Automatica 51 (2015) 40–48

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Approximate optimal trajectory tracking for continuous-time nonlinear systems[☆]

Rushikesh Kamalapurkar[a], Huyen Dinh[b], Shubhendu Bhasin[c], Warren E. Dixon[a]

[a] Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, USA

[b] Department of Mechanical Engineering, University of Transport and Communications, Hanoi, Viet Nam

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

ARTICLE INFO

Article history:
Received 21 September 2012
Received in revised form
23 May 2014
Accepted 29 September 2014
Available online 17 November 2014

Keywords:
Time-varying systems
Optimal control
Adaptive control
Tracking applications
Actor–critic

ABSTRACT

Adaptive dynamic programming has been investigated and used as a method to approximately solve optimal regulation problems. However, the extension of this technique to optimal tracking problems for continuous-time nonlinear systems has remained a non-trivial open problem. The control development in this paper guarantees ultimately bounded tracking of a desired trajectory, while also ensuring that the enacted controller approximates the optimal controller.

1. Introduction

Reinforcement learning (RL) is a concept that can be used to enable an agent to learn optimal policies from interaction with the environment. The objective of the agent is to learn the policy that maximizes or minimizes a cumulative long term reward. Almost all RL algorithms use some form of generalized policy iteration (GPI). GPI is a set of two simultaneous interacting processes, policy evaluation and policy improvement. Starting with an estimate of the state value function and an admissible policy, policy evaluation makes the estimate consistent with the policy and policy improvement makes the policy greedy with respect to the value function. These algorithms exploit the fact that the optimal value function satisfies Bellman's principle of optimality (Kirk, 2004; Sutton & Barto, 1998).

When applied to continuous-time systems the principle of optimality leads to the Hamilton–Jacobi–Bellman (HJB) equation which is the continuous-time counterpart of the Bellman equation (Doya, 2000). Similar to discrete-time adaptive dynamic programming (ADP), continuous-time ADP approaches aim at finding approximate solutions to the HJB equation. Various methods to solve this problem are proposed in Abu-Khalaf and Lewis (2002), Beard, Saridis, and Wen (1997), Bhasin et al. (2013), Jiang and Jiang (2012), Vamvoudakis and Lewis (2010), Vrabie and Lewis (2009) and Zhang, Luo, and Liu (2009) and the references therein. An infinite horizon regulation problem with a quadratic cost function is the most common problem considered in ADP literature. For these problems, function approximation techniques can be used to approximate the value function because it is time-invariant.

Approximation techniques like neural networks (NNs) are commonly used in ADP literature for value function approximation. ADP-based approaches are presented in results such as (Dierks & Jagannathan, 2010; Zhang, Cui, Zhang, & Luo, 2011) to address the tracking problem for continuous-time systems, where the value function, and the controller presented are time-varying functions of the tracking error. However, for the infinite horizon optimal control problem, time does not lie on a compact set, and NNs can only approximate functions on a compact domain. Thus, it is unclear how a NN with the tracking error as an input can approximate the time-varying value function and controller.

For discrete-time systems, several approaches have been developed to address the tracking problem. Park, Choi, and Lee

[☆] This research is supported in part by NSF award numbers 1161260, 1217908, ONR grant number N00014-13-1-0151, and a contract with the AFRL Mathematical Modeling and Optimization Institute. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsoring agency. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Antonio Loria under the direction of Editor Andrew R. Teel.

E-mail addresses: rkamalapurkar@ufl.edu (R. Kamalapurkar), huyendtt214@gmail.com (H. Dinh), sbhasin@ee.iitd.ac.in (S. Bhasin), wdixon@ufl.edu (W.E. Dixon).

http://dx.doi.org/10.1016/j.automatica.2014.10.103
0005-1098/© 2014 Elsevier Ltd. All rights reserved.

(1996) use generalized back-propagation through time to solve a finite horizon tracking problem that involves offline training of NNs. An ADP-based approach is presented in Dierks and Jagannathan (2009) to solve an infinite horizon optimal tracking problem where the desired trajectory is assumed to depend on the system states. Greedy heuristic dynamic programming based algorithms are presented in results such as (Luo & Liang, 2011; Wang, Liu, & Wei, 2012; Zhang, Wei, & Luo, 2008) which transform the nonautonomous system into an autonomous system, and approximate convergence of the sequence of value functions to the optimal value function is established. However, these results lack an accompanying stability analysis.

In this result, the tracking error and the desired trajectory both serve as inputs to the NN. This makes the developed controller fundamentally different from previous results, in the sense that a different HJB equation must be solved and its solution, i.e. the feedback component of the controller, is a time-varying function of the tracking error. In particular, this paper addresses the technical obstacles that result from the time-varying nature of the optimal control problem by including the partial derivative of the value function with respect to the desired trajectory in the HJB equation, and by using a system transformation to convert the problem into a time-invariant optimal control problem in such a way that the resulting value function is a time-invariant function of the transformed states, and hence, lends itself to approximation using a NN. A Lyapunov-based analysis is used to prove ultimately bounded tracking and that the enacted controller approximates the optimal controller. Simulation results are presented to demonstrate the applicability of the presented technique. To gauge the performance of the proposed method, a comparison with a numerical optimal solution is presented.

For notational brevity, unless otherwise specified, the domain of all the functions is assumed to be $\mathbb{R}_{\geq 0}$. Furthermore, time-dependence is suppressed while denoting trajectories of dynamical systems. For example, the trajectory $x : \mathbb{R}_{\geq 0} \to \mathbb{R}^n$ is defined by abuse of notation as $x \in \mathbb{R}^n$, and referred to as x instead of $x(t)$, and unless otherwise specified, an equation of the form $f + h(y, t) = g(x)$ is interpreted as $f(t) + h(y(t), t) = g(x(t))$ for all $t \in \mathbb{R}_{\geq 0}$.

2. Formulation of time-invariant optimal control problem

Consider a class of nonlinear control affine systems

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

where $x \in \mathbb{R}^n$ is the state, and $u \in \mathbb{R}^m$ is the control input. The functions $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are locally Lipschitz and $f(0) = 0$. The control objective is to track a bounded continuously differentiable signal $x_d \in \mathbb{R}^n$. To quantify this objective, a tracking error is defined as $e \triangleq x - x_d$. The open-loop tracking error dynamics can then be expressed as

$$\dot{e} = f(x) + g(x)u - \dot{x}_d. \tag{1}$$

The following assumptions are made to facilitate the formulation of an approximate optimal tracking controller.

Assumption 1. The function g is bounded, the matrix $g(x)$ has full column rank for all $x \in \mathbb{R}^n$, and the function $g^+ : \mathbb{R}^n \to \mathbb{R}^{m\times n}$ defined as $g^+ \triangleq \left(g^T g\right)^{-1} g^T$ is bounded and locally Lipschitz.

Assumption 2. The desired trajectory is bounded such that $\|x_d\| \leq d \in \mathbb{R}$, and there exists a locally Lipschitz function $h_d : \mathbb{R}^n \to \mathbb{R}^n$ such that $\dot{x}_d = h_d(x_d)$ and $g(x_d)g^+(x_d)(h_d(x_d) - f(x_d)) = h_d(x_d) - f(x_d)$, $\forall t \in \mathbb{R}_{\geq t_0}$.

The steady-state control policy $u_d : \mathbb{R}^n \to \mathbb{R}^m$ corresponding to the desired trajectory x_d is

$$u_d(x_d) = g_d^+ (h_d(x_d) - f_d), \tag{2}$$

where $g_d^+ \triangleq g^+(x_d)$ and $f_d \triangleq f(x_d)$. To transform the time-varying optimal control problem into a time-invariant optimal control problem, a new concatenated state $\zeta \in \mathbb{R}^{2n}$ is defined as (Zhang et al., 2008)

$$\zeta \triangleq \left[e^T, x_d^T\right]^T. \tag{3}$$

Based on (1) and Assumption 2, the time derivative of (3) can be expressed as

$$\dot{\zeta} = F(\zeta) + G(\zeta)\mu, \tag{4}$$

where the functions $F : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$, $G : \mathbb{R}^{2n} \to \mathbb{R}^{2n\times m}$, and the control $\mu \in \mathbb{R}^m$ are defined as

$$F(\zeta) \triangleq \begin{bmatrix} f(e + x_d) - h_d(x_d) + g(e + x_d)u_d(x_d) \\ h_d(x_d) \end{bmatrix},$$

$$G(\zeta) \triangleq \begin{bmatrix} g(e + x_d) \\ 0 \end{bmatrix}, \qquad \mu \triangleq u - u_d. \tag{5}$$

Local Lipschitz continuity of f and g, the fact that $f(0) = 0$, and Assumption 2 imply that $F(0) = 0$ and F is locally Lipschitz.

