This paper is a preprint of a paper submitted to IET Control Theory and Application. If accepted, the copy of record will be available at the IET Digital Library

IET Journals
The Institution of Engineering and Technology

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

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Amardeep Mishra[1] ✉, *Satadal Ghosh*[2]

[1] *Student, Department of Aerospace Engineering, IIT Madras, Chennai 600036, India, ae15d405@smail.iitm.ac.in*
[2] *Faculty, Department of Aerospace Engineering, IIT Madras, Chennai 600036, India, satadal@iitm.ac.in*
✉ *E-mail: ae15d405@smail.iitm.ac.in*

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) approximation error. 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. Numerical studies are then presented to validate the robust Reinforcement Learning control scheme in controlling a continuous time nonlinear system.

1 Introduction

Optimal Control as a part of Control Theory seeks to minimize the cost function subjected to system dynamics as constraints. In nature, oragnisms act on the environment and observe the resulting reward. Over time, the actions on the environment are tweaked to improve the likelihood of the reward. This process is sometimes referred to as reinforcement learning and it captures the notion of optimality [1]. Adaptive dynamic programming (ADP) refers to the mathematical formulation and solution of the reinforcement learning problem [1]. It is a practical way of implementing optimal control. The optimal control problem can be broadly classified into two major categories: regulation 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 equation is extremely difficult to solve. For linear systems though, HJB equation is transformed into Riccati equation.

In order to alleviate the challenge of solving HJB equation 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 [2]- [3]. Neural Network (NN) were used to deal with unknown functions. Sutton and Barto in 1995 proposed ADP for discrete time systems [4]. 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 [5] where both the NNs were tuned online. Werbos came up with a third NN to approximate the system dynamics [6]. All of the aforementioned works deal with DT systems and the first few works that appeared for generic nonlinear continuous time system are from [7], [8], [9], [10], [11] [12] [13] [14] [15]. The works mentioned above deal with continuous time (CT) nonlinear optimal regulation problem where states are driven to zero. Vamvoudakis and Lewis (2010) deal with CT nonlinear optimal regulation 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 [15] introduced a novel method of computing control action for regulation 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. [16] 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 unknown systems while implementing ADP. Integral Reinforcement Learning (IRL), first proposed by Vrabie et al. [17] 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 control dynamics 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. [14] in 2014 for continuous time nonlinear systems. A robust ADP algorithm was proposed by Jiang and Jiang [18] 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 methods introduced above, Jiang's formulation [18] required initial stabilizing controller.

Most of the aforementioned ADP schemes are for regulation problems. ADP formulations for optimal tracking control problem (OTCP) for CT nonlinear systems was initially proposed by Zhang et al. [19] 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 [20] in 2014 when they proposed a single network based critic structure to approximate the cost function. Thereafter, [21] 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 [14] and [21], however the knowledge of control dynamics 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 [22] as a single online approximator based system. Similarly, Yang el al. [23] proposed an ADP algorithm 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.

Inspired by [23] and [24], this paper addresses these concerns by proposing a ADP-based robust optimal tracking controller that is driven by a novel variable gain gradient descent tuning law. 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. 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) To the best of authors' knowledge, this is the first time when variable learning rate is leveraged in gradient descent to tune critic NN weights to solve robust optimal tracking problem for continuous time nonlinear systems with actuator constraints. The first term in the weight update law responsible for reducing the approximate HJB error is driven by variable learning rate gradient descent where, the variable learning rate 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 dampened, as the HJB error approaches zero.
(ii) Further, variable gain gradient descent leads to tighter residual set for critic NN weights thus resulting in approximated optimal controllers that are closer to ideal optimal controller. This in turn leads to improved tracking performance.

The rest of the paper is organized as follows, Section 2 introduces robust optimal tracking controller and its preliminaries. This section is followed by Section 3 that utilizes the concept of RL to solve optimal trajectory tracking problem for continuous time nonlinear system with actuator constraints. It is divided into two subsections (subsection 3.1 and 3.2) that delve into value function approximation using critic NN and existing parameter update law respectively. It is then followed by Section 4 that presents the primary contribution of this paper i.e., novel weight update law for critic NN. This section contains subsection 4.2 that discusses the stability proof of the update law presented in this paper in detail. Finally towards the end, the paper is concluded by 5 and 6 that discusses results and conclusions respectively.

2 Robust Optimal Tracking Controller

2.1 Problem Formulation

The uncertain nonlinear dynamics is given by the affine-in-control equation,

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

where $f(x)$ and $g(x)$ are known dynamics (drift and control coupling dynamics) and $\Delta f(x)$ is the unknown matching perturbation.

Assumption 1. The drift dynamics i.e., $f(x)$ is Lipschitz continuous in $x \in \Omega \subset \mathbb{R}^n$ and $g(x)$ is bounded such that, $\exists g_M > 0 \ni 0 < \|g(x)\| < g_M,\ \forall x \in \mathbb{R}^n$. It is also assumed that matching condition is satsified by the perturbation i.e., $\Delta f(x) = g(x)d(x)$, where $d(x) \in \mathbb{R}^m$ is an unknown function bounded by a known function $d_M(x) > 0$.

Assumption 2. The commanded trajectory, i.e, the $\dot{x}_d(t) : \mathbb{R} \rightarrow \mathbb{R}^n$ is bounded and is Lipschitz continuous satisfying $H(0) = 0, \ni, \dot{x}_d = H(x_d)$.

These two assumptions are in line with Assumption 2 of [23] and Assumptions 1, 2 and 3 of [24].

Objective of the control: It is required to derive a robust optimal tracking controller that makes the system trajectories x follow the desired reference trajectory x_d with state error in a sufficiently small neighborhood of the origin in the presence of unknown but bounded $d(x)$.

2.2 Preliminaries of Robust OTCP

In [24] the feedback controller was derived for constrained input case in the presence of unknown uncertainties for optimal regulation problem. In this paper optimal tracking problem is considered with actuator constraints and unknown uncertainties. In order to achieve the desired objective, an augmented system dynamics that consists of dynamics of errors ($\dot{e}$) and desired states ($\dot{x}_d$) is defined first. Using (1) and Assumptions (1) and (2), tracking error dynamics can be written as:

$$\begin{aligned} \dot{e} &= \dot{x} - \dot{x}_d \\ \dot{e} &= f(x_d + e) + g(x_d + e)u(t) - H(x_d(t)) + \Delta f(x_d + e) \end{aligned} \tag{2}$$

Therefore, the dynamics of augmented system, given as $z = [e^T, x_d^T]^T$, can compactly be written as:

$$\dot{z} = F(z) + G(z)u + \Delta F(z) \tag{3}$$

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

$$F(z) = \begin{pmatrix} f(e + x_d) - H(x_d) \\ H(x_d) \end{pmatrix},\ G(z) = \begin{pmatrix} g(e + x_d) \\ 0 \end{pmatrix} \tag{4}$$

$\Delta F(z) \in \mathbb{R}^{2n}$ and is defined as $\Delta F(z) = G(z)d(z)$ with $d(z) \in \mathbb{R}^m$ and $\|d(z)\| \leq d_M(z)$. Following Assumptions 1 and 2 and Eq. (4), $\|F(z)\| \leq L_f\|z\|$ and $\|G(z)\| \leq g_M$. In the subsequent analysis, $d_M \triangleq d_M(z)$.

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 [21] [25]. Nominal

augmented dynamics is given by:

$$\dot{z} = F(z) + G(z)u \tag{5}$$

The infinite horizon discounted cost function for (5) is considered as follows [21] :

$$V(z) = \int_t^\infty e^{-\gamma(\tau - t)}[d_M^2 + \bar{u}(z,u)]d\tau \tag{6}$$

where, $\bar{u} = z^T Q_1 z + C(u)$ is the utility function comprising of augmented state z and the control action u. The positive definite diagonal matrix $Q_1 \in \mathbb{R}^{2n\times 2n}$ is defined as,

$$Q_1 = \begin{pmatrix} Q & 0_{n\times n} \\ 0_{n\times n} & 0_{n\times n} \end{pmatrix} \tag{7}$$

where, $Q \in \mathbb{R}^{n\times n}$ is a positive definite diagonal matrix with non zero entries.

Remark 1. Following from [23] and [24], it can be shown that robust control problem for (3) can be transformed into optimal tracking control problem for nominal augmented system (5) with discounted cost function (6).

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, discounted cost function of the form (6) is chosen for trajectory tracking problems. Generally, the function $C(u)$ is quadratic in nature, however, it can be non-quadratic [26], [27], if, control constraints are taken into account, i.e, $|u_i| \leq u_m,\ i = 1, 2...m$. This corresponds to an input-constrained scenario, which is also considered in this paper. Thus, $C(u)$ is defined in this paper as follows [7],[26]-[28].

$$\begin{aligned} C(u) &= 2u_m \int_0^u (\psi^{-1}(\nu/u_m))^T R d\nu \\ &= 2u_m \sum_{i=1}^m \int_0^{u_i} (\psi^{-1}(\nu_i/u_m))^T R_i d\nu_i \end{aligned} \tag{8}$$

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 ($|\psi(.)| \leq 1$) that belongs to $C^p (p \geq 1)$. In literature dealing with constrained input, some of the possible candidates for ψ include, $tanh, erf, sigmoid$. In this paper $\psi^{-1}(.) = \tanh^{-1}(.)$. It can be clearly observed that $C(u)$ (as shown in Lemma 7.2 in appendix) is positive. The discount factor, $0 \leq \gamma$, defines the value of utility in future. The first term inside the integral caters to any perturbations or uncertainties that might appear in the plant dynamics.

