Model-based reinforcement learning for infinite-horizon approximate optimal tracking

Rushikesh Kamalapurkar, Lindsey Andrews, Patrick Walters, and Warren E. Dixon

***Abstract*—This paper provides an approximate online adaptive solution to the infinite-horizon optimal tracking problem for control-affine continuous-time nonlinear systems with unknown drift dynamics where model-based reinforcement learning is used to relax the persistence of excitation condition. Model-based reinforcement learning is implemented using a concurrent learning-based system identifier to simulate experience by evaluating the Bellman error over unexplored areas of the state space. Tracking of the desired trajectory and convergence of the developed policy to a neighborhood of the optimal policy is established via Lyapunov-based stability analysis.**

I. Introduction

In the past few decades, reinforcement learning (RL)-based techniques have been effectively utilized to obtain online approximate solutions to optimal control problems for systems with finite state-action spaces, and stationary environments (cf. [1], [2]). However, progress for systems with continuous state-action spaces has been slow due to many technical challenges [3], [4]. Various implementations of RL-based learning strategies to solve deterministic optimal regulation problems can be found in results such as [5]–[16].

RL in systems with continuous state and action spaces is realized via value function approximation, where the value function corresponding to the optimal control problem is approximated using a parametric universal approximator. The control policy is generally derived from the approximate value function, and hence, obtaining a good approximation of the value function is critical to the stability of the closed-loop system. In trajectory tracking problems, the value function depends explicitly on time. Since universal function approximators can approximate functions with arbitrary accuracy only on compact domains, value functions for infinite-horizon optimal tracking problems can not be approximated with arbitrary accuracy using universal function approximators [17], [18].

If the desired trajectory can be expressed as the output of an autonomous dynamical system, then the value function can be expressed as a stationary (time-independent) function of the state and the desired trajectory. Hence, universal function approximators can be employed to approximate the value function with arbitrary accuracy by using the system state, augmented with the desired trajectory, as the training input. The state augmentation-based approach is used to solve optimal tracking problems in [17]–[20].

Other offline and online approaches to solve infinite-horizon tracking problems are proposed in results such as [21]–[24], where the explicit dependence of the value function on time is not considered, and hence, the motivation behind using universal approximators to approximate the value function is unclear. Online techniques to solve infinite-horizon optimal control problems are presented in results such as [15], [18]–[21] for linear and nonlinear systems; however, these techniques require persistence of excitation (PE) of the error states to establish convergence. In general, it is impossible to guarantee PE a priori; hence, a probing signal designed using trial and error is added to the controller to ensure PE. The probing signal is not considered in the stability analysis; hence, stability of the closed-loop implementation can not be guaranteed.

In this paper, an infinite-horizon optimal tracking problem is solved for control-affine continuous-time nonlinear systems with unknown drift dynamics using model-based RL. Model-based RL is implemented using a concurrent learning (CL)-based system identifier (cf. [25]–[28]) to simulate experience by evaluating the Bellman error (BE) over unexplored areas of the state space (cf. [28]). The main contributions of this work are the following.

- Approximate model inversion using a CL-based system identifier to compute the desired steady-state controller in the presence of uncertainties in drift dynamics.
- Implementation of model-based RL to relax the PE condition, thereby eliminating the need for the addition of an ad-hoc probing signal.

II. Problem formulation and exact solution

Consider a nonlinear control affine system described by the differential equation

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

where $x \in \mathbb{R}^n$ denotes the state, $u \in \mathbb{R}^m$ denotes the control input, and $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are locally Lipschitz continuous functions that denote the drift dynamics, and the control effectiveness, respectively.

The control objective is to optimally track a time-varying desired trajectory $x_d \in \mathbb{R}^n$. To facilitate the subsequent

Rushikesh Kamalapurkar, Lindsey Andrews, Patrick Walters, and Warren E. Dixon are with the Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, FL, USA. Email: {rkamalapurkar, landr010, walters8, wdixon}@ufl.edu.

This research is supported in part by NSF award numbers 1161260 and 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.

control development, an error signal $e \in \mathbb{R}^n$ is defined as

$$e \triangleq x - x_d. \tag{2}$$

Since the steady-state control input that is required for the system in (1) to track a desired trajectory is, in general, not identically zero, an infinite-horizon optimal control problem formulated in terms of a quadratic cost function containing e and u is not well defined. To address this issue, an alternative cost function is formulated in terms of the tracking error and the mismatch between the actual control signal and the desired steady-state control. The following assumptions facilitate the determination of the desired steady-state control.

Assumption 1. The function g is bounded, the matrix $g\left(x\right)$ 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 $\dot{x}_d = h_d\left(x_d\right)$ and $g\left(x_d\right) g^+\left(x_d\right)\left(h_d\left(x_d\right) - f\left(x_d\right)\right) = h_d\left(x_d\right) - f\left(x_d\right)$, $\forall t \in \mathbb{R}_{\geq t_0}$.

