Approximate optimal trajectory tracking for continuous-time nonlinear systems

Rushikesh Kamalapurkar\textsuperscript{a}, Huyen Dinh\textsuperscript{b}, Shubhendu Bhasin\textsuperscript{c}, Warren E. Dixon\textsuperscript{a}

\textsuperscript{a} Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, USA
\textsuperscript{b} Department of Mechanical Engineering, University of Transport and Communications, Hanoi, Viet Nam
\textsuperscript{c} Department of Electrical Engineering, Indian Institute of Technology, Delhi, India

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 (Diersk & 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
Assumption 2. The proposed method, a comparison with a numerical optimal solvability of the presented technique. To gauge the performance of resulting value function is a time-invariant function of the trans-
mal problem by including the partial derivative of the value calibrate obstacles that result from the time-varying nature of the opti-
mate convergence of these sequence of value functions to the opti-
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 trans-
finite horizon tracking problem that involves offline training of NNs. An ADP-based approach is presented in Dierks and Jagan-
ics can then be expressed as

d= 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 \( e = x - x_d \). The open-loop tracking error dynamics can then be expressed as

\[
e = f (t) + h (y) + e = x \quad \text{for all} \quad t \in [0, T].
\]

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 \( g^T : \mathbb{R}^n \to \mathbb{R}^{m \times n} \) is defined as \( g^T \) and \( g^T \) 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 \( x_d = h_d (x_d) \) and \( g (x) g^T (x_d) (h_d (x_d) - f (x_d)) = h_d (x_d) - f (x_d) \), \( \forall t \in [0, T] \).

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

\[
u_d (x_d) = g_d^T (h_d (x_d) - f_d) ,
\]

where \( g_d^T \triangleq g^T (x_d) \) 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 \( x \in \mathbb{R}^{2n} \) is defined as \( (Zhang et al., 2008) \)

\[
\dot{\xi} = [e , x_d]^T .
\]

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

\[
\dot{\xi} = F (\xi) + G (\xi) \mu ,
\]

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

\[
F (\xi) \triangleq \begin{bmatrix} f (e + x_d) - h_d (x_d) + g (e + x_d) u_d (x_d) \end{bmatrix} ,
\]

\[
G (\xi) \triangleq \begin{bmatrix} g (e + x_d) \end{bmatrix} , \quad \mu \triangleq u - u_d .
\]

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 \( \mu : \mathbb{R}^{2n} \to \mathbb{R}^m \) such that the control law \( \mu = \mu^* (\xi) \) minimizes the cost functional

\[
J (\xi , \mu) = \int_{0}^{\infty} r (\xi (t) , \mu (t)) d t ,
\]

subject to the dynamic constraints in (4), where \( \Psi \) is the set of ad-
misible 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

\[
r (\xi , \mu) \triangleq \xi ^T \overline{Q} \xi + \mu ^T R \mu .
\]

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

\[
\overline{Q} = \begin{bmatrix} Q & 0_{n \times n} \\
0_{n \times n} & 0_{n \times n} \end{bmatrix} ,
\]

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

**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

\[
V^* (\xi) \triangleq \min_{\mu (t)} \int_{0}^{\infty} r (\phi (t ; t , \xi , \mu (t))) d t
\]

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

\[
H^* = V^* (\xi) \left( F (\xi) + G (\xi) \mu^* (\xi) \right) + r (\xi , \mu^* (\xi)) = 0 ,
\]

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

\[
\mu^* (\xi) = - \frac{1}{2} R^{-1} G^T (\xi) \left( V^* (\xi) \right) .
\]
The value function $V^*$ can be represented using a NN with $N$ neurons as
\begin{equation}
V^*(\zeta) = \hat{W}^T \sigma^T (\zeta) \varphi(\zeta) + \epsilon^T (\zeta),
\end{equation}
where $\varphi: \mathbb{R}^N \rightarrow \mathbb{R}$ is the constant ideal weight matrix bounded above by a known positive constant $\hat{W} \in \mathbb{R}$ in the sense that $\| \varphi \| \leq \hat{W}$. $\sigma: \mathbb{R}^2 \rightarrow \mathbb{R}^N$ is a bounded continuously differentiable nonlinear activation function, and $\epsilon: \mathbb{R}^2 \rightarrow \mathbb{R}$ is the function reconstruction error (Hornik, Stinchcombe, & White, 1990; Lewis, Selmic, & Campos, 2002).

