Variable Gain Gradient Descent-based Robust Reinforcement Learning for Optimal Tracking Control of Unknown Nonlinear System with Input-Constraints

Amardeep Mishra, Satadal Ghosh

Abstract—In recent times, a variety of Reinforcement Learning (RL) algorithms have been proposed for optimal tracking problem of continuous time nonlinear systems with input constraints. Most of these algorithms are based on the notion of uniform ultimate boundedness (UUB) stability, in which normally higher learning rates are avoided in order to restrict oscillations in state error to smaller values. However, this comes at the cost of higher convergence time of critic neural network weights. This paper addresses that problem by proposing a novel tuning law containing a variable gain gradient descent for critic neural network that can adjust the learning rate based on Hamilton-Jacobi-Bellman (HJB) error and instantaneous rate of variation of Lyapunov function along augmented system trajectories. By allowing high learning rate the proposed variable gain gradient descent tuning law could improve the convergence time of critic neural network weights. Simultaneously, it also results in tighter residual set, on which trajectories of augmented system converge to, leading to smaller oscillations in state error. A tighter bound for UUB stability of the proposed update mechanism is proved. In order to obviate the requirement of nominal dynamics, a neural network based identifier is chosen from existing literature that precedes the RL controller. Numerical studies are then presented to validate the effectiveness of the combined identifier and robust Reinforcement Learning control scheme in controlling a continuous time nonlinear system.

Index Terms—Adaptive Dynamic Programming, Reinforcement Learning, Optimal Tracking, Variable Gain Gradient Descent, Input Constraints

I. INTRODUCTION

OPTIMAL Control as a part of Control Theory seeks to minimize the cost function subjected to system dynamics as constraints. Optimal control can be broadly classified into two major categories: Regularization Problems (wherein states are driven to zero) and Trajectory Tracking Problems (wherein error between actual state and desired state is driven to zero). For a general nonlinear system, optimal control requires the solution of Hamilton Jacobi Bellman (HJB) equation (which is a nonlinear partial differential equation (PDE)) that yields the optimal cost function. The optimal value function is then used to generate optimal control action. The fundamental problem with this approach is that in even simplest of nonlinear cases, the HJB is extremely difficult to solve. For linear systems though, HJB is transformed into Riccati equation.

In order to alleviate the challenge of solving HJB directly, iterative Approximate Dynamic Programming methods were first proposed in the works of Werbos as a method to solve optimal control problem for discrete time (DT) systems in his seminal work [1]- [2]. Neural Network (NN) were used to deal with unknown functions. Sutton and Barto in 1995 proposed ADP for discrete time systems [3]. The usage of two distinct NNs (also known as Actor-Critic structure) to learn the cost function and the control action first appeared in the works of Barto [4] where both the NNs were tuned online. Werbos came up with a third NN to approximate the system dynamics [5]. All of the aforementioned works deal with DT systems and the first few works that appeared for generic nonlinear continuous time system are from [6], [7], [8], [9], [10], [11], [12], [13], [14]. The works mentioned above deal with CT nonlinear optimal regularization problem where states are driven to zero. Vamvoudakis and Lewis (2010) deal with CT nonlinear optimal regularization problem based on an online algorithm, which involved tuning of the critic and the actor weights in a synchronous fashion. The apriori knowledge of the CT nonlinear system dynamics is assumed in both Abu-Khalaf and Lewis (2005) and Vamvoudakis and Lewis (2010). Bhasin et al [14] introduced a novel method of computing control action for regularization problems for CT nonlinear system where partial knowledge of system dynamics exists. Their method demanded the knowledge of control gain matrix. The primary advantage of their methodology was simultaneous tuning of the actor and the critic. Nonetheless, a predefined convex set was required in their work for the implementation of the projection algorithm. This was done to force the NN weights to remain in the set. Identifier NNs were used in the works of Yang et al. [15] to obviate the requirement of knowledge of drift dynamics. This technique could generate the optimal control for nonlinear continuous time systems with unknown structures.

Use of identifiers is not the only method that has been proposed in the literature to deal with uncertain systems while implementing ADP. Integral Reinforcement Learning (IRL), first proposed by Vrabie et al. [16] is one such implementation of RL wherein the system dynamics knowledge is not required in policy evaluation step, i.e., the step involving the evaluation of cost function. However, it too requires the knowledge of

Amardeep Mishra is PhD candidate in the Department of Aerospace Engineering, Indian Institute of Technology, Madras, Tamilnada, 600036 India e-mail: (ae15d405@smail.iitm.ac.in).
Satadal Ghosh is Asst. Professor in Indian Institute of Technology, Madras, e-mail:(satadal@iitm.ac.in)
control dynamics \( g(x) \) in policy iteration step, i.e., the step involving generation of control action. Synchronous tuning of actor-critic NN, based on a novel IRL algorithm was first proposed by Modares et al. [13] in 2014 for continuous time nonlinear systems. A robust ADP algorithm was proposed by Jiang and Jiang [17] to derive the robust control for uncertain nonlinear systems. It was achieved by synthesizing the optimal control solution with infinite horizon cost for original uncertain nonlinear system. However, like most of the ADP based RL methods introduced above, Jiang’s formulation [17] required initial stabilizing controller.

Most of the aforementioned ADP based RL schemes are for regularization problems. ADP based RL for trajectory tracking problem for CT nonlinear systems was initially proposed by Zhang et al. [18] in 2011. Zhang’s method entailed two different controllers viz., the adaptive optimal control (for transient behaviour, i.e to stabilize the tracking error in transience in an optimal manner) and steady state controller (for steady state, i.e, to maintain the tracking error close to zero in steady state). However, the major limitation of his method was that it required the control gain matrix to be invertible in order to implement a steady state controller. This requirement was relaxed in the works of Heydari and Balakrishnan [19] in 2014 when they proposed a single network based critic structure to approximate the cost function. Thereafter, Modares and Lewis [20] proposed an algorithm that was used to analyze the constrained-input optimal tracking problem with a discounted value function for CT nonlinear systems. It should be mentioned that the knowledge of drift dynamics is not required in [13] and [20] (that is, the knowledge of \( f(x) \) presented in [11] is unknown, however the knowledge of control dynamics \( g(x) \) is assumed). Most of the schemes discussed above require an initial stabilizing control to initiate the process of policy iteration.

Finding an initial stabilizing controller to begin the policy improvement is often a very difficult task. Recently, a way to relax the criteria of initial stabilizing control for ADP based RL methods (policy iteration) was proposed by Dierks and Jagannathan [21] as a single online approximator based system. Similarly, Yang et al. [22] proposed an ADP-based RL for robust optimal tracking control of nonlinear systems in 2015. This formulation, did not require an initial stabilizing controller for robust optimal tracking control problem for nonlinear systems. In order to approximate the value function, a single critic NN was utilized in their paper. Tracking control action was generated by critic NN. However, their method requires the knowledge of nominal plant dynamics and does not include the input constraints. It is also noted that their method took a lot of time to achieve convergence of critic NN weights and reduction of oscillation magnitude in state error to a small residual set. These requirements might not be feasible for a lot of practical cases.

To the best of authors knowledge, most of the reinforcement learning schemes that have been recently proposed for tracking problem of nonlinear systems with input constraints do not focus on improving the convergence times of critic NN weights or the oscillations magnitude in state error.

Inspired by [23], [24], this paper addresses these concerns by proposing a robust ADP-based RL tracking controller that is driven by a novel variable gain gradient descent tuning law. In this paper, first, an identifier from existing literature is utilized to identify the nominal plant dynamics before the reinforcement learning controller can be initiated. The identifier design is based on the method proposed by Jin et al. [25] wherein the NN weights converge to a residual set around zero under a persistently excited (PE) condition. Then, Similar to [23] and [24], the critic update law is made up of three terms, the first term is responsible for reducing the HJB error, the second term is responsible for stability, i.e., it comes into effect when the Lyapunov function is growing along the augmented system trajectories and lastly the third term determines the size of the compact UUB set on which the augmented states finally converge to. However, unlike [23] and [24], the learning rate of gradient descent presented in this paper is a function of HJB error and instantaneous rate of variation of Lyapnov function along the augmented system trajectories. This leads to improved tracking performance in terms of faster convergence times of critic neural network weights and smaller oscillation magnitude of state error (error between actual state and desired state).

The salient features of the proposed variable gain gradient descent scheme for RL tracking controller are:

(i) The first term in the weight update law that is responsible for reducing the approximate HJB error, is driven by variable gain gradient descent. In this case, the variable gain is a function of HJB error. So, when HJB error is large, the learning rate gets scaled up proportionally which results in speedier reduction in HJB error, however the learning process is damped, as the HJB error approaches zero.

(ii) The stabilizing term in the update law that is responsible for reducing the rate of variation of Lyapnov along the augmented system trajectories is also driven by a variable gain gradient descent, wherein the variable gain is a function of rate of variation of Lyapnov function along the augmented system trajectories. This term speeds up the learning process when there is a big positive variation of Lyapnov function along the augmented system trajectories. This is done so that the critic weights can be quickly pushed in the direction where the Lyapnov is decreasing along the system trajectories thus resulting in stability.

(iii) The added advantage of variable gain gradient descent is that it provides additional control over the UUB set on which the augmented state converge to. Having the variable gain reduces the size of UUB set as will become clearer in the stability proofs.

The paper is organized as follows, Section II introduces system identification and controller synthesis. It is subdivided into two major subsections, i.e., Section II-A and II-B. The former discusses an existing identifier method whereas latter delves into reinforcement learning for optimal tracking problem of continuous time nonlinear systems with input constraints. It is Section II-B that also proposes a novel weight update law for critic neural network. This section is further subdivided into 3 distinct subsubsections namely, II-B1, II-B2, and II-B3. The first subsubsection, i.e., II-B1 introduces problem formulation and discusses preliminaries. The second subsubsection, i.e., II-B2...
introduces neural network approximation of value function. This subsection also discusses the novel weight update law for tuning of critic NN. Finally II-B3 provides the stability proof of the update law presented in this paper. Towards the end, the paper is concluded by III and IV that discusses results and conclusions respectively.

