J Control Theory Appl 2011 9 (3) 400–409
DOI 10.1007/s11768-011-0170-8

Asymptotic tracking by a reinforcement learning-based adaptive critic controller

Shubhendu BHASIN [1], Nitin SHARMA [2], Parag PATRE [3], Warren DIXON [1]

1.Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL 32611, U.S.A.;

2.Department of Physiology, University of Alberta, Edmonton, Alberta, Canada;

3.NASA Langley Research Center, Hampton, VA 23681, U.S.A.

Abstract: Adaptive critic (AC) based controllers are typically discrete and/or yield a uniformly ultimately bounded stability result because of the presence of disturbances and unknown approximation errors. A continuous-time AC controller is developed that yields asymptotic tracking of a class of uncertain nonlinear systems with bounded disturbances. The proposed AC-based controller consists of two neural networks (NNs) – an action NN, also called the actor, which approximates the plant dynamics and generates appropriate control actions; and a critic NN, which evaluates the performance of the actor based on some performance index. The reinforcement signal from the critic is used to develop a composite weight tuning law for the action NN based on Lyapunov stability analysis. A recently developed robust feedback technique, robust integral of the sign of the error (RISE), is used in conjunction with the feedforward action neural network to yield a semiglobal asymptotic result. Experimental results are provided that illustrate the performance of the developed controller.

Keywords: Adaptive critic; Reinforcement learning; Neural network-based control

1 Introduction

First used to explain animal behavior and psychology, reinforcement learning (RL) is now a useful computational tool for learning by experience in many engineering applications, such as computer game playing, industrial manufacturing, traffic management, robotics and control, etc. RL involves learning by interacting with the environment, sensing the states, and choosing actions based on these interactions, with the aim of maximizing a numerical reward [1]. Unlike supervised learning where learning is instructional and based on a set of examples of correct input/output behavior, RL is more evaluative and indicates only the measure of goodness of a particular action. Because interaction is done without a teacher, RL is particularly effective in situations where examples of desired behavior are not available but it is possible to evaluate the performance of actions based on some performance criterion.

Actor-critic or adaptive critic (AC) architectures have been proposed as models of RL [1, 2]. In AC-based RL, an actor network learns to select actions based on evaluative feedback from the critic in order to maximize future rewards. Because of the success of neural networks (NNs) as universal approximators [3, 4], they have become a natural choice in AC architectures for approximating unknown plant dynamics and cost functions [5, 6]. Typically, the AC architecture consists of two NNs – an action NN and a critic NN. The critic NN approximates the evaluation function, mapping states to an estimated measure of the value function, while the action NN approximates an optimal control law and generates actions or control signals. Following the works of Werbos [7], Watkins [8], Barto [9] and Sutton [10], current research focuses on the relationship between RL and dynamic programming (DP) [11] methods for solving optimal control problems. Because of the curse of dimensionality associated with using DP, Werbos [12] introduced an alternative approximate dynamic programming (ADP) approach that gives an approximate solution to the DP problem (or the Hamiltonian-Jacobi-Bellman equation for optimal control). A detailed review of AC designs can be found in [13]. Various modifications to ADP-based algorithms have since been proposed [14–16].

The performance of AC-based controllers has been successfully tested on various nonlinear plants with unknown dynamics. Venayagamoorthy et al. used AC for control of turbogenerators, synchronous generators, and power systems [17, 18]. Ferrari and Stengel [19] used a dual heuristic programming (DHP) based AC approach to control a nonlinear simulation of a jet aircraft in the presence of parameter variations and control failures. Jagannathan et al. [20] used ACs for grasping control of a three-finger-gripper. Some other interesting applications are missile control [21], HVAC control [22], and control of distributed parameter systems [23].

The convergence of algorithms for ADP-based RL controllers is studied in [14, 24–27]. Most of this work has been focused on convergence analysis for discrete-time systems. The fact that continuous-time ADP requires knowledge of the system dynamics has hampered the development of continuous-time extensions to ADP-based AC controllers. Recent results in [28–30] have made new inroads by addressing the problem for partially unknown nonlinear systems. However, the inherently iterative nature of the ADP

Received 21 July 2010; revised 22 March 2011.
This research was partly supported by the National Science Foundation (No.0901491).

algorithm has prevented the development of rigorous stability proofs of closed-loop controllers for continuous-time unknown nonlinear systems.

In this paper, a continuous asymptotic AC-based tracking controller is developed for a class of nonlinear systems with bounded disturbances. The approach is different from the optimal control-based ADP approaches proposed in literature [24–30], where the critic usually approximates a long-term cost function and the actor approximates the optimal control. However, the similarity with the ADP-based methods is in the use of the AC architecture, inherited from RL, where the critic, through a reinforcement signal, affects the behavior of the actor leading to an improved performance. The proposed robust adaptive controller consists of an NN feedforward term (actor NN) and a robust feedback term, where the weight update laws of the actor NN are designed as a composite of a tracking error term and a reinforcement learning term (from the critic), with the objective of minimizing the tracking error [31–33]. The robust term is designed to withstand the external disturbances and modeling errors in the plant. Typically, the presence of bounded disturbances and NN approximation errors lead to a uniformly ultimately bounded (UUB) result. The main contribution of this paper is the use of a recently developed continuous feedback technique, robust integral of sign of the error (RISE) [34, 35], in conjunction with the AC architecture to yield asymptotic tracking of an unknown nonlinear system subjected to bounded external disturbances. The use of RISE in conjunction with the action NN makes the design of the critic NN architecture challenging from a stability standpoint. To this end, the critic NN is combined with an additional RISE-like term to yield a reinforcement signal, which is used to update the weights of the action NN. A smooth projection algorithm is used to bound the NN weight estimates and a Lyapunov stability analysis guarantees closed-loop stability of the system. Experiments are performed to demonstrate the improved performance with the proposed RL-based AC method.

2 Dynamic model and properties

The mnth order MIMO Brunovsky form can be written as [31]

$$\begin{cases} \dot{x}_1 = x_2, \\ \vdots \\ \dot{x}_{n-1} = x_n, \\ \dot{x}_n = g(x) + u + d, \\ y = x_1, \end{cases} \tag{1}$$

where $x(t) \triangleq [x_1^{\mathrm{T}} \ x_2^{\mathrm{T}} \ \cdots \ x_n^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{mn}$ are the measurable system states, $u(t) \in \mathbb{R}^m$, $y \in \mathbb{R}^m$ are the control input and system output, respectively; $g(x) \in \mathbb{R}^m$ is an unknown smooth function, locally Lipschitz in x; and $d(t) \in \mathbb{R}^m$ is an external bounded disturbance.

Assumption 1 The function $g(x)$ is the second-order differentiable, i.e., $g(\cdot), \dot{g}(\cdot), \ddot{g}(\cdot) \in \mathcal{L}_\infty$ if $x^{(i)}(t) \in \mathcal{L}_\infty$, $i = 0, 1, 2$, where $(\cdot)^{(i)}(t)$ denotes the ith derivative with respect to time.

Assumption 2 The desired trajectory $y_{\mathrm{d}}(t) \in \mathbb{R}^m$ is designed such that $y_{\mathrm{d}}^{(i)}(t) \in \mathcal{L}_\infty$, $i = 0, 1, \ldots, n+1$.

Assumption 3 The disturbance term and its first and second time derivatives are bounded, i.e., $d(t), \dot{d}(t), \ddot{d}(t) \in \mathcal{L}_\infty$.

3 Control objective

The control objective is to design a continuous RL-based NN controller such that the output $y(t)$ tracks a desired trajectory $y_{\mathrm{d}}(t)$. To quantify the control objective, the tracking error $e_1(t) \in \mathbb{R}^m$ is defined as

$$e_1 \triangleq y - y_{\mathrm{d}}. \tag{2}$$

The following filtered tracking errors are defined to facilitate the subsequent stability analysis

$$\begin{cases} e_2 \triangleq \dot{e}_1 + \alpha_1 e_1, \\ e_i \triangleq \dot{e}_{i-1} + \alpha_{i-1} e_{i-1} + e_{i-2}, \ \ i = 3, \ldots, n, \end{cases} \tag{3}$$

$$r \triangleq \dot{e}_n + \alpha_n e_n, \tag{4}$$

where $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ are positive constant control gains. Note that the signals $e_1(t), \ldots, e_n(t) \in \mathbb{R}^m$ are measurable whereas the filtered tracking error $r(t) \in \mathbb{R}^m$ in (4) is not measurable because it depends on $\dot{x}_n(t)$. The filtered tracking errors in (3) can be expressed in terms of the tracking error $e_1(t)$ as

$$e_i = \sum_{j=0}^{i-1} a_{ij} e_1^{(j)}, \ \ i = 2, \ldots, n, \tag{5}$$

where $a_{ij} \in \mathbb{R}$ are positive constants obtained from substituting (5) in (3) and comparing coefficients [35]. It can be easily shown that

$$a_{ij} = 1, \ \ j = i - 1. \tag{6}$$

4 Action NN-based control

Using (2)–(6), the open loop error system can be written as

$$r = y^{(n)} - y_{\mathrm{d}}^{(n)} + f, \tag{7}$$

where $f(e_1, \dot{e}_1, \ldots, e_1^{(n-1)}) \in \mathbb{R}^m$ is a function of known and measurable terms, defined as

$$f = \sum_{j=0}^{n-2} a_{nj} (e_1^{(j+1)} + \alpha_n e_1^{(j)}) + \alpha_n e_1^{(n-1)}.$$

