumption 3 such that  $\|\epsilon_B\| \leq b_{\epsilon B}$ . Based on the definition of  $\epsilon_B$ , it follows that

$$\begin{aligned} e_c &= -\tilde{w}_c^T \sigma + \epsilon_B, \\ e_{c_j} &= -\tilde{w}_c^T \sigma_j + \epsilon_{B_j}. \end{aligned} \quad (26)$$

Thus, we obtain

$$\dot{\tilde{w}}_c = -\alpha_c \bar{D} \tilde{w}_c + \alpha_c D_1, \quad (27)$$

where  $\bar{D} = \bar{\sigma} \bar{\sigma}^T + \sum_{j=1}^{q_0} \bar{\sigma}_j \bar{\sigma}_j^T$ ,  $D_1 = (\bar{\sigma}/m_s)\epsilon_B + \sum_{j=1}^{q_0} (\bar{\sigma}_j/m_{s_j})\epsilon_{B_j}$ ,  $\bar{\sigma} = \sigma/(\sigma^T \sigma + 1)$ ,  $m_s = \sigma^T \sigma + 1$ ,  $\bar{\sigma}_j = \sigma_j/(\sigma_j^T \sigma_j + 1)$  and  $m_{s_j} = \sigma_j^T \sigma_j + 1$ . Note that, if Assumption 3 holds and  $\text{rank}(L) = N_c$ ,  $\bar{D}$  is positive and  $\|D_1\| \leq (1+q_0)b_{\epsilon B}$ .

Consider the Lyapunov function

$$L_2 = \frac{1}{2} \alpha_c^{-1} \tilde{w}_c^T \tilde{w}_c, \quad (28)$$

**Algorithm 1:** Robust near-optimal control algorithm based on IRL.

- 
- 1: Start with initial state  $x(0)$ , initial weights  $\hat{w}_c(0)$ ,  $\hat{w}_a(0)$
  - 2: **procedure**
  - 3: Propagate  $t$ ,  $x(t)$  using system (1) and  $u(t) = \hat{u}(x)$   
 $\triangleright \{x(t)\}$  is from integrating the system (1) with Runge-Kutta method, and  $t$  is from the integral interval  $[t_i, t_i + T]$ ,  $i \in N_{it}$  where  $T \in \mathbb{R}^+$  is the step size
  - 4: **if**  $\text{length}(L) < q_0$  or  $\text{rand}(L) \neq N_c$   $\triangleright$  {the history stack  $L = [\bar{\sigma}_1, \dots, \bar{\sigma}_{q_0}]$  must have  $N_c$  linearly independent element} **then**
  - 5: Select an arbitrary data point to be included in the history stack  $L$   $\triangleright$  {periodically update data in  $L$  (c.f. Remarks 4-5)}
  - 6: **end if**
  - 7: Update  $\alpha_c$  and  $\alpha_a$   $\triangleright$  {integrate  $\dot{\alpha}_c$  as in (23) and  $\dot{\alpha}_a$  as in (24) using Runge-Kutta method}
  - 8: Propagate  $\hat{w}_c$  and  $\hat{w}_a$   $\triangleright$  {integrate  $\dot{\hat{w}}_c$  as in (20) and  $\dot{\hat{w}}_a$  as in (21) with Runge-Kutta method}
  - 9: Compute  $\hat{V}(x)$  using (15)  $\triangleright$  output of the critic NN
  - 10: Compute  $\hat{u}(x)$  using (17)  $\triangleright$  output of the actor NN
  - 11: **end procedure**
- 

its time derivative is given by

$$\begin{aligned} \dot{L}_2 &= -\frac{1}{2} \alpha_c^{-2} \dot{\alpha}_c \tilde{w}_c^T \tilde{w}_c + \alpha_c^{-1} \tilde{w}_c^T \dot{\tilde{w}}_c \\ &= -\frac{1}{2} \alpha_c^{-2} \dot{\alpha}_c \tilde{w}_c^T \tilde{w}_c - \tilde{w}_c^T \bar{D} \tilde{w}_c + \tilde{w}_c^T D_1 \\ &\leq -\left( \lambda_{\min}(\bar{D}) + \frac{1}{2} \alpha_c^{-2} \dot{\alpha}_c \right) \|\tilde{w}_c\|^2 \\ &\quad + (1+q_0)b_{\epsilon B} \|\tilde{w}_c\|. \end{aligned}$$

Therefore,  $\dot{L}_2 < 0$  provided that

$$\|\tilde{w}_c\| > \frac{(1+q_0)b_{\epsilon B}}{\lambda_{\min}(\bar{D}) + \frac{1}{2} \alpha_c^{-2} \dot{\alpha}_c}. \quad (29)$$

It is shown that the critic weight estimation error  $\tilde{w}_c$  is UUB. This completes the proof.  $\square$

### 3.3. Stability and convergence analysis

The main result of this paper is presented by the following theorem.

**Theorem 2:** Suppose Assumptions 1-3 hold. The history stack  $L$  satisfies the condition  $\text{rank}(L) = N_c$ . Consider the system (3) and control policy (17). The update laws of the critic and actor NNs are described as (20) and (21). The critic and actor weight estimate errors are defined as  $\tilde{w}_c = w_c^* - \hat{w}_c$  and  $\tilde{w}_a = w_a^* - \hat{w}_a$ , respectively. Then,

- 1) the system state  $x$  and the weight estimate errors  $\tilde{w}_c$ ,  $\tilde{w}_a$  are UUB with the proper parameters and sufficiently large  $N_a$ ,

- 2) the approximate value function  $\hat{V}(x)$  and the approximate optimal control  $\hat{u}(x)$  converge to the optimal values within finite bounds, and
- 3) the adjustable gains  $\alpha_c$  and  $\alpha_a$  are bounded.

