Published in IET Control Theory and Applications
Received on 25th May 2013
Revised on 24th July 2013
Accepted on 2nd August 2013
doi: 10.1049/iet-cta.2013.0472

ISSN 1751-8644

Neural-network-based online optimal control for uncertain non-linear continuous-time systems with control constraints

Xiong Yang, Derong Liu, Yuzhu Huang

The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
E-mail: derong.liu@ia.ac.cn

Abstract: In this study, an online adaptive optimal control scheme is developed for solving the infinite-horizon optimal control problem of uncertain non-linear continuous-time systems with the control policy having saturation constraints. A novel identifier-critic architecture is presented to approximate the Hamilton–Jacobi–Bellman equation using two neural networks (NNs): an identifier NN is used to estimate the uncertain system dynamics and a critic NN is utilised to derive the optimal control instead of typical action–critic dual networks employed in reinforcement learning. Based on the developed architecture, the identifier NN and the critic NN are tuned simultaneously. Meanwhile, unlike initial stabilising control indispensable in policy iteration, there is no special requirement imposed on the initial control. Moreover, by using Lyapunov's direct method, the weights of the identifier NN and the critic NN are guaranteed to be uniformly ultimately bounded, while keeping the closed-loop system stable. Finally, an example is provided to demonstrate the effectiveness of the present approach.

1 Introduction

Saturation, backlash and dead zone are common features in real engineering applications. During the past several decades, controller design of non-linear systems with saturating actuators had drawn intensive attention in the control community [1–4]. The objective of designing a controller is generally to develop stable control schemes for non-linear systems [5, 6]. Nevertheless, stability is only a bare minimum requirement in a system design. The control scheme is often required to guarantee the stability of the closed-loop system, while keeping the prescribed cost function as small as possible. Or rather, optimality is more preferred for controller design of non-linear dynamic systems than stability alone.

From a mathematical point of view, the solution of the optimal control problem can be obtained by solving the Hamilton–Jacobi–Bellman (HJB) equation, which guarantees the sufficient condition for existence of optimality [7]. For linear dynamic systems and quadratic costs, the HJB equation reduces to the Riccati equation, which can accurately be solved by analytical or numerical methods [8]. In case of non-linear systems, the HJB equation is actually a non-linear partial differential equation that is intractable to be solved. Although dynamic programming (DP) provides a way to cope with optimal control problems, a serious drawback about it is that the computation is untenable to be run with the increasing dimension of the non-linear system, which is referred to the well-known 'curse of dimensionality' [9]. In addition, the backward direction of the search obviously prohibits the use of DP in the real-time control. In order to overcome the difficulties for the use of DP, adaptive dynamic programming (ADP) algorithms were developed by Werbos [10–12]. The ADP approaches employ NNs to derive optimal control forward-in-time. After that, large amounts of ADP methods were developed [13–17]. However, most of the ADP algorithms are implemented either by an offline process via iterative schemes or need a priori knowledge of system dynamics. In light of the exact knowledge of non-linear dynamic systems generally unavailable, it is intractable to implement these algorithms. Consequently, reinforcement learning (RL) methods are introduced. RL is a class of approaches used in machine learning to methodically revise the actions of an agent based on responses from its environment [18]. A typical structure of implementing RL algorithm is the actor–critic architecture, where the actor performs actions by interacting with its surroundings, and the critic evaluates actions and offers feedback information to the actor, leading to the improvement in performance of the subsequent actor [19].

Up to now, many researchers have studied optimal control problems for non-linear systems based on RL methods [19–22]. Abu-Khalaf and Lewis [20] presented an offline algorithm based on RL to solve the HJB equation of optimal control of fcontinuous-time (CT) non-linear systems with input having saturation constraints. By using the algorithm, the actor and the critic were sequently tuned and the solution of the HJB equation was successively approximated. In order to derive online optimal control for CT non-linear systems, Vamvoudakis and Lewis [21] proposed a novel

algorithm based on RL to synchronously tune the critic and the actor. However, the exact knowledge of CT non-linear systems is indispensable in [20, 21]. After that, Bhasin *et al.* [22] presented a projection algorithm to derive the optimal control of uncertain non-linear CT systems. Based on the algorithm, the requirement of the prior knowledge of non-linear dynamics was relaxed. Meanwhile, the actor, the critic, and the identifier were all simultaneously tuned. Nevertheless, a shortcoming of the method is that the use of the projection algorithm demands the selection of a predefined convex set so as to make the target NN weights remain in the set, which is a challenge. In addition, unfortunately, the algorithm proposed in [20, 21] and [22], all required the initial stabilising control. There is no a general approach developed to derive such a control. From mathematical perspectives, the initial stabilising policy is actually a suboptimal control. The suboptimal control is generally difficult to obtain since it is often impossible to give analytical solutions for partial differential equations. Accordingly, the initial stabilising control is a rather restrictive condition. Recently, Dierks and Jagannathan [23] relaxed the requirement of initial stabilising control using a single online approximator-based framework. However, the exact knowledge of the system dynamics is still required. In addition, the saturation of the control input is not taken into consideration. As mentioned before, plenty of real-world applications of feedback control involve control actuators with amplitude limitations. The control design techniques that ignore the actuators' limitation may give birth to undesirable transient response and cause system instability.

Motivated by the above work, in this paper, an online adaptive optimal control scheme is developed for solving the infinite-horizon optimal control problem of uncertain non-linear CT systems with the control policy having saturation constraints. A novel identifier–critic architecture is presented to approximate the HJB equation using two neural networks NNs: an identifer NN is used to estimate the uncertain system dynamics and a critic NN is utilised to derive the optimal control instead of typical action–critic dual networks employed in RL. Based on the developed architecture, the identifier NN and the critic NN are tuned simultaneously. Meanwhile, unlike initial stabilising control indispensable in policy iteration, there is no special requirement imposed on the initial control. Moreover, by using Lyapunov's direct method, the weights of the identifier NN and the critic NN are guaranteed to be uniformly ultimately bounded (UUB), while keeping the closed-loop system stable.

The main contributions of this paper include the following:

1. To the best of our knowledge, it is the first time that an identifier–critic architecture is developed to derive optimal control of uncertain non-linear CT systems with input constrains. Based on the architecture, the identifier NN and the critic NN are tuned simultaneously.
2. Compared with [21, 22], a clear advantage of the developed control scheme in this paper is that no initial stabilising control is required. Meanwhile, the optimal control is derived by using only one critic network, instead of the action–critic dual networks. In addition, another obvious advantage of this paper as compared with [21] lies in that the knowledge of the internal system dynamics is not required, that is, $f(x)$ is unknown in system (1).
3. This paper extends the work of [23] to derive optimal control for uncertain non-linear CT systems with saturating actuator. The stability analysis of non-linear CT systems with saturating actuator is more difficult than those regardless of saturating actuator. By using a novel modified weight tuning law for the critic NN and employing Taylor series, the difficulties of stability analysis is successfully overcome.
4. Unlike most identifier using a linear-in-parameter (LP) NN [24, 25], the presented identifier utilises a non-linear-in-parameters (NLP) NN. NLP NN is generally considered as more powerful than LP NN used to estimate the system dynamics [26].

The rest of this paper is organised as follows. The problem statement and preliminaries are presented in Section 2. Identifier design is proposed in Section 3. Online optimal neuro-controller design with constrained controls is developed in Section 4. Stability analysis and performance of the closed-loop system is indicated in Section 5. Simulation results are provided to show the effectiveness of the proposed control scheme in Section 6. Finally, several concluding remarks are given in Section 7.

For convenience, notations are listed here, which will be used throughout the paper.

- $\mathbb{R}$ denotes the real number, $\mathbb{R}^m$ and $\mathbb{R}^{m\times n}$ denote the real m-vector and the real $m \times n$ matrix, respectively. I_n represents $n \times n$ identity matrix. If there is no special explanation, T is a transposition symbol.
- $\|\cdot\|$ stands for any suitable norm. When z is a vector, $\|z\|$ denotes the Euclidean norm of z. When z is a matrix, and $\|z\|$ denotes Frobenius norm of z.
- Ω is a compact subset of $\mathbb{R}^n$ and $\mathfrak{A}$ is a subset of $\mathbb{R}^m$, $\mathrm{C}^m(\Omega) = \{f^{(m)} \in \mathrm{C} | f : \Omega \to \mathbb{R}\}$.

2 Problem statement and preliminaries

For purpose of the present paper, we consider the non-linear CT system given in the form

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

with state $x(t) \in \Omega \subseteq \mathbb{R}^n$ and control $u(t) \in \mathfrak{A} \subseteq \mathbb{R}^m$. $\mathfrak{A} = \{u = [u_1, u_2, \ldots, u_m]^{\mathsf{T}} : |u_i| \le \alpha_i, i = 1, \ldots, m\}$, where α_i is the saturating bound for the ith actuator. $f(x) \in \mathbb{R}^n$ is an unknown non-linear function, $g(x) \in \mathbb{R}^{n\times m}$ is a matrix of non-linear functions. It is assumed that $f(0) = 0$ and $f(x) + g(x)u$ is Lipschitz continuous on Ω containing the origin, such that the solution of system (1) is unique for any given initial state $x_0 \in \Omega$ and control $u \in \mathfrak{A}$. System (1) is stabilisable in the sense that there exists a continuous control $u \in \mathfrak{A}$ that asymptotically stabilises the system on Ω.

The value function for system (1) is generally described by

$$V(x(t)) = \int_t^\infty r(x(s), u(s))\,\mathrm{d}s \tag{2}$$

where $r(x, u) = Q(x) + \mathcal{W}(u)$, and $Q(x)$ is continuously differentiable and positive definite, that is, $\forall x \neq 0, Q(x) > 0$ and $x = 0 \Leftrightarrow Q(x) = 0$, and $\mathcal{W}(u)$ is positive definite. In order to confront bounded controls in system (1) and inspired by the work of [20, 27], we define $\mathcal{W}(u)$ as

$$\mathcal{W}(u) = 2\lambda \int_0^u \boldsymbol{\psi}^{-\mathsf{T}}(\upsilon/\lambda)\,\mathrm{d}\upsilon$$

where $\boldsymbol{\psi}^{-1}(\upsilon/\lambda) = [\psi^{-1}(\upsilon_1/\lambda), \psi^{-1}(\upsilon_2/\lambda), \ldots, \psi^{-1}(\upsilon_m/\lambda)]^{\mathsf{T}}$, λ is a positive constant, $\boldsymbol{\psi} \in \mathbb{R}^m$, $\boldsymbol{\psi}^{-\mathsf{T}}$ denotes $(\boldsymbol{\psi}^{-1})^{\mathsf{T}}$,

doi: 10.1049/iet-cta.2013.0472

and $\psi(\cdot)$ is a bounded one-to-one function satisfying $|\psi(\cdot)| \leq 1$ and belonging to $C^p(p \geq 1)$ and $L_2(\Omega)$. Meanwhile, $\psi(\cdot)$ is a monotonic odd function with its first derivative bounded. It is significant to state that $\mathcal{W}(u)$ is positive definite since $\psi^{-1}(\cdot)$ is a monotonic odd function. Without loss of generality, in this paper, we choose $\boldsymbol{\psi}(\cdot) = \tanh(\cdot)$.