II. System Identification and Controller Synthesis

As discussed in the Introduction, the robust RL controller designed by Yang et al. [22] required the knowledge of nominal plant dynamics along with the knowledge of bound of uncertainties appearing in the nominal plant dynamics. This might not be a feasible proposition in several cases, where the knowledge of nominal plant might not be available. In order to address that issue, an online adaptive identifier, as presented in [25], is introduced to augment the robust RL controller in Section II-A. Then, the robust ADP-based RL optimal tracking controller is devised leveraging variable gain-based gradient descent in Section II-B.

A. Selection of Online Identifier

This section deals with the design of adaptive online identifier to estimate the drift and control dynamics based on input-output data measurement in real time. In this paper the identifiers structure designed in [25] would be used for this purpose. A basic discussion on this identifier design is given in this section for the sake of completeness, while elaborated proofs could be found in [25].

Note that in this paper the system dynamics is assumed to be affine in control. Let the system dynamics be written in the following form:

\[ \dot{x} = f(x) + g(x)u \]

where \( x \in \mathbb{R}^n, u \in \mathbb{R}^m, f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^n \) and \( g(x) : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m} \). It is assumed that both drift and control dynamics are Lipschitz continuous and can be approximated by NNs [26]. Let there exist optimal NN weights that can accurately approximate both \( f(x) \) and \( g(x) \) as:

\[ f(x) = w_1 \xi_1(x) + \epsilon_f \]
\[ g(x) = w_2 \xi_2(x) + \epsilon_g \]

where, \( w_1 \in \mathbb{R}^{n \times k_{w_1}} \) and \( w_2 \in \mathbb{R}^{n \times k_{w_2}} \) are the unknown optimal weights that can accurately approximate the unknown dynamics and \( \xi_1 \in \mathbb{R}^{k_{w_1}}, \xi_2 \in \mathbb{R}^{k_{w_2} \times m} \) are the regressor vectors and \( \epsilon_f \in \mathbb{R}^n, \epsilon_g \in \mathbb{R}^{n \times m} \) are the approximation errors. According to Weierstrass higher-order approximation theory [27], as the size of regressor vector increase, i.e. \( k_{w_1} \rightarrow \infty, k_{w_2} \rightarrow \infty \), the approximation error goes to zero. Using (1) in (1) one gets.

\[ \dot{x} = w_1 \xi_1(x) + w_2 \xi_2(x)u + \epsilon_f + \epsilon_g \]

Eq. (3) can be re-written in compact form as.

\[ \dot{x} = W_f^T \Phi_f(x,u) + \epsilon_T \]

where, \( W_f = [w_1^T, w_2^T]^T \in \mathbb{R}^{(k_{w_1} + k_{w_2}) \times n} \) is the combined weight matrix for drift and control dynamics and \( \Phi(x,u) = [\xi_1^T(x), u^T \xi_2(x)]^T \in \mathbb{R}^{k_{w_1} + k_{w_2}} \) is the combined regressor. And, \( \epsilon_T \) represents the combined approximation error. There are various online parameter update schemes in literature for ADP-based RL controllers, like the one that minimizes the residual identifier output error (the error between actual state \( x \) and identifier state \( \hat{x} \)) [13]. Similarly, Bhasin et al. [14] proposed modified robust integral of sign of the error (RISE) algorithm for estimating \( W_f \). The identifier state \( \hat{x} \) converges to actual state \( x \), however, the convergence of estimated weights \( \hat{W}_f \) to true weights \( W_f \) is not guaranteed in these research works. In 2016, Jin Na et al. [25] proposed a novel online identification method that ensured convergence to true NN weights based on the weights error, which, in essence, led to a convergence in states.

The present paper leverages the identifier structure proposed by Jin Na et al. [25], in which low-pass-filtered versions of regressor vector \( \Phi_f \) and state vector \( x_f \) were defined as follows.

\[ k\dot{\Phi}_f + \Phi_f = \Phi \]
\[ k\dot{x}_f + x_f = x \]

where \( k > 0 \). In order to derive the update law, two matrices \( \Pi \) and \( K \) were defined as,

\[ \dot{\Pi} + \Pi = \Phi_f \Phi_f^T \]
\[ \dot{K} + lK = \Phi_f \dot{x}_f \]

where, \( l > 0 \). The convergence to true parameters is guaranteed if the update law is given by.

\[ \dot{W}_f = -\Gamma_1 M_f \]

where,

\[ M_1 = \Pi \dot{W}_f - K \]

and \( \Gamma_1 > 0 \) is the learning rate that determines how fast or slow the weight parameter will converge to its true value.

B. Robust ADP-based RL Optimal Tracking Controller

1) Problem Formulation and Preliminaries: The identified dynamics (in the form of \( \hat{f} \) and \( \hat{g} \)) is utilized in this section to synthesize the robust RL tracking controller that is driven by variable gain gradient descent. Thus,

\[ \dot{x} = \hat{f}(x) + \hat{g}(x)u + \Delta f(x) \]
where $\Delta f(x)$ is the deviation between identified plant dynamics and actual plant dynamics, $\hat{f}(x) = \hat{w}_1 \xi_1(x) : \mathbb{R}^n \to \mathbb{R}^n$ and $\hat{g}(x) = \hat{w}_2 \xi_2(x) : \mathbb{R}^n \to \mathbb{R}^{n \times m}$. Note that the identified dynamics, i.e., the $\hat{f}(x) + \hat{g}(u)$ is also Lipschitz continuous.

**Assumption 1.** It is assumed in the following analysis that, $\exists G_M > 0 \ni 0 < \|g(x)\| < G_M, \forall x \in \mathbb{R}^n$. It is also assumed that $\Delta f(x) = 0$, where $d(x) \in \mathbb{R}^m$ is an unknown function bounded by a known function $d_M(x) > 0$. Initial values of both $d$ and $d_M$ are zero.

**Assumption 2.** It is assumed in the following analysis that the commanded trajectory, i.e., the $\dot{x}_d(t) : \mathbb{R} \to \mathbb{R}^n$ is Lipschitz continuous and satisfies $H(0) = 0, \exists \dot{x}_d = H(x_d)$.

**Objective of Control:** The prime objective of the control system is to maintain the error, i.e., $e = x - x_d$, to a small neighborhood of the origin without any prior knowledge of the nominal dynamics and in presence of unknown term $d(x)$. 

**Assumption 3.** The robust RL controller is turned on when the identifier has learned the unknown plant dynamics with a significant accuracy. This implies that $\|f(x) - \hat{f}(x)\| \to 0$ and $\|g(x) - \hat{g}(x)\| \to 0$ and as a consequence $\|x - \hat{x}\| \to 0$. An illustration of which is shown in Fig. 2a and Fig. 2c.

In order to achieve the desired objective, one needs to define an augmented system dynamics that consists of dynamics of errors ($\hat{e}$) and desired states ($\dot{x}_d$). Using Assumptions (1, 2) and (3), tracking error dynamics can be written as:

$$
\dot{e} = \dot{x} - \dot{x}_d
$$

$$
\dot{e} = \hat{f}(x + d) + \hat{g}(x + d)u(t) - H(x_d(t)) + \Delta f(x + d)
$$

(10)

Therefore, the dynamics of augmented system, given as $z = [e^T, \dot{x}_d^T]^T$, can compactly be written as:

$$
z = \dot{F}(z) + \dot{G}(z)u + \Delta F(z)
$$

(11)

where, $u \in \mathbb{R}^m$, $\dot{F} : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ and $\dot{G} : \mathbb{R}^{2n} \to \mathbb{R}^{2n \times m}$ are given by:

$$
\dot{F}(z) = \begin{pmatrix} \hat{f}(e + x_d) - H(x_d) \\ H(x_d) \end{pmatrix}, \ \ \ \dot{G}(z) = \begin{pmatrix} \hat{g}(e + x_d) \\ 0 \end{pmatrix}
$$

(12)

$\Delta F(z) \in \mathbb{R}^{2n}$ and is defined as $\Delta F(z) = \hat{G}d(z)$ with $d(z) \in \mathbb{R}^m$ and $\|d(z)\| \leq d_M(z)$.

One of the prime advantages of creating an augmented system, is that, the controller does not require invertibility of control gain matrix and a single controller comprising of both steady state controller and transient control can be synthesized [20, 28]. Nominal augmented dynamics is given by:

$$
z = \dot{F}(z) + \dot{G}(z)u
$$

(13)

The infinite horizon discounted cost function for (13) is considered as follows [20]:

$$
V(z) = \int_0^\infty e^{-\gamma(t)}[d_M^2 + \bar{u}(z, u)]d\tau
$$

(14)

where, $\bar{u} = z^TQ_1z + C(u)$ is the utility function comprising of augmented state $z$ and the control action $u$. In trajectory tracking problems, $x_d$ contained in $z$ might not go to 0 in steady state and $u$, encapsulates both optimal part and steady state part, hence, infinite horizon cost index comprising of $z, u$ might blow up and become infinite. Hence, in order to make $V$ finite and bounded, discounted cost function of the form (14) is chosen for trajectory tracking problems. Here, $Q_1 \in \mathbb{R}^{2n \times 2n}$ is a positive definite matrix.

