Approximately Optimal Trajectory Tracking for Continuous Time Nonlinear Systems

Rushikesh Kamalapurkar, Huyen Dinh, Shubhendu Bhasin, and Warren Dixon

Abstract—Approximate dynamic programming has been investigated and used as a tool 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 uniformly ultimately bounded (UUB) tracking of a desired trajectory, while also asymptotically converging to an approximate optimal policy.

I. INTRODUCTION

Reinforcement learning (RL) is a concept that can be used to enable an agent to learn optimal policies from its 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 (also referred to as the critic) and policy improvement (also referred to as the actor). Starting with an estimate of the state cost function and an admissible policy, the critic makes the estimate consistent with the policy and the actor makes the policy greedy with respect to the cost function. These algorithms exploit the fact that the optimal cost function satisfies Bellman’s principle of optimality [1], [2].

The principle of optimality leads to a wide range of algorithms that focus on finding solutions to the Bellman equation (BE) or approximations of the BE. For discrete time systems, BE-based policy evaluation methods do not require a model of the environment, and hence, have been central to the development of RL [2]. Approximate dynamic programming (ADP) consists of algorithms that facilitate the solution of the approximate BE for problems with a continuous state space or an infinite discrete state space by utilizing a function approximation structure to approximate the state cost function [3].

The iterative nature of ADP has made it an attractive method for solving discrete time optimal regulation problems. Within ADP literature, neural networks (NNs) are the most popular tool for cost function approximation [4]–[9]. For continuous time systems, solutions to the optimal regulation problem using ADP have been proposed using discretization of the dynamical system, but these solutions become computationally prohibitive as the dimensionality of the problem increases. Advantage updating was proposed in [10] as a Q-learning method for continuous time systems, but it lacks an accompanying stability analysis.

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 BE [11]. Similar to discrete time ADP, continuous time ADP approaches aim at finding approximate solutions to the HJB equation. Various methods to solve this problem are proposed in [12]–[18] 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 cost function because it is a time invariant function. In tracking problems, the tracking error, and hence the cost function, is a function of the state and an explicit function of time. Approximation techniques like NNs are commonly used in ADP literature for cost function approximation. However, NNs can only approximate functions on compact domains, thus leading to a technical challenge to approximate the cost function for a tracking problem because the infinite horizon nature of the problem implies that time does not lie on a compact set.

For discrete time systems, several approaches have been developed to address the tracking problem. Park et.al. [19] use generalized backpropagation through time to solve a finite horizon tracking problem that involves offline training of NNs. An ADP-based approach is presented in [20] to solve an infinite horizon optimal tracking problem where the desired trajectory is assumed to depend on the system states. A greedy heuristic dynamic programming based algorithm is presented in [21] which uses a system transformation to transform the nonautonomous system into an autonomous system. However, this result lacks an accompanying stability analysis.

ADP-based approaches are presented in [22], [23] for continuous time systems. In both the results, the cost function (i.e. the critic), and the controller (i.e. the actor) presented are time-varying functions of the tracking error. However, as the problem being solved is an infinite horizon optimal control problem, time does not lie on a compact set. NNs can only approximate functions on a compact domain. Thus, it is unclear how a NN that only takes the tracking error as an input can approximate the time varying cost function and controller.

This paper presents an approach to solve the continuous-
time optimal tracking problem online using a system transformation to convert the problem into an optimal regulation problem in such a way that the resulting cost function is a stationary function of the transformed states, and hence, lends itself to approximation using a NN. The desired trajectory is assumed to be the output of a nonlinear dynamical system. A Lyapunov-based analysis is used to prove uniformly ultimately bounded (UUB) tracking and convergence to the approximate optimal control.

II. FORMULATION OF STATIONARY OPTIMAL CONTROL PROBLEM

Consider a class of nonlinear control affine systems

\[ \dot{x} = f + gu, \]

where \( x(t) \in \mathbb{R}^n \) is the state, and \( u(x(t), t) \in \mathcal{U} \subset \mathbb{R}^m \) is the control input. The functions \( f(x(t)) : \chi \to \mathbb{R}^n \) and \( g(x(t)) : \chi \to \mathbb{R}^{m \times m} \) are Lipschitz continuous functions on \( \chi \), where \( f(0) = 0 \), and the solution of the system is unique for any finite initial condition \( x_0 \in \chi \) and control \( u(x(t), t) \in \mathcal{U} \). The control objective is to track a bounded continuously differentiable signal \( \tilde{x}_d(t) : \mathbb{R}^+ \to \chi \). To quantify this objective, a tracking error is defined as \( e(x(t), x_d(t)) : = x(t) - x_d(t) \). The open-loop tracking error dynamics can then be written as

\[ \dot{e} = f + gu - \dot{x}_d. \]  