Using (10) and (11) the optimal policy can be represented as
\begin{equation}
\mu^*(\zeta) = -\frac{1}{2} \hat{R}^{-1} \sigma^T (\zeta) \varphi(\zeta) W^T + \epsilon^T (\zeta).
\end{equation}

Based on (11) and (12), the NN approximations to the value function and the optimal policy are given by
\begin{equation}
\hat{V}(\zeta, \hat{W}_e) \approx \hat{W}^T \sigma (\zeta),
\end{equation}
\begin{equation}
\mu (\zeta, \hat{W}_e) = -\frac{1}{2} \hat{R}^{-1} \sigma^T (\zeta) \varphi(\zeta) \hat{W}_e,
\end{equation}
where $\hat{W}_e \in \mathbb{R}^N$ and $\hat{W}_e \in \mathbb{R}^N$ are estimates of the ideal neural network weights $W$. The use of two separate sets of weight estimates $\hat{W}_e$ and $\hat{W}_e$ 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
\begin{equation}
u = -\frac{1}{2} \hat{R}^{-1} \sigma^T (\zeta) \varphi(\zeta) \hat{W}_e + \frac{1}{4} \hat{g}^T (\hat{h}_u (\hat{a}_d) - f_d).
\end{equation}

Using the approximations $\mu$ and $\hat{V}$ for $\mu^*$ 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
\begin{equation}
\delta = \hat{V}(\hat{\zeta}, \hat{W}_e) + r(\hat{\zeta}, \mu (\hat{\zeta}, \hat{W}_e)).
\end{equation}

The value function weights are updated to minimize $\int_0^T \sigma^T (\zeta) \varphi(\zeta) \hat{W}_e + \frac{1}{4} \hat{g}^T (\hat{h}_u (\hat{a}_d) - f_d) \sigma^T (\zeta) \varphi(\zeta) \hat{W}_e$ using a normalized least squares update law
\footnote{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.} with an exponential forgetting factor as (Ioannou & Sun, 1996)
\begin{equation}
\dot{\hat{W}}_e = -\eta_e \Gamma \frac{\sigma^T (\zeta) \varphi(\zeta) \hat{W}_e + \frac{1}{4} \hat{g}^T (\hat{h}_u (\hat{a}_d) - f_d)}{1 + \frac{1}{4} \frac{\sigma^T (\zeta) \varphi(\zeta) \hat{W}_e}{\Gamma}}
\end{equation}
\begin{equation}
\dot{\Gamma} = -\eta_c \Gamma \frac{\sigma^T (\zeta) \varphi(\zeta) \hat{W}_e + \frac{1}{4} \hat{g}^T (\hat{h}_u (\hat{a}_d) - f_d)}{1 + \frac{1}{4} \frac{\sigma^T (\zeta) \varphi(\zeta) \hat{W}_e}{\Gamma}}
\end{equation}
where $\eta_e, \eta_c \in \mathbb{R}$ are constant positive adaptation gains, $\omega: \mathbb{R}^2 \rightarrow \mathbb{R}^N$ is defined as $\omega \triangleq \sigma^T (\zeta) \varphi(\zeta)$, and $\lambda \in (0, 1)$ is the constant forgetting factor for the estimation gain matrix $\Gamma \in \mathbb{R}^N \times \mathbb{R}^N$. The policy weights are updated to follow the critic weights
\footnote{The least-squares approach cannot be used to update the policy weights because the BE is a nonlinear function of the policy weights.} as
\begin{equation}
\dot{\hat{W}}_e = -\eta_{a1} (\hat{W}_e - \hat{W}_e) - \eta_{a2} \hat{W}_e,
\end{equation}
where $\eta_{a1}, \eta_{a2} \in \mathbb{R}$ are constant positive adaptation gains. The following assumption facilitates the stability analysis using PE.

