JAISCR, 2011, Vol.1, No.4, pp. 315–332

ONLINE LEARNING ALGORITHM FOR ZERO-SUM GAMES WITH INTEGRAL REINFORCEMENT LEARNING

Kyriakos G. Vamvoudakis [1], Draguna Vrabie [2], Frank L. Lewis [1]

[1]*Automation and Robotics Research Institute, University of Texas at Arlington, Texas, USA*
e-mail:kyriakos@arri.uta.edu

[2]*United Technologies Research Center, Connecticut, USA*

Abstract

In this paper we introduce an online algorithm that uses integral reinforcement knowledge for learning the continuous-time zero sum game solution for nonlinear systems with infinite horizon costs and partial knowledge of the system dynamics. This algorithm is a data based approach to the solution of the Hamilton-Jacobi-Isaacs equation and it does not require explicit knowledge on the system's drift dynamics. A novel adaptive control algorithm is given that is based on policy iteration and implemented using an actor/disturbance/critic structure having three adaptive approximator structures. All three approximation networks are adapted simultaneously. A persistence of excitation condition is required to guarantee convergence of the critic to the actual optimal value function. Novel adaptive control tuning algorithms are given for critic, disturbance and actor networks. The convergence to the Nash solution of the game is proven, and stability of the system is also guaranteed. Simulation examples support the theoretical result.

1 Introduction

The H_∞ control problem is a *minimax* optimization problem [3], and hence a zero-sum game where the controller is a minimizing player and the disturbance a maximizing one.

Game theory [1] and H-infinity solutions rely on solving the Hamilton-Jacobi-Isaacs (HJI) equations, which in the zero-sum linear quadratic case reduce to the generalized game algebraic Riccati equation (GARE). In the nonlinear case the HJI equations are difficult or impossible to solve, and may not have global analytic solutions even in simple cases (e.g. scalar system, bilinear in input and state). Solution methods are generally offline.

In this paper we use Reinforcement Learning (RL) [10] methods, specifically a new Integral Reinforcement Learning (IRL) approach, to provide a learning algorithm for the solution of two-player zero-sum infinite horizon games *online*. This algorithm does not need any knowledge of the drift dynamics of the system. A novel adaptive control technique is given that is based on reinforcement learning techniques, whereby the control and disturbance policies are tuned online using data generated in real time along the system trajectories. Also tuned is a 'critic' approximator structure whose function is to identify the value or outcome of the current control and disturbance policies. Based on this value estimate, the policies are continuously updated. This is a sort of indirect adaptive control algorithm, yet, due to the direct form dependence of the policies on the learned value, it is affected online as direct ('optimal') adaptive control.

Reinforcement learning (RL) is a class of methods used in machine learning to methodically modify the actions of an agent based on observed responses from its environment [11, 12, 14, 16]. The RL methods have been developed starting from learning mechanisms observed in mammals. Every decision-making organism interacts with its environment and uses those interactions to improve its own actions in order to maximize the positive effect of its limited available resources; this in turn leads to better survival chances. RL is a means of *learning optimal behaviors by observing the response from the environment to non-optimal control policies*. In engineering termsPersonName, RL refers to the learning approach of an actor or agent which modifies its actionsPersonName, or control policiesPersonName, based on stimuli received in response to its interaction with its environment.

In view of the advantages offered by the RL methodsPersonName, a recent objective of control systems researchers is to introduce and develop RL techniques which result in optimal feedback controllers for dynamical systems that can be described in terms of ordinary differential or difference equations. These involve a computational intelligence technique known as Policy Iteration (PI) [8, 11, 12, 14, 16], which refers to a class of algorithms built as a two-step iteration: *policy evaluation* and *policy improvement*. PI provides effective means of learning solutions to HJ equations online. In control theoretic terms, the PI algorithm amounts to learning the solution to a nonlinear Lyapunov equation, and then updating the policy through minimizing a Hamiltonian function. PI has primarily been developed for discrete-time systems, and online implementation for control systems has been developed through approximation of the value function based on work by [9, 10,13]. Recently, online policy iteration methods for continuous-time systems have been developed by [17, 18, 20].

We present an online integral reinforcement algorithm that combines the advantages of [19] and [20]. These include *simultaneous tuning* of disturbance, actor and critic neural networks (i.e. all neural networks are tuned at the same time) and no need for the drift term in the dynamics. Simultaneous tuning of actor/disturbance/critic structures was introduced by [18, 19], and has been the idea of recent papers in the area, however in most of these papers the authors either designed model-based controllers [22] or used dynamic neural networks to identify a model for the unknown nonlinear plant [23, 24]. Our algorithm avoids full knowledge of the plant and uses only three neural networks by designing a hybrid controller as in [20]. This paper generalizes the method given in [19] to solve the 2-player zero-sum game problem for nonlinear continuous-time systems without knowledge of the drift dynamics.

The contributions of this paper are a new direct adaptive control structure with three parametric approximation structures that converges to the solution of the zero-sum game problem without knowing the system drift dynamics term. The adaptive structure converges to the solutions to the HJI equation without ever explicitly solving either the HJI equation or nonlinear Lyapunov equations. The three approximation structures are tuned simultaneously in real time using data measured along the system trajectories.

The paper is organized as follows. Section 2 reviews the formulation of the two-player zero-sum differential game. A policy iteration algorithm is given to solve the HJI equation by successive solutions on nonlinear Lyapunov-like equations. This essentially extends Kleinman's algorithm to nonlinear zero-sum differential games. Section 3 develops the online zero-sum game PI algorithm with integral reinforcement learning. Care is needed to develop suitable approximator structures for online solution of zero-sum games. First a suitable 'critic' approximator structure is developed for the value function and its tuning method is pinned down. A persistence of excitation is needed to guarantee proper convergence. Next, suitable 'actor' approximator structures are developed for the control and disturbance policies. Finally the main result is presented in Theorem 1, which shows how to tune all three approximators simultaneously by using measurements along the system trajectories in real time. Proofs using Lyapunov techniques guarantee convergence and closed-loop stability. Section 4 presents simulation examples that show the effectiveness of the online synchronous zero-sum game CT PI algorithm in learning the zero sum game solution for both linear and nonlinear systems.

2 Background on Zero Sum Games

In this section is presented a background review of 2-player zero-sum differential games. The objective is to lay a foundation for the structure needed in subsequent sections for online solution of these problems in real-time. In this regard, the Policy Iteration Algorithm for 2-player games presented at the end of this section is key.

Consider the nonlinear time-invariant affine in the input dynamical system given by

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

where state $x(t) \in \mathbb{R}^n$, $f(x(t)) \in \mathbb{R}^n$, $g(x(t)) \in \mathbb{R}^{nxm}$,control $u(x(t)) \in \mathbb{R}^m$, $k(x(t)) \in \mathbb{R}^{nxq}$and disturbance $d(x(t)) \in \mathbb{R}^q$, Assume that $f(x)$ is locally Lipschitz, $f(0) = 0$ so that $x = 0$ is an equilibrium point of the system.

Define the performance index [26]
(2)

for$Q(x) \geq 0$,$R = R^T > 0$, $r(x,u,d,T) = Q(x) + u^T \frac{R}{T} u - \frac{\gamma^2}{T} \|d\|^2$, and $T > 0$a parameter to be defined and $\gamma \geq \gamma^* \geq 0$, where γ^* is the smallest γ for which the system is stabilized [4]. For feedback policies $u(x)$ and disturbance policies $d(x)$, define the value or cost of the policies as

$$V(x(t),u,d) = \int_t^\infty \left(Q(x) + u^T \tfrac{R}{T} u - \tfrac{\gamma^2}{T} \|d\|^2 \right) dt \tag{3}$$

When the value is finite, a differential equivalent to this is the nonlinear Lyapunov-like equation

$$\begin{aligned} 0 &= r(x,u,d,T) \\ &+ (\nabla V)^T \left(f(x) + g(x)u(x) + k(x)d(x) \right), \ V(0) = 0 \end{aligned} \tag{4}$$

where $\nabla V = \partial V / \partial x \in R^n$ is the (transposed) gradient and the Hamiltonian is

$$\begin{aligned} H(x, \nabla V, u, d) &= r(x,u,d,T) \\ &+ (\nabla V)^T \left(f(x) + g(x)u(x) + k(x)d \right) \end{aligned} \tag{5}$$

For feedback policies [3], a solution $V(x) \geq 0$ to (4) is the value (3) for given feedback policy $u(x)$ and disturbance policy $d(x)$.

2.1 Two-Player Zero-Sum Differential Games and Nash Equilibrium

Define the 2-player zero-sum differential game [2], [3], [25]
(6)

subject to the dynamical constraints (1). Thus, u is the minimizing player and d is the maximizing one. This 2-player optimal control problem has a unique solution if a game theoretic saddle point exists, i.e., if the Nash condition holds

$$\min_u \max_d J(x(0),u,d) = \max_d \min_u J(x(0),u,d) \tag{7}$$

To this game is associated the Hamilton-Jacobi-Isaacs (HJI) equation
(8)

Given a solution $V^*(x) \geq 0 : \mathbb{R}^n \to \mathbb{R}$ to the HJI (8), denote the associated control and disturbance by employing the stationarity conditions as

$$\frac{\partial H}{\partial u} = 0 \Rightarrow u^* = -\tfrac{1}{2} T R^{-1} g^T(x) \nabla V^* \tag{9}$$

$$\frac{\partial H}{\partial d} = 0 \Rightarrow d^* = \frac{1}{2\gamma^2} T k^T(x) \nabla V^* \tag{10}$$

and write
(11)

Note that global solutions to the HJI (11) may not exist. Moreover, if they do, they may not be smooth. See [3] for a discussion on viscosity solutions to the HJI. The HJI equation (11) may have more than one nonnegative local smooth solution $V(x) \geq 0$. A minimal nonnegative solution $V_a(x) \geq 0$ is one such that there exists no other nonnegative solution $V(x) \geq 0$ such that $V_a(x) \geq V(x) \geq 0$. Of the nonnegative solutions to the GARE, select the one corresponding to the stable invariant manifold of the Hamiltonian matrix. Then, the minimum nonnegative solution of the HJI is the one having this stabilizing GARE solution as its Hessian matrix evaluated at the origin [4].

It is shown in [3] that if $V^*(x)$ is the minimum nonnegative solution to the HJI (11) and (1) is locally detectable, then (9), (10) given in terms of $V^*(x)$ are in Nash equilibrium solution to the zero-sum game and $V^*(x)$ is its value.

2.2 Policy Iteration

The HJI equation (11) is usually intractable to solve directly. One can solve the HJI iteratively using one of several algorithms that are built on iterative solutions of the Lyapunov equation (4). Included are [4] which uses an inner loop with iterations on the control, and [6] which uses an inner loop with iterations on the disturbance. These are in effect extensions of Kleinman's algorithm [27] to nonlinear 2-player games. Here, we shall use the latter algorithm.

$$J(x(0),u,d) = \int_0^\infty \left(Q(x) + u^T \tfrac{R}{T} u - \tfrac{\gamma^2}{T} \|d\|^2 \right) dt \equiv \int_0^\infty r(x,u,d,T)\, dt \tag{2}$$

$$V^*(x(0)) = \min_u \max_d J(x(0),u,d) = \min_u \max_d \int_0^\infty \left(Q(x) + u^T \tfrac{R}{T} u - \tfrac{\gamma^2}{T} \|d\|^2 \right) dt \tag{6}$$

Policy Iteration (PI) Algorithm for 2-Player Zero-Sum Differential Games

Initialization: Start with a stabilizing admissible control policy u_0

1. For $j = 0,1,...$ given u_j
2. For $i = 0,1,...$ set $d^0 = 0$,
   solve for $V_j^i(x(t))$, d^{i+1} using (12), (13)

   On convergence, set $V_{j+1}(x) = V_j^i(x)$
3. Update the control policy using (14)

Go to 1.