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

Assumption 1. The function \( g(x(t)) \) is bounded and has full column rank, and the function \( g^+(x(t)) : \chi \to \mathbb{R}^{m \times m} \) defined as \( g^+(x(t)) : = \left( g(x(t))^T g(x(t)) \right)^{-1} g(x(t))^T \) is bounded and Lipschitz continuous.

Note that \( g^+(x(t)) g(x(t)) = I_{m \times m} \), where \( I_{m \times m} \in \mathbb{R}^{m \times m} \) denotes the identity matrix.

Assumption 2. There exists a Lipschitz continuous function \( h_d(x_d(t)) : \chi \to \mathbb{R}^n \) such that \( \tilde{x}_d = h_d(x_d(t)) \) and \( h_d(0) = 0 \).

The steady state control input \( u_d(x_d(t)) \) corresponding to the desired trajectory \( x_d(t) \) is

\[ u_d = g^+_d \left( h_d - f_d \right), \]  

where the signals \( f_d(t) \in \mathbb{R}^n \) and \( g_d^+(t) \in \mathbb{R}^{n \times m} \) are defined as \( g_d^+(t) : = g^+(x_d(t)) \) and \( f_d(t) : = f(x_d(t)) \), respectively. To transform the non-stationary optimal control problem into a stationary optimal control problem, a new concatenated state \( \zeta(t) \in \chi \times \chi \subset \mathbb{R}^{2n} \) is defined as \( \zeta = \left[ e^T, x^T \right]^T \).

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

\[ \dot{\zeta} = F + G\mu, \]  

where the functions \( F(\zeta(t)) : \chi \times \chi \to \mathbb{R}^{2n} \), \( G(\zeta(t)) : \chi \times \chi \to \mathbb{R}^{2n \times m} \), and the policy \( \mu(\zeta(t)) : \chi \times \chi \to \mathbb{R}^m \) are defined as

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

\[ G(\zeta) \triangleq \begin{bmatrix} g(e + x_d) \\ 0 \end{bmatrix}, \quad \mu \triangleq u - u_d. \]

Lipschitz continuity of \( f(x(t)) \) and \( g(x(t)) \), the fact that \( f(0) = 0 \) and Assumption 2 imply that \( F(0) = 0 \) and \( F(\zeta(t)) \) is Lipschitz continuous in the sense that \( \|F\| \leq L_F \|\zeta\| \), where \( L_F \in \mathbb{R} \) is a positive constant. The optimal control problem can now be formulated as the need to design a policy \( \mu(\zeta(t)) \in \Psi(\chi \times \chi) \) that minimizes the cost functional \( V(\zeta(t), \mu(\zeta(t))) : \chi \times \chi \times \Psi \to \mathbb{R}^+ \) defined as

\[ V \triangleq \int_0^\infty \left( r(\zeta(t), \mu(t)) \right) d\rho, \]

subject to the dynamic constraints in (11) where \( \Psi(\chi \times \chi) \) is the set of admissible policies, and \( r(\zeta, \mu) \in \mathbb{R} \) is the local cost defined as

\[ r \triangleq \zeta^T Q \zeta + \mu^T R \mu. \]

In (11), \( R \in \mathbb{R}^{n \times n} \) is a positive definite symmetric matrix of constants, and \( Q \in \mathbb{R}^{2n \times 2n} \) is defined as

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

where \( Q \in \mathbb{R}^{n \times n} \) is a positive definite symmetric matrix of constants with the minimum eigenvalue \( \lambda_{\text{min}} \{Q\} \), and \( 0_{n \times n} \in \mathbb{R}^{n \times n} \) is a matrix of zeros.