\section{Stability analysis}
Before stating the main result of the paper, three supplementary technical lemmas are stated. To facilitate the discussion, let $\beta \in \mathbb{R}^n \times \mathbb{R}^n$ be a compact set, and let $z \triangleq \beta \cap \mathbb{R}^n \times \mathbb{R}^n$. Using the universal approximation property of NNs, on the compact set $\beta \cap \mathbb{R}^n \times \mathbb{R}^n$, the NN approximation errors can be bounded such that sup $| \varepsilon (\zeta) | \leq \epsilon $ and sup $| \hat{\varepsilon} (\zeta) | \leq \hat{\epsilon}$, where $\varepsilon \in \mathbb{R}$ and $\hat{\varepsilon} \in \mathbb{R}$ are positive constants, and there exists a positive constant $L_f \in \mathbb{R}$ such that sup $| \varepsilon (\zeta) | \leq L_f \| \zeta \|$. Using Assumptions 1 and 2 the following bounds are developed on the compact set $\beta \cap \mathbb{R}^n \times \mathbb{R}^n$ to aid the subsequent stability analysis:
\begin{equation}
\| \frac{1}{4} e' + \frac{1}{4} \frac{G^T \sigma^T (\zeta)}{2} \hat{g}^T \| \zeta \| \leq \lambda_1, \quad \| \hat{g} \| \leq \hat{\lambda}_2.
\end{equation}
\begin{equation}
\| \frac{1}{4} e' \hat{g}^T \| \leq \lambda_3, \quad \| \frac{1}{4} \frac{G^T \sigma^T (\zeta)}{2} \hat{g}^T \leq \lambda_4.
\end{equation}
\begin{equation}
\| \frac{1}{4} e' \hat{g}^T + \frac{1}{2} \frac{G^T \sigma \hat{g}^T}{2} \leq \lambda_5,
\end{equation}
where $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 \in \mathbb{R}$ are positive constants.

\subsection{Supporting lemmas}
The contribution in the previous section was the development of a transformation that enables the optimal policy and the optimal...
value function to be expressed as a time-invariant function of $\xi$. 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 $\xi$, because the matrix $Q$ is positive semi-definite. In this section, this technical obstacle is addressed by exploiting the fact that the time-invariant optimal value function $V^* : \mathbb{R}^n \to \mathbb{R}$ can be interpreted as a time-varying mapping $V^*_t : \mathbb{R}^n \to \mathbb{R}$, such that

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

(23)

for all $e \in \mathbb{R}^n$ and for all $t \in \mathbb{R}_{>0}$. 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 $V^*_t : \mathbb{R}^n \to \mathbb{R}$ 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 $V^*_t : \mathbb{R}^n \to \mathbb{R}$ satisfies the following properties

$$
V^*_t(e, t) \geq \psi(||e||),
$$

(24a)

$$
V^*_t(0, t) = 0,
$$

(24b)

$$
V^*_t(e, t) \leq \psi(||e||),
$$

(24c)

for all $t \in \mathbb{R}_{>0}$ and $\forall e \in B_a$ where $\psi : [0, a] \to \mathbb{R}_{>0}$ and $\psi$ are class $\mathcal{K}$ functions.

**Proof.** See Appendix.

**Lemma 2.** Let $Z(t) : = [e^T \ 0 \ 0 \ 0]^T$, and suppose that $Z(t) \in Z$, for all $t \in [t, t + T]$. Then the NN weights and the tracking errors satisfy

$$
\inf_{t \in [t, t + T]} \|e(t)\|^2 
\leq -\sigma_0 \sup_{t \in [t, t + T]} \|e(t)\|^2 + \sigma_1 T \sup_{t \in [t, t + T]} \|\hat{w}_a(t)\|^2 + \sigma_2 
- \inf_{t \in [t, t + T]} \|\hat{w}_e(t)\|^2 
\leq -\sigma_3 \sup_{t \in [t, t + T]} \|\hat{w}_a(t)\|^2 + \sigma_4 \inf_{t \in [t, t + T]} \|\hat{w}_e(t)\|^2 + \sigma_5 
+ \sigma_6, \tag{25}
\]

where

$$
\sigma_0 = \frac{(1 - 6N^2T^2L_z^2)}{2}, \quad \sigma_1 = \frac{3n}{4} \sup_{t} \|g \hat{g} + \hat{g} \hat{g}^T (h_3 - f_b) - \frac{1}{2} \hat{g} R G \hat{g} - h_3 \|^2,
$$

$$
\sigma_2 = \frac{(1 - 6N (\eta_1 + \eta_2) T^2)}{2},
$$

$$
\sigma_3 = \frac{6N \eta_1 T^2}{2 (1 - 6N (\eta_1 + \eta_2) T^2)^2},
$$

$$
\sigma_4 = \frac{18 (n \eta_1 \eta_2 \psi^2 L^2 T^2)}{\psi^2 (1 - 6N (\eta_1 + \eta_2) T^2)^2},
$$

$$
\sigma_5 = \frac{\psi 18 (N \eta_1 \eta_2 \psi^2 L^2 T^2)}{\psi^2 (1 - 6N (\eta_1 + \eta_2) T^2)^2} + \frac{3N (\eta_2 \psi L^2 T^2)}{\psi^2 (1 - 6N (\eta_1 + \eta_2) T^2)^2},
$$