Substituting the dynamics from (1) into (7) yields

$$r = g(x) + d - y_{\mathrm{d}}^{(n)} + f + u. \tag{8}$$

Adding and subtracting $g(x_{\mathrm{d}}) : \mathbb{R}^{mn} \to \mathbb{R}^m$, where $g(x_{\mathrm{d}})$ is a smooth unknown function of the desired trajectory $x_{\mathrm{d}}(t) \triangleq [y_{\mathrm{d}}^{\mathrm{T}} \ \dot{y}_{\mathrm{d}}^{\mathrm{T}} \ \cdots \ (y_{\mathrm{d}}^{(n-1)})^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{mn}$, the expression in (8) can be written as

$$r = g(x_{\mathrm{d}}) + S + d + Y + u, \tag{9}$$

where $Y(e_1, \dot{e}_1, \ldots, e_1^{(n-1)}, y_{\mathrm{d}}^{(n)}) \in \mathbb{R}^m$ contains known and measurable terms and is defined as

$$Y \triangleq -y_{\mathrm{d}}^{(n)} + f, \tag{10}$$

and the auxiliary function $S(x, x_{\mathrm{d}}) \in \mathbb{R}^m$ is defined as

$$S \triangleq g(x) - g(x_{\mathrm{d}}).$$

The unknown nonlinear term $g(x_{\mathrm{d}})$ can be represented by a multilayer NN as

$$g(x_{\mathrm{d}}) = W_{\mathrm{a}}^{\mathrm{T}} \sigma(V_{\mathrm{a}}^{\mathrm{T}} x_{\mathrm{a}}) + \varepsilon(x_{\mathrm{a}}), \tag{11}$$

where $x_{\rm a}(t) \in \mathbb{R}^{mn+1} \triangleq [1 \;\; x_{\rm d}^{\rm T}]^{\rm T}$ is the input to the NN, $W_{\rm a} \in \mathbb{R}^{(N_{\rm a}+1)\times m}$ and $V_{\rm a} \in \mathbb{R}^{(mn+1)\times N_{\rm a}}$ are the constant bounded ideal weights for the output and hidden layers respectively with $N_{\rm a}$ being the number of neurons in the hidden layer, $\sigma(\,\cdot\,) \in \mathbb{R}^{N_{\rm a}+1}$ is the bounded activation function, and $\varepsilon(x_{\rm a}) \in \mathbb{R}^m$ is the function reconstruction error.

Remark 1 The NN used in (11) is referred to as the action NN or the associative search element (ASE) [9], and it is used to approximate the system dynamics and generate appropriate control signals.

Since the desired trajectory is bounded (from Assumption 2), the following inequalities hold:

$$\begin{cases} \|\varepsilon_{\rm a}(x_{\rm a})\| \leqslant \varepsilon_{\rm a_1}, \;\; \|\dot{\varepsilon}_{\rm a}(x_{\rm a}, \dot{x}_{\rm a})\| \leqslant \varepsilon_{\rm a_2}, \\ \|\ddot{\varepsilon}_{\rm a}(x_{\rm a}, \dot{x}_{\rm a}, \ddot{x}_{\rm a})\| \leqslant \varepsilon_{\rm a_3}, \end{cases} \tag{12}$$

where $\varepsilon_{\rm a_1}, \varepsilon_{\rm a_2}, \varepsilon_{\rm a_3} \in \mathbb{R}$ are known positive constants. Also, the ideal weights are assumed to exist and be bounded by known positive constants [6], such that

$$\|V_{\rm a}\| \leqslant \bar{V}_{\rm a}, \;\; \|W_{\rm a}\| \leqslant \bar{W}_{\rm a}. \tag{13}$$

Substituting (11) in (9), the open loop error system can now be written as

$$r = W_{\rm a}^{\rm T}\sigma(V_{\rm a}^{\rm T}x_{\rm a}) + \varepsilon(x_{\rm a}) + S + d + Y + u. \tag{14}$$

The NN approximation for $g(x_{\rm d})$ can be represented as

$$\hat{g}(x_{\rm d}) = \hat{W}_{\rm a}^{\rm T}\sigma(\hat{V}_{\rm a}^{\rm T}x_{\rm a}),$$

where $\hat{W}_{\rm a}(t) \in \mathbb{R}^{(N_{\rm a}+1)\times m}$ and $\hat{V}_{\rm a}(t) \in \mathbb{R}^{(mn+1)\times N_{\rm a}}$ are the subsequently designed estimates of the ideal weights. The control input $u(t)$ in (14) can now be designed as

$$u \triangleq -Y - \hat{g}(x_{\rm d}) - \mu_{\rm a}, \tag{15}$$

where $\mu_{\rm a}(t) \in \mathbb{R}^m$ denotes the RISE feedback term defined as [34, 35]

$$\mu_{\rm a} \triangleq (k_{\rm a}+1)e_n(t) - (k_{\rm a}+1)e_n(0) + v, \tag{16}$$

where $v(t) \in \mathbb{R}^m$ is the generalized solution (in Filippov's sense [36]) to

$$\dot{v} = (k_{\rm a}+1)\alpha_n e_n + \beta_1 {\rm sgn}(e_n), \;\; v(0) = 0, \tag{17}$$

where $k_{\rm a}, \beta_1 \in \mathbb{R}$ are constant positive control gains, and ${\rm sgn}(\,\cdot\,)$ denotes a vector signum function.

Remark 2 Typically, the presence of the function reconstruction error and disturbance terms in (14) would lead to a UUB stability result. The RISE term used in (15) robustly accounts for these terms guaranteeing asymptotic tracking with a continuous controller [37] (i.e., compared with similar results that can be obtained by discontinuous sliding mode control). The derivative of the RISE structure includes a ${\rm sgn}(\,\cdot\,)$ term in (17) that allows it to implicitly learn and cancel terms in the stability analysis that are $\mathbb{C}^2$ with bounded time derivatives.

Substituting the control input (15) into (14) yields

$$r = W_{\rm a}^{\rm T}\sigma(V_{\rm a}^{\rm T}x_{\rm a}) - \hat{W}_{\rm a}^{\rm T}\sigma(\hat{V}_{\rm a}^{\rm T}x_{\rm a}) + S + d + \varepsilon_{\rm a} - \mu_{\rm a}. \tag{18}$$

To facilitate the subsequent stability analysis, the time derivative of (18) is expressed as

$$\begin{aligned} \dot{r} = {} & \hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\tilde{V}_{\rm a}^{\rm T}\dot{x}_{\rm a} + \tilde{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\hat{V}_{\rm a}^{\rm T}\dot{x}_{\rm a} \\ & + W_{\rm a}^{\rm T}\sigma'(V_{\rm a}^{\rm T}x_{\rm a})V_{\rm a}^{\rm T}\dot{x}_{\rm a} - W_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\hat{V}_{\rm a}^{\rm T}\dot{x}_{\rm a} \\ & - \hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\tilde{V}_{\rm a}^{\rm T}\dot{x}_{\rm a} - \dot{\hat{W}}_{\rm a}^{\rm T}\sigma(\hat{V}_{\rm a}^{\rm T}x_{\rm a}) \\ & - \hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\dot{\hat{V}}_{\rm a}^{\rm T}x_{\rm a} + \dot{S} + \dot{d} + \dot{\varepsilon}_{\rm a} - \dot{\mu}_{\rm a}, \end{aligned} \tag{19}$$

where $\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a}) \equiv \left.\dfrac{{\rm d}\sigma(V_{\rm a}^{\rm T}x_{\rm a})}{{\rm d}(V_{\rm a}^{\rm T}x_{\rm a})}\right|_{V_{\rm a}^{\rm T}x_{\rm a}=\hat{V}_{\rm a}^{\rm T}x_{\rm a}}$, and $\tilde{W}_{\rm a}(t) \in \mathbb{R}^{(N_{\rm a}+1)\times m}$ and $\tilde{V}_{\rm a}(t) \in \mathbb{R}^{(mn+1)\times N_{\rm a}}$ are the mismatch between the ideal and the estimated weights, and are defined as

$$\tilde{V}_{\rm a} \triangleq V_{\rm a} - \hat{V}_{\rm a}, \;\; \tilde{W}_{\rm a} \triangleq W_{\rm a} - \hat{W}_{\rm a}.$$