Thus, $C(u)$ is defined in this paper as follows [6, 29-31].

$$
C(u) = 2\mu_1 \int_0^\infty (\psi^{-1}(\nu / \mu_1))^T R_1 d\nu
$$

$$
= 2\mu_1 \sum_{i=1}^{\mu_2} \int_0^{\mu_1} (\psi^{-1}(\nu_i / \mu_1))^T R_1 d\nu_i
$$

(15)

where, $R \in \mathbb{R}^{m \times m}$ is a positive definite matrix, $\psi \in \mathbb{R}^m$ is a function possessing following properties

(i) It is odd and monotonically increasing

(ii) It is bounded function ($\psi(\cdot) \leq 1$) that belongs to $C^p (p \geq 1)$. In literature dealing with constrained input, some of the possible candidates for $\psi$ include, tanh, erf, sigmoid. One can clearly observe that $C(u)$ (as shown in Lemma A.1 in appendix) is positive. The discount factor, $0 \leq \gamma$, defines the value of utility in future. A small value of $\gamma$ denotes greedy approach and a large value shows that the agent (controller) cares about future value of rewards. The first term inside the integral caters to any perturbations or uncertainties that might appear in the plant dynamics. It counteracts the deviation between identified dynamics and the true plant dynamics.

Differentiating the (13) along the system trajectories the following can be obtained [13]:

$$
\nabla_z V(\hat{F}(z) + \hat{G}(z)u + \Delta F(z)) - \gamma V(z) + d_M^2 + \bar{u}(z, u)
$$

$$
= \mathcal{H}(z, u, \nabla_z V) = 0
$$

(16)

where, $\mathcal{H}(\cdot)$ represents the Hamiltonian. Let $V^*(z)$ be the optimal cost function that satisfies $\mathcal{H}(\cdot) = 0$ and is given by:

$$
V^*(z) = \min_u \int_0^\infty e^{-\gamma(t)}[d_M^2 + \bar{u}(z, u)]d\tau
$$

(17)

Thus, $\mathcal{H}(\cdot) = 0$ can be re-written in terms of optimal cost as:

$$
\nabla_z V^*(\hat{F}(z) + \hat{G}(z)u + \Delta F(z)) - \gamma V^*(z) + d_M^2 + \bar{u}(z, u) = 0
$$

(18)

Differentiating (18) with respect to $u$, i.e, $\partial \mathcal{H}/\partial u = 0$, closed form of optimal control action $u^*$ as mentioned in [15, 23] is obtained as:

$$
u^* = -u_m \tanh \left( \frac{1}{2u_m} \hat{G}(z)^T \nabla_z V^* \right)
$$

(19)
Substituting (19) in (18) the HJB equation is formulated as:

\[ V^*_z F(z) - 2u^*_m A^T(z) \tanh(A(z)) + d^2_M + z^T Q_1 z + \gamma V^* \]

where \( V^*_z = \nabla_z V^* \), and \( A = (1/2u_m)R^{-1}\hat{G}(z)^T V_z \in \mathbb{R}^m \). The \( C(u) \) or last but one term in left hand side of (20) can be simplified as:

\[
2u_m \int_0^{-u_m} \tanh^{-1}(\nu/\mu_m)^T R d\nu - \gamma V^* = 0
\]

Let \( \hat{z}_v^* \) or last but one term in left hand side of (20) can

The last term in (20) can be re-written as:

\[
\sum_{i=1}^{m} (\tanh^{-1}(u_i^*/u_{m}))^2 = \tanh^{-T}(u_i^*/u_{m}) \tanh^{-1}(u_i^*/u_{m})
\]

Now, utilizing (25) and (27) in (25):

\[
V^* = -d^2_M - z^T Q_1 z + d^2_T(z) d(z) - [d(z) + u_m \tanh^{-1}(u_i^*/u_{m})]^T [d(z) + u_m \tanh^{-1}(u_i^*/u_{m})] + \mathcal{L}_2 + \gamma V^*
\]

where \( \tanh^{-T} \) represents inverse transpose and \( \mathcal{L}_2 = 2u_m \sum_{i=1}^{m} \int_{0}^{\tanh^{-1}(u_i^*/u_{m})} \mu_i \tanh^2(\mu_i) \). Now, using the integral mean value theorem on \( \mathcal{L}_2 \), it can be concluded that there exists a \( \sigma \) satisfying \( 0 \leq \sigma \leq \tanh^{-1}(u_i^*/u_{m}) \) such that

\[
\mathcal{L}_2 = 2u_m \sum_{i=1}^{m} \tanh^{-1}(u_i^*/u_{m}) \sigma_i \tanh^2(\sigma_i)
\]

Eq. (31) is a nonlinear PDE in optimal cost function.

**Assumption 4.** It is assumed that the optimal cost and its gradient is finite and bounded, such that \( ||V^*|| \leq \kappa_V \) and \( ||V^*_z|| \leq \delta_M \).

**Theorem II.1.** For augmented system defined in (11), and its associated discounted cost function defined in (14), the optimal controller described by (19) ensures stability of error dynamics (77) in the sense of uniform ultimate boundedness (UUB) under Assumptions 1-4.

**Proof.** Let \( V^* \) and \( u^* \) represent optimal value function and optimal control function, respectively, such that, \( V^* > 0 \ \forall x \neq 0 \) and \( V^* = 0 \) iff \( x = 0 \). Taking the derivative of \( V^*(z) \) along the system trajectories defined by (11):

\[
\dot{V}^*(z) = V^*_z (F(z) + \hat{G}(z) + V^* \Delta F(z)
\]

From (19) note that \( V^*_z \hat{G}(z) = -2u_m \tanh^{-1}(u_i^*/u_{m}) \). This and Assumption 1 lead to the following:

\[
2u_m \tanh^{-1}(u_i^*/u_{m}) d(z)
\]

Using (20) and (24), Eq. (23) can be rewritten as:

\[
\dot{V}^*(z) = -d^2_M - z^T Q_1 z + L_1 - 2u_m \tanh^{-1}(u_i^*/u_{m}) d(z) + \gamma V^*
\]

where \( L_1 = -2u_m \sum_{i=1}^{m} \int_{0}^{\tanh^{-1}(\mu_i/\mu_m)} \mu_i d\mu_i \). In order to simplify \( L_1 \), let \( \mu_i \equiv \tanh^{-1}(\mu_i/\mu_m) \). Then,

\[
L_1 = -2u_m \sum_{i=1}^{m} \int_{0}^{\tanh^{-1}(u_i^*/u_{m})} \mu_i (1 - \tanh^2(\mu_i)) d\mu_i
\]

Upon further simplification:

\[
\dot{V}^* \leq -\lambda_{min}(Q_1) ||z||^2 - d^2_M ||d_M||^2 - ||d(z)|| + u_m \tanh^{-1}(u_i^*/u_{m})^T [d(z) + u_m \tanh^{-1}(u_i^*/u_{m})] + \frac{1}{2} g^2 M^2 + \frac{1}{2} \gamma^2 + \frac{1}{2} \kappa_V^2
\]

Now in order for \( \dot{V}^* \) to be negative definite, following inequality should always hold:

\[
-\lambda_{min}(Q_1) ||z||^2 + \frac{1}{2} g^2 M^2 + \frac{1}{2} \gamma^2 + \frac{1}{2} \kappa_V^2 < 0
\]

or

\[
||z|| > \sqrt{\frac{\lambda_{min}(Q_1)}{2\kappa_V}} \left( \frac{g^2 M^2 + \gamma^2 + \kappa_V^2}{\lambda_{min}(Q_1)} \right)
\]

This implies that the augmented system will be UUB stable if \( z \) stays out of the ball described by:

\[
\Omega_z = \left\{ x : ||z|| \leq \sqrt{\frac{g^2 M^2 + \gamma^2 + \kappa_V^2}{2\lambda_{min}(Q_1)}} \right\}
\]
This completes the proof.

Note that the set $\Omega_z$ depends on minimum eigenvalue of positive definite matrix $Q_1$ and the discount factor $\gamma$. The set $\Omega_z$ can be made arbitrarily small by choosing the positive definite matrix $Q_1$ such that minimum eigenvalue is large. If the discount factor $\gamma$ is chosen to be 0, then the form of set $\Omega_z$ turn out to be similar to the one mentioned in (19).

2) Approximation of Value function using Critic NN: For applying the optimal controller (19), $V^*$ must be calculated first. This is difficult to achieve because requires solution to (18), which is a nonlinear PDE. In order to by-pass solving the HJB directly, an NN will be utilized to approximate the value function.

Let there exist ideal weight parameter vector $W$ that can accurately approximate the value function as:

$$V^*(z) = W^T \vartheta(z) + \varepsilon$$

(36)

where, $W \in \mathbb{R}^N$ ($N$ being the size of the regressor vector) denotes the ideal weight vector that can closely approximate the value function. And, $\vartheta(z) = [\vartheta_1(z), \vartheta_2(z), ..., \vartheta_N(z)]^T \in \mathbb{R}^N$ represents a set of regressor functions, with following properties such as: $\vartheta_j(z) \in C^1$ and $\vartheta_j(0) = 0$ and $\vartheta_j$s are linearly independent of each other. Substituting (36) in (19),

$$u^*(z) = -u_m \tanh \left( \frac{1}{2u_m} R^{-1} \hat{G}(z) \nabla \vartheta^T W + \varepsilon_{u^*} \right)$$

(37)

where, $\varepsilon_{u^*} = (1/2u_m)R^{-1} \hat{G}(z) \nabla \varepsilon \in \mathbb{R}^m$. Next, substituting (36) in (22), the HJB equation can be written as,

$$W^T \nabla \vartheta \hat{F}(z) - \gamma W^T \vartheta + z^T Q_1 z + d_M^2 + u_m^2 \sum_{i=1}^m \log[1 - \tanh^2 (\tau_1 i + \epsilon_{u^*})] + \nabla \tau^T \hat{F}(z) = 0$$

