Automatica 49 (2013) 82–92

Contents lists available at SciVerse ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems[☆]

S. Bhasin[a,1], R. Kamalapurkar[b], M. Johnson[b], K.G. Vamvoudakis[c], F.L. Lewis[d], W.E. Dixon[b]

[a] *Department of Electrical Engineering, Indian Institute of Technology, Delhi, India*

[b] *Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA*

[c] *Center for Control, Dynamical Systems, and Computation (CCDC), University of California Santa Barbara, CA 93106-9560, USA*

[d] *Automation and Robotics Research Institute, The University of Texas at Arlington, 7300 Jack Newell Blvd. S., Ft. Worth, TX 76118, USA*

ARTICLE INFO

Article history:
Received 4 May 2011
Received in revised form
23 July 2012
Accepted 1 August 2012
Available online 23 October 2012

Keywords:
Learning control
Adaptive control
Optimal control
Approximate dynamic programming
Actor–critic–identifier

ABSTRACT

An online adaptive reinforcement learning-based solution is developed for the infinite-horizon optimal control problem for continuous-time uncertain nonlinear systems. A novel actor–critic–identifier (ACI) is proposed to approximate the Hamilton–Jacobi–Bellman equation using three neural network (NN) structures—actor and critic NNs approximate the optimal control and the optimal value function, respectively, and a robust dynamic neural network identifier asymptotically approximates the uncertain system dynamics. An advantage of using the ACI architecture is that learning by the actor, critic, and identifier is continuous and simultaneous, without requiring knowledge of system drift dynamics. Convergence of the algorithm is analyzed using Lyapunov-based adaptive control methods. A persistence of excitation condition is required to guarantee exponential convergence to a bounded region in the neighborhood of the optimal control and uniformly ultimately bounded (UUB) stability of the closed-loop system. Simulation results demonstrate the performance of the actor–critic–identifier method for approximate optimal control.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Reinforcement learning (RL) uses evaluative feedback from the environment to take appropriate actions (Sutton & Barto, 1998). One of the most widely used architectures to implement RL algorithms is the actor–critic architecture, where an actor performs certain actions by interacting with its environment, the critic evaluates the actions and gives feedback to the actor, leading to improvement in performance of subsequent actions (Barto, Sutton, & Anderson, 1983; Sutton & Barto, 1998; Widrow, Gupta, & Maitra, 1973). Actor–critic algorithms are pervasive in machine learning and are used to learn the optimal policy online for finite-space discrete-time Markov decision problems (Barto et al., 1983; Konda & Tsitsiklis, 2004; Prokhorov, Wunsch, & C, 1997; Sutton & Barto, 1998; Werbos, 1990).

Similar to RL, optimal control involves selection of an optimal policy based on some long-term performance criteria. Dynamic Programming (DP) provides a means to solve optimal control problems (Kirk, 2004); however, DP is implemented backward in time, making it offline and computationally expensive for complex systems. Owing to the similarities between optimal control and RL (Sutton, Barto, & Williams, 1992), Werbos (1990) introduced RL-based actor–critic methods for optimal control, called Approximate Dynamic Programming (ADP). ADP uses neural networks (NNs) to approximately solve DP forward-in-time, thus avoiding the *curse of dimensionality*. A detailed discussion of ADP-based designs is found in Bertsekas and Tsitsiklis (1996), Prokhorov et al. (1997) and Si, Barto, Powell, and Wunsch (2004). The success of ADP prompted a major research effort towards designing ADP-based optimal feedback controllers. The discrete/iterative nature of the ADP formulation lends itself naturally to the design of discrete-time optimal controllers (Al-Tamimi, Lewis, & Abu-Khalaf, 2008; Balakrishnan & Biega, 1996; Dierks, Thumati, & Jagannathan, 2009; Ferrari & Stengel, 2002; He & Jagannathan, 2007; Lendaris, Schultz, & Shannon, 2000; Padhi, Unnikrishnan, Wang, & Balakrishnan, 2006).

Extensions of ADP-based controllers to continuous-time systems entails challenges in proving stability, and convergence, and

[☆] This research is supported in part by the NSF CAREER award no. 0547448, NSF award no. 0901491, and the Department of Energy, grant no. DE-FG04-86NE37967 as part of the DOE University Research Program in Robotics (URPR). The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Raul Ordonez under the direction of Editor Miroslav Krstic.

E-mail addresses: sbhasin@ee.iitd.ac.in (S. Bhasin), rkamalapurkar@ufl.edu (R. Kamalapurkar), marc1518@ufl.edu (M. Johnson), kyriakos@ece.ucsb.edu (K.G. Vamvoudakis), lewis@uta.edu (F.L. Lewis), wdixon@ufl.edu (W.E. Dixon).

[1] Tel.: +91 11 26591124; fax: +91 11 26581606.

0005-1098/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2012.09.019

ensuring the algorithm is online and model-free. Early solutions to the problem consisted of using a discrete-time formulation of time and state, and then applying an RL algorithm on the discretized system. Discretizing the state space for high dimensional systems requires a large memory space and a computationally prohibitive learning process. Baird (1993) proposed *Advantage Updating*, an extension of the Q-learning algorithm which could be implemented in continuous-time and provided faster convergence. Doya (2000) used a *Hamilton–Jacobi–Bellman* (HJB) framework to derive algorithms for value function approximation and policy improvement, based on a continuous-time version of the temporal difference error. Murray, Cox, Lendaris, and Saeks (2002) also used the HJB framework to develop a *stepwise stable* iterative ADP algorithm for continuous-time input-affine systems with an input quadratic performance measure. In Beard, Saridis, and Wen (1997), Galerkin's spectral method is used to approximate the solution to the generalized HJB (GHJB), using which a stabilizing feedback controller was computed offline. Similar to Beard et al. (1997), Abu-Khalaf and Lewis (2005) proposed a least-squares successive approximation solution to the GHJB, where an NN is trained offline to learn the GHJB solution.

All of the aforementioned approaches for continuous-time nonlinear systems are offline and/or require complete knowledge of system dynamics. One of the contributions in Vrabie and Lewis (2009) is that only partial knowledge of the system dynamics is required, and a hybrid continuous-time/discrete-time sampled data controller is developed based on policy iteration (PI), where the feedback control operation of the actor occurs at a faster time scale than the learning process of the critic. Vamvoudakis and Lewis (2010) extended the idea by designing a model-based online algorithm called *synchronous PI* which involved synchronous, continuous-time adaptation of both actor and critic neural networks. Inspired by the work in Vamvoudakis and Lewis (2010), a novel actor–critic–identifier architecture is proposed in this paper to approximately solve the continuous-time infinite horizon optimal control problem for uncertain nonlinear systems; however, unlike Vamvoudakis and Lewis (2010), the developed method does not require knowledge of the system drift dynamics. The actor and critic NNs approximate the optimal control and the optimal value function, respectively, whereas the identifier dynamic neural network (DNN) estimates the system dynamics online. The integral RL technique in Vrabie and Lewis (2009) leads to a hybrid continuous-time/discrete-time controller with two time-scale actor–critic learning process, whereas the approach in Vamvoudakis and Lewis (2010), although continuous-time, requires complete knowledge of system dynamics. A contribution of this paper is the use of a novel actor–critic–identifier architecture, which obviates the need to know the system drift dynamics, and where the learning of the actor, critic and identifier is continuous and simultaneous. Moreover, the actor–critic–identifier method utilizes an identification-based online learning scheme, and hence is the first ever indirect adaptive control approach to RL. The idea is similar to the *Heuristic Dynamic Programming* (HDP) algorithm (Werbos, 1992), where Werbos suggested the use of a model network along with the actor and critic networks. Because of the generality of the considered system and objective function, the solution approach in this paper can be used in a wide range of applications in different fields, e.g., optimal control of space/air vehicles, chemical and manufacturing processes, robotics, financial systems, etc.

In the developed method, the actor and critic NNs use gradient and least-squares-based update laws, respectively, to minimize the Bellman error, which is the difference between the exact and the approximate HJB equation. The identifier DNN is a combination of a Hopfield-type (Hopfield, 1984) component, in parallel configuration with the system (Poznyak, Sanchez, & Yu, 2001), and a novel RISE (Robust Integral of Sign of the Error) component. The Hopfield component of the DNN learns the system dynamics based on online gradient-based weight tuning laws, while the RISE term robustly accounts for the function reconstruction errors, guaranteeing asymptotic estimation of the state and the state derivative. The online estimation of the state derivative allows the actor–critic–identifier architecture to be implemented without knowledge of system drift dynamics; however, knowledge of the input gain matrix is required to implement the control policy. While the design of the actor and critic are coupled through the HJB equation, the design of the identifier is decoupled from actor–critic, and can be considered as a modular component in the actor–critic–identifier architecture. Convergence of the actor–critic–identifier-based algorithm and stability of the closed-loop system are analyzed using Lyapunov-based adaptive control methods, and a *persistence of excitation* (PE) condition is used to guarantee exponential convergence to a bounded region in the neighborhood of the optimal control and uniformly ultimately bounded (UUB) stability of the closed-loop system. The PE condition is equivalent to the exploration paradigm in RL (Sutton & Barto, 1998) and ensures adequate sampling of the system's dynamics, required for convergence to the optimal policy.

2. Actor–critic–identifier architecture for HJB approximation

Consider a continuous-time nonlinear system

$$\dot{x} = F(x, u),$$

where $x(t) \in \mathcal{X} \subseteq \mathbb{R}^n$, $u(t) \in \mathcal{U} \subseteq \mathbb{R}^m$ is the control input, $F : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}^n$ is Lipschitz continuous on $\mathcal{X} \times \mathcal{U}$ containing the origin, such that the solution $x(t)$ of the system is unique for any finite initial condition x_0 and control $u \in \mathcal{U}$. The optimal value function can be defined as

$$V^*(x(t)) = \min_{\substack{u(\tau)\in\Psi(\mathcal{X})\\ t\le\tau<\infty}} \int_t^\infty r(x(s), u(x(s)))\, ds, \tag{1}$$

where $\Psi(\mathcal{X})$ is a set of admissible policies, and $r(x, u) \in \mathbb{R}$ is the immediate or local cost, defined as

$$r(x, u) = Q(x) + u^T Ru, \tag{2}$$

where $Q(x) \in \mathbb{R}$ is continuously differentiable and positive definite, and $R \in \mathbb{R}^{m\times m}$ is a positive-definite symmetric matrix. For the local cost in (2), which is convex in the control, and control-affine dynamics of the form

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

where $f(x) \in \mathbb{R}^n$ and $g(x) \in \mathbb{R}^{n\times m}$, the closed-form expression for optimal control is derived as Kirk (2004)

$$u^*(x) = -\frac{1}{2}R^{-1}g^T(x)\frac{\partial V^*(x)}{\partial x}^T, \tag{4}$$

where it is assumed that the value function $V^*(x)$ is continuously differentiable and satisfies $V^*(0) = 0$.