The objective of the optimal control problem is to design a policy $\mu^* : \mathbb{R}^{2n} \to \mathbb{R}^m \in \Psi$ such that the control law $\mu = \mu^*(\zeta)$ minimizes the cost functional

$$J(\zeta, \mu) \triangleq \int_0^\infty r(\zeta(\rho), \mu(\rho))\, d\rho,$$

subject to the dynamic constraints in (4), where Ψ is the set of admissible policies (Beard et al., 1997), and $r : \mathbb{R}^{2n} \times \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ is the local cost defined as

$$r(\zeta, \mu) \triangleq \zeta^T \overline{Q} \zeta + \mu^T R \mu. \tag{6}$$

In (6), $R \in \mathbb{R}^{m\times m}$ is a positive definite symmetric matrix of constants, and $\overline{Q} \in \mathbb{R}^{2n\times 2n}$ is defined as

$$\overline{Q} \triangleq \begin{bmatrix} Q & 0_{n\times n} \\ 0_{n\times n} & 0_{n\times n} \end{bmatrix}, \tag{7}$$

where $Q \in \mathbb{R}^{n\times n}$ is a positive definite symmetric matrix of constants with the minimum eigenvalue $\underline{q} \in \mathbb{R}_{>0}$, and $0_{n\times n} \in \mathbb{R}^{n\times n}$ is a matrix of zeros. For brevity of notation, let $(\cdot)'$ denote $\partial(\cdot)/\partial\zeta$.

3. Approximate optimal solution

Assuming that a minimizing policy exists and that the optimal value function $V^* : \mathbb{R}^{2n} \to \mathbb{R}_{\geq 0}$ defined as

$$V^*(\zeta) \triangleq \min_{\mu(\tau)|\tau\in\mathbb{R}_{\geq t}} \int_t^\infty r(\phi^\mu(\tau; t, \zeta), \mu(\tau))\, d\tau \tag{8}$$

is continuously differentiable, the HJB equation for the optimal control problem can be written as

$$H^* = V^{*\prime}(\zeta)\left(F(\zeta) + G(\zeta)\mu^*(\zeta)\right) + r\left(\zeta, \mu^*(\zeta)\right) = 0, \tag{9}$$

for all ζ, with the boundary condition $V^*(0) = 0$, where H^* denotes the Hamiltonian, and $\mu^* : \mathbb{R}^{2n} \to \mathbb{R}^m$ denotes the optimal policy. In (8) $\phi^\mu(\tau; t, \zeta)$ denotes the trajectory of (4) under the controller μ starting at initial time t and initial state ζ. For the local cost in (6) and the dynamics in (4), the optimal policy can be obtained in closed-form as (Kirk, 2004)

$$\mu^*(\zeta) = -\frac{1}{2} R^{-1} G^T(\zeta)\left(V^{*\prime}(\zeta)\right)^T. \tag{10}$$

The value function V^* can be represented using a NN with N neurons as

$$V^*(\zeta) = W^T\sigma(\zeta) + \epsilon(\zeta), \tag{11}$$

where $W \in \mathbb{R}^N$ is the constant ideal weight matrix bounded above by a known positive constant $\bar{W} \in \mathbb{R}$ in the sense that $\|W\| \leq \bar{W}$, $\sigma : \mathbb{R}^{2n} \to \mathbb{R}^N$ is a bounded continuously differentiable nonlinear activation function, and $\epsilon : \mathbb{R}^{2n} \to \mathbb{R}$ is the function reconstruction error (Hornik, Stinchcombe, & White, 1990; Lewis, Selmic, & Campos, 2002).

Using (10) and (11) the optimal policy can be represented as

$$\mu^*(\zeta) = -\frac{1}{2}R^{-1}G^T(\zeta)\left(\sigma'^T(\zeta)W + \epsilon'^T(\zeta)\right). \tag{12}$$

Based on (11) and (12), the NN approximations to the optimal value function and the optimal policy are given by

$$\hat{V}\left(\zeta, \hat{W}_c\right) = \hat{W}_c^T\sigma(\zeta),$$

$$\mu\left(\zeta, \hat{W}_a\right) = -\frac{1}{2}R^{-1}G^T(\zeta)\sigma'^T(\zeta)\hat{W}_a, \tag{13}$$

where $\hat{W}_c \in \mathbb{R}^N$ and $\hat{W}_a \in \mathbb{R}^N$ are estimates of the ideal neural network weights W. The use of two separate sets of weight estimates $\hat{W}_a$ and $\hat{W}_c$ for W is motivated by the fact that the Bellman error (BE) is linear with respect to the value function weight estimates and nonlinear with respect to the policy weight estimates. Use of a separate set of weight estimates for the value function facilitates least squares-based adaptive updates.

The controller is obtained from (2), (5), and (13) as

$$u = -\frac{1}{2}R^{-1}G^T(\zeta)\sigma'^T(\zeta)\hat{W}_a + g_d^+\left(h_d(x_d) - f_d\right). \tag{14}$$

Using the approximations μ and $\hat{V}$ for μ^* and V^* in (9), respectively, the error between the approximate and the optimal Hamiltonian, called the BE $\delta \in \mathbb{R}$, is given in a measurable form by

$$\delta \triangleq \hat{V}'\left(\zeta, \hat{W}_c\right)\dot{\zeta} + r\left(\zeta, \mu\left(\zeta, \hat{W}_a\right)\right). \tag{15}$$

The value function weights are updated to minimize $\int_0^t \delta^2(\rho)\,d\rho$ using a normalized least squares update law[1] with an exponential forgetting factor as (Ioannou & Sun, 1996)

$$\dot{\hat{W}}_c = -\eta_c\Gamma\frac{\omega}{1+\nu\omega^T\Gamma\omega}\delta, \tag{16}$$

$$\dot{\Gamma} = -\eta_c\left(-\lambda\Gamma + \Gamma\frac{\omega\omega^T}{1+\nu\omega^T\Gamma\omega}\Gamma\right), \tag{17}$$

where $\nu, \eta_c \in \mathbb{R}$ are constant positive adaptation gains, $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}^N$ is defined as $\omega \triangleq \sigma'(\zeta)\dot{\zeta}$, and $\lambda \in (0, 1)$ is the constant forgetting factor for the estimation gain matrix $\Gamma \in \mathbb{R}^{N\times N}$. The policy weights are updated to follow the critic weights[2] as

$$\dot{\hat{W}}_a = -\eta_{a1}\left(\hat{W}_a - \hat{W}_c\right) - \eta_{a2}\hat{W}_a, \tag{18}$$

where $\eta_{a1}, \eta_{a2} \in \mathbb{R}$ are constant positive adaptation gains. The following assumption facilitates the stability analysis using PE.

Assumption 3. The regressor $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}^N$ defined as $\psi \triangleq \frac{\omega}{\sqrt{1+\nu\omega^T\Gamma\omega}}$ is persistently exciting (PE). Thus, there exist $T, \underline{\psi} > 0$ such that $\underline{\psi}I \leq \int_t^{t+T}\psi(\tau)\psi(\tau)^T\,d\tau$.[3]

Using Assumption 3 and Corollary 4.3.2 in Ioannou and Sun (1996) it can be concluded that

$$\underline{\varphi}I_{N\times N} \leq \Gamma \leq \overline{\varphi}I_{N\times N}, \quad \forall t \in \mathbb{R}_{\geq 0} \tag{19}$$

where $\overline{\varphi}, \underline{\varphi} \in \mathbb{R}$ are constants such that $0 < \underline{\varphi} < \overline{\varphi}$.[4] Based on (19), the regressor vector can be bounded as

$$\|\psi\| \leq \frac{1}{\sqrt{\nu\underline{\varphi}}}, \quad \forall t \in \mathbb{R}_{\geq 0}. \tag{20}$$

For notational brevity, state-dependence of the functions $h_d,\ F,\ G,\ V^*,\ \mu^*,\ \sigma$, and ϵ is suppressed hereafter.