**Proof:** Note that the system (3) with control (17) constitutes the closed-loop system. Thus, the closed-loop system involves  $x$ ,  $\tilde{w}_c$  and  $\tilde{w}_a$ . The Lyapunov function candidate is given as follows:

$$L_4 = \int_t^{t+T} V^*(x) d\tau + \frac{1}{2} \alpha_c^{-1} \tilde{w}_c^T \tilde{w}_c + \frac{1}{2} \alpha_a^{-1} \tilde{w}_a^T \tilde{w}_a + \frac{1}{2} x^T x, \quad (30)$$

where  $V^*(x)$  is the optimal value function,  $x$  is the system states,  $\alpha_c$  and  $\alpha_a$  are adjustable gains,  $\tilde{w}_c$  and  $\tilde{w}_a$  are weight estimation errors. Note that the terms of the Lyapunov function candidate  $L_4$  are relevant to the optimal value function, weight errors and system states.

Firstly, consider  $L_1 = \int_t^{t+T} V^* d\tau$ . The time derivative of  $L_1$  is

$$\begin{aligned} \dot{L}_1 &= \int_t^{t+T} \dot{V}^* d\tau \\ &= \int_t^{t+T} (\nabla_x \Theta_c^T w_c^* + \nabla_x \mathcal{E}_c)^T (f + g\hat{u}) d\tau. \end{aligned} \quad (31)$$

Define HJB approximation error as

$$\epsilon_H = w_c^{*T} \nabla_x \Theta_c (f + g\hat{u}^*) + \rho^2 \|x\|^2 + Q(x) + U(u^*), \quad (32)$$

with  $\|\epsilon_H\| \leq b_H$ .

Together with (31), (32), and the relationship

$$\|\nabla_x \mathcal{E}_c^T (f + g\hat{u})\| \leq b_{\epsilon c x} b_f \|x\| + b_{\epsilon c x} b_g b_{\hat{u}},$$

the  $\dot{L}_1$  can be written as

$$\begin{aligned} \dot{L}_1 &= \int_t^{t+T} \left( w_c^{*T} \nabla_x \Theta_c g (\hat{u} - u^*) - \rho^2 \|x\|^2 - Q(x) \right) d\tau \\ &\quad + \int_t^{t+T} (-U(u^*) + \epsilon_H + \nabla_x \mathcal{E}_c^T (f + g\hat{u})) d\tau \\ &\leq \int_t^{t+T} \left( -\rho^2 \|x\|^2 - x^T S x + b_{11} \|x\| + b_{12} \right) d\tau \\ &\leq \int_t^{t+T} \left( -b_{10} \|x\|^2 + b_{11} \|x\| + b_{12} \right) d\tau, \end{aligned}$$

where  $b_{10} = \rho^2 + \lambda_{\min}(S)$ ,  $b_{11} = b_{\epsilon c x} b_f$  and  $b_{12} = b_{w_c}^* b_{\epsilon c x} b_g (b_{\hat{u}} + b_{u^*}) + b_H + b_{\epsilon c x} b_g b_{\hat{u}}$ .

Secondly, consider  $L_2 = 1/2 \alpha_c^{-1} \tilde{w}_c^T \tilde{w}_c$ . The result of  $\dot{L}_2$  can be written as

$$\dot{L}_2 \leq -b_{20} \|\tilde{w}_c\|^2 + b_{21} \|\tilde{w}_c\|, \quad (33)$$

where  $b_{20} = \lambda_{\min}(\bar{D}) + (1/2) \alpha_c^{-2} \dot{\alpha}_c$  and  $b_{21} = (1 + q_0) b_{\epsilon B}$ .

Thirdly, consider  $L_3 = (1/2) \alpha_a^{-1} \tilde{w}_a^T \tilde{w}_a$ . The time derivative of  $L_3$  is

$$\begin{aligned} \dot{L}_3 &= -\frac{1}{2} \alpha_a^{-2} \dot{\alpha}_a \tilde{w}_a^T \tilde{w}_a + \alpha_a^{-1} \tilde{w}_a^T \dot{\tilde{w}}_a \\ &= -\frac{1}{2} \alpha_a^{-2} \dot{\alpha}_a \tilde{w}_a^T \tilde{w}_a + \underbrace{\tilde{w}_a^T \Theta_a \tanh'(\hat{A}) e_a}_{C_1} \\ &\quad + \underbrace{\Upsilon \tilde{w}_a^T \Theta_a \tanh'(\hat{A}) g^T x}_{C_2}. \end{aligned} \quad (34)$$

According to the results in [25], we have

$$\begin{aligned} C_1 &= -\eta^2 \hat{A}^T \tanh'(\hat{A}) \tanh(\hat{A}) - \tilde{w}_a^T \Theta_a \tanh'(\hat{A}) \epsilon_a \\ &\quad - \eta^2 A^{*T} \tanh'(\hat{A}) \tanh(A^*), \end{aligned} \quad (35)$$

where  $A^* = \Theta_a^T(x) w_a^* / \eta$  and  $\hat{A} = \Theta_a^T(x) \hat{w}_a / \eta$ . Since  $\tanh(\mathbf{0}) = \mathbf{0}$ ,  $\tanh'(\cdot) > 0$ ,  $v^T \tanh(v) > 0$  and  $\tanh(\cdot)$  is a strictly monotonic odd function,  $C_1$  is negative for sufficiently large number of basis functions in (17). Thus, the following inequality holds

$$\dot{L}_3 \leq -\frac{1}{2} \alpha_a^{-2} \dot{\alpha}_a \tilde{w}_a^T \tilde{w}_a + C_2. \quad (36)$$

Finally, the time derivative of Lyapunov function  $L_4$  can be given by

$$\begin{aligned} \dot{L}_4 &\leq \int_t^{t+T} \left( -b_{10} \|x\|^2 + b_{11} \|x\| + b_{12} \right) d\tau \\ &\quad - b_{20} \|\tilde{w}_c\|^2 + b_{21} \|\tilde{w}_c\| \\ &\quad - \frac{1}{2} \alpha_a^{-2} \dot{\alpha}_a \tilde{w}_a^T \tilde{w}_a + C_2 \\ &\quad + x^T (f + g\hat{u}). \end{aligned} \quad (37)$$