$$
\sigma_6 = \frac{(1 - 6N (\eta_1 + \eta_2) T^2)}{2 \sqrt{\eta_1 \eta_2 \psi^2 L^2 T^2}},
$$

\[\]
where \( r \triangleq \frac{1}{1} \text{sup}_{z,y \in Z} \| z - y \| \) denotes the radius of Z. Since the Lipschitz constant and the bounds on NN approximation error depend on the size of the compact set Z, the constant \( Z \) 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 \( V_t \) is

\[
\dot{V}_t = V^* F + V^* G \mu + \dot{W}_t \Gamma^{-1} \dot{W}^\top_t - \dot{W}^\top_t \dot{W}_t = \frac{1}{2} \dot{W}_t^\top \Gamma^{-1} \text{tr} \Gamma^{-1} \dot{W}_t.
\]

Provided the sufficient conditions in (27) are satisfied, (16), (21), the bounds in (20)–(22), and the facts that \( V^* F = - \dot{W}^\top_t \dot{W}_t \) and \( V^* G = -2 \mu^* R \) yield

\[
\dot{V}_t \leq \frac{q}{2} \| e \|^2 - \frac{1}{2} \eta_1 \| \dot{W}_t \|^2 - \eta_2 \| e \|^2 + t.
\]

The inequality in (29) is valid provided \( Z (t) \in Z \).

Integrating (29), using the facts that \( - \int_{r}^{r+T} \| e (r) \|^2 \, dr \leq -T \inf_{0 \leq t \leq T} \| e (t) \|^2 \) and \( - \int_{r}^{r+T} \| W_t (r) \|^2 \, dr \leq -T \inf_{0 \leq t \leq T} \| W_t (t) \|^2 \), Lemmas 2 and 3, and the gain conditions in (27) yields

\[
V_t (Z (t + T) , t + T) - V_t (Z (t) , t) \leq \frac{- \eta_1 \eta_2 T}{16} \| e (t) \|^2 - \frac{\sigma_0 q T}{8} \| e (t) \|^2 + \sigma_0 T.
\]

Provided \( Z (t) \in Z \), \( \forall r \in [t, t + T] \). Thus, \( V_t (Z (t + T) , t + T) - V_t (Z (t) , t) < 0 \) provided \( Z (t) \| \geq \sqrt{r} \) and \( Z (t) \in Z \), \( \forall r \in [t, t + T] \). The bounds on the Lyapunov function in (25) yield \( V_t (Z (t + T) , t + T) - V_t (Z (t) , t) < 0 \) provided \( V_t (Z (t) , t) > \sqrt{r} \) and \( Z (t) \in Z \), \( \forall r \in [t, t + T] \).

Since \( Z (t) \in Z \), (29) can be used to conclude that \( \dot{V}_t (Z (t_0) , t_0) \leq 0 \). The sufficient condition in (28) ensures that \( \int_{t_0}^{t_1} (V_t (Z (t_0) , t_0)) \, ds \leq r \); hence, \( Z (t) \in Z \) for all \( t \in [t_0, t_0 + T] \). If \( V_t (Z (t_0) , t_0) > \sqrt{r} \), then \( Z (t) \in Z \) for all \( t \in [t_0, t_0 + T] \) implies \( V_t (Z (t_0 + T) , t_0 + T) - V_t (Z (t_0) , t_0) < 0 \); hence, \( V_t (Z (t_0 + T) , t_0 + T) \leq r \). Thus, \( Z (t) \in Z \) for all \( t \in [t_0, t_0 + T + \tau] \). Inductively, the system is bounded such that \( \sup_{t \in [0, T]} \| Z (t) \| \leq r \) and ultimately bounded\(^7\) such that

\[
\lim_{t \to \infty} \sup_{t \in [0, T]} \| Z (t) \| \leq r.
\]

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

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

\[
M \ddot{q} + V_m \ddot{q} + F \ddot{q} + F_s = u,
\]

where \( q = [q_1, q_2] \) and \( \dot{q} = [q_1, q_2]^T \) are the angular positions in radians and the angular velocities in radian/s respectively. In (30), \( M \in \mathbb{R}^{3 \times 3} \) denotes the inertia matrix, and \( V_m \in \mathbb{R}^{3 \times 2} \) denotes the centripetal–Coriolis matrix given by

\[
V_m \triangleq \begin{bmatrix}
-p_{32} q_2 & -p_{32} (q_1 + q_2) \\
p_{32} q_1 & p_{32} (q_1 + q_2)
\end{bmatrix},
\]

where \( c_2 = \cos (q_2), s_2 = \sin (q_2), p_1 = 3.473 \text{ kg m}^2, p_2 = 0.196 \text{ kg m}^2, \) and \( p_3 = 0.242 \text{ kg m}^2, \) and \( F_d = \text{diag} [5.3, 1.1] \) N m and \( F_s (\dot{q}) = [8.45 \ tanh (q_1), 2.35 \ tanh (q_1)]^T \) N m are the models for the static and the dynamic friction, respectively.