The weight update laws for the action NN are designed based on the subsequent stability analysis as

$$\begin{cases} \dot{\hat{W}}_{\rm a} \triangleq {\rm proj}(\Gamma_{\rm aw}\alpha_n\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\hat{V}_{\rm a}^{\rm T}\dot{x}_{\rm a}e_n^{\rm T} \\ \qquad\qquad + \Gamma_{\rm aw}\sigma(\hat{V}_{\rm a}^{\rm T}x_{\rm a})R\hat{W}_{\rm c}^{\rm T}\sigma'(\hat{V}_{\rm c}^{\rm T}e_n)\hat{V}_{\rm c}^{\rm T}), \\ \dot{\hat{V}}_{\rm a} = {\rm proj}(\Gamma_{\rm av}\alpha_n\dot{x}_{\rm a}e_n^{\rm T}\hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a}) \\ \qquad\qquad + \Gamma_{\rm av}x_{\rm a}R\hat{W}_{\rm c}^{\rm T}\sigma'(\hat{V}_{\rm c}^{\rm T}e_n)\hat{V}_{\rm c}^{\rm T}\hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})), \end{cases} \tag{20}$$

where $\Gamma_{\rm aw} \in \mathbb{R}^{(N_{\rm a}+1)\times(N_{\rm a}+1)}, \Gamma_{\rm av} \in \mathbb{R}^{(mn+1)\times(mn+1)}$ are constant, positive definite, symmetric gain matrices, $R(t) \in \mathbb{R}$ is the subsequently designed reinforcement signal, ${\rm proj}(\,\cdot\,)$ is a smooth projection operator utilized to guarantee that the weight estimates $\hat{W}_{\rm a}(t)$ and $\hat{V}_{\rm a}(t)$ remain bounded [38, 39], and $\hat{V}_{\rm c}(t) \in \mathbb{R}^{m\times N_{\rm c}}$ and $\hat{W}_{\rm c}(t) \in \mathbb{R}^{(N_{\rm c}+1)\times 1}$ are the subsequently introduced weight estimates for the critic NN. The NN weight update law in (20) is composite in the sense that it consists of two terms, one of which is affine in the tracking error $e_n(t)$ and the other in the reinforcement signal $R(t)$.

The update law in (20) can be decomposed into two terms

$$\dot{\hat{W}}_{\rm a}^{\rm T} = \chi_{e_n}^{\rm W} + \chi_{\rm R}^{\rm W}, \;\; \dot{\hat{V}}_{\rm a}^{\rm T} = \chi_{e_n}^{\rm V} + \chi_{\rm R}^{\rm V}. \tag{21}$$

Using Assumption 2, (13) and the projection algorithm in (20), the following bounds can be established

$$\begin{cases} \|\chi_{e_n}^{\rm W}\| \leqslant \gamma_1\|e_n\|, \;\; \|\chi_{\rm R}^{\rm W}\| \leqslant \gamma_2|R|, \\ \|\chi_{e_n}^{\rm V}\| \leqslant \gamma_3\|e_n\|, \;\; \|\chi_{\rm R}^{\rm V}\| \leqslant \gamma_4|R|, \end{cases} \tag{22}$$

where $\gamma_1, \gamma_2, \gamma_3$, and $\gamma_4 \in \mathbb{R}$ are known positive constants. Substituting (16), (20), and (21) in (19), and grouping terms, the following expression is obtained

$$\dot{r} = \tilde{N} + N_{\rm R} + N - e_n - (k_{\rm a}+1)r - \beta_1{\rm sgn}(e_n), \tag{23}$$

where the unknown auxiliary terms $\tilde{N}(t) \in \mathbb{R}^m$ and $N_{\rm R}(t) \in \mathbb{R}^m$ are defined as

$$\tilde{N} \triangleq \dot{S} + e_n - \chi_{e_n}^{\rm W}\sigma(\hat{V}_{\rm a}^{\rm T}x_{\rm a}) - \hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\chi_{e_n}^{\rm V}x_{\rm a}, \tag{24}$$

$$N_{\rm R} \triangleq -\chi_{\rm R}^{\rm W}\sigma(\hat{V}_{\rm a}^{\rm T}x_{\rm a}) - \hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\chi_{\rm R}^{\rm V}x_{\rm a}. \tag{25}$$

The auxiliary term $N(t) \in \mathbb{R}^m$ is segregated into two terms as

$$N = N_{\rm d} + N_{\rm B}, \tag{26}$$

where $N_{\rm d}(t) \in \mathbb{R}^m$ is defined as

$$N_{\rm d} \triangleq W_{\rm a}^{\rm T}\sigma'(V_{\rm a}^{\rm T}x_{\rm a})V_{\rm a}^{\rm T}\dot{x}_{\rm a} + \dot{d} + \dot{\varepsilon}_{\rm a}, \tag{27}$$

and $N_{\rm B}(t) \in \mathbb{R}^m$ is further segregated into two terms as

$$N_{\rm B} = N_{\rm B1} + N_{\rm B2}, \tag{28}$$

where $N_{\rm B1}(t), N_{\rm B2}(t) \in \mathbb{R}^m$ are defined as

$$\begin{cases} N_{\rm B1} \triangleq -W_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\hat{V}_{\rm a}^{\rm T}\dot{x}_{\rm a} - \hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\tilde{V}_{\rm a}^{\rm T}\dot{x}_{\rm a}, \\ N_{\rm B2} \triangleq \tilde{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\hat{V}_{\rm a}^{\rm T}\dot{x}_{\rm a} + \hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\tilde{V}_{\rm a}^{\rm T}\dot{x}_{\rm a}. \end{cases} \tag{29}$$

Using the mean value theorem, the following upper bound

can be developed [35, 37]

$$\|\tilde{N}(t)\| \leqslant \rho_1(\|z\|)\|z\|, \tag{30}$$

where $z(t) \in \mathbb{R}^{(n+1)m}$ is defined as

$$z \triangleq [e_1^{\mathrm{T}} \ e_2^{\mathrm{T}} \ \cdots \ e_n^{\mathrm{T}} \ r^{\mathrm{T}}]^{\mathrm{T}}, \tag{31}$$

and the bounding function $\rho_1(\cdot) \in \mathbb{R}$ is a positive, globally invertible, nondecreasing function. Using Assumption 2, Assumption 3, (12), (13), and (20), the following bounds can be developed for (25)–(29):

$$\begin{cases} \|N_{\mathrm{d}}\| \leqslant \zeta_1, \\ \|N_{\mathrm{B1}}\| \leqslant \zeta_2, \ \ \|N_{\mathrm{B2}}\| \leqslant \zeta_3, \\ \|N\| \leqslant \zeta_1 + \zeta_2 + \zeta_3, \\ \|N_{\mathrm{R}}\| \leqslant \zeta_4 |R|. \end{cases} \tag{32}$$

The bounds for the time derivative of (27) and (28) can be developed using Assumption 2, Assumption 3, (12) and (20)

$$\|\dot{N}_{\mathrm{d}}\| \leqslant \zeta_5, \ \ \|\dot{N}_{\mathrm{B}}\| \leqslant \zeta_6 + \zeta_7\|e_n\| + \zeta_8|R|, \tag{33}$$

where $\zeta_i \in \mathbb{R}$ $(i = 1, 2, \ldots, 8)$ are computable positive constants.

Remark 3 The segregation of the auxiliary terms in (21) and (23) follows a typical RISE strategy [37] that is motivated by the desire to separate terms that can be upper bounded by state-dependent terms from terms that can be upper bounded by constants. Specifically, $\tilde{N}(t)$ contains terms upper bounded by tracking error state-dependent terms, $N(t)$ has terms bounded by a constant, and is further segregated into $N_{\mathrm{d}}(t)$ and $N_{\mathrm{B}}(t)$ whose derivatives are bounded by a constant and linear combination of tracking error states, respectively. Similarly, $N_{\mathrm{R}}(t)$ contains reinforcement signal dependent terms. The terms in (28) are further segregated because $N_{\mathrm{B1}}(t)$ will be rejected by the RISE feedback, whereas $N_{\mathrm{B2}}(t)$ will be partially rejected by the RISE feedback and partially canceled by the NN weight update law.

5 Critic NN architecture

In RL literature [1], the critic generates a scalar evaluation signal which is then used to tune the action NN. The critic itself consists of an NN that approximates an evaluation function based on some performance measure. The proposed AC architecture is shown in Fig. 1. The filtered tracking error $e_n(t)$ can be considered as an instantaneous utility function of the plant performance [31, 32].

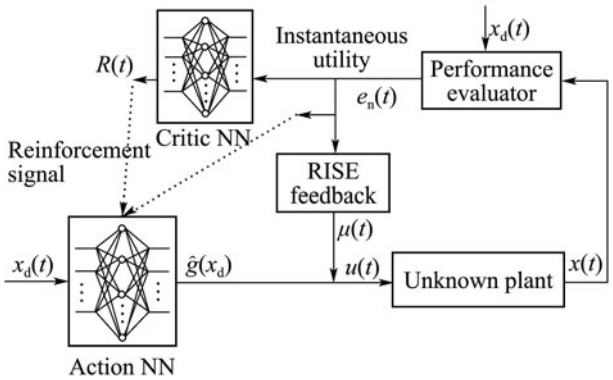

Fig. 1 Architecture of the RISE-based adaptive critic controller.