Definition 1: (UUB [28]) The equilibrium point x_e of system (1) is said to be UUB, if there exist positive constants b and c, independent of $t_0 \geq 0$, and for every $a \in (0, c)$, there is $T = T(a, b) > 0$, independent of t_0, such that

$$\|x(t_0) - x_e\| \leq a \Rightarrow \|x(t) - x_e\| \leq b, \quad \forall\ t \geq t_0 + T$$

Objective of control: The control objective is to derive an online adaptive control not only stabilises system (1) but also minimises the value function (2), while ensuring that all the signals involved in the closed-loop system are UUB.

Prior to continuing our discussion, we present the following required assumption.

Assumption 1: The control matrix $g(x)$ is known and bounded, that is, there exist positive constants g_m and $g_M (g_m < g_M)$, such that $g_m \leq \|g(x)\| \leq g_M$, for $\forall\ x \in \Omega$.

3 Identifier design

In control engineering, NNs are considered as powerful tools for approximating non-linear functions owing to their properties of non-linearity, adaptivity, self-learning and fault tolerance. In this section, a single-hidden layer feedforward NN is applied to approximate $\mathcal{F}(x) \in C^n(\Omega)$ ($\mathcal{F}(x)$ is a non-linear function to be detailed subsequently) as follows [29]

$$\mathcal{F}(x) = W_1^{\mathsf{T}} \sigma(V_1^{\mathsf{T}} x) + \varepsilon_1(x) \tag{3}$$

where $\sigma(\cdot) \in \mathbb{R}^{N_1}$ is the activation function, $\varepsilon_1(x) \in \mathbb{R}^n$ is the NN function reconstruction error, $V_1 \in \mathbb{R}^{n \times N_1}$ and $W_1 \in \mathbb{R}^{N_1 \times n}$ are the weights for the input layer to the hidden layer and the hidden layer to the output layer, respectively. The number of the hidden layer nodes is denoted as N_1. In general, activation functions for $\sigma(\cdot)$ are bounded, measurable, non-decreasing functions from the real numbers onto $[-1, 1]$, which include, for instance, hyperbolic tangent function $\sigma(x) = (e^x - e^{-x})/(e^x + e^{-x})$ etc. Without loss of generality, in this paper, we employ hyperbolic tangent function $\sigma(x)$ as the activation function.

Since the dynamics of system (1) is uncertain, we need to identify the system for deriving the optimal control. From plant (1), we have that

$$\begin{aligned} \dot{x}(t) &= f(x) + g(x)u \\ &= Ax + \mathcal{F}(x) + g(x)u \end{aligned} \tag{4}$$

where $\mathcal{F}(x) = f(x) - Ax$, and $A \in \mathbb{R}^{n \times n}$ is a known constant matrix. By using (3), (4) can be developed by

$$\dot{x}(t) = Ax + W_1^{\mathsf{T}} \sigma(V_1^{\mathsf{T}} x) + g(x)u + \varepsilon_1(x) \tag{5}$$

The NN identifier approximates system (1) as

$$\dot{\hat{x}}(t) = A\hat{x} + \hat{W}_1^{\mathsf{T}} \sigma(\hat{V}_1^{\mathsf{T}} \hat{x}) + g(\hat{x})u + \nu(t) \tag{6}$$

where $\hat{x}(t) \in \mathbb{R}^n$ is the identifier NN state, $\hat{W}_1 \in \mathbb{R}^{N_1 \times n}$, $\hat{V}_1 \in \mathbb{R}^{n \times N_1}$ are weight estimates, and $\nu(t) \in \mathbb{R}^n$ is the robust feedback term defined as $\nu(t) = \chi\tilde{x}$ with the identification error $\tilde{x}(t) \triangleq x(t) - \hat{x}(t)$ and the design matrix $\chi \in \mathbb{R}^{n \times n}$ selected such that $A - \chi$ is a Hurwitz matrix.

By using (5) and (6), the identification error dynamics is given by

$$\dot{\tilde{x}}(t) = (A - \chi)\tilde{x}(t) + \tilde{W}_1^{\mathsf{T}} \sigma(\hat{V}_1^{\mathsf{T}} \hat{x}) + \delta(x) \tag{7}$$

where $\tilde{W}_1 = W_1 - \hat{W}_1, \delta(x) = W_1^{\mathsf{T}}[\sigma(V_1{}^{\mathsf{T}} x) - \sigma(\hat{V}_1^{\mathsf{T}} \hat{x})] + [g(x) - g(\hat{x})]u + \varepsilon_1(x)$

Before showing the stability of the identification error $\tilde{x}(t)$, we need to present some mild assumptions and facts. It is worth pointing out that these assumptions are common techniques, which have been used in [7, 22, 25, 30].

Assumption 2: The NN weights W_1 and V_1 are bounded over the compact set Ω by known positive constants W_{M_1} and V_{M_1}, respectively. That is

$$\|W_1\| \leq W_{M_1}, \quad \|V_1\| \leq V_{M_1}$$

Assumption 3: The NN function reconstruction error $\varepsilon_1(x)$ is bounded over the compact set Ω as $\|\varepsilon_1(x)\| \leq \varepsilon_{M_1}$, where ε_{M_1} is a known positive constant.

Fact 1: The NN activation function is bounded by a known positive constant over the compact set Ω, that is, there exists $\sigma_M > 0$, such that $\|\sigma(x)\| \leq \sigma_M$, for $\forall\ x \in \Omega$.

Fact 2: Since $A - \chi$ is a Hurwitz matrix, there exists a unique positive-definite symmetric matrix $P \in \mathbb{R}^{n \times n}$ satisfying the Lyapunov equation

$$(A - \chi)^{\mathsf{T}} P + P(A - \chi) = -\alpha I_n$$

where $\alpha > 0$ is a design parameter.

Theorem 1: Let Assumptions 1–3 hold, if NN weight estimates $\hat{W}_1$ and $\hat{W}_2$ are updated as

$$\dot{\hat{W}}_1 = -l_1 \sigma(\hat{V}_1^{\mathsf{T}} \hat{x})\tilde{x}^{\mathsf{T}}(A - \chi)^{-1} - \kappa_1 \|\tilde{x}\| \hat{W}_1 \tag{8}$$

$$\dot{\hat{V}}_1 = -l_2 \mathrm{sgn}(\hat{x})\tilde{x}^{\mathsf{T}}(A - \chi)^{-1} \hat{W}_1^{\mathsf{T}} \left(I_{N_1} - \Phi(\hat{V}_1^{\mathsf{T}} \hat{x})\right) - \kappa_2 \|\tilde{x}\| \hat{V}_1 \tag{9}$$

with design parameters $l_i > 0$ and $\kappa_i > 0$ $(i = 1, 2)$, $\Phi(\hat{V}_1^{\mathsf{T}} \hat{x}) = \mathrm{diag}\{\sigma_j^2(V_{1j}^{\mathsf{T}} \hat{x})\}$ $(j = 1, \ldots, N_1)$, and $\mathrm{sgn}(\hat{x}) = [\mathrm{sgn}(\hat{x}_1), \ldots, \mathrm{sgn}(\hat{x}_n)]^{\mathsf{T}}$, where $\mathrm{sgn}(\hat{x}_k)$ $(k = 1, \ldots, n)$ is a sign function with respect to $\hat{x}_k$ [32]. Then, the NN identifier developed in (6) can ensure that the identification error $\tilde{x}(t)$ converges to the small compact set

$$\Omega_{\tilde{x}} = \{\tilde{x} : \|\tilde{x}\| \leq 2\mathfrak{B}/\alpha\} \tag{10}$$

where $\mathfrak{B}$ is defined as in (16). In addition, the error dynamics of NN weight estimates $\tilde{W}_1 = W_1 - \hat{W}_1$ and $\tilde{V}_1 = V_1 - \hat{V}_1$ are all guaranteed to be UUB.

Proof: Consider the Lyapunov function candidate

$$J(t) = J_1(t) + J_2(t) \tag{11}$$

where

$$J_1(t)=\frac{1}{2}\tilde{x}^{\mathrm{T}}P\tilde{x}\quad J_2(t)=\frac{1}{2}\mathrm{tr}\left(\tilde{W}_1^{\mathrm{T}}l_1^{-1}\tilde{W}_1\right)+\frac{1}{2}\mathrm{tr}\left(\tilde{V}_1^{\mathrm{T}}l_2^{-1}\tilde{V}_1\right)$$

Taking the time derivative of $J_1(t)$ in (11) and using Facts 1–2, we have that

$$\begin{aligned}\dot{J}_1(t)&=-\frac{\alpha}{2}\tilde{x}^{\mathrm{T}}\tilde{x}+\tilde{x}^{\mathrm{T}}P\left[\tilde{W}_1^{\mathrm{T}}\sigma(\hat{V}_1^{\mathrm{T}}\hat{x})+\delta(x)\right]\\&\le-\frac{\alpha}{2}\|\tilde{x}\|^2+\|\tilde{x}\|\|P\|\left(\|\tilde{W}_1\|\sigma_M+\delta_M\right)\end{aligned}\tag{12}$$

where δ_M is the upper bound of $\delta(x)$, that is, $\|\delta(x)\|\le\delta_M$. Actually, by Assumptions 1–3 and Fact 1, one can easily draw the conclusion that $\delta(x)$ in (7) is a bounded function since the control is constrained.