Differentiating (6) along the nominal system trajectories the following can be obtained [14]:

$$\begin{aligned} V_z(z)(F(z) + G(z)u) - \gamma V(z) + d_M^2 + \bar{u}(z,u) \\ = \mathcal{H}(z, u, V_z(z)) = 0 \end{aligned} \tag{9}$$

where, $\mathcal{H}(.)$ represents the Hamiltonian and $V_z(z) \triangleq \nabla_z V(z)$. Let $V^*(z) \in C^1$ be the optimal cost function that satisfies $\mathcal{H}(.) = 0$ and is given by:

$$V^*(z) = \min_u \int_t^\infty e^{-\gamma(\tau - t)}[d_M^2 + \bar{u}(z,u)]d\tau \tag{10}$$

Also in the subsequent analysis, $V \triangleq V(z)$, $V^* \triangleq V^*(z)$ and $V_z^* \triangleq V_z^*(z)$. Thus, $\mathcal{H}(.) = 0$ can be re-written in terms of optimal cost as:

$$\nabla_z V^*(F(z) + G(z)u) - \gamma V^*(z) + d_M^2 + \bar{u}(z,u) = 0 \tag{11}$$

Differentiating (11) with respect to (w.r.t.) u, i.e, $\partial\mathcal{H}/\partial u = 0$, closed form of optimal control action u^* is obtained as:

$$u^* = -u_m \tanh\left(\frac{1}{2u_m} R^{-1} G(z)^T \nabla_z V^*\right) \tag{12}$$

Substituting (12) in (11) the HJB equation is formulated as:

$$\begin{aligned} V_z^* F(z) - 2u_m^2 A^T(z)\tanh(A(z)) + d_M^2 + z^T Q_1 z + \\ 2u_m \int_0^{u^*} \tanh^{-1}(\nu/u_m)^T R d\nu - \gamma V^* = 0 \end{aligned} \tag{13}$$

where $V_z^* = \nabla_z V^*$, and $A = (1/2u_m)R^{-1}G(z)^T V_z^* \in \mathbb{R}^m$. The $C(u)$ or last but one term in left hand side of (13) can be simplified as:

$$\begin{aligned} &2u_m \int_0^{-u_m \tanh A(z)} \tanh^{-1}(\nu/u_m)^T R d\nu \\ &= 2u_m^2 A^T(z) R \tanh A(z) + u_m^2 \sum_{i=1}^m R_i \ln[1 - \tanh^2 A_i(z)] \end{aligned} \tag{14}$$

Eq. (14) follows from Lemma 7.1 given in Appendix 7. Now, using (14), Eq. (13) can further be simplified into:

$$V_z^* F(z) + d_M^2 + z^T Q_1 z + u_m^2 \sum_{i=1}^m R_i \ln[1 - \tanh^2 A_i(z)] - \gamma V^* = 0 \tag{15}$$

Eq. (15) is the HJB equation which is a nonlinear PDE in optimal cost function. Note that $\ln(.)$ used in this paper is natural log with base e. Now, the optimality of u^* (defined in (12)) and asymptotic stability of ($e = x - x_d$) would be discussed, which follows the line of logic as Theorem 1 of [21].

Theorem 2.1. *For augmented system defined in (3), and its associated discounted cost function defined in (6) with V^* being the solution of the HJB equation, the controller (u^*) described as in (12) minimizes the performance index (6) over all control policies constrained to $|u_i| \leq u_m$. Further, it also ensures asymptotic stability of error dynamics (2) in the limiting sense when $\gamma \to 0$.*

Proof: In the formulation of robust OTCP, it can be observed from (6) and (9) that both $V(z)$ and $\mathcal{H}$ contain d_m (upper bound of unknown perturbation $d(z)$) in their expressions, respectively. This is the difference w.r.t. the expressions of $V(z)$ and $\mathcal{H}$ corresponding to OTCP in [21]. It is evident that the presence of the robust term d_M^2 in the performance index does not alter the proof for optimality of u^*. Therefore, for details of this part of the proof refer to the proof of Theorem 1 in [21].

However, the presence of the robust term d_M^2 in the performance index affect the proof of the stability in the following way.

Differentiating $V^*(z)$ along the augmented system trajectories,

$$\nabla_z V^*(F(z)+G(z)u+\Delta F(z))-\gamma V^*(z)+d_M^2+ \bar{u}(z,u)=0 \tag{16}$$

Multiplying $e^{-\gamma t}$ to both sides of Eq. (16),

$$\frac{de^{-\gamma t}V(z(t))}{dt}=-e^{-\gamma t}[d_M^2+\bar{u}(z,u^*)]\leq 0 \tag{17}$$

Note that, (17) has an additional term (d_M^2) on the right hand side (RHS) of the equation compared to the Eq. (43) of [21]. Therefore, following the same methodology as [21], it can be observed that, tracking error is asymptotically stable when $\gamma=0$. However, when $\gamma\neq 0$, the considering V^* and $C(u^*)$ to be finite and using the fact that $z^TQ_1z=e^TQe$, the stability can be analyzed in two cases, namely,

Case (a): When $A<0\Rightarrow\gamma<\frac{d_M^2+C(u^*)}{V^*}$, then in this case, $\dot{V}<0$, for all values of e implying asymptotic stability. In order to ensure asymptotic stability, use of sufficiently small value of γ is suggested.

Case (b): When $A>0\Rightarrow\gamma>\frac{d_M^2+C(u^*)}{V^*}$, then $\dot{V}<0$, only when following inequality is satisfied, i.e.,

$$\|e\|>\sqrt{\frac{\gamma V^*-d_M^2-C(u^*)}{\lambda_{min}(Q)}} \tag{18}$$

The RHS of Eq. (18) gives the UUB bound for state error e, which is valid only when $A>0$.

□

3 Background of Optimal Tracking Using RL

3.1 Approximation of Value function using Critic NN

For applying the optimal controller (12), V^* must be calculated first. This is difficult to achieve because it requires solution to (11), which is a nonlinear PDE. In order to by-pass solving the HJB equation directly, an NN will be utilized to approximate the value function. For that, in this paper, the value function is assumed to be smooth. Let there exist ideal weight parameter vector W that can accurately approximate the value function as:

$$V^*(z)=W^T\vartheta(z)+\varepsilon \tag{19}$$

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 ϑ_js are linearly independent of each other. Substituting (19) in (12),

$$u^*(z)=-u_m\tanh\left(\frac{1}{2u_m}R^{-1}G(z)^T\nabla\vartheta^TW+\varepsilon_{uu}\right) \tag{20}$$

where, $\varepsilon_{u^*}=(1/2u_m)R^{-1}G^T(z)\nabla\varepsilon\in\mathbb{R}^m$. Next, substituting (19) in (15), the HJB equation can be written as,

$$W^T\nabla\vartheta F(z)-\gamma W^T\vartheta+z^TQ_1z+d_M^2+ u_m^2\sum_{i=1}^{m}\ln[1-\tanh^2(\tau_{1i}+\epsilon_{u_i^*})]+\nabla\epsilon^TF(z)=0 \tag{21}$$

where, $\tau_1=(1/2u_m)R^{-1}G(z)^T\nabla\vartheta^TW=[\tau_{11},...,\tau_{1m}]^T\in\mathbb{R}^m$, $\varepsilon_{u^*}=[\varepsilon_{u_{11}^*},\varepsilon_{u_{12}^*},...,\varepsilon_{u_{1m}^*}]^T$. Upon using Mean value theorem [29], Eq. (21) becomes:

$$W^T\nabla\vartheta F(z)-\gamma W^T\vartheta+z^TQ_1z+d_M^2+ u_m^2\sum_{i=1}^{m}\ln[1-\tanh^2(\tau_{1i})]+\epsilon_{HJB}=0 \tag{22}$$

where, ϵ_{HJB} represents the HJB approximation error [7],[14] having a form similar to the one in [24] and is given as,

$$\epsilon_{HJB}=\nabla\epsilon^TF(z)+\sum_{i=1}^{m}\frac{2u_m^2}{p_{1i}}\tanh p_{2i}(\tanh^2 p_{2i}-1)\epsilon_{u_i^*} \tag{23}$$

where, $p_{1i}\in\mathbb{R}$ and $p_{2i}\in\mathbb{R}$ considered between $1-\tanh^2 A_i(z)$ and $1-\tanh^2\tau_i$. Now, using (19) and mean value theorem, the optimal control can be re-written as:

$$u^*=-u_m\tanh(\tau_1(z))+\epsilon_u \tag{24}$$

where $\tau_1(z)=(1/2u_m)R^{-1}\hat{G}^T\nabla\vartheta^TW=[\tau_{11},...,\tau_{1m}]^T\in\mathbb{R}^m$ and $\epsilon_u=-(1/2)((I_m-diag(\tanh^2(q)))R^{-1}\hat{G}^T\nabla\epsilon)$ with $q\in\mathbb{R}^m$ and $q_i\in\mathbb{R}$ considered between $\tau_{1i}+\varepsilon_{uui}$ and ϵ_{uui} i.e., i^{th} element of $\tau_1+\varepsilon_{uu}$ and ϵ_{uu}, respectively such that tangent of $\tanh(q)$ is equal to the slope of the line joining $\tanh(\tau_1+\epsilon_{uu})$ and $\tanh\varepsilon_{uu}$. For the detailed proof, refer to Lemma 7.3. In the subsequent analysis, $\tau_1\triangleq\tau_1(z)$.