(i) When  $\Upsilon = 0$ , that is  $C_2 = 0$  and  $x^T(t+T)x(t+T) - x^T(t)x(t) \leq 0$ . For  $t > T$ , we obtain  $1/2(x^T(t+T)x(t+T) - x^T(t)x(t)) = \int_t^{t+T} x^T \dot{x} d\tau = \int_t^{t+T} x^T (f + g\hat{u}) d\tau \leq 0$  which implies that  $x^T (f + g\hat{u}) < 0$  for  $\|x\| > 0$ . Then, (37) satisfies the following relationship

$$\begin{aligned} \dot{L}_4 &\leq \int_t^{t+T} \left( -b_{10} \|x\|^2 + b_{11} \|x\| + b_{12} \right) d\tau \\ &\quad - b_{20} \|\tilde{w}_c\|^2 + b_{21} \|\tilde{w}_c\| \\ &\quad - \frac{1}{2} \alpha_a^{-2} \dot{\alpha}_a \tilde{w}_a^T \tilde{w}_a. \end{aligned} \quad (38)$$

It is observed that  $\dot{L}_4 < 0$  provided that

$$\|x\| > \frac{b_{11}}{2b_{10}} + \sqrt{\frac{b_{11}^2 + 4b_{10}b_{12}}{4b_{10}^2}}, \quad (39)$$

$$\|\tilde{w}_c\| > \frac{b_{21}}{b_{20}}. \quad (40)$$

Inspired by [26], the upper and lower bounds on the Lyapunov function candidate  $L_4$  can be found, i.e., there exist class  $k$  functions  $k_1$  and  $k_2$  such that

$$k_1(\|z\|) \leq L_4 \leq k_2(\|z\|), \quad (41)$$

where  $z = [x^T, \tilde{w}_c^T, \tilde{w}_a^T]^T$ . Moreover, based on this notation, the first line in (38) is negative if (39) holds. The second line in (38) will be negative if (40) holds. And the last line in (38) is also negative. Thus, it is concluded that  $\dot{L}_4 < 0$  if  $\|z\| > z_1$  for some  $z_1$ . Because  $L_4$  in (30) is positive and  $\dot{L}_4$  is negative for sufficiently large  $\|z\|$ , it follows that  $z$  is UUB.

(ii) When  $\Upsilon = 1$ , that is  $x^T(f + g\hat{u}) \geq 0$ . Let  $W = C_2 + x^T(f + g\hat{u})$ , we have

$$W = x^T g \tanh'(\hat{A}) \Theta_a^T \tilde{w}_a + x^T f + x^T g \hat{u}. \quad (42)$$

Based on Taylor series, the following equality holds.

$$\begin{aligned} \hat{u} - u^* &= \eta \tanh(\hat{A}) - \eta \tanh(A^*) - \varepsilon_a \\ &= -\eta \tanh'(\hat{A})(A^* - \hat{A}) - O((A^* - \hat{A})^2) - \varepsilon_a \\ &= -\tanh'(\hat{A}) \Theta_a^T \tilde{w}_a - O((\Theta_a^T \tilde{w}_a / \eta)^2) - \varepsilon_a, \end{aligned}$$

where  $O((\Theta_a^T \tilde{w}_a / \eta)^2)$  is the high order term in Taylor series with  $\|O((\Theta_a^T \tilde{w}_a / \eta)^2)\| \leq b_{T1} + b_{T2} \|\tilde{w}_a\|$ .

According to the assumptions in [15], for the system (3) and the optimal control policy  $u^*(x)$ , there exists a symmetric positive definite matrix  $\Gamma(x)$  such that  $x^T(f + gu^*) = -x^T \Gamma(x)x$ . Hence,  $W$  can be written as

$$\begin{aligned} W &= x^T f + x^T g (u^* - O((\Theta_a^T \tilde{w}_a / \eta)^2) - \varepsilon_a) \\ &= x^T (f + gu^*) - x^T g O((\Theta_a^T \tilde{w}_a / \eta)^2) - x^T g \varepsilon_a \\ &= -x^T \Gamma(x)x - x^T g O((\Theta_a^T \tilde{w}_a / \eta)^2) - x^T g \varepsilon_a \\ &\leq -b_{30} \|x\|^2 + b_{31} \|x\| + b_{32} \|\tilde{w}_a\| \|x\| \\ &= -\iota_1 b_{30} \|x\|^2 + b_{32} \|\tilde{w}_a\| \|x\| \\ &\quad - \iota_2 b_{30} \|x\|^2 + b_{31} \|x\| \\ &= -\iota_1 b_{30} \left( \|x\| - \frac{b_{32}}{2\iota_1 b_{30}} \|\tilde{w}_a\| \right)^2 + \frac{b_{32}^2}{4\iota_1 b_{30}} \|\tilde{w}_a\|^2 \\ &\quad - \iota_2 b_{30} \|x\|^2 + b_{31} \|x\| \\ &\leq b_{33} \|\tilde{w}_a\|^2 - \iota_2 b_{30} \|x\|^2 + b_{31} \|x\|, \end{aligned} \quad (43)$$

where  $b_{30} = \lambda_{\min}(\Gamma)$ ,  $b_{31} = b_g(b_{T1} + b_{\varepsilon a})$ ,  $b_{32} = b_g b_{T2}$  and  $b_{33} = b_{32}^2 / (4\iota_1 b_{30})$ . The notations  $\iota_1$  and  $\iota_2$  satisfy  $0 < \iota_1 < 1$ ,  $0 < \iota_2 < 1$  and  $\iota_1 + \iota_2 = 1$ .