The reinforcement signal $R(t) \in \mathbb{R}$ is defined as [31]

$$R \triangleq \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n) + \psi, \tag{34}$$

where $\hat{V}_{\mathrm{c}} \in \mathbb{R}^{m\times N_{\mathrm{c}}}$, $\hat{W}_{\mathrm{c}} \in \mathbb{R}^{(N_{\mathrm{c}}+1)\times 1}$, $\sigma(\cdot) \in \mathbb{R}^{N_{\mathrm{c}}+1}$ is the nonlinear activation function, N_{c} are the number of hidden layer neurons of the critic NN, and the performance measure $e_n(t)$ defined in (3) is the input to the critic NN, and $\psi \in \mathbb{R}$ is an auxilliary term generated as

$$\dot{\psi} = \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)\hat{V}_{\mathrm{c}}^{\mathrm{T}}(\mu_{\mathrm{a}} + \alpha_n e_n) - k_{\mathrm{c}}R - \beta_2\mathrm{sgn}(R), \tag{35}$$

where $k_{\mathrm{c}}, \beta_2 \in \mathbb{R}$ are constant positive control gains. The weight update law for the critic NN is generated based on the subsequent stability analysis as

$$\begin{cases} \dot{\hat{W}}_{\mathrm{c}} = \mathrm{proj}(-\Gamma_{\mathrm{cw}}\sigma(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)R - \Gamma_{\mathrm{cw}}\hat{W}_{\mathrm{c}}), \\ \dot{\hat{V}}_{\mathrm{c}} = \mathrm{proj}(-\Gamma_{\mathrm{cv}}e_n\hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)R - \Gamma_{\mathrm{cv}}\hat{V}_{\mathrm{c}}), \end{cases} \tag{36}$$

where $\Gamma_{\mathrm{cw}}, \Gamma_{\mathrm{cv}} \in \mathbb{R}$ are constant positive control gains.

Remark 4 The structure of the reinforcement signal $R(t)$ in (34) is motivated by works such as [31–33], where the reinforcement signal is typically the output of a critic NN that tunes the actor based on a performance measure. The performance measure considered in this paper is the tracking error $e_n(t)$, and the critic weight update laws are designed using a gradient algorithm to minimize the tracking error, as seen from the subsequent stability analysis. The auxiliary term $\psi(t)$ in (33) is a RISE-like robustifying term that is added to account for certain disturbance terms that appear in the error system of the reinforcement learning signal. Specifically, the inclusion of $\psi(t)$ is used to implicitly learn and compensate for disturbances and function reconstruction errors in the reinforcement signal dynamics, yielding an asymptotic tracking result.

To aid the subsequent stability analysis, the time derivative of the reinforcement signal in (34) is obtained as

$$\begin{aligned} \dot{R} = {} & \dot{\hat{W}}_{\mathrm{c}}^{\mathrm{T}}\sigma(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n) + \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)\dot{\hat{V}}_{\mathrm{c}}^{\mathrm{T}}e_n \\ & + \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)\hat{V}_{\mathrm{c}}^{\mathrm{T}}\dot{e}_n + \dot{\psi}. \end{aligned} \tag{37}$$

Using (18), (35), (36) and the Taylor series expansion [6]

$$\sigma(V_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}}) = \sigma(\hat{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}}) + \sigma'(\hat{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}})\tilde{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}} + \mathrm{O}(\tilde{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}})^2,$$

where $\mathrm{O}(\cdot)^2$ represents higher order terms, the expression in (37) can be written as

$$\begin{aligned} \dot{R} = {} & \dot{\hat{W}}_{\mathrm{c}}^{\mathrm{T}}\sigma(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n) + \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)\dot{\hat{V}}_{\mathrm{c}}^{\mathrm{T}}e_n + N_{\mathrm{dc}} + N_{\mathrm{s}} \\ & + \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)\hat{V}_{\mathrm{c}}^{\mathrm{T}}\tilde{W}_{\mathrm{a}}^{\mathrm{T}}\sigma(\hat{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}}) \\ & + \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)\hat{V}_{\mathrm{c}}^{\mathrm{T}}\hat{W}_{\mathrm{a}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}})\tilde{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}} \\ & - k_{\mathrm{c}}R - \beta_2\mathrm{sgn}(R), \end{aligned} \tag{38}$$

where the auxiliary terms $N_{\mathrm{dc}}(t) \in \mathbb{R}$ and $N_{\mathrm{s}}(t) \in \mathbb{R}$ are unknown functions defined as

$$\begin{cases} N_{\mathrm{dc}} \triangleq \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)\hat{V}_{\mathrm{c}}^{\mathrm{T}}\tilde{W}_{\mathrm{a}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}})\tilde{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}} \\ \qquad\quad + W_{\mathrm{a}}^{\mathrm{T}}\mathrm{O}(\tilde{V}_{\mathrm{a}}^{\mathrm{T}}x_{\mathrm{a}})^2 + d + \varepsilon_{\mathrm{a}}, \\ N_{\mathrm{s}} \triangleq \hat{W}_{\mathrm{c}}^{\mathrm{T}}\sigma'(\hat{V}_{\mathrm{c}}^{\mathrm{T}}e_n)\hat{V}_{\mathrm{c}}^{\mathrm{T}}S. \end{cases} \tag{39}$$

Using Assumptions 2 and 3, (12), (36), and the mean value theorem, the following bounds can be developed for (39)

$$\|N_{\mathrm{dc}}\| \leqslant \zeta_9, \ \ \|N_{\mathrm{s}}\| \leqslant \rho_2(\|z\|)\|z\|, \tag{40}$$

where $\zeta_9 \in \mathbb{R}$ is a computable positive constant, and $\rho_2(\cdot) \in \mathbb{R}$ is a positive, globally invertible, nondecreasing function.

6 Stability analysis

Theorem 1 The RISE-based AC controller given in (15) and (34) along with the weight update laws for the action and critic NN given in (20) and (36), respectively, ensure that all system signals are bounded under closed-loop operation and that the tracking error is regulated in the sense that

$$\|e_1(t)\| \to 0 \text{ as } t \to \infty$$

provided the control gains $k_{\rm a}$ and $k_{\rm c}$ are selected sufficiently large based on the initial conditions of the states, α_{n-1}, α_n, β_2, and $k_{\rm c}$, are chosen according to the following sufficient conditions

$$\alpha_{n-1} > \frac{1}{2},\ \alpha_n > \beta_3 + \frac{1}{2},\ \beta_2 > \zeta_9,\ k_{\rm c} > \beta_4, \quad (41)$$

and $\beta_1, \beta_3, \beta_4 \in \mathbb{R}$, introduced in (46), are chosen according to the sufficient conditions [1]

$$\begin{cases} \beta_1 > \max(\zeta_1 + \zeta_2 + \zeta_3,\ \zeta_1 + \zeta_2 + \dfrac{\zeta_5}{\alpha_n} + \dfrac{\zeta_6}{\alpha_n}), \\ \beta_3 > \zeta_7 + \dfrac{\zeta_8}{2}, \\ \beta_4 > \dfrac{\zeta_8}{2}. \end{cases} \quad (42)$$

Proof Let $\mathcal{D} \subset \mathbb{R}^{(n+1)m+3}$ be a domain containing $y(t) = 0$, where $y(t) \in \mathbb{R}^{(n+1)m+3}$ is defined as

$$y \triangleq [z^{\rm T}\ \ R\ \ \sqrt{P}\ \ \sqrt{Q}]^{\rm T}, \quad (43)$$

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

$$Q \triangleq \frac{1}{2}{\rm tr}(\tilde{W}_{\rm a}^{\rm T}\Gamma_{\rm aw}^{-1}\tilde{W}_{\rm a}) + \frac{1}{2}{\rm tr}(\tilde{V}_{\rm a}^{\rm T}\Gamma_{\rm av}^{-1}\tilde{V}_{\rm a}) + \frac{1}{2}{\rm tr}(\hat{W}_{\rm c}^{\rm T}\hat{W}_{\rm c}) + \frac{1}{2}{\rm tr}(\hat{V}_{\rm c}^{\rm T}\hat{V}_{\rm c}), \quad (44)$$

where ${\rm tr}(\cdot)$ is the trace of a matrix. The auxiliary function $P(t) \in \mathbb{R}$ in (43) is the generalized solution to the differential equation

$$\begin{cases} \dot{P} = -L, \\ P(0) = \beta_1 \sum_{i=1}^{m} |e_{ni}(0)| - e_n(0)^{\rm T}N(0), \end{cases} \quad (45)$$

where the subscript $i = 1, 2, \ldots, m$ denotes the ith element of the vector, and the auxiliary function $L(t) \in \mathbb{R}$ is defined as

$$L \triangleq r^{\rm T}(N_{\rm d} + N_{\rm B1} - \beta_1 {\rm sgn}(e_n)) + \dot{e}_n^{\rm T}N_{\rm B2} - \beta_3\|e_n\|^2 - \beta_4|R|^2, \quad (46)$$

where $\beta_1, \beta_3, \beta_4 \in \mathbb{R}$ are chosen according to the sufficient conditions in (42). Provided the sufficient conditions introduced in (42) are satisfied, then $P(t) \geqslant 0$. From (23), (32), (38) and (40), some disturbance terms in the closed-loop error systems are bounded by a constant. Typically, such terms (e.g., NN reconstruction error) lead to a UUB stability result. The definition of $P(t)$ is motivated by the RISE control structure to compensate for such disturbances so that an asymptotic tracking result is obtained.