Since ideal weights that can accurately approximate the value function are unknown, their estimates will be used instead as follows.

$$V(z)=\hat{W}^T\vartheta(z) \tag{25}$$

Error in critic weights is given by $\tilde{W}=W-\hat{W}$. Using (25) the estimated optimal control action can be described as:

$$\hat{u}(z)=-u_m\tanh\left(\frac{1}{2u_m}G^T(z)\nabla\vartheta^T\hat{W}\right) \tag{26}$$

From (15) and (25) the HJB approximation error is obtained as follows.

$$\hat{H}(z,\hat{W})=\hat{W}^T\nabla\vartheta F(z)-\gamma\hat{W}^T\vartheta+z^TQ_1z+d_M^2+ u_m^2\sum_{i=1}^{m}\ln[1-\tanh^2(\tau_{2i})]\triangleq e(z,\hat{W}) \tag{27}$$

where, $e(z,\hat{W})$ is the HJB error (referred to as $\hat{e}$ in subsequent discussion) and $\tau_2(z)=(1/2u_m)G^T(z)\nabla\vartheta^T\hat{W}=[\tau_{21}(z),...,\tau_{2m}(z)]^T\in\mathbb{R}^m$. Next, from (22) and (27) the HJB error can be expressed in terms of ($\tilde{W}$ which is $W-\hat{W}$) and W as [24]:

$$e=-\tilde{W}^T\nabla\vartheta F(z)+\gamma\tilde{W}^T\vartheta+\sum_{i=1}^{m}u_m^2[\Gamma(\tau_{2i})-\Gamma(\tau_{1i})]-\epsilon_{HJB} \tag{28}$$

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

$$\Gamma(\tau_{\iota i})=-2\ln[1+exp(-2\tau_{\iota i}sgn(\tau_{\iota i}))]-2\tau_{\iota i}sgn(\tau_{\iota i})+\ln(4) \tag{29}$$

where, sgn is signum function. Also note that:

$$\sum_{i=1}^{m}\Gamma(\tau_{\iota i})=-2\sum_{i=1}^{m}\ln[1+exp(-2\tau_{\iota i}sgn(\tau_{\iota i}))]- 2\tau_\iota^Tsgn(\tau_\iota)+m\ln(4) \tag{30}$$

Therefore, using (28) and (30), e in terms of $\tilde{W}$, is obtained as [24]:

$$\begin{aligned}\hat{e} &= 2u_m^2[\tau_1^T sgn(\tau_1) - \tau_2^T sgn(\tau_2)] - \tilde{W}\nabla\vartheta F(z) + u_m^2\Delta\tau \\ &\quad - \epsilon_{HJB} \\ &= u_m[W^T\nabla\vartheta G(z)sgn(\tau_1(z)) - \\ &\hat{W}^T\nabla\vartheta G(z)sgn(\tau_2(z))]\tilde{W}^T\nabla\vartheta F(z) + u_m^2\Delta\tau - \epsilon_{HJB} \\ &= -\tilde{W}^T[\nabla\vartheta F(z) - u_m\nabla\vartheta G(z)sgn(\tau_2)] + \rho(z)\end{aligned} \tag{31}$$

where,

$$\begin{aligned}\Delta\tau &= 2\sum_{i=1}^{m}\ln\left(\frac{1+exp[-2\tau_{1i}(z)sgn(\tau_{1i}(z))]}{1+exp[-2\tau_{2i}(z)sgn(\tau_{2i}(z))]}\right) \\ \rho(z) &= u_m W^T\nabla\vartheta G(z)[sgn(\tau_1(z) - sgn(\tau_2(z)))] + u_m^2\Delta\tau \\ &\quad - \epsilon_{HJB}\end{aligned} \tag{32}$$

3.2 Existing update laws in literature

In traditional RL literature for continuous time nonlinear systems, a quadratic cost function of the form, $E = (1/2)\hat{e}^2$ is chosen, and then gradient descend (GD) is used to drive the parameters $\hat{W}$ so as to minimize this cost E and thus to minimize the HJB error. The following tuning law has been proposed in [10],[15],[16],[19],[21],[30].

$$\dot{\hat{W}} = -\frac{\alpha}{(1+\phi^T\phi)^2}\frac{\partial E}{\partial\hat{W}} = -\frac{\alpha\phi}{(1+\phi^T\phi)^2}\hat{e} \tag{33}$$

where, $\phi = \nabla\vartheta(F(z) + G(z)\hat{u})$, $\alpha > 0$ is the learning rate, and $1 + \phi^T\phi$ is the normalization factor. Then in 2015, Yang et al. [24] proposed a modified version of (33) for optimal regulation problems wherein they used constant learning rate in their gradient descent formulation. Their update mechanism was given as below.

$$\begin{aligned}\dot{\hat{W}} &= -\alpha\bar{\phi}\left(Y(x) + d_M^2(x) + u_m^2\sum_{i=1}^{m}\ln[1-\tanh^2(\tau_{2i}(x))]\right) \\ &\quad + \frac{\alpha}{2}\Xi(x,\hat{u})\nabla\vartheta G(x)[I_m - \mathcal{B}(\tau_2(x))]G^T(x)L_{2x} \\ &\quad + \alpha\Big((K_1\varphi^T - K_2)\hat{W} + u_m\nabla\vartheta G(x)[\tanh(\tau_2(x)) - \\ &\quad sgn(\tau_2(x))]\frac{\varphi^T}{m_s}\hat{W}\Big)\end{aligned} \tag{34}$$

where, x is the actual state of the system (not the augmented state), $\alpha > 0$, $\phi = \nabla\vartheta(F(x) + G(x)\hat{u})$, $\bar{\phi} = \phi/m_s^2$, $\varphi = \phi/m_s$, $m_s = 1 + \phi\phi^T$, $Y(x) = \hat{W}^T\nabla\vartheta F + x^T Q_1 x$, $\mathcal{B} = diag\{\tanh^2(\tau_{2i}(x))\}$, $i = 1, 2..., m$.

4 Variable gain-based update law

4.1 Update law

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 [24] or the error in states $(x - x_d)$ [23] 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.

$$\begin{aligned}\dot{\hat{W}} &= -\alpha(|e(z,\hat{W})|^{k_2} + l)\bar{\phi}e(z,\hat{W}) \\ &\quad + \frac{\alpha}{2}\Xi(z,\hat{u})\nabla\vartheta G(z)[I_m - \mathcal{B}(\tau_2(z))]G^T(z)L_{2z} \\ &\quad + \alpha(|e(z)|^{k_2} + l)\Big((K_1\varphi^T - K_2)\hat{W} \\ &\quad + u_m\nabla\vartheta G(z)[\tanh(\tau_2(z)) - sgn(\tau_2(z))]\frac{\varphi^T}{m_s}\hat{W}\Big)\end{aligned} \tag{35}$$

where, $\alpha > 0$ is the learning rate, l is a small positive constant, and $e(z,\hat{W})$ is the HJB error as mentioned in (27). In the subsequent analysis, $g_1 \triangleq |\hat{e}|^{k_2} + l$ further, to ease the development of stability proof, $k_2 = 1$, however it can be set to any positive value. In (35), the term ϕ is defined as $\phi = \nabla\vartheta(F(z) + G(z)\hat{u}) - \gamma\vartheta(z)$, $\bar{\phi} = \phi/m_s^2$, $\varphi = \phi/m_s$, $m_s = 1 + \phi\phi^T$, $\mathcal{B} = diag\{\tanh^2(\tau_{2i}(z))\}$, $i = 1, 2..., m$. The term $\Xi(z,\hat{u})$ is a piecewise continuous indicator function defined as in [24].

$$\Xi(z,\hat{u}) = \begin{cases}0, & if\ \Sigma(z(t)) < 0 \\ 1, & otherwise\end{cases} \tag{36}$$

where, $\Sigma(z(t)) = L_{2z}^T(F(z) + G(z)\hat{u})$ denotes the rate of variation of Lyapunov function along the system trajectories. It is to be noted that, $L_2 = (1/2)z^T z$ and hence $L_{2z} = z$. The constants, $k_2 > 0$ provide an augmentation to the controller by enabling accelerated learning, when the HJB error $(e(z,\hat{W}))$ is large. On the other hand, it dampen the learning process when $e(z,\hat{W})$ diminishes to a small quantity. Proper choice of this 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 5. Thus, the controller can bring the error within a small residual set around origin much quickly without any significant oscillations.

Note that the form of (35) is different from (34) that was presented in literature [24] in following ways.

- The ϕ in (35) has an additional term $\gamma\vartheta(z)$ and $e(z,\hat{W})$ has $-\gamma\hat{W}\vartheta(z)$. Both these terms arise because of the discounted cost function that was used in (35) compared to (34).
- The variable gain in first term of (35) is chosen to be a function of HJB error. This has been done in order to accelerate the reduction of HJB error when it is large and dampen the reduction process when the HJB error becomes small. The added benefit of variable gain is that it shrinks the size of the residual sets for both error in state and error in parameter as will become clear in the stability proof.
- The second term in (35) is dependent on the variation of Lyapunov function along the system trajectories. It is 0, when the Lypunov 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 leads to growth in Lyapunov function along the augmented system trajectories. 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 Σ denote the variation of Lyapunov function along the sytem trajectories as $\Sigma = L_{2z}(F(z) - u_m G(z)\tanh\tau_2(z))$
  ○ Gradient descent is utilized in [24] to drive the weights in direction such that Σ can be reduced and eventually made negative.

$$\begin{aligned}-\alpha\frac{\partial\Sigma}{\partial\hat{W}} &= -\alpha\frac{\partial[L_{2z}(F(z) - u_m G(z)\tanh\tau_2(z))]}{\partial\hat{W}} \\ &= \alpha\left(\frac{\partial\tau_2(z)}{\partial\hat{W}}\right)^T\frac{\partial[u_m L_{2z}G(z)\tanh\tau_2(z)]}{\partial\hat{W}} \\ &= \frac{\alpha}{2}\nabla\vartheta G(z)[I_m - \mathcal{B}(\tau_2(z))]G^T(z)L_{2z}\end{aligned} \tag{37}$$

- The last term in (35) provides control over the UUB sets as mentioned in [24]. Proper choice of gains K_1 and K_2 can shrink the UUB ball close to the origin.