The Hamiltonian of the system in (3) is given by

$$H(x, u, V_x) \triangleq V_x F_u + r_u,$$

where $V_x \triangleq \frac{\partial V}{\partial x} \in \mathbb{R}^{1\times n}$ denotes the gradient of the value function $V(x)$, $F_u(x, u) \triangleq f(x) + g(x)u \in \mathbb{R}^n$ denotes the system dynamics with control $u(x)$, and $r_u \triangleq r(x, u)$ denotes the local cost with control $u(x)$. The optimal value function $V^*(x)$ in (1) and the associated optimal policy $u^*(x)$ in (4) satisfy the HJB equation

$$H(x, u^*, V_x^*) = V_x^* F_{u^*} + r_{u^*} = 0. \tag{5}$$

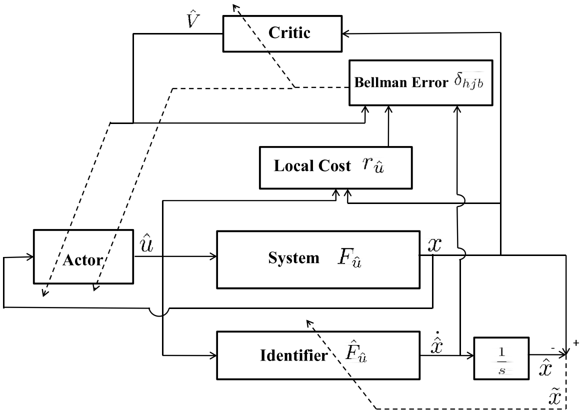

Fig. 1. Actor–critic–identifier architecture to approximate the HJB.

Replacing $u^*(x)$, $V_x^*(x)$, and $F_{u^*}(x, u^*)$ in (5) by their approximations, $\hat{u}(x)$ (actor), $\hat{V}(x)$ (critic), and $\hat{F}_{\hat{u}}(x, \hat{x}, \hat{u})$ (identifier), respectively, the approximate HJB equation is given by

$$\hat{H}(x, \hat{x}, \hat{u}, \hat{V}_x) = \hat{V}_x \hat{F}_{\hat{u}} + r_{\hat{u}}, \tag{6}$$

where $\hat{x}(t)$ is the state of the identifier, and $\hat{H}(\cdot)$ is the approximate Hamiltonian. Using (5) and (6), the error between the actual and the approximate HJB equation is given by the Bellman residual error $\delta_{hjb}(x, \hat{x}, \hat{u}, \hat{V}_x)$, defined as

$$\delta_{hjb} \triangleq \hat{H}(x, \hat{x}, \hat{u}, \hat{V}_x) - H(x, u^*, V_x^*). \tag{7}$$

Since $H(x, u^*, V_x^*) \equiv 0$, the Bellman error can be written in a measurable form as

$$\delta_{hjb} = \hat{H}(x, \hat{x}, \hat{u}, \hat{V}_x) = \hat{V}_x \hat{F}_{\hat{u}} + r(x, \hat{u}). \tag{8}$$

The actor and critic learn based on the Bellman error $\delta_{hjb}(\cdot)$, whereas the identifier estimates the system dynamics online using the identification error $\tilde{x}(t) \triangleq x(t) - \hat{x}(t)$, and hence is decoupled from the actor–critic design. The block diagram of the ACI architecture is shown in Fig. 1.

The following assumptions are made about the control-affine system in (3).

Assumption 1. The functions $f(x)$ and $g(x)$ are second-order differentiable.

Assumption 2. The input gain matrix $g(x)$ is known and bounded, i.e. $0 < \|g(x)\| \leq \bar{g}$, where $\bar{g}$ is a known positive constant.

Assuming the optimal control, the optimal value function and the system dynamics are continuous and defined on compact sets, NNs can be used to approximate them (Cybenko, 1989; Hornik, Stinchcombe, & White, 1985). Some standard NN assumptions which will be used throughout the paper are:

Assumption 3. Given a continuous function $\Upsilon : \mathbb{S} \to \mathbb{R}^n$, where $\mathbb{S}$ is a compact simply connected set, there exists ideal weights W, V such that the function can be represented by a NN as

$$\Upsilon(x) = W^T \sigma(V^T x) + \varepsilon(x),$$

where $\sigma(\cdot)$ is the nonlinear activation function, and $\varepsilon(x)$ is the function reconstruction error.

Assumption 4. The ideal NN weights are bounded by known positive constants, i.e. $\|W\| \leq \bar{W}$, $\|V\| \leq \bar{V}$ (Lewis, Selmic, & Campos, 2002).

Assumption 5. The NN activation function $\sigma(\cdot)$ and its derivative with respect to its arguments, $\sigma'(\cdot)$, are bounded.

Assumption 6. Using the NN universal approximation property (Cybenko, 1989; Hornik et al., 1985), the function reconstruction errors and its derivative with respect to its arguments are bounded (Lewis et al., 2002) as $\|\varepsilon(\cdot)\| \leq \bar{\varepsilon}$, $\|\varepsilon'(\cdot)\| \leq \bar{\varepsilon}'$.

3. Actor–critic design

Using Assumption 3 and (4), the optimal value function and the optimal control can be represented by NNs as

$$V^*(x) = W^T \phi(x) + \varepsilon_v(x),$$
$$u^*(x) = -\frac{1}{2} R^{-1} g^T(x) (\phi'(x)^T W + \varepsilon_v'(x)^T), \tag{9}$$

where $W \in \mathbb{R}^N$ are unknown ideal NN weights, N is the number of neurons, $\phi(x) \triangleq [\phi_1(x)\phi_2(x) \cdots \phi_N(x)]^T \in \mathbb{R}^N$ and $\phi'(x) \triangleq \frac{\partial \phi}{\partial x} \in \mathbb{R}^{N \times n}$, such that $\phi_i(0) = 0$ and $\phi_i'(0) = 0$ $\forall i = 1 \ldots N$, and $\varepsilon_v(\cdot) \in \mathbb{R}$ is the function reconstruction error.

Assumption 7. The NN activation functions $\{\phi_i(x) : i = 1 \ldots N\}$ are selected so that as $N \to \infty$, $\phi(x)$ provides a complete independent basis for $V^*(x)$.

Using Assumption 7 and the Weierstrass higher-order approximation theorem, both $V^*(x)$ and $\frac{\partial V^*(x)}{\partial x}$ can be uniformly approximated by NNs in (9), i.e. as $N \to \infty$, the approximation errors $\varepsilon_v(x), \varepsilon_v'(x) \to 0$ (Abu-Khalaf & Lewis, 2005). The critic $\hat{V}(x)$ and the actor $\hat{u}(x)$ approximate the optimal value function and the optimal control in (9), and are given by

$$\hat{V}(x) = \hat{W}_c^T \phi(x); \qquad \hat{u}(x) = -\frac{1}{2} R^{-1} g^T(x) \phi'^T(x) \hat{W}_a, \tag{10}$$

where $\hat{W}_c(t) \in \mathbb{R}^N$ and $\hat{W}_a(t) \in \mathbb{R}^N$ are estimates of the ideal weights of the critic and actor NNs, respectively. The weight estimation errors for the critic and actor NNs are defined as $\tilde{W}_c(t) \triangleq W - \hat{W}_c(t) \in \mathbb{R}^N$ and $\tilde{W}_a(t) \triangleq W - \hat{W}_a(t) \in \mathbb{R}^N$, respectively.

Remark 1. Since the optimal control is determined using the gradient of the optimal value function in (9), the critic NN in (10) may be used to determine the actor without using another NN for the actor. However, for ease in deriving weight update laws and subsequent stability analysis, separate NNs are used for the actor and the critic (Vamvoudakis & Lewis, 2010).

The actor and critic NN weights are both updated based on the minimization of the Bellman error $\delta_{hjb}(\cdot)$ in (8), which can be rewritten by substituting $\hat{V}(x)$ from (10) as

$$\delta_{hjb} = \hat{W}_c^T \omega + r(x, \hat{u}), \tag{11}$$

where $\omega(x, \hat{x}, \hat{u}) \triangleq \phi'(x) \hat{F}_{\hat{u}}(x, \hat{x}, \hat{u}) \in \mathbb{R}^N$ is the critic NN regressor vector.

3.1. *Least-squares update for the critic*

Let $E_c(\delta_{hjb}) \in \mathbb{R}^+$ denote the integral squared Bellman error as

$$E_c = \int_0^t \delta_{hjb}^2(\tau) d\tau. \tag{12}$$

The least-squares (LS) update law for the critic is generated by minimizing (12) as

$$\frac{\partial E_c}{\partial \hat{W}_c} = 2 \int_0^t \delta_{hjb}(\tau) \frac{\partial \delta_{hjb}(\tau)}{\partial \hat{W}_c(t)} d\tau = 0. \tag{13}$$

Using $\frac{\partial \delta_{hjb}}{\partial \hat{W}_c} = \omega^T$ from (11), the batch LS critic weight estimate is determined from (13) as Sastry and Bodson (1989)

$$\hat{W}_c(t) = -\left(\int_0^t \omega(\tau)\omega(\tau)^T d\tau\right)^{-1} \int_0^t \omega(\tau) r(\tau) d\tau, \tag{14}$$

provided the inverse $\left(\int_0^t \omega(\tau)\omega(\tau)^T d\tau\right)^{-1}$ exists. For online implementation, a normalized recursive formulation of the LS algorithm is developed by taking the time derivative (14) and normalizing as Sastry and Bodson (1989)

$$\dot{\hat{W}}_c = -\eta_c \Gamma \frac{\omega}{1 + \nu\omega^T \Gamma \omega} \delta_{hjb}, \tag{15}$$

where $\nu, \eta_c \in \mathbb{R}$ are constant positive gains, and $\Gamma(t) \triangleq \left(\int_0^t \omega(\tau) \omega(\tau)^T d\tau\right)^{-1} \in \mathbb{R}^{N \times N}$ is a symmetric estimation gain matrix generated as

$$\dot{\Gamma} = -\eta_c \Gamma \frac{\omega\omega^T}{1 + \nu\omega^T \Gamma \omega} \Gamma; \qquad \Gamma(t_r^+) = \Gamma(0) = \varphi_0 I, \tag{16}$$

where t_r^+ is the resetting time at which $\lambda_{min}\{\Gamma(t)\} \leq \varphi_1, \varphi_0 > \varphi_1 > 0$. The covariance resetting ensures that $\Gamma(t)$ is positive-definite for all time and prevents its value from becoming arbitrarily small in some directions, thus avoiding slow adaptation in some directions (also called the covariance wind-up problem) (Sastry & Bodson, 1989). From (16), it is clear that $\dot{\Gamma} \leq 0$, which means that the covariance matrix $\Gamma(t)$ can be bounded as