Considering (37) and (43), we obtain

$$\begin{aligned} \dot{L}_4 &\leq \int_t^{t+T} \left( -b_{10} \|x\|^2 + b_{11} \|x\| + b_{12} \right) d\tau \\ &\quad - \iota_2 b_{30} \|x\|^2 + b_{31} \|x\| \\ &\quad - b_{20} \|\tilde{w}_c\|^2 + b_{21} \|\tilde{w}_c\| \\ &\quad + \left( -\frac{1}{2} \alpha_a^{-2} \dot{\alpha}_a + b_{33} \right) \|\tilde{w}_a\|^2. \end{aligned} \quad (44)$$

When  $\Upsilon = 1$ , it is obvious that the control policy  $\hat{u}(x)$  does not converge to the optimal value  $u^*(x)$ , and the inequality  $\|e_a\| \geq e_{\zeta a}$  holds, i.e.,  $\dot{\alpha}_a > 0$ . As a result, by choosing the parameters appropriately, the term

$-\frac{1}{2} \alpha_a^{-2} \dot{\alpha}_a + b_{33}$  can be negative. Then,  $\dot{L}_4 < 0$  provided that

$$\|x\| > \max \left\{ \frac{b_{11}}{2b_{10}} + \sqrt{\frac{b_{11}^2 + 4b_{10}b_{12}}{4b_{10}^2}}, \frac{b_{31}}{\iota_2 b_{30}} \right\}, \quad (45)$$

$$\|\tilde{w}_c\| > \frac{b_{21}}{b_{20}}, \quad (46)$$

$$-\frac{1}{2} \alpha_a^{-2} \dot{\alpha}_a + b_{33} < 0. \quad (47)$$

Similarly, let  $z = [x^T, \tilde{w}_c^T, \tilde{w}_a^T]^T$  and the relationship (41) holds. Then, both the first line and second line in (44) are negative if (45) holds. The third line is negative if (46) holds. And the last line will be also negative if (47) holds. Thus, it is clear that  $\dot{L}_4 < 0$  if  $\|z\| > z_2$  for some  $z_2$ . Because  $L_4$  in (30) is positive and  $\dot{L}_4$  is negative for sufficiently large  $\|z\|$ , it follows that  $z$  is UUB.

Combining (i) and (ii) and using the standard Lyapunov extension theorem [33], it is concluded that the system state  $x$  and weight estimation errors  $\tilde{w}_c, \tilde{w}_a$  are UUB with ultimate bounds.

Based on the analysis above, the following condition holds:  $\|\hat{V} - V^*\| \leq \frac{(1+q_0)b_{\varepsilon B}b_c}{\lambda_{\min}(D) + \frac{1}{2}\alpha_c^{-2}\dot{\alpha}_c} + b_{\varepsilon c} = \gamma_V$ . On the other hand, the term with regard to  $\tilde{w}_a$  is negative for sufficiently large  $N_a$  and relationship (47), thus  $\|\hat{u} - u^*\| \leq \gamma_u$ , where  $\gamma_u$  is a small positive constant. This indicates that the approximate value function and approximate optimal control can converge to the optimal values within finite bounds.

Furthermore, if  $|e_c| \geq e_{\zeta c}$  and  $\|e_a\| \geq e_{\zeta a}$ , the state  $x$  and weight estimate errors  $\tilde{w}_c, \tilde{w}_a$  will approach the residual set. Therefore, the adjustable gains  $\alpha_c$  and  $\alpha_a$  are bounded based on (23) and (24). This completes the proof.  $\square$

**Remark 8:** According to Theorems 1 and 2, the control law in (17) can guarantee the state of the system (1) to be UUB under constraints and uncertainties.

## 4. SIMULATION RESULTS

In this section, the effectiveness of the proposed algorithm is validated by using two numerical cases. The first case is used to show that the proposed method obtains the near-optimal controller for a nominal nonlinear system. The second case is carried out to demonstrate the feasibility for a nonlinear system in the presence of constraints and uncertainties. In the simulations below, since the history stack is empty at the beginning, a small probing noise in the form of  $n_e(t) = 0.1 * (\sin(0.3\pi t) + \cos(0.3\pi t))$  will be added to the control input for the first second.

Consider a torsional pendulum system, the dynamics are described as follows [34]:

$$\begin{cases} \frac{d\theta}{dt} = \omega, \\ J \frac{d\omega}{dt} = \tau - Mgl \sin \theta - f_d \frac{d\theta}{dt}, \end{cases}$$

where the pendulum's mass  $M = (1/3)$  kg, its length  $l = (2/3)$  m, the rotary inertia  $J = (4/3)$  kg·m<sup>2</sup>, the gravitational acceleration  $g = 9.8$  m/s<sup>2</sup> and the frictional factor  $f_d = 0.2$ . The system state is  $x = [\theta, \omega]^T = [x_1, x_2]^T$  denoted by the angle and angular velocity, and the control input is  $u = \tau$ . Then the nonlinear system is given by

$$\dot{x} = \begin{bmatrix} x_2 \\ -\frac{1}{J}(Mgl \sin(x_1) + f_d x_2) \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{J} \end{bmatrix} u. \quad (48)$$

In simulations, the initial state  $x(0) = [0.8, -0.8]^T$ ,  $S = I_2$ ,  $R = 1$ ,  $e_{\zeta_c} = e_{\zeta_a} = 10^{-4}$ ,  $k_c = 10$ ,  $k_a = 5$ ,  $T = 0.01$ s, and the initial adjustable gains  $\alpha_c(0) = 150$ ,  $\alpha_a(0) = 10$ . The basis functions of critic and actor NNs are chosen as  $\Theta_c(x) = [x_1^2, x_1 x_2, x_2^2]^T$  and  $\Theta_a(x) = [x_1, x_2, x_1^2, x_1 x_2, x_2^2]^T$ , respectively.

#### 4.1. Case 1

The first case is intended to show that the proposed method can obtain a near-optimal controller for nonlinear system without constraints and uncertainties. For (48), the constraint bounds are chosen large enough to make sure that the states and control do not exceed these bounds. Thus, the following performance index is given

$$J = \int_0^{\infty} (x^T S x + Q_s + U) d\tau, \quad (49)$$

where  $Q_s = \alpha_s \sum_{i=1}^2 |x_i|^2 \ln \left( \frac{\beta^2}{\beta^2 - x_i^2} \right)$  with  $\alpha_s = 10^{-5}$ ,  $\beta = 10$  and  $U = 2 \int_0^u (\eta \tanh^{-1}(\xi/\eta))^T R d\xi$  with  $\eta = 10$ .