On the other hand, taking the time derivative of $J_2(t)$ in (11) and using weights update laws (8) and (9), we obtain that

$$\begin{aligned}\dot{J}_2(t)=&\,\mathrm{tr}\left\{\tilde{W}_1^{\mathrm{T}}\sigma(\hat{V}_1^{\mathrm{T}}\hat{x})\tilde{x}^{\mathrm{T}}(A-\chi)^{-1}+\frac{\kappa_1}{l_1}\|\tilde{x}\|\tilde{W}_1^{\mathrm{T}}(W_1-\tilde{W}_1)\right\}\\&+\mathrm{tr}\left\{\tilde{V}_1^{\mathrm{T}}\mathrm{sgn}(\hat{x})\tilde{x}^{\mathrm{T}}(A-\chi)^{-1}(W_1-\tilde{W}_1)^{\mathrm{T}}\right.\\&\left.\times\left(I_{N_1}-\Phi(\hat{V}_1^{\mathrm{T}}\hat{x})\right)+\frac{\kappa_2}{l_2}\|\tilde{x}\|\tilde{V}_1^{\mathrm{T}}(V_1-\tilde{V}_1)\right\}\end{aligned}\tag{13}$$

Observing that $\mathrm{tr}(XY)=\mathrm{tr}(YX)=YX$, for $\forall X\in\mathbb{R}^{n\times1}$, $Y\in\mathbb{R}^{1\times n}$ and $\mathrm{tr}[\tilde{Z}^{\mathrm{T}}(Z-\tilde{Z})]\le\|\tilde{Z}\|\|Z\|-\|\tilde{Z}\|^2$, for $\forall\ Z,\tilde{Z}\in\mathbb{R}^{m\times n}$, we can rewrite (13) as

$$\begin{aligned}\dot{J}_2(t)=&\,\tilde{x}^{\mathrm{T}}(A-\chi)^{-1}\tilde{W}_1^{\mathrm{T}}\sigma(\hat{V}_1^{\mathrm{T}}\hat{x})+\frac{\kappa_1}{l_1}\|\tilde{x}\|\mathrm{tr}\left(\tilde{W}_1^{\mathrm{T}}(W_1-\tilde{W}_1)\right)\\&+\tilde{x}^{\mathrm{T}}(A-\chi)^{-1}(W_1-\tilde{W}_1)^{\mathrm{T}}\left(I_{N_1}-\Phi(\hat{V}_1^{\mathrm{T}}\hat{x})\right)\tilde{V}_1^{\mathrm{T}}\mathrm{sgn}(\hat{x})\\&+\frac{\kappa_2}{l_2}\|\tilde{x}\|\mathrm{tr}\left(\tilde{V}_1^{\mathrm{T}}(V_1-\tilde{V}_1)\right)\\\le&\,\beta\sigma_M\|\tilde{x}\|\|\tilde{W}_1\|+\frac{\kappa_1}{l_1}\|\tilde{x}\|\left(W_{M_1}\|\tilde{W}_1\|-\|\tilde{W}_1\|^2\right)\\&+\beta\left\|I_{N_1}-\Phi(\hat{V}_1^{\mathrm{T}}\hat{x})\right\|\|\tilde{x}\|\left(W_{M_1}+\|\tilde{W}_1\|\right)\|\tilde{V}_1\|\\&+\frac{\kappa_2}{l_2}\|\tilde{x}\|\left(V_{M_1}\|\tilde{V}_1\|-\|\tilde{V}_1\|^2\right)\end{aligned}\tag{14}$$

where $\beta=\|(A-\chi)^{-1}\|$.

Combining (12) with (14) and noting that $\|I_{N_1}-\Phi(\hat{V}_1^{\mathrm{T}}\hat{x})\|\le1$, we obtain that

$$\begin{aligned}\dot{J}(t)=&-\frac{\alpha}{2}\|\tilde{x}\|^2+\left\{\delta_M\|P\|+\left((\|P\|+\beta)\sigma_M+\frac{\kappa_1}{l_1}W_{M_1}\right)\|\tilde{W}_1\|\right.\\&+\left(\beta W_{M_1}+\frac{\kappa_2}{l_2}V_{M_1}\right)\|\tilde{V}_1\|-\left(\frac{\kappa_1}{l_1}-\frac{\beta^2}{4}\right)\|\tilde{W}_1\|^2\\&\left.-\left(\frac{\kappa_2}{l_2}-1\right)\|\tilde{V}_1\|^2-\left(\frac{\beta}{2}\|\tilde{W}_1\|-\|\tilde{V}_1\|\right)^2\right\}\|\tilde{x}\|\\=&-\frac{\alpha}{2}\|\tilde{x}\|^2+\left\{\delta_M\|P\|+\left(\frac{\kappa_1}{l_1}-\frac{\beta^2}{4}\right)\gamma_1^2\right.\\&+\left(\frac{\kappa_2}{l_2}-1\right)\gamma_2^2-\left(\frac{\kappa_1}{l_1}-\frac{\beta^2}{4}\right)\|\tilde{W}_1+\gamma_1\|^2\\&\left.-\left(\frac{\kappa_2}{l_2}-1\right)\|\tilde{V}_1+\gamma_2\|^2-\left(\frac{\beta}{2}\|\tilde{W}_1\|-\|\tilde{V}_1\|\right)^2\right\}\|\tilde{x}\|\end{aligned}\tag{15}$$

where

$$\gamma_1=\frac{2l_1(\beta+\|P\|)\sigma_M+2\kappa_1W_{M_1}}{\beta^2l_1-4\kappa_1},\quad\gamma_2=\frac{\beta l_2W_{M_1}+\kappa_2V_{M_1}}{2(l_2-\kappa_2)}$$

Selecting $\kappa_1>\beta^2l_1/4,\kappa_2>l_2$ and from (15), we derive that

$$\begin{aligned}\dot{J}(t)\le&-\frac{\alpha}{2}\|\tilde{x}\|^2+\left\{\delta_M\|P\|+\left(\frac{\kappa_1}{l_1}-\frac{\beta^2}{4}\right)\gamma_1^2\right.\\&\left.+\left(\frac{\kappa_2}{l_2}-1\right)\gamma_2^2\right\}\|\tilde{x}\|\\=&-\left(\frac{\alpha}{2}\|\tilde{x}\|-\mathfrak{B}\right)\|\tilde{x}\|\end{aligned}\tag{16}$$

where $\mathfrak{B}=\delta_M\|P\|+\left(\frac{\kappa_1}{l_1}-\frac{\beta^2}{4}\right)\gamma_1^2+\left(\frac{\kappa_2}{l_2}-1\right)\gamma_2^2$

Therefore $\dot{J}(t)$ is negative as long as $\|\tilde{x}(t)\|>2\mathfrak{B}/\alpha$, where $\mathfrak{B}$ is defined as in (16). That is, the system identification error $\tilde{x}(t)$ converges to the compact set $\Omega_{\tilde{x}}$ defined as in (10). Meanwhile, according to the standard Lyapunov extension theorem [30], this demonstrates the uniformly ultimate boundedness of the NN weight estimates error $\tilde{W}_1$ and $\tilde{V}_1$. □

4 Online optimal neuro-controller design with constrained controls

This section consists of two subsections. In the first subsection, the HJB equation for constrained non-linear CT systems is developed. Then, an online NN-based optimal control scheme is presented.

4.1 HJB equation for constrained non-linear CT systems

Since system (1) can be approximated by (6) outside of the small compact set $\Omega_{\tilde{x}}$, we replace system (1) with (6) in the subsequent discussion. Meanwhile, system state $x(t)$ is replaced by $\hat{x}(t)$, and (6) is represented by

$$\dot{\hat{x}}(t)=h(\hat{x})+g(\hat{x})u\tag{17}$$

where $h(\hat{x})=A\hat{x}+\hat{W}_1^{\mathrm{T}}\sigma(\hat{V}_1^{\mathrm{T}}\hat{x})+\chi\tilde{x}(t)$. The value function (2) is rewritten as

$$V(\hat{x}(t))=\int_t^{\infty}\left(Q(\hat{x}(s))+2\lambda\int_0^u\tanh^{-\mathrm{T}}(\upsilon/\lambda)\,\mathrm{d}\upsilon\right)\mathrm{d}s\tag{18}$$

Definition 2: (Admissible control [32]) A control $u(\hat{x}):\mathbb{R}^n\to\mathbb{R}^m$ is said to be admissible with respect to (18) on Ω, written as $u(\hat{x})\in\mathscr{A}(\Omega)$, if $u(\hat{x})$ is continuous on Ω, $u(0)=0,u(\hat{x})$ stabilises system (17) on Ω, and $V(\hat{x}_0)$ is finite for every $\hat{x}\in\Omega$.

Given a control $u(\hat{x})\in\mathscr{A}(\Omega)$, if the associated value function $V(\hat{x})\in\mathrm{C}^1(\Omega)$, its infinitesimal version of (18) is the so-called Lyapunov equation

$$V_{\hat{x}}^{\mathrm{T}}(h(\hat{x})+g(\hat{x})u)+Q(\hat{x})+2\lambda\int_0^u\tanh^{-\mathrm{T}}(\upsilon/\lambda)\,\mathrm{d}\upsilon=0$$

where $V_{\hat{x}}\in\mathbb{R}^n$ denotes the partial derivative of $V(\hat{x})$ with respect to $\hat{x}$. Define the Hamiltonian for the control

doi: 10.1049/iet-cta.2013.0472

$u(\hat{x}) \in \mathscr{A}(\Omega)$ and the associated value function $V(\hat{x})$ by

$$H(\hat{x}, V_{\hat{x}}, u) = V_{\hat{x}}^{\mathsf{T}}(h(\hat{x}) + g(\hat{x})u) + Q(\hat{x}) + 2\lambda \int_0^u \tanh^{-\mathsf{T}}(\upsilon/\lambda)\, \mathrm{d}\upsilon$$

Then, the optimal cost $V^*(\hat{x})$ can be obtained by solving the HJB equation

$$\min_{u(\hat{x}) \in \mathscr{A}(\Omega)} H(\hat{x}, V_{\hat{x}}^*, u) = 0. \tag{19}$$

Suppose that the minimum value on the left-hand side of (19) exists and is unique. Then, the closed-form expression for constrained optimal control is derived as

$$u^*(\hat{x}) = -\lambda \tanh\left(\frac{1}{2\lambda} g^{\mathsf{T}}(\hat{x}) V_{\hat{x}}^*\right) \tag{20}$$

Substituting (20) into (19), we derive the HJB equation for constrained non-linear systems as

$$V_{\hat{x}}^{*\mathsf{T}} h(\hat{x}) - 2\lambda^2 \mathfrak{C}^{\mathsf{T}}(\hat{x}) \tanh(\mathfrak{C}(\hat{x})) + Q(\hat{x}) + 2\lambda \int_0^{-\lambda \tanh(\mathfrak{C}(\hat{x}))} \tanh^{-\mathsf{T}}(\upsilon/\lambda)\, \mathrm{d}\upsilon = 0 \tag{21}$$

where $\mathfrak{C}(\hat{x}) = \frac{1}{2\lambda} g^{\mathsf{T}}(\hat{x}) V_{\hat{x}}^*$. Observe that

$$2\lambda \int_0^{-\lambda \tanh(\mathfrak{C}(\hat{x}))} \tanh^{-\mathsf{T}}(\upsilon/\lambda)\, \mathrm{d}\upsilon = 2\lambda^2 \mathfrak{C}^{\mathsf{T}}(\hat{x}) \tanh(\mathfrak{C}(\hat{x})) + \lambda^2 \ln[1 - \tanh^2(\mathfrak{C}(\hat{x}))]$$

Then, we can rewrite (21) as

$$V_{\hat{x}}^{*\mathsf{T}} h(\hat{x}) + Q(\hat{x}) + \lambda^2 \ln[1 - \tanh^2(\mathfrak{C}(\hat{x}))] = 0 \tag{22}$$

where $\mathfrak{C}(x)$ is defined as in (21).