$$\varphi_1 I \leq \Gamma(t) \leq \varphi_0 I. \tag{17}$$

3.2. *Gradient update for the actor*

The actor update, like the critic update in Section 3.1, is based on the minimization of the Bellman error $\delta_{hjb}(\cdot)$. However, unlike the critic weights, the actor weights appear nonlinearly in $\delta_{hjb}(\cdot)$, making it problematic to develop a LS update law. Hence, a gradient update law is developed for the actor which minimizes the squared Bellman error $E_a(t) \triangleq \delta_{hjb}^2$, whose gradient is given by

$$\begin{aligned}\frac{\partial E_a}{\partial \hat{W}_a} &= 2\frac{\partial \delta_{hjb}}{\partial \hat{W}_a}\delta_{hjb} \\ &= 2\left(\hat{W}_c^T \phi' \frac{\partial \hat{F}_{\hat{u}}}{\partial \hat{u}} \frac{\partial \hat{u}}{\partial \hat{W}_a} + 2\hat{u}^T R \frac{\partial \hat{u}}{\partial \hat{W}_a}\right)\delta_{hjb},\end{aligned} \tag{18}$$

where (11) is used, and $G(x) \triangleq g(x)R^{-1}g(x)^T \in \mathbb{R}^{n\times n}$ is a symmetric matrix. Using (18), the gradient-based update law for the actor NN is given by

$$\begin{aligned}\dot{\hat{W}}_a = proj\Bigg\{&-\frac{2\eta_{a1}}{\sqrt{1+\omega^T\omega}}\left(\hat{W}_c^T \phi' \frac{\partial \hat{F}_{\hat{u}}}{\partial \hat{u}} \frac{\partial \hat{u}}{\partial \hat{W}_a}\right)^T \delta_{hjb} \\ &- \frac{4\eta_{a1}}{\sqrt{1+\omega^T\omega}} \frac{\partial \hat{u}}{\partial \hat{W}_a}^T R\hat{u}\delta_{hjb} - \eta_{a2}(\hat{W}_a - \hat{W}_c)\Bigg\}\end{aligned} \tag{19}$$

where $proj\{\cdot\}$ is a smooth projection operator used to bound the weight estimates (Dixon, Behal, Dawson, & Nagarkatti, 2003; Krstic, Kokotovic, & Kanellakopoulos, 1995), $\eta_{a1}, \eta_{a2} \in \mathbb{R}$ are positive adaptation gains, $\frac{1}{\sqrt{1+\omega^T\omega}}$ is the normalization term, and the last term in (19) is added for stability (based on the subsequent stability analysis).

Remark 2. A recursive least-squares update law with covariance resetting is developed for the critic in (15), which exploits the fact that the critic weights appear linearly in the Bellman error $\delta_{hjb}(\cdot)$. This is in contrast to the modified Levenberg–Marquardt algorithm in Vamvoudakis and Lewis (2010) which is similar to the normalized gradient update law. The actor update law in (19) also differs in the sense that the update law in Vamvoudakis and Lewis (2010) is purely motivated by the stability analysis whereas the proposed actor update law is based on the minimization of the Bellman error with an additional term for stability. Heuristically, these differences in the update law development could lead to improved performance in terms of faster convergence of the actor and critic weights, as seen from the simulation results in Section 6.

4. Identifier design

The following assumption is made for the identifier design:

Assumption 8. The control input is bounded, i.e. $u(t) \in \mathcal{L}_\infty$. Using Assumptions 2 and 5 and the projection algorithm in (19), this assumption holds for the control design $u(t) = \hat{u}(x)$ in (10).

Using Assumption 3, the dynamic system in (3), with control $\hat{u}(x)$, can be represented using a multi-layer NN as

$$\dot{x} = F_{\hat{u}}(x, \hat{u}) = W_f^T \sigma(V_f^T x) + \varepsilon_f(x) + g(x)\hat{u}, \tag{20}$$

where $W_f \in \mathbb{R}^{L_f+1\times n}, V_f \in \mathbb{R}^{n\times L_f}$ are the unknown ideal NN weights, $\sigma_f \triangleq \sigma(V_f^T x) \in \mathbb{R}^{L_f+1}$ is the NN activation function, and $\varepsilon_f(x) \in \mathbb{R}^n$ is the function reconstruction error. The following multi-layer dynamic neural network (MLDNN) identifier is used to approximate the system in (20)

$$\dot{\hat{x}} = \hat{F}_{\hat{u}}(x, \hat{x}, \hat{u}) = \hat{W}_f^T \hat{\sigma}_f + g(x)\hat{u} + \mu, \tag{21}$$

where $\hat{x}(t) \in \mathbb{R}^n$ is the DNN state, $\hat{\sigma}_f \triangleq \sigma(\hat{V}_f^T \hat{x}) \in \mathbb{R}^{L_f+1}$, $\hat{W}_f(t) \in \mathbb{R}^{L_f+1\times n}$ and $\hat{V}_f(t) \in \mathbb{R}^{n\times L_f}$ are weight estimates, and $\mu(t) \in \mathbb{R}^n$ denotes the RISE feedback term defined as Patre, MacKunis, Kaiser, and Dixon (2008) and Xian, Dawson, de Queiroz, and Chen (2004)

$$\mu \triangleq k\tilde{x}(t) - k\tilde{x}(0) + v, \tag{22}$$

where $\tilde{x}(t) \triangleq x(t) - \hat{x}(t) \in \mathbb{R}^n$ is the identification error, and $v(t) \in \mathbb{R}^n$ is the Filippov generalized solution (Filippov, 1988) to the differential equation

$$\dot{v} = (k\alpha + \gamma)\tilde{x} + \beta_1 sgn(\tilde{x}); \qquad v(0) = 0,$$

where $k, \alpha, \gamma, \beta_1 \in \mathbb{R}$ are positive constant control gains, and $sgn(\cdot)$ denotes a vector signum function. The identification error dynamics can be written as

$$\dot{\tilde{x}} = \tilde{F}_u(x, \hat{x}, u) = W_f^T \sigma_f - \hat{W}_f^T \hat{\sigma}_f + \varepsilon_f(x) - \mu, \tag{23}$$

where $\tilde{F}_{\hat{u}}(x, \hat{x}, \hat{u}) \triangleq F_{\hat{u}}(x, \hat{u}) - \hat{F}_{\hat{u}}(x, \hat{x}, \hat{u}) \in \mathbb{R}^n$. A filtered identification error is defined as

$$e_f \triangleq \dot{\tilde{x}} + \alpha\tilde{x}. \tag{24}$$

Taking the time derivative of (24) and using (23) yields

$$\begin{aligned}\dot{e}_f &= W_f^T \sigma_f' V_f^T \dot{x} - \dot{\hat{W}}_f^T \hat{\sigma}_f - \hat{W}_f^T \hat{\sigma}_f' \dot{\hat{V}}_f^T \hat{x} - \hat{W}_f^T \hat{\sigma}_f' \hat{V}_f^T \dot{\hat{x}} \\ &\quad + \dot{\varepsilon}_f(x) - ke_f - \gamma\tilde{x} - \beta_1 sgn(\tilde{x}) + \alpha\dot{\tilde{x}}.\end{aligned} \tag{25}$$

Based on (25) and the subsequent stability analysis, the weight update laws for the DNN are designed as

$$\begin{aligned}\dot{\hat{W}}_f &= proj(\Gamma_{wf}\hat{\sigma}_f' \hat{V}_f^T \dot{\hat{x}}\tilde{x}^T), \\ \dot{\hat{V}}_f &= proj(\Gamma_{vf}\dot{\hat{x}}\tilde{x}^T \hat{W}_f^T \hat{\sigma}_f'),\end{aligned} \tag{26}$$

where $\Gamma_{wf} \in \mathbb{R}^{L_f+1\times L_f+1}$, $\Gamma_{vf} \in \mathbb{R}^{n\times n}$ are positive constant adaptation gain matrices. The expression in (25) can be rewritten as

$$\dot{e}_f = \tilde{N} + N_{B1} + \hat{N}_{B2} - ke_f - \gamma\tilde{x} - \beta_1 sgn(\tilde{x}), \tag{27}$$

where the auxiliary signals, $\tilde{N}(x, \tilde{x}, e_f, \hat{W}_f, \hat{V}_f, t)$, $N_{B1}(x, \hat{x}, \hat{W}_f, \hat{V}_f, t)$, and $\hat{N}_{B2}(\hat{x}, \dot{\hat{x}}, \hat{W}_f, \hat{V}_f, t) \in \mathbb{R}^n$ are defined as

$$\begin{aligned}\tilde{N} &\triangleq \alpha\dot{\tilde{x}} - \dot{\hat{W}}_f^T\hat{\sigma}_f - \hat{W}_f^T\hat{\sigma}_f'\dot{\hat{V}}_f^T\hat{x} + \frac{1}{2}W_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{\tilde{x}} \\ &\quad + \frac{1}{2}\hat{W}_f^T\hat{\sigma}_f'V_f^T\dot{\tilde{x}},\end{aligned} \tag{28}$$

$$N_{B1} \triangleq W_f^T\sigma_f'V_f^T\dot{x} - \frac{1}{2}W_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{x} - \frac{1}{2}\hat{W}_f^T\hat{\sigma}_f'V_f^T\dot{x} + \dot{\varepsilon}_f(x), \tag{29}$$

$$\hat{N}_{B2} \triangleq \frac{1}{2}\tilde{W}_f^T\hat{\sigma}_f'\hat{V}_f^T\dot{\hat{x}} + \frac{1}{2}\hat{W}_f^T\hat{\sigma}_f'\tilde{V}_f^T\dot{\hat{x}}, \tag{30}$$

where $\tilde{W}_f \triangleq W_f - \hat{W}_f(t) \in \mathbb{R}^{L_f+1\times n}$ and $\tilde{V}_f \triangleq V_f - \hat{V}_f(t) \in \mathbb{R}^{n\times L_f}$. To facilitate the subsequent stability analysis, an auxiliary term $N_{B2}(\hat{x}, \dot{x}, \hat{W}_f, \hat{V}_f, t) \in \mathbb{R}^n$ is defined by replacing $\dot{\hat{x}}(t)$ in $\hat{N}_{B2}(\cdot)$ by $\dot{x}(t)$, and $\tilde{N}_{B2}(\hat{x}, \dot{\tilde{x}}, \hat{W}_f, \hat{V}_f, t) \triangleq \hat{N}_{B2}(\cdot) - N_{B2}(\cdot)$. The terms $N_{B1}(\cdot)$ and $N_{B2}(\cdot)$ are grouped as $N_B \triangleq N_{B1} + N_{B2}$. Using Assumptions 2 and 4–6, and (24), (26), (29) and (30), the following bounds can be obtained