Note that this algorithm relies on successive solutions of nonlinear Lyapunov-like equations (12). As such, the discussion surrounding (4) shows that the algorithm finds the value $V_j^i(x(t))$ of successive control/disturbance policy pairs.

3 Online Solution of Zero Sum Games with Integral Reinforcement Learning

The online solution of zero-sum games in real time cannot be accomplished by simply throwing in the standard NN structures and adaptation approaches. E.g., for one thing, approximation is required of both the value function *and its gradient*. Second, one requires learning of the cost or value associated with the current control and disturbance policies. Therefore, in this section we first carefully develop proper approximator structures which lead to solution of the problem.

3.1 Value Function Approximation

A practical method for implementing PI for continuous-time systems involves two aspects: value function approximation (VFA) and integral reinforcement learning (IRL). This section discusses VFA, and the next presents IRL. In VFA, the Critic value and the Actor control function are approximated by neural networks, and the PI algorithm consists in tuning alternatively each of the three neural networks.

It is important to approximate $V(x)$ in Sobolev space, since both the value $V(x)$ and its gradient must be approximated. This machinery is provided by the Weierstrass higher-order [7] approximation theorem. Thus, assume there exist a weight parameter matrix W_1 such that the value $V(x)$ is approximated by a neural network as

$$V(x) = W_1^T \phi(x) + \varepsilon(x) \tag{15}$$

where $\phi(x) : \mathbb{R}^n \to \mathbb{R}^N$ is the activation function vector, N the number of neurons in the hidden layer, and $\varepsilon(x)$ the NN approximation error. It is known that $\varepsilon(x)$ is bounded by a constant on a compact set. Select the activation functions to provide a *complete* basis set such that $V(x)$ and its derivative

$$\frac{\partial V}{\partial x} = \nabla\phi^T W_1 + \frac{\partial \varepsilon}{\partial x} \tag{16}$$

are uniformly approximated. According to the Weierstrass higher-order approximation theorem [7], such a basis exists if $V(x)$ is sufficiently smooth. Then, as the number of hidden-layer neurons $N \to \infty$, the approximation error $\varepsilon \to 0$ uniformly.

3.2 Integral Reinforcement Learning

The PI algorithm given above requires full system dynamics, since $f(x), g(x), k(x)$appear in the Bellman equation (12). In order to find an equivalent formulation of the Bellman equation that does not involve the dynamics, we note that for any time t_0 and time interval *T ¿ 0* the value function (3) satisfies (17)

In [20] it is shown that (17) and (12) are equivalent, i.e., they both have the same solution. Therefore, (17) can be viewed as a Bellman equation for CT systems. Note that this form does not involve the system dynamics. We call this the integral re-

$$0 = Q(x) + \nabla V^T(x)f(x) - \frac{1}{4}\nabla V^T(x)g(x)TR^{-1}g^T(x)\nabla V(x) \quad + \frac{1}{4\gamma^2}\nabla V^T(x)Tkk^T\nabla V(x), \qquad V(0) = 0 \quad (8)$$

$$0 = H(x, \nabla V, u^*, d^*) = Q(x) + \nabla V^T(x)f(x) - \frac{1}{4}\nabla V^T(x)g(x)TR^{-1}g^T(x)\nabla V(x) + \frac{1}{4\gamma^2}\nabla V^T(x)Tkk^T\nabla V(x) \quad (11)$$

inforcement learning (IRL) form of the Bellman equation.

Therefore, by using a critic NN for VFA, the Bellman error based on (17) becomes (18)

where the parameter in the control weighting term is selected as the time *T>0.* We define the integral reinforcement as

$$p(t) = \int_{t-T}^{t} \left(Q(x) + u^T \tfrac{R}{T} u - \tfrac{\gamma^2}{T} \|d\|^2 \right) d\tau \quad (19)$$

Now (18) can be written as

$$\varepsilon_B - p = W_1^T \Delta\phi(x(t)) \quad (20)$$

where

$$\Delta\phi(x(t)) \equiv \phi(x(t)) - \phi(x(t-T)). \quad (21)$$

Under the Lipschitz assumption on the dynamics, the residual error ε_B is bounded on a compact set.

Remark 1. Note that, as $N \to \infty$, $\varepsilon_B \to 0$ uniformly [1].

3.3 Online Integral Reinforcement Learning Algorithm for Zero Sum Games

Standard PI algorithms for CT systems are offline methods that require complete knowledge on the system dynamics to obtain the solution (*i.e.* the functions $f(x), g(x), k(x)$ in Bellman equation (12) need to be known). It is desired to change the offline character of PI for CT systems and implement it online in real-time as in adaptive control mechanisms. Therefore, we present an adaptive learning algorithm that uses simultaneous continuous-time tuning for the actor and critic neural networks and does not need the drift term $f(x)$ in the dynamics. We term this the *online integral reinforcement learning algorithm for zero sum games*.

3.3.1 Critic Neural Network

The weights of the critic NN,W_1 that solve (18) are unknown. The output of the critic neural network is

$$\hat{V}(x) = \hat{W}_1^T \phi(x) \quad (22)$$

where $\hat{W}_1$ are the current known values of the critic NN weights. Recall that $\phi(x) : \mathbb{R}^n \to \mathbb{R}^N$ is the activation functions vector, with N the number of neurons in the hidden layer. The approximate Bellman error is then (23)

$$\int_{t-T}^{t} \left(Q(x) + u^T \tfrac{R}{T} u - \tfrac{\gamma^2}{T} \|d\|^2 \right) d\tau$$
$$+\hat{W}_1^T \phi(x(t)) - \hat{W}_1^T \phi(x(t-T)) = e_1 \quad (23)$$

which according to (19) can be written as

$$\hat{W}_1^T \Delta\phi(x(t)) = e_1 - p \quad (24)$$

It is desired to select $\hat{W}_1$ to minimize the squared residual error

$$E_1 = \tfrac{1}{2} e_1^T e_1 \quad (25)$$

Then$\hat{W}_1(t) \to W_1$. We select the tuning law for the critic weights as the normalized gradient descent algorithm

$$\dot{\hat{W}}_1 = -a_1 \frac{\Delta\phi(x(t))^T}{(1 + \Delta\phi(x(t))^T \Delta\phi(x(t)))^2} \bullet$$
$$[\int_{t-T}^{t} \left(Q(x) + u^T \tfrac{R}{T} u - \tfrac{\gamma^2}{T} \|d\|^2 \right) d\tau + \Delta\phi(x(t))^T \hat{W}_1] \quad (26)$$

Note that the data required in this tuning algorithm at each time are $(\Delta\phi(t), p(t))$. The system dynamics $f(x), g(x)$ are not needed. Note for future use that

$$\begin{aligned} \Delta\phi(t) &\equiv \Delta\phi(x(t)) = \textstyle\int_{t-T}^{t} \nabla\phi(x)\dot{x} d\tau \\ &= \textstyle\int_{t-T}^{t} \nabla\phi(f + gu + kd)\, d\tau = \int_{t-T}^{t} \sigma_1\, d\tau \end{aligned} \quad (27)$$

Define the critic weight estimation error $\tilde{W}_1 = W_1 - \hat{W}_1$ and substitute (18) in (26). Then, with the notation $\Delta\bar{\phi}(t) = \Delta\phi(t)/(\Delta\phi(t)^{\mathrm{T}}\Delta\phi(t) + 1)$ and $m_s =$

$$0 = Q(x) + \nabla V_j^{iT}(x)(f + gu_j + kd^i) + u_j^T Ru_j - \gamma^2 \left\|d^i\right\|^2 \tag{12}$$

$$d^{i+1} = \underset{d\in\Psi(\Omega)}{\arg\max}[H(x,\nabla V_j^i,u_j,d)] = \frac{1}{2\gamma^2}k^T(x)\nabla V_j^i \tag{13}$$

$$u_{j+1} = \underset{u\in\Psi(\Omega)}{\arg\min}[H(x,\nabla V_{j+1}),u,d] = -\tfrac{1}{2}R^{-1}g^T(x)\nabla V_{j+1} \tag{14}$$

$1+\Delta\phi(t)^T\Delta\phi(t)$, we obtain the dynamics of the critic weight estimation error as

$$\dot{\tilde{W}}_1 = -a_1\Delta\bar{\phi}(t)\,\Delta\bar{\phi}(t)^{\mathrm{T}}\tilde{W}_1 + a_1\Delta\bar{\phi}(t)\,\frac{\varepsilon_B}{m_s} \tag{28}$$

Though it is traditional to use critic tuning algorithms of the form (26), it is not generally understood when convergence of the critic weights can be guaranteed. In this paper, we address this issue in a formal manner. To guarantee convergence of $\hat{W}_1$ to W_1, the next Persistence of Excitation (PE) assumption is required. Note from (24) that the regression vector $\Delta\phi(t)$, or equivalently the normalized vector $\Delta\bar{\phi}(t)$, must be persistently exciting to solve for $\hat{W}_1$in a least squares sense.

Persistence of Excitation (PE) Assumption. Let the signal $\Delta\bar{\phi}(t)$ be persistently exciting over the interval $[t-T_{PE},t]$, *i.e.* there exist constants $\beta_1>0$, $\beta_2>0$, $T_{PE}>0$ such that, for all t,

$$\beta_1\mathrm{I} \le S_0 \equiv \int_{t-T_{PE}}^{t}\Delta\bar{\phi}(\tau)\Delta\bar{\phi}^{\mathrm{T}}(\tau)d\tau \le \beta_2\mathrm{I} \tag{29}$$

Technical Lemma 1. Consider the error dynamics (28) with output

$$y_1 = \Delta\bar{\phi}(t)^{\mathrm{T}}\tilde{W}_1$$

Assume $\Delta\bar{\phi}(t)$ is PE according to (29).

Let$||\varepsilon_B|| \le \varepsilon_{\max}$and $||y_1|| \le y_{\max}$. Then $||\tilde{W}_1||$converges exponentially to the residual set

$$\tilde{W}_1(t) \le \frac{\sqrt{\beta_2 T_{PE}}}{\beta_1}\left\{[y_{\max} + \delta\beta_2 a_1(\varepsilon_{\max}+y_{\max})]\right\}.$$

where δ is a positive constant of the order of 1.

Proof: [14].

3.3.2 Action and Disturbance Neural Networks

The policy improvements step in PI are given approximately as

$$u(x) = -\tfrac{1}{2}TR^{-1}g^T(x)\nabla\phi^T W_1 \tag{30}$$

$$d(x) = \tfrac{1}{2\gamma^2}Tk^T(x)\nabla\phi^T W_1 \tag{31}$$

with critic weights W_1 unknown. Therefore, define the control and disturbance policy in the form of action neural networks which compute the control and the disturbance input in the structured form

$$u_2(x) = -\tfrac{1}{2}TR^{-1}g^T(x)\nabla\phi^T\hat{W}_2 \tag{32}$$

$$d_3(x) = \tfrac{1}{2\gamma^2}Tk^T(x)\nabla\phi^T\hat{W}_3 \tag{33}$$

where $\hat{W}_2$, $\hat{W}_3$ denote the current known values of the actor and disturbance NN weights respectively.

Based on (30), (31) and (18), define the approximate HJI equation (34) with the notations
$\bar{D}_1(x) = \nabla\phi(x)g(x)TR^{-1}g^T(x)\nabla\phi^T(x)$,
$\bar{E}_1(x) = \frac{T}{\gamma^2}\nabla\phi(x)k(x)k^T(x)\nabla\phi^T(x)$
where W_1 denotes the ideal unknown weights of the critic, actor and disturbance neural networks which solve the HJI. The error $\varepsilon_{HJI}(x)$ has components arising from the NN approximation error and its gradient. We now present the main results, which provide tuning laws for the actor, disturbance and critic neural networks that guarantee convergence to the Nash solution of the game with closed-loop stability. The next notion of practical stability is needed.