Nevertheless, one shall find that (22) is intractable to solve since it is actually a partial differential equation with respect to $V_{\hat{x}}^*$. In order to overcome the difficulty, an online NN-based optimal control scheme is developed. Prior to presenting the optimal control scheme, we develop the following required lemma.

Lemma 1: Consider system (17) with the associated value function $V(\hat{x})$ as in (18) and the optimal control (20), let $L_1(\hat{x})$ be a continuously differentiable radially unbounded Lyapunov candidate function such that $\dot{L}_1(\hat{x}) = \dot{L}_{1\hat{x}}(h(\hat{x}) + g(\hat{x})u) < 0$ with $\dot{L}_{1\hat{x}}$ the partial derivative of $L_1(\hat{x})$ with respect to $\hat{x}$. Meanwhile, let $\Lambda(\hat{x}) \in \mathbb{R}^{n\times n}$ be positive definite on Ω, that is, for $\forall \hat{x} \neq 0$, $\Lambda(\hat{x}) > 0$ and $\hat{x} = 0 \Leftrightarrow \Lambda(\hat{x}) = 0$, and there exist positive constants such that $\rho I_n \leq \Lambda(\hat{x}) \leq \varrho I_n$ ($\rho < \varrho$), for $\forall\, \hat{x} \in \Omega$ [30]. In addition, let $\Lambda(\hat{x}) = \infty$ when $\|\hat{x}\| = \infty$ and there exists

$$V_{\hat{x}}^{*\mathsf{T}} \Lambda(\hat{x}) L_{1\hat{x}} = Q(\hat{x}) + 2\lambda \int_0^{u^*} \tanh^{-\mathsf{T}}(\upsilon/\lambda)\, \mathrm{d}\upsilon \tag{23}$$

Then, one can derive that

$$L_{1\hat{x}}^{\mathsf{T}}(h(\hat{x}) + g(\hat{x})u^*) = -L_{1\hat{x}}^{\mathsf{T}} \Lambda(\hat{x}) L_{1\hat{x}} \tag{24}$$

Proof: Taking the time derivative of (18) and employing (20), we have that

$$\dot{V}(\hat{x}) = V_{\hat{x}}^{*\mathsf{T}}\left(h(\hat{x}) + g(\hat{x})u^*\right) = -Q(\hat{x}) - 2\lambda \int_0^{u^*} \tanh^{-\mathsf{T}}(\upsilon/\lambda)\, \mathrm{d}\upsilon \tag{25}$$

By utilising (23) and (25), we derive that

$$h(\hat{x}) + g(\hat{x})u^* = -\left(V_{\hat{x}}^* V_{\hat{x}}^{*\mathsf{T}}\right)^{-1}\left(V_{\hat{x}}^* V_{\hat{x}}^{*\mathsf{T}}\right) \Lambda(\hat{x}) L_{1\hat{x}} = -\Lambda(\hat{x}) L_{1\hat{x}} \tag{26}$$

Hence, one can obtain (24) by premultiplying both sides of (26) with $L_{1\hat{x}}^{\mathsf{T}}$. □

4.2 Online NN-based optimal control scheme

The purpose of this subsection is to design an online optimal control scheme by using a unique critic NN. According to the universal approximation property of feedforward NNs [29], $V(\hat{x})$ in (18) can accurately be represented as

$$V(\hat{x}) = W_2^{\mathsf{T}} \sigma(V_2^{\mathsf{T}} \hat{x}) + \varepsilon_2(\hat{x}) \tag{27}$$

where $V_2 \in \mathbb{R}^{n\times N_2}$ and $W_2 \in \mathbb{R}^{N_2}$ are the weights for the input layer to the hidden layer and the hidden layer to the output layer, respectively. N_2 is the number of the neurons. Since the inner weights are generally initialised randomly and kept constant, the activation function $\sigma(V_2^{\mathsf{T}}\hat{x})$ is written as $\sigma(\hat{x})$ for briefly. $\sigma(\hat{x}) = [\sigma_1(\hat{x}), \sigma_2(\hat{x}), \ldots, \sigma_{N_2}(\hat{x})]^{\mathsf{T}} \in \mathbb{R}^{N_2}$ with $\sigma_j(\hat{x}) \in \mathrm{C}^1(\Omega)$, $\sigma_j(0) = 0$, and the set $\{\sigma_j(\hat{x})\}_1^{N_2}$ is selected to be linearly independent. $\varepsilon_2(\hat{x})$ is the NN function reconstruction error. The derivative of $V(\hat{x})$ with respect to $\hat{x}$ is given by

$$V_{\hat{x}} = \nabla\sigma^{\mathsf{T}}(\hat{x}) W_2 + \nabla\varepsilon_2 \tag{28}$$

with $\nabla\sigma(\hat{x}) = \partial\sigma(\hat{x})/\partial\hat{x}$ and $\nabla\sigma(0) = 0$.

By using (28), (20) can be represented as

$$u^*(\hat{x}) = -\lambda \tanh\left(\frac{1}{2\lambda} g^{\mathsf{T}}(\hat{x}) \nabla\sigma^{\mathsf{T}} W_2\right) + \varepsilon_{u^*} \tag{29}$$

where

$$\varepsilon_{u^*} = -\frac{1}{2}(1 - \tanh^2(\xi_1)) g^{\mathsf{T}}(\hat{x}) \nabla\varepsilon_2$$

with $\xi_1 \in \mathbb{R}^m$ selected between $\frac{1}{2\lambda} g^{\mathsf{T}}(\hat{x}) \nabla\sigma^{\mathsf{T}} W_2$ and $\frac{1}{2\lambda} g^{\mathsf{T}}(\hat{x})(\nabla\sigma^{\mathsf{T}} W_2 + \nabla\varepsilon_2)$.

Similarly, with the aid of (28), (22) can be rewritten as

$$W_2^{\mathsf{T}} \nabla\sigma h(\hat{x}) + Q(\hat{x}) + \lambda^2 \ln\left[1 - \tanh^2(\mathfrak{L}_1(\hat{x}))\right] + \varepsilon_{\mathrm{HJB}} = 0 \tag{30}$$

where

$$\mathfrak{L}_1(\hat{x}) = \frac{1}{2\lambda} g^{\mathsf{T}}(\hat{x}) \nabla\sigma^{\mathsf{T}} W_2$$

$$\varepsilon_{\mathrm{HJB}} = \nabla\varepsilon_2^{\mathsf{T}}\left[h(\hat{x}) + g(\hat{x}) \frac{\lambda}{\xi_2} \tanh(\xi_3)(\tanh^2(\xi_3) - 1)\right]$$

with $\xi_2 \in \mathbb{R}$ selected between $1 - \tanh^2(\mathfrak{C}(\hat{x}))$ and $1 - \tanh^2(\mathfrak{L}_1(\hat{x}))$, and $\xi_3 \in \mathbb{R}^m$ chosen between $\mathfrak{C}(\hat{x})$ and $\mathfrak{L}_1(\hat{x})$.

Remark 1: It was shown in [20] that the HJB approximation error $\varepsilon_{\mathrm{HJB}}$ converges to zero as the number of hidden layer neurons N_2 increases. That is, for $\forall\ \varepsilon_h > 0$, there exists a positive N_0 (depending only on ε_h) such that $N_2 > N_0$ implies $\|\varepsilon_{\mathrm{HJB}}\| \leq \varepsilon_h$.

Since the ideal NN weight W_2 is typically unknown, (27) cannot be implemented in real control process. Therefore we employ $\hat{V}(\hat{x})$ to approximate the value function in (18) as

$$\hat{V}(\hat{x}) = \hat{W}_2^{\mathrm{T}}\sigma(\hat{x}) \tag{31}$$

where $\hat{W}_2$ is the estimation of W_2. The weight estimates error for the critic NN is defined as

$$\tilde{W}_2 = W_2 - \hat{W}_2 \tag{32}$$

By utilising (31), the estimate of (20) is written as

$$\hat{u}(\hat{x}) = -\lambda \tanh\left(\frac{1}{2\lambda} g^{\mathrm{T}}(\hat{x})\nabla\sigma^{\mathrm{T}}\hat{W}_2\right) \tag{33}$$

From (31) and (33), we derive the approximate Hamiltonian as

$$\begin{aligned} H(\hat{x}, \hat{W}_2) &= \hat{W}_2^{\mathrm{T}}\nabla\sigma h(\hat{x}) + Q(\hat{x}) \\ &\quad + \lambda^2 \ln\left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right] \triangleq e \end{aligned} \tag{34}$$

where $\mathfrak{L}_2(\hat{x}) = \frac{1}{2\lambda} g^{\mathrm{T}}(\hat{x})\nabla\sigma^{\mathrm{T}}\hat{W}_2$.

Combining (30), (32) and (34), we derive that

$$e = -\tilde{W}_2^{\mathrm{T}}\nabla\sigma h(\hat{x}) + \lambda^2\left[\Gamma(\mathfrak{L}_2(\hat{x})) - \Gamma(\mathfrak{L}_1(\hat{x}))\right] - \varepsilon_{\mathrm{HJB}} \tag{35}$$

where $\Gamma(\mathfrak{L}_i(\hat{x})) = \ln[1 - \tanh^2(\mathfrak{L}_i(\hat{x}))]$ $(i = 1, 2)$. Observe that, for $\forall\ \xi \in \mathbb{R}$, $\Gamma(\xi)$ can accurately be represented by

$$\begin{aligned} \Gamma(\xi) &= \ln(1 - \tanh^2(\xi)) \\ &= \begin{cases} \ln 4 - 2\xi - 2\ln(1 + \exp(-2\xi)), & \xi > 0 \\ \ln 4 + 2\xi - 2\ln(1 + \exp(2\xi)), & \xi < 0 \end{cases} \end{aligned}$$

That is,

$$\Gamma(\xi) = \ln 4 - 2\xi\,\mathrm{sgn}(\xi) - 2\ln\left[1 + \exp(-2\xi\,\mathrm{sgn}(\xi))\right] \tag{36}$$