Based on Assumptions 1 and 2, the steady-state control policy $u_d : \mathbb{R}^n \to \mathbb{R}^m$ required for the system in (1) to track the desired trajectory x_d can be expressed as

$$u_d\left(x_d\right) = g_d^+\left(h_d\left(x_d\right) - f_d\right), \tag{3}$$

where $f_d \triangleq f\left(x_d\right)$ and $g_d^+ \triangleq g^+\left(x_d\right)$. The error between the actual control signal and the desired steady-state control signal is defined as

$$\mu \triangleq u - u_d\left(x_d\right). \tag{4}$$

Using (1), (2), and (4), the system dynamics can be expressed in the autonomous form

$$\dot{\zeta} = F\left(\zeta\right) + G\left(\zeta\right)\mu, \tag{5}$$

where the concatenated state $\zeta \in \mathbb{R}^{2n}$ is defined as

$$\zeta \triangleq \left[e^T, \, x_d^T\right]^T, \tag{6}$$

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

$$F\left(\zeta\right) \triangleq \begin{bmatrix} f\left(e + x_d\right) - h_d + g\left(e + x_d\right) u_d\left(x_d\right) \\ h_d \end{bmatrix},$$
$$G\left(\zeta\right) \triangleq \begin{bmatrix} g\left(e + x_d\right) \\ 0 \end{bmatrix}.$$

The control error μ is treated hereafter as the design variable. The control objective is to solve the infinite-horizon optimal regulation problem online, i.e., to simultaneously synthesize and utilize a control signal μ online to minimize the cost functional

$$J\left(\zeta, \mu\right) \triangleq \int_{t_0}^{\infty} r\left(\zeta\left(\tau\right), \mu\left(\tau\right)\right) d\tau, \tag{7}$$

under the dynamic constraint in (5) while tracking the desired trajectory, where $r : \mathbb{R}^{2n} \times \mathbb{R}^m \to \mathbb{R}$ is the local cost defined as

$$r\left(\zeta, \xi\right) \triangleq Q\left(e\right) + \xi^T R \xi. \tag{8}$$

In (8), $R \in \mathbb{R}^{m \times m}$ is a positive definite symmetric matrix of constants, and $Q : \mathbb{R}^n \to \mathbb{R}$ is a continuous positive definite function that satisfies

$$\underline{q}\left(\|e\|\right) \leq Q\left(e\right) \leq \overline{q}\left(\|e\|\right),$$

where $\underline{q} : \mathbb{R} \to \mathbb{R}$ and $\overline{q} : \mathbb{R} \to \mathbb{R}$ are class $\mathcal{K}$ functions.

Assuming that an optimal policy exists, the optimal policy can be characterized in terms of the value function $V^* : \mathbb{R}^{2n} \to \mathbb{R}$ defined as [1]

$$V^*(\zeta) \triangleq \min_{\xi(\tau) | \tau \in \mathbb{R}_{\geq t}} \int_{t}^{\infty} r\left(\phi^{\xi}(\tau, t, \zeta), \xi\left(\phi^{\xi}(\tau, t, \zeta)\right)\right) d\tau, \tag{9}$$

where the notation $\phi^{\xi}\left(t; t_0, \zeta_0\right)$ denotes a trajectory of the system in (5) under the control signal $\xi : \mathbb{R}_{\geq 0} \to \mathbb{R}^m$ with the initial condition $\zeta_0 \in \mathbb{R}^{2n}$ and initial time $t_0 \in \mathbb{R}_{\geq 0}$. Assuming that a minimizing policy exists and that V^* is continuously differentiable, a closed-form solution for the optimal policy can be obtained as [29]

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

where $\nabla_{\zeta}\left(\cdot\right) \triangleq \frac{\partial(\cdot)}{\partial \zeta}$. The policy in (10) and the value function in (9) satisfy the Hamilton-Jacobi-Bellman (HJB) equation [29]

$$\nabla_{\zeta} V^*(\zeta)(F(\zeta) + G(\zeta)\,\mu^*(\zeta)) + \overline{Q}(\zeta) + \mu^{*T}(\zeta) R \mu^*(\zeta) = 0, \tag{11}$$

$\forall \zeta \in \mathbb{R}^{2n}$, with the initial condition $V^*\left(0\right) = 0$. In (11), the function $\overline{Q} : \mathbb{R}^{2n} \to \mathbb{R}$ is defined as

$$\overline{Q}\left(\begin{bmatrix} e \\ x_d \end{bmatrix}\right) = Q\left(e\right), \, \forall x_d \in \mathbb{R}^n. \tag{12}$$

III. Bellman Error

Since a closed-form solution of the HJB in (11) is generally infeasible to obtain, an approximate solution is sought. In an approximate actor-critic-based solution, the optimal value function V^* is replaced by a parametric estimate $\hat{V}\left(\zeta, \hat{W}_c\right)$ and the optimal policy u^* by a parametric estimate $\hat{u}\left(\zeta, \hat{W}_a\right)$ where $\hat{W}_c \in \mathbb{R}^L$ and $\hat{W}_a \in \mathbb{R}^L$ denote vectors of estimates of the ideal parameters. The objective of the critic is to learn the parameters $\hat{W}_c$, and the objective of the actor is to learn the parameters $\hat{W}_a$. Substituting the estimates $\hat{V}$ and $\hat{u}$ for V^* and u^* in (11), respectively, a residual error $\delta : \mathbb{R}^{2n} \times \mathbb{R}^L \times \mathbb{R}^L \to \mathbb{R}$, called the BE, is defined as

$$\delta\left(\zeta, \hat{W}_c, \hat{W}_a\right) = \overline{Q}\left(\zeta\right) + \hat{\mu}^T\left(\zeta, \hat{W}_a\right) R \hat{\mu}\left(\zeta, \hat{W}_a\right)$$

[1] Since the closed-loop system corresponding to (5) under a feedback policy is autonomous, the cost-to-go, i.e., the integral in (9) is independent of initial time. Hence, the value function is only a function of ζ.

$$+\nabla_{\zeta}\hat{V}\left(\zeta,\hat{W}_c\right)\left(F\left(\zeta\right)+G\left(\zeta\right)\hat{\mu}\left(\zeta,\hat{W}_a\right)\right). \quad (13)$$

To solve the optimal control problem, the critic aims to find a set of parameters $\hat{W}_c$ and the actor aims to find a set of parameters $\hat{W}_a$ such that $\delta\left(\zeta,\hat{W}_c,\hat{W}_a\right)=0$, and $\hat{u}\left(\zeta,\hat{W}_a\right)=-\frac{1}{2}R^{-1}G^T(\zeta)\left(\nabla_{\zeta}\hat{V}\left(\zeta,\hat{W}_a\right)\right)^T$, $\forall\zeta\in\mathbb{R}^{2n}$. Since an exact basis for value function approximation is generally not available, an approximate set of parameters that minimizes the BE is sought.