Definition 2. [26] (UUB) A time signal $\zeta(t)$is said to be uniformly ultimately bounded (UUB) if there exists a compact set $S\subset\mathbb{R}^n$so that for all $\zeta(0)\in S$there exists a bound B and a time $T_B(B,\zeta(0))$ such that $\|\zeta(t)\|\le B$ for all $t\ge t_0+T_B$.

Definition 3. [27] A continuous function $\alpha:[\alpha,0)\to[0,\infty)$is said to belong to class Kif it is strictly increasing and $\alpha(0)=0$. It is said to belong to class K_∞ if $\alpha=\infty$and $\alpha(r)=\infty$ as $r\to\infty$.

Facts 1. For a given compact set $\Omega\subset\mathbb{R}^n$:

1. *f(.)* is Lipschitz so that $\|f(x)\|\le b_f\|x\|$

$$V(x_{t_0}) = \int_{t_0-T}^{t_0} r(x(\tau), u(x(\tau)), d(x(\tau)))d\tau + V(x_{t_0-T}) \tag{17}$$

$$\int_{t-T}^{t} \left(Q(x) + u^T \tfrac{R}{T} u - \tfrac{\gamma^2}{T} \|d\|^2 \right) d\tau + W_1^T \phi(x(t)) - W_1^T \phi(x(t-T)) \equiv \varepsilon_B \tag{18}$$

2. *g(.), k(.)* are bounded by constants:

$$\|g(x)\| < b_g, \quad \|k(x)\| < b_k$$

3. The NN approx error and its gradient are bounded so that

$$\|\varepsilon_1\| < b_{\varepsilon 1}, \|\nabla\varepsilon_1\| < b_{\varepsilon_{1x}}$$

4. The NN activation function and its gradients are bounded so that

$$\|\phi(x)\| < b_\phi, \ \|\nabla\phi(x)\| < b_{\phi_x}$$

5. The critic NN weights are bounded by a constant

$\|W_1\| < W_{\max} \, \|\phi_1(x)\| < b_{\phi_1},$
$\|\nabla\phi_1(x)\| < b_{\phi_{1x}},$
$\|\phi_2(x)\| < b_{\phi_2},$
$\|\nabla\phi_2(x)\| < b_{\phi_{2x}}$

Theorem 1. Adaptive Tuning Algorithm for zero sum games. Let the system dynamics be given by (1), tuning for the critic NN be provided by (35) where (36) and assume that $\Delta\bar{\phi}_2(t)$ is persistently exciting (which means u_2, $d_3$are persistently exciting). Let the actor NN be tuned as (37) and the disturbance NN be tuned as (38) where $F_1 > 0$, $F_2 > 0$, $F_3 > 0$, $F_4 > 0$ are tuning parameters chosen as in the proof. Then there exists a N_0 and a time T_0 such that, for the number of hidden layer units $N > N_0$ and the time interval $T < T_0$, the closed-loop system state, the critic NN error $\tilde{W}_1$, the actor NN error $\tilde{W}_2$, and the disturbance NN error $\tilde{W}_3$are UUB.

Proof: See appendix.

Remark 2. Note that the data required in the critic tuning algorithm (35) at each time are $\Delta\phi_2(t)$ and the integral reinforcement. The system dynamics $f(x), g(x), k(x)$ are not needed. The input coupling dynamics $g(x), k(x)$ are needed for the actor and disturbance tuning algorithm (37).

Remark 3. The tuning parameters F_1,F_2,F_3,F_4 are selected appropriately to ensure stability as detailed in the proof of Theorem 1.

Remark 4. The proof reveals that the time interval T cannot be selected too large nor the number of hidden layer units N too small.

Remark 5. The assumption $Q(x) > 0$ is sufficient but not necessary for this result

Theorem 2. Nash Solution. Suppose the hypotheses of Theorem 1, hold. Then:

1. $H(x, \hat{W}_1, \hat{u}_1, \hat{d}_1)$ is UUB. That is, $\hat{W}_1$ converges to the approximate HJI solution, the value of the ZS game. Where

$$\hat{u}_1 = -\frac{1}{2} R^{-1} g^T(x) \nabla\phi_1^T(x) \hat{W}_1 \tag{39}$$

$$\hat{d}_1 = \frac{1}{2\gamma^2} k^T(x) \nabla\phi_1^T(x) \hat{W}_1 \tag{40}$$

2. $u_2(x), d_3(x)$ (see (32) and (33)) converges to the approximate Nash equilibrium solution of the ZS game.

Proof: See [19].

Remark 6. The theorems show that PE is needed for proper identification of the value function by the critic NN.

4 Simulations

To support the new online algorithm for zero sum games we offer two simulation examples, one linear and one nonlinear. In both cases we observe convergence to the Nash solution of the game without knowing the system drift dynamics. In these simulations, exponentially decreasing noise is added to the control and disturbance inputs to ensure PE until convergence is obtained.

4.1 Linear System

Consider the continuous-time F16 aircraft plant with quadratic cost function used in [28]. The system state vector is $x = [\ \alpha \quad q \quad \delta_e\ ]$, where α denotes the angle of attack, q is the pitch rate and δ_e is the elevator deflection angle. The control input is

$$\int_{t-\mathrm{T}}^{t}\left(-Q(x)-\frac{1}{4}W_1^T\overline{D}_1(x)W_1+\frac{1}{4}W_1^T\bar{E}_1(x)W_1+\varepsilon_{HJI}(x)\right)d\tau \ = W_1^T\Delta\phi(x(t)) \tag{34}$$

$$\dot{\hat{W}}_1=-a_1\tfrac{\Delta\phi_2(t)}{(\Delta\phi_2(t)^T\Delta\phi_2(t)+1)^2}\left(\Delta\phi_2(t)^T\hat{W}_1+\int_{t-T}^{t}\left(Q(x)+\tfrac{1}{4}\hat{W}_2^T\overline{D}_1\hat{W}_2-\tfrac{1}{4}\hat{W}_3^T\bar{E}_1\hat{W}_3\right)d\tau\right) \tag{35}$$

$$\Delta\phi_2(x(t))=\int_{t-T}^{t}\nabla\phi(f+gu_2+kd_3)\,d\tau=\int_{t-T}^{t}\sigma_2\,d\tau\equiv\phi(x(t))-\phi(x(t-T))\equiv\Delta\phi_2(t) \tag{36}$$

$$\dot{\hat{W}}_2=-a_2\left\{\left(F_2\hat{W}_2-F_1T\Delta\bar{\phi}_2^T\hat{W}_1\right)-\tfrac{1}{4m_s}\overline{D}_1(x)\hat{W}_2\Delta\bar{\phi}_2^T\hat{W}_1\right\} \tag{37}$$

$$\dot{\hat{W}}_3=-a_3\left\{\left(F_4\hat{W}_3-F_3T\Delta\bar{\phi}_2^T\hat{W}_1\right)+\frac{1}{4m_s}\bar{\mathrm{E}}_1(x)\hat{W}_3\Delta\bar{\phi}_2^T\hat{W}_1\right\} \tag{38}$$

$$f(x)=\begin{bmatrix}-x_1+x_2\\ -x_1^3-x_2^3+0.25x_2(\cos(10x_1)+2)^2-0.25x_2\frac{1}{\gamma^2}(\sin(x_1)+2)^2\end{bmatrix} \tag{41}$$

$$\dot{x}=\begin{bmatrix}-1.01887 & 0.90506 & -0.00215 & & \\ 0.82225 & -1.07741 & -0.175550 & 0 & -1\end{bmatrix}x+\begin{bmatrix}0\\0\\1\end{bmatrix}u+\begin{bmatrix}1\\0\\0\end{bmatrix}d \tag{42}$$

the elevator actuator voltage and the disturbance is wind gusts on angle of attack.

One has the dynamics $\dot{x}=Ax+Bu+Kd$,(42) where Q and R in the cost function are identity matrices of appropriate dimensions and $\gamma=5$. Also $a_1=10$, $a_2=a_3=1$, $F_1=I$, $F_2=10I$, $F_3=I$, $F_4=10I$ where I is an identity matrix of appropriate dimensions and $T=0.01$. In this linear case the solution of the HJI equation is given by the solution of the game algebraic Riccati equation (GARE) [25]. Solving the GARE gives the parameters of the optimal critic as $W_1^*=[1.6573\ 1.3954\ -0.1661\ 1.6573\ -0.1804\ 0.4371]^{\mathrm{T}}$ which are the components of the Riccati solution matrix P.

The online integral reinforcement zero-sum game algorithm is implemented as in Theorem 1. Figure 1 shows the critic parameters, denoted by $\hat{W}_1=[\ W_{c1}\ \ W_{c2}\ \ W_{c3}\ \ W_{c4}\ \ W_{c5}\ \ W_{c6}\ ]^T$ converging to the optimal values. In fact after 300s the critic parameters converged to $\hat{W}_1(t_f)=[1.7408\ 1.2247\ -0.2007\ 1.5247\ -0.1732\ 0.4585]^T$ The actor and disturbance parameters after 300s converge to the values of

$$\hat{W}_3(t_f)=\hat{W}_2(t_f)=\hat{W}_1(t_f).$$

Then, the actor NN is given as

$$\hat{u}_2(x)=-\tfrac{T}{2}R^{-1}\begin{bmatrix}0\\0\\1\end{bmatrix}^T\nabla\phi_1^T(x)\hat{W}_2(t_f).$$

Then, the disturbance NN is given as

$$\hat{d}(x)=\tfrac{T}{2\gamma^2}\begin{bmatrix}0\\0\\1\end{bmatrix}^T\nabla\phi_1^T(x)\hat{W}_3(t_f)$$

The evolution of the system states is presented in Figure 2. One can see that after 300s convergence of the NN weights in critic, actor and disturbance has occurred.

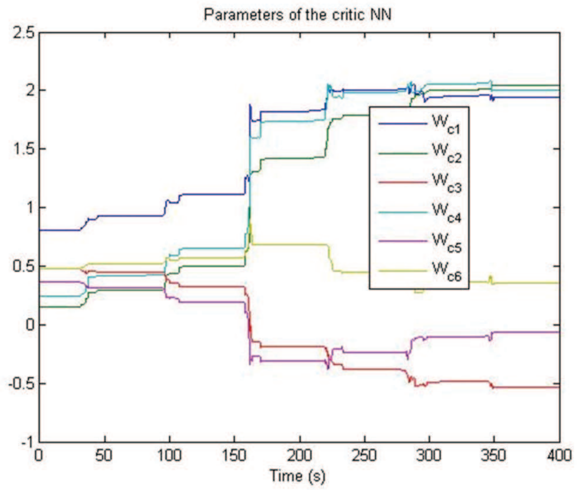

Figure 1. Convergence of the critic parameters to the parameters of the optimal critic.

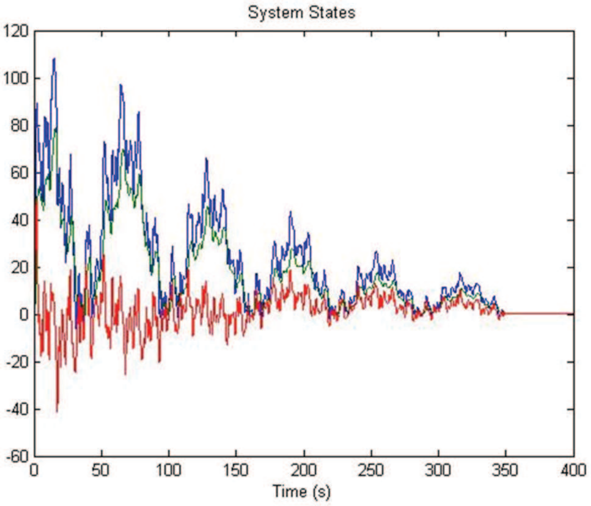

Figure 2. Evolution of the states.