Replacing ξ with $\mathfrak{L}_i(\hat{x})$ in (36) and observing $\mathfrak{L}_i(\hat{x}) \in \mathbb{R}^m$, we obtain that

$$\begin{aligned} \Gamma(\mathfrak{L}_i(\hat{x})) &= \ln 4 - 2\mathfrak{L}_i^{\mathrm{T}}(\hat{x})\mathrm{sgn}(\mathfrak{L}_i(\hat{x})) \\ &\quad - 2\ln\left[1 + \exp\left(-2\mathfrak{L}_i^{\mathrm{T}}(\hat{x})\mathrm{sgn}(\mathfrak{L}_i(\hat{x}))\right)\right] \end{aligned} \tag{37}$$

where $\mathrm{sgn}(\mathfrak{L}_i(\hat{x})) \in \mathbb{R}^m$ is a sign vector-valued function [32]. Therefore by (35) and (37), we obtain that

$$\begin{aligned} e &= 2\lambda^2\left[\mathfrak{L}_1^{\mathrm{T}}(\hat{x})\mathrm{sgn}(\mathfrak{L}_1(\hat{x})) - \mathfrak{L}_2^{\mathrm{T}}(\hat{x})\mathrm{sgn}(\mathfrak{L}_2(\hat{x}))\right] \\ &\quad - \tilde{W}_2^{\mathrm{T}}\nabla\sigma h(\hat{x}) + \lambda^2\Delta_{\mathfrak{L}} - \varepsilon_{\mathrm{HJB}} \\ &= \lambda\left[W_2^{\mathrm{T}}\nabla\sigma g(\hat{x})\mathrm{sgn}(\mathfrak{L}_1(\hat{x})) - \hat{W}_2^{\mathrm{T}}\nabla\sigma g(\hat{x})\mathrm{sgn}(\mathfrak{L}_2(\hat{x}))\right] \\ &\quad - \tilde{W}_2^{\mathrm{T}}\nabla\sigma h(\hat{x}) + \lambda^2\Delta_{\mathfrak{L}} - \varepsilon_{\mathrm{HJB}} \\ &= -\tilde{W}_2^{\mathrm{T}}\nabla\sigma h(\hat{x}) + \lambda\tilde{W}_2^{\mathrm{T}}\nabla\sigma g(\hat{x})\mathrm{sgn}(\mathfrak{L}_2(\hat{x})) + \mathfrak{D}_1(\hat{x}) \end{aligned} \tag{38}$$

where

$$\begin{aligned} \Delta_{\mathfrak{L}} &= 2\ln\frac{1 + \exp\left[-2\mathfrak{L}_1^{\mathrm{T}}(\hat{x})\mathrm{sgn}(\mathfrak{L}_1(\hat{x}))\right]}{1 + \exp\left[-2\mathfrak{L}_2^{\mathrm{T}}(\hat{x})\mathrm{sgn}(\mathfrak{L}_2(\hat{x}))\right]} \\ \mathfrak{D}_1(\hat{x}) &= \lambda W_2^{\mathrm{T}}\nabla\sigma g(\hat{x})\left[\mathrm{sgn}(\mathfrak{L}_1(\hat{x})) - \mathrm{sgn}(\mathfrak{L}_2(\hat{x}))\right] \\ &\quad + \lambda^2\Delta_{\mathfrak{L}} - \varepsilon_{\mathrm{HJB}} \end{aligned}$$

Remark 2: From the expression of $\Delta_{\mathfrak{L}}$, one can obtain that $\Delta_{\mathfrak{L}} \in [-\ln 4, \ln 4]$. Meanwhile, by Remark 1, one can conclude that $\mathfrak{D}_1(\hat{x})$ is a bounded function since W_2 is generally set to be bounded and λ is a constant.

In order to obtain the minimum value of e, it is desired to minimise the objective function $E = \frac{1}{2}e^{\mathrm{T}}e$ with the gradient descent algorithm. However, tuning the critic NN weights to minimise E alone does not guarantee the stability of system (17) during the learning process of NNs. For ensuring the algorithm can be implemented online, a novel weight update law for the critic NN is developed by

$$\begin{aligned} \dot{\hat{W}}_2 &= -\eta\bar{\phi}\left(Q(\hat{x}) + \hat{W}_2^{\mathrm{T}}\nabla\sigma h(\hat{x}) + \lambda^2\ln\left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right]\right) \\ &\quad + \frac{\eta}{2}\Sigma(\hat{x}, \hat{u})\nabla\sigma g(\hat{x})\left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right]g^{\mathrm{T}}(\hat{x})L_{1\hat{x}} \\ &\quad + \eta\left(\lambda\nabla\sigma g(\hat{x})\left[\tanh(\mathfrak{L}_2(\hat{x})) - \mathrm{sgn}(\mathfrak{L}_2(\hat{x}))\right]\frac{\varphi^{\mathrm{T}}}{m_s}\hat{W}_2\right. \\ &\quad \left.-(F_2\hat{W}_2 - F_1\varphi^{\mathrm{T}}\hat{W}_2)\right) \end{aligned} \tag{39}$$

where $\bar{\phi} = \phi/m_s^2, \varphi = \phi/m_s, m_s = 1 + \phi^{\mathrm{T}}\phi, \phi = \nabla\sigma h(\hat{x}) - \lambda\nabla\sigma g(\hat{x})\tanh(\mathfrak{L}_2(\hat{x}))$, $\eta > 0$ is a design parameter, $L_{1\hat{x}}$ is defined as in Lemma 1, F_1 and F_2 are tuning parameters with suitable dimensions, and $\Sigma(\hat{x}, \hat{u})$ is described by

$$\Sigma(\hat{x}, \hat{u}) = \begin{cases} 0, & \text{if } L_{1\hat{x}}h(\hat{x}) - \lambda L_{1\hat{x}}g(\hat{x})\tanh(\mathfrak{L}_2(\hat{x})) < 0 \\ 1, & \text{otherwise} \end{cases} \tag{40}$$

Remark 3: From the expression of (34) and (39), one shall find that $\hat{x} = 0$ gives birth to $H(\hat{x}, \hat{W}_2) = e = 0$ and $\dot{\hat{W}}_2 = 0$. That is, when the system state $\hat{x}(t)$ goes to zero, the approximated value function $\hat{V}(\hat{x})$ will no longer be updated. However, the optimal control might not be obtained at finite time t_0, which makes $x(t_0) = 0$. In order to avoid this circumstance from occurring, probing noise is added to the control input, that is, persistency of excitation (PE) condition is required.

By the definition of ϕ in (39), we have that $\nabla\sigma h(\hat{x}) = \phi + \lambda\nabla\sigma g(\hat{x})\tanh(\mathfrak{L}_2(\hat{x}))$. Therefore (38) can be represented by

$$e = -\tilde{W}_2^{\mathrm{T}}\phi + \lambda\tilde{W}_2^{\mathrm{T}}\nabla\sigma g(\hat{x})\mathfrak{N}(\hat{x}) + \mathfrak{D}_1(\hat{x}) \tag{41}$$

where $\mathfrak{N}(\hat{x}) = \mathrm{sgn}(\mathfrak{L}_2(\hat{x})) - \tanh(\mathfrak{L}_2(\hat{x}))$. From (32), (34), (39) and (41), we derive that

$$\begin{aligned} \dot{\tilde{W}}_2 &= \eta\frac{\varphi}{m_s}\left[-\tilde{W}_2^{\mathrm{T}}\phi + \lambda\tilde{W}_2^{\mathrm{T}}\nabla\sigma g(\hat{x})\mathfrak{N}(\hat{x}) + \mathfrak{D}_1(\hat{x})\right] \\ &\quad - \frac{\eta}{2}\Sigma(\hat{x}, \hat{u})\nabla\sigma g(\hat{x})\left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right]g^{\mathrm{T}}(\hat{x})L_{1\hat{x}} \\ &\quad + \eta\left[\lambda\nabla\sigma g(\hat{x})\mathfrak{N}(\hat{x})\frac{\varphi^{\mathrm{T}}}{m_s}\hat{W}_2 + (F_2\hat{W}_2 - F_1\varphi^{\mathrm{T}}\hat{W}_2)\right] \end{aligned} \tag{42}$$

A general schematic programming of the proposed control algorithm is presented in Fig. 1.

5 Stability analysis and performance of the closed-loop system

The purpose of this section is to present our main results by employing Lyapunov's direct method. Before engaging

doi: 10.1049/iet-cta.2013.0472

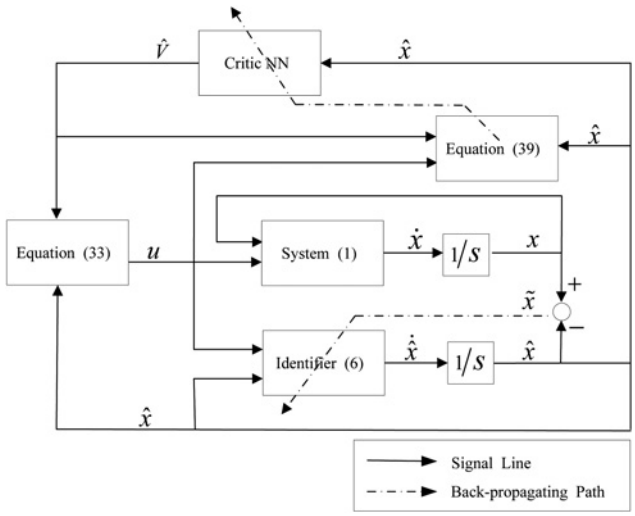

Fig. 1 *Proposed control algorithm for uncertain non-linear CT systems.*

in developing the main theorem, we need to provide a mild assumption as follows:

Assumption 4: The critic NN weight W_2 is bounded by a known positive constant W_{M_2} over the compact set Ω, that is, $\|W_2\| \leq W_{M_2}$. Meanwhile, the critic NN approximation error $\varepsilon_2(\hat{x})$ is bounded by a known positive constant ε_{M_2} over Ω, that is, $\|\varepsilon_2(\hat{x})\| \leq \varepsilon_{M_2}$, for $\forall\, \hat{x} \in \Omega$. In addition, ε_{u^*} in (29) is upper bounded by ε_a over Ω, that is $\|\varepsilon_{u^*}\| \leq \varepsilon_a$.

With the aid of Assumptions 1–4 and Facts 1–2, our main theorem is derived as follows:

Theorem 2: Consider the non-linear CT system described by (1) with associated HJB equation (22). Let Assumptions 1–4 hold and take the control input for system (1) as in (33). Moreover, let weights update laws for the identifier NN be (8) and (9), and let the critic NN weight tuning law be given by (39). Then, the system identifier error $\tilde{x}(t)$, the NN weight estimates errors $\tilde{W}_1$, $\tilde{V}_1$, and $\tilde{W}_2$ are guaranteed to be UUB.

Proof: Consider the Lyapunov function candidate

$$L(t) = L_1(t) + L_2(t) + L_3(t) \tag{43}$$

where $L_1(t)$ is defined as in Lemma 1, $L_2(t) = J(t)$ with $J(t)$ defined as in Theorem 1, and $L_3(t) = \frac{1}{2}\mathrm{tr}\left(\tilde{W}_2^{\mathsf{T}}\eta^{-1}\tilde{W}_2\right)$.