III. APPROXIMATE SOLUTION

Assuming that the minimizing policy exists, the HJB equation for the optimal control problem can be written as

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

where \( H^* \left( V^*_\zeta(\zeta(t)), \zeta(t), \mu^*(\zeta(t)) \right) \) is the Hamiltonian, \( V^*_\zeta(\zeta(t)) \triangleq \frac{\partial V^*_\zeta(\zeta(t))}{\partial \zeta} \), and \( \mu^*(\zeta(t)) \) denotes the optimal policy. For the local cost in (11) and the dynamics in (11), the optimal policy can be obtained in closed-form as (11)

\[ \mu^* = \frac{1}{2} R^{-1} G^T V^*_{\zeta^T}, \]

assuming that the optimal cost function \( V^*(\zeta(t)) \) satisfies \( V^*(\zeta(t)) \in C^1 \) and \( V^*(0) = 0 \). The following assumptions are made to facilitate the use of NNs to approximate the optimal policy and the optimal cost function.

Assumption 3. The set \( \chi \) is compact. Based on the subsequent stability analysis, this assumption holds as long as the initial condition \( x(0) \) is bounded. See Remark 2 in the subsequent stability analysis.
Assumption 4. The cost function $V^*(\zeta(t))$ can be represented using a NN with $N$ neurons as

$$V^* = W^T \sigma + \epsilon,$$

where $W \in \mathbb{R}^N$ is the ideal weight matrix bounded above by a known constant $W \in \mathbb{R}$ in the sense that $\|W\|_2 \leq W_\phi$. 

$W(\zeta(t)) : \chi \times \chi \rightarrow \mathbb{R}^N \triangleq \left[ \sigma_1(\zeta(t)) \cdots \sigma_N(\zeta(t)) \right]^T$ is a bounded continuously differentiable nonlinear activation function, and $\epsilon(\zeta(t)) : \chi \times \chi \rightarrow \mathbb{R}$ is the function reconstruction error such that $|\epsilon(\zeta(t))| \leq \tilde{\epsilon}$ and $|\epsilon'(\zeta(t))| \leq \tilde{\epsilon}'$, where $\tilde{\epsilon} \in \mathbb{R}$ and $\tilde{\epsilon}' \in \mathbb{R}$ are positive constants [24, 25].

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

$$\mu^* = -\frac{1}{2} R^{-1} G^T (\sigma^T W + \epsilon^T).$$

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

$$\hat{V} = \hat{W}_c^T \sigma,$$

$$\mu = -\frac{1}{2} R^{-1} G^T \sigma^T \hat{W}_a,$$

where $\hat{W}_c(t) \in \mathbb{R}^N$ and $\hat{W}_a(t) \in \mathbb{R}^N$ are estimates of the ideal neural network weights $W$. The use of two separate sets of weight estimates $\hat{W}_a(t)$ and $\hat{W}_c(t)$ to approximate the same ideal weights $W$ is motivated by the fact that the Bellman error is linear with respect to the cost function weight estimates and nonlinear with respect to the policy weight estimates. Use of a separate set of weight estimates for the cost function facilitates the application of recursive least squares technique for adaptive updates. The controller $u(x(t), t)$ is obtained from (2), (3), and (13) as

$$u = -\frac{1}{2} R^{-1} G^T \sigma^T \hat{W}_a + g_d^T (h_d - f_d).$$

Remark 1. Similar NNs have been previously developed to approximate the cost function and the policy (e.g., 22, 23). However, the tracking error is considered as the only input to the NNs, which implies that the cost function is considered to be a function of the tracking error alone. The presence of a term similar to $u_d(x_d)$ in the cost function definition in the aforementioned results makes the cost function a time-varying function of the tracking error. For an infinite horizon optimal control problem, time does not lie on a compact set, whereas NNs can only approximate functions on a compact domain. Thus, it is unclear how a NN with only the tracking error as the input can approximate the time-varying cost function. In this result, the tracking error and the desired trajectory both serve as inputs to the NN. This makes the controller in (14) fundamentally different, in the sense that a different HJB equation must be solved and its solution, the feedback component $\mu(\zeta)$, is a time-varying function of the tracking error as opposed to a time-invariant function of the tracking error in results such as 22, 23. In particular, this paper addresses the technical obstacles that result from the non-stationary nature of the optimal control problem by including the term $\frac{\partial V}{\partial x_d}$ in the HJB equation.