4.2 Nonlinear System

Consider the following affine in control and disturbance inputs nonlinear system, with a quadratic cost constructed as in [29]

$$\dot{x} = f(x) + g(x)u + k(x)d,\ x \in \mathbb{R}^2$$

where (41)

$$g(x) = \begin{bmatrix} 0 \\ \cos(10x_1)+2 \end{bmatrix},$$

$$k(x) = \begin{bmatrix} 0 \\ (\sin(x_1)+2) \end{bmatrix}.$$

One selects $Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $R = 1$, $\gamma = 8$. Also $a_1 = 50, a_2 = a_3 = 1$, $F_1 = I$, $F_2 = 10I$, $F_3 = I$, $F_4 = 10I$ where I is an identity matrix of appropriate dimensions and $T = 0.01$.

The optimal value function is

$$V^*(x) = \frac{1}{4}x_1^4 + \frac{1}{2}x_2^2$$

the optimal control signal is

$$u^*(x) = -\tfrac{T}{2}(\cos(10x_1)+2)x_2$$

and

$$d^*(x) = \frac{T}{2\gamma^2}(\sin(x_1)+2)x_2$$

One selects the critic NN vector activation function as

$$\varphi_1(x) = [x_1^2 \quad x_2^2 \quad x_1^4\, x_2^4]$$

Figure 3 shows the critic parameters, denoted by

$$\hat{W}_1 = [\ W_{c1} \quad W_{c2} \quad W_{c3} \quad W_{c4}]^T$$

by using the synchronous zero-sum game algorithm. After convergence at about 50s have

$$\hat{W}_1(t_f) = [0.0036 \quad 0.5045 \quad 0.2557 \quad 0.0006]^T$$

The actor and disturbance parameters after 80s converge to the values of

$$\hat{W}_3(t_f) = \hat{W}_2(t_f) = \hat{W}_1(t_f).$$

So that the actor NN

$$\hat{u}_2(x) = -\tfrac{T}{2}R^{-1}\begin{bmatrix} 0 \\ \cos(2x_1)+2 \end{bmatrix}^T \nabla\phi_1^T(x)\hat{W}_2(t_f)$$

also converged to the optimal control, and the disturbance NN

$$\hat{d}(x) = \tfrac{T}{2\gamma^2}\begin{bmatrix} 0 \\ \sin(4x_1)+2 \end{bmatrix}^T \nabla\phi_1^T(x)\hat{W}_3(t_f)$$

also converged to the optimal disturbance.

The evolution of the system states is presented in Figure 4.

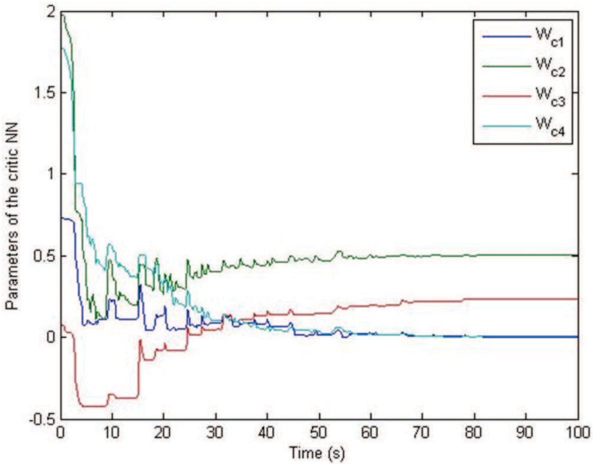

Figure 3. Convergence of the critic parameters to the parameters of the optimal critic.

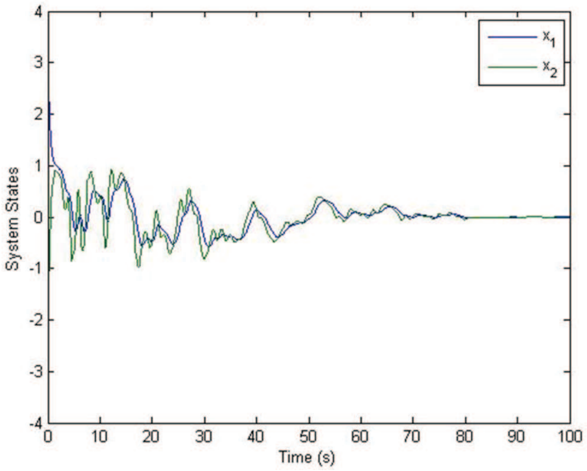

Figure 4. Evolution of the states.

5 Conclusion

In this paper we have proposed a new adaptive algorithm which solves the continuous-time zero-sum game problem for nonlinear systems. The importance of this algorithm relies on the partial need of dynamics, only $g(x)$ is needed, the simultaneous tuning of the actor, disturbance and critic neural networks and the convergence to HJI and saddle point without solving these equations.

Appendix

Proof for Theorem 1: Let $\tilde{W}_1 = W_1 - \hat{W}_1$, $\tilde{W}_2 = W_1 - \hat{W}_2$ and $\tilde{W}_3 = W_1 - \hat{W}_3$denote the errors between the weights. We consider the Lyapunov function candidate (A.1) The derivative of the Lyapunov function is given by(A.2).

Next we will evaluate each one of the three terms of $\dot{L}(x)$. The first term is $\dot{L}_V(x) = \int_{t-T}^{t} \dot{V}(x(\tau))d\tau = \int_{t-T}^{t} (W_1^T \nabla\phi_1(x)\dot{x} + \dot{\varepsilon}(x))d\tau$ With the control computed by the actor approximator (32) the system dynamics are given by

$$\dot{x} = f(x) - \tfrac{1}{2} g(x) T R^{-1} g^T(x) \nabla\phi_1^T(x)\hat{W}_2 \\ + \tfrac{1}{2}\tfrac{T}{\gamma^2} kk^T \nabla\phi_1^T(x)\hat{W}_3$$

The first term in (A.2) becomes (A.3).

Now we want to obtain a representation of $\dot{L}_V(x)$ in terms of the parameters of the optimal value functionW_1, and the parameter errors $\tilde{W}_1$, $\tilde{W}_2$ and $\tilde{W}_3$. Thus, by adding and subtracting$\tfrac{1}{2} W_1^T \bar{D}_1(x) W_1$, we obtain (A.3.1) and using the notation (A.3.2). From the HJI equation (34) one has (A.4) then (A.5). Using the tuning law for the critic, the second term in (A.1) becomes (A.5.1) Adding (A.3) to the integral in the right hand side, using the notation $m_s = \Delta\phi_2(t)^T \Delta\phi_2(t) + 1$and (36) we obtain (A.6)

Using (27) and (36) the first integral in (A.6) becomes (A6.1)

Then (A.5) becomes

$$\dot{L}_1 = \tilde{W}_1^T \tfrac{\Delta\phi_2(x(t),T)}{m_s^2} \\ \left(\int_{t-T}^{t} \left(-\tfrac{1}{2}\hat{W}_2^T \bar{D}_1(x)\hat{W}_1 + \tfrac{1}{2}\hat{W}_3^T \bar{E}_1(x)\hat{W}_1 \right.\right. \\ \left. + \tfrac{1}{4} W_1^T \bar{D}_1(x) W_1 - \tfrac{1}{4} W_1^T \bar{E}_1(x) W_1 \right) d\tau \\ + \int_{t-T}^{t} \left(\tfrac{1}{4}\hat{W}_2^T \bar{D}_1(x)\hat{W}_2 - \tfrac{1}{4}\hat{W}_3^T \bar{E}_1(x)\hat{W}_3 \right. \\ \left.\left. - \tilde{W}_1^T \nabla\phi_1(x) f(x) \; + \varepsilon_{HJI}(x)\right) d\tau\right)$$

Using the definition for the parameter error $W_1 = \hat{W}_2 + \tilde{W}_2$, $W_1 = \hat{W}_3 + \tilde{W}_3$the first six terms under the integral can be written as

$$-\tfrac{1}{4} W_1^T \bar{E}_1(x) W_1 + \tfrac{1}{2}\hat{W}_3^T \bar{E}_1(x)\hat{W}_1 - \tfrac{1}{4}\hat{W}_3^T \bar{E}_1(x)\hat{W}_3 \\ \tfrac{1}{4} W_1^T \bar{D}_1(x) W_1 - \tfrac{1}{2}\hat{W}_2^T \bar{D}_1(x)\hat{W}_1 + \tfrac{1}{4}\hat{W}_2^T \bar{D}_1(x)\hat{W}_2 \\ = \tfrac{1}{4}(\hat{W}_2 + \tilde{W}_2)^T \bar{D}_1(x)(\hat{W}_2 + \tilde{W}_2) \\ - \tfrac{1}{4}(\hat{W}_3 + \tilde{W}_3)^T \bar{E}_1(x)(\hat{W}_3 + \tilde{W}_3) \\ - \tfrac{1}{2}\hat{W}_2^T \bar{D}_1(x)\hat{W}_1 + \tfrac{1}{4}\hat{W}_2^T \bar{D}_1(x)\hat{W}_2 \\ + \tfrac{1}{2}\hat{W}_3^T \bar{E}_1(x)\hat{W}_1 - \tfrac{1}{4}\hat{W}_3^T \bar{D}_1(x)\hat{W}_3$$

Developing the parenthesis, and making use of the definition $W_1 = \hat{W}_1 + \tilde{W}_1$ we obtain

$$L(t) = \int_{t-T}^{t} V(x(\tau))d\tau + \tfrac{1}{2}\tilde{W}_1^T(t)a_1^{-1}\tilde{W}_1(t) + \tfrac{1}{2}\int_{t-T}^{t} \tilde{W}_2^T(\tau)a_2^{-1}\tilde{W}_2(\tau)d\tau + \tfrac{1}{2}\int_{t-T}^{t} \tilde{W}_3^T(\tau)a_3^{-1}\tilde{W}_3(\tau)d\tau \triangleq\triangleq L_V(x) + L_1(x) + L_2(x) + L_3(x) \tag{A.1}$$

$$\dot{L}(x) = \int_{t-T}^{t} \dot{V}(x(\tau))d\tau + \tilde{W}_1^T(t)a_1^{-1}\dot{\tilde{W}}_1(t) + \int_{t-T}^{t} \tilde{W}_2^T(\tau)a_2^{-1}\dot{\tilde{W}}_2(\tau)d\tau + \int_{t-T}^{t} \tilde{W}_3^T(\tau)a_3^{-1}\dot{\tilde{W}}_3(\tau)d\tau \tag{A.2}$$

$$\begin{gathered}\dot{L}_V(x) = \int_{t-T}^{t} W_1^T\left(\nabla\phi_1(x)f(x) - \tfrac{1}{2}\overline{D}_1(x)\hat{W}_2 + \tfrac{1}{2}\bar{E}_1(x)\hat{W}_3\right)d\tau + \varepsilon_1(x) \\ \text{where } \varepsilon_1(x(t)) = \int_{t-T}^{t} \nabla\varepsilon^T(x)\,(f(x) - \tfrac{1}{2}g(x)R^{-1}g^T(x)\nabla\phi_1^T\hat{W}_2 + \tfrac{1}{2\gamma^2}kk^T\nabla\phi_1^T\hat{W}_3)d\tau\end{gathered} \tag{A.3}$$

$$\begin{gathered}\dot{L}_V(x) = \int_{t-T}^{t}\left(W_1^T\nabla\phi_1 f(x) + \tfrac{1}{2}W_1^T\overline{D}_1(x)\left(W_1 - \hat{W}_2\right)\right. \\ \left.-\tfrac{1}{2}W_1^T\bar{E}_1(x)\left(W_1 - \hat{W}_3\right) - \tfrac{1}{2}W_1^T\overline{D}_1(x)W_1 + \tfrac{1}{2}W_1^T\bar{E}_1(x)W_1\right)d\tau + \varepsilon_1(x) \\ = \int_{t-T}^{t}\left(W_1^T\nabla\phi_1 f(x) + \tfrac{1}{2}W_1^T\overline{D}_1(x)\tilde{W}_2 - \tfrac{1}{2}W_1^T\bar{E}_1(x)\tilde{W}_3\right. \\ \left.-\tfrac{1}{2}W_1^T\overline{D}_1(x)W_1 + \tfrac{1}{2}W_1^T\bar{E}_1(x)W_1\right)d\tau + \varepsilon_1(x)\end{gathered} \tag{A.3.1}$$