The simulation results throughout learning using the proposed method are shown in Figs. 1-3. Fig 1 depicts the phase portrait of state evolution and the control input, demonstrating that the system states are convergent. The critic and actor NNs weights converge to  $\hat{w}_c = [2.5560, 0.2904, 1.5331]^T$  and  $\hat{w}_a = [-0.0556, -1.0389, -0.1594, 0.1563, 0.1205]^T$  respectively, as shown in Figs. 2 and 3.

Since there is no known analytic optimal solution for (48) with (49), the hp-pseudospectral method [35], which is known as an accurate numerical method for nonlinear optimal control problems, is utilized to obtain the optimal controller.

The performance of the final controller found at the end of learning process with the proposed method is compared with the performance of the optimal controller obtained by the hp-pseudospectral method. Fig. 4 and Table 1 show the system states, control input and optimal value functions for the two controllers.

From Fig. 4, we can see that the performance of the final controller obtained by the proposed algorithm is very close to that of the hp-pseudospectral method. The state and control trajectories almost overlap all the time. The difference between the optimal value functions given by the proposed approach and the hp-pseudospectral method

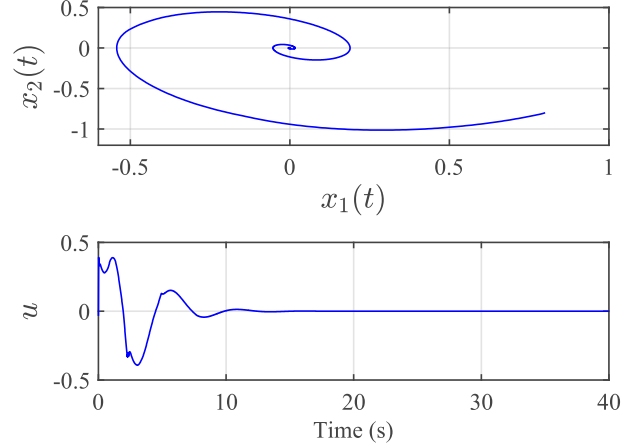


Fig. 1. Evolution of the phase plot of states (Top) and the control input (Bottom) throughout learning for Case 1.

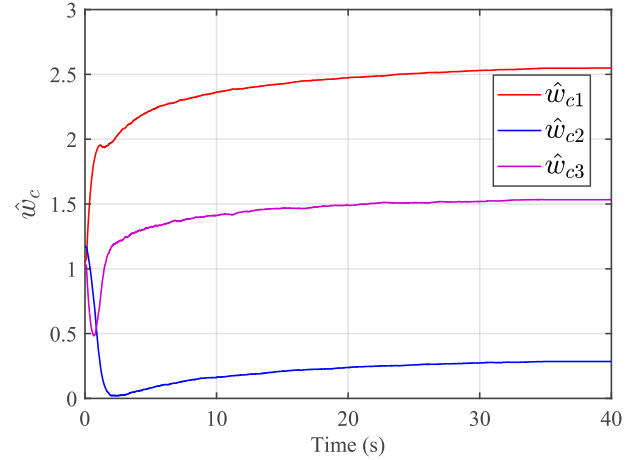


Fig. 2. Convergence of critic NN weights for Case 1.

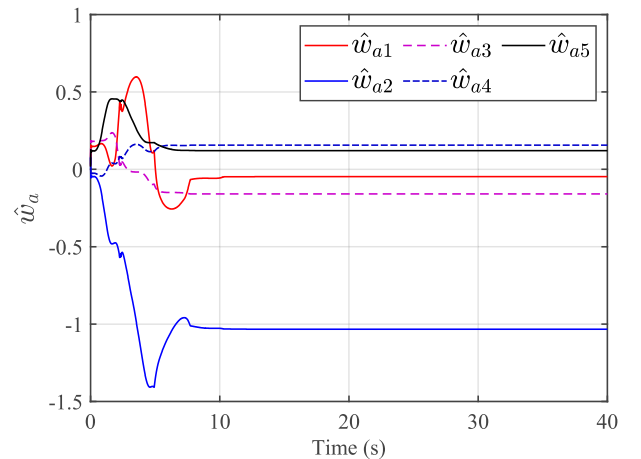


Fig. 3. Convergence of actor NN weights for Case 1.



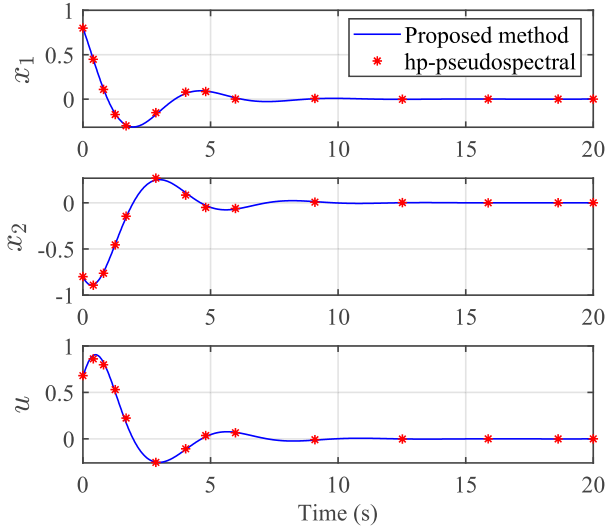


Fig. 4. System states  $x_1$ ,  $x_2$  and control  $u$  calculated by the proposed method and the hp-pseudospectral method for Case 1.

Table 1. The optimal value functions for Case 1.

Method	Optimal value function
Proposed method	2.1607
hp-pseudospectral method	2.1545

is about  $6.2 \times 10^{-3}$ , which confirms that the proposed method obtains a near-optimal controller for nonlinear system without constraints and uncertainties.