(38)

where, $\tau_1 = (1/2u_m)R^{-1} \hat{G}(z) \nabla \vartheta^T W = [\tau_{11}, ..., \tau_{1m}]^T \in \mathbb{R}^m$, $\varepsilon_{u^*} = [\epsilon_{u^*}^{e_{u^*}}]$. Upon using Mean value theorem (32), Eq. (38) becomes:

$$W^T \nabla \vartheta \hat{F}(z) - \gamma W^T \vartheta + z^T Q_1 z + d_M^2 + u_m^2 \sum_{i=1}^m \log[1 - \tanh^2 (\tau_1 i)] + \epsilon_{HJB} = 0$$

(39)

where, $\epsilon_{HJB}$ represents the HJB approximation error (5, 13) having a form similar to the one in (23) and is given as,

$$\epsilon_{HJB} = \nabla e^T \hat{F}(z) + \sum_{i=1}^m \frac{2u_m^2}{p_{2i}} \tanh p_{2i}(\tanh^2 p_{2i} - 1)\epsilon_{u^*}$$

(40)

where, $p_{1i} \in \mathbb{R}$ and $p_{2i} \in \mathbb{R}$ considered between $1 - \tanh^2 A_i(z)$ and $1 - \tanh^2 \gamma$. Now, using (39) and mean value theorem, the optimal control can be re-written as:

$$u^* = -u_m \tanh (\tau_1(z)) + \epsilon_{u^*}$$

(41)

where, $\epsilon_{u^*} = -(1/2)(\vartheta_1 - \tanh^2 (\xi) \hat{G}(z) \nabla \varepsilon)$ with $\xi \in \mathbb{R}^m$ considered between $\tau_1$ and $A(z)$ and $\vartheta_1 = [1, 1, ..., 1]^T \in \mathbb{R}^m$. Since ideal weights that can accurately approximate the value function are unknown, their estimates will be used instead as follows.

$$\hat{V}(z) = \hat{W}^T \vartheta(z)$$

(42)

Error in critic weights is given by $\hat{W} = W - \tilde{W}$. Using (42) the estimated optimal control action can be described as:

$$\hat{u}(z) = -u_m \tanh \left( \frac{1}{2u_m} \hat{G}^T(z) \nabla \vartheta^T \hat{W} \right)$$

(43)

From (32) and (42) the HJB approximation error is obtained as follows.

$$H(z, \tilde{W}) = \tilde{W}^T \nabla \vartheta \hat{F}(z) - \gamma \tilde{W}^T \vartheta + z^T Q_1 z + d_M^2 + u_m^2 \sum_{i=1}^m \log[1 - \tanh^2 (\tau_2 i)] \equiv \epsilon(z, \tilde{W})$$

(44)

where, $\epsilon(z, \tilde{W})$ is the HJB error (referred to as $\hat{e}$ in subsequent discussion) and $\tau_2(z) = (1/2u_m)\hat{G}^T(z) \nabla \vartheta^T \tilde{W}^T = [\tau_{21}(z), ..., \tau_{2m}(z)]^T \in \mathbb{R}^m$. Next, from (39) and (44) the HJB error can be expressed in terms of $W$ as (23):

$$\epsilon = -\tilde{W}^T \nabla \vartheta \hat{F}(z) + \sum_{i=1}^m u_m^2 [\Gamma(\tau_{2i}) - \Gamma(\tau_{1i})] - \epsilon_{HJB}$$

(45)

where, $\Gamma(\tau_{1i}) = \log[1 - \tanh^2 \tau_{1i}]$, $i = 1, 2$. It is observed that for all $\tau_{1i}(z) \in \mathbb{R}$, $\Gamma(\tau_{1i})$ can be represented by:

$$\Gamma(\tau_{1i}) = -2\log[1 + \exp(-2\tau_{1i} \text{sgn}(\tau_{1i}))] - 2\tau_{1i} \text{sgn}(\tau_{1i}) + \log(4)$$

(46)

Therefore, using (43) and (47), $\epsilon$ in terms of $\tilde{W}$, is obtained as (23):

$$\hat{e} = 2u_m^2 (\tau_{1i} \text{sgn}(\tau_{1i}) - \tau_{2i} \text{sgn}(\tau_{2i})) - \tilde{W} \nabla \vartheta \hat{F}(z) + u_m^2 \Delta \tau - \epsilon_{HJB}$$

where,

$$\Delta \tau = \sum_{i=1}^m \log \left( \frac{1 + \exp[-2\tau_{1i}(z) \text{sgn}(\tau_{1i}(z))]}{1 + \exp[-2\tau_{2i}(z) \text{sgn}(\tau_{2i}(z))]} \right)$$

(48)

$$\rho(z) = u_m W^T \nabla \vartheta \hat{G}(z) [\text{sgn}(\tau_{1}(z) - \text{sgn}(\tau_{2}(z))] + u_m^2 \Delta \tau - \epsilon_{HJB}$$

(49)

In traditional RL literature for continuous time nonlinear systems, a quadratic cost function of the form, $E = (1/2)e^2$ is chosen, and then gradient descend (GD) is used to drive the parameters $W$ so as to minimize this cost $E$ and thus
to minimize the difference between ideal HJB and the estimated HJB. The following tuning law has been proposed in [9], [14], [15], [18], [20], [33].

$$\dot{W} = -\frac{\alpha}{(1 + \phi^T \phi)} \frac{\partial E}{\partial W} = -\frac{\alpha \phi}{(1 + \phi^T \phi)}^T$$

(50)

where, $\phi = \nabla \theta(\hat{F}(z) + \hat{G}(z)\tilde{u})$, $\alpha > 0$ is the learning rate, and $1 + \phi^T \phi$ is the normalization factor. Then in 2015, Yang et al. [23] proposed a modified version of (50) for optimal regularization problems wherein they used constant learning rate in their gradient descent formulation. Their update mechanism was given as below.

$$\dot{W} = -\alpha \phi \left( Y(x) + d^2_{\hat{f}}(x) + u_m^2 \sum_{i=1}^m \log[1 - \tanh^2(\tau_2(x))] \right)$$

$$+ \frac{\alpha}{2} \Xi(x, \hat{u}) \nabla \theta \hat{G}(x)[I_m - B(\tau_2(x))]\hat{G}^T(x)L_{2x}$$

$$+ \alpha \left( (K_1 \varphi^T - K_2) \hat{W} + u_m \nabla \theta \hat{G}(x)[\tanh(\tau_2(x)) - \text{sgn}(\tau_2(x))] \right) \frac{\varphi^T m_s}{m_s} \hat{W}$$

(51)

where, $x$ is the actual state of the system (not the augmented state), $\alpha > 0$, $\phi = \nabla \theta(\hat{F}(x) + \hat{G}(z)\tilde{u})$, $\varphi = \phi/m_s^2$, $\varphi = \phi/m_s$, $m_s = 1 + \phi \phi^T$, $Y(x) = \hat{W}^T \nabla \hat{F} + x^T Q_1 x$, $B = \text{diag}(\tanh^2(\tau_2(x)))$, $i = 1, 2, ..., m$ and $L_{2x}$ is gradient of Lyapunov.

It can be observed in [23] and [24] that significantly high amount of time is taken by the approximate optimal controller to bring the states [23] or the error in states ($x - x_d$) [24] to a small residual set around origin. In both the above papers, a smaller learning rate was selected to avoid oscillations. However, small values of learning rate results in longer learning phase. In order to address this issue, in this paper, a tuning law with variable learning rate gradient descent is proposed and expressed as follows.

$$\dot{W} = -\alpha \left( |e(z, \hat{W})|^2 \right) \dot{\theta}(e(z, \hat{W}))$$

$$+ \alpha \left( |\Sigma|^2 \right) \Xi(z, \hat{u}) \nabla \theta \hat{G}(z)[I_m - B(\tau_2(z))]\hat{G}^T(z)L_{2x}$$

$$+ \alpha \left( |e(z)|^2 \right) \left( (K_1 \varphi^T - K_2) \hat{W} \right.$$  

$$\left. + u_m \nabla \theta \hat{G}(z)[\tanh(\tau_2(z)) - \text{sgn}(\tau_2(z))] \right) \frac{\varphi^T m_s}{m_s} \hat{W}$$

(52)

where, $\alpha > 0$ is the learning rate, $e(z, \hat{W})$ is the HJB error as mentioned in [24]. The term $\Xi(z, \hat{u})$ is a piece-wise continuous indicator function defined as in [23].

$$\Xi(z, \hat{u}) = \begin{cases} 0, & \text{if } \Sigma(z(t)) < 0 \\ 1, & \text{otherwise} \end{cases} \quad (53)$$

where, $\Sigma(z(t)) = L_{2x}^T \hat{F}(z) + \hat{G}(z)\tilde{u}$ denotes the rate of variation of Lyapunov function along the system trajectories. The constants, $k_2 > 0$ and $q_2 > 0$ provide an augmentation to the controller in the following way. They enable accelerating the learning process, when the HJB error ($e(z, \hat{W})$) is large or when the rate of variation of Lyapunov function along the system trajectories ($\Sigma(z(t))$) is positive with high magnitude, respectively. On the other hand, they dampen the learning process when $e(z, \hat{W})$ and $\Sigma(z(t))$ diminishes to a small quantity. Proper choice of these constants allows for the use of higher value of learning rate without significant oscillations as will be observed in the simulation results presented in Section III. Thus, the controller can bring the error within a small residual set around origin much quickly without any significant oscillations.

In (52), the term $\phi$ is defined as $\phi = \nabla \theta(\hat{F}(z) + \hat{G}(z)\tilde{u}) - |\gamma \theta(z)|$, $\varphi = \phi/m_s^2$, $\varphi = \phi/m_s$, $m_s = 1 + \phi \phi^T$, $B = \text{diag}(\tanh^2(\tau_2(z)))$, $i = 1, 2, ..., m$. Note that the form of (52) is different from (51) that was presented in literature in following ways.