Since the BE in (13) requires model knowledge, a dynamic system identifier is developed to generate a parametric estimate $\hat{F}\left(\zeta,\hat{\theta}\right)$ of the drift dynamics F, where $\hat{\theta}$ denotes the estimate of the matrix of unknown parameters. Given $\hat{F}$, $\hat{V}$, and $\hat{\mu}$, an estimate of the BE can be evaluated at any $\zeta\in\mathbb{R}^{2n}$. That is, using $\hat{F}$, experience can be simulated by extrapolating the BE over unexplored off-trajectory points in the operating domain. Hence, if an identifier can be developed such that $\hat{F}$ approaches F exponentially fast, learning laws for the optimal policy can utilize simulated experience along with experience gained and stored along the state trajectory.

If parametric approximators are used to approximate F, convergence of $\hat{F}$ to F is implied by convergence of the parameters to their unknown ideal values. It is well known that adaptive system identifiers require PE to achieve parameter convergence. To relax the PE condition, a CL-based (cf. [25]–[28]) system identifier that uses recorded data for learning is developed in the following section.

IV. System Identification

On any compact set $\mathcal{C}\subset\mathbb{R}^n$ the function f can be represented using a neural network (NN) as

$$f\left(x\right)=\theta^T\sigma_f\left(Y^Tx_1\right)+\epsilon_\theta\left(x\right), \quad (14)$$

where $x_1\triangleq\begin{bmatrix}1\\x\end{bmatrix}\in\mathbb{R}^{n+1}$, $\theta\in\mathbb{R}^{p+1\times n}$ and $Y\in\mathbb{R}^{n+1\times p}$ denote the unknown output-layer and hidden-layer NN weights, $\sigma_f:\mathbb{R}^p\to\mathbb{R}^{p+1}$ denotes a bounded NN basis function, $\epsilon_\theta:\mathbb{R}^n\to\mathbb{R}^n$ denotes the function reconstruction error, and $p\in\mathbb{N}$ denotes the number of NN neurons. Using the universal function approximation property of single layer NNs, given a constant matrix Y such that the rows of $\sigma_f\left(Y^Tx_1\right)$ form a proper basis, there exist constant ideal weights θ and known constants $\overline{\theta}$, $\overline{\epsilon_\theta}$, and $\overline{\epsilon_\theta'}\in\mathbb{R}$ such that $\left\|\theta\right\|_F\leq\overline{\theta}<\infty$, $\sup_{x\in\mathcal{C}}\left\|\epsilon_\theta\left(x\right)\right\|\leq\overline{\epsilon_\theta}$, and $\sup_{x\in\mathcal{C}}\left\|\nabla_x\epsilon_\theta\left(x\right)\right\|\leq\overline{\epsilon_\theta'}$, where $\left\|\cdot\right\|_F$ denotes the Frobenius norm.

Using an estimate $\hat{\theta}\in\mathbb{R}^{p+1\times n}$ of the weight matrix θ, the function f can be approximated by the function $\hat{f}:\mathbb{R}^{2n}\times\mathbb{R}^{p+1\times n}\to\mathbb{R}^n$ defined as

$$\hat{f}\left(\zeta,\hat{\theta}\right)\triangleq\hat{\theta}^T\sigma_\theta\left(\zeta\right), \quad (15)$$

where $\sigma_\theta:\mathbb{R}^{2n}\to\mathbb{R}^{p+1}$ is defined as $\sigma_\theta\left(\zeta\right)=\sigma_f\left(Y^T\begin{bmatrix}1 & e^T+x_d^T\end{bmatrix}^T\right)$. Based on (14), an estimator for online identification of the drift dynamics is developed as

$$\dot{\hat{x}}=\hat{\theta}^T\sigma_\theta\left(\zeta\right)+g\left(x\right)u+k\tilde{x}, \quad (16)$$

where $\tilde{x}\triangleq x-\hat{x}$, and $k\in\mathbb{R}$ is a positive constant learning gain. The following assumption facilitates CL-based system identification.

Assumption 3. [26] A history stack containing recorded state-action pairs $\left\{x_j,u_j\right\}_{j=1}^M$ along with numerically computed state derivatives $\left\{\dot{\bar{x}}_j\right\}_{j=1}^M$ that satisfies

$$\lambda_{\min}\left(\sum_{j=1}^M\sigma_{fj}\sigma_{fj}^T\right)=\underline{\sigma_\theta}>0,\quad\left\|\dot{\bar{x}}_j-\dot{x}_j\right\|<\overline{d},\ \forall j \quad (17)$$

is available a priori. In (17), $\sigma_{fj}\triangleq\sigma_f\left(Y^T\begin{bmatrix}1\\x_j\end{bmatrix}\right)$, $\overline{d}\in\mathbb{R}$ is a known positive constant, and $\lambda_{\min}\left(\cdot\right)$ denotes the minimum eigenvalue.

The weight estimates $\hat{\theta}$ are updated using the following CL-based update law:

$$\dot{\hat{\theta}}=\Gamma_\theta\sigma_f\left(Y^Tx_1\right)\tilde{x}^T+k_\theta\Gamma_\theta\sum_{j=1}^M\sigma_{fj}\left(\dot{\bar{x}}_j-g_ju_j-\hat{\theta}^T\sigma_{fj}\right)^T, \quad (18)$$

where $k_\theta\in\mathbb{R}$ is a constant positive CL gain, and $\Gamma_\theta\in\mathbb{R}^{p+1\times p+1}$ is a constant, diagonal, and positive definite adaptation gain matrix.