#### 4.2. Case 2

In this scenario, the effectiveness of the proposed method is illustrated for constrained nonlinear system in the presence of uncertainties. The nonlinear system with constraints and uncertainties is given by

$$\dot{x} = \begin{bmatrix} x_2 \\ -\frac{1}{j}(Mgl \sin(x_1) + f_d x_2) \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{j} \end{bmatrix} (u + d(x)), \quad (50)$$

with  $|x_1| \leq 1$ ,  $|x_2| \leq 1.5$  and  $|u| \leq 0.1$ , where the uncertainty term  $d(x) = 2x_1 \sin^2(x_2)$  and  $\|d\| \leq 2\|x\|$ . The appropriate performance index is given by

$$J = \int_0^\infty (4\|x\|^2 + x^T Sx + Q_s + U) d\tau, \quad (51)$$

where  $Q_s = \alpha_s \sum_{i=1}^2 |x_i|^2 \ln\left(\frac{\beta_i^2}{\beta_i^2 - x_i^2}\right)$  with  $\alpha_s = 10^{-5}$ ,  $\beta_1 = 1$ ,  $\beta_2 = 1.5$  and  $U = 2 \int_0^u (\eta \tanh^{-1}(\xi/\eta))^T R d\xi$  with  $\eta = 0.1$ .

Figs. 5-7 show the simulation results throughout learning using the proposed method. Fig. 5 presents the phase portrait of states and the control. The evolution of the

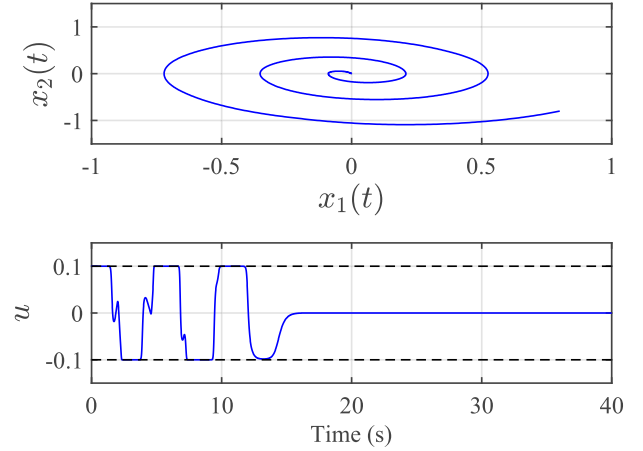


Fig. 5. Evolution of the phase plot of states (Top) and the control input (Bottom) using the proposed method throughout learning for Case 2.

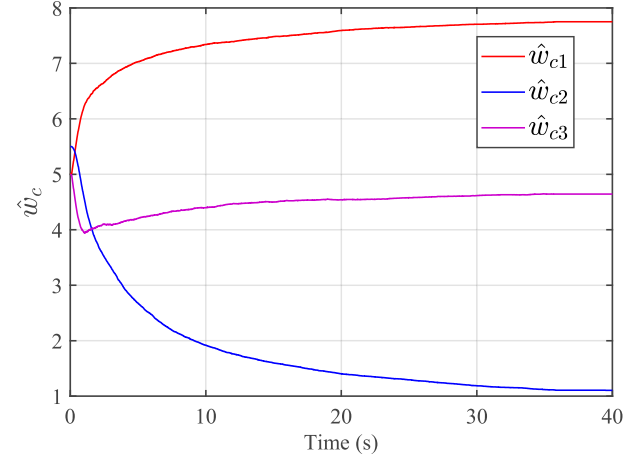


Fig. 6. Convergence of critic NN weights using the proposed method for Case 2.

states and control does not exceed the boundary of prescribed region, and the states converge to the origin. Figs. 6 and 7 indicate that the critic NN weights converge to  $\hat{w}_c = [7.7512, 1.1064, 4.6441]^T$  and the convergence of the actor NN weights are denoted by  $\hat{w}_a = [-0.5200, -4.5739, -0.0686, 0.3075, 0.2913]^T$ . The initial control (i.e.,  $u(0) = 0.1$ ) cannot bring the states to zero. The proposed method ensures that the system is stable throughout learning, which verifies the effect of the second term in (21).

The update laws (20) and (21) without the second terms (classical update laws), as described in [25], are implemented to compare with our results. Only the probing noise  $n_e$  is considered for the first second in simulation. The convergence of the critic and actor weights using the classical update laws is shown in Fig. 8. It is obvious that the critic and actor weights get stuck in a local minimum

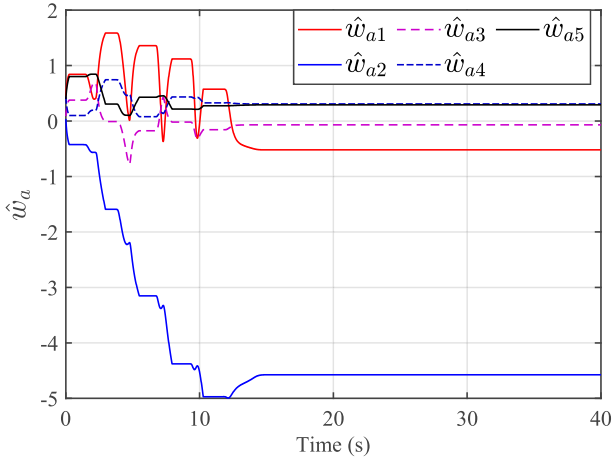


Fig. 7. Convergence of actor NN weights using the proposed method for Case 2.