- The $\phi$ in (52) has an additional term $\gamma \theta(z)$ and $e(z, \hat{W})$ has $-\gamma \hat{W}\theta(z)$. Both these terms arise because of the discounted cost function that was used in (52) compared to (51).

- The variable gain in first term of (52) is chosen to be $|\gamma \theta(z)|$ and dampen the reduction process when the HJB error becomes small. The added benefit of variable gain is that it provides control over UUB set as can be seen from (68) that in turn leads to performance improvement in Fig. (52). The error dynamics settle to a much more compact set due to variable gain.

- The second term in (52) is dependent on the variation of Lyapunov function along the system trajectories. It is 0, when the Lyapunov function is strictly decreasing along the system trajectories as shown by the piece-wise indicator function $\Xi(z, \hat{u})$. However it comes into effect when the Lyapunov function is non-decreasing along the system trajectories. It implies that the control action generated at any time step during policy improvement is destabilizing. The second terms starts pulling the critic weights in the direction where the Lyapunov is no more increasing along the system trajectories. In order to fully understand it, let $\Sigma$ denote the variation of Lyapunov function along the system trajectories as $\Sigma = L_{2x} \hat{F}(z) - u_m \hat{G}(z) \tanh(\tau_2(z))$

- If the system is unstable at any time step during policy improvement, then $\Sigma \geq 0$, i.e., the Lyapunov function is non-decreasing along the system trajectories.

- Gradient descent is utilized in [23] to drive the weights in direction such that $\Sigma$ can be reduced and eventually made negative.

$$-\frac{\partial \Sigma}{\partial \hat{W}} = -\frac{\partial [L_{2x} \hat{F}(\hat{z}) - u_m \hat{G}(\hat{z}) \tanh(\tau_2(z))]}{\partial \hat{W}}$$

$$= \alpha \left( \frac{\partial \tau_2(z)}{\partial \hat{W}} \right)^T \frac{\partial [u_m L_{2x} \hat{G}(z) \tanh(\tau_2(z))]}{\partial \hat{W}}$$

$$= \alpha \left( \frac{\partial \tau_2(z)}{\partial \hat{W}} \right)^T \frac{\partial [u_m L_{2x} \hat{G}(z) \tanh(\tau_2(z))]}{\partial \hat{W}}$$

(54)

- In this paper, the constant learning rate ($\alpha$) in second
term of (51) is modified to \((\alpha(|\Sigma|^{q_2}))\) in (52).

\[
-\alpha(|\Sigma|^{q_2}) \frac{\partial \Sigma}{\partial W} = \alpha \left( |\Sigma|^{q_2} \right) \nabla \theta \hat{G}(z) \left[ I_m - \mathcal{B}(\tau_2(z)) \right] \hat{G}^T(z) L_{2z}.
\]  

The significance of this modification lies in the fact that, in the second term of (52), the learning rate is directly contingent upon how far away is \(\Sigma\) from 0. If it is far off, then our second term will try to aggressively bring the \(\Sigma\) to 0. As \(\Sigma\) approaches 0, the learning rate is dampened to reduce oscillations. The constant \(q_2\) is selected based on trial and error.

- The last term in (52) provides control over the UUB sets as mentioned in (23). Proper choice of gains \(K_1\) and \(K_2\) can shrink the UUB ball close to the origin.

Using (48) and (52) the dynamics of error in critic weights \((\dot{W} = W - \hat{W})\) is then given as,

\[
\dot{W} = \alpha(|\dot{\varepsilon}|^{q_2}) \frac{\varphi}{m_s} \left[ -\hat{W}^T \phi + u_m \hat{W}^T \nabla \theta \hat{G}(z) F(z) + \rho(z) \right] - \alpha(\varphi) \frac{1}{2} \hat{\Sigma}(\varphi) \nabla \theta \hat{G}(z) \left[ I_m - \mathcal{B}(\tau_2(z)) \right] \hat{G}(z) L_{2z} + \alpha(|\dot{\varepsilon}|^{q_2}) \nabla \theta \hat{G}(z) F(z) \left\{ \frac{\varphi}{m_s} \hat{W} + (K_2 - K_1 \varphi) T \right\}
\]  

where, \(F(z) = \text{sgn}(\tau_2(z)) - \tanh(\tau_2(z)).\)

3) Proof of Stability of Tuning Law:

**Assumption 5.** Ideal NN weight vector \(W\) is considered to be bounded by a positive constant \(W_M > 0\) such that \(|W| \leq W_M\). There exists positive constants \(b_1\) and \(b_2\) that bound the approximation error and its gradient such that \(|\varepsilon(z)| \leq b_1\) and \(\|\nabla \varepsilon\| \leq b_2\).

**Assumption 6.** Critic regressors are considered to be bounded as well: \(|\vartheta(z)| \leq b_\vartheta\) and \(\|\nabla \vartheta(z)\| \leq b_\vartheta T\).

**Assumption 7.** HJB approximation error and rate of variation of Lyapunov along the system trajectories are assumed to be finite for the time instant when reinforcement learning tracking controller is applied to the system, i.e., \(\dot{\varepsilon} \leq E_{HJB} \leq \infty\) and \(|L_2| \leq Ld < \infty\).

**Assumption 8.** Let \(L_2 \in C^1\) be a continuously differentiable and radially bounded Lyapunov candidate for (13) and satisfies \(L_2 = L_{2z}(\hat{F}(z) + \hat{G} u^*) < 0\). Furthermore, a symmetric and positive definite \(u_m(z) \in \mathbb{R}^{n \times n}\) can be found, such that \(L_{2z}(\hat{F}(z) + \hat{G} u^*) = -L_{2z}^T u_m(z) L_{2z}\), where \(L_{2z}\) is the partial derivative of \(L_2\) wrt \(z\). This assumption is in line with Assumption 4 mentioned in (23).

**Assumption 9.** The approximation error in control (ref. to (41)) is bounded, such that, \(|e_u| \leq b_{e_u}\).

**Theorem II.2.** Let the CT nonlinear augmented system be described by (13) with associated HJB as (22) and approximate optimal control as (19), then the tuning law (52) is stable in the sense of UUB.

**Proof.** Let the Lyapunov candidate be \(L = L_2 + \frac{1}{2}\lambda_1 \hat{W}^T \hat{W}\) (Where \(L_2\) is a positive definite function of augmented state).

Also in subsequent analysis, the terms \(|\dot{\varepsilon}(t)|^{q_2}\) (\(\dot{\varepsilon}(t)\)) is obtained from (44) will be referred to as \(g_1(t)\) and \(|\Sigma(z(t))|^{q_2}\) as \(g(t)\). It is also assumed that \(|g_1(t)| \leq G_1\) and \(|g(t)| \leq G_1\), where \(G_1 < \infty\). Derivative of \(L\) w.r.t. time is obtained as below.

\[
\dot{L} = L_{2z}(\hat{F}(z) + \hat{G} u) + \hat{W} \alpha^{-1} \hat{W} = L_{2z}(\hat{F}(z) - u_m \hat{G}(z) \text{tanh}(\tau_2(z))) + \hat{W} \alpha^{-1} \hat{W}
\]

Utilizing error dynamics of weights, i.e (56) and using the fact that \(\dot{z} = \hat{F}(z) + \hat{G}(z) \dot{u}\), the last term of Lyapunov derivative becomes:

\[
\dot{W} = \alpha^{-1} \dot{W} = \left[ -W^T \phi + u_m W^T \nabla \theta \hat{G}(z) F(z) + \rho(z) \right] g_1 \frac{\varphi}{m_s} \hat{W} - \frac{1}{2} g \hat{\Sigma}(z, \dot{u}) L_{2z}^T \hat{G}(z) \left[ I_m - \mathcal{B}(\tau_2(z)) \right] \hat{G}(z) \nabla \theta W + g_1 u_m W^T \nabla \theta \hat{G}(z) F(z) \left\{ \frac{\varphi}{m_s} \hat{W} + (K_2 - K_1 \varphi) T \right\}
\]

where \(\delta(z) = \rho(z) / m_s\).

Similarly, \(\beta(z) = u_m \nabla \theta \hat{G}(z) F(z) (\varphi / m_s) W\) The last term in (58) can be expressed as:

\[
W^T (K_2 \hat{W} - K_1 \varphi^T \hat{W}) = W^T K_2 W - W^T K_2 \hat{W} - W^T K_1 \varphi^T \hat{W} + W^T K_1 \varphi^T \hat{W}
\]

Let,

\[
J = [\sqrt{g_1} \varphi^T \vartheta, \sqrt{g_1} \hat{W}^T \hat{W}]^T
\]

then (58) can be re-written as:

\[
\dot{W} = -J T M J + J T N - \frac{1}{2} g \hat{\Sigma}(z, \dot{u}) L_{2z}^T \hat{G}(z) \left[ I_m - B(\tau_2(z)) \right] \hat{G}(z) \nabla \theta W
\]

where, \(M\) and \(N\) are defined in the manner similar to (23):

\[
M = \begin{pmatrix}
I & \frac{1}{2} \hat{K}_1^T \\
-K_1^T & K_2
\end{pmatrix}
\]

\[
N = \begin{pmatrix}
\sqrt{g_1} \alpha(z) \\
\sqrt{g_1} (\beta(z) + K_2 \hat{W} - K_1 \varphi^T \hat{W})
\end{pmatrix}
\]

Therefore, the Lyapunov derivative can be rendered in the following inequality:

\[
\dot{\hat{L}} \leq L_{2z}(\hat{F}(z) + \hat{G} u) - \lambda_{min}(M) \|J\|^2 + b_N \|\hat{J}\|^2 - \frac{1}{2} g \hat{\Sigma}(z, \dot{u}) L_{2z}^T \hat{G}(z) \left[ I_m - B(\tau_2(z)) \right] \hat{G}(z) \nabla \theta \hat{W}
\]

where, \(b_N\) is the upper bound of \(N\) which is given by the expression:

\[
\|N\| \leq b_N = \sqrt{g_1} \left( \sqrt{\alpha^2_M + (K_2 \hat{W} - K_1 \varphi^T \hat{W})^2} \right)
\]
where, $\|\alpha(z)\| \leq \alpha_M, \|\beta(z)\| \leq \beta_M, \|\varphi\| \leq \varphi_M, \|W\| \leq W_M$. Based on the variation of Lyapunov function along the system trajectories, which is captured by the value of the piecewise continuous function, $\Xi(z, \dot{u})$, (63) can be explained in two cases:

**Case(i):** When $\Xi(z, \dot{u}) = 0$.

By definition, in this case, $L_{2z}^T \dot{z} < 0$ (where $\dot{z} = \dot{F}(z) + \dot{G}(z)\dot{u}$). Then, considering the density property of real numbers it can be said that there exists $\beta$ such that $0 < \beta \leq \|\dot{z}\|$. Therefore,

$$L_{2z}^T \dot{z} \leq -\|L_{2z}\|\beta < 0 \quad (65)$$

Now, using (65) in (63), $\dot{L}$ is obtained as,

$$\dot{L} \leq L_{2z}^T \dot{z} - \lambda_{\min}(M)\|J\|^2 + b_N\|J\|$$

$$\leq -\|L_{2z}\|\beta + \frac{b_N^2}{\lambda_{\min}(M)} - \lambda_{\min}(M)\left(\|J\|^2\right)$$

$$- \frac{b_N}{\lambda_{\min}(M)} \quad (66)$$

Form the first line of (66), a sufficient condition to ensure negative definiteness of $\dot{L}$ for the above system to be stable can be obtained as $\|J\|^2 > b_N/\lambda_{\min}(M)$. Also, recall from the definition of $\dot{J}$ above (60), the upper bound of $\|J\|$ can be obtained as,

$$\|J\| \leq \sqrt{1 + \|\varphi\|^2} \sqrt{\|G_1\|W} \quad (67)$$

Therefore, using (64) and the bounds on $\|J\|$ obtained above, $\|W\| > \sqrt{\alpha_M + (\beta_M + K_2\varphi_MW_M)^2} \lambda_{\min}(M)\sqrt{1 + \|\varphi\|^2}$ (68).

Again, from the second line of (64), for ensuring stability,

$$-\|L_{2z}\|\beta + \frac{b_N^2}{\lambda_{\min}(M)} < 0 \quad (69)$$

This implies from (64) that

$$\|L_{2z}\| > \frac{G_1^2(\alpha_M + (\beta_M + K_2\varphi_MW_M)^2}{4\beta\lambda_{\min}(M)} \quad (70)$$

This proves that $W$ and $L_{2z}$ are UUB with corresponding sets described by (68) and (70), respectively. Note that since $L_2$ is positive definite function of augmented state and if $L_2 = (1/2)z^T z$, then $L_{2z} = z$ and a tighter UUB bound over $L_{2z}$ implies a tighter UUB bound over $z$. It is to be noted that both $G_1(t)$ and $G(t)$ are time varying upper bound functions, that bound $\dot{e}(t)$ and $\Sigma(t)$, respectively. Furthermore, as the time proceeds and the time varying bound on the HJB error ($G_1(t)$) decrease, it leads to the shrinking of the UUB set for $L_{2z}$ as can be inferred from (70).

This can be observed in simulation studies given in Figs. 35, 36. As one can clearly observe (see Figs. 35, 45) that variable gain gradient descent leads to smaller steady state HJB error bound vis-a-vis a case when variable gain was not utilized.

**Case(ii):** If $\Xi(z, \dot{u}) = 1$.

By definition, in this case, the Lyapunov function is non-decreasing along the system trajectories. The analysis of this case follows similarly as in the previous one, except, the last term in the right hand side (RHS) of (63) also needs to be considered. For that, (19), (63) and Assumption 8 would be utilized.

$$\dot{L} \leq L_{2z}^T \dot{F}(z) - u_mL_{2z}^T \dot{G}(z) \left[\tanh(\tau_2(z)) + \frac{G}{2u_m}[I_m - B(\tau_2(z))]\dot{G}^T \tanh\dot{W}\right] - \lambda_{\min}(M)\|J\|^2 + b_N\|J\|$$

$$\leq -\lambda_{\min}(M)\|J\|^2 \quad (71)$$

Now, adding and subtracting $L_{2z}^T (\dot{G}(z)\dot{u}^*)$ one gets:

$$\dot{L} \leq L_{2z}^T (\dot{F}(z) + \dot{G}u^*) - u_mL_{2z}^T \dot{G}(z) \left[\tanh(\tau_2(z)) + \frac{G}{2u_m}[I_m - B(\tau_2(z))]\dot{G}^T \tanh\dot{W}\right] - \lambda_{\min}(M)\|J\|^2$$

$$+ b_N\|J\| - L_{2z}^T \dot{G}(z)(-u_m \tanh(\tau_1(z)) + \epsilon_u^*)$$

$$\leq -\lambda_{\min}(M)\|J\|^2 \leq -\lambda_{\min}(M)\|J\|^2 + T_M\lambda g \quad (72)$$

$\lambda_{\min}(M) \leq \frac{2\lambda_{\min}(M)}{\lambda_{\min}(M)} = 1 - \tanh^2(\tau_1(z))$. Now, following similar method as in (23), two positive constant numbers $n_1$ and $n_2$ are defined such that $n_1 + n_2 = 1$. In the following analysis, $\|W\|^2 \leq \|J\|^2$ is also utilized.

Therefore the inequality in (73) can be developed as follows:

$$\dot{L} \leq -n_1\lambda_{\min}(M)\|L_{2z}\|^2 + \|L_{2z}\|T_m u_m g \quad (74)$$

$$- n_2\lambda_{\min}(M)\left(\|L_{2z}\|^2 - \frac{\|G/2N_1\|\|\dot{G}^T\|\|\dot{W}\|}{2n_2\lambda_{\min}(M)}\right)^2$$

$$\leq -n_1\lambda_{\min}(M)\|L_{2z}\|^2 + \|L_{2z}\|T_m u_m g \quad (75)$$

$$- n_2\lambda_{\min}(M)\left(\|L_{2z}\|^2 - \frac{\|G/2N_1\|\|\dot{G}^T\|\|\dot{W}\|}{2n_2\lambda_{\min}(M)}\right)^2$$

$$+ \frac{\|G/2N_1\|\|\dot{G}^T\|\|\dot{W}\|}{4n_2\lambda_{\min}(M)}\|J\|^2 + b_N\|J\| - \lambda_{\min}(M)\|J\|^2 \leq -n_1\lambda_{\min}(M)(\|L_{2z}\|^2 - Q^2) + Q_1$$

$$- (\lambda_{\min}(M) - Q)\left(\|J\| - \frac{b_N}{2\lambda_{\min}(M) - Q}\right)^2 \quad (76)$$

where, $Q \leq \frac{\|G/2N_1\|\|\dot{G}^T\|\|\dot{W}\|}{4n_2\lambda_{\min}(M)}$

$$Q_1 \leq \frac{T_m u_m + kg^2 b_{\epsilon_u}}{4n_1\lambda_{\min}(M)} \quad (77)$$

$$Q_2 \leq \frac{T_m u_m + kg^2 b_{\epsilon_u}}{2n_1\lambda_{\min}(M)}$$
In order to facilitate the proof of stability, the last inequality of \((74)\) can be split into two inequalities for \(L_{2z}^T\) and \(J\) (or equivalently \(\bar{W}\)) as follows.

\[
-n_1\lambda_{\min}(\Lambda)(\|L_{2z}^T\| - Q_2)^2 + Q_1 < 0 \tag{76}
\]

or

\[
Q_1 - (\lambda_{\min}(M) - Q)(\|J\| - \frac{b_N}{2(\lambda_{\min}(M) - Q)})^2 < 0 \tag{77}
\]

Therefore, (76) leads to,

\[
\|L_{2z}^T\| > Q_2 + \sqrt{\frac{Q_1}{n_1\lambda_{\min}(\Lambda)}} \tag{78}
\]

Augmented system will be stable if the \(L_{2z}\) stays outside the ball described by,

\[
\Omega_{L_{2z}} = \left\{ L_{2z} : \|L_{2z}\| \leq \frac{T_m g_{M} u_m + kg_{2z} b_{z}}{2n_1\lambda_{\min}(\Lambda)} + \frac{(T_m g_{M} u_m + kg_{2z} b_{z})^2}{(2n_1\lambda_{\min}(\Lambda))^2} + \frac{G_2(\alpha_{\min}^2 + (\beta M + K_2 W M - K_1 \varphi M W)^2)}{4(n_1\lambda_{\min}(\Lambda))(\lambda_{\min}(M) - Q)} \right\} \tag{79}
\]

During the non-decreasing phase of the Lyapunov function, in this on-policy (policy iteration) algorithm, the size of UUB set for \(L_{2z}\) decreases as the time varying bound over HJB error decrease as can be inferred from (79) as it contains \(G_1\) in the square root term. However, without \(\bar{W}\), where the time varying bound over the HJB (\(G_1\)) appears directly, in (79) the same appears inside a square root. Hence, the bound in the case(ii) as described by (78) is larger compared to (70). Similarly, for \(J\), (77) can be re-written as,

\[
\|J\| > \frac{b_N}{2(\lambda_{\min}(M) - Q)} + \frac{Q_1}{\lambda_{\min}(M) - Q} \tag{80}
\]

However, if \(\lambda_{\min}(\Lambda)\) is large enough, then,

\[
Q_1 \approx \frac{b_N^2}{4(\lambda_{\min}(M) - Q)} \tag{81}
\]

Using (81), Eq. (80) can be re-written as,

\[
\|J\| > \frac{b_N}{\lambda_{\min}(M) - Q} \tag{82}
\]

From (67) and (82) it can be inferred that,

\[
\frac{b_N}{\lambda_{\min}(M) - Q} < \left(\sqrt{1 + \|\varphi\|^2}\right)\sqrt{G_1}\tilde{W} \tag{83}
\]

Utilizing (64) for \(b_N\), the \(||\tilde{W}||\) is found to be UUB with the set defined by,

\[
||\tilde{W}|| > \sqrt{\frac{\alpha_{\min}^2 + (\beta M + K_2 W M - K_1 \varphi M W)^2}{\lambda_{\min}(M) - Q}} \sqrt{1 + ||\varphi||^2} \tag{84}
\]

It can be observed from (65) and (84) that the UUB bound over \(\tilde{W}\) in the case when Lyapunov is strictly decreasing along augmented system trajectories is tighter than in the case when Lyapunov is non-decreasing. Similar to the Case(i), it is now shown that \(L_{2z}^T\) is UUB with the set defined by (78) and \(\bar{W}\) is UUB with set defined by (84).