Using (31) and (35) the dynamics of error in critic weights is then given as,

$$\begin{aligned}\dot{\tilde{W}} &= \alpha g_1 \frac{\varphi}{m_s}\Big[-\tilde{W}^T\phi + u_m\tilde{W}^T\nabla\vartheta G(z)\mathcal{F}(z)+\rho(z)\Big]\\ &-\frac{\alpha}{2}\Xi(z,\hat{u})\nabla\vartheta G(z)\Big[I_m-\mathcal{B}(z)\Big]G^T(z)L_{2z}\\ &+\alpha g_1\Big[\nabla\vartheta G(z)\mathcal{F}(z)\frac{\varphi^T}{m_s}\hat{W}+(K_2-K_1\varphi^T)\hat{W}\Big]\end{aligned} \tag{38}$$

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

4.2 Proof of Stability of Online Tuning Law

Assumption 3. Ideal NN weight vector W is considered to be bounded, i.e., $\|W\| \leq W_M$. There exists positive constants b_ϵ and $b_{\epsilon z}$ that bound the approximation error and its gradient such that $\|\varepsilon(z)\| \leq b_\epsilon$ and $\|\nabla\varepsilon\| \leq b_{\epsilon z}$. This is in line with Assumptions 3b of [10], Assumption 5 of [24] and Assumptions made in Section 4.1 in [21].

Assumption 4. Critic regressors are considered to be bounded as well: $\|\vartheta(z)\| \leq b_\vartheta$ and $\|\nabla\vartheta(z)\| \leq b_{\vartheta z}$. This is in line with Assumption 4 of [23], Assumption 6 of [24] and Assumption 4 of [21].

In this paper, both the assumptions hold true because, there exists a stabilizing term (second term) in the update law (35) that comes into effect when Lyapunov function starts growing along the system trajectories. This term helps in pulling the system out of region where Lyapunov function is growing thus ensuring that the trajectories remain finite within a region $\Omega_1 \subset R^{2n}$.

Assumption 5. Let $L_2 \in C^1$ be a continuously differentiable and radially unbounded Lyapunov candidate for (5) and satisfies $\dot{L}_2 = L_{2z}^T(F(z)+Gu^*) < 0$. Furthermore, a symmetric and positive definite $\Lambda(z) \in \mathbb{R}^{n\times n}$ can be found, such that, $L_{2z}^T(F(z)+Gu^*) = -L_{2z}^T\Lambda(z)L_{2z}$, where L_{2z} is the partial derivative of L_2 wrt z. In the subsequent analysis, $\Lambda \triangleq \Lambda(z)$.

Following Lipschitz continuity of $(F(z)+Gu^*)$ in z, this assumption can be shown reasonable. It is also in line with Assumption 4 mentioned in [24] and [22].

Theorem 4.1. *Let the CT nonlinear augmented system be described by (5) with associated HJB as (15) and approximate optimal control as (26) and let the Assumptions 1-5 hold true, then the tuning law (35) makes L_{2z} and $\tilde{W}$ uniform ultimate boundedness (UUB) stable. Further, the UUB set could be made arbitrary small by proper selection of gains K_1, K_2 and exponent k_2 in (35).*

Proof: Let the Lyapunov candidate be $L = L_2 + (1/2\alpha)\tilde{W}^T\tilde{W}$ (Where L_2 is a positive definite function of augmented state as defined after (35)). Derivative of L w.r.t. time is obtained as follows.

$$\begin{aligned}\dot{L} &= L_{2z}(F(z)+G(z)\hat{u}) + \dot{\tilde{W}}\alpha^{-1}\tilde{W}\\ &= L_{2z}(F(z)-u_mG(z)\tanh(\tau_2(z))) + \dot{\tilde{W}}\alpha^{-1}\tilde{W}\end{aligned} \tag{39}$$

Utilizing error dynamics of weights, i.e (38) and using the fact that $\dot{z} = F(z)+G(z)\hat{u}$, the last term of Lyapunov derivative becomes:

$$\begin{aligned}&\dot{\tilde{W}}\alpha^{-1}\tilde{W} = \Big[-\tilde{W}^T\phi + u_m\tilde{W}^T\nabla\vartheta G(z)\mathcal{F}(z)+\rho(z)\Big]g_1\frac{\varphi^T}{m_s}\tilde{W}\\ &-\frac{1}{2}g_2\Xi(z,\hat{u})L_{2z}^TG(z)\Big[I_m-\mathcal{B}(\tau_2(z))\Big]G^T(z)\nabla\vartheta^T\tilde{W}\\ &+g_1u_m\tilde{W}\nabla\vartheta G(z)\mathcal{F}(z)\frac{\varphi^T}{m_s}\hat{W}+g_1\tilde{W}^T(K_2\hat{W}-K_1\varphi^T\hat{W})\\ &=-g_1\tilde{W}\varphi\varphi^T\tilde{W}+g_1\delta(z)\varphi^T\tilde{W}+g_1\tilde{W}^T\beta(z)+g_1\tilde{W}^T(K_2\hat{W}\\ &-K_1\varphi^T\hat{W})\underbrace{-\frac{1}{2}\Xi(z,\hat{u})L_{2z}^TG[I_m-\mathcal{B}(\tau_2(z))]G^T(z)\nabla\vartheta^T\tilde{W}}_{S}\end{aligned} \tag{40}$$

where $\delta(z) \triangleq \rho(z)/m_s$ and $\beta(z) \triangleq u_m\nabla\vartheta G(z)\mathcal{F}(z)(\varphi^T/m_s)W$. Let, $A \triangleq -g_1\tilde{W}\varphi\varphi^T\tilde{W} + g_1\delta(z)\varphi^T\tilde{W} + g_1\tilde{W}^T\beta(z)$ $+ g_1\tilde{W}^T(K_2\hat{W}-K_1\varphi^T\hat{W})$. The last term in (40) can be expressed as:

$$\begin{aligned}\tilde{W}^T(K_2\hat{W}-K_1\varphi^T\hat{W}) = \tilde{W}^TK_2W - \tilde{W}^TK_2\tilde{W} -\\ \tilde{W}^TK_1\varphi^TW + \tilde{W}^TK_1\varphi^T\tilde{W}\end{aligned} \tag{41}$$

Let,

$$\mathcal{J} \triangleq [\tilde{W}^T\varphi, \tilde{W}^T]^T \tag{42}$$

then (40) can be re-written as:

$$\dot{\tilde{W}}^T\tilde{W}/\alpha = A + S \leq g_1(-\lambda_{min}(M)\|\mathcal{J}\|^2 + b_N\|\mathcal{J}\|) + S \tag{43}$$

where, M and N are defined as:

$$M = \begin{pmatrix}1 & -\frac{1}{2}K_1^T\\ -\frac{1}{2}K_1 & K_2\end{pmatrix};\ N = \begin{pmatrix}\delta(z)\\ (\beta(z)+K_2W-K_1\varphi^TW)\end{pmatrix} \tag{44}$$

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

$$\|N\| \leq b_N = \max(\|N\|) \tag{45}$$

In (44), if K_1 and K_2 are chosen such that K_2 is symmetric, then M becomes symmetric. Further, in order to ensure that $\lambda_{min}(M)$ is real and positive, K_1 and K_2 should be selected such that M is positive definite. Further, A can be developed by leveraging g_1 as a function of $\tilde{W}$. From (31), g_1 as a function of $\tilde{W}$ is, $g_1 = |\hat{e}(\tilde{W})|^{k_2} + l$ (with $k_2 = 1$).

$$\begin{aligned}g_1 &= |-\tilde{W}^T\phi + u_m\tilde{W}^T\nabla\vartheta G(z)\mathcal{F}(z)+\rho(z)| + l\\ &\leq \|\rho\| + \|\tilde{W}\|\|\phi\| + u_m\|\tilde{W}\|b_{\vartheta z}g_M2\sqrt{m} + l\\ &\leq \|\tilde{W}\|\underbrace{(\|\phi\| + u_mb_{\vartheta z}g_M2\sqrt{m})}_{A_1} + \underbrace{(\|\rho\|+l)}_{A_2}\end{aligned} \tag{46}$$

where, $\|\mathcal{F}\| \leq 2\sqrt{m}$. It could be noted that $\delta(z) = \rho(z)/m_s$ is one of the component of vector N, and by appropriately selecting K_1 and K_2 in N, and selecting a very small offset l, it can be ensured that $A_2 = \|\rho\| + l \leq b_N$. Also, from (42), $\|\tilde{W}\| \leq \|\mathcal{J}\|$, therefore,

$$g_1 \leq A_1\|\mathcal{J}\| + b_N \tag{47}$$

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

$$\begin{aligned}\dot{L} &\leq L_{2z}(F(z)+G(z)\hat{u}) +\\ &(A_1\|\mathcal{J}\| + b_N)(-\lambda_{min}(M)\|\mathcal{J}\|^2 + b_N\|\mathcal{J}\|) + S\end{aligned} \tag{48}$$

Based on the variation of Lyapunov function along the system trajectories, which is captured by the value of the piecewise continuous

function, $\Xi(z,\hat{u})$, (48) can be explained in two cases:
Case(i): When $\Xi(z,\hat{u})=0 \Rightarrow S=0$.
By definition, in this case, $L_{2z}^T\dot{z}<0$ (where $\dot{z}=F(z)+G(z)\hat{u}$). Therefore,

$$\dot{L} \leq L_{2z}^T\dot{z} + \underbrace{(A_1\|\mathcal{J}\|+b_N)(-\lambda_{min}(M)\|\mathcal{J}\|^2+b_N\|\mathcal{J}\|)}_{A} \tag{49}$$

In order for $\dot{L}$ to be negative definite, A should be negative, now for $\|\mathcal{J}\|\neq 0$, $A<0$ when,

$$\begin{aligned} &-A_1\lambda_{min}(M)\|\mathcal{J}\|^2 + \underbrace{(b_NA_1-\lambda_{min}(M)b_N)}_{\triangleq B_1}\|\mathcal{J}\| + b_N^2 < 0 \\ &\Rightarrow \|\mathcal{J}\| > \frac{B_1}{2A_1\lambda_{min}(M)} + \sqrt{\frac{B_1^2}{4A_1^2\lambda_{min}^2(M)} + \frac{b_N^2}{A_1\lambda_{min}(M)}} \\ &\Rightarrow \|\mathcal{J}\| > \frac{b_N}{\lambda_{min}(M)} \underbrace{\left[\frac{1}{2}(1-\gamma_1)+\sqrt{\frac{1}{4}(1-\gamma_1)^2-\gamma_1}\right]}_{\triangleq \Gamma} \end{aligned} \tag{50}$$

where, $\gamma_1 \triangleq \frac{\lambda_{min}(M)}{A_1}$, therefore, if, $0\leq\gamma_1\leq 3-\sqrt{8}\approx 0.17$, then $.478\leq\Gamma\leq 1$. Also, recall from the definition of $\mathcal{J}$ in (42), the upper bound of $\|\mathcal{J}\|$ can be obtained as,

$$\|\mathcal{J}\| \leq \left(\sqrt{1+\|\varphi\|^2}\right)\|\tilde{W}\| \tag{51}$$

Therefore, from lower and upper bounds of $\mathcal{J}$ in (50) and (51), respectively, the bound over $\|\tilde{W}\|$ becomes,

$$\|\tilde{W}\| > \frac{\frac{b_N}{\lambda_{min}(M)}\Gamma}{\sqrt{1+\|\varphi\|^2}} \tag{52}$$

It could be seen that $\tilde{W}$ is UUB stable with bound given in the RHS of (52). Also, note that if Eq. (52) holds, then the negative definiteness of $\dot{L}$ is ensured.