Taking the time derivative of (43) and using Theorem 1, we obtain that

$$\begin{aligned}\dot{L}(t) &= \dot{L}_1(t) + \dot{L}_2(t) + \dot{L}_3(t)\\ &\leq L_{1\hat{x}}^{\mathsf{T}}[h(\hat{x}) - \lambda g(\hat{x})\tanh(\mathfrak{L}_2(\hat{x}))]\\ &\quad - \frac{\alpha}{2}\|\tilde{x}\|^2 + \mathfrak{B}\|\tilde{x}\| + \dot{\tilde{W}}_2^{\mathsf{T}}\eta^{-1}\tilde{W}_2\end{aligned} \tag{44}$$

By using (42), we derive the last term in (44) as

$$\begin{aligned}\dot{\tilde{W}}_2^{\mathsf{T}}\eta^{-1}\tilde{W}_2 &= \left[-\tilde{W}_2^{\mathsf{T}}\phi + \lambda\tilde{W}_2^{\mathsf{T}}\nabla\sigma g(\hat{x})\mathfrak{N}(\hat{x}) + \mathfrak{D}_1(\hat{x})\right]\frac{\varphi^{\mathsf{T}}}{m_s}\tilde{W}_2\\ &\quad - \frac{1}{2}\Sigma(\hat{x},\hat{u})L_{1\hat{x}}^{\mathsf{T}}g(\hat{x})\left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right]\\ &\quad \times g^{\mathsf{T}}(\hat{x})\nabla\sigma^{\mathsf{T}}\tilde{W}_2\\ &\quad + \lambda\tilde{W}_2^{\mathsf{T}}\nabla\sigma g(\hat{x})\mathfrak{N}(\hat{x})\frac{\varphi^{\mathsf{T}}}{m_s}\hat{W}_2\\ &\quad + \tilde{W}_2^{\mathsf{T}}(F_2\hat{W}_2 - F_1\varphi^{\mathsf{T}}\hat{W}_2)\\ &= -\tilde{W}_2^{\mathsf{T}}\varphi\varphi^{\mathsf{T}}\tilde{W}_2 + \bar{\mathfrak{D}}_1(\hat{x})\varphi^{\mathsf{T}}\tilde{W}_2 + \tilde{W}_2^{\mathsf{T}}\bar{\mathfrak{D}}_2(\hat{x})\\ &\quad - \frac{1}{2}\Sigma(\hat{x},\hat{u})L_{1\hat{x}}^{\mathsf{T}}g(\hat{x})\left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right]\\ &\quad \times g^{\mathsf{T}}(\hat{x})\nabla\sigma^{\mathsf{T}}\tilde{W}_2\\ &\quad + \tilde{W}_2^{\mathsf{T}}(F_2\hat{W}_2 - F_1\varphi^{\mathsf{T}}\hat{W}_2)\end{aligned} \tag{45}$$

where $\bar{\mathfrak{D}}_1(\hat{x}) = \frac{\mathfrak{D}_1(\hat{x})}{m_s}$ and $\bar{\mathfrak{D}}_2(\hat{x}) = \lambda\nabla\sigma g(\hat{x})\mathfrak{N}(\hat{x})\frac{\varphi^{\mathsf{T}}}{m_s}W_2$.

Observe that

$$\begin{aligned}&\tilde{W}_2^{\mathsf{T}}(F_2\hat{W}_2 - F_1\varphi^{\mathsf{T}}\hat{W}_2)\\ &\quad = \tilde{W}_2^{\mathsf{T}}F_2W_2 - \tilde{W}_2^{\mathsf{T}}F_2\tilde{W}_2 - \tilde{W}_2^{\mathsf{T}}F_1\varphi^{\mathsf{T}}W_2 + \tilde{W}_2^{\mathsf{T}}F_1\varphi^{\mathsf{T}}\tilde{W}_2\end{aligned}$$

Denote $\mathcal{Z}^{\mathsf{T}} = [\tilde{W}_2^{\mathsf{T}}\varphi, \tilde{W}_2^{\mathsf{T}}]$. Then, (45) can be developed by

$$\begin{aligned}\dot{\tilde{W}}_2^{\mathsf{T}}\eta^{-1}\tilde{W}_2 &= -\mathcal{Z}^{\mathsf{T}}K\mathcal{Z} + \mathcal{Z}^{\mathsf{T}}G - \frac{1}{2}\Sigma(\hat{x},\hat{u})L_{1\hat{x}}^{\mathsf{T}}g(\hat{x})\\ &\quad \times \left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right]g^{\mathsf{T}}(\hat{x})\nabla\sigma^{\mathsf{T}}\tilde{W}_2\end{aligned} \tag{46}$$

where

$$K = \begin{bmatrix} I & -\frac{F_1^{\mathsf{T}}}{2} \\ -\frac{F_1}{2} & F_2 \end{bmatrix} \quad G = \begin{bmatrix} \bar{\mathfrak{D}}_1(x) \\ \bar{\mathfrak{D}}_2(x) + F_2W_2 - F_1\varphi^{\mathsf{T}}W_2 \end{bmatrix}$$

Combining (44) with (46) and selecting F_1 and F_2 such that K is positive definite, we obtain that

$$\begin{aligned}\dot{L}(t) &\leq L_{1\hat{x}}^{\mathsf{T}}[h(\hat{x}) - \lambda g(\hat{x})\tanh(\mathfrak{L}_2(\hat{x}))]\\ &\quad - \frac{\alpha}{2}\|\tilde{x}\|^2 + \mathfrak{B}\|\tilde{x}\| - \sigma_{\min}(K)\|\mathcal{Z}\|^2 + \vartheta_M\|\mathcal{Z}\|\\ &\quad - \frac{1}{2}\Sigma(\hat{x},\hat{u})L_{1\hat{x}}^{\mathsf{T}}g(\hat{x})\left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right]g^{\mathsf{T}}(\hat{x})\nabla\sigma^{\mathsf{T}}\tilde{W}_2\end{aligned} \tag{47}$$

where $\sigma_{\min}(K)$ denotes the minimum eigenvalue of K, and ϑ_M is the upper bound of $\|G\|$, that is, $\|G\| \leq \vartheta_M$.

Owing to the definition of $\Sigma(\hat{x},\hat{u})$ in (40), (47) is divided into the following two cases for discussion:

Case 1: $\Sigma(\hat{x},\hat{u}) = 0$. By the definition of $\Sigma(\hat{x},\hat{u})$ in (40), we know that the first term in (47) is negative. Since $\|\hat{x}\| > 0$ is guaranteed by persistent excitation, one can draw the conclusion that there exists a constant τ, such that $0 < \tau \leq \|\dot{\hat{x}}\|$ implies $-\|L_{1\hat{x}}\|\tau \geq L_{1\hat{x}}^{\mathsf{T}}\dot{\hat{x}}$ based on Archimedean property of

$\mathbb{R}$ [32]. Then, (47) can be developed by

$$\dot{L}(t) \le -\tau\|L_{1\hat{x}}\| - \frac{\alpha}{2}\left(\|\tilde{x}\| - \mathfrak{B}/\alpha\right)^2 - \sigma_{\min}(K)\left(\|\mathcal{Z}\| - \frac{1}{2}\vartheta_M/\sigma_{\min}(K)\right)^2 + \frac{1}{2}\mathfrak{B}^2/\alpha + \frac{1}{4}\vartheta_M^2/\sigma_{\min}(K) \quad (48)$$

Therefore we can derive that (48) implies that $\dot{L}(t) < 0$ as long as one of the following conditions holds

$$\|L_{1\hat{x}}\| > \frac{2\sigma_{\min}(K)\mathfrak{B}^2 + \alpha\vartheta_M^2}{4\tau\alpha\sigma_{\min}(K)}$$

or

$$\|\tilde{x}(t)\| > \frac{\mathfrak{B}}{\alpha} + \frac{1}{\alpha}\sqrt{\mathfrak{B}^2 + \frac{\alpha\vartheta_M^2}{2\sigma_{\min}(K)}}$$

or

$$\|\mathcal{Z}\| > \frac{1}{2\sigma_{\min}(K)}\left(\vartheta_M + \sqrt{\vartheta_M^2 + 2\mathfrak{B}^2\sigma_{\min}(K)/\alpha}\right)$$

Case 2: $\Sigma(\hat{x}, \hat{u}) = 1$. By the definition of $\Sigma(\hat{x}, \hat{u})$ in (40), we know that the first term in (47) is non-negative, which implies that the control (33) may not stabilise system (17). Under this circumstance, (47) becomes

$$\dot{L}(t) \le L_{1\hat{x}}^{\mathsf{T}} h(\hat{x}) - \lambda L_{1\hat{x}}^{\mathsf{T}} g(\hat{x})\left[\tanh(\mathfrak{L}_2(\hat{x})) + \frac{1}{2\lambda}[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))]g^{\mathsf{T}}(\hat{x})\nabla\sigma^{\mathsf{T}}\tilde{W}_2\right] - \sigma_{\min}(K)\left(\|\mathcal{Z}\| - \frac{1}{2}\vartheta_M/\sigma_{\min}(K)\right)^2 - \frac{\alpha}{2}\left(\|\tilde{x}\| - \mathfrak{B}/\alpha\right)^2 + \frac{1}{2}\mathfrak{B}^2/\alpha + \frac{1}{4}\vartheta_M^2/\sigma_{\min}(K) \quad (49)$$

Denote $\mathscr{B}(\mathfrak{L}_i(\hat{x})) = \tanh(\mathfrak{L}_i(\hat{x}))$, $i = 1, 2$. By using Taylor series, we have that

$$\begin{aligned}&\mathscr{B}(\mathfrak{L}_1(\hat{x})) - \mathscr{B}(\mathfrak{L}_2(\hat{x}))\\ &= \dot{\mathscr{B}}(\mathfrak{L}_2(\hat{x}))\left(\mathfrak{L}_1(\hat{x}) - \mathfrak{L}_2(\hat{x})\right) + O\left((\mathfrak{L}_1(\hat{x}) - \mathfrak{L}_2(\hat{x}))^2\right)\\ &= \frac{1}{2\lambda}\left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right]g^{\mathsf{T}}(\hat{x})\nabla\sigma^{\mathsf{T}}\tilde{W}_2\\ &\quad + O\left((\mathfrak{L}_1(\hat{x}) - \mathfrak{L}_2(\hat{x}))^2\right)\end{aligned} \quad (50)$$

From (50), we obtain that

$$\begin{aligned}&\tanh(\mathfrak{L}_2(\hat{x})) + \frac{1}{2\lambda}\left[1 - \tanh^2(\mathfrak{L}_2(\hat{x}))\right]g^{\mathsf{T}}(\hat{x})\nabla\sigma^{\mathsf{T}}\tilde{W}_2\\ &= \tanh(\mathfrak{L}_1(\hat{x})) - O\left((\mathfrak{L}_1(\hat{x}) - \mathfrak{L}_2(\hat{x}))^2\right)\end{aligned} \quad (51)$$