Using (9), (15), and (16), an unmeasurable form of the BE can be written as

$$\begin{aligned}\delta = {} & -\tilde{W}_c^T\omega + \frac{1}{4}\tilde{W}_a^T\mathcal{G}_\sigma\tilde{W}_a + \frac{1}{4}\epsilon'\mathcal{G}\epsilon'^T \\ & + \frac{1}{2}W^T\sigma'\mathcal{G}\epsilon'^T - \epsilon'F,\end{aligned} \tag{21}$$

where $\mathcal{G} \triangleq GR^{-1}G^T$ and $\mathcal{G}_\sigma \triangleq \sigma'GR^{-1}G^T\sigma'^T$. The weight estimation errors for the value function and the policy are defined as $\tilde{W}_c \triangleq W - \hat{W}_c$ and $\tilde{W}_a \triangleq W - \hat{W}_a$, respectively.

4. Stability analysis

Before stating the main result of the paper, three supplementary technical lemmas are stated. To facilitate the discussion, let $\mathcal{Y} \in \mathbb{R}^{2n+2N}$ be a compact set, and let $\mathcal{Z} \triangleq \mathcal{Y} \cap \mathbb{R}^{n+2N}$. Using the universal approximation property of NNs, on the compact set $\mathcal{Y} \cap \mathbb{R}^{2n}$, the NN approximation errors can be bounded such that $\sup|\epsilon(\zeta)| \leq \bar{\epsilon}$ and $\sup\left|\epsilon'(\zeta)\right| \leq \bar{\epsilon}'$, where $\bar{\epsilon} \in \mathbb{R}$ and $\bar{\epsilon}' \in \mathbb{R}$ are positive constants, and there exists a positive constant $L_F \in \mathbb{R}$ such that[5] $\sup\|F(\zeta)\| \leq L_F\|\zeta\|$. Using Assumptions 1 and 2 the following bounds are developed on the compact set $\mathcal{Y} \cap \mathbb{R}^{2n}$ to aid the subsequent stability analysis:

$$\begin{gathered}\left\|\left(\frac{\epsilon'}{4} + \frac{W^T\sigma'}{2}\right)\mathcal{G}\epsilon'^T\right\| + \bar{\epsilon}'L_F\|x_d\| \leq \iota_1, \qquad \|\mathcal{G}_\sigma\| \leq \iota_2, \\ \left\|\epsilon'\mathcal{G}\epsilon'^T\right\| \leq \iota_3, \qquad \left\|\frac{1}{2}W^T\mathcal{G}_\sigma + \frac{1}{2}\epsilon'\mathcal{G}\sigma'^T\right\| \leq \iota_4, \\ \left\|\frac{1}{4}\epsilon'\mathcal{G}\epsilon'^T + \frac{1}{2}W^T\sigma'\mathcal{G}\epsilon'^T\right\| \leq \iota_5,\end{gathered} \tag{22}$$

where $\iota_1, \iota_2, \iota_3, \iota_4, \iota_5 \in \mathbb{R}$ are positive constants.

4.1. Supporting lemmas

The contribution in the previous section was the development of a transformation that enables the optimal policy and the optimal

[1] The least-squares approach is motivated by faster convergence. With minor modifications to the stability analysis, the result can also be established for a gradient descent update law.

[2] The least-squares approach cannot be used to update the policy weights because the BE is a nonlinear function of the policy weights.

[3] The regressor is defined here as a trajectory indexed by time. It should be noted that different initial conditions result in different regressor trajectories; hence, the constants T and $\underline{\psi}$ depend on the initial values of ζ and $\hat{W}_a$. Hence, the final result is not uniform in the initial conditions.

[4] Since the evolution of ψ is dependent on the initial values of ζ and $\hat{W}_a$, the constants $\overline{\varphi}$ and $\underline{\varphi}$ depend on the initial values of ζ and $\hat{W}_a$.

[5] Instead of using the fact that locally Lipschitz functions on compact sets are Lipschitz, it is possible to bound the function F as $\|F(\zeta)\| \leq \rho(\|\zeta\|)\|\zeta\|$, where $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is non-decreasing. This approach is feasible and results in additional gain conditions.

value function to be expressed as a time-invariant function of ζ. The use of this transformation presents a challenge in the sense that the optimal value function, which is used as the Lyapunov function for the stability analysis, is not a positive definite function of ζ, because the matrix $\overline{Q}$ is positive semi-definite. In this section, this technical obstacle is addressed by exploiting the fact that the time-invariant optimal value function $V^* : \mathbb{R}^{2n} \to \mathbb{R}$ can be interpreted as a time-varying map $V_t^* : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}$, such that

$$V_t^* (e, t) = V^* \left(\begin{bmatrix} e \\ x_d (t) \end{bmatrix} \right) \tag{23}$$

for all $e \in \mathbb{R}^n$ and for all $t \in \mathbb{R}_{\geq 0}$. Specifically, the time-invariant form facilitates the development of the approximate optimal policy, whereas the equivalent time-varying form can be shown to be a positive definite and decrescent function of the tracking error. In the following, Lemma 1 is used to prove that $V_t^* : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ is positive definite and decrescent, and hence, a candidate Lyapunov function.

Lemma 1. *Let B_a denote a closed ball around the origin with the radius $a \in \mathbb{R}_{>0}$. The optimal value function $V_t^* : \mathbb{R}^n \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ satisfies the following properties*

$$V_t^* (e, t) \geq \underline{v} (\|e\|), \tag{24a}$$

$$V_t^* (0, t) = 0, \tag{24b}$$

$$V_t^* (e, t) \leq \overline{v} (\|e\|), \tag{24c}$$

$\forall t \in \mathbb{R}_{\geq 0}$ and $\forall e \in B_a$ where $\underline{v} : [0, a] \to \mathbb{R}_{\geq 0}$ and $\overline{v} : [0, a] \to \mathbb{R}_{\geq 0}$ are class $\mathcal{K}$ functions.

Proof. See Appendix.

Lemma 2. *Let $Z \triangleq \begin{bmatrix} e^T & \tilde{W}_c^T & \tilde{W}_a^T \end{bmatrix}^T$, and suppose that $Z(\tau) \in \mathcal{Z}$, for all $\tau \in [t, t + T]$. Then, the NN weights and the tracking errors satisfy*

$$\begin{aligned} &- \inf_{\tau \in [t, t+T]} \|e(\tau)\|^2 \\ &\leq -\varpi_0 \sup_{\tau \in [t, t+T]} \|e(\tau)\|^2 + \varpi_1 T^2 \sup_{\tau \in [t, t+T]} \left\| \tilde{W}_a (\tau) \right\|^2 + \varpi_2 \\ &\quad - \inf_{\tau \in [t, t+T]} \left\| \tilde{W}_a (\tau) \right\|^2 \leq -\varpi_3 \sup_{\tau \in [t, t+T]} \left\| \tilde{W}_a (\tau) \right\|^2 \\ &\quad + \varpi_4 \inf_{\tau \in [t, t+T]} \left\| \tilde{W}_c (\tau) \right\|^2 + \varpi_5 \sup_{\tau \in [t, t+T]} \|e(\tau)\|^2 + \varpi_6, \end{aligned}$$

where

$$\varpi_0 = \frac{\left(1 - 6nT^2 L_F^2\right)}{2}, \qquad \varpi_1 = \frac{3n}{4} \sup_t \left\| g R^{-1} G^T \sigma'^T \right\|^2,$$

$$\varpi_2 = \frac{3n^2 T^2 \left(d L_F + \sup_t \left\| g g_d^+ (h_d - f_d) - \frac{1}{2} g R^{-1} G^T \sigma'^T W - h_d \right\| \right)^2}{n},$$

$$\varpi_3 = \frac{\left(1 - 6N (\eta_{a1} + \eta_{a2})^2 T^2\right)}{2},$$

$$\varpi_4 = \frac{6N \eta_{a1}^2 T^2}{\left(1 - 6N (\eta_c \overline{\varphi} T)^2 / \left(\nu \underline{\varphi} \right)^2 \right)},$$

$$\varpi_5 = \frac{18 \left(\eta_{a1} N \eta_c \overline{\varphi} \bar{\epsilon}' L_F T^2 \right)^2}{\nu \underline{\varphi} \left(1 - 6N (\eta_c \overline{\varphi} T)^2 / \left(\nu \underline{\varphi} \right)^2 \right)},$$

$$\varpi_6 = \frac{18 \left(N \eta_{a1} \eta_c \overline{\varphi} \left(\bar{\epsilon}' L_F d + \iota_5 \right) T^2 \right)^2}{\nu \underline{\varphi} \left(1 - 6N (\eta_c \overline{\varphi} T)^2 / \left(\nu \underline{\varphi} \right)^2 \right)} + 3N \left(\eta_{a2} \overline{W} T \right)^2.$$

Proof. The proof is omitted due to space constraints, and is available in Kamalapurkar, Dinh, Bhasin, and Dixon (2013).