$$\left\|\tilde{N}\right\| \le \rho_1(\|z\|)\|z\|, \tag{31}$$

$$\|N_{B1}\| \le \zeta_1, \qquad \|N_{B2}\| \le \zeta_2,$$

$$\left\|\dot{N}_B\right\| \le \zeta_3 + \zeta_4\rho_2(\|z\|)\|z\|, \tag{32}$$

$$\left\|\dot{\tilde{x}}^T\tilde{N}_{B2}\right\| \le \zeta_5\left\|\tilde{x}\right\|^2 + \zeta_6\left\|e_f\right\|^2, \tag{33}$$

where $z \triangleq \left[\tilde{x}^T e_f^T\right]^T \in \mathbb{R}^{2n}$, $\rho_1(\cdot), \rho_2(\cdot) \in \mathbb{R}$ are positive, globally invertible, non-decreasing functions, and $\zeta_i \in \mathbb{R}, i = 1, \ldots, 6$ are computable positive constants. To facilitate the subsequent stability analysis, let $\mathcal{D} \subset \mathbb{R}^{2n+2}$ be a domain containing $y(t) = 0$, where $y(t) \in \mathbb{R}^{2n+2}$ is defined as

$$y \triangleq \left[\tilde{x}^T\ e_f^T\ \sqrt{P}\ \sqrt{Q}\right]^T, \tag{34}$$

where the auxiliary function $P(z, t) \in \mathbb{R}$ is the Filippov generalized solution (Filippov, 1988) to the differential equation

$$\dot{P} = -L, \qquad P(0) = \beta_1\sum_{i=1}^{n}\left|\tilde{x}_i(0)\right| - \tilde{x}^T(0)N_B(0), \tag{35}$$

where the auxiliary function $L(z, t) \in \mathbb{R}$ is defined as

$$L \triangleq e_f^T(N_{B1} - \beta_1 sgn(\tilde{x})) + \dot{\tilde{x}}^T N_{B2} - \beta_2\rho_2(\|z\|)\|z\|\left\|\tilde{x}\right\|, \tag{36}$$

where $\beta_1, \beta_2 \in \mathbb{R}$ are chosen according to the following sufficient conditions to ensure $P(t) \ge 0$ (Patre et al., 2008)

$$\beta_1 > \max\left(\zeta_1 + \zeta_2, \zeta_1 + \frac{\zeta_3}{\alpha}\right), \qquad \beta_2 > \zeta_4. \tag{37}$$

The auxiliary function $Q(\tilde{W}_f, \tilde{V}_f) \in \mathbb{R}$ in (34) is defined as

$$Q \triangleq \frac{1}{4}\alpha\left[tr(\tilde{W}_f^T\Gamma_{wf}^{-1}\tilde{W}_f) + tr(\tilde{V}_f^T\Gamma_{vf}^{-1}\tilde{V}_f)\right],$$

where $tr(\cdot)$ denotes the trace of a matrix.

Theorem 1. *For the system in (3), the identifier developed in (21) along with the weight update laws in (26) ensures asymptotic identification of the state and its derivative, in the sense that*

$$\lim_{t\to\infty}\left\|\tilde{x}(t)\right\| = 0 \quad and \quad \lim_{t\to\infty}\left\|\dot{\tilde{x}}(t)\right\| = 0,$$

provided the control gains k and γ are chosen sufficiently large based on the initial conditions of the states[2] *and satisfy the following sufficient conditions*

$$\gamma > \frac{\zeta_5}{\alpha}, \qquad k > \zeta_6, \tag{38}$$

where ζ_5 and ζ_6 are introduced in (33), and β_1, β_2 introduced in (36), are chosen according to the sufficient conditions in (37).

Proof. See the Appendix. □

Using the developed identifier in (21), the actor weight update law can now be simplified using (19) as

$$\begin{aligned}\dot{\hat{W}}_a = proj\Bigg\{&-\frac{\eta_{a1}}{\sqrt{1+\omega^T\omega}}\phi'G\phi'^T\left(\hat{W}_a - \hat{W}_c\right)\delta_{hjb} \\ &- \eta_{a2}(\hat{W}_a - \hat{W}_c)\Bigg\}.\end{aligned} \tag{39}$$

5. Convergence and stability analysis

The unmeasurable form of the Bellman error can be written using (5)–(8) and (11), as

$$\begin{aligned}\delta_{hjb} &= \hat{W}_c^T\omega - W_c^T\phi'F_{u*} + \hat{u}^TR\hat{u} - u^{*^T}Ru^* - \varepsilon_v'F_{u*}\cdot \\ &= -\tilde{W}_c^T\omega - W^T\phi'\tilde{F}_{\hat{u}} + \frac{1}{4}\tilde{W}_a^T\phi'G\phi'^T\tilde{W}_a \\ &\quad - \frac{1}{4}\varepsilon_v'G\varepsilon_v'^T - \varepsilon_v'F_{u*},\end{aligned} \tag{40}$$

where (9) and (10) are used. The dynamics of the critic weight estimation error $\tilde{W}_c(t)$ can now be developed by substituting (40) in (15), as

$$\begin{aligned}\dot{\tilde{W}}_c &= -\eta_c\Gamma\psi\psi^T\tilde{W}_c + \eta_c\Gamma\frac{\omega}{1+\nu\omega^T\Gamma\omega}\Bigg[-W^T\phi'\tilde{F}_{\hat{u}} \\ &\quad + \frac{1}{4}\tilde{W}_a^T\phi'G\phi'^T\tilde{W}_a - \frac{1}{4}\varepsilon_v'G\varepsilon_v'^T - \varepsilon_v'F_{u*}\Bigg],\end{aligned} \tag{41}$$

where $\psi(t) \triangleq \frac{\omega(t)}{\sqrt{1+\nu\omega(t)^T\Gamma(t)\omega(t)}} \in \mathbb{R}^N$ is the normalized critic regressor vector, bounded as

$$\|\psi\| \le \frac{1}{\sqrt{\nu\varphi_1}}, \tag{42}$$

where φ_1 is introduced in (17). The error system in (41) can be represented by the following perturbed system

$$\dot{\tilde{W}}_c = \Omega_{nom} + \Delta_{per}, \tag{43}$$

where $\Omega_{nom}(\tilde{W}_c, t) \triangleq -\eta_c\Gamma\psi\psi^T\tilde{W}_c \in \mathbb{R}^N$ denotes the nominal system, and $\Delta_{per}(t) \triangleq \eta_c\Gamma\frac{\omega}{1+\nu\omega^T\Gamma\omega}\left[-W^T\phi'\tilde{F}_{\hat{u}} + \frac{1}{4}\tilde{W}_a^T\phi'G\phi'^T\tilde{W}_a - \frac{1}{4}\varepsilon_v'G\varepsilon_v'^T - \varepsilon_v'F_{u*}\right] \in \mathbb{R}^N$ denotes the perturbation. Using Theorem 2.5.1 in Sastry and Bodson (1989), the nominal system

$$\dot{\tilde{W}}_c = -\eta_c\Gamma\psi\psi^T\tilde{W}_c \tag{44}$$

is globally exponentially stable, if the bounded signal $\psi(t)$ is PE, i.e.

$$\mu_2 I \ge \int_{t_0}^{t_0+\delta}\psi(\tau)\psi(\tau)^T d\tau \ge \mu_1 I \quad \forall t_0 \ge 0,$$

[2] See subsequent semi-global stability analysis.

for some positive constants $\mu_1, \mu_2, \delta \in \mathbb{R}$. Since $\Omega_{nom}(\tilde{W}_c, t)$ is continuously differentiable and the Jacobian $\frac{\partial \Omega_{nom}}{\partial \tilde{W}_c} = -\eta_c \Gamma \psi \psi^T$ is bounded for the exponentially stable system in (44), the converse Lyapunov Theorem 4.14 in Khalil (2002) can be used to show that there exists a function $V_c : \mathbb{R}^N \times [0, \infty) \to \mathbb{R}$, which satisfies the following inequalities

$$c_1 \left\|\tilde{W}_c\right\|^2 \le V_c(\tilde{W}_c, t) \le c_2 \left\|\tilde{W}_c\right\|^2$$

$$\frac{\partial V_c}{\partial t} + \frac{\partial V_c}{\partial \tilde{W}_c} \Omega_{nom}(\tilde{W}_c, t) \le -c_3 \left\|\tilde{W}_c\right\|^2 \tag{45}$$

$$\left\|\frac{\partial V_c}{\partial \tilde{W}_c}\right\| \le c_4 \left\|\tilde{W}_c\right\|,$$

for some positive constants $c_1, c_2, c_3, c_4 \in \mathbb{R}$. Using Assumptions 2, 4–6 and 8, the projection bounds in (19), the fact that $F_{u^*} \in \mathcal{L}_\infty$ (using (4), Assumptions 2–6, and (9)), and provided the conditions of Theorem 1 hold (required to prove that $\tilde{F}_{\hat{u}} \in \mathcal{L}_\infty$), the following bounds can be developed:

$$\left\|\tilde{W}_a\right\| \le \kappa_1, \qquad \left\|\phi' G \phi'^T\right\| \le \kappa_2,$$

$$\left\|\frac{1}{4}\tilde{W}_a^T \phi' G \phi'^T \tilde{W}_a - \frac{1}{4}\varepsilon'_v G \varepsilon_v'^T - W^T \phi' \tilde{F}_{\hat{u}} - \varepsilon'_v F_{u^*}\right\| \le \kappa_3,$$

$$\left\|\frac{1}{2}W^T \phi' G \varepsilon_v'^T + \frac{1}{2}\varepsilon'_v G \varepsilon_v'^T + \frac{1}{2}W^T \phi' G \phi'^T \tilde{W}_a + \frac{1}{2}\varepsilon'_v G \phi'^T\right\| \le \kappa_4, \tag{46}$$

where $\kappa_1, \kappa_2, \kappa_3, \kappa_4 \in \mathbb{R}$ are computable positive constants.