$$\sigma_1(x) = \nabla\phi_1 f(x) - \frac{1}{2}\overline{D}_1(x)W_1 + \frac{1}{2}\bar{E}_1(x)W_1$$

$$\dot{L}_V(x) = \int_{t-\mathrm{T}}^{t}\left(W_1^T\sigma_1 + \tfrac{1}{2}W_1^T\overline{D}_1(x)\tilde{W}_2 - \tfrac{1}{2}W_1^T\bar{E}_1(x)\tilde{W}_3\right)d\tau + \varepsilon_1(x) \tag{A.3.2}$$

$$\int_{t-T}^{t} W_1^T\sigma_1 d\tau = \int_{t-\mathrm{T}}^{t}\left(-Q(x) - \tfrac{1}{4}W_1^T\overline{D}_1(x)W_1 + \tfrac{1}{4}W_1^T\bar{E}_1(x)W_1 + \varepsilon_{HJI}(x)\right)d\tau\,. \tag{A.4}$$

$$\begin{gathered}\dot{L}_V(x) = \int_{t-\mathrm{T}}^{t}\left(-Q(x) - \tfrac{1}{4}W_1^T\overline{D}_1(x)W_1 + \tfrac{1}{4}W_1^T\bar{E}_1(x)W_1 + \varepsilon_{HJI}(x)\right)d\tau \quad + \\ \int_{t-\mathrm{T}}^{t}\left(\tfrac{1}{2}W_1^T\bar{D}_1(x)\tilde{W}_2 - \tfrac{1}{2}W_1^T\bar{E}_1(x)\tilde{W}_3\right)d\tau + \varepsilon_1(x)\end{gathered} \tag{A.5}$$

$$\begin{gathered}\dot{L}_1 = \tilde{W}_1^T(t)a_1^{-1}\dot{\tilde{W}}_1(t) \\ = \tilde{W}_1^T\frac{\Delta\phi_2(t)}{(\Delta\phi_2(t)^T\Delta\phi_2(t)+1)^2}\{\Delta\phi_2(t)^T\hat{W}_1 + \int_{t-T}^{t}\left(Q(x) + \tfrac{1}{4}\hat{W}_2^T\overline{D}_1\hat{W}_2 - \tfrac{1}{4}\hat{W}_3^T\bar{E}_1\hat{W}_3\right)d\tau\}\end{gathered} \tag{A.5.1}$$

$$\begin{gathered}\dot{L}_1 = \tilde{W}_1^T\frac{\Delta\phi_2(t)}{m_s^2}\left(\int_{t-T}^{t}(\sigma_2^T(x)\hat{W}_1 - \sigma_1^T(x)W_1\ )d\tau + \int_{t-T}^{t}\left(\tfrac{1}{4}\hat{W}_2^T\bar{D}_1(x)\hat{W}_2 - \tfrac{1}{4}\hat{W}_3^T\bar{E}_1(x)\hat{W}_3\right.\right. \\ \left.\left.-\tfrac{1}{4}W_1^T\bar{D}_1(x)W_1 + \tfrac{1}{4}W_1^T\bar{E}_1(x)W_1 + \varepsilon_{HJI}(x)\right)d\tau\right)\end{gathered} \tag{A.6}$$

$$\begin{gathered}\int_{t-T}^{t}(\sigma_2^T(x)\,\hat{W}_1 - \sigma_1^T(x)W_1)d\tau = \\ = \int_{t-T}^{t}(-\tilde{W}_1^T\nabla\phi_1(x)f(x) - \tfrac{1}{2}\hat{W}_2^T\bar{D}_1(x)\hat{W}_1 + \tfrac{1}{2}\hat{W}_3^T\bar{E}_1(x)\hat{W}_1 + \tfrac{1}{2}W_1^T\bar{D}_1(x)W_1 - \tfrac{1}{2}W_1^T\bar{E}_1(x)W_1)d\tau\end{gathered} \tag{A.6.1}$$

$$\begin{aligned}
&\tfrac{1}{4}W_1^T\bar{D}_1(x)W_1\\
&-\tfrac{1}{2}\hat{W}_2^T\bar{D}_1(x)\hat{W}_1+\tfrac{1}{4}\hat{W}_2^T\bar{D}_1(x)\hat{W}_2\\
&-\tfrac{1}{4}W_1^T\bar{E}_1(x)W_1\\
&+\tfrac{1}{2}\hat{W}_3^T\bar{E}_1(x)\hat{W}_1-\tfrac{1}{4}\hat{W}_3^T\bar{E}_1(x)\hat{W}_3\\
&=\tfrac{1}{2}\hat{W}_2^T\bar{D}_1(x)\tilde{W}_1+\tfrac{1}{4}\tilde{W}_2^T\bar{D}_1(x)\tilde{W}_2\\
&-\tfrac{1}{2}\hat{W}_3^T\bar{E}_1(x)\tilde{W}_1-\tfrac{1}{4}\tilde{W}_3^T\bar{E}_1(x)\tilde{W}_3
\end{aligned}$$

Using this last relation and (36), we obtain (A.7). Inserting the results (A.5) and (A.7) in (A.2), and using the notation $\frac{\Delta\phi_2(x(t-T),T,\hat{u}_2,\hat{d}_3)}{m_s}=\Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3)$, equation (A.2) becomes (A7.1). Using the relation $\frac{\Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3)}{m_s}=\int_{t-T}^{t}\frac{1}{m_s}\bar{\sigma}_2(x(\tau))d\tau\equiv\bar{\Phi}_2(t)$ and making use of weight error definitions it can be written as (A.7.2).

Using the dynamics of the actor parameters (A.7.2.1) the dynamics of the disturbance parameters (A.7.2.2) and the weight error definitions and rearranging the terms, (A.2) becomes (A.7.3)

According to Facts 1, we can write (A.3) as (A.7.2.3)

Also since $Q(x)>0$ there exist $q$such that $x^Tqx<Q(x)$and for $x\in\Omega$.

Now (A.2) becomes (A.8).

Select $\varepsilon>0$ and $N_0(\varepsilon)$ such that$\sup\limits_{x\in\Omega}\|\varepsilon_{HJI}\|<\varepsilon$.Then, assuming $N>N_0$ we define

$$\tilde{Z}(t,\tau)=\begin{bmatrix} x(\tau)\\ \Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3)^T\tilde{W}_1(t)\\ \tilde{W}_2(\tau)\\ \tilde{W}_3(\tau)\end{bmatrix}$$

and

$$z(t,\tau)=\begin{bmatrix} \Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3)^T\tilde{W}_1(t)\\ \Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3)^T\tilde{W}_1(\tau)\\ \tilde{W}_2(\tau)\\ \tilde{W}_2(\tau)^T\bar{D}_1(x(\tau))\tilde{W}_2(\tau)\\ \tilde{W}_3(\tau)\\ \tilde{W}_3(\tau)^T\bar{E}_1(x(\tau))\tilde{W}_3(\tau)\end{bmatrix}$$

then (A.8) becomes (A.8.1) where (A.8.2)

$$W=\begin{bmatrix} 0 & 0 & W_{13} & W_{14} & W_{15} & W_{16}\\ 0 & 0 & W_{23} & W_{24} & W_{25} & W_{26}\\ W_{31} & W_{32} & 0 & 0 & 0 & 0\\ W_{41} & W_{42} & 0 & 0 & 0 & 0\\ W_{51} & W_{52} & 0 & 0 & 0 & 0\\ W_{61} & W_{62} & 0 & 0 & 0 & 0\end{bmatrix}$$

with (A.8.3) and

$$D_1(x)=\nabla\phi(x)g(x)R^{-1}g^T(x)\nabla\phi^T(x),$$

$$E_1(x)=\tfrac{1}{\gamma^2}\nabla\phi(x)k(x)k^T(x)\nabla\phi^T(x).$$

Also

$$M=\begin{bmatrix} M_{11} & 0 & 0 & 0\\ 0 & M_{22} & M_{23} & M_{24}\\ 0 & M_{32} & M_{33} & M_{34}\\ 0 & M_{42} & M_{43} & M_{44}\end{bmatrix}$$

with $M_{11}=qI$, $M_{22}=I$,

$$M_{32}=M_{23}^T=-\frac{1}{2}TF_1-\left(\frac{1}{8m_s(t)}\bar{D}_1(\tau)W_1(\tau)\right),$$

$$M_{42}=M_{24}^T=-\frac{1}{2}TF_3+\left(\frac{1}{8m_s(t)}\bar{E}_1(\tau)W_1(\tau)\right)$$

$$\begin{aligned}
&M_{33}=F_2\\
&-\tfrac{1}{4}(W_1(\tau)^T\bar{\Phi}_2(\tau)^T-W_1(t)^T\bar{\Phi}_2(t)^T)\bar{D}_1(x(\tau))\\
&-\tfrac{1}{8}\left(\bar{D}_1(\tau)W_1(t)m^{\mathrm{T}}(t)+m(t)W_1(t)^T\bar{D}_1(\tau)\right)
\end{aligned}$$

$$\begin{aligned}
&M_{44}=F_4\\
&+\tfrac{1}{4}(W_1(\tau)^T\bar{\Phi}_2(\tau)^T+W_1(t)^T\bar{\Phi}_2(t)^T)\bar{E}_1(x(\tau))\\
&+\tfrac{1}{8}\left(\bar{E}_1(\tau)W_1(t)m^{\mathrm{T}}(t)+m(t)W_1(t)^T\bar{E}_1(\tau)\right)
\end{aligned}$$

and

$$d=\begin{bmatrix} d_1\\ d_2\\ d_3\\ d_4\end{bmatrix}$$

with (A.8.4) and $m(t)\equiv\frac{\Delta\phi_2(t)}{(\Delta\phi_2(t)^T\Delta\phi_2(t)+1)^2}$

After taking norms and using the relations

$$\begin{aligned}
\|z\|^2&=\left\|\Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3)^T\tilde{W}_1(t)\right\|^2\\
&+\left\|\Delta\bar{\phi}_2(x(\tau-T),T,\hat{u}_2,\hat{d}_3)^T\tilde{W}_1(\tau)\right\|^2\\
&+\left\|\tilde{W}_2(\tau)\right\|^2+\left\|\tilde{W}_2(\tau)^T\bar{D}_1(x(\tau))\tilde{W}_2(\tau)\right\|^2\\
&+\left\|\tilde{W}_3(\tau)\right\|^2+\left\|\tilde{W}_3(\tau)^T\bar{E}_1(x(\tau))\tilde{W}_3(\tau)\right\|^2
\end{aligned}$$

and for appropriate selection of ρ_1, ρ_2 one has

$$\|z(t,\tau)\|\le\rho_1\left\|\tilde{Z}(t,\tau)\right\|+\rho_2T\left\|\tilde{Z}(t,\tau)\right\|^2$$

$$\dot{L}_1 = \tilde{W}_1^T \tfrac{\Delta\phi_2(x(t-T),T,\hat{u}_2,\hat{d}_3)}{m_s^2} \int_{t-T}^{t} \left(\tfrac{1}{4}\tilde{W}_2^T \bar{D}_1(x)\tilde{W}_2 - \tfrac{1}{4}\tilde{W}_3^T \bar{E}_1(x)\tilde{W}_3 - \sigma_2^T \tilde{W}_1 + \varepsilon_{HJI}(x)\right) d\tau \tag{A.7}$$