Lemma 3. *Let $Z \triangleq \begin{bmatrix} e^T & \tilde{W}_c^T & \tilde{W}_a^T \end{bmatrix}^T$, and suppose that $Z(\tau) \in \mathcal{Z}$, for all $\tau \in [t, t + T]$. Then, the critic weights satisfy*

$$\begin{aligned} - \int_t^{t+T} \left\| \tilde{W}_c^T \psi \right\|^2 d\tau \leq -\underline{\psi} \varpi_7 \left\| \tilde{W}_c \right\|^2 + \varpi_8 \int_t^{t+T} \|e\|^2 d\tau \\ + 3\iota_2^2 \int_t^{t+T} \left\| \tilde{W}_a (\sigma) \right\|^4 d\sigma + \varpi_9 T, \end{aligned}$$

where $\varpi_7 = \frac{\nu^2 \underline{\varphi}^2}{2\left(\nu^2 \underline{\varphi}^2 + \eta_c^2 \overline{\varphi}^2 T^2\right)}$, $\varpi_8 = 3\bar{\epsilon}'^2 L_F^2$, and $\varpi_9 = 2(\iota_5^2 + \bar{\epsilon}'^2 L_F^2 d^2)$.

Proof. The proof is omitted due to space constraints, and is available in Kamalapurkar et al. (2013).

4.2. Gain conditions and gain selection

The following section details sufficient gain conditions derived based on a stability analysis performed using the candidate Lyapunov function $V_L : \mathbb{R}^{n+2N} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ defined as $V_L (Z, t) \triangleq V_t^* (e, t) + \frac{1}{2} \tilde{W}_c^T \Gamma^{-1} \tilde{W}_c + \frac{1}{2} \tilde{W}_a^T \tilde{W}_a$. Using Lemma 1 and (19),

$$\underline{v_l} (\|Z\|) \leq V_L (Z, t) \leq \overline{v_l} (\|Z\|), \tag{25}$$

$\forall Z \in B_b$, $\forall t \in \mathbb{R}_{\geq 0}$, where $\underline{v_l} : [0, b] \to \mathbb{R}_{\geq 0}$ and $\overline{v_l} : [0, b] \to \mathbb{R}_{\geq 0}$ are class $\mathcal{K}$ functions, and $B_b \subset \mathbb{R}^{n+2N}$ denotes a ball of radius $b \in \mathbb{R}_{>0}$ around the origin, containing $\mathcal{Z}$.

To facilitate the discussion, define $\eta_{a12} \triangleq \eta_{a1} + \eta_{a2}$, $Z \triangleq [e^T \; \tilde{W}_c^T \; \tilde{W}_a^T]^T$, $\iota \triangleq \frac{\left(\eta_{a2} \overline{W} + \iota_4\right)^2}{\eta_{a12}} + 2\eta_c (\iota_1)^2 + \frac{1}{4} \iota_3$, $\varpi_{10} \triangleq \frac{\varpi_6 \eta_{a12} + 2\varpi_2 \underline{q} + \eta_c \varpi_9}{8} + \iota$, and $\varpi_{11} \triangleq \frac{1}{16} \min(\eta_c \underline{\psi} \varpi_7, 2\varpi_0 \underline{q} T, \varpi_3 \eta_{a12} T)$. Let $Z_0 \in \mathbb{R}_{\geq 0}$ denote a known constant bound on the initial condition such that $\|Z(t_0)\| \leq Z_0$, and let

$$\overline{Z} \triangleq \underline{v_l}^{-1} \left(\overline{v_l} \left(\max \left(Z_0, \sqrt{\frac{\varpi_{10} T}{\varpi_{11}}} \right) \right) + \iota T \right). \tag{26}$$

The sufficient gain conditions for the subsequent Theorem 1 are given by[6]

$$\begin{aligned} &\eta_{a12} > \max \left(\eta_{a1} \xi_2 + \frac{\eta_c \iota_2}{4} \sqrt{\frac{\overline{Z}}{\nu \underline{\varphi}}}, 3\eta_c \iota_2^2 \overline{Z} \right), \\ &\xi_1 > 2\bar{\epsilon}' L_F, \qquad \eta_c > \frac{\eta_{a1}}{\lambda \underline{\gamma} \xi_2}, \qquad \underline{\psi} > \frac{2\varpi_4 \eta_{a12}}{\eta_c \varpi_7} T, \\ &\underline{q} > \max \left(\frac{\varpi_5 \eta_{a12}}{\varpi_0}, \frac{1}{2} \eta_c \varpi_8, \eta_c L_F \bar{\epsilon}' \xi_1 \right), \\ &T < \min \left(\frac{1}{\sqrt{6N} \eta_{a12}}, \frac{\nu \underline{\varphi}}{\sqrt{6N} \eta_c \overline{\varphi}}, \frac{1}{2\sqrt{n} L_F}, \right. \\ &\qquad \left. \sqrt{\frac{\eta_{a12}}{6N \eta_{a12}^3 + 8\underline{q} \varpi_1}} \right), \end{aligned} \tag{27}$$

where $\xi_1, \; \xi_2 \in \mathbb{R}$ are known adjustable positive constants. Furthermore, the compact set $\mathcal{Z}$ satisfies the sufficient condition

$$\overline{Z} \leq r, \tag{28}$$

[6] Similar conditions on $\underline{\psi}$ and T can be found in PE-based adaptive control in the presence of bounded or Lipschitz uncertainties (cf. Misovec, 1999 and Narendra & Annaswamy, 1986).

where $r \triangleq \frac{1}{2}\sup_{z,y\in\mathcal{Z}}\|z-y\|$ denotes the radius of $\mathcal{Z}$. Since the Lipschitz constant and the bounds on NN approximation error depend on the size of the compact set $\mathcal{Z}$, the constant $\overline{Z}$ depends on r; hence, feasibility of the sufficient condition in (28) is not apparent. Algorithm 1 in the Appendix details an iterative gain selection process in order to ensure satisfaction of the sufficient condition in (28).

4.3. Main result

Theorem 1. *Provided that the sufficient conditions in (27) and (28) are satisfied and Assumptions 1–3 hold, the controller in (14) and the update laws in (16)–(18) guarantee that the tracking error is ultimately bounded, and the error $\|\mu(t)-\mu^*(\zeta(t))\|$ is ultimately bounded as $t\to\infty$.*

Proof. The time derivative of V_L is

$$\dot{V}_L = V^{*\prime}F + V^{*\prime}G\mu + \tilde{W}_c^T\Gamma^{-1}\dot{\tilde{W}}_c - \tilde{W}_a^T\dot{\hat{W}}_a - \frac{1}{2}\tilde{W}_c^T\Gamma^{-1}\dot{\Gamma}\Gamma^{-1}\tilde{W}_c.$$

Provided the sufficient conditions in (27) are satisfied, (16), (21), the bounds in (20)–(22), and the facts that $V^{*\prime}F = -V^{*\prime}G\mu^* - r(\zeta,\mu^*)$ and $V^{*\prime}G = -2\mu^{*T}R$ yield

$$\dot{V}_L \leq -\frac{\underline{q}}{2}\|e\|^2 - \frac{1}{8}\eta_c\left\|\tilde{W}_c^T\psi\right\|^2 - \frac{\eta_{a12}}{4}\left\|\tilde{W}_a\right\|^2 + \iota. \tag{29}$$

The inequality in (29) is valid provided $Z(t)\in\mathcal{Z}$.