Substituting (51) into (49) and adding and subtracting $L_{1\hat{x}}^{\mathsf{T}} g(\hat{x})\varepsilon_{u^*}$ to the right-hand side of (49), we derive that

$$\dot{L}(t) \le L_{1\hat{x}}^{\mathsf{T}}\left(h(\hat{x}) + g(\hat{x})u^*\right) - L_{1\hat{x}}^{\mathsf{T}} g(\hat{x})\varepsilon_{u^*} + \lambda L_{1\hat{x}}^{\mathsf{T}} g(\hat{x}) O\left((\mathfrak{L}_1(\hat{x}) - \mathfrak{L}_2(\hat{x}))^2\right) - \sigma_{\min}(K)\left(\|\mathcal{Z}\| - \frac{1}{2}\vartheta_M/\sigma_{\min}(K)\right)^2 - \frac{\alpha}{2}\left(\|\tilde{x}\| - \mathfrak{B}/\alpha\right)^2 + \frac{1}{2}\mathfrak{B}^2/\alpha + \frac{1}{4}\vartheta_M^2/\sigma_{\min}(K) \quad (52)$$

By using Lemma 1, we can rewrite (52) as

$$\dot{L}(t) \le -\sigma_{\min}(\Lambda(\hat{x}))\left(\|L_{1\hat{x}}\| - \frac{1}{2}\theta_M/\sigma_{\min}(\Lambda(\hat{x}))\right)^2 - \sigma_{\min}(K)\left(\|\mathcal{Z}\| - \frac{1}{2}\vartheta_M/\sigma_{\min}(K)\right)^2 - \frac{\alpha}{2}\left(\|\tilde{x}\| - \mathfrak{B}/\alpha\right)^2 + \mu \quad (53)$$

where $\theta_M = g_M(\varepsilon_a + \lambda\varepsilon_b)$, ε_b is the upper bound of $O((\mathfrak{L}_1(\hat{x}) - \mathfrak{L}_2(\hat{x}))^2)$, and μ is defined as follows

$$\mu = \frac{1}{2}\mathfrak{B}^2/\alpha + \frac{1}{4}\theta_M^2/\sigma_{\min}(\Lambda(\hat{x})) + \frac{1}{4}\vartheta_M^2/\sigma_{\min}(K)$$

Consequently, we can obtain that (53) implies that $\dot{L}(t) < 0$ as long as one of the following conditions holds

$$\|L_{1\hat{x}}\| > \frac{\theta_M}{2\sigma_{\min}(\Lambda(\hat{x}))} + \sqrt{\frac{\mu}{\sigma_{\min}(\Lambda(\hat{x}))}}$$

or

$$\|\tilde{x}(t)\| > \frac{\mathfrak{B}}{\alpha} + \sqrt{\frac{2\mu}{\alpha}}$$

or

$$\|\mathcal{Z}\| > \frac{\vartheta_M}{2\sigma_{\min}(K)} + \sqrt{\frac{\mu}{\sigma_{\min}(K)}}$$

Combining Cases 1 and 2 and using the standard Lyapunov extension theorem [30], one can come to the conclusion that the system identifier error $\tilde{x}(t)$, NN weight estimates errors $\tilde{W}_1$, $\tilde{V}_1$ and $\tilde{W}_2$ are UUB. □

6 Simulation results

Consider the affine non-linear CT system described by [21]

$$\dot{x}(t) = f(x) + g(x)u \quad (54)$$

where

$$f(x) = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2 + 0.5x_2[\cos(2x_1) + 2]^2 \end{bmatrix}$$

$$g(x) = \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix}$$

The objective is to control the system with control limits of $|u| \le 1.2$. The non-quadratic cost function is given by

$$V(x) = \int_0^\infty \left(Q(x) + 2\lambda\int_0^u \tanh^{-\mathsf{T}}(\upsilon/\lambda)\,\mathrm{d}\upsilon\right)\mathrm{d}t$$

where $Q(x) = x_1^2 + x_2^2$. Meanwhile, the priori knowledge of $f(x)$ is assumed to be unavailable, which is distinctly different from it employed in [21]. For the sake of deriving the knowledge of system dynamics, an identifier NN as in (6)

doi: 10.1049/iet-cta.2013.0472

is employed. The identifier gains are selected as

$$A = [-1\ \ 1; -0.5\ \ -0.5], \quad \chi = [1\ \ 0; -0.5\ \ 0], \quad l_1 = 40$$
$$l_2 = 40, \quad k_1 = 15, \quad k_2 = 45, \quad N_1 = 8$$

and the gains for the critic NN are chosen as

$$\eta = 38, \quad \lambda = 1.2, \quad N_2 = 8$$

The activation function for the critic NN is chosen as

$$\sigma(x) = [x_1^2 \quad x_2^2 \quad x_1x_2 \quad x_1^4 \quad x_2^4 \quad x_1^3x_2 \quad x_1^2x_2^2 \quad x_1x_2^3]^{\mathrm{T}}$$

and the critic NN weights are denoted as $\hat{W}_2 = [W_2^1 \cdots W_2^8]^{\mathrm{T}}$. The initial weights $\hat{W}_1$ and $\hat{V}_1$ for the identifier NN are selected randomly within an interval of $[-10, 10]$ and $[-5, 5]$, respectively. Meanwhile, the initial weights for the critic NN are chosen to be zeros, and the initial system state is selected to be $x_0 = [3\ \ -0.5]^{\mathrm{T}}$. It is significant to point out that, under this circumstance, the initial control can not stabilise system (54). That is, no initial stabilising control is required for implementation of the algorithm. From this fact, one shall find that a distinct advantage of the developed algorithm in this paper as compared with the method proposed in [21] lies in that there is no special requirement imposed on the initial control rather than the initial stabilising control is indispensable in [21]. To guarantee the PE qualitatively, a small exploratory signal $n(t) = \sin^5(t)\cos(t) + \sin^5(2t)\cos(0.2t)$ is added to control policy $u(t)$ for the first 10 s.

The computer simulation results are presented in Figs. 2–8. Fig. 2 presents the trajectories of system state $x(t)$. Fig. 3 indicates the performance of the system identification error. Fig. 4 shows the 2-norm of the weights for the identifier NN. Fig. 5 provides the performance of the convergence of the critic NN weights. Fig. 6 presents the optimal controller with actuator saturation. In order to make comparison with the controller without considering the actuator saturation, we use Figs. 7 and 8 to show the system state and the controller designed separately regardless of the actuator saturation. The actuator saturation actually exists; therefore, the control input is limited to the bounded value when it overruns the saturation bound. From Figs. 2–8, it is observed that the identifier NN weights and the critic NN weights

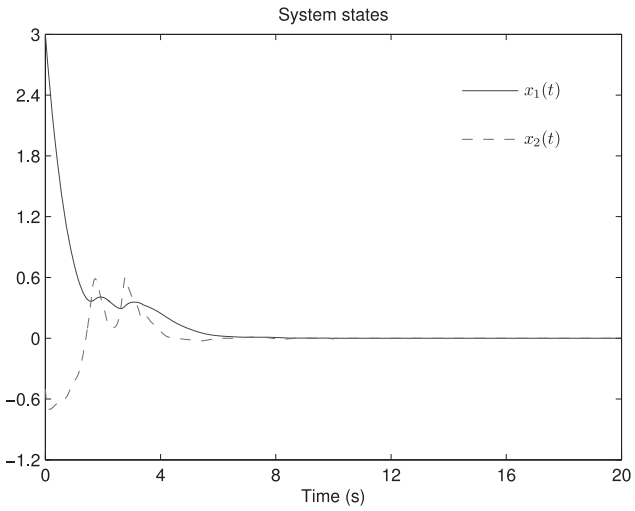

Fig. 2 *Evolution of state $x(t)$ for learning process*

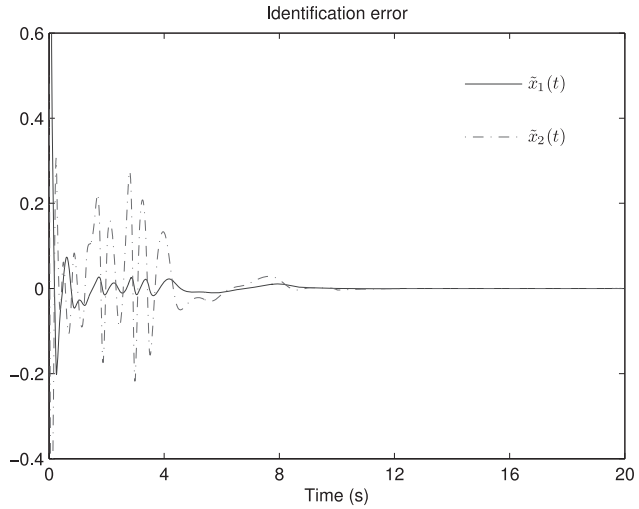

Fig. 3 *System identification error $\tilde{x}(t)$*

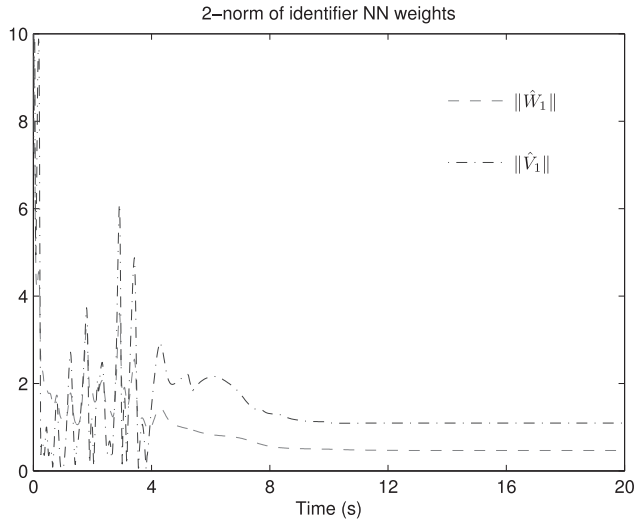

Fig. 4 *Two-norm of identifier NN weights $\|\hat{W}_1\|$ and $\|\hat{V}_1\|$*

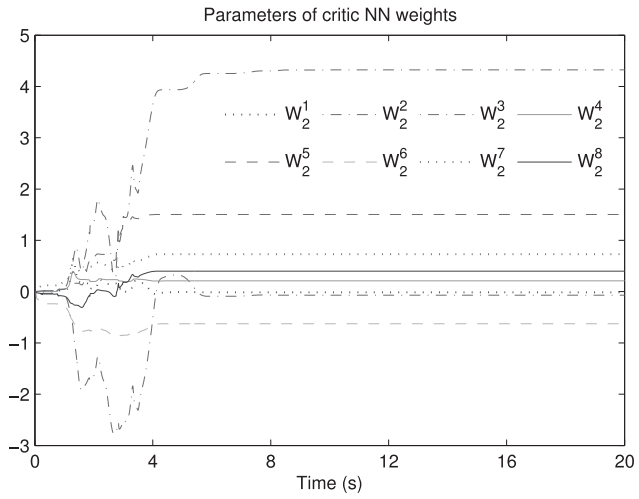

Fig. 5 *Convergence of critic NN weights $\hat{W}_2$*

are tuned simultaneously. It is also observed that the system states, and the estimates weights of both the identifier NN and the critic NN are all guaranteed to be UUB, while