To facilitate the subsequent stability analysis, a candidate Lyapunov function $V_0:\mathbb{R}^n\times\mathbb{R}^{p+1\times n}\to\mathbb{R}$ is selected as

$$V_0\left(\tilde{x},\tilde{\theta}\right)=\frac{1}{2}\tilde{x}^T\tilde{x}+\frac{1}{2}\text{tr}\left(\tilde{\theta}^T\Gamma_\theta^{-1}\tilde{\theta}\right), \quad (19)$$

where $\tilde{\theta}\triangleq\theta-\hat{\theta}$ and $\text{tr}\left(\cdot\right)$ denotes the trace of a matrix. Using (16)-(18), the following bound on the time derivative of V_0 is established:

$$\dot{V}_0\leq-k\left\|\tilde{x}\right\|^2-k_\theta\underline{\sigma_\theta}\left\|\tilde{\theta}\right\|_F^2+\overline{\epsilon_\theta}\left\|\tilde{x}\right\|+k_\theta\overline{d_\theta}\left\|\tilde{\theta}\right\|_F, \quad (20)$$

where $\overline{d_\theta}\triangleq\overline{d}\sum_{j=1}^M\left\|\sigma_{\theta j}\right\|+\sum_{j=1}^M\left(\left\|\epsilon_{\theta j}\right\|\left\|\sigma_{\theta j}\right\|\right)$. Using (19) and (20) a Lyapunov-based stability analysis can be used to show that $\hat{\theta}$ converges exponentially to a neighborhood around θ.

Using (15), the BE in (13) can be approximated as

$$\hat{\delta}\left(\zeta,\hat{\theta},\hat{W}_c,\hat{W}_a\right)=\overline{Q}\left(\zeta\right)+\hat{\mu}^T\left(\zeta,\hat{W}_a\right)R\hat{\mu}\left(\zeta,\hat{W}_a\right),$$
$$+\nabla_\zeta\hat{V}\left(\zeta,\hat{W}_a\right)\left(F_\theta\left(\zeta,\hat{\theta}\right)+F_1\left(\zeta\right)+G\left(\zeta\right)\hat{\mu}\left(\zeta,\hat{W}_a\right)\right) \quad (21)$$

where

$$F_\theta\left(\zeta,\hat{\theta}\right)\triangleq\begin{bmatrix}\hat{\theta}^T\sigma_\theta\left(\zeta\right)-g\left(x\right)g^+\left(x_d\right)\hat{\theta}^T\sigma_\theta\left(\begin{bmatrix}\mathbf{0}_{n\times1}\\x_d\end{bmatrix}\right)\\0\end{bmatrix},$$

and $F_1\left(\zeta\right)\triangleq\begin{bmatrix}-h_d+g\left(e+x_d\right)g^+\left(x_d\right)h_d\\h_d\end{bmatrix}$.

V. Value function approximation

Since V^* and μ^* are functions of the state ζ, the minimization problem stated in Section II is infinite-dimensional, and hence, intractable. To obtain a finite-dimensional minimization problem, the optimal value function is represented over any compact operating domain $\mathcal{C} \subset \mathbb{R}^{2n}$ using a NN as

$$V^*\left(\zeta\right) = W^T\sigma\left(\zeta\right) + \epsilon\left(\zeta\right), \tag{22}$$

where $W \in \mathbb{R}^L$ denotes a vector of unknown NN weights, $\sigma : \mathbb{R}^{2n} \to \mathbb{R}^L$ denotes a bounded NN basis function, $\epsilon : \mathbb{R}^{2n} \to \mathbb{R}$ denotes the function reconstruction error, and $L \in \mathbb{N}$ denotes the number of NN neurons. Using the universal function approximation property of single layer NNs, for any compact set $\mathcal{C} \subset \mathbb{R}^{2n}$, there exist constant ideal weights W and known constants $\overline{W}$, $\overline{\epsilon}$, and $\overline{\epsilon'} \in \mathbb{R}$ such that $\|W\| \leq \overline{W} < \infty$, $\sup_{\zeta\in\mathcal{C}} \|\epsilon\left(\zeta\right)\| \leq \overline{\epsilon}$, and $\sup_{\zeta\in\mathcal{C}} \|\nabla_\zeta \epsilon\left(\zeta\right)\| \leq \overline{\epsilon'}$.

Using (10), a NN representation of the optimal policy is obtained as $\mu^*\left(\zeta\right) = -\frac{1}{2}R^{-1}G^T\left(\zeta\right)\left(\sigma'^T\left(\zeta\right)W + \epsilon'^T\left(\zeta\right)\right)$, where $\sigma' \triangleq \frac{\partial\sigma}{\partial\zeta}$ and $\epsilon' \triangleq \frac{\partial\epsilon}{\partial\zeta}$. Using estimates $\hat{W}_c$ and $\hat{W}_a$ for the ideal weights W, the optimal value function and the optimal policy are approximated as

$$\begin{aligned}\hat{V}\left(\zeta, \hat{W}_c\right) &\triangleq \hat{W}_c^T\sigma\left(\zeta\right),\\ \hat{\mu}\left(\zeta, \hat{W}_a\right) &\triangleq -\frac{1}{2}R^{-1}G^T\left(\zeta\right)\sigma'^T\left(\zeta\right)\hat{W}_a.\end{aligned} \tag{23}$$