Theorem 2. *If Assumptions 1–8 hold, the regressor* $\psi(t) \triangleq \frac{\omega}{\sqrt{1+\omega^T \Gamma \omega}}$ *is PE (persistently exciting), and provided (37), (38) and the following sufficient gain condition is satisfied*[3]

$$\frac{c_3}{\eta_{a1}} > \kappa_1 \kappa_2, \tag{47}$$

where $\eta_{a1}, c_3, \kappa_1, \kappa_2$ *are introduced in (19), (45) and (46), then the controller in (10), the actor–critic weight update laws in (15), (16) and (39), and the identifier in (21) and (26) guarantee that the state of the system* $x(t)$*, and the actor–critic weight estimation errors* $\tilde{W}_a(t)$ *and* $\tilde{W}_c(t)$ *are UUB.*

Proof. To investigate the stability of (3) with control $\hat{u}(x)$, and the perturbed system in (43), consider $V_L : \mathcal{X} \times \mathbb{R}^N \times \mathbb{R}^N \times [0, \infty) \to \mathbb{R}$ as the continuously differentiable, positive-definite Lyapunov function candidate defined as

$$V_L(x, \tilde{W}_c, \tilde{W}_a, t) \triangleq V^*(x) + V_c(\tilde{W}_c, t) + \frac{1}{2}\tilde{W}_a^T \tilde{W}_a,$$

where $V^*(x)$ (the optimal value function), is the Lyapunov function for (3), and $V_c(\tilde{W}_c, t)$ is the Lyapunov function for the exponentially stable system in (44). Since $V^*(x)$ is continuously differentiable and positive-definite from (1) and (2), there exist class $\mathcal{K}$ functions α_1 and α_2 defined on $[0, a]$, where $B_a \subset \mathcal{X}$ (see Lemma 4.3 in Khalil, 2002), such that

$$\alpha_1(\|x\|) \le V^*(x) \le \alpha_2(\|x\|) \quad \forall x \in B_a. \tag{48}$$

Using (45) and (48), $V_L(x, \tilde{W}_c, \tilde{W}_a, t)$ can be bounded as

$$\alpha_1(\|x\|) + c_1 \left\|\tilde{W}_c\right\|^2 + \frac{1}{2}\left\|\tilde{W}_a\right\|^2 \le V_L(x, \tilde{W}_c, \tilde{W}_a, t)$$
$$\le \alpha_2(\|x\|) + c_2 \left\|\tilde{W}_c\right\|^2 + \frac{1}{2}\left\|\tilde{W}_a\right\|^2,$$

which can be written as

$$\alpha_3(\left\|\tilde{z}\right\|) \le V_L(x, \tilde{W}_c, \tilde{W}_a, t) \le \alpha_4(\left\|\tilde{z}\right\|) \qquad \forall \tilde{z} \in B_s,$$

where $\tilde{z}(t) \triangleq [x(t)^T \tilde{W}_c(t)^T \tilde{W}_a(t)^T]^T \in \mathbb{R}^{n+2N}$, α_3 and α_4 are class $\mathcal{K}$ functions defined on $[0, s]$, where $B_s \subset \mathcal{X} \times \mathbb{R}^N \times \mathbb{R}^N$. Taking the time derivative of $V_L(\cdot)$ yields

$$\dot{V}_L = \frac{\partial V^*}{\partial x} f + \frac{\partial V^*}{\partial x} g \hat{u} + \frac{\partial V_c}{\partial t} + \frac{\partial V_c}{\partial \tilde{W}_c}\Omega_{nom} + \frac{\partial V_c}{\partial \tilde{W}_c}\Delta_{per} - \tilde{W}_a^T \dot{\hat{W}}_a, \tag{49}$$

where the time derivative of $V^*(\cdot)$ is taken along the trajectories of the system (3) with control $\hat{u}(\cdot)$ and the time derivative of $V_c(\cdot)$ is taken along the trajectories of the perturbed system (43). To facilitate the subsequent analysis, the HJB in (5) is rewritten as $\frac{\partial V^*}{\partial x} f = -\frac{\partial V^*}{\partial x} g u^* - Q(x) - u^{*T} R u^*$. Substituting for $\frac{\partial V^*}{\partial x} f$ in (49), using the fact that $\frac{\partial V^*}{\partial x} g = -2u^{*T} R$ from (4), and using (19) and (45), (49) can be upper bounded as

$$\dot{V}_L \le -Q - u^{*T} R u^* - c_3 \left\|\tilde{W}_c\right\|^2 + c_4 \left\|\tilde{W}_c\right\| \left\|\Delta_{per}\right\| + 2u^{*T} R(u^* - \hat{u}) + \eta_{a2} \tilde{W}_a^T (\hat{W}_a - \hat{W}_c) + \frac{\eta_{a1}}{\sqrt{1+\omega^T \omega}} \tilde{W}_a^T \phi' G \phi'^T (\hat{W}_a - \hat{W}_c) \delta_{hjb}. \tag{50}$$

Substituting for u^*, $\hat{u}$, δ_{hjb}, and Δ_{per} using (4), (10), (40) and (43), respectively, and using (17) and (42) in (50), yields

$$\begin{aligned}
\dot{V}_L \le\ & -Q - c_3 \left\|\tilde{W}_c\right\|^2 - \eta_{a2} \left\|\tilde{W}_a\right\|^2 + \frac{1}{2}W^T \phi' G \varepsilon_v'^T \\
& + \frac{1}{2}\varepsilon'_v G \varepsilon_v'^T + \frac{1}{2}W^T \phi' G \phi'^T \tilde{W}_a + \frac{1}{2}\varepsilon'_v G \phi'^T \tilde{W}_a \\
& + c_4 \frac{\eta_c \varphi_0}{2\sqrt{\nu \varphi_1}} \left\| -W^T \phi' \tilde{F}_{\hat{u}} + \frac{1}{4}\tilde{W}_a^T \phi' G \phi'^T \tilde{W}_a - \frac{1}{4}\varepsilon'_v G \varepsilon_v'^T - \varepsilon'_v F_{u^*}\right\| \left\|\tilde{W}_c\right\| + \eta_{a2} \left\|\tilde{W}_a\right\| \left\|\tilde{W}_c\right\| \\
& + \frac{\eta_{a1}}{\sqrt{1+\omega^T \omega}} \tilde{W}_a^T \phi' G \phi'^T (\tilde{W}_c - \tilde{W}_a) \Big(- \tilde{W}_c^T \omega \\
& - W^T \phi' \tilde{F}_{\hat{u}} + \frac{1}{4}\tilde{W}_a^T \phi' G \phi'^T \tilde{W}_a - \frac{1}{4}\varepsilon'_v G \varepsilon_v'^T - \varepsilon'_v F_{u^*} \Big).
\end{aligned} \tag{51}$$

Using the bounds developed in (46), (51) can be further upper bounded as

$$\begin{aligned}
\dot{V}_L \le\ & -Q - (c_3 - \eta_{a1}\kappa_1\kappa_2) \left\|\tilde{W}_c\right\|^2 - \eta_{a2} \left\|\tilde{W}_a\right\|^2 \\
& + \left(\frac{c_4 \eta_c \varphi_0}{2\sqrt{\nu \varphi_1}}\kappa_3 + \eta_{a1}\kappa_1\kappa_2\kappa_3 + \eta_{a1}\kappa_1^2\kappa_2 + \eta_{a2}\kappa_1\right) \left\|\tilde{W}_c\right\| \\
& + \eta_{a1}\kappa_1^2\kappa_2\kappa_3 + \kappa_4.
\end{aligned}$$

[3] Since c_3 is a function of the critic adaptation gain η_c, η_{a1} is the actor adaptation gain, and κ_1, κ_2 are known constants, the sufficient gain condition in (47) can be easily satisfied.

Provided $c_3 > \eta_{a1}\kappa_1\kappa_2$, and completing the square yields

$$\begin{aligned}\dot{V}_L \leq{}& -Q - (1-\theta)(c_3 - \eta_{a1}\kappa_1\kappa_2)\left\|\tilde{W}_c\right\|^2 - \eta_{a2}\left\|\tilde{W}_a\right\|^2 \\ &+ \frac{1}{4\theta(c_3 - \eta_{a1}\kappa_1\kappa_2)}\left[\frac{c_4\eta_c\varphi_0}{2\sqrt{\nu\varphi_1}}\kappa_3 + \eta_{a1}\kappa_1\kappa_2\kappa_3 \right. \\ &\left. + \eta_{a1}\kappa_1^2\kappa_2 + \eta_{a2}\kappa_1\right]^2 + \eta_{a1}\kappa_1^2\kappa_2\kappa_3 + \kappa_4, \end{aligned} \tag{52}$$

where $0 < \theta < 1$. Since $Q(x)$ is positive definite, Lemma 4.3 in Khalil (2002) indicates that there exist class $\mathcal{K}$ functions α_5 and α_6 such that

$$\begin{aligned}\alpha_5(\left\|\tilde{z}\right\|) &\leq Q + (1-\theta)(c_3 - \eta_{a1}\kappa_1\kappa_2)\left\|\tilde{W}_c\right\|^2 + \eta_{a2}\left\|\tilde{W}_a\right\|^2 \\ &\leq \alpha_6(\left\|\tilde{z}\right\|) \quad \forall \upsilon \in B_s. \end{aligned} \tag{53}$$

Using (53), the expression in (52) can be further upper bounded as

$$\begin{aligned}\dot{V}_L \leq{}& -\alpha_5(\left\|\tilde{z}\right\|) + \frac{1}{4\theta(c_3 - \eta_{a1}\kappa_1\kappa_2)}\left[\frac{c_4\eta_c\varphi_0}{2\sqrt{\nu\varphi_1}}\kappa_3 \right. \\ &\left. + \eta_{a1}\kappa_1\kappa_2\kappa_3 + \eta_{a1}\kappa_1^2\kappa_2 + \eta_{a2}\kappa_1\right]^2 + \eta_{a1}\kappa_1^2\kappa_2\kappa_3 + \kappa_4, \end{aligned}$$

which proves that $\dot{V}_L(\cdot)$ is negative whenever $\tilde{z}(t)$ lies outside the compact set $\Omega_{\tilde{z}} \triangleq \Big\{\tilde{z} : \left\|\tilde{z}\right\| \leq \alpha_5^{-1}\Big(\frac{1}{4\theta(c_3-\eta_{a1}\kappa_1\kappa_2)}\big[\frac{c_4\eta_c\varphi_0}{2\sqrt{\nu\varphi_1}}\kappa_3 + \eta_{a1}\kappa_1\kappa_2\kappa_3 + \eta_{a1}\kappa_1^2\kappa_2 + \eta_{a2}\kappa_1\big]^2 + \eta_{a1}\kappa_1^2\kappa_2\kappa_3 + \kappa_4\Big)\Big\}$, and hence, $\left\|\tilde{z}(t)\right\|$ is UUB (see Theorem 4.18 in Khalil, 2002). The bounds in (46) depend on the actor NN approximation error ε_υ', which can be reduced by increasing the number of neurons N, thereby reducing the size of the residual set $\Omega_{\tilde{z}}$. From Assumption 7, as the number of neurons of the actor and critic NNs $N \to \infty$, the reconstruction error $\varepsilon_\upsilon' \to 0$. □

Remark 3. Since the actor, critic and identifier are continuously updated, the developed RL algorithm can be compared to fully optimistic PI in machine learning literature (Bertsekas & Tsitsiklis, 1996), where policy evaluation and policy improvement are done after every state transition, unlike traditional PI, where policy improvement is done after convergence of the policy evaluation step. Proving convergence of optimistic PI is complicated and is an active area of research in machine learning (Bertsekas & Tsitsiklis, 1996; Busoniu, Babuska, De Schutter, & Ernst, 2010). By considering an adaptive control framework, this result investigates the convergence and stability behavior of fully optimistic PI in continuous-time.