Case(ii): If $\Xi(z,\hat{u})=1$
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 (48) also needs to be considered. For that, (12), (48) and Assumption 5 would be utilized.

$$\begin{aligned} \dot{L} \leq\ & L_{2z}^TF(z) - u_mL_{2z}^TG(z)\Big[\tanh(\tau_2(z)) + \frac{2}{2u_m}[I_m \\ & - \mathcal{B}(\tau_2(z))]G^T\nabla\vartheta^T\tilde{W}\Big] + (A_1\|\mathcal{J}\|+b_N)(-\lambda_{min}(M)\|\mathcal{J}\|^2+b_N\|\mathcal{J}\|) \end{aligned} \tag{53}$$

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

$$\begin{aligned} \dot{L} \leq\ & L_{2z}^T(F(z)+Gu^*) - u_mL_{2z}^TG(z)\Big[\tanh(\tau_2(z)) \\ & + \frac{g_2}{2u_m}[I_m-\mathcal{B}(\tau_2(z))]G^T\nabla\vartheta^T\tilde{W}\Big] + (A_1\|\mathcal{J}\|+b_N) \\ & \times(-\lambda_{min}(M)\|\mathcal{J}\|^2+b_N\|\mathcal{J}\|) - L_{2z}^TG(z)(-u_m\tanh(\tau_1(z))+\epsilon_u) \end{aligned} \tag{54}$$

Using the inequality $\|\tanh(\tau_1(z))-\tanh(\tau_2(z))\|\leq T_m$ (see Lemma 7.4 in Appendix 7), Assumption 3, 4 and 5, Inequality (54) can be re-written as:

$$\begin{aligned} \dot{L} \leq\ & -L_{2z}^T\Lambda L_{2z} - L_{2z}^TG(z)\epsilon_u + u_m\|L_{2z}^T\|g_M\|\tanh(\tau_1(z) \\ & - \tanh\tau_2(z)\| + (A_1\|\mathcal{J}\|+b_N)(-\lambda_{min}(M)\|\mathcal{J}\|^2+b_N\|\mathcal{J}\|) \\ & + \frac{g_2}{2}\|L_{2z}^T\|\|\mathcal{N}_1\nabla^T\vartheta\tilde{W}\| \\ \leq\ & -\lambda_{min}(\Lambda)\|L_{2z}\|^2 + \|L_{2z}\|(T_mu_mg_M + \frac{g_2}{2}\|\mathcal{N}_1\nabla^T\vartheta\tilde{W}\|) \\ & + (A_1\|\mathcal{J}\|+b_N)(-\lambda_{min}(M)\|\mathcal{J}\|^2+b_N\|\mathcal{J}\|) + \frac{1}{2}\|L_{2z}^T\|g_M^2b_{\epsilon z} \end{aligned} \tag{55}$$

where, $\mathcal{N}_1 \triangleq G(z)[\mathcal{B}(\tau_2(z))-I_m]G^T(z)$. Now, two positive constant numbers n_1 and n_2 are defined such that $n_1+n_2=1$. In the following analysis, $\|\tilde{W}\|^2\leq\|\mathcal{J}\|^2$ is also utilized. Therefore the inequality in (55) can be developed as follows:

$$\begin{aligned} \dot{L} \leq\ & -n_1\lambda_{min}(\Lambda)\|L_{2z}^T\|^2 + \|L_{2z}\|T_mu_mg_M + \frac{\|g_2/2\mathcal{N}_1\nabla^T\vartheta\|^2\|\tilde{W}\|^2}{4n_2\lambda_{min}(\Lambda)} \\ & - n_2\lambda_{min}(\Lambda)\Big(\|L_{2z}^T\| - \frac{\|g_2/2\mathcal{N}_1\nabla^T\vartheta\|\|\tilde{W}\|}{2n_2\lambda_{min}(\Lambda)}\Big)^2 + A \\ \leq\ & -\underbrace{n_1\lambda_{min}(\Lambda)}_{\triangleq a}\|L_{2z}^T\|^2 + \|L_{2z}\|\underbrace{(T_mu_mg_M + \frac{1}{2}g_M^2b_{\epsilon z})}_{\triangleq b} \\ & + \underbrace{\left(\frac{\|g_2/2\mathcal{N}_1\nabla^T\vartheta\|^2}{4n_2\lambda_{min}(\Lambda)}\right)}_{\triangleq c}\|\mathcal{J}\|^2 + (A_1\|\mathcal{J}\|+b_N)(-\lambda_{min}(M)\|\mathcal{J}\|^2+b_N\|\mathcal{J}\|) \\ \leq\ & -a\left(L_{2z}-\frac{b}{2a}\right)^2 + \frac{b^2}{4a} + \|\mathcal{J}\|\Big(-A_1\lambda_{min}(M)\|\mathcal{J}\|^2 \\ & + (A_1b_N - b_N\lambda_{min}(M)+c)\|\mathcal{J}\| + b_N^2\Big) \end{aligned} \tag{56}$$

In order for $\dot{L}$ to be negative definite,

$$-a\left(L_{2z}-\frac{b}{2a}\right)^2 + \frac{b^2}{4a} < 0 \Rightarrow \|L_{2z}\| > \frac{b}{a} \tag{57}$$

and

$$\begin{aligned} &\|\mathcal{J}\|\Big(-A_1\lambda_{min}(M)\|\mathcal{J}\|^2 + (A_1b_N - b_N\lambda_{min}(M)+c)\|\mathcal{J}\| \\ &+ b_N^2\Big) < 0 \\ &\Rightarrow \|\mathcal{J}\| > \frac{b_N}{\lambda_{min}(M)} \underbrace{\left[\frac{1}{2}(1-\gamma_1+\alpha_2)+\sqrt{\left(\frac{1}{2}(1-\gamma_1+\alpha_2)\right)^2+\gamma_1}\right]}_{\triangleq \Gamma'} \end{aligned} \tag{58}$$

where, $\alpha_2 = c/(A_1b_N)\geq 0$ and Γ' is a fractional scaling factor that can scale the term $\frac{b_N}{\lambda_{min}(M)}$. Now, in order for Γ' to lie between $[c_1,c_2]$ such that $\frac{1}{2}<c_1<c_2<1$, γ_1 must lie between,

$$\frac{2c_2(1+\alpha_2)-c_2^2}{2c_2-1} \leq \gamma_1 \leq \frac{2c_1(1+\alpha_2)-c_1^2}{2c_1-1} \tag{59}$$

From (51) and (58), UUB set for $\tilde{W}$ is,

$$\|\tilde{W}\| > \frac{\frac{b_N}{\lambda_{min}(M)}\Gamma'}{\sqrt{1+\|\varphi\|^2}} \tag{60}$$

Therefore, for $\dot{L}$ to be negative definite, both (57) and (60) should hold true.

This completes the stability proof of the update mechanism (35). □

Remark 2. Note that for **Case (i)**, if $\tilde{W}$ satisfies (52), and for **Case (ii)**, if L_{2z} and $\tilde{W}$ satisfy (57) and (60), respectively, then it leads to decreasing $\tilde{W}$ and L_{2z}. It is evident that when the Lyapunov function is decreasing along the augmented state trajectory, variable learning rate has a direct influence over UUB bound for error in critic NN weights ($\tilde{W}$). By suitable selection of K_1 and K_2, the scaling factor Γ in (50) can be varied between .478 and 1 or $\Gamma^{'}$ in (58) can be varied between $\frac{1}{2}$ and 1 and accordingly the UUB bound of $\tilde{W}$ gets scaled down compared to that $\left(= \frac{b_N}{\lambda_{min}(M)}\right)$ with constant learning rate (also derived in Eq. (45) of [24]). The UUB set for $\tilde{W}$ for constant learning rate gradient descent is $\|\tilde{W}\| > \frac{b_N}{\lambda_{min}(M)}$ for both Case (i) and (ii). This leads to critic NN weights converging close to their ideal weights in finite time.

Remark 3. Further, variable learning can scale the learning speed based on the instantaneous value of the HJB approximation error. This leads to faster convergence time as compared to constant learning rate gradient descents

These advantages are exemplified in the following section.

5 Results and Simulation

In this section we will consider the numerical simulation of the parameter update law presented in this paper.

1. At first the parameter update law is validated on a generic nonlinear system with two different actuator limits and the results are contrasted against the constant learning parameter update law.
2. Thereafter, the variable gain update law is validated on a full 6-DoF nonlinear model of UAV and the result is contrasted against the constant learning-based parameter update law.

5.1 Nonlinear system

Consider a continuous time nonlinear system $\dot{x} = f(x) + g(x)u$ as mentioned in [23],

$$f = \begin{pmatrix} -x_1 + x_2 \\ -(x_1+1)x_2 - 49x_1 + .5((\cos(x_1))^3 \sin(x_2)) \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$
$$g = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \tag{61}$$

Drift dynamics f_1, f_2 and control coupling dynamics (g_1, g_2) are as mentioned in (61).