The optimal control problem is thus reformulated as the need to find a set of weights $\hat{W}_c$ and $\hat{W}_a$ online, to minimize the error

$$\hat{E}_{\hat{\theta}}\left(\hat{W}_c, \hat{W}_a\right) \triangleq \sup_{\zeta\in\chi}\left|\hat{\delta}\left(\zeta, \hat{\theta}, \hat{W}_c, \hat{W}_a\right)\right|, \tag{24}$$

for a given $\hat{\theta}$, while simultaneously improving $\hat{\theta}$ using (18), and ensuring stability of the system in (1) using the control law

$$u = \hat{\mu}\left(\zeta, \hat{W}_a\right) + \hat{u}_d\left(\zeta, \hat{\theta}\right), \tag{25}$$

where

$$\hat{u}_d\left(\zeta, \hat{\theta}\right) \triangleq g_d^+\left(h_d - \hat{\theta}^T\sigma_{\theta d}\right), \tag{26}$$

and $\sigma_{\theta d} \triangleq \sigma_\theta\left(\begin{bmatrix}\mathbf{0}_{n\times 1}\\ x_d\end{bmatrix}\right)$. Using (3), (25), and (26), the virtual controller μ for the concatenated system in (5) can be expressed as[2]

$$\mu = \hat{\mu}\left(\zeta, \hat{W}_a\right) + g_d^+\tilde{\theta}^T\sigma_{\theta d} + g_d^+\epsilon_{\theta d}, \tag{27}$$

where $\epsilon_{\theta d} \triangleq \epsilon_\theta\left(x_d\right)$.

VI. Simulation of experience

Since computation of the supremum in (24) is intractable in general, simulation of experience is implemented by minimizing a summation of BEs over finitely many points in the state space. The following assumption facilitates the aforementioned approximation.

Assumption 4. [28] There exists a finite set of points $\left\{\zeta_i \in \mathcal{C} \mid i = 1, \cdots, N\right\}$ such that $\forall t \in \mathbb{R}$

$$0 < \underline{c} \triangleq \frac{1}{N}\left(\inf_{t\in\mathbb{R}_{\geq t_0}}\left(\lambda_{min}\left\{\sum_{i=1}^{N}\frac{\omega_i\omega_i^T}{\rho_i}\right\}\right)\right), \tag{28}$$

where $\rho_i \triangleq 1 + \nu\omega_i^T\Gamma\omega_i \in \mathbb{R}$, and $\omega_i \triangleq \sigma'\left(\zeta_i\right)\left(F_\theta\left(\zeta_i, \hat{\theta}\right) + F_1\left(\zeta_i\right) + G\left(\zeta_i\right)\hat{\mu}\left(\zeta_i, \hat{W}_a\right)\right)$.

Using Assumption 4, simulation of experience is implemented by the weight update laws

$$\dot{\hat{W}}_c = -\eta_{c1}\Gamma\frac{\omega}{\rho}\hat{\delta}_t - \frac{\eta_{c2}}{N}\Gamma\sum_{i=1}^{N}\frac{\omega_i}{\rho_i}\hat{\delta}_{ti}, \tag{29}$$

$$\dot{\Gamma} = \left(\beta\Gamma - \eta_{c1}\Gamma\frac{\omega\omega^T}{\rho^2}\Gamma\right)\mathbf{1}_{\left\{\|\Gamma\|\leq\overline{\Gamma}\right\}}, \ \|\Gamma\left(t_0\right)\| \leq \overline{\Gamma}, \tag{30}$$

$$\begin{aligned}\dot{\hat{W}}_a = &-\eta_{a1}\left(\hat{W}_a - \hat{W}_c\right) - \eta_{a2}\hat{W}_a\\ &+ \left(\frac{\eta_{c1}G_\sigma^T\hat{W}_a\omega^T}{4\rho} + \sum_{i=1}^{N}\frac{\eta_{c2}G_{\sigma i}^T\hat{W}_a\omega_i^T}{4N\rho_i}\right)\hat{W}_c,\end{aligned} \tag{31}$$

where $\omega \triangleq \nabla\sigma\left(\zeta\right)\left(F_\theta\left(\zeta, \hat{\theta}\right) + F_1\left(\zeta\right) + G\left(\zeta\right)\hat{\mu}\left(\zeta, \hat{W}_a\right)\right)$, $\Gamma \in \mathbb{R}^{L\times L}$ is the least-squares gain matrix, $\overline{\Gamma} \in \mathbb{R}$ denotes a positive saturation constant, $\beta \in \mathbb{R}$ denotes the forgetting factor, $\eta_{c1}, \eta_{c2}, \eta_{a1}, \eta_{a2} \in \mathbb{R}$ denote constant positive adaptation gains, $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function of the set $\{\cdot\}$, $G_\sigma \triangleq \sigma'\left(\zeta\right)G\left(\zeta\right)R^{-1}G^T\left(\zeta\right)\left(\sigma'\left(\zeta\right)\right)^T$, and $\rho \triangleq 1 + \nu\omega^T\Gamma\omega$, where $\nu \in \mathbb{R}$ is a positive normalization constant. In (29)-(31) and in the subsequent development, for any function $\xi\left(\zeta, \cdot\right)$, the notation ξ_i, is defined as $\xi_i \triangleq \xi\left(\zeta_i, \cdot\right)$, and the instantaneous BEs $\hat{\delta}_t$ and $\hat{\delta}_{ti}$ are defined as

$$\hat{\delta}_t\left(t\right) \triangleq \hat{\delta}\left(\zeta\left(t\right), \hat{W}_c\left(t\right), \hat{W}_a\left(t\right), \hat{\theta}\left(t\right)\right) \tag{32}$$

and $\hat{\delta}_{ti}\left(t\right) \triangleq \hat{\delta}\left(\zeta_i, \hat{W}_c\left(t\right), \hat{W}_a\left(t\right), \hat{\theta}\left(t\right)\right)$. The saturated least-squares update law in (30) ensures that there exist positive constants $\underline{\gamma}, \overline{\gamma} \in \mathbb{R}$ such that

$$\underline{\gamma} \leq \left\|\left(\Gamma\left(t\right)\right)^{-1}\right\| \leq \overline{\gamma}, \ \forall t \in \mathbb{R}. \tag{33}$$