Using the approximations $\mu(\zeta(t))$ and $\hat{V}(\zeta(t))$ for $\mu^*(\zeta(t))$ and $V^*(\zeta(t))$ in (9), respectively, the approximate Hamiltonian $\hat{H}(\hat{V}(\zeta(t)), \zeta(t), \mu(\zeta(t)))$ can be obtained as

$$\hat{H} = \hat{V}(F + G\mu) + r(\zeta, \mu),$$

where $\hat{V}(\zeta(t)) \triangleq \frac{\partial V(\zeta(t))}{\partial \zeta(t)}$. Using (9), the error between the approximate and the optimal Hamiltonian, called the Bellman Error $\delta \big(\hat{V}(\zeta(t)), \zeta(t), \mu(\zeta(t)) \big) \in \mathbb{R}$, is given in a measurable form by

$$\delta \triangleq \hat{H} - H^* = \hat{V}(F + G\mu) + r(\zeta, \mu).$$

The cost function weights are updated to minimize $\int_0^T \delta^2(\rho) d\rho$ using a normalized least squares update law with an exponential forgetting factor as [26, 27]

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

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

where $\nu, \eta_c \in \mathbb{R}$ are positive adaptation gains, $\omega(\zeta(t), \mu(\zeta(t))) \in \mathbb{R}^N$ is defined as $\omega \triangleq \sigma^T (F + G\mu)$, and $\lambda \in (0, 1)$ is the forgetting factor for the estimation gain matrix $\Gamma(t) \in \mathbb{R}^{N \times N}$. The policy weights are updated to minimize $\delta^2$ using a gradient descent update law as

$$\dot{\hat{W}}_a = \text{proj} \left\{ -\eta_{a1} \frac{\sigma G \Gamma^{-1} \sigma^T}{\sqrt{1 + \omega^T \omega}} \left( \hat{W}_a - \hat{W}_c \right) \delta 

- \eta_{a2} \left( \hat{W}_a - \hat{W}_c \right) \right\},$$

where $\eta_{a1}, \eta_{a2} \in \mathbb{R}$ are positive adaptation gains, and $\text{proj} \{ \cdot \}$ is a smooth projection operator [28]. The use of a forgetting factor ensures that [26, 27]

$$\hat{\Gamma} \leq \Gamma(t) \leq \bar{\Gamma} \hat{\Gamma},$$

where $\bar{\Gamma} \in \mathbb{R}$ are constants such that $0 < \omega < \bar{\omega}$. Using (9), (15), and (16), an unmeasurable form of the BE can be written as

$$\delta = -\hat{W}_c^T \omega + \frac{1}{4} \hat{W}_a^T G \sigma \hat{W}_a + \frac{1}{4} \epsilon' \hat{G} \epsilon' + \frac{1}{2} W^T \sigma^T \epsilon' F,$$

where $\hat{G} \triangleq G \Gamma^{-1} G^T$ and $G \sigma \triangleq \sigma G \Gamma^{-1} G^T \sigma^T$. The weight estimation errors for the cost function and the policy are defined as $\hat{W}_c(t) \triangleq W - \hat{W}_c(t)$ and $\hat{W}_a(t) \triangleq W - \hat{W}_a(t)$, respectively. Using (20), the weight estimation error dynamics for the cost function are

$$\dot{\hat{W}}_c = -\eta_c \psi \psi^T \hat{W}_c + \eta_c \frac{\omega}{1 + \nu \omega^T \hat{\Gamma} \omega} \left( \frac{1}{4} \hat{W}_c^T G \sigma \hat{W}_a 

+ \frac{1}{4} \epsilon' \hat{G} \epsilon' + \frac{1}{2} W^T \sigma^T \epsilon' F \right),$$

where $\psi(\zeta(t), \mu(\zeta(t)), \Gamma(t)) \triangleq \frac{\omega}{\sqrt{1 + \nu \omega^T \hat{\Gamma} \omega}} \in \mathbb{R}^N$ is the regressor vector. Based on (19), the regressor vector can be bounded as

$$\|\psi\| \leq \frac{1}{\sqrt{\nu} \omega}. $$
The dynamics in (21) can be regarded as a perturbed form of the nominal system
\[ \dot{\hat{W}}_c = -\eta c \Gamma \psi \psi^T \hat{W}_c. \]
(23)

It is shown in [26, 27] that this is globally exponentially stable if the regressor vector \( \psi(\zeta(t)) \) is persistently exciting. Given [14, 22, and 24], Theorem 4.14 in [29] can be used to show that there exists a nonautonomous function \( V_c(W_c(t), t) : \mathbb{R}^N \times [0, \infty) \to \mathbb{R} \) and positive constants \( v_{c1}, v_{c2}, v_{c3}, \) and \( v_{c4} \) such that
\[ v_{c1} \| \dot{W}_c \|^2 \leq V_c \leq v_{c2} \| \dot{W}_c \|^2, \]
\[ \frac{\partial V_c}{\partial W_c} (-\eta c \Gamma \psi \psi^T \hat{W}_c) + \frac{\partial V_c}{\partial t} \leq -v_{c1} \| \dot{W}_c \|^2, \]
\[ \frac{\partial V_c}{\partial W_c} \leq v_{c2} \| \dot{W}_c \|. \]
(24)
(25)
(26)