The objective is to find a policy \( \mu \) that ensures that the state \( x \triangleq [q_1, q_2, \dot{q}_1, \dot{q}_2] \) tracks the desired trajectory \( \dot{x}_d (t) = [0.5 \ cos (2t), 0.33 \ cos (3t), - \ sin (2t), - \ sin (3t)]^T \), while minimizing the cost \( J = \int_{0}^{\infty} (\dot{e}^T q \ddot{e} + \mu^T \mu) \, dt \), where \( \dot{e} = \text{diag} [10, 10, 2, 2] \). Using (2)–(5) and the definitions

\[
f (x_3, x_4, -M^{-1} (V_m - F_d) \begin{bmatrix} x_3 \\ x_4 \end{bmatrix})^T,
\]

\[
g_d (0, 0, 0)^T,
\]

\[
h_d (x_3, x_4, -4x_3, -9x_4)^T
\]

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

\[
\sigma (\zeta) = \frac{1}{2} \begin{bmatrix}
\zeta_1^2 & \zeta_1 \zeta_2 & \zeta_1 \zeta_3 & \zeta_1 \zeta_4 & \zeta_2 \zeta_3 & \zeta_2 \zeta_4 & \zeta_3 \zeta_4 & \zeta_1^3 & \zeta_2^3 & \zeta_3^3 & \zeta_4^3 & \zeta_1^2 \zeta_2 & \zeta_1 \zeta_2^2 & \zeta_2 \zeta_3 & \zeta_2 \zeta_4 & \zeta_3 \zeta_4 & \zeta_1 \zeta_2 \zeta_3 & \zeta_1 \zeta_2 \zeta_4 & \zeta_1 \zeta_3 \zeta_4 & \zeta_2 \zeta_3 \zeta_4
\end{bmatrix}^T.
\]

The control gains are selected as \( \eta_{a1} = 5, \eta_{a2} = 0.001, \eta_c = 1.25, \lambda = 0.01, \) and \( \nu = 0.005 \), and the initial conditions are \( x (0) = [1.8, 1.6, 0, 0]^T, \dot{W}_t (0) = 10 \times 1_{23 	imes 1}, W_s (0) = 6 \times 1_{23 	imes 1}, \) and \( \Gamma (0) = 2000 \times 1_{23 	imes 1} \), where \( 1_{23 	imes 1} \) is a vector of ones. To ensure PE, a probing signal

\[
p (t) = \begin{bmatrix}
2.55 \ tanh (2t) \sin (\sqrt{232} \pi t) \cos (\sqrt{20} \pi t) \\
+ 6 \sin (18 \pi t^2) + 20 \ cos (40 \pi t) \cos (21t)
\end{bmatrix}
\]

is added to the control signal for the first 30 s of the simulation (Namvoudakis & 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.
The NN weights converge to the following values

\[ \hat{W}_c = \hat{W}_a = \begin{bmatrix} 83.36 & 2.37 & 27.0 & 2.78 & -2.83 & 0.20 & 14.13 \\ 29.81 & 18.87 & 4.11 & 3.47 & 6.69 & 9.71 & 15.58 & 4.97 & 12.42 \\ 11.31 & 3.29 & 1.19 & -1.99 & 4.55 & -0.47 & 0.56 \end{bmatrix}^T. \] (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 \( \mu \) 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 non-zero. Since these costate variables represent the sensitivity of the cost with respect to the desired trajectories, this
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 an increased computational cost, by adding more basis functions to the existing set in (32). The total cost \[ J = \int_0^{100} (e(t)^T \dot{Q} e(t) + \mu(t)^T \dot{R} u(t) + \mu(t)^T R \mu(t)) \, dt \] obtained using the numerical solution is found to be 75.42 and the total cost \[ \int_0^{100} (e(t)^T \dot{Q} e(t) + \mu(t)^T \dot{R} u(t) + \mu(t)^T R \mu(t)) \, dt \] 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 approximate the infinite horizon 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 \( \{ \sigma \} \) for any parameter \( \sigma \) denotes the value of \( \sigma \) computed in the \( i \)th iteration. Algorithm 1 ensures satisfaction of the sufficient condition in (28).