Remark 4. The PE condition in Theorem 2 is equivalent to the exploration paradigm in RL which ensures sufficient sampling of the state space and convergence to the optimal policy (Sutton & Barto, 1998).

6. Simulation

The following nonlinear system is considered (Vamvoudakis & Lewis, 2010)

$$\begin{aligned}\dot{x} ={}& \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2(1 - (\cos(2x_1) + 2)^2) \end{bmatrix} \\ &+ \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix} u, \end{aligned} \tag{54}$$

where $x(t) \triangleq [x_1(t)x_2(t)]^T \in \mathbb{R}^2$ and $u(t) \in \mathbb{R}$. The state and control penalties are chosen as

$$Q(x) = x^T\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}x; \qquad R = 1.$$

The optimal value function and optimal control for the system in (54) are known, and given by Vamvoudakis and Lewis (2010)

$$V^*(x) = \frac{1}{2}x_1^2 + x_2^2; \qquad u^*(x) = -(\cos(2x_1) + 2)x_2,$$

which can be used to find the optimal weights $W = [0.501]^T$. The activation function for the critic NN is selected with $N = 3$ neurons as

$$\phi(x) = [x_1^2 \; x_1x_2 \; x_2^2]^T,$$

while the activation function for the identifier DNN is selected as a symmetric sigmoid with $L_f = 5$ neurons in the hidden layer.

Remark 5. The choice of a good basis for the value function and control policy is critical for convergence. For a general nonlinear system, choosing a suitable basis can be a challenging problem without any prior knowledge about the system. This is an active area of research in machine learning.

The identifier gains are selected as

$$k = 800, \qquad \alpha = 300, \qquad \gamma = 5, \qquad \beta_1 = 0.2,$$
$$\Gamma_{wf} = 0.1\mathbb{I}_{6\times 6}, \qquad \Gamma_{vf} = 0.1\mathbb{I}_{2\times 2},$$

and the gains for the actor–critic learning laws are selected as

$$\eta_{a1} = 10, \qquad \eta_{a2} = 50, \qquad \eta_c = 20, \qquad \nu = 0.005.$$

The covariance matrix is initialized to $\Gamma(0) = 5000$, all the NN weights are randomly initialized in $[-1, 1]$, and the states are initialized to $x(0) = [3, -1]$. An implementation issue in using the developed algorithm is to ensure PE of the critic regressor vector. Unlike linear systems, where PE of the regressor translates to sufficient richness of the external input, no verifiable method exists to ensure PE in nonlinear regulation problems. To ensure PE qualitatively, a small exploratory signal consisting of sinusoids of varying frequencies, $n(t) = \sin^2(t)\cos(t) + \sin^2(2t)\cos(0.1t) + \sin^2(-1.2t)\cos(0.5t) + \sin^5(t)$, is added to the control $u(t)$ for the first 3 s (Vamvoudakis & Lewis, 2010). The proposed control algorithm is implemented using (10), (11), (15), (16), (21), (22) and (39). The evolution of states is shown in Fig. 2. The identifier approximates the system dynamics, and the state derivative estimation error is shown in Fig. 3.

Remark 6. As compared to discontinuous sliding mode robust identifiers which require infinite bandwidth and exhibit chattering, the RISE-based identifier developed in (21) is continuous, and thus, mitigates chattering to a large extent, as seen in Fig. 3.

Persistence of excitation ensures that the weights converge close to their optimal values, i.e., $\hat{W}_c = [0.5004 \; 0.0005 \; 0.9999]^T \; (\approx \hat{W}_a)$ in approximately 2 s, as seen from the evolution of actor–critic weights in Figs. 4 and 5. The improved actor–critic weight update laws, based on minimization of the Bellman error, led to faster convergence of weights as compared to (Vamvoudakis & Lewis, 2010). The errors in approximating the optimal value function and optimal control at steady state ($t = 10$ s.) are plotted against the states in Figs. 6 and 7, respectively.

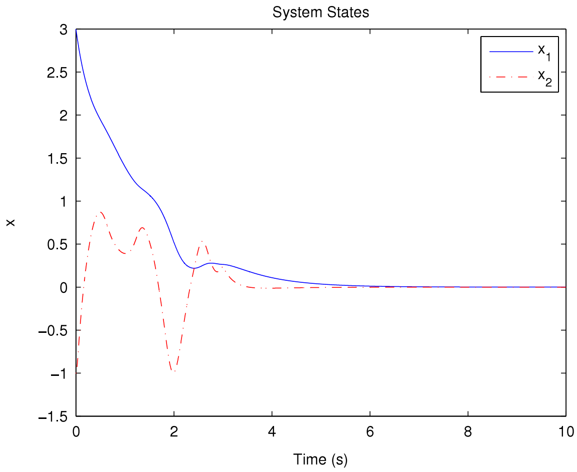

Fig. 2. System states $x(t)$ with persistently excited input for the first 3 s.

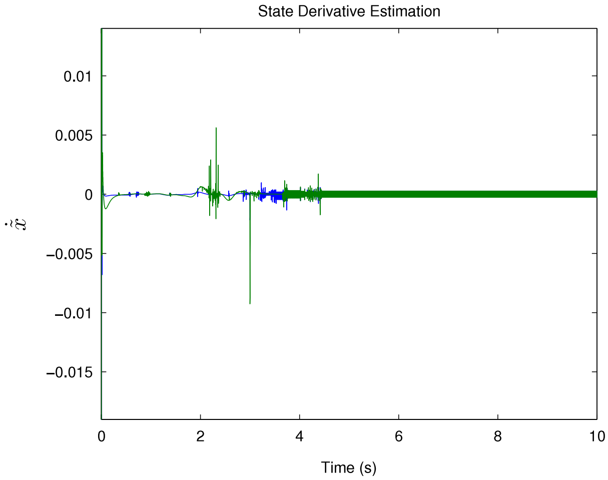

Fig. 3. Error in estimating the state derivative $\dot{\tilde{x}}(t)$ by the identifier.

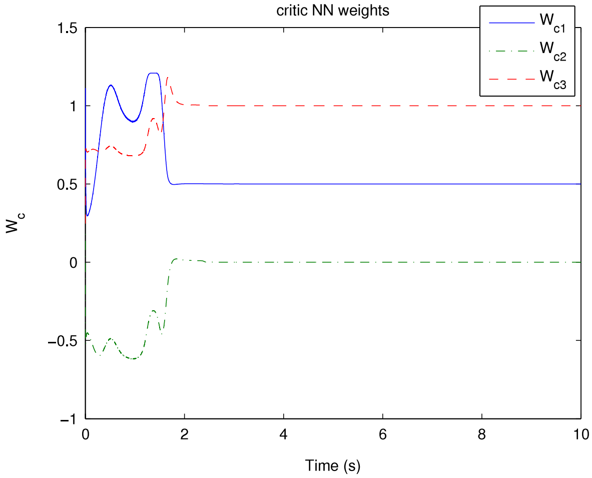

Fig. 4. Convergence of critic weights $\hat{W}_c(t)$.

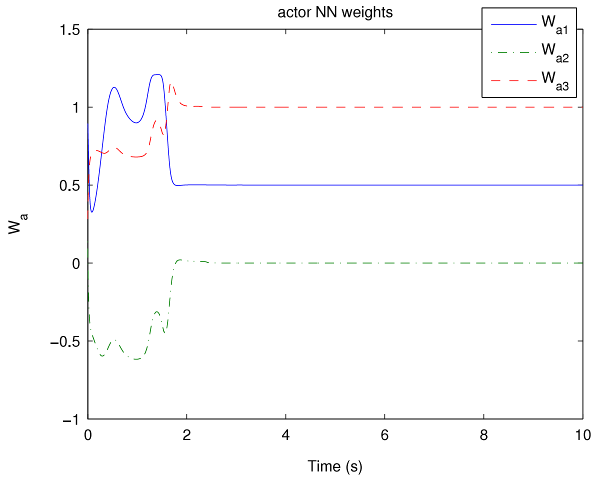

Fig. 5. Convergence of actor weights $\hat{W}_a(t)$.

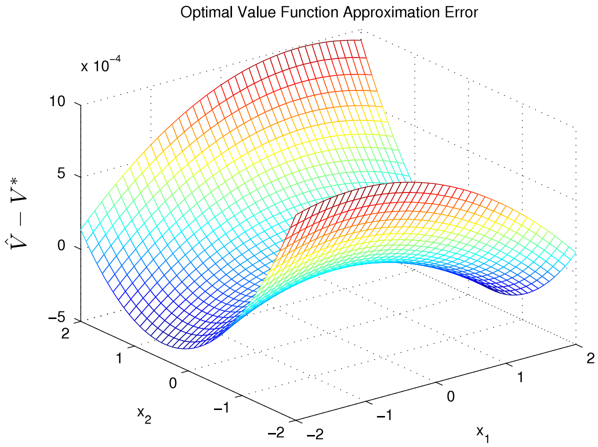

Fig. 6. Error in approximating the optimal value function by the critic at steady state.

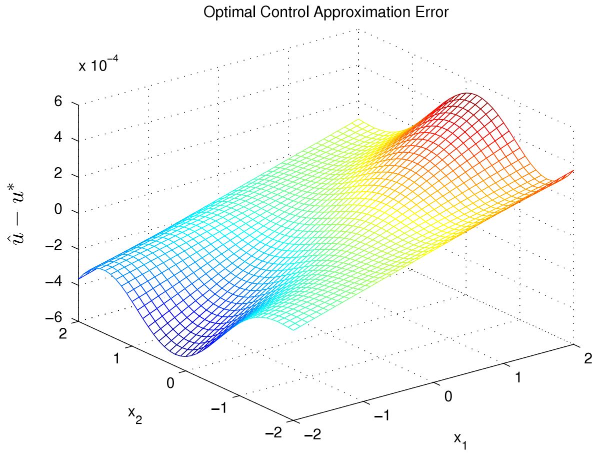

Fig. 7. Error in approximating the optimal control by the actor at steady state.

7. Conclusion

An actor–critic–identifier architecture is proposed to learn the approximate solution to the HJB equation for infinite-horizon optimal control of uncertain nonlinear systems. The online method is the first ever indirect adaptive control approach to continuous-time RL. The learning by the actor, critic and identifier is continuous and simultaneous, and the novel addition of the identifier to the traditional actor–critic architecture eliminates the need to know the system drift dynamics. The actor and critic minimize the Bellman error using gradient and least-squares update laws, respectively, and provide online approximations to the optimal control and the optimal value function, respectively. The identifier estimates the system dynamics online and asymptotically converges to the system state and its derivative. A PE condition is required to ensure exponential convergence to a bounded region in the neighborhood of the optimal control and UUB stability of the closed-loop system. Simulation results demonstrate the performance of the actor–critic–identifier-based method. A limitation of the method is the requirement of the knowledge of the input gain matrix. Future efforts will investigate ways to overcome this limitation, e.g., using methods similar to the model-free Q-learning methods (Bradtke, Ydstie, & Barto, 1994; Mehta, Barooah, & Dixon, 2009; Watkins & Dayan, 1992).