Let $V(y) : \mathcal{D} \times [0, \infty) \to \mathbb{R}$ be a Lipschitz continuous regular positive definite function defined as

$$V \triangleq \frac{1}{2}z^{\rm T}z + \frac{1}{2}R^2 + P + Q, \quad (47)$$

which satisfies the following inequalities:

$$U_1(y) \leqslant V(y) \leqslant U_2(y), \quad (48)$$

where $U_1(y)$, $U_2(y) \in \mathbb{R}$ are continuous positive definite functions. From (3), (4), (23), (38), (44), and (45), the differential equations of the closed-loop system are continuous except in the set $\{y | e_n = 0 \text{ or } R = 0\}$. Using Filippov's differential inclusion [36, 40–42], the existence of solutions can be established for $\dot{y} = f(y)$, where $f(y) \in \mathbb{R}^{(n+1)m+3}$ denotes the right-hand side of the the closed-loop error signals. Under Filippov's framework, a generalized Lyapunov stability theory can be used (see [42–45] for further details) to establish strong stability of the closed-loop system. The generalized time derivative of (47) exists almost everywhere (a.e.), and $\dot{V}(y) \in \dot{\tilde{V}}^{\rm a.e.}(y)$, where

$$\dot{\tilde{V}} = \bigcap_{\xi \in \partial V(y)} \xi^{\rm T}K[\dot{z}^{\rm T}\ \ \dot{R}\ \ \frac{1}{2}P^{-\frac{1}{2}}\dot{P}\ \ \frac{1}{2}Q^{-\frac{1}{2}}\dot{Q}]^{\rm T},$$

where ∂V is the generalized gradient of V [43], and $K[\cdot]$ is defined as [44, 45]

$$K[f](y) \triangleq \bigcap_{\delta > 0} \bigcap_{\mu N = 0} \overline{\rm co} f(B(y, \delta) - N),$$

where $\bigcap_{\mu N = 0}$ denotes the intersection of all sets N of Lebesgue measure zero, $\overline{\rm co}$ denotes convex closure, and $B(y, \delta)$ represents a ball of radius δ around y. Because $V(y)$ is a Lipschitz continuous regular function,

$$\begin{aligned} \dot{\tilde{V}} &= \nabla V^{\rm T}K[\dot{z}^{\rm T}\ \ \dot{R}\ \ \frac{1}{2}P^{-\frac{1}{2}}\dot{P}\ \ \frac{1}{2}Q^{-\frac{1}{2}}\dot{Q}]^{\rm T} \\ &= [z^{\rm T}\ \ R\ \ 2P^{\frac{1}{2}}\ \ 2Q^{\frac{1}{2}}]K[\dot{z}^{\rm T}\ \ \dot{R}\ \ \frac{1}{2}P^{-\frac{1}{2}}\dot{P}\ \ \frac{1}{2}Q^{-\frac{1}{2}}\dot{Q}]^{\rm T}. \end{aligned}$$

Using the calculus for $K[\cdot]$ from [45] and the dynamics from (23), (38), (44), and (45), and splitting $k_{\rm c}$ as $k_{\rm c} = k_{\rm c1} + k_{\rm c2}$, yields

$$\begin{aligned} \dot{\tilde{V}} \subset\ & r^{\rm T}(\tilde{N} + N_{\rm R} + N - e_n - (k_{\rm a}+1)r - \beta_1 K[{\rm sgn}(e_n)]) \\ & + \sum_{i=1}^{n} e_i^{\rm T}\dot{e}_i + R(\dot{\hat{W}}_{\rm c}^{\rm T}\sigma(\hat{V}_{\rm c}^{\rm T}e_n) + N_{\rm dc} + N_{\rm s}) \\ & + \hat{W}_{\rm c}^{\rm T}\sigma'(\hat{V}_{\rm c}^{\rm T}e_n)\dot{\hat{V}}_{\rm c}^{\rm T}e_n R - \beta_2 R K[{\rm sgn}(R)] \\ & + \hat{W}_{\rm c}^{\rm T}\sigma'(\hat{V}_{\rm c}^{\rm T}e_n)\hat{V}_{\rm c}^{\rm T}\tilde{W}_{\rm a}^{\rm T}\sigma(\hat{V}_{\rm a}^{\rm T}x_{\rm a})R - k_{\rm c}R^2 \\ & + \hat{W}_{\rm c}^{\rm T}\sigma'(\hat{V}_{\rm c}^{\rm T}e_n)\hat{V}_{\rm c}^{\rm T}\hat{W}_{\rm a}^{\rm T}\sigma'(\hat{V}_{\rm a}^{\rm T}x_{\rm a})\tilde{V}_{\rm a}^{\rm T}x_{\rm a}R \\ & - r^{\rm T}(N_{\rm d} + N_{\rm B1} - \beta_1 K[{\rm sgn}(e_n)]) \\ & - \dot{e}_n(t)^{\rm T}N_{\rm B2} + \beta_3\|e_n\|^2 + \beta_4|R|^2 \\ & - \frac{1}{2}{\rm tr}(\tilde{W}_{\rm a}^{\rm T}\Gamma_{\rm aw}^{-1}\dot{\hat{W}}_{\rm a}) - \frac{1}{2}{\rm tr}(\tilde{V}_{\rm a}^{\rm T}\Gamma_{\rm av}^{-1}\dot{\hat{V}}_{\rm a}) \\ & - \frac{1}{2}{\rm tr}(\hat{W}_{\rm c}^{\rm T}\dot{\hat{W}}_{\rm c}) - \frac{1}{2}{\rm tr}(\hat{V}_{\rm c}^{\rm T}\dot{\hat{V}}_{\rm c}) \\ =\ & -\sum_{i=1}^{n}\alpha_i\|e_i\|^2 + e_{n-1}^{\rm T}e_n - \|r\|^2 - (k_{\rm c1} + k_{\rm c2})|R|^2 \\ & + r^{\rm T}(\tilde{N} + N_{\rm R} - k_{\rm a}r) + R(N_{\rm dc} + N_{\rm s} - k_{\rm c}R) \\ & - \beta_2|R| - \Gamma_{\rm cw}|R|^2\|\sigma(\hat{V}_{\rm c}^{\rm T}e_n)\|^2 - \Gamma_{\rm cw}\|\hat{W}_{\rm c}\|^2 \\ & + 2\Gamma_{\rm cw}|R|\|\hat{W}_{\rm c}\sigma(\hat{V}_{\rm c}^{\rm T}e_n)\| + \beta_3\|e_n\|^2 + \beta_4|R|^2 \\ & - \Gamma_{\rm cv}(\|\hat{V}_{\rm c}\|^2 - 2\|\hat{W}_{\rm c}^{\rm T}\sigma'(\hat{V}_{\rm c}^{\rm T}e_n)\|\|e_n\|\ \|\hat{V}_{\rm c}\||R|) \\ & - \Gamma_{\rm cv}\|\hat{W}_{\rm c}^{\rm T}\sigma'(\hat{V}_{\rm c}^{\rm T}e_n)\|^2\|e_n\|^2|R|^2, \end{aligned} \quad (49)$$

where the NN weight update laws from (20), (36), and the fact that $(r^{\rm T} - r^{\rm T})_i{\rm SGN}(e_{ni}) = 0$ is used (the subscript i denotes the ith element), where $K[{\rm sgn}(e_n)] = {\rm SGN}(e_n)$ [45], such that ${\rm SGN}(e_{ni}) = 1$ if $e_{ni} > 0$, $[-1, 1]$ if $e_{ni} = 0$,

[1] The derivation of the sufficient conditions in (42) is provided in Appendix.

and -1 if $e_{ni} < 0$. Upper bounding the expression in (49) using (30), (32), and (40), yields

$$\begin{aligned}\dot{\tilde{V}} \leqslant & -\sum_{i=1}^{n-2}\alpha_i\|e_i\|^2 - (\alpha_{n-1}-\frac{1}{2})\|e_{n-1}\|^2 - \|r\|^2 \\ & -(\alpha_n-\beta_3-\frac{1}{2})\|e_n\|^2 - (k_{c1}-\beta_4)|R|^2 \\ & +(\zeta_9-\beta_2)|R| - [k_a\|r\|^2 - \rho_1(\|z\|)\|z\|\,\|r\|] \\ & -[k_{c2}|R|^2 - (\rho_2(\|z\|)+\zeta_4)|R|\|z\|]. \end{aligned} \quad (50)$$

Provided the gains are selected according to (41), (50) can be further upper bounded by completing the squares as

$$\begin{aligned}\dot{\tilde{V}} &\leqslant -\lambda\|z\|^2 + \frac{\rho^2(\|z\|)\|z\|^2}{4k} - (k_{c1}-\beta_4)|R|^2 \\ &\leqslant -U(y),\ \ \forall y\in\mathcal{D}, \end{aligned} \quad (51)$$

where $k \triangleq \min(k_a, k_{c2})$ and $\lambda \in \mathbb{R}$ is a positive constant defined as

$$\lambda = \min\{\,\alpha_1,\alpha_2,\ldots,\alpha_{n-2},\alpha_{n-1}-\frac{1}{2},\ \alpha_n-\beta_3-\frac{1}{2},1\}.$$