This continuous time nonlinear system is required to track a desired reference system given as [23].

$$\begin{pmatrix} \dot{x}_{d1} \\ \dot{x}_{d2} \end{pmatrix} = \begin{pmatrix} x_{d2} \\ -49x_{d1} \end{pmatrix} \tag{62}$$

The augmented state vector $z = [e_{x1}, e_{x2}, x_{d1}, x_{d2}]^T$. The Lyapunov function L_2 is selected as $L_2 = 1/2 z^T z$. Also, $R = 1$, and Q_1 (refer to Eqs. (6), (8)) is selected as,

$$Q_1 = \begin{pmatrix} I_2 & 0 \\ 0 & 0_{2\times 2} \end{pmatrix} \tag{63}$$

where, $I_2 = diag(10, 10)$. Regressor vector for critic network is selected as [23]. The larger the size of the regressor vector with multiple polynomial powers of augmented state z, the accurate will be the results [31].

$$\vartheta(z) = [z_1^2, z_2^2, z_3^2, z_4^2, z_1z_2, z_1z_3, z_1z_4, z_2z_3, z_2z_4, z_3z_4]^T \tag{64}$$

Initial state of the system is chosen to be, $x(0) = [1.5, 1.5]^T$. Critic weights are initialized to 0, i.e, $\hat{W}(0) = 0$. A dithering noise of the form $n(t) = 2e^{-0.009t}(\sin(11.9t)^2\cos(19.5t) + \sin(2.2t)^2\cos(5.8t) + \sin(1.2t)^2\cos(9.5t) + \sin(2.4t)^5)$ is added to maintain the persistent excitation (PE) condition [32]. 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. 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. 3 and 4 correspond to the case with input bound of 1.8, while Figs. 1 and 2 correspond to the input bound of 9. Constant learning rate (α) for $u_m = 9$ is selected to be 35.9 and discount ($\gamma = .1$), similarly for $u_m = 1.8$, the constant learning rate α was 92.9 and discount being $\gamma = .1$. Constants used in variable gain gradient descent are $k_2 = 1.4$ for $u_m = 9$ and $k_2 = .7$ for $u_m = 1.8$ (see Eq. (35)). Simulations have been run till the critic NN weights have converged in respective cases.

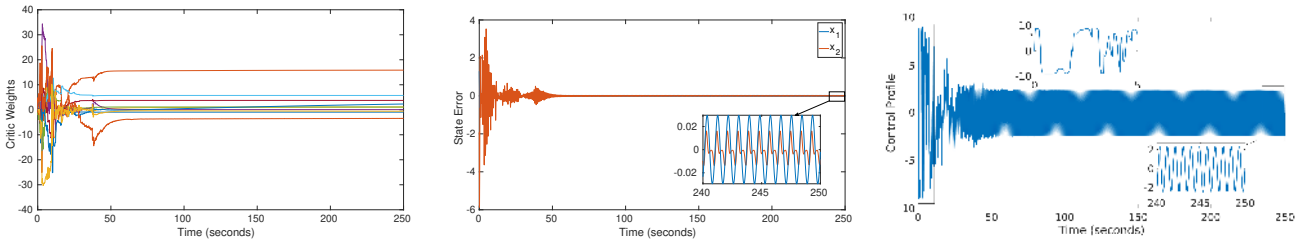
Comparing Figs. 1a and 2a, it can be observed that the application of variable gain gradient descent leads to faster and efficient learning of critic NN weights. In Fig. 1a when the variable gain gradient descent algorithm was utilized, the critic NN weights converged much before 250s, whereas in Fig. 2a, they took approximately 1300s. Observe that the update law presented in this paper is able to bring the state error to a much tighter residual set compared to constant learning rate-based gradient descents as can be seen from Figs. 1b and 2b. This is due to the fact that, the evolution of weight vectors with variable gain gradient descent converges to a smaller neighborhood about the ideal weight vector than with constant learning rate gradient descent.

It is noted from Figs. 1c and 2c that the optimal control commands generated were within the saturation limit of $[-9, 9]$. Also, most of the time, the control effort was well within this bounded interval $[-2.2, 2.2]$. Hence, in order to study the performance of the proposed adaptation scheme in a more stringent setting, a tighter control saturation limit $u_m = 1.8 < 2.2$ would be considered next.

A tighter input saturation has an adverse effect on learning as it takes more time to achieve convergence of critic NN weights as can be seen from Figs. 3a and 4a. In contrast to constant learning gradient descents, variable gain gradient descent-based update law presented in this paper is able to not only achieve convergence of critic NN weights within 1200s (refer to Fig. 3a) but also bring the state error (refer to Fig. 3b) to a tight residual set comparable to Fig 1b. On the other hand, under constant learning-based update law, some of the critic NN weights were not able to converge properly even after 3000s (refer to Fig. 4a). This results in larger state error as can be seen in Fig. 4b.

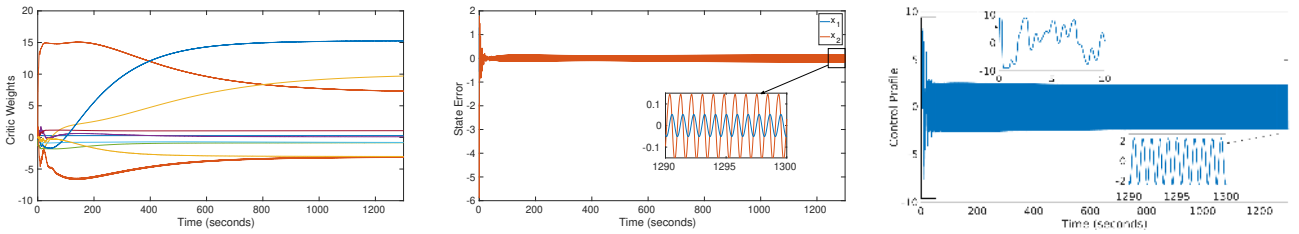
It is because of these reasons, that the update law presented in this paper yields improved tracking performance even with tight actuator constraints. The control effort was limited to $[-1.8, 1.8]$ for both with/without variable gain gradient descent update law as can be seen in Figs. 3c and 4c.

In [23], 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 [23]. Also in [23], the control effort during the initial phases touches $[-20, 20]$ and does not incorporate actuator constraints. As it can be inferred from Figs. 2b and 4b that a high constant learning rate leads to larger oscillation bound on states error compared to the case when variable gain gradient descent (see Figs. 1b and 3b) was utilized. It can be clearly concluded from Figs. 1 and 3 that, the variable gain gradient descent based tuning law proposed in this paper yields faster learning and is able to successfully bring the state error to a much tighter residual set than constant learning rate for both actuator contraints limits considered in this paper, i.e., $u_m = 9$ and $u_m = 1.8$.



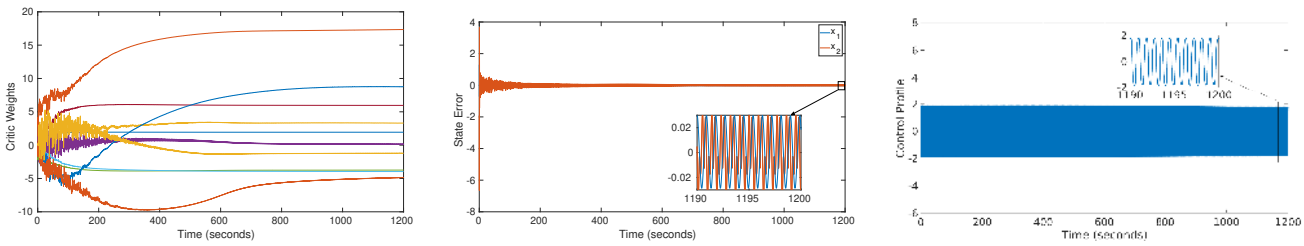
(a) Online training of Critic Weights (b) State error (c) Control action

Fig. 1: Critic NN, state error and control profile with variable gain gradient descent (for $u_m = 9$)



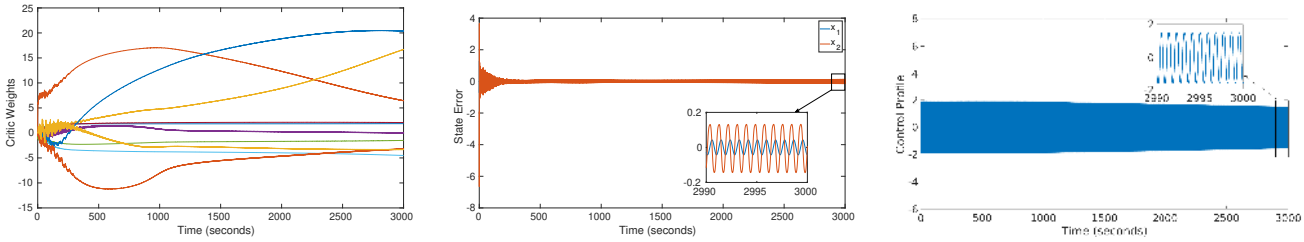
(a) Online training of critic weights (b) State error (c) Control action

Fig. 2: Critic NN, state error and control profile with constant learning gradient descent (for $u_m = 9$)



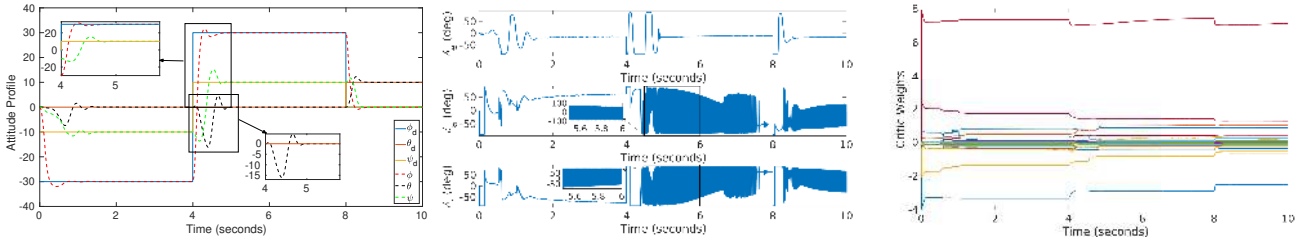
(a) Online training of critic weights (b) State error (c) Control action