$$\begin{gathered}
\dot{L}(x) = \int_{t-\mathrm{T}}^{t} \left(-Q(x) - \tfrac{1}{4}W_1(\tau)^T \overline{D}_1(x(\tau))W_1(\tau) + \tfrac{1}{4}W_1(\tau)^T \overline{E}_1(x(\tau))W_1(\tau) + \varepsilon_{HJI}(x)\right) d\tau + \varepsilon_1(x) \\
-\tilde{W}_1(t)^T \Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3) \left(\Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3)\right)^T \tilde{W}_1(t) \\
+\tilde{W}_1(t)^T \Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3) \int_{t-T}^{t} \varepsilon_{HJI}(x(\tau))d\tau \\
+ \int_{t-\mathrm{T}}^{t} \left(\tfrac{1}{2}W_1(\tau)^T \bar{D}_1(x)\tilde{W}_2(\tau)\right) d\tau + \int_{t-T}^{t} \tilde{W}_2^T(\tau)\alpha_2^{-1}\dot{\tilde{W}}_2(\tau)d\tau + \int_{t-T}^{t} \tilde{W}_3^T(\tau)\alpha_3^{-1}\dot{\tilde{W}}_3(\tau)d\tau \\
+\tilde{W}_1(t)^T \tfrac{\Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3)}{m_s} \quad \int_{t-T}^{t} \left(\tfrac{1}{4}\tilde{W}_2(\tau)^T \bar{D}_1(x(\tau))\tilde{W}_2(\tau)\right) d\tau \\
-\tilde{W}_1(t)^T \tfrac{\Delta\bar{\phi}_2(x(t-T),T,\hat{u}_2,\hat{d}_3)}{m_s} \quad \int_{t-T}^{t} \left(\tfrac{1}{4}\tilde{W}_3(\tau)^T \bar{E}_1(x(\tau))\tilde{W}_3(\tau)\right) d\tau
\end{gathered} \tag{A.7.1}$$

$$\begin{gathered}
\dot{L}(x) = \int_{t-\mathrm{T}}^{t} \left(-Q(x) - \tfrac{1}{4}W_1(\tau)^T \overline{D}_1(x(\tau))W_1(\tau) + \tfrac{1}{4}W_1(\tau)^T \overline{E}_1(x(\tau))W_1(\tau) + \varepsilon_{HJI}(x)\right) d\tau + \varepsilon_1(x(t)) \\
-\tilde{W}_1(t)^T \Delta\overline{\varphi}_2(t)\Delta\overline{\varphi}_2(t)^T \tilde{W}_1(t) + \tilde{W}_1(t)^T \bar{\Phi}_2(t) \int_{t-T}^{t} \varepsilon_{HJI}(x(\tau))d\tau + \int_{t-\mathrm{T}}^{t} \left(\tfrac{1}{2}W_1(\tau)^T \bar{D}_1(x)\tilde{W}_2(\tau)\right) d\tau \\
- \int_{t-\mathrm{T}}^{t} \left(\tfrac{1}{2}W_1(\tau)^T \bar{E}_1(x)\tilde{W}_3(\tau)\right) d\tau - W_1(t)^T \bar{\Phi}_2(t) \int_{t-T}^{t} \left(\tfrac{1}{4}W_1(\tau)^T \bar{D}_1(x)\tilde{W}_2(\tau)\right) d\tau \\
+W_1(t)^T \bar{\Phi}_2(t) \int_{t-T}^{t} \left(\tfrac{1}{4}W_1(\tau)^T \bar{E}_1(x)\tilde{W}_3(\tau)\right) d\tau + \tilde{W}_1(t)^T \bar{\Phi}_2(t) \int_{t-T}^{t} \left(\tfrac{1}{4}W_1(\tau)^T \bar{D}_1(x)\tilde{W}_2(\tau)\right) d\tau \\
-\tilde{W}_1(t)^T \bar{\Phi}_2(t) \int_{t-T}^{t} \left(\tfrac{1}{4}W_1(\tau)^T \bar{E}_1(x)\tilde{W}_3(\tau)\right) d\tau + W_1(t)^T \bar{\Phi}_2(t) \int_{t-T}^{t} \left(\tfrac{1}{4}\tilde{W}_2(\tau)^T \bar{D}_1(x)\tilde{W}_2(\tau)\right) d\tau \\
-W_1(t)^T \bar{\Phi}_2(t) \int_{t-T}^{t} \left(\tfrac{1}{4}\tilde{W}_3(\tau)^T \bar{E}_1(x)\tilde{W}_3(\tau)\right) d\tau + \hat{W}_1(t)^T \bar{\Phi}_2(t)^T \tfrac{1}{4} \int_{t-T}^{t} \hat{W}_2(\tau)^T \bar{D}_1(x(\tau))\tilde{W}_2(\tau)d\tau \\
-\tfrac{1}{4} \int_{t-T}^{t} \hat{W}_1(\tau)^T \bar{\Phi}_2(\tau)^T \hat{W}_2(\tau)^T \bar{D}_1(x(\tau))\tilde{W}_2(\tau)d\tau - \hat{W}_1(t)^T \bar{\Phi}_2(t)^T \tfrac{1}{4} \int_{t-T}^{t} \hat{W}_3(\tau)^T \bar{E}_1(x(\tau))\tilde{W}_3(\tau)d\tau \\
+\tfrac{1}{4} \int_{t-T}^{t} \hat{W}_1(\tau)^T \bar{\Phi}_2(\tau)^T \hat{W}_3(\tau)^T \bar{E}_1(x(\tau))\tilde{W}_3(\tau)d\tau \\
+ \int_{t-T}^{t} \left[\dot{\hat{W}}_2(\tau)^T a_2^{-1} + \tfrac{1}{4}\hat{W}_1(\tau)^T \bar{\Phi}_2(\tau)^T \hat{W}_2(\tau)^T \bar{D}_1(x)\right] \tilde{W}_2(\tau)d\tau \\
+ \int_{t-T}^{t} \left[\dot{\hat{W}}_3(\tau)^T a_3^{-1} - \tfrac{1}{4}\hat{W}_1(\tau)^T \bar{\Phi}_2(\tau)^T \hat{W}_3(\tau)^T \bar{E}_1(x)\right] \tilde{W}_3(\tau)d\tau
\end{gathered} \tag{A.7.2}$$

$$\dot{\hat{W}}_2 = -a_2 \left\{ \left(F_2\hat{W}_2 - F_1 T \Delta\overline{\varphi}_2^T \hat{W}_1\right) - \tfrac{1}{4m_s}\overline{D}_1(x)\hat{W}_2\Delta\overline{\varphi}_2^T \hat{W}_1 \right\} \tag{A.7.2.1}$$

$$\dot{\hat{W}}_2 = -a_2 \left\{ \left(F_2\hat{W}_2 - F_1 T \Delta\overline{\varphi}_2^T \hat{W}_1\right) - \tfrac{1}{4m_s}\overline{D}_1(x)\hat{W}_2\Delta\overline{\varphi}_2^T \hat{W}_1 \right\} \tag{A.7.2.2}$$

$$\|\varepsilon_1(x)\| < \int_{t-T}^{t} \left(b_{\varepsilon_x} b_f \|x\| + \tfrac{1}{2} T b_{\varepsilon_x} b_g^2 b_{\phi_x} \sigma_{\min}(R) \left(W_{\max} + \left\|\tilde{W}_2\right\|\right) + \tfrac{1}{2\gamma^2} T b_{\varepsilon_x} b_\kappa^2 b_{\phi_x} \left(W_{\max} + \left\|\tilde{W}_3\right\|\right)\right) d\tau \tag{A.7.2.3}$$

$$\begin{array}{l}
\dot{L}(x)=\int_{t-\mathrm{T}}^{t}\left(-Q(x)-\frac{1}{4}W_1(\tau)^T\overline{D}_1(x(\tau))W_1(\tau)+\frac{1}{4}W_1(\tau)^T\overline{E}_1(x(\tau))W_1(\tau)+\varepsilon_{HJI}(x)\right)d\tau \\
+\varepsilon_1(x(t))-\tilde{W}_1(t)^T\Delta\overline{\phi}_2(t)\Delta\overline{\phi}_2(t)^T\tilde{W}_1(t)+\int_{t-T}^{t}(\tilde{W}_2(\tau)^TF_2W_1(\tau)-T\tilde{W}_2(\tau)^TF_1\Delta\overline{\phi}_2(\tau)^TW_1(\tau) \\
-\tilde{W}_2(\tau)^TF_2\tilde{W}_2(\tau)+T\tilde{W}_2(\tau)^TF_1\Delta\overline{\phi}_2(\tau)^T\tilde{W}_1(\tau))d\tau+\int_{t-T}^{t}(\tilde{W}_3(\tau)^TF_4W_1(\tau)-T\tilde{W}_3(\tau)^TF_2\Delta\overline{\phi}_2(\tau)^TW_1(\tau) \\
-\tilde{W}_3(\tau)^TF_4\tilde{W}_3(\tau)+T\tilde{W}_3(\tau)^TF_3\Delta\overline{\phi}_2(\tau)^T\tilde{W}_1(\tau))d\tau+\tilde{W}_1(t)^T\bar{\Phi}_2(t)\int_{t-T}^{t}\varepsilon_{HJI}(x(\tau))d\tau \\
+\int_{t-\mathrm{T}}^{t}\left(\frac{1}{2}W_1(\tau)^T\bar{D}_1(x)\tilde{W}_2(\tau)\right)d\tau-\int_{t-\mathrm{T}}^{t}\left(\frac{1}{2}W_1(\tau)^T\bar{E}_1(x)\tilde{W}_3(\tau)\right)d\tau \\
+\tilde{W}_1(t)^T\bar{\Phi}_2(t)\int_{t-T}^{t}\left(\frac{1}{4}W_1(\tau)^T\bar{D}_1(x)\tilde{W}_2(\tau)\right)d\tau-\tilde{W}_1(t)^T\bar{\Phi}_2(t)\int_{t-T}^{t}\left(\frac{1}{4}W_1(\tau)^T\bar{E}_1(x)\tilde{W}_3(\tau)\right)d\tau \\
+W_1(t)^T\bar{\Phi}_2(t)\int_{t-T}^{t}\left(\frac{1}{4}\tilde{W}_2(\tau)^T\bar{D}_1(x)\tilde{W}_2(\tau)\right)d\tau-W_1(t)^T\bar{\Phi}_2(t)\int_{t-T}^{t}\left(\frac{1}{4}\tilde{W}_3(\tau)^T\bar{E}_1(x)\tilde{W}_3(\tau)\right)d\tau \\
+\hat{W}_1(t)^T\bar{\Phi}_2(t)^T\frac{1}{4}\int_{t-T}^{t}\hat{W}_2(\tau)^T\bar{D}_1(x(\tau))\tilde{W}_2(\tau)d\tau-\frac{1}{4}\int_{t-T}^{t}\hat{W}_1(\tau)^T\bar{\Phi}_2(\tau)^T\hat{W}_2(\tau)^T\bar{D}_1(x(\tau))\tilde{W}_2(\tau)d\tau \\
-\hat{W}_1(t)^T\bar{\Phi}_2(t)^T\frac{1}{4}\int_{t-T}^{t}\hat{W}_3(\tau)^T\bar{E}_1(x(\tau))\tilde{W}_3(\tau)d\tau+\frac{1}{4}\int_{t-T}^{t}\hat{W}_1(\tau)^T\bar{\Phi}_2(\tau)^T\hat{W}_3(\tau)^T\bar{E}_1(x(\tau))\tilde{W}_3(\tau)d\tau \\
-\frac{1}{4}\int_{t-T}^{t}\hat{W}_1(\tau)^T\bar{\Phi}_2(\tau)^T\hat{W}_2(\tau)^T\bar{D}_1(x(\tau))\tilde{W}_2(\tau)d\tau+\frac{1}{4}\int_{t-T}^{t}\hat{W}_1(\tau)^T\bar{\Phi}_2(\tau)^T\hat{W}_3(\tau)^T\bar{E}_1(x(\tau))\tilde{W}_3(\tau)d\tau
\end{array} \tag{A.7.3}$$

So (A.8) becomes (A.8.5)

It is now desired to show that for small enough T $\left\|\tilde{Z}\right\|$ is UUB. Select $p_B>0$large as detailed subsequently. Select T such that $Tf(t,\tau)<\varepsilon_f$ for some fixed $\varepsilon_f>0\forall\left\|\tilde{Z}\right\|<p_B$

Complete the squares to see that (A.9)

Implies $W_3(\left\|\tilde{Z}\right\|)>0$and $\dot{L}<0$.