VII. Stability analysis

If the state penalty function $\overline{Q}$ is positive definite, then the optimal value function V^* is positive definite, and serves as a Lyapunov function for the system in (5) under the optimal control policy μ^*; hence, V^* is used (cf. [11], [12], [30]) as a candidate Lyapunov function for the closed-loop system under the policy $\hat{\mu}$. Based on the definition in (12),

[2] The expression in (27) is developed to facilitate the stability analysis, whereas the equivalent expression in (25) is implemented in practice.

the function $\overline{Q}$, and hence, the function V^* are positive semidefinite; hence, the function V^* is not a valid candidate Lyapunov function. However, the results in [18] can be used to show that a nonautonomous form of the optimal value function denoted by $V_t^* : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, defined as

$$V_t^*\left(e,t\right) = V^*\left(\begin{bmatrix} e \\ x_d\left(t\right)\end{bmatrix}\right),\ \forall e \in \mathbb{R}^n,\ t \in \mathbb{R},$$

is positive definite and decrescent. Hence, $V_t^*\left(0,t\right) = 0, \forall t \in \mathbb{R}$ and there exist class $\mathcal{K}$ functions $\underline{v} : \mathbb{R} \to \mathbb{R}$ and $\overline{v} : \mathbb{R} \to \mathbb{R}$ such that

$$\underline{v}\left(\|e\|\right) \leq V_t^*\left(e,t\right) \leq \overline{v}\left(\|e\|\right), \tag{34}$$

for all $e \in \mathbb{R}^n$ and for all $t \in \mathbb{R}$.

To facilitate the stability analysis, a concatenated state $Z \in \mathbb{R}^{2n+2L+n(p+1)}$ is defined as

$$Z \triangleq \begin{bmatrix} e^T & \tilde{W}_c^T & \tilde{W}_a^T & \tilde{x}^T & \left(\mathrm{vec}\left(\tilde{\theta}\right)\right)^T\end{bmatrix}^T,$$

and a candidate Lyapunov function is defined as

$$V_L(Z,t) \triangleq V_t^*(e,t) + \frac{1}{2}\tilde{W}_c^T\Gamma^{-1}\tilde{W}_c + \frac{1}{2}\tilde{W}_a^T\tilde{W}_a + V_0\left(\tilde{\theta},\tilde{x}\right), \tag{35}$$

where $\mathrm{vec}\left(\cdot\right)$ denotes the vectorization operator and V_0 is defined in (19). Using (19), the bounds in (33) and (34), and the fact that $\mathrm{tr}\left(\tilde{\theta}^T\Gamma_\theta^{-1}\tilde{\theta}\right) = \left(\mathrm{vec}\left(\tilde{\theta}\right)\right)^T\left(\Gamma_\theta^{-1}\otimes\mathbb{I}_{p+1}\right)\left(\mathrm{vec}\left(\tilde{\theta}\right)\right)$, the candidate Lyapunov function in (35) can be bounded as

$$\underline{v_l}\left(\|Z\|\right) \leq V_L\left(Z,t\right) \leq \overline{v_l}\left(\|Z\|\right), \tag{36}$$

for all $Z \in \mathbb{R}^{2n+2L+n(p+1)}$ and for all $t \in \mathbb{R}$, where $\underline{v_l} : \mathbb{R} \to \mathbb{R}$ and $\overline{v_l} : \mathbb{R} \to \mathbb{R}$ are class $\mathcal{K}$ functions.

For notational brevity, the dependence of the functions F, G, σ, σ', ϵ, ϵ', σ_θ, ϵ_θ, and g on the system states is suppressed hereafter. To facilitate the stability analysis, the approximate BE in (32) is expressed in terms of the weight estimation errors as

$$\hat{\delta}_t = -\omega^T\tilde{W}_c - W^T\sigma' F_{\tilde{\theta}} + \frac{1}{4}\tilde{W}_a^T G_\sigma \tilde{W}_a + \Delta, \tag{37}$$

where $F_{\tilde{\theta}} \triangleq F_\theta\left(\zeta,\tilde{\theta}\right)$ and $\Delta = O\left(\overline{\epsilon}, \overline{\epsilon'}, \overline{\epsilon_\theta}\right)$. Given any compact set $\chi \subset \mathbb{R}^{2n+2L+n(p+1)}$ containing an open ball of radius $\rho \in \mathbb{R}$ centered at the origin, a positive constant $\iota \in \mathbb{R}$ is defined as

$$\iota \triangleq \frac{3\left(\frac{(\eta_{c1}+\eta_{c2})\overline{W}^2\overline{\|G_\sigma\|}}{16\sqrt{\nu\underline{\Gamma}}} + \frac{\overline{\|(W^TG_\sigma+\epsilon' G_r\sigma'^T)\|}}{4} + \frac{\eta_{a2}\overline{W}}{2}\right)^2}{\left(\eta_{a1}+\eta_{a2}\right)} + \frac{3\left(\left(\overline{\left\|W^T\sigma' Gg_d^+\right\|} + \overline{\left\|\epsilon' Gg_d^+\right\|}\right)\overline{\sigma_g} + k_\theta\overline{d_\theta}\right)^2}{4k_\theta\underline{\sigma_\theta}} + \frac{\left(\eta_{c1}+\eta_{c2}\right)^2\overline{\|\Delta\|}^2}{4\nu\underline{\Gamma}\eta_{c2}\underline{c}} + \frac{\overline{\epsilon_\theta}^2}{2k} + \overline{\left\|\epsilon' Gg_d^+\epsilon_{\theta d}\right\|}$$