Using Assumptions 1, 2, and 4 and the fact the \( \hat{W}_a(t) \) is bounded by projection, the following bounds are developed to aid the subsequent stability analysis:
\[ \left\| \frac{1}{4} \hat{W}_a^T \hat{G}_a \hat{W}_a + \frac{1}{2} e^T + \frac{1}{2} W^T \sigma^T \right\| \leq \epsilon, \]
\[ \| \hat{G}_a \| \leq \bar{v}_1, \]
\[ \| \dot{W}_a \| \leq \bar{v}_2, \]
\[ \left\| \frac{1}{4} e^T \sigma^T + \frac{1}{2} W^T \sigma^T \right\| \leq \bar{v}_3, \]
(27)

where \( \bar{v}_1, \bar{v}_2, \bar{v}_3, \bar{v}_4 \in \mathbb{R} \) are positive constants.

IV. Stability Analysis

The contribution in the previous section was the development of a transformation that enables the optimal policy and the optimal cost function to be expressed as a stationary function of \( \zeta(t) \). The use of this transformation presents a challenge in the sense that the optimal cost function, which is used as the Lyapunov function for the stability analysis, is not a positive definite function of \( \zeta(t) \) because the matrix \( \mathcal{Q} \) is positive semidefinite. In this section, we address this technical obstacle by exploiting the fact that the stationary Lyapunov function \( V^*(\zeta(t)) : \chi \times \mathbb{R} \to \mathbb{R} = V^*(e(t), t) : \chi \times [0, \infty) \to \mathbb{R} \). Specifically, the use of \( V^*(\zeta(t)) \) facilitates the development of the approximate optimal policy, whereas the equivalent non-stationary form \( V^*(e(t), t) \) can be shown to be a positive definite and decrescent function of the tracking error \( e(t) \). In the following, Lemma 1 and Lemma 2 are used to prove that, written as a nonautonomous function, \( V^*(e(t), t) \) is positive definite, and hence, a Lyapunov candidate. Theorem 1 then states the main result of the paper.

Lemma 1. Let \( D \subseteq \mathbb{R}^n \) contain the origin, and let \( (x, t) \in D \times \mathbb{R}^+ \). Any uniformly bounded, uniformly continuous, positive definite nonautonomous function \( \Xi(x, t) : D \times \mathbb{R}^+ \to \mathbb{R}^+ \) is decrescent in \( D \).

Proof: Uniform boundedness of \( \Xi(x, t) \) implies that \( \forall (x, t) \in D \times \mathbb{R}^+, \sup_{t \in \mathbb{R}^+} \{ \Xi(x, t) \} \) exists and is unique. Let the function \( \alpha(x) : D \to \mathbb{R}^+ \) be defined as
\[ \alpha(x) = \sup_{t \in \mathbb{R}^+} \{ \Xi(x, t) \}, \]
with the property
\[ \Xi(x, t) \leq \alpha(x), \forall t \in \mathbb{R}^+. \]
(28)

Uniform continuity of \( \Xi(x, t) \) implies that \( \forall \epsilon > 0, \exists \delta > 0 \) such that \( \forall (x, t), (y, t) \in D \times \mathbb{R}^+, \)
\[ d_{D \times \mathbb{R}^+}((x, t), (y, t)) < \delta \implies d_{\Xi}((x, t), (y, t)) < \epsilon, \]
(29)

where \( d_M(\cdot, \cdot) \) denotes the standard Euclidean metric on the metric space \( M \). By the definition of \( d_M(\cdot, \cdot) \),
\[ d_{D \times \mathbb{R}^+}((x, t), (y, t)) = d_D(x, y). \]
(30)

From (30) and (31),
\[ d_D(x, y) < \delta \implies |\Xi(x, t) - \Xi(y, t)| < \epsilon. \]
(32)