Integrating (29), using the facts that $-\int_t^{t+T}\|e(\tau)\|^2 d\tau \leq -T\inf_{\tau\in[t,t+T]}\|e(\tau)\|^2$ and $-\int_t^{t+T}\|\tilde{W}_a(\tau)\|^2 d\tau \leq -T\inf_{\tau\in[t,t+T]}\|\tilde{W}_a(\tau)\|^2$, Lemmas 2 and 3, and the gain conditions in (27) yields

$$\begin{aligned} V_L(Z(t+T),t+T) - V_L(Z(t),t) \\ \leq -\frac{\eta_c\underline{\psi}\varpi_7}{16}\left\|\tilde{W}_c(t)\right\|^2 - \frac{\varpi_0\underline{q}T}{8}\|e(t)\|^2 \\ - \frac{\varpi_3\eta_{a12}T}{16}\left\|\tilde{W}_a(t)\right\|^2 + \varpi_{10}T, \end{aligned}$$

provided $Z(\tau)\in\mathcal{Z},\ \forall\tau\in[t,t+T]$. Thus, $V_L(Z(t+T),t+T) - V_L(Z(t),t) < 0$ provided $\|Z(t)\| > \sqrt{\frac{\varpi_{10}T}{\varpi_{11}}}$ and $Z(\tau)\in\mathcal{Z},\ \forall\tau\in[t,t+T]$. The bounds on the Lyapunov function in (25) yield $V_L(Z(t+T),t+T) - V_L(Z(t),t) < 0$ provided $V_L(Z(t),t) > \overline{v_l}\left(\sqrt{\frac{\varpi_{10}T}{\varpi_{11}}}\right)$ and $Z(\tau)\in\mathcal{Z},\ \forall\tau\in[t,t+T]$.

Since $Z(t_0)\in\mathcal{Z}$, (29) can be used to conclude that $\dot{V}_L(Z(t_0),t_0) \leq \iota$. The sufficient condition in (28) ensures that $\underline{v_l}^{-1}(V_L(Z(t_0),t_0)+\iota T) \leq r$; hence, $Z(t)\in\mathcal{Z}$ for all $t\in[t_0,t_0+\overline{T}]$. If $V_L(Z(t_0),t_0) > \overline{v_l}\left(\sqrt{\frac{\varpi_{10}T}{\varpi_{11}}}\right)$, then $Z(t)\in\mathcal{Z}$ for all $t\in[t_0,t_0+T]$ implies $V_L(Z(t_0+T),t_0+T) - V_L(Z(t_0),t_0) < 0$; hence, $\underline{v_l}^{-1}(V_L(Z(t_0+T),t_0+T)+\iota T) \leq r$. Thus, $Z(t)\in\mathcal{Z}$ for all $\overline{t}\in[t_0+T,t_0+2T]$. Inductively, the system state is bounded such that $\sup_{t\in[0,\infty)}\|Z(t)\| \leq r$ and ultimately bounded[7] such that

$$\limsup_{t\to\infty}\|Z(t)\| \leq \underline{v_l}^{-1}\left(\overline{v_l}\left(\sqrt{\frac{\varpi_{10}T}{\varpi_{11}}}\right)+\iota T\right).$$

5. Simulation

Simulations are performed on a two-link manipulator to demonstrate the ability of the presented technique to approximately optimally track a desired trajectory. The two link robot manipulator is modeled using Euler–Lagrange dynamics as

$$M\ddot{q} + V_m\dot{q} + F_d\dot{q} + F_s = u, \tag{30}$$

where $q = \begin{bmatrix}q_1 & q_2\end{bmatrix}^T$ and $\dot{q} = \begin{bmatrix}\dot{q}_1 & \dot{q}_2\end{bmatrix}^T$ are the angular positions in radians and the angular velocities in radian/s respectively. In (30), $M\in\mathbb{R}^{2\times 2}$ denotes the inertia matrix, and $V_m\in\mathbb{R}^{2\times 2}$ denotes the centripetal–Coriolis matrix given by $M \triangleq \begin{bmatrix}p_1+2p_3c_2 & p_2+p_3c_2\\ p_2+p_3c_2 & p_2\end{bmatrix}$, $V_m \triangleq \begin{bmatrix}-p_3s_2\dot{q}_2 & -p_3s_2(\dot{q}_1+\dot{q}_2)\\ p_3s_2\dot{q}_1 & 0\end{bmatrix}$, where $c_2 = \cos(q_2)$, $s_2 = \sin(q_2)$, $p_1 = 3.473$ kg m^2, $p_2 = 0.196$ kg m^2, and $p_3 = 0.242$ kg m^2, and $F_d = \operatorname{diag}\begin{bmatrix}5.3, & 1.1\end{bmatrix}$ N m s and $F_s(\dot{q}) = [8.45\tanh(\dot{q}_1), 2.35\tanh(\dot{q}_2)]^T$ N m are the models for the static and the dynamic friction, respectively.

The objective is to find a policy μ that ensures that the state $x \triangleq [q_1,\ q_2,\ \dot{q}_1,\ \dot{q}_2]^T$ tracks the desired trajectory $x_d(t) = [0.5\cos(2t),\ 0.33\cos(3t),\ -\sin(2t),\ -\sin(3t)]^T$, while minimizing the cost $\int_0^\infty\left(e^TQe+\mu^T\mu\right)dt$, where $Q = \operatorname{diag}[10, 10, 2, 2]$. Using (2)–(5) and the definitions

$$\begin{aligned} f &\triangleq \left[x_3,\quad x_4,\quad \left(M^{-1}(-V_m-F_d)\begin{bmatrix}x_3\\x_4\end{bmatrix}-F_s\right)^T\right]^T, \\ g &\triangleq \left[\begin{bmatrix}0, & 0\end{bmatrix}^T,\quad \begin{bmatrix}0, & 0\end{bmatrix}^T,\quad \left(M^{-1}\right)^T\right]^T, \\ g_d^+ &\triangleq \left[\begin{bmatrix}0, & 0\end{bmatrix}^T,\quad \begin{bmatrix}0, & 0\end{bmatrix}^T,\quad M(x_d)\right], \\ h_d &\triangleq \begin{bmatrix}x_{d3}, & x_{d4}, & -4x_{d1}, & -9x_{d2}\end{bmatrix}^T, \end{aligned} \tag{31}$$

the optimal tracking problem can be transformed into the time-invariant form in (5).

In this effort, the basis chosen for the value function approximation is a polynomial basis with 23 elements given by

$$\begin{aligned} \sigma(\zeta) = \frac{1}{2}[\zeta_1^2 \quad \zeta_2^2 \quad \zeta_1\zeta_3 \quad \zeta_1\zeta_4 \quad \zeta_2\zeta_3 \quad \zeta_2\zeta_4 \quad \zeta_1^2\zeta_2^2 \quad \zeta_1^2\zeta_5^2 \\ \zeta_1^2\zeta_6^2 \quad \zeta_1^2\zeta_7^2 \quad \zeta_1^2\zeta_8^2 \quad \zeta_2^2\zeta_5^2 \quad \zeta_2^2\zeta_6^2 \quad \zeta_2^2\zeta_7^2 \quad \zeta_2^2\zeta_8^2 \quad \zeta_3^2\zeta_5^2 \\ \zeta_3^2\zeta_6^2 \quad \zeta_3^2\zeta_7^2 \quad \zeta_3^2\zeta_8^2 \quad \zeta_4^2\zeta_5^2 \quad \zeta_4^2\zeta_6^2 \quad \zeta_4^2\zeta_7^2 \quad \zeta_4^2\zeta_8^2]^T. \end{aligned} \tag{32}$$

The control gains are selected as $\eta_{a1} = 5,\ \eta_{a2} = 0.001,\ \eta_c = 1.25$, $\lambda = 0.001$, and $\nu = 0.005$, and the initial conditions are $x(0) = \begin{bmatrix}1.8 & 1.6 & 0 & 0\end{bmatrix}^T$, $\hat{W}_c(0) = 10\times\mathbf{1}_{23\times 1}$, $\hat{W}_a(0) = 6\times\mathbf{1}_{23\times 1}$, and $\Gamma(0) = 2000\times I_{23\times 23}$, where $\mathbf{1}_{23\times 1}$ is vector of ones. To ensure PE, a probing signal

$$p(t) = \begin{bmatrix} 2.55\tanh(2t)\Big(20\sin\left(\sqrt{232}\pi t\right)\cos\left(\sqrt{20}\pi t\right) \\ +6\sin\left(18e^2t\right)+20\cos(40t)\cos(21t)\Big) \\ 0.01\tanh(2t)\Big(20\sin\left(\sqrt{132}\pi t\right)\cos\left(\sqrt{10}\pi t\right) \\ +6\sin(8et)+20\cos(10t)\cos(11t)\Big) \end{bmatrix} \tag{33}$$

is added to the control signal for the first 30 s of the simulation (Vamvoudakis & Lewis, 2010).