$$+ \overline{\left\|\frac{1}{2}G_\epsilon\right\|} + \overline{\left\|\frac{1}{2}W^T\sigma' G_r\epsilon'^T\right\|} + \overline{\left\|W^T\sigma' Gg_d^+\epsilon_{\theta d}\right\|}, \tag{38}$$

where $G_r \triangleq GR^{-1}G^T$, and $G_\epsilon \triangleq \epsilon' G_r\left(\epsilon'\right)^T$. Let $v_l : \mathbb{R} \to \mathbb{R}$ be a class $\mathcal{K}$ function such that

$$v_l\left(\|Z\|\right) \leq \frac{\underline{q}\left(\|e\|\right)}{2} + \frac{\eta_{c2}\underline{c}}{8}\left\|\tilde{W}_c\right\|^2 + \frac{\left(\eta_{a1}+\eta_{a2}\right)}{6}\left\|\tilde{W}_a\right\|^2 + \frac{k}{4}\left\|\tilde{x}\right\|^2 + \frac{k_\theta\underline{\sigma_\theta}}{6}\left\|\mathrm{vec}\left(\tilde{\theta}\right)\right\|^2. \tag{39}$$

The sufficient gain conditions used in the subsequent Theorem 1 are

$$v_l^{-1}\left(\iota\right) < \overline{v_l}^{-1}\left(\underline{v_l}\left(\rho\right)\right) \tag{40}$$

$$\eta_{c2}\underline{c} > \frac{3\left(\eta_{c2}+\eta_{c1}\right)^2\overline{W}^2\overline{\|\sigma'\|}^2\overline{\sigma_g}^2}{4k_\theta\underline{\sigma_\theta}\nu\underline{\Gamma}} \tag{41}$$

$$\left(\eta_{a1}+\eta_{a2}\right) > \frac{3\left(\eta_{c1}+\eta_{c2}\right)\overline{W}\overline{\|G_\sigma\|}}{8\sqrt{\nu\underline{\Gamma}}} + \frac{3}{\underline{c}\eta_{c2}}\left(\frac{\left(\eta_{c1}+\eta_{c2}\right)\overline{W}\overline{\|G_\sigma\|}}{8\sqrt{\nu\underline{\Gamma}}} + \eta_{a1}\right)^2. \tag{42}$$

In (38)-(42), for any function $\varpi : \mathbb{R}^l \to \mathbb{R}$, $l \in \mathbb{N}$, the notation $\overline{\|\varpi\|}$, denotes $\sup_{y\in\chi\cap\mathbb{R}^l}\left\|\varpi\left(y\right)\right\|$, and $\overline{\sigma_g} \triangleq \overline{\|\sigma_\theta\|} + \overline{\left\|gg_d^+\right\|}\,\overline{\|\sigma_{\theta d}\|}$.

The sufficient condition in (40) requires the set χ to be large enough based on the constant ι. Since the NN approximation errors depend on the compact set χ, in general, for a fixed number of NN neurons, the constant ι increases with the size of the set χ. However, for a fixed set χ, the constant ι can be reduced by reducing function reconstruction errors, i.e., by increasing number of NN neurons, and by increasing the learning gains provided $\underline{\sigma_\theta}$ is large enough. Hence sufficient number of NN neurons and extrapolation points are required to satisfy the condition in (40).

Theorem 1. *Provided Assumptions 2-4 hold, and the control gains are selected based on (40)-(42), the controller in (25), along with the weight update laws (29)-(31), and the identifier in (16) along with the weight update law (18) ensure that the system states remain bounded, the tracking error is ultimately bounded, and that the control policy $\hat{\mu}$ converges to a neighborhood around the optimal control policy μ^*.*

Proof: Using (5) and the fact that $\dot{V}_t^*\left(e\left(t\right),t\right) = \dot{V}^*\left(\zeta\left(t\right)\right),\ \forall t \in \mathbb{R}$, the time-derivative of the candidate Lyapunov function in (35) is

$$\dot{V}_L = \nabla_\zeta V^*\left(F + G\mu^*\right) - \tilde{W}_c^T\Gamma^{-1}\dot{\hat{W}}_c - \frac{1}{2}\tilde{W}_c^T\Gamma^{-1}\dot{\Gamma}\Gamma^{-1}\tilde{W}_c - \tilde{W}_a^T\dot{\hat{W}}_a + \dot{V}_0 + \nabla_\zeta V^* G\mu - \nabla_\zeta V^* G\mu^*. \tag{43}$$

Provided the sufficient conditions in (41)-(42) are satisfied, using (11), (23), (27), (37) the bound in (20), and the update laws in (29)-(31) the expression in (43) can be bounded as

$$\dot{V}_L \leq -v_l\left(\|Z\|\right),\ \forall \|Z\| \geq v_l^{-1}\left(\iota\right),\ \forall Z \in \chi. \tag{44}$$

Using (36), (40), and (44) Theorem 4.18 in [31] can be invoked to conclude that every trajectory $Z(t)$ satisfying $\|Z(t_0)\| \leq \overline{v_l}^{-1}\left(\underline{v_l}(\rho)\right)$, is bounded for all $t \in \mathbb{R}$ and satisfies $\limsup_{t\to\infty}\|Z(t)\| \leq \underline{v_l}^{-1}\left(\overline{v_l}\left(v_l^{-1}(\iota)\right)\right)$. ■

VIII. Conclusion

A concurrent-learning based implementation of model-based RL is developed to obtain an approximate online solution to infinite-horizon optimal tracking problems for nonlinear continuous-time control-affine systems. The desired steady-state controller is used to facilitate the formulation of a feasible optimal control problem, and the system state is augmented with the desired trajectory to facilitate the formulation of a stationary optimal control problem. A CL-based system identifier is developed to remove the dependence of the desired steady-state controller on the system drift dynamics, and to facilitate simulation of experience via BE extrapolation. Ultimately bounded tracking and convergence of the developed policy to a neighborhood of the optimal policy is established using a Lyapunov-based analysis.