It will now be proved that the closed loop system i.e. the augmented state vector \((z)\) is stable in the sense of UUB during the learning phase of critic network. To this end the Lyapunov function candidate is considered as mentioned in (17). Taking its time derivative along the system trajectory during the learning phase of critic network, i.e., \(\dot{z} = \tilde{F}(z) + G(z)u\)

\[
\dot{V} = V_{z}^* \tilde{F}(z) + V_{z}^* \dot{G}(z)u \tag{85}
\]

where, from (19) and (20),

\[
V_{z}^* \tilde{F}(z) = -V_{z}^* \dot{G}(z)u - (\frac{b_N}{2} + \gamma V^*) \sum_{z=1}^{T} Q_1 z + \gamma V^* - 2u_m \int_{0}^{u_m} \tanh^{-}(\frac{\nu}{u_m}) d\nu \tag{86}
\]

Now, adding and subtracting \(V^* \tilde{F}(z)\) to both sides of the above equation one obtains:

\[
V_{z}^* \tilde{F}(z) + V_{z}^* \dot{G}(z)u = (\tilde{V}^* - u^*) - \sum_{z=1}^{T} Q_1 z + \gamma V^* - 2u_m \int_{0}^{u_m} \tanh^{-}(\frac{\nu}{u_m}) d\nu \tag{87}
\]

It can be seen that (85) and (87) are same. Now, the last term in (87) is positive definite as shown in Lemma A.2 Therefore, from (85) and (87) and using (66), an inequality of \(\tilde{V}^*\) is obtained as,

\[
\tilde{V}^* \leq -\sum_{z=1}^{T} Q_1 z + \gamma \sum_{z=1}^{T} (||\tilde{G}(z)||||\tilde{u} - u^*|| + \|\tilde{G}(z)||||\tilde{u} - u^*|| + \gamma V^*) \tag{88}
\]

Now, using Young’s Inequality on \(\gamma V^*\) along with Assumptions 1 2 4 8 and Lemma A.1

\[
\tilde{V}^* \leq -\lambda_{\min}(Q_1) \sum_{z=1}^{T} Q_1 z + \|\tilde{G}(z)||||\tilde{u} - u^*|| + \(\sum_{z=1}^{T} (\beta M + K_2 W M - K_1 \varphi M W)^2) \sqrt{1 + ||\varphi||^2} + \frac{b_N}{2} \tag{89}
\]

In order to have \(\tilde{V}^* < 0\) for the augmented system to be stable during the learning phase, (89) implies,

\[
\|z\| > \sqrt{\frac{(g_{M} u_m)(b_{z} W M + b_{z}) + \gamma^2 + \frac{k_{z}^2}{2}}{\lambda_{\min}(Q_1)}} \tag{90}
\]

Hence, the augmented state will be UUB stable during the online learning phase, if the augmented state remains outside of the ball described by:

\[
\Omega_z = \left\{ z : \|z\| \leq \sqrt{\frac{(g_{M} u_m)(b_{z} W M + b_{z}) + \gamma^2 + \frac{k_{z}^2}{2}}{\lambda_{\min}(Q_1)}} \right\} \tag{91}
\]
This ball can be made arbitrarily small by selecting a high \( \lambda_{\text{min}}(Q_1) \) and low \( \gamma \). This completes the stability proof of the update mechanism.

The expression for \( \Omega_2 \) derived in this paper in (91) is similar to the expression for bound over states in [23]. The difference comes from the discount factor \( \gamma \).

### III. Results and Simulation

Consider a continuous time nonlinear system \( \dot{x} = f(x) + g(x)u \) as mentioned in [24].

\[
\begin{align*}
    f &= \begin{pmatrix} -x_1 + x_2 \\
    -(x_1 + 1)x_2 - 49x_1 + 1.5(\cos(x_1))^3 \sin(x_2) \end{pmatrix} \\
    g &= \begin{pmatrix} 1 \\
    0 \end{pmatrix}
\end{align*}
\]

The parameters for online identifier are, \( k = .001, I = 3, \Gamma_1 = 170 \) (See Eqs. 5, 6, 7). As it can be seen, the identifier is able to quickly identify the dynamics accurately (ref. to Fig. 3). This continuous time nonlinear system is required to track a desired reference command. The augmented state vector \( z = [z_1, z_2, x_1, x_2, \cos(x_1)^3 \sin(x_2), u]^T \).

\[
\begin{align*}
    (\dot{x}_d1) &= (x_{d2}) \\
    (x_{d2}) &= 49x_{d1}
\end{align*}
\]

The Lyapunov function \( L_2 \) is selected as \( L_2 = 1/2z^Tz \). Also, \( R = 1 \), and \( Q_1 \) (refer to Eqs. 14, 15) is selected as,

\[
Q_1 = \begin{pmatrix} I_2 & 0 \\
0 & 0_{2 \times 2} \end{pmatrix}
\]

where, \( I_2 = \text{diag}(10, 0) \). Regressor vector for critic network is selected as [24].

\[
\theta(z) = [z_1^2, z_2^2, z_3^2, z_4^2, z_1, z_2, z_3, z_4, z_2, z_3, z_2 z_3, z_2 z_4, z_3 z_4]^T
\]

Discount factor \( \gamma = \frac{0.1}{1} \). Initial state of the system is chosen to be, \( x(0) = [1.5, 1.5]^T \). Critical weights are initialized to 0, i.e., \( W(0) = 0 \). A dithering noise of the form \( n(t) = 2e^{-0.0009t} (\sin(11.9t)^2 \cos(19.5t) + \sin(2.2t)^2 \cos(5.8t) + \sin(1.2t)^2 \cos(9.5t) + \sin(2.4t)^5) \) is added to maintain the persistent excitation (PE) condition.

Now, a comparative study of the variable gain gradient descent method presented in this paper w.r.t. constant gradient descent will be carried out. All other system parameters are kept same. Learning rate is selected to be \( \alpha = 35.9 \) for both. In order to validate the performance of the controller developed in this paper, two input bounds were selected, i.e \( u_m = 1.8 \) and \( u_m = 9 \). Figs. 5 and 6 correspond to the case with input bound of 1.8, while Figs. 4 and 3 correspond to the input bound of 9. Constants used in variable gain gradient descent are \( k_2 = 1.33, q_2 = .90 \) for \( u_m = 9 \) and \( k_2 = .001, q_2 = .01 \) for \( u_m = 1.8 \) (see Eq. 52).

It can be clearly inferred from Fig. 3 that, the variable gain gradient descent based tuning law proposed in this paper is able to successfully bring the state error to a much tighter residual set than constant learning rate.

Comparing Figs. 4a and 5a, one observes that the application of variable gain gradient descent leads to faster and efficient learning of critic NN weights. In Fig. 5a, learning process was allowed to continue for 250s, whereas in Fig. 4a, they were not able to converge within 250s. Control effort in both these cases was found to be well within the saturation limit for most of the time, except at the beginning because of the dithering noise that is added to the reinforcement learning control signal to maintain PE condition. It is noted from Figs. 4d and 5d that the actual control commands were within the band of \([-2, 2]\) for most of the time. Therefore, a more tight saturation limit \( u_m = 1.8 < 2 \) would be considered next.

Under the restricted saturation limit (\( u_m = 1.8 \)) the performance of both (with and without variable gain gradient descent) is expected to degrade, which can be seen from Figs. 5 (without variable gain gradient descent) and 6 (with variable gain gradient descent).

However, even in tight saturation limit on input, the variable gain gradient descent performs better than the constant learning rate gradient descent as can be inferred from Figs. 5c and 6c. The variable gain gradient descent allows for a tighter bound over state error vis-a-vis the case when constant learning rate was used in gradient descent. In Fig. 5c and 6c, the controller is able to limit the control effort to 1.8. However during the learning phase, i.e., for the time, the dithering noise is added to the reinforcement controller, the control efforts exceed the saturation limit. This is due to the fact that the dithering noise is chosen to be a slow decaying sinusoids, its magnitude is pretty significant during the initial stages of the learning process. The input saturation has an adverse effect on the tracking performance when variable gain gradient descent was not used as can be seen from Figs. 5b and 6b. This adverse effect has been alleviated to some extent using the variable gain gradient descent as can be inferred from Figs. 5d and 6d. In Figs. 5d and 6d, it can be clearly observed that both the HJB error and the state error are appreciably lesser than 5b and 6b. In [24], a smaller value of learning rate was selected to train the critic network online, however, following their formulation, it takes a lot more time for the controller to bring the oscillation magnitude of state error down to a small bound, which can be clearly seen in Fig. 1 in [24]. Also in [24], the control effort during the initial phases touched \([-20, 20]\). As it can be inferred from Figs. 5c and 6c, a high constant learning rate leads to larger oscillation bound on states error compared to the case when variable gain gradient descent (see Fig. 5c) was utilized. Thus, the prime advantage of variable gain gradient descent-based critic update law is the ability to select reasonably high learning rates without large magnitude of oscillation in error dynamics.