Appendix. Proof of Theorem 1

Let $V_I(y) : \mathcal{D} \to \mathbb{R}$ be a Lipschitz continuous regular positive definite function defined as

$$V_I \triangleq \frac{1}{2}e_f^T e_f + \frac{1}{2}\gamma\tilde{x}^T\tilde{x} + P + Q, \tag{A.1}$$

which satisfies the following inequalities:

$$U_1(y) \le V_I(y) \le U_2(y), \tag{A.2}$$

where $U_1(y), U_2(y) \in \mathbb{R}$ are continuous positive definite functions defined as

$$U_1 \triangleq \frac{1}{2}\min(1,\gamma)\|y\|^2 \qquad U_2 \triangleq \max(1,\gamma)\|y\|^2 .$$

Let $\dot{y} = h(y,t)$ represent the closed-loop differential equations in (23), (26), (27) and (35), where $h(y,t) \in \mathbb{R}^{2n+2}$ denotes the right-hand side of the closed-loop error signals. Using Filippov's theory of differential inclusions (Filippov, 1964; Smirnov, 2002), the existence of solutions can be established for $\dot{y} \in K[h](y,t)$, where $K[h] \triangleq \bigcap_{\delta>0}\bigcap_{\mu M=0}\overline{co}h(B(y,\delta)-M,t)$, where $\bigcap_{\mu M=0}$ denotes the intersection over all sets M of Lebesgue measure zero, $\overline{co}$ denotes convex closure, and $B(y,\delta) = \{x \in \mathbb{R}^{2n+2} |\, \|y-x\| < \delta\}$. The right hand side of the differential equation, $h(y,t)$, is continuous except for the Lebesgue measure zero set of times $t \in [t_0, t_f]$ when $\tilde{x}(t) = 0$. Hence, the set of time instances for which $\dot{y}(t)$ is not defined is Lebesgue negligible. The absolutely continuous solution $y(t) = y(t_0) + \int_{t_0}^{t}\dot{y}(t)dt$ does not depend on the value of $\dot{y}(t)$ on a Lebesgue negligible set of time instances (Leine & van de Wouw, 2008). Under Filippov's framework, a generalized Lyapunov stability theory can be used to establish strong stability of the closed-loop system $\dot{y} = h(y,t)$. The generalized time derivative of (A.1) exists almost everywhere (a.e.), i.e. for almost all $t \in [t_0, t_f]$, and $\dot{V}_I(y) \in^{a.e.} \dot{\tilde{V}}_I(y)$ where

$$\dot{\tilde{V}}_I = \bigcap_{\xi\in\partial V_I(y)} \xi^T K\left[\dot{e_f}^T \dot{\tilde{x}}^T \frac{1}{2}P^{-\frac{1}{2}}\dot{P}\frac{1}{2}Q^{-\frac{1}{2}}\dot{Q}\right]^T, \tag{A.3}$$

where ∂V_I is the generalized gradient of $V_I(y)$ (Clarke, 1990). Since $V_I\,(y)$ is a Lipschitz continuous regular function which is smooth in y, (A.3) can be simplified as Shevitz and Paden (1994)

$$\begin{aligned}\dot{\tilde{V}}_I &= \nabla V_I^T K\left[\dot{e_f}^T \dot{\tilde{x}}^T \frac{1}{2}P^{-\frac{1}{2}}\dot{P}\frac{1}{2}Q^{-\frac{1}{2}}\dot{Q}\right]^T \\ &= \left[e_f^T \gamma\tilde{x}^T 2P^{\frac{1}{2}} 2Q^{\frac{1}{2}}\right] K\left[\dot{e_f}^T \dot{\tilde{x}}^T \frac{1}{2}P^{-\frac{1}{2}}\dot{P}\frac{1}{2}Q^{-\frac{1}{2}}\dot{Q}\right]^T .\end{aligned}$$

Using the calculus for $K[\cdot]$ from Paden and Sastry (1987, Theorem 1, Properties (2,5,7)), and substituting the dynamics from (27) and (35), yields

$$\begin{aligned}\dot{\tilde{V}}_I \subset\ & e_f^T(\tilde{N} + N_{B1} + \hat{N}_{B2} - ke_f - \beta_1 K[sgn(\tilde{x})] - \gamma\tilde{x}) \\ & + \gamma\tilde{x}^T(e_f - \alpha\tilde{x}) - e_f^T(N_{B1} - \beta_1 K[sgn(\tilde{x})]) \\ & - \dot{\tilde{x}}^T N_{B2} + \beta_2\rho_2(\|z\|)\|z\|\|\tilde{x}\| \\ & - \frac{1}{2}\alpha\left[tr(\tilde{W}_f^T \Gamma_{wf}^{-1}\dot{\hat{W}}_f) + tr(\tilde{V}_f^T \Gamma_{vf}^{-1}\dot{\hat{V}}_f)\right] \qquad \text{(A.4)} \\ & - \frac{1}{2}\alpha\sum_{i=1}^{m}\left[tr(\tilde{W}_{gi}^T \Gamma_{wgi}^{-1}\dot{\hat{W}}_{gi}) + tr(\tilde{V}_{gi}^T \Gamma_{vgi}^{-1}\dot{\hat{V}}_{gi})\right]. \\ \overset{a.e.}{\le}\ & -\alpha\gamma\|\tilde{x}\|^2 - k\|e_f\|^2 + \rho_1(\|z\|)\|z\|\|e_f\| \\ & + \zeta_5\|\tilde{x}\|^2 + \zeta_6\|e_f\|^2 + \beta_2\rho_2(\|z\|)\|z\|\|\tilde{x}\|, \qquad \text{(A.5)}\end{aligned}$$

where (26), (31) and (33) are used, $K[sgn(\tilde{x})] = SGN(\tilde{x})$ (Paden & Sastry, 1987), such that $SGN(\tilde{x}_i) = 1$ if $\tilde{x}_i > 0$, $[-1, 1]$ if $\tilde{x}_i = 0$, and -1 if $\tilde{x}_i < 0$ (the subscript i denotes the i^{th} element). The set in (A.4) reduces to the scalar inequality in (A.5) since the RHS is continuous a.e., i.e., the RHS is continuous except for the Lebesgue measure zero set of times when $\tilde{x}(t) = 0$ (Leine & van de Wouw, 2008). Substituting for $k \triangleq k_1 + k_2$ and $\gamma \triangleq \gamma_1 + \gamma_2$, and completing the squares, the expression in (A.5) can be upper bounded as

$$\begin{aligned}\dot{\tilde{V}}_I \overset{a.e.}{\le}\ & -(\alpha\gamma_1 - \zeta_5)\|\tilde{x}\|^2 - (k_1 - \zeta_6)\|e_f\|^2 \\ & + \frac{\rho_1(\|z\|)^2}{4k_2}\|z\|^2 + \frac{\beta_2^2\rho_2(\|z\|)^2}{4\alpha\gamma_2}\|z\|^2 .\end{aligned} \tag{A.6}$$

Provided the sufficient conditions in (38) are satisfied, the expression in (A.6) can be rewritten as

$$\begin{aligned}\dot{\tilde{V}}_I \overset{a.e.}{\le}\ & -\lambda\|z\|^2 + \frac{\rho(\|z\|)^2}{4\eta}\|z\|^2 \\ \overset{a.e.}{\le}\ & -U(y)\forall y \in \mathcal{D}\end{aligned} \tag{A.7}$$

where $\lambda \triangleq \min\{\alpha\gamma_1 - \zeta_5, k_1 - \zeta_6\}$, $\eta \triangleq \min\left\{k_2, \frac{\alpha\gamma_2}{\beta_2^2}\right\}$, $\rho(\|z\|)^2 \triangleq \rho_1(\|z\|)^2 + \rho_2(\|z\|)^2$ is a positive, globally invertible, nondecreasing function, and $U(y) = c\|z\|^2$, for some positive constant c, is a continuous, positive semi-definite function defined on the domain $\mathcal{D} \triangleq \left\{y(t) \in \mathbb{R}^{2n+2} \mid \|y\| \le \rho^{-1}\left(2\sqrt{\lambda\eta}\right)\right\}$. The size of the domain $\mathcal{D}$ can be increased by increasing the gains k and γ. The inequalities in (A.2) and (A.7) can be used to show that $V_I(y) \in \mathcal{L}_\infty$ in $\mathcal{D}$; hence, $\tilde{x}(t), e_f(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$. Using (24), standard linear analysis can be used to show that $\dot{\tilde{x}}(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$, and since $\dot{x}(t) \in \mathcal{L}_\infty$, $\dot{\hat{x}}(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$. Since $\hat{W}_f(t) \in \mathcal{L}_\infty$ from the use of projection in (26), $\hat{\sigma}_f(t) \in \mathcal{L}_\infty$ from Assumption 5, and $\hat{u}(t) \in \mathcal{L}_\infty$ from Assumption 8, $\mu(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$ from (21). Using the above bounds, it can be shown from (27) that $\dot{e_f}(t) \in \mathcal{L}_\infty$ in $\mathcal{D}$. Since $\tilde{x}(t)$,$e_f(t) \in \mathcal{L}_\infty$, the definition of $U(y)$ can be used to show that it is uniformly continuous in $\mathcal{D}$. Let $\mathcal{S} \subset \mathcal{D}$ denote a set defined as $\mathcal{S} \triangleq \left\{y(t)\subset \mathcal{D} \mid U_2(y(t)) < \frac{1}{2}\left(\rho^{-1}\left(2\sqrt{\lambda\eta}\right)\right)^2\right\}$, where the region of attraction can be made arbitrarily large to include any initial conditions by increasing the control gain η (i.e. a semi-global type of stability result), and hence $c\|z\|^2 \to 0 as t \to \infty \forall y(0) \in \mathcal{S}$. Using the definition of $z(t)$, it can be shown that $\|\tilde{x}(t)\|$, $\left\|\dot{\tilde{x}}(t)\right\|$, $\|e_f\| \to 0 as t \to \infty \forall y(0) \in \mathcal{S}$.

References

Abu-Khalaf, M., & Lewis, F. (2005). Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. *Automatica*, *41*(5), 779–791.

Al-Tamimi, A., Lewis, F. L., & Abu-Khalaf, M. (2008). Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof. *IEEE Transactions on Systems, Man, and Cybernetics. Part B Cybernetics*, *38*, 943–949.

Baird, L. (1993). Advantage updating, Tech. rep., Wright Lab, Wright–Patterson Air Force Base, OH.