It is clear from Fig. 1 that the system states are bounded during the learning phase and the algorithm converges to a stabilizing controller in the sense that the tracking errors go to zero when the probing signal is eliminated. Furthermore, Fig. 2 shows that the weight estimates for the value function and the policy are bounded and they converge.

[7] If the regressor ψ satisfies a stronger u-PE assumption (cf. Loría & Panteley, 2002 and Panteley, Loria, & Teel, 2001), the tracking error and the weight estimation errors can be shown to be uniformly ultimately bounded.

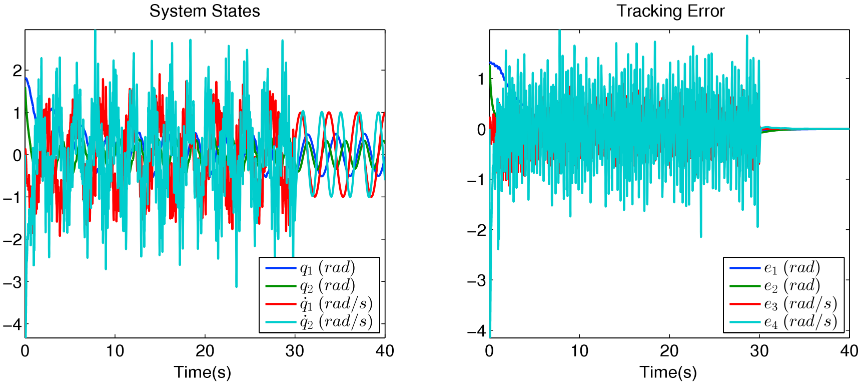

Fig. 1. State and error trajectories with probing signal.

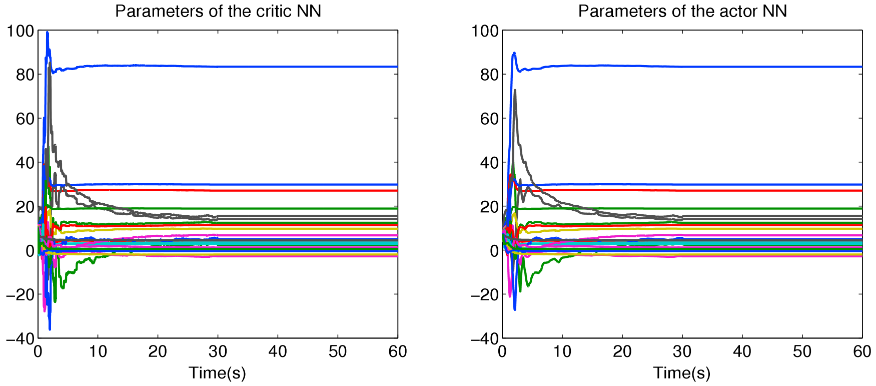

Fig. 2. Evolution of value function and policy weights.

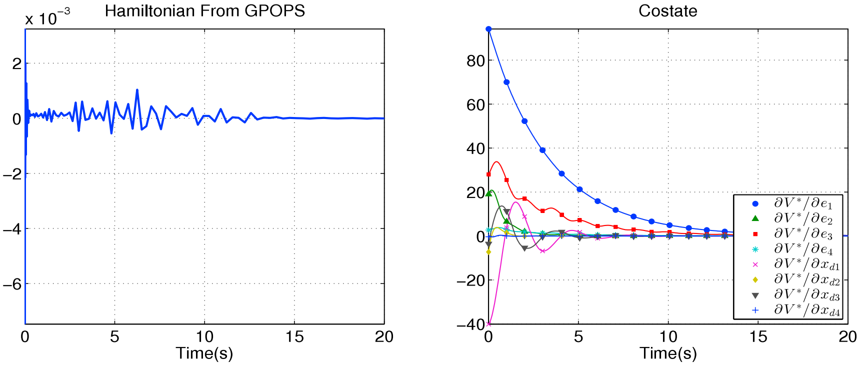

Fig. 3. Hamiltonian and costate of the numerical solution computed using GPOPS.

The NN weights converge to the following values

$$\hat{W}_c = \hat{W}_a = \big[83.36\ \ 2.37\ \ 27.0\ \ 2.78\ \ -2.83\ \ 0.20\ \ 14.13\ \ 29.81\ \ 18.87\ \ 4.11\ \ 3.47\ \ 6.69\ \ 9.71\ \ 15.58\ \ 4.97\ \ 12.42\ \ 11.31\ \ 3.29\ \ 1.19\ \ -1.99\ \ 4.55\ \ -0.47\ \ 0.56\big]^T. \tag{34}$$

Note that the last sixteen weights that correspond to the terms containing the desired trajectories $\zeta_5, \ldots, \zeta_8$ are non-zero. Thus, the resulting value function V and the resulting policy μ depend on the desired trajectory, and hence, are time-varying functions of the tracking error. Since the true weights are unknown, a direct comparison of the weights in (34) with the true weights is not possible. Instead, to gauge the performance of the presented technique, the state and the control trajectories obtained using the estimated policy are compared with those obtained using Radau-pseudospectral numerical optimal control computed using the GPOPS software (Rao et al., 2010). Since an accurate numerical solution is difficult to obtain for an infinite horizon optimal control problem, the numerical optimal control problem is solved over a finite horizon ranging over approximately 5 times the settling time associated with the slowest state variable. Based on the solution obtained using the proposed technique, the slowest settling time is estimated to be approximately 20 s. Thus, to approximate the infinite horizon solution, the numerical solution is computed over a 100 s time horizon using 300 collocation points.

As seen in Fig. 3, the Hamiltonian of the numerical solution is approximately zero. This supports the assertion that the optimal control problem is time-invariant. Furthermore, since the Hamiltonian is close to zero, the numerical solution obtained using GPOPS is sufficiently accurate as a benchmark to compare against the ADP-based solution obtained using the proposed technique. Note that in Fig. 3, the costate variables corresponding to the desired trajectories are nonzero. Since these costate variables represent the sensitivity of the cost with respect to the desired trajectories, this

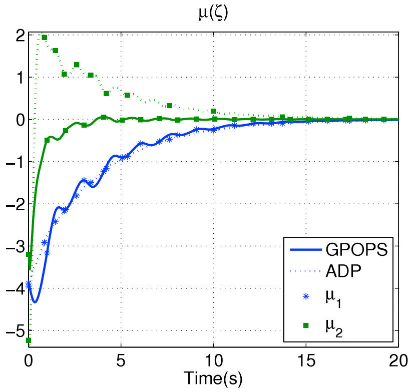

Fig. 4. Control trajectories $\mu(t)$ obtained from GPOPS and the developed technique.

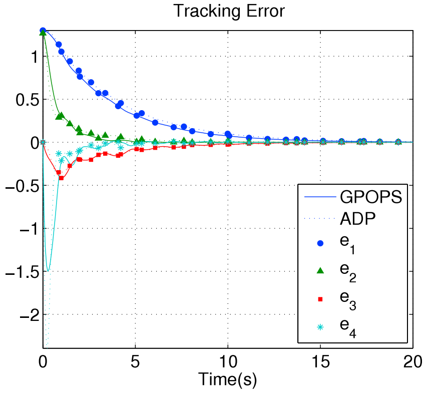

Fig. 5. Tracking error trajectories $e(t)$ obtained from GPOPS and the developed technique.

further supports the assertion that the optimal value function depends on the desired trajectory, and hence, is a time-varying function of the tracking error.