It is now desired to find upper and lower bounds on the Lyapunov function L. Define $w(t)=\begin{bmatrix} x(t) \\ \tilde{W}_1(t) \\ \tilde{W}_2(t) \\ \tilde{W}_3(t)\end{bmatrix}$according to

$$\begin{array}{l}
L(t)=\int_{t-T}^{t}V(x(\tau))d\tau+\frac{1}{2}\tilde{W}_1^T(t)a_1^{-1}\tilde{W}_1(t) \\
+\frac{1}{2}\int_{t-T}^{t}\tilde{W}_2^T(\tau)a_2^{-1}\tilde{W}_2(\tau)d\tau \\
+\frac{1}{2}\int_{t-T}^{t}\tilde{W}_3^T(\tau)a_3^{-1}\tilde{W}_3(\tau)d\tau
\end{array}$$

we can find class K (see Definition 3) functions k_j and write

$$\begin{array}{l}
k_3\left(\|x(t)\|\right)=\int_{t-T}^{t}k_1\left(\|x(\tau)\|\right)d\tau\le\int_{t-T}^{t}V(x(\tau))d\tau \\
\le\int_{t-T}^{t}k_2\left(\|x(\tau)\|\right)d\tau=k_4\left(\|x(t)\|\right)
\end{array}$$

$$k_5\left(\left\|\tilde{W}_2\right\|\right)\le\frac{1}{2}\int_{t-T}^{t}\tilde{W}_2^T(\tau)a_2^{-1}\tilde{W}_2(\tau)d\tau\le k_6\left(\left\|\tilde{W}_2\right\|\right)$$

$$k_8\left(\left\|\tilde{W}_3\right\|\right)\le\frac{1}{2}\int_{t-T}^{t}\tilde{W}_3^T(\tau)a_3^{-1}\tilde{W}_3(\tau)d\tau\le k_9\left(\left\|\tilde{W}_3\right\|\right)$$

We now need to find a relationship between $\|w(t)\|$ and $\left\|\tilde{Z}(t,\tau)\right\|$to apply the results of Theorem 4.18 in [27].

One has

$$\left\|\tilde{Z}(t,\tau)\right\|\le\|w(t)\|\left\|\Delta\bar{\phi}_2(t)\right\|\le\|w(t)\|$$

According to Technical Lemma 1, (A.9.1)

Now assume that we have enough hidden layer units $N>N_0$then $\varepsilon_B\to 0$according to Remark 1. So we can write

$$k_{10}\|w(t)\|\equiv\left(\frac{\sqrt{\beta_2T_{PE}}}{\beta_1}\left(1+\delta\beta_2a_1\right)\right)^{-1}\|w(t)\|$$
$$\le\left\|\tilde{Z}(t,\tau)\right\|\le\|w(t)\|$$

Finally, we can bound the Lyapunov function as

$$w^T\underline{S}w\le L\le w^T\bar{S}w$$

Therefore,

$$\left\|\tilde{Z}\right\|^2\underline{\sigma}\left(\underline{S}\right)\le\|w\|^2\underline{\sigma}\left(\underline{S}\right)\le L\le\|w\|^2\bar{\sigma}\left(\bar{S}\right)\le\left\|\tilde{Z}\right\|^2\bar{\sigma}\left(\bar{S}\right)$$

and

$$\underline{S}=\begin{bmatrix} k_3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & k_5 & 0 \\ 0 & 0 & 0 & k_8 \end{bmatrix}$$

and $\bar{S}=\begin{bmatrix} k_4 & 0 & 0 & 0 \\ 0 & k_{10} & 0 & 0 \\ 0 & 0 & k_6 & 0 \\ 0 & 0 & 0 & k_9 \end{bmatrix}$

Take $p_1>0$as defined in (A.9). Select $p_B>\sqrt{\frac{\bar{\sigma}(\bar{S})}{\underline{\sigma}(\underline{S})}}p_1$. Then according to Theorem 4.18 in [27],$\forall\tilde{Z}(0)$, $\left\|\tilde{Z}\right\|\le p_B$, $\forall 0\le t$, and $\left\|\tilde{Z}\right\|\le\sqrt{\frac{\bar{\sigma}(\bar{S})}{\underline{\sigma}(\underline{S})}}p_1$, $\forall t\le T_B$.

Now the Technical Lemma 1 and the persistence of excitation condition of $\Delta\bar{\phi}_2$show UUB of $\left\|\tilde{W}_1\right\|$.

References

[1] Tijs S. Introduction to Game Theory. Hindustan Book Agency, India, 2003.

$$
\begin{aligned}
&\dot{L}(x) \leq \tfrac{1}{4}\int_{t-T}^{t} W_{\max}^{2}\left\|\bar{D}_1(x)\right\| d\tau + \tfrac{1}{4}\int_{t-T}^{t} W_{\max}^{2}\left\|\bar{E}_1(x)\right\| d\tau + \int_{t-\mathrm{T}}^{t}\left(x(\tau)^T q x(\tau) + \varepsilon(\tau)\right) d\tau \\
&+ \int_{t-T}^{t}\left(b_{\varepsilon_x} b_f \|x\| + \tfrac{1}{2} T b_{\varepsilon_x} b_g^2 b_{\phi_x} \sigma_{\min}(R)\left(W_{\max} + \left\|\tilde{W}_2\right\|\right) + \tfrac{1}{2\gamma^2} T b_{\varepsilon_x} b_\kappa^2 b_{\phi_x}\left(W_{\max} + \left\|\tilde{W}_3\right\|\right)\right) d\tau \\
&-\tilde{W}_1(t)^T \Delta\bar{\phi}_2(t)\Delta\bar{\phi}_2(t)^T \tilde{W}_1(t) + \int_{t-T}^{t} \tilde{W}_2(\tau)^T F_2 W_1(\tau) - T\tilde{W}_2(\tau)^T F_1 \Delta\bar{\phi}_2(\tau)^T W_1(\tau)) \\
&-\tilde{W}_2(\tau)^T F_2 \tilde{W}_2(\tau)\big) d\tau + \int_{t-T}^{t}\big(\tilde{W}_3(\tau)^T F_4 W_1(\tau) - T\tilde{W}_3(\tau)^T F_3 \Delta\bar{\phi}_2(\tau)^T W_1(\tau) \\
&-\tilde{W}_3(\tau)^T F_4 \tilde{W}_3(\tau)\big) d\tau + \tilde{W}_1(t)^T \bar{\Phi}_2(t)\int_{t-T}^{t} \varepsilon_{HJI}(x(\tau)) d\tau + \int_{t-\mathrm{T}}^{t}\left(\tfrac{1}{2} W_1(\tau)^T \bar{D}_1(x)\tilde{W}_2(\tau)\right) d\tau \\
&-\int_{t-\mathrm{T}}^{t}\left(\tfrac{1}{2} W_1(\tau)^T \bar{E}_1(x)\tilde{W}_3(\tau)\right) d\tau + \tilde{W}_1(t)^T \bar{\Phi}_2(t)\int_{t-T}^{t}\left(\tfrac{1}{4} W_1(\tau)^T \bar{D}_1(x)\tilde{W}_2(\tau)\right) d\tau \\
&-\tilde{W}_1(t)^T \bar{\Phi}_2(t)\int_{t-T}^{t}\left(\tfrac{1}{4} W_1(\tau)^T \bar{E}_1(x)\tilde{W}_3(\tau)\right) d\tau + W_1(t)^T \bar{\Phi}_2(t)\int_{t-T}^{t}\left(\tfrac{1}{4}\tilde{W}_2(\tau)^T \bar{D}_1(x)\tilde{W}_2(\tau)\right) d\tau \\
&-W_1(t)^T \bar{\Phi}_2(t)\int_{t-T}^{t}\left(\tfrac{1}{4}\tilde{W}_3(\tau)^T \bar{E}_1(x)\tilde{W}_3(\tau)\right) d\tau + \hat{W}_1(t)^T \bar{\Phi}_2(t)^T \tfrac{1}{4}\int_{t-T}^{t} \hat{W}_2(\tau)^T \bar{D}_1(x(\tau))\tilde{W}_2(\tau) d\tau \\
&-\tfrac{1}{4}\int_{t-T}^{t} \hat{W}_1(\tau)^T \bar{\Phi}_2(\tau)^T \hat{W}_2(\tau)^T \bar{D}_1(x(\tau))\tilde{W}_2(\tau) d\tau - \hat{W}_1(t)^T \bar{\Phi}_2(t)^T \tfrac{1}{4}\int_{t-T}^{t} \hat{W}_3(\tau)^T \bar{E}_1(x(\tau))\tilde{W}_3(\tau) d\tau \\
&+\tfrac{1}{4}\int_{t-T}^{t} \hat{W}_1(\tau)^T \bar{\Phi}_2(\tau)^T \hat{W}_3(\tau)^T \bar{E}_1(x(\tau))\tilde{W}_3(\tau) d\tau - \tfrac{1}{4}\int_{t-T}^{t} \hat{W}_1(\tau)^T \bar{\Phi}_2(\tau)^T \hat{W}_2(\tau)^T \bar{D}_1(x(\tau))\tilde{W}_2(\tau) d\tau \\
&+\tfrac{1}{4}\int_{t-T}^{t} \hat{W}_1(\tau)^T \bar{\Phi}_2(\tau)^T \hat{W}_3(\tau)^T \bar{E}_1(x(\tau))\tilde{W}_3(\tau) d\tau \\
&+\int_{t-T}^{t}\left(\tilde{W}_2(\tau)^T T F_1\left(\Delta\bar{\phi}_2(\tau)^T \tilde{W}_1(\tau) - \Delta\bar{\phi}_2(t)^T \tilde{W}_1(t)\right)\right) d\tau \\
&+\int_{t-T}^{t} T\tilde{W}_2(\tau)^T F_1 \Delta\bar{\phi}_2(t)^T \tilde{W}_1(t) d\tau \\
&+\int_{t-T}^{t}\left(\tilde{W}_3(\tau)^T T F_3\left(\Delta\bar{\phi}_2(\tau)^T \tilde{W}_1(\tau) - \Delta\bar{\phi}_2(t)^T \tilde{W}_1(t)\right)\right) d\tau \\
&+\int_{t-T}^{t} T\tilde{W}_3(\tau)^T F_3 \Delta\bar{\phi}_2(t)^T \tilde{W}_1(t) d\tau
\end{aligned}
\tag{A.8}
$$

$$
\dot{L} \leq -\int_{t-T}^{t} \tilde{Z}^T M \tilde{Z} d\tau + \int_{t-T}^{t} d\tilde{Z} d\tau + \int_{t-T}^{t} (c+\varepsilon) d\tau + T\int_{t-T}^{t} z^T W z d\tau \tag{A.8.1}
$$

$$
c = \tfrac{1}{4} W_{\max}^{2}\left\|\bar{D}_1(x)\right\| + \tfrac{1}{4} W_{\max}^{2}\left\|\bar{E}_1(x)\right\| + \varepsilon + \tfrac{1}{2} W_{\max} T b_{\varepsilon_x} b_{\varphi_x} b_g^2 \sigma_{\min}(R) + \tfrac{1}{2\gamma^2} T b_{\varepsilon_x} b_\kappa^2 b_{\phi_x} W_{\max} \tag{A.8.2}
$$