In (51), $\rho(\cdot)\in\mathbb{R}$ is a positive, globally invertible, nondecreasing function defined as

$$\rho^2(\|z\|) = \rho_1^2(\|z\|) + (\rho_2(\|z\|)+\zeta_4)^2.$$

In (51), $U(y) = c\|[z^{\mathrm{T}}\ \ R]^{\mathrm{T}}\|^2$, for some positive constant c, is a continuous, positive semidefinite function defined on the domain

$$\mathcal{D} \triangleq \{y(t)\in\mathbb{R}^{(n+1)m+3}\big|\|y\| \leqslant \rho^{-1}(2\sqrt{\lambda k})\}.$$

The size of the domain $\mathcal{D}$ can be increased by increasing k. The result in (51) indicates that $\dot{V}(y) \leqslant -U(y), \forall \dot{V}(y) \in \dot{\tilde{V}}(y), \forall y \in \mathcal{D}$. The inequalities in (48) and (51) can be used to show that $V(y) \in \mathcal{L}_\infty$ in $\mathcal{D}$; hence, $e_1(t)$, $e_2(t),\ldots,e_n(t)$, $r(t)$ and $R(t)\in\mathcal{L}_\infty$ in $\mathcal{D}$. Standard linear analysis methods can be used along with (1)–(5) to prove that $\dot{e}_1(t)$, $\dot{e}_2(t),\ldots,\dot{e}_n(t)$, $x^{(i)}(t)\in\mathcal{L}_\infty$ $(i=0,1,2)$ in $\mathcal{D}$. Furthermore, Assumptions 1 and 3 can be used to conclude that $u(t)\in\mathcal{L}_\infty$ in $\mathcal{D}$. From these results, (12), (13), (19), (20), and (34)–(37) can be used to conclude that $\dot{r}(t),\psi(t),\dot{R}(t)\in\mathcal{L}_\infty$ in $\mathcal{D}$. Hence, $U(y)$ is uniformly continuous in $\mathcal{D}$. Let $\mathcal{S}\subset\mathcal{D}$ denote a set defined as follows:

$$\mathcal{S} \triangleq \{y(t)\subset\mathcal{D}\big|U_2(y(t)) < \lambda_1(\rho^{-1}(2\sqrt{\lambda k}))^2\}. \quad (52)$$

The region of attraction in (52) can be made arbitrarily large to include any initial condition by increasing the control gain k (i.e., a semiglobal type of stability result), and hence

$$\|e_1(t)\|,\ |R|\to 0\ \text{ as }\ t\to\infty,\ \forall y(0)\in\mathcal{S}.$$

7 Experimental results

To test the performance of the proposed AC-based approach, the controller in (15), (20), (34)–(36) was implemented on a two-link robot manipulator, where two aluminum links are mounted on a 240 N·m (first link) and a 20 N·m (second link) switched reluctance motor. The motor resolvers provide rotor position measurements with a resolution of 614400 pulses/revolution, and a standard backwards difference algorithm is used to numerically determine angular velocity from the encoder readings. The two-link revolute robot is modeled as an Euler-Lagrange system with the following dynamics

$$M(q)\ddot{q} + V_m(q,\dot{q})\dot{q} + F(\dot{q}) + \tau_d = \tau, \quad (53)$$

where $M(q)\in\mathbb{R}^{2\times2}$ denotes the inertia matrix, $V_m(q,\dot{q})\in\mathbb{R}^{2\times2}$ denotes the centripetal-Coriolis matrix, $F(\dot{q})\in\mathbb{R}^2$ denotes friction, $\tau_d(t)\in\mathbb{R}^2$ denotes an unknown external disturbance, $\tau(t)\in\mathbb{R}^2$ represents the control torque, and $q(t),\dot{q}(t),\ddot{q}(t)\in\mathbb{R}^2$ denote the link position, velocity and acceleration. The dynamics in (53) can be transformed into the Brunovsky form as

$$\begin{cases}\dot{x}_1 = x_2,\\ \dot{x}_2 = g(x)+u+d,\end{cases} \quad (54)$$

where

$$\begin{aligned}&x_1 \triangleq q,\ \ x_2\triangleq\dot{q},\ \ x=[x_1\ \ x_2]^{\mathrm{T}},\\ &g(x)\triangleq -M^{-1}(q)[V_m(q,\dot{q})\dot{q}+F(\dot{q})],\\ &u\triangleq M^{-1}(q)\tau(t),\end{aligned}$$

[2]

and $d \triangleq M^{-1}(q)\tau(t)$. The control objective is to track a desired link trajectory, selected as (in degrees)

$$q_d(t) = 60\sin(2.5t)(1-\mathrm{e}^{-0.01t^3}).$$

Two controllers are implemented on the system, both having the same expression for the control $u(t)$ as in (15); however, they differ in the NN weight update laws. The first controller (denoted by NN+RISE) employs a standard NN gradient-based weight update law that is affine in the tracking error, given as

$$\begin{aligned}\dot{\hat{W}}_a &= \mathrm{proj}(\Gamma_{aw}\alpha_n\sigma'(\hat{V}_a^{\mathrm{T}}x_a)\hat{V}_a^{\mathrm{T}}\dot{x}_a e_n^{\mathrm{T}}),\\ \dot{\hat{V}}_a &= \mathrm{proj}(\Gamma_{av}\alpha_n\dot{x}_a e_n^{\mathrm{T}}\hat{W}_a^{\mathrm{T}}\sigma'(\hat{V}_a^{\mathrm{T}}x_a)).\end{aligned}$$

The proposed AC-based controller (denoted by AC+RISE) uses a composite weight update law, consisting of a gradient-based term and a reinforcement-based term, as in (20), where the reinforcement term is generated from the critic architecture in (34). For the NN+RISE controller, the initial weights of the NN, $\hat{W}_a(0)$ is chosen to be zero, whereas $\hat{V}_a(0)$ is randomly initialized in $[-1,1]$, such that it forms a basis [46]. The input to the action NN is chosen as $x_a = [1\ \ q_d^{\mathrm{T}}\ \ \dot{q}_d^{\mathrm{T}}]$, and the number of hidden layer neurons are chosen by trial and error as $N_a = 10$. All other states are initialized to zero. A sigmoid activation function is chosen for the NN and the adaptation gains are selected as $\Gamma_{aw} = I_{11}, \Gamma_{av} = 0.1I_{11}$, with feedback gains selected as $\alpha_1 = \mathrm{diag}(10,15)$, $\alpha_2 = \mathrm{diag}(20,15)$, $k_a = (20,15)$ and $\beta_1 = \mathrm{diag}(2,1)$. For the AC+RISE controller, the critic is added to the NN+RISE by including an additional RL term in the weight update law of the action NN. The actor NN and the RISE term in AC+RISE use the same gains as NN+RISE. The number of hidden layer neurons for the critic are selected by trial and error as $N_c = 3$. The initial critic NN weights $\hat{W}_c(0)$ and $\hat{V}_c(0)$ are randomly chosen in $[-1,1]$. The control gains for the critic are selected as $k_c = 5, \beta_2 = 0.1, \Gamma_{cw} = 0.4, \Gamma_{cv} = 1$. Experiments for both controllers were repeated 10 consecutive times with the same gains to check the repeatability and accuracy of results. For each run, the RMS values of the tracking error $e_1(t)$ and torques $\tau(t)$ are calculated. A one-tailed unpaired t-test is performed with a significance level of $\alpha = 0.05$. A summary of comparative results with the two controllers are tabulated in Tables 1 and 2.

Table 1 Summarized experimental results and P values of one-tailed unpaired t-test for Link 1.

Experiment	RMS error (Link 1)		Torque (Link 1)/(N·m)	
	NN+RISE	AC+RISE	NN+RISE	AC+RISE
Maximum	0.143°	0.123°	15.937	16.013
Minimum	0.101°	0.098°	15.451	15.470
Mean	0.125°	0.108°	15.687	15.764
Standard deviation	0.014°	0.009°	0.152	0.148
$P\,(T \leqslant t)$	0.003^*		0.134	

* denotes statistically significant value.

Table 2 Summarized experimental results and P values of one-tailed unpaired t-test for Link 2.

Experiment	RMS error (Link 2)		Torque (Link 2)/(N·m)	
	NN+RISE	AC+RISE	NN+RISE	AC+RISE
Maximum	0.161°	0.138°	1.856	1.858
Minimum	0.112°	0.107°	1.717	1.670
Mean	0.137°	0.127°	1.783	1.753
Standard deviation	0.015°	0.010°	0.045	0.054
$P\,(T \leqslant t)$	0.046^*		0.098	

* denotes statistically significant value.

Tables 1 and 2 indicate that the AC+RISE controller has statistically smaller mean RMS errors for Link 1 ($P = 0.003$) and Link 2 ($P = 0.046$) as compared with the NN+RISE controller. The AC+RISE controller, while having a reduced error, uses approximately the same amount of control torque (statistically insignificant difference) as NN+RISE. The results indicate that the mean RMS position tracking errors for Links 1 and 2 are approximately 14% and 7% smaller for the proposed AC+RISE controller. The plots for tracking error and control torques are shown for a typical experiment in Figs. 2 and 3.