Algorithm 1 Gain Selection

First iteration:
Given \( Z_0 \in \mathbb{R}^{n+1} \) such that \( \| Z(t_0) \| < Z_0 \), let \( Z_1 = \{ \varrho \in \mathbb{R}^{n+1} \| \| \| \leq \beta_1 \| \bar{\vartheta}(\| Z_0 \|) \} \) for some \( \beta_1 > 1 \). Using \( Z_1 \), compute the bounds in (22) and (26), and select the gains according to (27). If \( \| Z \|_1 \leq \beta_1 \| \bar{\vartheta}(\| Z_0 \|) \), set \( Z = Z_1 \) and terminate.

Second iteration:
If \( \| Z \|_1 > \beta_1 \| \bar{\vartheta}(\| Z_0 \|) \), let \( Z_2 \equiv \{ \varrho \in \mathbb{R}^{n+1} \| \varrho \| \leq \beta_2 \| \bar{\vartheta}(\| Z_0 \|) \} \). Using \( Z_2 \), compute the bounds in (22) and (26) and select the gains according to (27). If \( \| Z \|_2 \leq \| Z \|_1 \), set \( Z = Z_2 \) and terminate.

Third iteration:
If \( \| Z \|_2 > \| Z \|_1 \), increase the number of NN neurons to \( N_1 \) to yield a lower function approximation error \( \{ \bar{\vartheta} \} \) such that \( \| L_{\bar{\vartheta}} \|_2 \| \bar{\vartheta} \|_1 \leq \beta_3 \| \bar{\vartheta} \|_1 \). The increase in the number of NN neurons ensures that \( t|_3 \leq t|_1 \). Furthermore, the assumption that the PE interval \( T_1 \) is small enough such that \( \{ L_{\bar{\vartheta}} \} \| T_1 \|_1 \leq \| T \|_1 \| N \|_1 \) ensures that \( \| \bar{\vartheta} \| \| T \|_1 \| N \|_1 \) such that \( \| \bar{\vartheta} \| \| N \|_1 \) and hence, \( \| Z \|_3 \leq \beta_2 \| Z \|_1 \). Set \( Z = \{ \varrho \in \mathbb{R}^{n+1} \| \varrho \| \leq \beta_2 \| Z \|_1 \} \) 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 \( \bar{\vartheta} : D \times \mathbb{R}^m \to \mathbb{R}^n \) be positive definite. If \( t \mapsto \bar{\vartheta}(x, t) \) is uniformly bounded for all \( x \in D \) and if \( x \mapsto \bar{\vartheta}(x, t) \) is continuous, uniformly in \( t \), then \( \bar{\vartheta} \) is decrescent in \( D \).
Proof. Since \( t \rightarrow \mathcal{S}(x, t) \) is uniformly bounded, for all \( x \in D, \) \( \sup_{t \in \mathbb{R}^n} \mathcal{S}(x, t) \) exists and is unique for all \( x \in D. \) Let the function \( \alpha : D \rightarrow \mathbb{R}_{\geq 0} \) be defined as

\[
\alpha(x) = \sup_{t \in \mathbb{R}^n} \{ \mathcal{S}(x, t) \}.
\]

Since \( x \rightarrow \mathcal{S}(x, t) \) is continuous, uniformly in \( t, \) \( \forall \varepsilon > 0, \exists \delta(\varepsilon) > 0 \) such that \( \forall y \in D, \)

\[
d_{D \times \mathbb{R}^n}((x, t), (y, t)) < \delta(\varepsilon) \implies d_{\mathbb{R}^n}(x, y) < \varepsilon,
\]

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

\[
d_D(x, y) < \varepsilon \implies \mathcal{S}(x, t) - \mathcal{S}(y, t) < \varepsilon.
\]

Given the fact that \( \mathcal{S} \) is positive, (37) implies \( \mathcal{S}(x, t) < \mathcal{S}(y, t) + \varepsilon \) and \( \mathcal{S}(y, t) < \mathcal{S}(x, t) + \varepsilon \) which from (35) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon, \) and hence, from (37), \( d_D(x, y) < \varepsilon \) implies \( \alpha(x) < \alpha(y) + \varepsilon \) and \( \alpha(y) < \alpha(x) + \varepsilon,
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.