Balakrishnan, S., & Biega, V. (1996). Adaptive-critic-based neural networks for aircraft optimal control. *Journal of Guidance, Control and Dynamics*, *19*(4), 893–898.

Barto, A., Sutton, R., & Anderson, C. (1983). Neuron-like adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*(5), 834–846.

Beard, R., Saridis, G., & Wen, J. (1997). Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation. *Automatica*, *33*, 2159–2178.

Bertsekas, D., & Tsitsiklis, J. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bradtke, S., Ydstie, B., & Barto, A. (1994). Adaptive linear quadratic control using policy iteration. In *Proc. Am. control conf* (pp. 3475–3479). IEEE.

Busoniu, L., Babuska, R., De Schutter, B., & Ernst, D. (2010). *Reinforcement learning and dynamic programming using function approximators*. CRC Press.

Clarke, F. H. (1990). *Optimization and nonsmooth analysis*. SIAM.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*, 303–314.

Dierks, T., Thumati, B., & Jagannathan, S. (2009). Optimal control of unknown affine nonlinear discrete-time systems using offline-trained neural networks with proof of convergence. *Neural Networks*, *22*(5–6), 851–860.

Dixon, W. E., Behal, A., Dawson, D. M., & Nagarkatti, S. (2003). *Nonlinear control of engineering systems: a Lyapunov-based approach*. Boston: Birkhuser.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, *12*(1), 219–245.

Ferrari, S., & Stengel, R. (2002). An adaptive critic global controller. In *Proc. Am. control conf. vol. 4*.

Filippov, A. (1964). Differential equations with discontinuous right-hand side. *American Mathematical Society Translations*, *42*(2), 199–231.

Filippov, A. F. (1988). *Differential equations with discontinuous right-hand sides*. Kluwer Academic Publishers.

He, P., & Jagannathan, S. (2007). Reinforcement learning neural-network-based controller for nonlinear discrete-time systems with input constraints. *IEEE Transactions on Systems, Man, and Cybernetics. Part B Cybernetics*, *37*(2), 425–436.

Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(10), 3088.

Hornik, K., Stinchcombe, M., & White, H. (1985). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Khalil, H. K. (2002). *Nonlinear systems* (3rd ed.). Prentice Hall.

Kirk, D. (2004). *Optimal control theory: an introduction*. Dover Pubns.

Konda, V., & Tsitsiklis, J. (2004). On actor–critic algorithms. *SIAM Journal on Control and Optimization*, *42*(4), 1143–1166.

Krstic, M., Kokotovic, P. V., & Kanellakopoulos, I. (1995). *Nonlinear and adaptive control design*. John Wiley & Sons.

Leine, R., & van de Wouw, N. (2008). Non-smooth dynamical systems. In *Lecture notes in applied and computational mechanics*: *vol. 36*. *Stability and convergence of mechanical systems with unilateral constraints* (pp. 59–77). Berlin/Heidelberg: Springer.

Lendaris, G., Schultz, L., & Shannon, T. (2000). Adaptive critic design for intelligent steering and speed control of a 2-axle vehicle. In *Int. joint conf. neural netw.* (pp. 73–78).

Lewis, F. L., Selmic, R., & Campos, J. (2002). *Neuro-fuzzy control of industrial systems with actuator nonlinearities*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics.

Mehta, S.S.S., Barooah, P., & Dixon, W.E. (2009). A novel algorithm for refinement of vision-based two-view pose estimates. In *IEEE conference on decision and control*.

Murray, J., Cox, C., Lendaris, G., & Saeks, R. (2002). Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics. Part C Applications and Review*, *32*(2), 140–153.

Paden, B., & Sastry, S. (1987). A calculus for computing Filippov's differential inclusion with application to the variable structure control of robot manipulators. *IEEE Transactions on Circuits and Systems*, *34*(1), 73–82.

Padhi, R., Unnikrishnan, N., Wang, X., & Balakrishnan, S. (2006). A single network adaptive critic (SNAC) architecture for optimal control synthesis for a class of nonlinear systems. *Neural Networks*, *19*(10), 1648–1660.

Patre, P. M., MacKunis, W., Kaiser, K., & Dixon, W. E. (2008). Asymptotic tracking for uncertain dynamic systems via a multilayer neural network feedforward and RISE feedback control structure. *IEEE Transactions on Automatic Control*, *53*(9), 2180–2185.

Poznyak, A., Sanchez, E., & Yu, W. (2001). *Differential neural networks for robust nonlinear control: identification, state estimation and trajectory tracking*. World Scientific Pub Co Inc.

Prokhorov, D. V., Wunsch, I., & C, D. (1997). Adaptive critic designs. *IEEE Transactions on Neural Networks*, *8*, 997–1007.

Sastry, S., & Bodson, M. (1989). *Adaptive control: stability, convergence, and robustness*. Upper Saddle River, NJ: Prentice-Hall.

Shevitz, D., & Paden, B. (1994). Lyapunov stability theory of nonsmooth systems. *IEEE Transactions on Automatic Control*, *39*(9), 1910–1914.

Si, J., Barto, A., Powell, W., & Wunsch, D. (Eds.) (2004). *Handbook of learning and approximate dynamic programming*. Wiley-IEEE Press.

Smirnov, G. V. (2002). *Introduction to the theory of differential inclusions*. American Mathematical Society.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. MIT Press.

Sutton, R., Barto, A., & Williams, R. (1992). Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems Magazin*, *12*(2), 19–22.

Vamvoudakis, K., & Lewis, F. (2010). Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, *46*, 878–888.

Vrabie, D., & Lewis, F. (2009). Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, *22*(3), 237–246.

Watkins, C., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3), 279–292.

Werbos, P. (1990). A menu of designs for reinforcement learning over time. *Neural Networks for Control*, 67–95.

Werbos, P. (1992). Approximate dynamic programming for real-time control and neural modeling. In D. A. White, & D. A. Sofge (Eds.), *Handbook of intelligent control: neural, fuzzy, and adaptive approaches*. New York: Van Nostrand Reinhold.

Widrow, B., Gupta, N., & Maitra, S. (1973). Punish/reward: Learning with a critic in adaptive threshold systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *3*(5), 455–465.

Xian, B., Dawson, D. M., de Queiroz, M. S., & Chen, J. (2004). A continuous asymptotic tracking control strategy for uncertain nonlinear systems. *IEEE Transactions on Automatic Control*, *49*, 1206–1211.

Shubhendu Bhasin received his Ph.D. in 2011 from the Department of Mechanical and Aerospace Engineering at the University of Florida. He is currently Assistant Professor in the Department of Electrical Engineering at the Indian Institute of Technology, Delhi. His research interests include reinforcement learning-based feedback control, approximate dynamic programming, neural network-based control, nonlinear system identification and parameter estimation, and robust and adaptive control of uncertain nonlinear systems.

Rushikesh Kamalapurkar received his Bachelor's degree in mechanical engineering from Visvesvaraya National Institute of Technology, India in 2007 and his Master's from the University of Florida in 2011. He is currently a Ph.D. student with the Nonlinear Control and Robotics group at the University of Florida. His research interests include the applications of reinforcement learning to feedback control of uncertain nonlinear systems, and differential game-based distributed control of multiple autonomous agents.

Marcus Johnson received his Ph.D. in 2011 from the Department of Mechanical and Aerospace Engineering at the University of Florida. He is currently working as a Research Aerospace Engineer at NASA Ames Research Center and his main research interest is the development of Lyapunov-based proofs for optimality of nonlinear adaptive systems.

Kyriakos G. Vamvoudakis was born in Athens Greece. He received the Diploma (5 year degree) in electronic and computer engineering from the Technical University of Crete, Greece in 2006 with highest honors, and the M.Sc. and Ph.D. degrees in electrical engineering from The University of Texas at Arlington in 2008 and 2011 respectively. From May 2011 to January 2012, he was working as an Adjunct Professor and Faculty Research Associate at The University of Texas at Arlington and at the Automation and Robotics Research Institute. He is currently working as a Project Research Scientist at the Center of Control, Dynamical Systems and Computation (CCDC) at the University of California, Santa Barbara. He is coauthor of 6 book chapters, 40 technical publications and the book *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. His research interests include approximate dynamic programming, game theory, neural network feedback control, and optimal control. Recently, his research has focused on network security and multi-agent optimization. He is a member of Tau Beta Pi, Eta Kappa Nu and Golden Key honor societies and is listed in *Who's Who in the World, Who's Who in Science and Engineering*, and *Who's Who in America*. He received the Best Paper Award for Autonomous/Unmanned Vehicles at the 27 th Army Science Conference in 2010, the Best Presentation Award at the World Congress of Computational Intelligence in 2010 and the Best Researcher Award, UTA Automation & Robotics Research Institute in 2011. He has organized special sessions for several international conferences. Dr. Vamvoudakis is a registered Electrical/Computer engineer (PE) and member of the Technical Chamber of Greece.

F.L. Lewis, IEEE Fellow, IFAC Fellow, Fellow Inst. Measurement & Control, PE Texas, UK. Chartered Engineer. Distinguished Scholar Professor and Moncrief–O'Donnell Chair at The University of Texas at Arlington. Works in feedback control and intelligent systems. Author of 6 US patents, books, and several journal papers. Awards include Fulbright Research Award, NSF Research Initiation Grant, ASEE Terman Award, and International Neural Network Society Gabor Award and Neural Network Pioneer Award. Selected as Engineer of the Year by Ft. Worth IEEE Section.

Warren Dixon received his Ph.D. in 2000 from the Department of Electrical and Computer Engineering from Clemson University. After completing his doctoral studies he was selected as an Eugene P. Wigner Fellow at Oak Ridge National Laboratory (ORNL). In 2004, Dr. Dixon joined the faculty of the University of Florida in the Mech. and Aero. Eng. Dept. His research focus is the development and application of Lyapunov-based control techniques for uncertain nonlinear systems. He has published 3 books, an edited collection, 9 chapters, and over 250 refereed journal and conference papers. His work has been recognized by the 2011 American Society of Mechanical Engineers (ASME) Dynamics Systems and Control Division Outstanding Young Investigator Award, 2009 American Automatic Control Council (AACC) O. Hugo Schuck Award, 2006 IEEE Robotics and Automation Society (RAS) Early Academic Career Award, an NSF CAREER Award (2006–2011), 2004 DOE Outstanding Mentor Award, and the 2001 ORNL Early Career Award for Engineering Achievement. Dr. Dixon is a senior member of IEEE. He serves or has served as a member of numerous technical, conference program, and organizing committees. He served as an appointed member to the IEEE CSS Board of Governors (BoG) in 2008, and now serves as the Director of Operations for the Executive Committee of the BoG. He is currently or formerly an associate editor for *ASME Journal of Dynamic Systems, Measurement and Control, Automatica, IEEE Transactions on Systems Man and Cybernetics: Part B Cybernetics, and the International Journal of Robust and Nonlinear Control*.