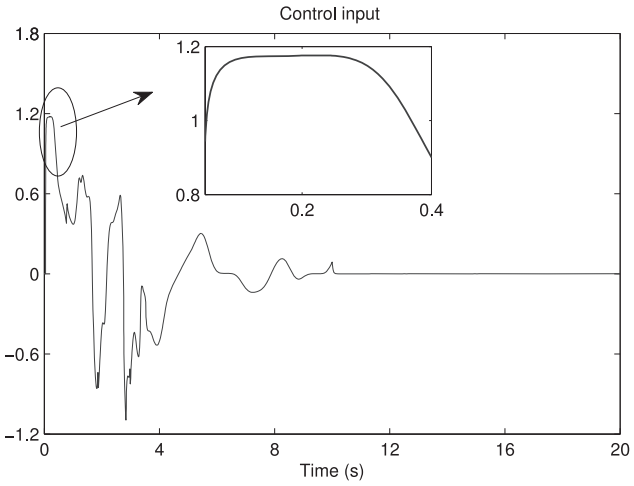

Fig. 6 *Control input u*

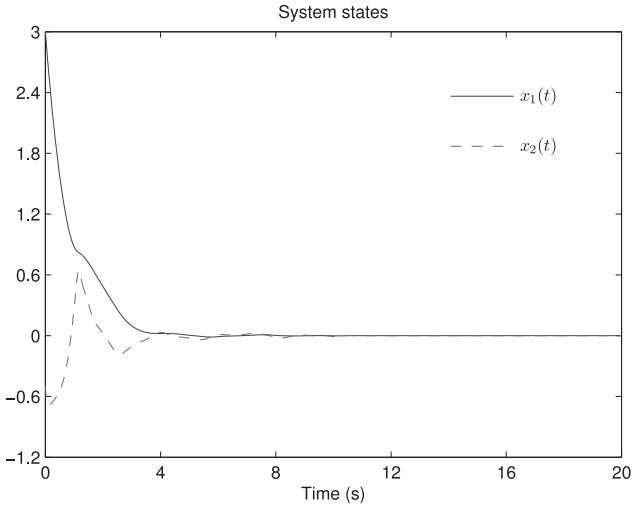

Fig. 7 *Evolution of state $x(t)$ without considering actuator saturation*

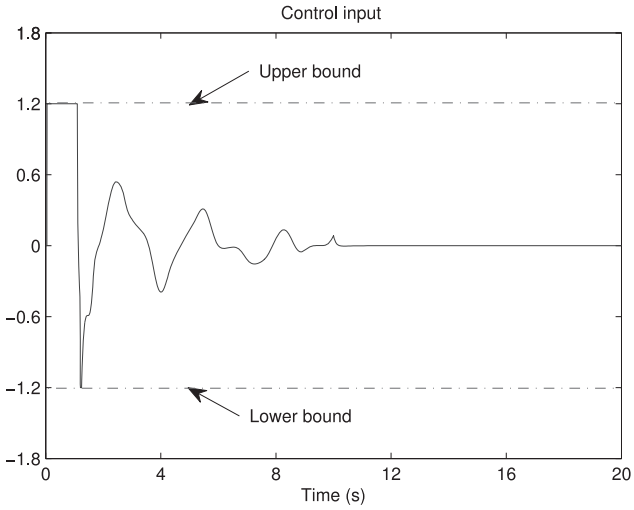

Fig. 8 *Control input u without considering actuator saturation*

keeping closed-loop system stable. In addition, comparing Fig. 6 with Fig. 8, one shall find that the restriction of actuator saturation has been successfully overcome.

7 Conclusions

This paper develops an online adaptive optimal control scheme for solving the infinite-horizon optimal control problem of uncertain non-linear CT systems with saturation actuators. An identifier–critic architecture is first time used to approximate the HJB equation. Two NNs are employed in this architecture: a robust NN is utilised to estimate the uncertain system dynamics and a critic NN is used to derive the optimal control instead of typical action–critic dual networks. Based on the developed architecture, the identifier NN and the critic NN are tuned simultaneously. Meanwhile, no initial stabilising control is required. By using Lyapunov's direct method, the weights of the identifier NN and the critic NN are guaranteed to be UUB, while keeping the closed-loop system stable. A limitation of the method is that the structure of non-linearities should be known first. In our future work, we will focus on designing optimal controllers for unknown non-affine non-linear CT systems without initial stabilising control policy.

8 Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61034002, 61233001 and 61273140.

9 References

1 Mahmoud, M.S.: 'Stabilization of dynamical systems with nonlinear actuators', *J. Franklin Inst.*, 1997, **334**, (3), pp. 357–375
2 Haidar, A., Boukas, E.K., Xu, S., Lam, J.: 'Exponential stability and static output feedback stabilisation of singular time-delay systems with saturating actuators', *IET Control Theory Appl.*, 2009, **3**, (9), pp. 1293–1305
3 Song, G., Zhang, Y., Xu, S.: 'Stability and l_2-gain analysis for a class of discrete-time nonlinear Markovian jump systems with actuator saturation and incomplete knowledge of transition probabilities', *IET Control Theory Appl.*, 2012, **6**, (17), pp. 2716–2723
4 Fan, J.H., Zhang, Y.M., Zheng, Z.Q.: 'Robust fault-tolerant control against time-varying actuator faults and saturation', *IET Control Theory Appl.*, 2012, **6**, (14), pp. 2198–2208
5 Kokotović, P., Arcak, M.: 'Constructive nonlinear control: a historical perspective', *Automatica*, 2001, **37**, (5), pp. 637–662
6 Abdollahi, F., Talebi, H.A., Patel, R.V.: 'A stable neural network-based observer with application to flexible-joint manipulators', *IEEE Trans. Neural Netw.*, 2006, **17**, (1), pp. 118–129
7 Lewis, F.L., Syrmos, V.L.: 'Optimal Control' (John Wiley & Sons, 1995)
8 Laub, A.J.: 'A Schur method for solving algebraic Riccati equations', *IEEE Trans. Autom. Control*, 1979, **24**, (6), pp. 913–921
9 Bellman, R.E.: 'Dynamic Programming' (Princeton University Press, 1957)
10 Werbos, P.J.: 'Beyond regression: new tools for prediction and analysis in the Behavioral Sciences'. PhD thesis, Harvard University, 1974
11 Werbos, P.J.: 'Advanced forecasting methods for global crisis warning and models of intelligence', *Gen. Syst. Yearbook*, 1977, **22**, pp. 25–38
12 Werbos, P.J.: 'Approximate dynamic programming for real-time control and neural modeling', in: White, D.A. and Sofge, D.A. (Ed.): 'Handbook of intelligent control: neural, fuzzy, and adaptive approaches' (Van Nostrand Reinhold, 1992)
13 Bertsekas, D.P., Tsitsiklis, J.N.: 'Neuro-Dynamic Programming' (Athena Scientific, 1996)
14 Prokhorov, D.V., Wunsch, D.C.: 'Adaptive critic designs', *IEEE Trans. Neural Netw.*, 1997, **8**, (5), pp. 997–1007
15 Murray, J.J., Cox, C.J., Lendaris, G.G., Saeks, R.: 'Adaptive dynamic programming', *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, 2002, **32**, (2), pp. 140–153
16 Wang, F.Y., Zhang, H., Liu, D.: 'Adaptive dynamic programming: an introduction', *IEEE Comput. Intell. Mag.*, 2009, **4**, (2), pp. 39–47
17 Li, H., Liu, D.: 'Optimal control for discrete-time affine nonlinear systems using general value iteration', *IET Control Theory Appl.*, 2012, **6**, (18), pp. 2725–2736

doi: 10.1049/iet-cta.2013.0472

18 Sutton, R.S., Barto, A.G.: 'Reinforcement learning – an introduction' (MIT Press, 1998)
19 Lewis, F.L., Vrabie, D.: 'Reinforcement learning and adaptive dynamic programming for feedback control', *IEEE Circuits Syst. Mag.*, 2009, **9**, (3), pp. 32–50
20 Abu-Khalaf, M., Lewis, F.L.: 'Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach', *Automatica*, 2005, **41**, (5), pp. 779–791
21 Vamvoudakis, K.G., Lewis, F.L.: 'Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem', *Automatica*, 2010, **46**, (5), pp. 878–888
22 Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K.G., Lewis, F.L., Dixon, W.E.: 'A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems', *Automatica*, 2013, **49**, (1), pp. 82–92
23 Dierks, T., Jagannathan, S.: 'Optimal control of affine nonlinear continuous-time systems'. American Control Conf., Baltimore, MD, USA, June–July 2010, pp. 1568–1573
24 Zhang, H., Cui, L., Zhang, X., Luo, Y.: 'Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method', *IEEE Trans. Neural Netw.*, 2011, **22**, (12), pp. 2226–2236
25 Yu, W.: 'Recent advances in intelligent control systems' (Springer–Verlag, 2009)
26 Haykin, S.: 'Neural networks and learning machines' (Prentice-Hall, 2008)
27 Modares, H., Lewis, F.L., Sistani, M.: 'Online solution of nonquadratic two-player zero-sum games arising in the H_∞ control of constrained input systems', *Int. J. Adapt. Control Signal Process.*, 2012, doi: 10.1002/acs.2348
28 Khalil, H.: 'Nonlinear systems' (Prentice-Hall, 2002, 3rd edn.)
29 Hornic, K., Stinchombe, M.: 'Multilayer feedforward neural networks are universal approximators', *Neural Netw.*, 1989, **2**, (5), pp. 359–366
30 Lewis, F.L., Jagannathan, S., Yesildirek, A.: 'Neural network control of robot manipulators and nonlinear systems' (Taylor & Francis, 1999)
31 Beard, R., Saridis, G., Wen, J.: 'Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation', *Automatica*, 1997, **33**, (12), pp. 2159–2177
32 Rudin, W.: 'Principles of mathematical analysis' (McGraw-Hill, 1976, 3rd edn.)