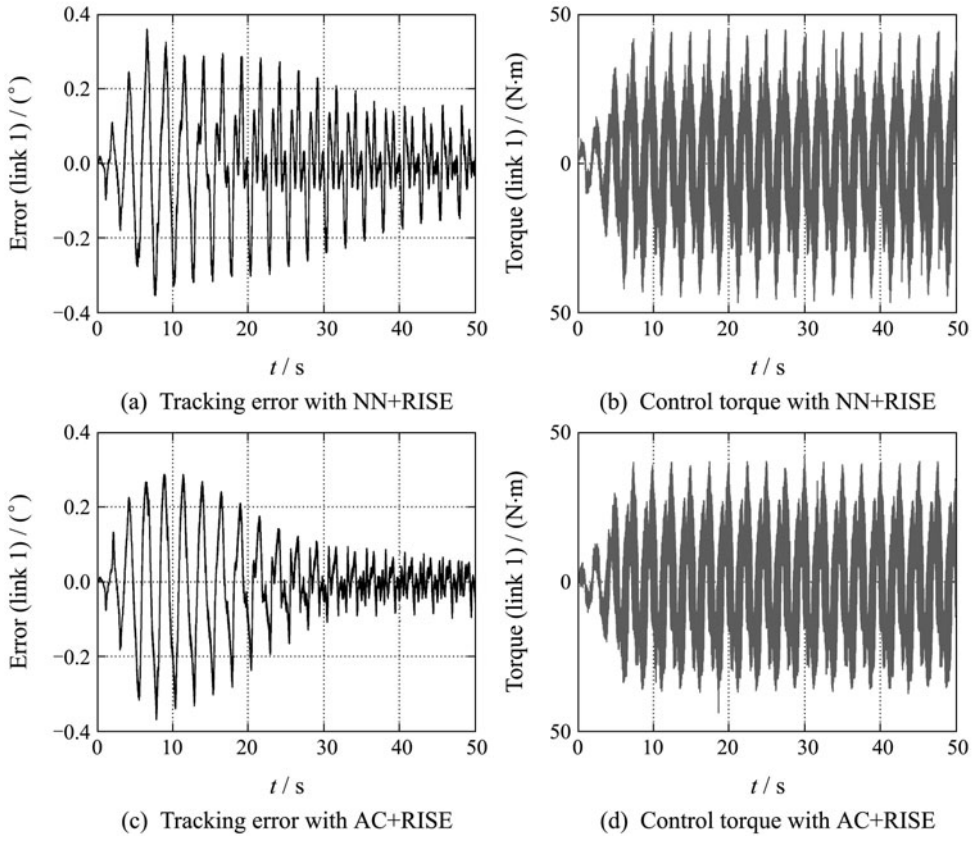

Fig. 2 Comparison of tracking errors and torques between NN+RISE and AC+RISE for Link 1.

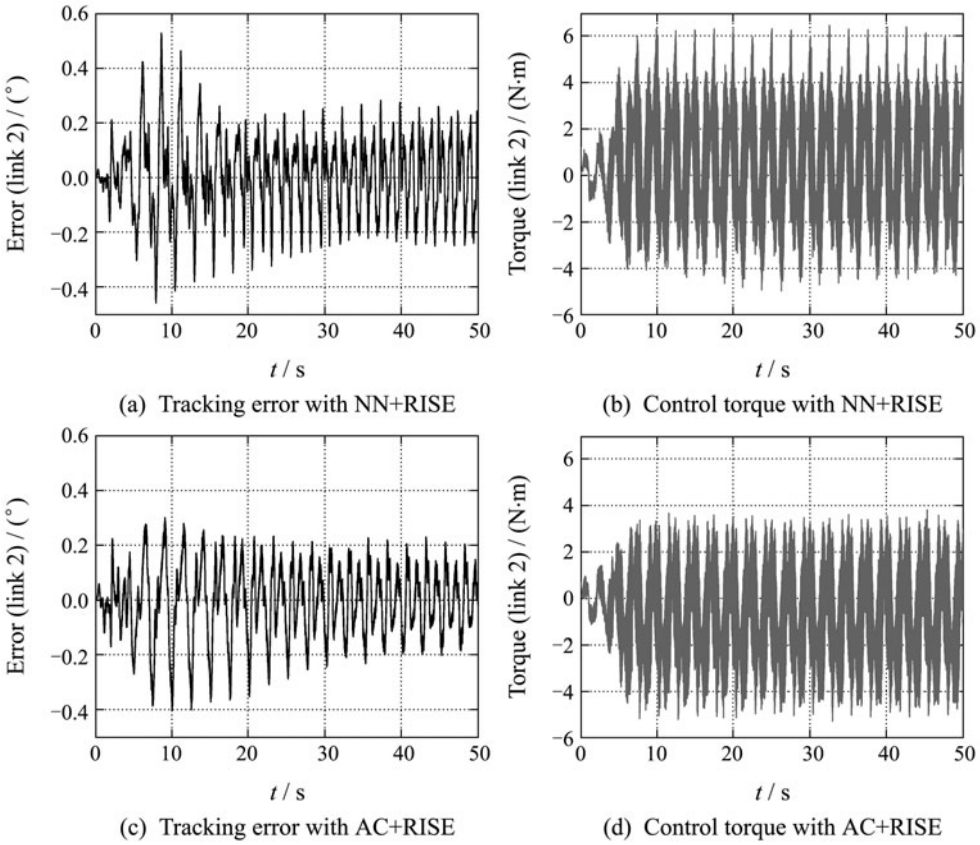

Fig. 3 Comparison of tracking errors and torques between NN+RISE and AC+RISE for Link 2.

8 Conclusions

In this paper, a non-DP based AC controller is developed for a class of uncertain nonlinear systems with additive bounded disturbances. The main contribution of the paper is the combination of the continuous RISE feedback with the AC architecture to guarantee the asymptotic tracking of the nonlinear system. The feedforward action NN approximates the nonlinear system dynamics and the robust feedback (RISE) rejects the NN functional reconstruction error and disturbances. In addition, the action NN is trained online using a combination of tracking error and a reinforcement signal, generated by the critic. Experimental results and t-test analysis demonstrate faster convergence of the tracking error when a reinforcement learning term is included in the NN weight update laws. Although the proposed method guarantees asymptotic tracking, a limitation of the controller is that it does not ensure optimality, which is a common feature (at least approximate optimal control) of other DP-based RL controllers. Future work will focus on developing optimal closed-loop stable RL controllers.

References

[1] R. S. Sutton, A. G. Barto. *Introduction to Reinforcement Learning*. Cambridge: MIT Press, 1998.

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

[3] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics Control Signals System*, 1989, 2(4): 303 – 314.

[4] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 1993, 39(3): 930 – 945.

[5] P. J. Werbos. A menu of designs for reinforcement learning over time. *Neural Networks for Control*, Cambridge: MIT Press, 1990: 67 – 95.

[6] F. L. Lewis, R. Selmic, J. Campos. *Neuro-fuzzy Control of Industrial Systems with Actuator Nonlinearities*. Philadelphia: SIAM, 2002.

[7] P. J. Werbos. Building and understanding adaptive systems: a statistical/numerical approach to factory automation and brain research. *IEEE Transactions on Systems, Man, and Cybernetics*, 1987, 17(1): 7 – 20.

[8] C. J. C. H. Watkins, P. Dayan. Q-learning. *Machine Learning*, 1992, 8(3): 279 – 292.

[9] A. G. Barto, R. S. Sutton, C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 1983, 13(5): 834 – 846.

[10] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 1988, 3(1): 9 – 44.

[11] R. Bellman. *Dynamic Programming*. New York: Dover Publications Inc., 2003.

[12] P. J. Werbos. Approximate dynamic programming for real-time control and neural modeling. D. A. White, D. A. Sofge, eds. *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, New York: Van Nostrand Reinhold, 1992: 493 – 525.

[13] D. V. Prokhorov, D. C. Wunsch. Adaptive critic designs. *IEEE Transactions on Neural Networks*, 1997, 8(5): 997 – 1007.

[14] S. Ferrari, R. F. Stengel. An adaptive critic global controller. *American Control Conference*, New York: IEEE, 2002: 2665 – 2670.

[15] J. Si, Y. Wang. On-line learning control by association and reinforcement. *IEEE Transactions on Neural Networks*, 2001, 12(2): 264 – 276.

[16] J. Si, A. Barto, W. Powell, et al. *Handbook of Learning and Approximate Dynamic Programming*. Piscataway: Wiley-IEEE Press, 2004.

[17] G. K. Venayagamoorthy, R. G. Harley, D. C. Wunsch. Comparison of heuristic dynamic programming and dual heuristic programming adaptive critics for neurocontrol of a turbogenerator. *IEEE Transactions on Neural Networks*, 2002, 13(3): 764 – 773.

[18] G. K. Venayagamoorthy, R. G. Harley, D. C. Wunsch. Dual heuristic programming excitation neurocontrol for generators in a multimachine power system. *IEEE Transactions on Industry Applications*, 2003, 39(2): 382 – 394.

[19] S. Ferrari, R. F. Stengel. Online adaptive critic flight control. *Journal of Guidance Control and Dynamics*, 2004, 27(5): 777 – 786.

[20] S. Jagannathan, G. Galan. Adaptive critic neural network-based object grasping control using a three-finger gripper. *IEEE Transactions on Neural Networks*, 2004, 15(2): 395 – 407.