Given the fact that \( \Xi(x, t) \) is a positive function, (32) implies \( \Xi(x, t) < \Xi(y, t) + \epsilon \) and \( \Xi(y, t) < \Xi(x, t) + \epsilon \) which from (28) implies \( \alpha(x) < \alpha(y) + \epsilon \) and \( \alpha(y) < \alpha(x) + \epsilon \), and hence, from (32),
\[ d_D(x, y) < \delta \implies |\alpha(x) - \alpha(y)| < \epsilon. \]
(33)

Since \( \Xi(x, t) \) is positive definite, (28) can be used to conclude
\[ \alpha(0) = 0. \]
(34)

From (28), (32), and (34), the function \( \Xi(x, t) \) is bounded above by a uniformly continuous positive definite function, and hence, is decrescent in \( D \).

Lemma 2. Let \( B_a \) denote a closed ball around the origin with the radius \( a \in \mathbb{R}^+ \). The optimal cost function \( V^*(e(t), t) : \chi \times \mathbb{R}^+ \to \mathbb{R} \) satisfies the following properties
\[ V^*(e(t), t) \geq v(||e||), \forall t \in \mathbb{R}^+, \]
(35a)
\[ V^*(0, t) = 0, \forall t \in \mathbb{R}^+, \]
(35b)
\[ V^*(e(t), t) \leq \mathcal{V}(||e||), \forall t \in \mathbb{R}^+, \]
(35c)

for all \( e(t) \in B_a \) where \( v(||e||) : [0, a] \to [0, \infty) \) and \( \mathcal{V}(||e||) : [0, a] \to [0, \infty) \) are class \( K \) functions, and \( B_a \subset \chi \).

Proof: a) By the definition of \( V(e(t), t) \) in (6), and \( \mathcal{Q} \) in (5),
\[ V^*(e(t), t) = \int_{t}^{\infty} \left( e^T(\rho) Q e(\rho) + \mu^T(\rho) R \mu^*(\rho) \right) d\rho, \]
\[ \geq V_c(e(t)), \forall t \in \mathbb{R}^+, \]
(36)

where \( V_c(e(t)) \) is the positive definite function. According to Lemma 4.3 in [29] there exists a class \( K \) function \( v(||e||) \) such that \( v(||e||) \leq V_c(e(t), t) \), which along with (36), implies (35a).
b) Since $V^*(e(t), t)$ depends explicitly on time only through the desired trajectory $x_d(t)$, it is sufficient to prove (35) for all $x_d(t) \in \chi$. From (36),

$$V^*(0, x_d(t)) = \int_{t}^{\infty} (\mu^*(\rho) R_\mu(\rho)) \, d\rho,$$  

(37)

where

$$\mu^*(t) = \arg \min_{\mu \in \Psi} \int_{t}^{\infty} (\mu(\rho) R_\mu(\rho)) \, d\rho.$$  

The policy

$$\mu^*(t) = 0, \forall t \in \mathbb{R}^+$$  

(38)

minimizes $V^*(0, x_d(t))$. Furthermore, $V^*(0, x_d(t))$ is the cost incurred when starting with $e(t) = 0$ and following the optimal policy thereafter for any arbitrary desired trajectory $x_d(t)$ (cf. Section 3.7 of [2]). Substituting $x(0) = x_d(0)$, $\mu(0) = 0$ and $e(0) = 0$. Thus, when starting from $e(t) = 0$, the optimal policy in [38] satisfies the dynamic constraints in [4]. Substituting (35) into (37), the optimal cost is $V^*(0, x_d(t)) = 0, \forall x_d(t) \in \chi$ which implies (35).

Using Lemma 1 and the bounds in (23) - (27) the Lyapunov derivative in (41) can be bounded above as

$$\dot{V}_L \leq -K_{e^2} ||e||^2 - K_{\hat{w}_2} ||\hat{w}_2||^2 - \eta_{a_2} ||\hat{w}_a||^2$$