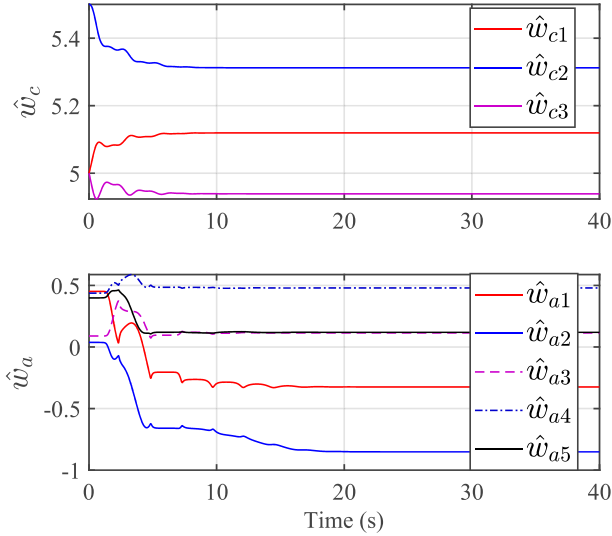


Fig. 8. Convergence of critic (Top) and actor (Bottom) NNs weights using the classical update laws for Case 2.

using the classical update laws. This happens since the PE condition is violated. The proposed method with the historical data and the stabilization term avoids this issue.

Furthermore, for system (50), the performance of the final robust near-optimal controller found at the end of learning process with the proposed method is compared with the performance of the final controller obtained by the classical update laws. Figs. 9 and 10 present the system states and control inputs under the two final controllers. It is clear that the performance of the final robust near-optimal controller obtained by the proposed method is superior to that of the controller acquired using the classical update laws, since both the states and control for the proposed method converge to zero faster under uncertainties. As a result, the developed robust near-optimal control

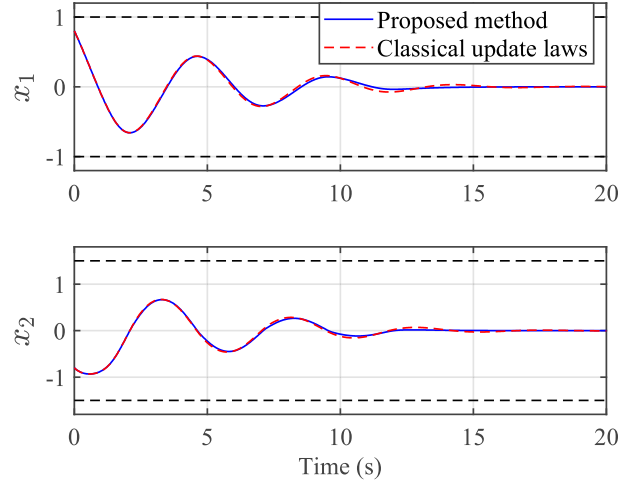


Fig. 9. System states  $x_1$  and  $x_2$  calculated by the proposed method and the classical update laws for Case 2.

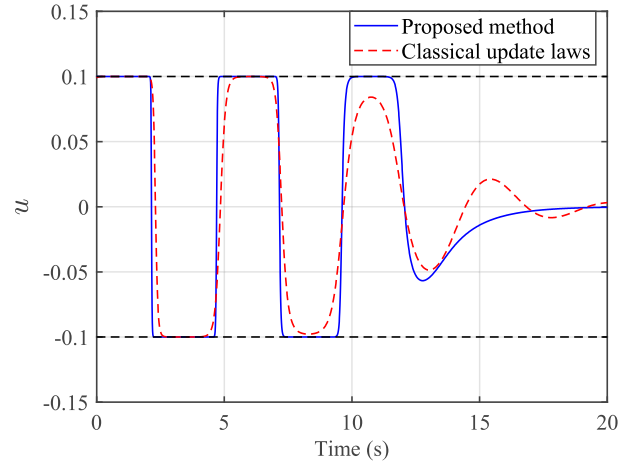


Fig. 10. Control input  $u$  calculated by the proposed method and the classical update laws for Case 2.

method is effective for nonlinear systems subject to state constraints, input saturation and uncertainties.

## 5. CONCLUSION

A robust near-optimal controller for a class of partially-unknown nonlinear systems with both constraints and uncertainties is proposed in this paper. The state constraints and input saturation are handled by a barrier function and a non-quadratic penalty term. To approximate the optimal value function and control policy online, an actor-critic architecture is adopted. The novel gain-adjustable update law is developed, in which the critic weights are updated with the historical and current data to relax the PE condition, and the actor weights with a stabilization term ensure the system stability throughout learning without an initial admissible control. The stability and convergence analy-

sis of the proposed method is conducted. Numerical studies verify the effectiveness and feasibility of the proposed control scheme.