Figs. 4 and 5 show the control and the tracking error trajectories obtained from the developed technique (dashed lines) plotted alongside the numerical solution obtained using GPOPS (solid lines). The trajectories obtained using the developed technique are close to the numerical solution. The inaccuracies are a result of the facts that the set of basis functions in (32) is not exact, and the proposed method attempts to find the weights that generate the least total cost for the given set of basis functions. The accuracy of the approximation can be improved by choosing a more appropriate set of basis functions, or at an increased computational cost, by adding more basis functions to the existing set in (32). The total cost $\int_0^{100}\left(e(t)^T Qe(t) + \mu(t)^T R\mu(t)\right)dt$ obtained using the numerical solution is found to be 75.42 and the total cost $\int_0^{\infty}(e(t)^T Qe(t) + \mu(t)^T R\mu(t))dt$ obtained using the developed method is found to be 84.31. Note that from Figs. 4 and 5, it is clear that both the tracking error and the control converge to zero after approximately 20 s, and hence, the total cost obtained from the numerical solution is a good approximation of the infinite horizon cost.

6. Conclusion

An ADP-based approach using the policy evaluation and policy improvement architecture is presented to approximately solve the infinite horizon optimal tracking problem for control affine nonlinear systems with quadratic cost. The problem is solved by transforming the system to convert the tracking problem that has a time-varying value function, into a time-invariant optimal control problem. The ultimately bounded tracking and estimation result was established using Lyapunov analysis for nonautonomous systems. Simulations are performed to demonstrate the applicability and the effectiveness of the developed method. The accuracy of the approximation depends on the choice of basis functions and the result hinges on the system states being PE. Furthermore, computation of the desired control in (2) requires exact model knowledge.

A solution to the tracking problem without using the desired control while employing a multi-layer neural network that can approximate the basis functions remains a future challenge. In adaptive control, it is generally possible to formulate the control problem such that PE along the desired trajectory is sufficient to achieve parameter convergence. In the ADP-based tracking problem, PE along the desired trajectory would be sufficient to achieve parameter convergence if the BE can be formulated in terms of the desired trajectories. Achieving such a formulation is not trivial, and is a subject for future research.

Appendix

The proofs for the technical lemmas and the gain selection algorithm are detailed in this section.

Algorithm for selection of NN architecture and learning gains

Since the gains depend on the initial conditions and on the compact sets used for function approximation and the Lipschitz bounds, an iterative algorithm is developed to select the gains. In Algorithm 1, the notation $\{\varpi\}_i$ for any parameter ϖ denotes the value of ϖ computed in the ith iteration. Algorithm 1 ensures satisfaction of the sufficient condition in (28).

Algorithm 1 Gain Selection

First iteration:

Given $Z_0 \in \mathbb{R}_{\geq 0}$ such that $\|Z(t_0)\| < Z_0$, let $\mathcal{Z}_1 = \{\varrho \in \mathbb{R}^{n+2\{N\}_1} \mid \|\varrho\| \leq \beta_1 \underline{v_l}^{-1}(\overline{v_l}(Z_0))\}$ for some $\beta_1 > 1$. Using $\mathcal{Z}_1$, compute the bounds in (22) and (26), and select the gains according to (27). If $\left\{\overline{Z}\right\}_1 \leq \beta_1 \underline{v_l}^{-1}\left(\overline{v_l}\left(\|Z_0\|\right)\right)$, set $\mathcal{Z} = \mathcal{Z}_1$ and terminate.

Second iteration:

If $\{\overline{Z}\}_1 > \beta_1 \underline{v_l}^{-1}(\overline{v_l}(\|Z_0\|))$, let $\mathcal{Z}_2 \triangleq \{\varrho \in \mathbb{R}^{n+2\{N\}_1} \mid \|\varrho\| \leq \beta_2 \{\overline{Z}\}_1\}$. Using $\mathcal{Z}_2$, compute the bounds in (22) and (26) and select the gains according to (27). If $\left\{\overline{Z}\right\}_2 \leq \left\{\overline{Z}\right\}_1$, set $\mathcal{Z} = \mathcal{Z}_2$ and terminate.

Third iteration:

If $\left\{\overline{Z}\right\}_2 > \left\{\overline{Z}\right\}_1$, increase the number of NN neurons to $\{N\}_3$ to yield a lower function approximation error $\left\{\overline{\epsilon}'\right\}_3$ such that $\{L_F\}_2\left\{\overline{\epsilon}'\right\}_3 \leq \{L_F\}_1\left\{\overline{\epsilon}'\right\}_1$. The increase in the number of NN neurons ensures that $\{\iota\}_3 \leq \{\iota\}_1$. Furthermore, the assumption that the PE interval $\{T\}_3$ is small enough such that $\{L_F\}_2\{T\}_3 \leq \{T\}_1\{L_F\}_1$ and $\{N\}_3\{T\}_3 \leq \{T\}_1\{N\}_1$ ensures that $\left\{\frac{\varpi_{10}}{\varpi_{11}}\right\}_3 \leq \left\{\frac{\varpi_{10}}{\varpi_{11}}\right\}_1$, and hence, $\left\{\overline{Z}\right\}_3 \leq \beta_2\left\{\overline{Z}\right\}_1$. Set $\mathcal{Z} = \left\{\varrho \in \mathbb{R}^{n+2\{N\}_3} \mid \|\varrho\| \leq \beta_2\left\{\overline{Z}\right\}_1\right\}$ and terminate.

Proof of Lemma 1

The following supporting technical lemma is used to prove Lemma 1.

Lemma 4. *Let $D \subseteq \mathbb{R}^n$ contain the origin and let $\Xi : D \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be positive definite. If $t \longmapsto \Xi(x, t)$ is uniformly bounded for all $x \in D$ and if $x \longmapsto \Xi(x, t)$ is continuous, uniformly in t, then Ξ is decrescent in D.*

Proof. Since $t \longmapsto \Xi(x, t)$ is uniformly bounded, for all $x \in D$, $\sup_{t \in \mathbb{R}_{\geq 0}} \{\Xi(x, t)\}$ exists and is unique for all $x \in D$. Let the function $\alpha : D \to \mathbb{R}_{\geq 0}$ be defined as

$$\alpha(x) \triangleq \sup_{t \in \mathbb{R}_{\geq 0}} \{\Xi(x, t)\}. \tag{35}$$

Since $x \to \Xi(x, t)$ is continuous, uniformly in t, $\forall \varepsilon > 0, \exists \varsigma(x) > 0$ such that $\forall y \in D$,

$$\begin{aligned} d_{D \times \mathbb{R}_{\geq 0}}((x, t), (y, t)) &< \varsigma(x) \\ &\Longrightarrow d_{\mathbb{R}_{\geq 0}}(\Xi(x, t), \Xi(y, t)) < \varepsilon, \end{aligned} \tag{36}$$

where $d_M(\cdot, \cdot)$ denotes the standard Euclidean metric on the metric space M. By the definition of $d_M(\cdot, \cdot)$, $d_{D \times \mathbb{R}_{\geq 0}}((x, t), (y, t)) = d_D(x, y)$. Using (36),

$$d_D(x, y) < \varsigma(x) \Longrightarrow |\Xi(x, t) - \Xi(y, t)| < \varepsilon. \tag{37}$$

Given the fact that Ξ is positive, (37) implies $\Xi(x, t) < \Xi(y, t) + \varepsilon$ and $\Xi(y, t) < \Xi(x, t) + \varepsilon$ which from (35) implies $\alpha(x) < \alpha(y) + \varepsilon$ and $\alpha(y) < \alpha(x) + \varepsilon$, and hence, from (37), $d_D(x, y) < \varsigma(x) \Longrightarrow |\alpha(x) - \alpha(y)| < \varepsilon$. Since Ξ is positive definite, (35) can be used to conclude $\alpha(0) = 0$. Thus, Ξ is bounded above by a continuous positive definite function; hence, Ξ is decrescent in D. □

Based on the definitions in (8)–(7) and (23), $V_t^*(e, t) > 0$, $\forall t \in \mathbb{R}_{\geq 0}$ and $\forall e \in B_a \setminus \{0\}$. The optimal value function $V^*\left(\left[0, x_d^T\right]^T\right)$ is the cost incurred when starting with $e = 0$ and following the optimal policy thereafter for an arbitrary desired trajectory x_d. Substituting $x(t_0) = x_d(t_0)$, $\mu(t_0) = 0$ and (2) in (4) indicates that $\dot{e}(t_0) = 0$. Thus, when starting from $e = 0$, a policy that is identically zero satisfies the dynamic constraints in (4). Furthermore, the optimal cost is $V^*\left(\left[0, x_d^T(t_0)\right]^T\right) = 0, \forall x_d(t_0)$ which, from (23), implies (24b). Since the optimal value function V_t^* is strictly positive everywhere but at $e = 0$ and is zero at $e = 0$, V_t^* is a positive definite function. Hence, Lemma 4.3 in Khalil (2002) can be invoked to conclude that there exists a class $\mathcal{K}$ function $\underline{v} : [0, a] \to \mathbb{R}_{\geq 0}$ such that $\underline{v}(\|e\|) \leq V_t^*(e, t)$, $\forall t \in \mathbb{R}_{\geq 0}$ and $\forall e \in B_a$.