+ $K_e ||e|| + K_{\hat{w}_2} ||\hat{w}_2|| + K,$

(42)

where,

$$K_{e^2} = \lambda_{\min} \{Q\} - \frac{c^2 L_F}{2} \left( \frac{\eta_{a_1}^2}{\sqrt{\nu_p}} + \frac{\eta_{a_2}^2}{\sqrt{\nu_p}} \right),$$

$$K_{\hat{w}_2} = \left( v_{c_1} - \eta_{a_1}^2 \right) - \frac{c^2 L_F}{2} \left( \frac{\eta_{a_1}^2}{\sqrt{\nu_p}} + \frac{\eta_{a_2}^2}{\sqrt{\nu_p}} \right),$$

$$K_c = \eta_{a_1} \eta_{a_2} \frac{c^2 L_F}{\sqrt{\nu_p}},$$

$$K_{\hat{w}_2} = \eta_{a_1} \eta_{a_2} \frac{c^2 L_F}{\sqrt{\nu_p}},$$

$$K = \mu_4 + \eta_{a_1} \eta_{a_2} \frac{c^2 L_F}{\sqrt{\nu_p}}.$$

Provided the sufficient conditions (35) are satisfied, Lemma 4.3 in [29] along with completion of the squares on $||e||$ and $||\dot{w}_a||$ in (42) yields

$$\dot{V}_L \leq -v_1 ||Z|| \cdot \forall ||Z|| \geq \tau_5 > 0,$$

(43)

where $\tau_5 = v_1^{-1} \left( \frac{K_2^2}{2 \lambda_{\min} \{Q\}} + \frac{K_3^2}{2 \lambda_{\min} \{Q\}} + K \right)$, and $v_1 \left( ||Z(t)|| \right) : [0, b) \rightarrow [0, \infty)$ is a class $\mathcal{K}$ function. Using (40), (43), and Theorem 4.18 in [29], $Z(t)$ is UUB.}

**Remark 2.** If $||Z(0)|| \geq \tau_5$, and if the gain conditions in (35) are satisfied then $\dot{V}_L(Z(0), 0) < 0$. Thus, $V_L(Z(t), t)$ is decreasing at $t = 0$. Thus, $Z(t) \in \mathcal{L}_{2\infty}$, and hence, $\zeta(t) \in \mathcal{L}_{2\infty}$ at $t = 0$. Thus all conditions of Theorem 1 are satisfied at $t = 0$. As a result, $V_L(Z(t), t)$ is decreasing at $t = 0$. By induction, $||Z(0)|| \geq \tau_5$, $\Rightarrow V_L(Z(t), t) \leq V_L(Z(0), 0), \forall t \in \mathbb{R}^+$. Thus, from (40), $\|e(t)\| \leq \|Z(t)\| \leq v_1^{-1}(\mathcal{L}(\|Z(0)\|))$ which implies
The problem was solved by transforming the system to convert the tracking problem, that relaxes the persistence of excitation (14) being persistently excited. Furthermore, the control policy in \( \text{(13)} \) requires exact model knowledge. A model-free solution to the tracking problem, that relaxes the persistence of excitation condition, will be the focus of future research.

V. CONCLUSION

An ADP-based approach using the policy evaluation (Critic) and policy improvement (Actor) architecture is presented to approximately solve the infinite horizon optimal tracking problem for control affine nonlinear systems with quadratic cost. The problem was solved by transforming the system to convert the tracking problem that has a non-stationary cost function, into a regulation problem that has a stationary cost function. The UUB tracking and estimation result was established using approximate dynamic programming: Convergence proof, in Proc. IEEE Conf. Decis. Control, Las Vegas, NV, 2002, pp. 943–948. K. Vamvoudakis and F. Lewis, “Online synchronous policy iteration method for optimal control,” in Recent Advances in Intelligent Control Systems. Springer, 2009, pp. 357–374. D. Vrabie and F. Lewis, “Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems,” Neural Networks, vol. 22, no. 3, pp. 237 – 246, 2009. K. Vamvoudakis and F. Lewis, “Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem,” Automatica, vol. 46, pp. 878–888, 2010. 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. Y. Jiang and Z.-P. Jiang, “Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics,” Automatica, vol. 48, no. 10, pp. 2699 – 2704, 2012. Y. M. Park, M. S. Choi, and K. Y. Lee, “An optimal tracking neurocontroller for nonlinear dynamic systems,” IEEE Trans. Neural Networks, vol. 7, no. 5, pp. 1099–1110, 1996. T. Diersk 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. 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. T. Diersk and S. Jagannathan, “Optimal control of affine nonlinear time-varying systems,” in Proc. Am. Control Conf., 2010, pp. 1568–1573. 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. K. Hornik, M. Stinchcombe, and H. White, “Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks,” Neural Networks, vol. 3, no. 5, pp. 551 – 560, 1990. F. L. Lewis, R. Selmic, and J. Campos, Neuro-Fuzzy Control of Industrial Systems with Actuator Nonlinearities. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2003.