Fig. 3: Critic NN, state error and control profile with variable gain gradient descent (for $u_m = 1.8$)



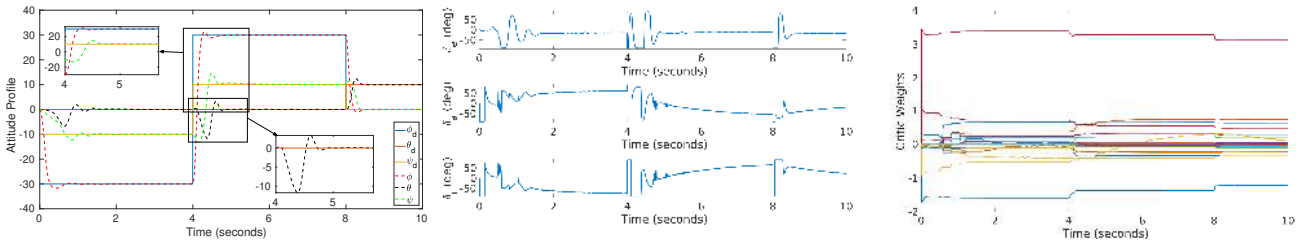
(a) Online training of critic weights (b) State error (c) Control action

Fig. 4: Critic NN, State error and Control Profiles with constant learning gradient descent (for $u_m = 1.8$)



(a) Desired and actual attitude profile (b) Control profile representing elevator, eileron and rudder (c) Online training of critic NN weights

Fig. 5: Performance of Aerosonde UAV under constant learning-based update law



(a) Desired and actual attitude profile (b) Control profile representing elevator, eileron and rudder (c) Online training of critic NN weights

Fig. 6: Performance of Aerosonde UAV under variable gain gradient descent-based update law

Thus, the prime advantage of variable gain gradient descent-based critic update law is the ability to select reasonably high learning rates without large steady state errors.

5.2 Full 6-DoF nonlinear model of UAV

In this subsection, the variable gain gradient descent update law presented in this paper is validated on the full 6-DoF nonlinear model of the Aerosonde UAV (refer to Pages 61 and 276 of [33]), and its performance is compared against that of the constant learning-based update law. The requirement of the control scheme is to track the desired attitude angles of the UAV. Desired set points for ϕ, θ, ψ (roll, pitch and yaw angle, respectively) were set to $(-30, 0, -10)$ degrees for first 4 seconds, then $(30, 0, 10)$ degrees from 4-8 seconds and finally $(0, 0, 10)$ degrees.

The control implementation is made up of two cascaded loops, the the first loop, i.e., outer loop converts the desired Euler angle information to desired rates, the inner loop uses the developed control algorithm to track the desired rates in an optimal way. Desired Euler angle rates are given by, $p_{des} = \dot{\phi}_{des} - 8e_{\phi}$, $q_{des} = \dot{\theta}_{des} - 10e_{\theta}$, $r_{des} = \dot{\psi}_{des} - 12e_{\psi}$, where ϕ, θ, ψ are roll, pitch and yaw angles, respectively. The deflection of elevator, aileron and rudder forms the control input to the UAV (represented by $\delta_e, \delta_a, \delta_r$ respectively). The control deflections are limited to ± 90 degrees.

The augmented state is, $z = [e_p, e_q, e_r, p_{des}, q_{des}, r_{des}]^T \in \mathbb{R}^6$ where $e = x - x_{des}$ and $x = [p, q, r]^T$. The regressor vector for critic NN is chosen to be, $\vartheta = [z_1, z_2, z_3, z_4, z_5, z_6, z_1^2, z_2^2, z_3^2, z_4^2, z_5^2, z_6^2, z_1z_2, z_1z_3, z_1z_4, z_1z_5, z_1z_6, z_2z_3, z_2z_4, z_2z_5, z_2z_6, z_3z_4, z_3z_5, z_3z_6, z_4z_5, z_4z_6, z_5z_6]^T$. The discount factor was selected as $\gamma = 0.1$, the weight matrix for augmented states and control are $Q_1 = diag(10, 10, 50, 0, 0, 0)$ and $R = I_3$, respectively. The baseline learning rate $\alpha = 14.7$, parameters for variable gain gradient descent are $k_2 = .1$. A dithering noise of the form, $n(t) = 2e^{-0.009t}(\sin(11.9t)^2\cos(19.5t) + \sin(2.2t)^2\cos(5.8t) + \sin(1.2t)^2\cos(9.5t) + \sin(2.4t)^5)$ is added to maintain the persistent excitation (PE) condition as demonstrated in [32]. All the critic weights were initialized to 0, i.e., $\hat{W}(0) = 0$. Both the update laws, i.e., variable gain gradient descent and constant learning rate update law were run with same set of parameters except the exponents in variable gain term and are able to track the desired reference set point. However, it can be seen from Figs. 6c and 5c that the critic weights undergo spike at the times when reference command for attitude changes i.e., at 4 and 8 second. At these junctures it can be clearly noticed that the critic weights when updated via the variable gain gradient descent-update law converge properly within sufficiently short time-span and before the next attitude reference signal changes, i.e., during the intervals, $0 - 4$ sec, $4 - 8$ sec and then finally $8 - 10$ sec. However, on the other hand, the critic weights are not able to converge properly in such short time-span between changes in the reference attitude when updated via constant learning-based update law. As critic weights have converged close to their ideal values in very small time when updated via the variable gain gradient descent update law, overshoots in state errors in this case are smaller in comparison with that in case of constant gain gradient descent, as can be seen from Figs. 6a and 5a. This effect is especially prominent in pitch (θ) dynamics. Additionally, the optimal control action (refer to Fig. 6b) generated via the variable gain gradient descent is much smoother compared to the control action (refer to 5b) generated via the constant learning based method, which is found to lead to persistent chattering in control command. The control effort in both these cases is bounded within ± 90 degrees.

Based on the above discussion it can be inferred that the variable gain gradient descent-based update law leads to faster convergence of critic weights closer to their ideal values. This in turn leads to achieving the ideal optimal tracking controller faster resulting into smaller overshoot in state error. Further, the control action generated by variable gain update law is devoid of chattering for the same set of actuator constraints.

6 Conclusion

The paper presents a variable gain gradient descent based update law for robust optimal tracking for continuous time nonlinear systems using reinforcement learning. The critic neural network (NN) is utilized to approximate the value function which is also the solution of the tracking HJB equation. It is this critic NN that is tuned online using the update law presented 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. 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. In addition to this, 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. The update law presented in this paper forms the basis of future scope of research using which model-free online update law to solve optimal tracking problem will be developed.

7 Appendices

Lemma 7.1. *Following equality holds true,*

$$2u_m\int_0^{-u_m\tanh A(z)}\tanh^{-1}(\nu/u_m)^T Rd\nu = 2u_m^2A^T(z)R\tanh A(z)+u_m^2\sum_{i=1}^{m}R_i\ln[1-\tanh^2 A_i(z)] \tag{65}$$

Proof:

$$\int\tanh^{-1}\left(\frac{x}{a}\right)=\frac{1}{2}a\ln(a^2-x^2)+x\tanh^{-1}\left(\frac{x}{a}\right)+I \tag{66}$$

Therefore,

$$\int_0^u\tanh^{-1}\left(\frac{\nu}{u_m}\right)d\nu=\frac{1}{2}u_m\ln(u_m^2-\nu^2)+\nu\tanh^{-1}\left(\frac{\nu}{u_m}\right)\Big|_0^u$$
$$2u_m\int_0^u\tanh^{-1}\left(\frac{\nu}{u_m}\right)d\nu=u_m^2\ln(u_m^2-\nu^2)+2u_m\nu\tanh^{-1}\left(\frac{\nu}{u_m}\right)\Big|_0^u$$
$$=u_m^2\ln(1-\frac{u^2}{u_m^2})+2u_m^2\tanh A(z)$$
$$=u_m^2\ln(1-\tanh^2 A(z))+2u_m^2\tanh A(z) \tag{67}$$

where, $u=-u_m\tanh A(z)$ is scalar. Now if u is a vector, then,

$$2u_m\int_0^u\tanh^{-1}\left(\frac{\nu}{u_m}\right)Rd\nu = 2u_m^2A^T(z)R\tanh A(z)+u_m^2\sum_{i=1}^{m}R_i\ln[1-\tanh^2 A_i(z)] \tag{68}$$

□

Lemma 7.2. *Following inequality holds true:*

$$C(u_i)=2u_m\int_0^{u_i}\psi^{-1}(\frac{\nu}{u_m})R_id\nu\geq 0 \tag{69}$$

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

Proof: If ψ^{-1} is monotonic odd and increasing, then,

$$\left(\frac{\nu}{u_m}\right)\psi^{-1}\left(\frac{\nu}{u_m}\right)\geq 0 \tag{70}$$

or

$$\nu\psi^{-1}\left(\frac{\nu}{u_m}\right)\geq 0 \tag{71}$$

where $\nu\in\mathbb{R}$ and $u_m>0$. Let $\theta=1/u_m$. In order to prove that, $2u_m\int_0^{u_i}\psi^{-1}(\nu/u_m)R_id\nu\geq 0$, 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, $\mathcal{K}\in[0,\theta]$ is assumed. Therefore,

$$\int_0^{u_i}\psi^{-1}(\nu\theta)d\nu=\frac{1}{\theta}\int_0^{u_i\theta}\psi^{-1}(l)dl \tag{72}$$

where $l=\nu\theta$. Similarly,

$$\frac{1}{\theta}\int_0^{u_i\theta}\psi^{-1}(l)dl=\frac{1}{\theta}\int_0^{\theta}\psi^{-1}(u_i\mathcal{K})u_id\mathcal{K} \tag{73}$$

by utilizing $l=u_i\mathcal{K}$
Since, $\psi^{-1}(u_i\mathcal{K})u_i\geq 0$, which implies,

$$\frac{1}{\theta}\int_0^{\theta}\psi^{-1}(u_i\mathcal{K})u_id\mathcal{K}\geq 0 \tag{74}$$