[21] D. Han, S. N. Balakrishnan. State-constrained agile missile control with adaptive-critic-based neural networks. *IEEE Transactions on Control Systems Technology*, 2002, 10(4): 481 – 489.

[22] C. W. Anderson, D. Hittle, M. Kretchmar, et al. Robust reinforcement learning for heating, ventilation, and air conditioning control of buildings. J. Si, A. G. Barto, W. B. Powell, et al., eds. *Handbook of Learning and Approximate Dynamic Programming*, Piscataway: Wiley-IEEE Press, 2004: 517 – 529.

[23] R. Padhi, S. N. Balakrishnan, T. Randolph. Adaptive-critic based optimal neuro control synthesis for distributed parameter systems. *Automatica*, 2001, 37(8): 1223 – 1234.

[24] D. V. Prokhorov, R. A. Santiago, D. C. Wunsch. Adaptive critic designs: a case study for neurocontrol. *Neural Networks*, 1995, 8(9): 1367 – 1372.

[25] J. J. Murray, C. J. Cox, G. G. Lendaris, et al. Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, 2002, 32(2): 140 – 15.

[26] X. Liu, S. N. Balakrishnan. Convergence analysis of adaptive critic based optimal control. *American Control Conference*, New York: IEEE, 2000: 1929 – 1933.

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

[28] D. Vrabie, M. Abu-Khalaf, F. L. Lewis, et al. Continuous-time ADP for linear systems with partially unknown dynamics. *Proceedings of IEEE International Symposium Approximately Dynanic Programming Reinforcement Learning*, New York: IEEE, 2007: 247 – 253.

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

[30] K. G. Vamvoudakis, F. L. Lewis. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. *Automatica*, 2010, 46(5): 878 – 888.

[31] J. Campos, F. L. Lewis. Adaptive critic neural network for feedforward compensation. *American Control Conference*, New York: IEEE, 1999: 2813 – 2818.

[32] O. Kuljaca, F. L. Lewis. Adaptive critic design using non-linear network structures. *International Journal of Adaptive Control and Signal Processing*, 2003, 17(6): 431 – 445.

[33] Y. H. Kim, F. L. Lewis. *High-level Feedback Control with Neural Networks*. Hackensack: World Scientific Publishing Co., 1998.

[34] P. M. Patre, W. MacKunis, C. Makkar, et al. Asymptotic tracking for systems with structured and unstructured uncertainties. *IEEE Transactions on Control Systems Technology*, 2008, 16(2): 373 – 379.

[35] B. Xian, D. M. Dawson, M. S. de Queiroz, et al. A continuous asymptotic tracking control strategy for uncertain nonlinear systems. *IEEE Transactions on Automatic Control*, 2004, 49(7): 1206 – 1211.

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

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

[38] M. Krstic, P. V. Kokotovic, I. Kanellakopoulos. *Nonlinear and Adaptive Control Design*, New York: John Wiley & Sons, 1995.

[39] W. E. Dixon, A. Behal, D. M. Dawson, et al. *Nonlinear Control of Engineering Systems: A Lyapunov-based Approach*. Boston: Birkhuser, 2003.

[40] A. Filippov. *Differential Equations with Discontinuous Right-hand Side*. Netherlands: Kluwer Academic Publishers, 1988.

[41] G. V. Smirnov. *Introduction to the Theory of Differential Inclusions*. New York: American Mathematical Society, 2002.

[42] J. P. Aubin, H. Frankowska. *Set-valued Analysis*. Boston: Birkhuser, 2008.

[43] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Philadelphia: SIAM, 1990.

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

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

[46] F. L. Lewis. Nonlinear network structures for feedback control. *Asian Journal of Control*, 1999, 1(4): 205 – 228.

Appendix

Derivation of sufficient conditions in (42) Integrating (46), the following expression is obtained:

$$\int_0^t L(\tau)\mathrm{d}\tau = \int_0^t \{r^{\mathrm{T}}(N_{\mathrm{d}} + N_{\mathrm{B1}} - \beta_1 \mathrm{sgn}(e_n)) + \dot{e}_n^{\mathrm{T}} N_{\mathrm{B2}} - \beta_3 \|e_n\|^2 - \beta_4 |R|^2\}\mathrm{d}\tau.$$

Using (4), integrating the first integral by parts, and integrating the second integral yields

$$\begin{aligned}\int_0^t L(\tau)\mathrm{d}\tau = {} & e_n^{\mathrm{T}} N - e_n^{\mathrm{T}}(0)N(0) - \int_0^t e_n^{\mathrm{T}}(\dot{N}_{\mathrm{B}} + \dot{N}_{\mathrm{d}})\mathrm{d}\tau \\ & + \beta_1 \sum_{i=1}^{m} |e_{ni}(0)| - \beta_1 \sum_{i=1}^{m} |e_{ni}(t)| \\ & + \int_0^t \alpha_n e_n^{\mathrm{T}}(N_{\mathrm{d}} + N_{\mathrm{B1}} - \beta_1 \mathrm{sgn}(e_n))\mathrm{d}\tau \\ & - \int_0^t (\beta_3 \|e_n\|^2 + \beta_4 |R|^2)\mathrm{d}\tau.\end{aligned}$$

Using the fact that $\|e_n\| \leqslant \sum_{i=1}^{m} |e_{ni}|$, and using the bounds in (32) and (33), yields

$$\begin{aligned}\int_0^t L(\tau)\mathrm{d}\tau \leqslant {} & \beta_1 \sum_{i=1}^{m} |e_{ni}(0)| - e_n^{\mathrm{T}}(0)N(0) \\ & -(\beta_1 - \zeta_1 - \zeta_2 - \zeta_3)\|e_n\| \\ & - \int_0^t (\beta_3 - \zeta_7 - \frac{\zeta_8}{2})\|e_n\|^2 \mathrm{d}\tau \\ & - \int_0^t (\beta_4 - \frac{\zeta_8}{2})|R|^2 \mathrm{d}\tau \\ & + \int_0^t \alpha_n \|e_n\|(\zeta_1 + \zeta_2 + \frac{\zeta_5}{\alpha_n} + \frac{\zeta_6}{\alpha_n} - \beta_1)\mathrm{d}\tau.\end{aligned}$$

If the sufficient conditions in (42) are satisfied, then the following inequality holds

$$\begin{cases} \int_0^t L(\tau)\mathrm{d}\tau \leqslant \beta_1 \sum_{i=1}^{m} |e_{ni}(0)| - e_n(0)^{\mathrm{T}} N(0), \\ \int_0^t L(\tau)\mathrm{d}\tau \leqslant P(0). \end{cases} \tag{a1}$$

Using (a1) and (45), it can be shown that $P(t) \geqslant 0$.

Shubhendu BHASIN received his B.E. degree in Manufacturing Processes and Automation in 2004 from NSIT, University of Delhi, India, and M.S. degree in Mechanical Engineering in 2009 from the University of Florida. He is currently a Ph.D. candidate at the Nonlinear Control and Robotics Lab at the University of Florida. 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. E-mail: sbhasin@ufl.edu.

Nitin SHARMA received his Ph.D. in 2010 from the Department of Mechanical and Aerospace Engineering at the University of Florida. He is a recipient of the 2009 O. Hugo Schuck Award and Best Student Paper Award in Robotics at the 2009 ASME Dynamic Systems and Controls Conference. He was also a finalist for the Best Student Paper Award at the 2008 IEEE Multi-Conference on Systems and Control. Currently, he is a postdoctoral fellow in the Department of Physiology at the University of Alberta, Edmonton, Canada. His research interests include intelligent and robust control of functional electrical stimulation (FES), modeling, optimization, and control of FES-elicited walking, and control of uncertain nonlinear systems with input and state delays. E-mail: nitin2@ualberta.ca.

Parag PATRE received his B.Tech. degree in Mechanical Engineering from the Indian Institute of Technology Madras, India, in 2004. Following this he was with Larsen and Toubro Limited, India until 2005, when he joined the Graduate School at the University of Florida. He received his M.S. and Ph.D. degrees in Mechanical Engineering in 2007 and 2009, respectively, and is currently a NASA postdoctoral program fellow at the NASA Langley Research Center, Hampton, VA. His areas of research interest are Lyapunov-based design and analysis of control methods for uncertain nonlinear systems, robust and adaptive control, control of robots, aircraft control, control in the presence of actuator failures, decentralized adaptive control, and neural networks. E-mail: parag.patre@gmail.com.

Warren DIXON received his Ph.D. in 2000 from the Department of Electrical and Computer Engineering from Clemson University. He was a Eugene P. Wigner Fellow at Oak Ridge National Laboratory (ORNL) until joining the University of Florida Mechanical and Aerospace Engineering Department in 2004. His research interests include the development and application of Lyapunov-based control techniques for uncertain nonlinear systems. His work has been recognized by the 2009 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. He is a senior member of IEEE, and an associate editor for ASME Journal of Dynamic Systems, Measurement and Control, Automatica, IEEE Transactions on Systems, Man and Cybernetics – Part B: Cybernetics, International Journal of Robust and Nonlinear Control, and Journal of Robotics. E-mail: wdixon@ufl.edu.