### IV. Conclusion

The paper presents two stage tracking control scheme comprising of identification followed by reinforcement learning for continuous time nonlinear system with input constraints.
(a) Identifier Weights  
(b) Difference between identified and actual drift dynamics  
(c) Difference between identified and actual control coupling dynamics

Figure 2: Identifier

(a) Critic Weights with variable gain gradient descent  
(b) HJB error with variable gain gradient descent  
(c) State error with variable gain gradient descent  
(d) Control action with variable gain gradient descent

Figure 3: Critic NN, HJB error, State error and Control profiles with variable gain gradient descent (for $u_{tm} = 9$)

(a) Critic Weights without variable gain gradient descent  
(b) HJB error without variable gain gradient descent  
(c) State error without variable gain gradient descent  
(d) Control action without variable gain gradient descent

Figure 4: Critic NN, HJB error, State error and Control profiles without variable gain gradient descent (for $u_{tm} = 9$)

(a) Critic Weights without variable gain gradient descent  
(b) HJB error without variable gain gradient descent  
(c) State error without variable gain gradient descent  
(d) Control action without variable gain gradient descent

Figure 5: Critic NN, HJB error, State error and Control profiles without variable gain gradient descent (for $u_{tm} = 1.8$)

(a) Critic Weights with variable gain gradient descent  
(b) HJB error with variable gain gradient descent  
(c) State error with variable gain gradient descent  
(d) Control action with variable gain gradient descent

Figure 6: Critic NN, HJB error, State error and Control profiles with variable gain gradient descent (for $u_{tm} = 1.8$)
In order to identify the unknown dynamics, a neural network based identifier has been chosen from the existing literature. Subsequently, an infinite horizon discounted cost function has been considered to obtain optimal control action for the augmented system containing error dynamics and desired dynamics. A critic neural network is brought to bear to approximate the value function which is also the solution of the tracking HJB equation. The critic neural network has been tuned online using a novel variable gain gradient descent scheme proposed in this paper. The hallmarks of this update law stems from the fact that it can adjust its learning rate based on the HJB error and instantaneous rate of variation of Lyapunov function along the system trajectories. The tuning law speeds up the learning process if the HJB error is large and it slows it down as the HJB error becomes small. Similarly, the second term of the parameter update law accelerates the learning rate when the rate of variation of Lyapunov is a high positive number. This has the effect of faster tuning towards stable region when the Lyapunov is increasing along the system trajectories. Thus, the parameter update law presented in this paper leads to smaller convergence times of critic NN weights and tighter residual set over which the augmented system trajectories converge to. One limitation of this control scheme arises from the usage of identifier to learn the nominal plant dynamics first, i.e., identifier is run for some amount of time to learn the system dynamics online, subsequently the identifier weights are frozen before critic tuning can be initiated. The update law presented in this paper forms the basis of future scope of research using which this limitation will be addressed.

**APPENDIX A**

**LEMMA**

**Lemma A.1.** Following vector inequality holds true:

\[ \| \tanh (\tau_1(z)) - \tanh (\tau_2(z)) \| \leq T_m \leq 2\sqrt{m} \]  

where \( T_m = \sqrt{\sum_{i=1}^{m} \min(|\tau_{i1} - \tau_{i2}|, 2, 4)} \), \( \tau_1(z) \) and \( \tau_2(z) \) both belong in \( \mathbb{R}^m \), therefore, \( \tanh (\tau_1(z)) \in \mathbb{R}^m \), \( i = 1, 2 \).

**Proof:** Since, \( \tanh(\cdot) \) is 1-Lipschitz, one can write,

\[ |\tanh (\tau_{i1}) - \tanh (\tau_{i2})| \leq |\tau_{i1} - \tau_{i2}| \]  

Therefore using the above inequality and the fact that, \( -1 \leq \tanh (\cdot) \leq 1 \)

\[ \| \tanh (\tau_1(z)) - \tanh (\tau_2(z)) \|^2 \leq \sum_{i=1}^{m} |\tanh (\tau_{i1}) - \tanh (\tau_{i2})|^2 \leq \sum_{i=1}^{m} \min(|\tau_{i1} - \tau_{i2}|, 2)^2 \leq \sum_{i=1}^{m} \min(|\tau_{i1} - \tau_{i2}|, 4)^2 \]  

One can also see, using the absolute upper bound of \( \tanh(\cdot) \),

\[ \sum_{i=1}^{m} \min(|\tau_{i1} - \tau_{i2}|, 4) \leq 2\sqrt{m} \]

Which implies,

\[ \| \tanh (\tau_1(z)) - \tanh (\tau_2(z)) \| \leq T_m \leq 2\sqrt{m} \]

**Lemma A.2.** Following inequality holds true:

\[ C(u_i) = 2u_m \int_0^{u_i} \psi^{-1}(\frac{\nu}{u_m})R_i d\nu \geq 0 \]

if \( \psi^{-1} \) is monotonic odd and increasing and \( R_i > 0 \). Where \( u_i \in \mathbb{R}, i = 1, 2, ..., m \)

**Proof:** If \( \psi^{-1} \) is monotonic odd and increasing, then,

\[ \left( \frac{\nu}{u_m} \right)^{\psi^{-1}(\frac{\nu}{u_m})} \geq 0 \]

or

\[ \nu \psi^{-1}(\frac{\nu}{u_m}) \geq 0 \]

where \( \nu \in \mathbb{R} \) and \( u_m > 0 \). Let \( \theta = 1/u_m \). In order to prove that, \( 2u_m \int_0^{u_i} \psi^{-1}(\nu/u_m)R_i d\nu \geq 0 \), it is enough to prove that, \( \int_0^{u_i} \psi^{-1}(\nu\theta) d\nu \geq 0 \). In order to prove this inequality, a variable, \( K \in [0, \theta] \) is assumed. Therefore,

\[ \int_0^{u_i} \psi^{-1}(\nu\theta) d\nu = \frac{1}{\theta} \int_0^{u_i\theta} \psi^{-1}(l)dl \]

where \( l = \nu\theta \). Similarly,

\[ \frac{1}{\theta} \int_0^{u_i\theta} \psi^{-1}(l)dl = \frac{1}{\theta} \int_0^{u_i}(\nu\theta)K u_i dK \]

by utilizing \( l = u_i K \).

Since, \( \psi^{-1}(u_i(K)u_i \geq 0 \), which implies,

\[ \frac{1}{\theta} \int_0^{u_i} \psi^{-1}(u_i(K)u_i \geq 0 \]

**REFERENCES**

[1] P. Werbos, “Beyond regression:” new tools for prediction and analysis in the behavioral sciences,” Ph. D. dissertation, Harvard University, 1974.

[2] ——, “Advanced forecasting methods for global crisis warning and models of intelligence,” General System Yearbook, pp. 25–38, 1977.

[3] A. G. Barto, “1” 1 adaptive critics and the basal ganglia,” Models of information processing in the basal ganglia, p. 215, 1995.

[4] A. G. Barto, R. S. Sutton, and C. W. Anderson, “Neuronlike adaptive elements that can solve difficult learning control problems,” IEEE transactions on systems, man, and cybernetics, no. 5, pp. 834–846, 1983.

[5] P. J. Werbos, “Neural networks for control and system identification,” in Proceedings of the 28th IEEE Conference on Decision and Control, IEEE, 1989, pp. 260–265.

[6] M. Abu-Khalaf and F. L. Lewis, “Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach,” Automatica, vol. 41, no. 5, pp. 779–791, 2005.

[7] W.-S. Lin, “Optimality and convergence of adaptive optimal control by reinforcement synthesis,” Automatica, vol. 47, no. 5, pp. 1047–1052, 2011.

[8] D. Liu, X. Yang, and H. Li, “Adaptive optimal control for a class of continuous-time affine nonlinear systems with unknown internal dynamics,” Neural Computing and Applications, vol. 23, no. 7–8, pp. 1843–1830, 2013.

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

[10] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, “Adaptive dynamic programming,” IEEE Transactions on Systems, Man, and Cybernetics, Part C, vol. 32, no. 2, pp. 140–153, 2002.

[11] X. Yang, D. Liu, and Q. Wei, “Online approximate optimal control for affine non-linear systems with unknown internal dynamics using adaptive dynamic programming,” IET Control Theory & Applications, vol. 8, no. 16, pp. 1676–1688, 2014.
[12] D. Zhao and Y. Zhu, “Meca near-optimal online reinforcement learning algorithm for continuous deterministic systems,” IEEE transactions on neural networks and learning systems, vol. 26, no. 2, pp. 346–356, 2014.

[13] 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.

[14] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, “A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems,” Automatica, vol. 49, no. 1, pp. 82–92, 2013.

[15] X. Yang, D. Liu, and D. Wang, “Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints,” International Journal of Control, vol. 87, no. 3, pp. 553–566, 2014.

[16] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, “Adaptive optimal control for continuous-time linear systems based on policy iteration,” Automatica, vol. 45, no. 2, pp. 477–484, 2009.

[17] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, “A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems,” Automatica, vol. 49, no. 1, pp. 82–92, 2013.

[18] X. Yang, D. Liu, and Q. Wei, “Robust tracking control of uncertain nonlinear systems using adaptive dynamic programming,” in Proceedings of the 2010 American Control Conference. IEEE, 2010, pp. 1568–1573.

[19] S. Arik, T. Huang, W. Lai, and Q. Liu, “Neural information processing: 22nd international conference,” in ICONIP, 2015, pp. 9–12.

[20] B. A. Finlayson, The method of weighted residuals and variational principles. SIAM, 2013, vol. 73.

[21] 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, vol. 50, no. 4, pp. 1167–1175, 2014.