## REFERENCES

- [1] G. A. Rovithakis, "Robust redesign of a neural network controller in the presence of unmodeled dynamics," *IEEE Transactions on Neural Networks*, vol. 15, no. 6, pp. 1482-1490, November 2004.
- [2] J. L. Chang and T. C. Wu, "Disturbance observer based output feedback controller design for systems with mismatched disturbance," *International Journal of Control, Automation, and Systems*, vol. 16, no. 4, pp. 1775-1782, August 2018.
- [3] F. Lin and R. D. Brandt, "An optimal control approach to robust control of robot manipulators," *IEEE Transactions on Robotics and Automation*, vol. 14, no. 1, pp. 69-77, February 1998.
- [4] D. Wang, D. R. Liu, and H. L. Li, "Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 627-632, April 2014.
- [5] D. Wang, D. R. Liu, H. L. Li, and H. W. Ma, "Neural-network-based robust optimal control design for a class of uncertain nonlinear systems via adaptive dynamic programming," *Information Sciences*, vol. 282, pp. 167-179, October 2014.
- [6] D. Wang, D. R. Liu, Q. C. Zhang, and D. B. Zhao, "Data-based adaptive critic designs for nonlinear robust optimal control with uncertain dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 11, pp. 1544-1555, November 2016.
- [7] H. G. Zhang, K. Zhang, G. Y. Xiao, and H. Jiang, "Robust optimal control scheme for unknown constrained-input nonlinear systems via a plug-n-play event-sampled critic-only algorithm," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 9, pp. 3169-3180, September 2020.
- [8] Y. L. Yang, Y. X. Yin, W. He, K. G. Vamvoudakis, H. Modares, and D. C. Wunsch, "Safety-aware reinforcement learning framework with an actor-critic-barrier structure," *Proc. of the 2019 American Control Conference*, pp. 2352-2358, August 2019.
- [9] J. Lee and R. S. Sutton, "Policy iterations for reinforcement learning problems in continuous time and space—Fundamental theory and methods," *Automatica*, vol. 126, 109421, April 2021.
- [10] Q. L. Wei, L. Y. Han, and T. L. Zhang, "Spiking adaptive dynamic programming based on poisson process for discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 1846-1856, May 2022.
- [11] K. G. Vamvoudakis and F. L. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878-888, May 2010.
- [12] Y. L. Yang, D. W. Ding, H. Y. Xiong, Y. X. Yin, and D. C. Wunsch, "Online barrier-actor-critic learning for  $H_\infty$  control with full-state constraints and input saturation," *Journal of the Franklin Institute*, vol. 357, no. 6, pp. 3316-3344, April 2020.
- [13] H. Y. Dong, X. W. Zhao, and H. Y. Yang, "Reinforcement learning-based approximate optimal control for attitude re-orientation under state constraints," *IEEE Transactions on Control Systems Technology*, vol. 29, no. 4, pp. 1664-1673, July 2021.
- [14] X. X. Guo, W. S. Yan, and R. X. Cui, "Reinforcement learning-based nearly optimal control for constrained-input partially unknown systems using differentiator," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4713-4725, November 2020.
- [15] D. R. Liu, X. Yang, D. Wang, and Q. L. Wei, "Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints," *IEEE Transactions on Cybernetics*, vol. 45, no. 7, pp. 1372-1385, July 2015.
- [16] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237-246, April 2009.
- [17] 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, July 2014.
- [18] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 916-932, May 2015.
- [19] R. Z. Song, F. L. Lewis, Q. Wei, and H. G. Zhang, "Off-policy actor-critic structure for optimal control of unknown systems with disturbances," *IEEE Transactions on Cybernetics*, vol. 46, no. 5, pp. 1041-1050, May 2016.
- [20] K. G. Vamvoudakis, D. Vrabie, and F. L. Lewis, "Online adaptive algorithm for optimal control with integral reinforcement learning," *International Journal of Robust and Nonlinear Control*, vol. 24, no. 17, pp. 2686-2710, November 2014.
- [21] Q. L. Wei, H. Y. Li, X. Yang, and H. B. He, "Continuous-time distributed policy iteration for multicontroller nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 51, no. 5, pp. 2372-2383, May 2021.
- [22] C. Liu, H. G. Zhang, H. Ren, and Y. L. Liang, "An analysis of IRL-based optimal tracking control of unknown nonlinear systems with constrained input," *Neural Processing Letters*, vol. 50, no. 3, pp. 2681-2700, December 2019.
- [23] Z. L. Zhang, R. Z. Song, and M. Cao, "Synchronous optimal control method for nonlinear systems with saturating actuators and unknown dynamics using off-policy integral reinforcement learning," *Neurocomputing*, vol. 356, pp. 162-169, September 2019.

- [24] X. Yang, D. R. Liu, B. Luo, and C. Li, "Data-based robust adaptive control for a class of unknown nonlinear constrained-input systems via integral reinforcement learning," *Information Sciences*, vol. 369, pp. 731-747, November 2016.
- [25] F. A. Yaghmaie and D. J. Braun, "Reinforcement learning for a class of continuous-time input constrained optimal control problems," *Automatica*, vol. 99, pp. 221-227, January 2019.
- [26] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193-202, January 2014.
- [27] A. Mishra and S. Ghosh, "Simultaneous identification and optimal tracking control of unknown continuous-time systems with actuator constraints," *International Journal of Control*, pp. 1-19, March 2021.
- [28] A. Mishra and S. Ghosh, " $H_\infty$  tracking control via variable gain gradient descent-based integral reinforcement learning for unknown continuous time non-linear system," *IET Control Theory & Applications*, vol. 14, no. 20, pp. 3476-3489, December 2020.
- [29] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 10, pp. 1513-1525, October 2013.
- [30] B. A. Finlayson, *The Method of Weighted Residuals and Variational Principles*, Society for Industrial and Applied Mathematics, 2013.
- [31] K. G. Vamvoudakis, M. F. Miranda, and J. P. Hespanha, "Asymptotically stable adaptive-optimal control algorithm with saturating actuators and relaxed persistence of excitation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2386-2398, November 2016.
- [32] X. Yang and Q. L. Wei, "Adaptive critic learning for constrained optimal event-triggered control with discounted cost," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 91-104, January 2021.
- [33] F. W. Lewis, S. Jagannathan, and A. Yesildirak, *Neural Network Control of Robot Manipulators and Non-linear Systems*, CRC press, 2020.
- [34] B. Zhao, D. R. Liu, and C. M. Luo, "Reinforcement learning-based optimal stabilization for unknown nonlinear systems subject to inputs with uncertain constraints," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 10, pp. 4330-4340, October 2020.
- [35] C. L. Darby, *hp-Pseudospectral Method for Solving Continuous-time Nonlinear Optimal Control Problems*, University of Florida, 2011.



**Yu-Qing Qiu** received his B.S. degree in Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2017, an M.S. degree in Northwestern Polytechnical University, Xi'an, China, in 2020, where he is currently pursuing a Ph.D. degree in navigation, guidance, and control. His research interests include robust control, optimal control and their applications

to aerial vehicles.



**Yan Li** received his B.S. and M.S. degrees in automatic control from Northwestern Polytechnical University, Xi'an, China, in 1995 and 1998, respectively, and a Ph.D. degree in Nanyang Technological University, Singapore in 2001. He is currently a Professor with the Department of Navigation, Guidance and Control, Northwestern Polytechnical University. His current research interests include robust control, optimal control, fault-tolerant control, and flight control theory.



**Zhong Wang** received his B.S., M.S., and Ph.D. degrees in automatic control from Northwestern Polytechnical University, Xi'an, China, in 2013, 2016, and 2021, respectively. His current research interests include flight control, nonlinear control, and Bayesian inference.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.