Lemma 7.3. *Following equation holds true,*

$$u=-u_m\tanh\left(\frac{1}{2u_m}R^{-1}\hat{G}^T\nabla\vartheta^TW+\epsilon_{uu}\right)=-u_m\tanh(\tau_1(z))+\epsilon_u \tag{75}$$

where, $\epsilon_{uu}=(1/2u_m)R^{-1}\hat{G}^T\nabla\varepsilon(z)=[\varepsilon_{uu_{11}},\varepsilon_{uu_{12}},...,\varepsilon_{uu_{1m}}]^T\in\mathbb{R}^m$. $\tau_1(z)=(1/2u_m)R^{-1}\hat{G}^T\nabla\vartheta^TW=[\tau_{11},...,\tau_{1m}]^T\in\mathbb{R}^m$ and $\epsilon_u=-(1/2)((I_m-diag(\tanh^2(q)))R^{-1}\hat{G}^T\nabla\epsilon)$ with $q\in\mathbb{R}^m$ and $q_i\in\mathbb{R}$ considered between $\tau_{1i}+\varepsilon_{uui}$ and ϵ_{uui} i.e., i^{th} element of ϵ_{uu}.

Proof:

$$u=-u_m\tanh(\tau_1+\varepsilon_{uu}) \tag{76}$$

Using mean value theorem,

$$\tanh(\tau_1+\varepsilon_{uu})-\tanh(\tau_1)=\tanh'(q)\varepsilon_{uu}$$
$$=(I_m-diag(\tanh^2(q)))\varepsilon_{uu} \tag{77}$$

where, $q\in\mathbb{R}^m$ and $q_i\in\mathbb{R}$ lying between τ_{1i} and $\tau_{1i}+\varepsilon_{uui}$.

Now, using the expression for ε_{uu} in (77), $\tanh(\tau_1+\varepsilon_{uu})$ can be rewritten as,

$$\tanh(\tau_1+\varepsilon_{uu})=\tanh(\tau_1)+(I_m-diag(\tanh^2(q)))\times\left(\frac{1}{2u_m}R^{-1}\hat{G}^T\nabla\varepsilon(z)\right) \tag{78}$$

Multiplying $-u_m$ on both sides,

$$-u_m\tanh(\tau_1+\varepsilon_{uu})=-u_m\tanh(\tau_1)-\frac{1}{2}(I_m-diag(\tanh^2(q)))\times\left(R^{-1}\hat{G}^T(z)\nabla\varepsilon(z)\right) \tag{79}$$

Hence proved. □

Lemma 7.4. *Following vector inequality holds true:*

$$\|\tanh(\tau_1(z))-\tanh(\tau_2(z))\|\leq T_m\leq 2\sqrt{m} \tag{80}$$

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

Proof: Since, $\tanh(.)$ is 1-Lipschitz, one can write,

$$|\tanh(\tau_{1i}) - \tanh(\tau_{2i})| \leq |\tau_{1i} - \tau_{2i}| \tag{81}$$

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

$$\begin{aligned} \|\tanh(\tau_1(z)) - \tanh(\tau_2(z))\|^2 &= \sum_{i=1}^{m} |\tanh\tau_{1i} - \tanh\tau_{2i}|^2 \\ &\leq \sum_{i=1}^{m} min(|\tau_{1i} - \tau_{2i}|, 2)^2 \\ &\leq \sum_{i=1}^{m} min(|\tau_{1i} - \tau_{2i}|^2, 4) \end{aligned} \tag{82}$$

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

$$\sum_{i=1}^{m} min(|\tau_{1i} - \tau_{2i}|^2, 4) \leq 2\sqrt{m} \tag{83}$$

Which implies,

$$\|\tanh(\tau_1(z)) - \tanh(\tau_2(z))\| \leq T_m \leq 2\sqrt{m} \tag{84}$$

□

8 References

1 Lewis, F.L. and Vrabie, D.: 'Reinforcement learning and adaptive dynamic programming for feedback control', *IEEE circuits and systems magazine*, 2009, **9**, (3), pp. 32–50

2 Werbos, P.: 'Beyond regression:" new tools for prediction and analysis in the behavioral sciences', *Ph D dissertation, Harvard University*, 1974,

3 Werbos, P.: 'Advanced forecasting methods for global crisis warning and models of intelligence', *General System Yearbook*, 1977, pp. 25–38

4 Barto, A.G.: '1" 1 adaptive critics and the basal ganglia,âĂİ', *Models of information processing in the basal ganglia*, 1995, p. 215

5 Barto, A.G., Sutton, R.S. and Anderson, C.W.: 'Neuronlike adaptive elements that can solve difficult learning control problems', *IEEE transactions on systems, man, and cybernetics*, 1983, , (5), pp. 834–846

6 Werbos, P.J.: 'Neural networks for control and system identification'. Proceedings of the 28th IEEE Conference on Decision and Control,, 1989. pp. 260–265

7 Abu.Khalaf, M. and Lewis, F.L.: 'Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach', *Automatica*, 2005, **41**, (5), pp. 779–791

8 Lin, W.S.: 'Optimality and convergence of adaptive optimal control by reinforcement synthesis', *Automatica*, 2011, **47**, (5), pp. 1047–1052

9 Liu, D., Yang, X. and Li, H.: 'Adaptive optimal control for a class of continuous-time affine nonlinear systems with unknown internal dynamics', *Neural Computing and Applications*, 2013, **23**, (7-8), pp. 1843–1850

10 Vamvoudakis, K.G. and Lewis, F.L.: 'Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem', *Automatica*, 2010, **46**, (5), pp. 878–888

11 Murray, J.J., Cox, C.J., Lendaris, G.G. and Saeks, R.: 'Adaptive dynamic programming', *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 2002, **32**, (2), pp. 140–153

12 Yang, X., Liu, D. and Wei, Q.: 'Online approximate optimal control for affine non-linear systems with unknown internal dynamics using adaptive dynamic programming', *IET Control Theory & Applications*, 2014, **8**, (16), pp. 1676–1688

13 Zhao, D. and Zhu, Y.: 'MecâĂŤa near-optimal online reinforcement learning algorithm for continuous deterministic systems', *IEEE transactions on neural networks and learning systems*, 2014, **26**, (2), pp. 346–356

14 Modares, H., Lewis, F.L. and Naghibi.Sistani, M.B.: 'Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems', *Automatica*, 2014, **50**, (1), pp. 193–202

15 Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K.G., Lewis, F.L. and Dixon, W.E.: 'A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems', *Automatica*, 2013, **49**, (1), pp. 82–92

16 Yang, X., Liu, D. and Wang, D.: 'Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints', *International Journal of Control*, 2014, **87**, (3), pp. 553–566

17 Vrabie, D., Pastravanu, O., Abu.Khalaf, M. and Lewis, F.L.: 'Adaptive optimal control for continuous-time linear systems based on policy iteration', *Automatica*, 2009, **45**, (2), pp. 477–484

18 Jiang, Y. and Jiang, Z.P.: 'Robust adaptive dynamic programming and feedback stabilization of nonlinear systems', *IEEE Transactions on Neural Networks and Learning Systems*, 2014, **25**, (5), pp. 882–893

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

20 Heydari, A. and Balakrishnan, S.N.: 'Fixed-final-time optimal tracking control of input-affine nonlinear systems', *Neurocomputing*, 2014, **129**, pp. 528–539

21 Modares, H. and Lewis, F.L.: 'Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning', *Automatica*, 2014, **50**, (7), pp. 1780–1792

22 Dierks, T. and Jagannathan, S.: 'Optimal control of affine nonlinear continuous-time systems'. Proceedings of the 2010 American Control Conference, 2010. pp. 1568–1573

23 Yang, X., Liu, D. and Wei, Q.: 'Robust tracking control of uncertain nonlinear systems using adaptive dynamic programming'. International Conference on Neural Information Processing, 2015. pp. 9–16

24 Liu, D., Yang, X., Wang, D. and Wei, Q.: 'Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints', *IEEE transactions on cybernetics*, 2015, **45**, (7), pp. 1372–1385

25 Kiumarsi, B., Lewis, F.L., Modares, H., Karimpour, A. and Naghibi.Sistani, M.B.: 'Reinforcement q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics', *Automatica*, 2014, **50**, (4), pp. 1167–1175

26 Abu.Khalaf, M., Lewis, F.L. and Huang, J.: 'Neurodynamic programming and zero-sum games for constrained control systems', *IEEE Transactions on Neural Networks*, 2008, **19**, (7), pp. 1243–1252

27 Modares, H., Lewis, F.L. and Naghibi.Sistani, M.B.: 'Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks', *IEEE Transactions on Neural Networks and Learning Systems*, 2013, **24**, (10), pp. 1513–1525

28 Lyashevskiy, S.: 'Constrained optimization and control of nonlinear systems: new results in optimal control'. Proceedings of 35th IEEE Conference on Decision and Control. vol. 1, 1996. pp. 541–546

29 Rudin, W., et al.: 'Principles of mathematical analysis'. vol. 3. (McGraw-hill New York, 1964)

30 Lewis, F.L. and Liu, D.: 'Reinforcement learning and approximate dynamic programming for feedback control'. vol. 17. (John Wiley & Sons, 2013)

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

32 Vamvoudakis, K.G., Vrabie, D. and Lewis, F.L.: 'Online adaptive algorithm for optimal control with integral reinforcement learning', *International Journal of Robust and Nonlinear Control*, 2014, **24**, (17), pp. 2686–2710

33 Beard, R.W. and McLain, T.W.: 'Small unmanned aircraft: Theory and practice'. (Princeton university press, 2012)