Similar to the PE condition, the condition in (28) can not, in general, be guaranteed a priori. However, the condition in (28) can be heuristically met by oversampling, i.e., by selecting $N \gg L$. Furthermore, unlike PE, the condition in (28) can be monitored online; hence, threshold-based algorithms can be employed to ensure (28) by selecting new points if the minimum singular value in (28) falls below a certain threshold. Provided the minimum singular value does not decrease during a switch, the trajectories of the resulting switched system can be shown to be uniformly bounded using a common Lyapunov function. Formulation of sufficient conditions for (28) that can be verified a priori is a topic for future research.

References

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

[2] D. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 2007.

[3] P. Mehta and S. Meyn, "Q-learning and pontryagin's minimum principle," in *Proc. IEEE Conf. Decis. Control*, Dec. 2009, pp. 3598 –3605.

[4] M. P. Deisenroth, *Efficient reinforcement learning using Gaussian processes*. KIT Scientific Publishing, 2010.

[5] K. Doya, "Reinforcement learning in continuous time and space," *Neural Comput.*, vol. 12, no. 1, pp. 219–245, 2000.

[6] M. Abu-Khalaf, F. Lewis, and J. Huang, "Policy iterations on the hamilton ndash;jacobi ndash;isaacs equation for h_∞ state feedback control with input saturation," *IEEE Trans. Automat. Control*, vol. 51, no. 12, pp. 1989–1995, Dec 2006.

[7] R. Padhi, N. Unnikrishnan, X. Wang, and S. Balakrishnan, "A single network adaptive critic (SNAC) architecture for optimal control synthesis for a class of nonlinear systems," *Neural Netw.*, vol. 19, no. 10, pp. 1648–1660, 2006.

[8] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 38, pp. 943–949, 2008.

[9] Z. Chen and S. Jagannathan, "Generalized Hamilton-Jacobi-Bellman formulation -based neural network control of affine nonlinear discrete-time systems," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 90–106, Jan. 2008.

[10] T. Dierks, B. Thumati, and S. Jagannathan, "Optimal control of unknown affine nonlinear discrete-time systems using offline-trained neural networks with proof of convergence," *Neural Netw.*, vol. 22, no. 5-6, pp. 851–860, 2009.

[11] K. Vamvoudakis and F. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[12] S. Bhasin, R. Kamalapurkar, M. Johnson, K. Vamvoudakis, F. L. Lewis, and W. Dixon, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 89–92, 2013.

[13] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control Algorithms and Stability*, ser. Communications and Control Engineering. London: Springer-Verlag, 2013.

[14] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 621–634, Mar. 2014.

[15] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, 2014.

[16] X. Yang, D. Liu, and D. Wang, "Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints," *Int. J. Control*, vol. 87, no. 3, pp. 553–566, 2014.

[17] H. Zhang, Q. Wei, and Y. Luo, "A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy hdp iteration algorithm," *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 38, no. 4, pp. 937–942, 2008.

[18] R. Kamalapurkar, H. Dinh, S. Bhasin, and W. E. Dixon, "Approximate optimal trajectory tracking for continuous-time nonlinear systems," *Automatica*, to appear (see also arXiv:1301.7664).

[19] B. Kiumarsi, F. L. Lewis, H. Modares, A. Karimpour, and M.-B. Naghibi-Sistani, "Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, 2014.

[20] C. Qin, H. Zhang, and Y. Luo, "Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming," *International Journal of Control*, vol. 87, no. 5, pp. 1000–1009, 2014.

[21] T. Dierks and S. Jagannathan, "Optimal tracking control of affine nonlinear discrete-time systems with unknown internal dynamics," in *Proc. IEEE Conf. Decis. Control*, 2009, pp. 6750–6755.

[22] Y. Luo and M. Liang, "Approximate optimal tracking control for a class of discrete-time non-affine systems based on gdhp algorithm," in *IWACI Int. Workshop Adv. Comput. Intell.*, 2011, pp. 143–149.

[23] H. Zhang, L. Cui, X. Zhang, and Y. Luo, "Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 2226–2236, 2011.

[24] Q. Wei and D. Liu, "Optimal tracking control scheme for discrete-time nonlinear systems with approximation errors," in *Advances in Neural Networks ISNN 2013*, ser. Lecture Notes in Computer Science, C. Guo, Z.-G. Hou, and Z. Zeng, Eds. Springer Berlin Heidelberg, 2013, vol. 7952, pp. 1–10.

[25] G. Chowdhary, "Concurrent learning adaptive control for convergence without persistencey of excitation," Ph.D. dissertation, Georgia Institute of Technology, December 2010.

[26] G. V. Chowdhary and E. N. Johnson, "Theory and flight-test validation of a concurrent-learning adaptive controller," *J. Guid. Control Dynam.*, vol. 34, no. 2, pp. 592–607, March 2011.

[27] G. Chowdhary, T. Yucelen, M. Mühlegg, and E. N. Johnson, "Concurrent learning adaptive control of linear systems with exponentially convergent bounds," *Int. J. Adapt. Control Signal Process.*, vol. 27, no. 4, pp. 280–301, 2013.

[28] R. Kamalapurkar, P. Walters, and W. E. Dixon, "Concurrent learning-based approximate optimal regulation," in *Proc. IEEE Conf. Decis. Control*, Florence, IT, Dec. 2013, pp. 6256–6261.

[29] D. Kirk, *Optimal Control Theory: An Introduction*. Dover, 2004.

[30] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*, 3rd ed. Wiley, 2012.

[31] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Prentice Hall, 2002.