Admissibility of the optimal policy implies that $V^*(\zeta)$ is bounded over all compact subsets $K \subset \mathbb{R}^{2n}$. Since the desired trajectory is bounded, $t \longmapsto V_t^*(e, t)$ is uniformly bounded for all $e \in B_a$. To establish that $e \longmapsto V_t^*(e, t)$ is continuous, uniformly in t, let $\chi_{e_o} \subset \mathbb{R}^n$ be a compact set containing e_o. Since x_d is bounded, $x_d \in \chi_{x_d}$, where $\chi_{x_d} \subset \mathbb{R}^n$ is compact. Since $V^* : \mathbb{R}^{2n} \to \mathbb{R}_{\geq 0}$ is continuous, and $\chi_{e_o} \times \chi_{x_d} \subset \mathbb{R}^{2n}$ is compact, V^* is uniformly continuous on $\chi_{e_o} \times \chi_{x_d}$. Thus, $\forall \varepsilon > 0$, $\exists \varsigma > 0$, such that $\forall([e_o^T, x_d^T]^T, [e_1^T, x_d^T]^T) \in \chi_{e_o} \times \chi_{x_d}$, $d_{\chi_{e_o} \times \chi_{x_d}}([e_o^T, x_d^T]^T, [e_1^T, x_d^T]^T) < \varsigma \Longrightarrow d_{\mathbb{R}}(V^*([e_o^T, x_d^T]^T), V^*([e_1^T, x_d^T]^T)) < \varepsilon$. Thus, for each $e_o \in \mathbb{R}^n$, there exists a $\varsigma > 0$ independent of x_d, that establishes the continuity of $e \longmapsto V^*([e^T, x_d^T]^T)$ at e_o. Thus, $e \longmapsto V^*([e^T, x_d^T]^T)$ is continuous, uniformly in x_d, and hence, using (23) $e \longmapsto V_t^*(e, t)$ is continuous, uniformly in t. Using Lemma 4 and (24a) and (24b), there exists a positive definite function $\alpha : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ such that $V_t^*(e, t) < \alpha(e)$, $\forall (e, t) \in \mathbb{R}^n \times \mathbb{R}_{\geq 0}$. Lemma 4.3 in Khalil (2002) indicates that there exists a class $\mathcal{K}$ function $\overline{v} : [0, a] \to \mathbb{R}_{\geq 0}$ such that $\alpha(e) \leq \overline{v}(\|e\|)$, which implies (24c).

References

Abu-Khalaf, M., & Lewis, F. (2002). Nearly optimal HJB solution for constrained input systems using a neural network least-squares approach. In *Proc. IEEE conf. decis. control* (pp. 943–948). Las Vegas, NV.

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

Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K., Lewis, F. L., & Dixon, W. (2013). A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems. *Automatica*, *49*(1), 89–92.

Dierks, T., & Jagannathan, S. (2009). Optimal tracking control of affine nonlinear discrete-time systems with unknown internal dynamics. In *Proc. IEEE conf. decis. control* (pp. 6750–6755).

Dierks, T., & Jagannathan, S. (2010). Optimal control of affine nonlinear continuous-time systems. In *Proc. Am. control conf.* (pp. 1568–1573).

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

Hornik, K., Stinchcombe, M., & White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, *3*(5), 551–560.

Ioannou, P., & Sun, J. (1996). *Robust adaptive control*. Prentice Hall.

Jiang, Y., & Jiang, Z.-P. (2012). Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, *48*(10), 2699–2704.

Kamalapurkar, R., Dinh, H., Bhasin, S., & Dixon, W. (2013). Approximately optimal trajectory tracking for continuous time nonlinear systems. arXiv:1301.7664.

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

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

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.

Loría, A., & Panteley, E. (2002). Uniform exponential stability of linear time-varying systems: revisited. *Systems & Control Letters*, *47*(1), 13–24.

Luo, Y., & Liang, M. (2011). Approximate optimal tracking control for a class of discrete-time non-affine systems based on GDHP algorithm. In *IWACI int. workshop adv. comput. intell.* (pp. 143–149).

Misovec, K. M. (1999). Friction compensation using adaptive non-linear control with persistent excitation. *International Journal of Control*, *72*(5), 457–479.

Narendra, K., & Annaswamy, A. (1986). Robust adaptive control in the presence of bounded disturbances. *IEEE Transactions on Automatic Control*, *31*(4), 306–315.

Panteley, E., Loria, A., & Teel, A. (2001). Relaxed persistency of excitation for uniform asymptotic stability. *IEEE Transactions on Automatic Control*, *46*(12), 1874–1886.

Park, Y. M., Choi, M. S., & Lee, K. Y. (1996). An optimal tracking neuro-controller for nonlinear dynamic systems. *IEEE Transactions on Neural Networks*, *7*(5), 1099–1110.

Rao, A. V., Benson, D. A., Darby, C. L., Patterson, M. A., Francolin, C., & Huntington, G. T. (2010). Algorithm 902: GPOPS, a MATLAB software for solving multiple-phase optimal control problems using the Gauss pseudospectral method. *ACM Transactions on Mathematical Software*, *37*(2), 1–39.

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

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

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

Wang, D., Liu, D., & Wei, Q. (2012). Finite-horizon neuro-optimal tracking control for a class of discrete-time nonlinear systems using adaptive dynamic programming approach. *Neurocomputing*, *78*(1), 14–22.

Zhang, H., Cui, L., Zhang, X., & Luo, Y. (2011). Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method. *IEEE Transactions on Neural Networks*, *22*(12), 2226–2236.

Zhang, H., Luo, Y., & Liu, D. (2009). Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints. *IEEE Transactions on Neural Networks*, *20*(9), 1490–1503.

Zhang, H., Wei, Q., & Luo, Y. (2008). A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy hdp iteration algorithm. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *38*(4), 937–942.

Rushikesh Kamalapurkar received his Bachelor of Technology degree in Mechanical Engineering from Visvesvaraya National Institute of Technology, Nagpur, India. He worked for two years as a Design Engineer at Larsen and Toubro Ltd., Mumbai, India. He received his Master of Science degree and his Doctor of Philosophy degree from the Department of Mechanical and Aerospace Engineering at the University of Florida under the supervision of Dr. Warren E. Dixon. He is currently a postdoctoral researcher with the Nonlinear Controls and Robotics lab at the University of Florida. His research interests include dynamic programming, optimal control, reinforcement learning, and data-driven adaptive control for uncertain nonlinear dynamical systems.

Huyen Dinh received a B.S. Degree in Mechatronics from Hanoi University of Science and Technology, Hanoi, Vietnam in 2006, and M.Eng. and Ph.D. Degrees in Mechanical Engineering from University of Florida in 2010 and 2012, respectively. She currently works as an Assistant Professor in the Department of Mechanical Engineering at University of Transport and Communications, Hanoi, Vietnam. Her primary research interest is the development of Lyapunov-based control and applications for uncertain nonlinear systems. Current research interests include Learning-based Control, Adaptive Control for uncertain nonlinear systems.

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.

Warren E. 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, he joined the University of Florida in the Mechanical and Aerospace Engineering Department. His main research interest has been the development and application of Lyapunov-based control techniques for uncertain nonlinear systems. He has published over 300 refereed papers and several books in this area. His work has been recognized by the 2013 Fred Ellersick Award for Best Overall MILCOM Paper, 2012–2013 University of Florida College of Engineering Doctoral Dissertation Mentoring Award, 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 (Best Paper) 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 an IEEE Control Systems Society (CSS) Distinguished Lecturer. He currently serves as a member of the US Air Force Science Advisory Board and as the Director of Operations for the Executive Committee of the IEEE CSS Board of Governors. He has formerly served as an associate editor for several journals, and is currently an associate editor for Automatica and the International Journal of Robust and Nonlinear Control.