$$
\begin{gathered}
W_{31} = W_{13}^T = -W_{32} = -W_{23}^T = \left(\tfrac{1}{8m_s} W_{\max} D_1(x(\tau)) + \tfrac{F_1}{2}\right)\ , \\
W_{51} = W_{15}^T = -W_{52} = -W_{25}^T = \left(-\tfrac{1}{8m_s} W_{\max} E_1(x(\tau)) + \tfrac{F_3}{2}\right)\ , \\
W_{41} = W_{14}^T = -W_{42} = -W_{24}^T = \left(-\frac{1}{8m_s} D_1(x(\tau))\right), \\
W_{61} = W_{16}^T = -W_{62} = -W_{26}^T = \left(\frac{1}{8m_s} E_1(x(\tau))\right).
\end{gathered}
\tag{A.8.3}
$$

$$\begin{gathered} d_1 = b_{\varepsilon_x} b_f, \\ d_2 = \frac{\varepsilon_{HJI}(x(\tau))}{m_s(t)}, \end{gathered}$$

$$\begin{aligned} & d_3 = \tfrac{1}{4}(W_1(t)^T \bar{\Phi}_2(t) - W_1(\tau)^T \bar{\Phi}_2(\tau)) W_1(\tau)^T \bar{D}_1(x(\tau)) \\ & +(\tfrac{1}{2}\bar{D}_1(x(\tau)) + F_2 - T F_1 \Delta\overline{\phi}_2(\tau)^T - \tfrac{1}{4}\bar{D}_1(x(t)) W_1(t) m(t)^T) W_1(\tau) \\ & +\tfrac{1}{2} b_{\varepsilon_x} b_{g^2} b_{\phi_x} \sigma_{\min}(R) \end{aligned}$$

$$\begin{aligned} & d_4 = -\tfrac{1}{4}(W_1(t)^T \bar{\Phi}_2(t) - W_1(\tau)^T \bar{\Phi}_2(\tau)) W_1(\tau)^T \bar{E}_1(x(\tau)) \\ & +(-\tfrac{1}{2}\bar{E}_1(x(\tau)) + F_4 - T F_3 \Delta\overline{\phi}_2(\tau)^T + \tfrac{1}{4}\bar{E}_1(x(t)) W_1(t) m(t)^T) W_1(\tau) \\ & +\tfrac{1}{2\gamma^2} b_{\varepsilon_x} b_{k^2} b_{\phi_x}. \end{aligned}$$

(A.8.4)

$$\begin{aligned} & \dot{L} \le -\textstyle\int_{t-T}^{t} \left\|\tilde{Z}\right\|^2 \sigma_{\min}(M) d\tau + \int_{t-T}^{t} \|d\| \left\|\tilde{Z}\right\| d\tau + \int_{t-T}^{t} (c+\varepsilon)\, d\tau \\ & +\textstyle\int_{t-T}^{t} \left(\rho_1 \left\|\tilde{Z}(t,\tau)\right\| + \rho_2 T \left\|\tilde{Z}(t,\tau)\right\|^2\right)^2 T\sigma_{\max}(W) d\tau \end{aligned}$$

$$\begin{aligned} & \dot{L} \le -\textstyle\int_{t-T}^{t} \left\|\tilde{Z}\right\|^2 \sigma_{\min}(M) d\tau + \int_{t-T}^{t} \|d\| \left\|\tilde{Z}\right\| d\tau + \int_{t-T}^{t} (c+\varepsilon)\, d\tau \\ & +\textstyle\int_{t-T}^{t} \left(\rho_1^2 T \left\|\tilde{Z}(t,\tau)\right\|^2 + \rho_2^2 T^2 \left\|\tilde{Z}(t,\tau)\right\|^4 + 2\rho_1\rho_2 T \left\|\tilde{Z}(t,\tau)\right\|^3\right) \quad \sigma_{\max}(W) d\tau \end{aligned}$$

$$\begin{aligned} & \dot{L} \le -\textstyle\int_{t-T}^{t} \left(\sigma_{\min}(M) - \rho_1^2 T \sigma_{\max}(W)\right) \left\|\tilde{Z}\right\|^2 d\tau + \int_{t-T}^{t} \|d\| \left\|\tilde{Z}\right\| d\tau + \int_{t-T}^{t} (c+\varepsilon)\, d\tau \\ & +\textstyle\int_{t-T}^{t} \left(\rho_2^2 T^2 \left\|\tilde{Z}(t,\tau)\right\|^4 + 2\rho_1\rho_2 T \left\|\tilde{Z}(t,\tau)\right\|^3\right) \sigma_{\max}(W) d\tau \end{aligned}$$

$$\begin{aligned} & \dot{L} \le \textstyle\int_{t-T}^{t} \left(\rho_2^2 T^2 \left\|\tilde{Z}(t,\tau)\right\|^4 + 2\rho_1\rho_2 T \left\|\tilde{Z}(t,\tau)\right\|^3\right) \sigma_{\max}(W) d\tau \\ & +\textstyle\int_{t-T}^{t} \left(\left(-\sigma_{\min}(M) + \rho_1^2 T \sigma_{\max}(W)\right) \left\|\tilde{Z}\right\|^2 + \|d\| \left\|\tilde{Z}\right\| + (c+\varepsilon)\right) d\tau \\ & \equiv \left(T \textstyle\int_{t-T}^{t} f(t,\tau) + \int_{t-T}^{t} g(t,\tau)\right) d\tau \end{aligned}$$

where

$$\begin{aligned} & f(t,\tau) = \rho_2^2 T \left\|\tilde{Z}(t,\tau)\right\|^4 \sigma_{\max}(W) + 2\rho_1\rho_2 \left\|\tilde{Z}(t,\tau)\right\|^3 \sigma_{\max}(W) + \rho_1^2 \sigma_{\max}(W) \left\|\tilde{Z}\right\|^2 \\ & g(t,\tau) = -\sigma_{\min}(M) \left\|\tilde{Z}\right\|^2 + \|d\| \left\|\tilde{Z}\right\| + (c+\varepsilon) \end{aligned}$$

(A.8.5)

$$\dot{L} \le \int_{t-T}^{t} \left(-\sigma_{\min}(M) \left\| \tilde{Z} \right\|^2 + \|d\| \left\| \tilde{Z} \right\| + (c + \varepsilon + \varepsilon_f) \right) d\tau$$
$$\dot{L} \le -\int_{t-T}^{t} \left(\left\| \tilde{Z} \right\| - \tfrac{\|d\|}{2\sigma_{\min}(M)} \right)^2 d\tau + \int_{t-T}^{t} \left(\tfrac{\|d\|}{2\sigma_{\min}(M)} \right)^2 d\tau + \int_{t-T}^{t} \left(\tfrac{c+\varepsilon+\varepsilon_f}{\sigma_{\min}(M)} \right) d\tau \equiv -\int_{t-T}^{t} W_3(\left\| \tilde{Z} \right\|) d\tau$$

Then

$$\left\| \tilde{Z} \right\| > \frac{\|d\|}{2\sigma_{\min}(M)} + \sqrt{\frac{d^2}{4\sigma_{\min}^2(M)} + \frac{c+\varepsilon+\varepsilon_f}{\sigma_{\min}(M)}} \equiv p_1. \tag{A.9}$$

$$\left\| \tilde{W}_1(t) \right\| \le \tfrac{\sqrt{\beta_2 T_{PE}}}{\beta_1} \left\{ \left[(1 + \delta\beta_2 a_1) \left\| \Delta\bar{\phi}_2{}^{\mathrm{T}} \tilde{W}_1 \right\| + \delta\beta_2 a_1 \varepsilon_B \right] \right\} \quad \equiv \tfrac{\sqrt{\beta_2 T_{PE}}}{\beta_1} (1 + \delta\beta_2 a_1) \left\| \Delta\bar{\phi}_2{}^{\mathrm{T}} \tilde{W}_1 \right\| + \varepsilon_3 \tag{A.9.1}$$

[2] Baar T., Olsder G. J. Dynamic Noncooperative Game Theory, 2nd ed. Philadelphia, PA: SIAM, 1999, vol. 23, SIAM's Classic in Applied Mathematics.

[3] Baar T., Bernard P. Optimal Control and Related Minimax Design Problems. Boston, MA: Birkhuser, 1995.

[4] Van Der Shaft A. J. L2-gain analysis of nonlinear systems and nonlinear state feedback control. IEEE Transactions on Automatic Control 1992; 37(6): 770-784.

[5] Abu-Khalaf M., Lewis F. L. Neurodynamic Programming and Zero-Sum Games for Constrained Control Systems. IEEE Transactions on Neural Networks 2008; 19(7); 1243-1252.

[6] Abu-Khalaf M., Lewis F. L., Huang J. Policy Iterations on the Hamilton-Jacobi- Isaacs Equation for H? State Feedback Control With Input Saturation. IEEE Transactions on Automatic Control 2006; 51(12); 1989–1995.

[7] Abu-Khalaf M., Lewis F. L., Nearly Optimal Control Laws for Nonlinear Systems with Saturating Actuators Using a Neural Network HJB Approach, *Automatica* 2005, **41** (5), 779–791.

[8] Murray J. J., Cox C. J., Lendaris G. G., Saeks R., Adaptive Dynamic Programming, *IEEE Trans. on Systems, Man and Cybernetics* 2002, **32** (2), 140-153.

[9] Bertsekas D. P., Tsitsiklis J. N. Neuro-Dynamic Programming. Athena Scientific: MA, 1996.

[10] Si J., Barto A., Powel W., Wunch D., *Handbook of Learning and Approximate Dynamic Programming*, John Wiley, New Jersey, 2004.

[11] Sutton R. S., Barto A. G., *Reinforcement Learning – An Introduction*, MIT Press, Cambridge, Massachusetts, 1998.

[12] Howard R. A., *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, Massachusetts, 1960.

[13] Werbos P.J. Beyond Regression: New Tools for Prediction and Analysis in the Behavior Sciences. Ph.D. Thesis. 1974.

[14] Werbos P. J. Approximate dynamic programming for real-time control and neural modeling. Handbook of Intelligent Control. ed. D.A. White and D.A. Sofge, New York: Van Nostrand Reinhold, 1992.

[15] Werbos, Neural networks for control and system identification, *IEEE Proc. CDC89.*

[16] Prokhorov D., Wunsch D., Adaptive critic designs," *IEEE Trans. on Neural Networks* 1997, **8**(5), 997–1007.

[17] Baird III L. C., "Reinforcement Learning in Continuous Time: Advantage Updating", *Proc. Of ICNN*, 1994.

[18] Vamvoudakis K. G., Lewis F. L. Online Actor-Critic Algorithm to Solve the Continuous-Time Infinite Horizon Optimal Control Problem. Automatica 2010; 46(5): 878–888.

[19] Vamvoudakis K. G., Lewis F. L. Online Neural Network Solution of Nonlinear Two-Player Zero-Sum Games Using Synchronous Policy Iteration. to appear in International Journal of Robust and Nonlinear Control, 2011.

[20] Vrabie D., Pastravanu O., Lewis F., Abu-Khalaf M., Adaptive Optimal Control for Continuous-Time Linear Systems Based on Policy Iteration,*Automatica* 2009, 42(2), 477–484.

[21] Kleinman D. On an Iterative Technique for Riccati Equation Computations. *IEEE Transactions on Automatic Control* 1968; 13; 114–115.

[22] Dierks T., Jagannathan S., Optimal Control of Affine Nonlinear Continuous-time systems Using an Online Hamilton-Jacobi-Isaacs Formulation, *Proc. 49th IEEE Conference on Decision and Control* 2010, 3048–3053.

[23] Johnson M., Bhasin S., Dixon W. E. Nonlinear Two-player Zero-sum Game Approximate Solution Using a Policy Iteration Algorithm. to appear *IEEE Conference on Decision and Control*, Orlando, FL, 2011.

[24] Bhasin S., Johnson M., Dixon W. E. A model free robust policy iteration algorithm for optimal control of nonlinear systems. *Proc. 49th IEEE Conference on Decision and Control* 2010; 3060–3065.

[25] Lewis F. L., Syrmos V. L. Optimal Control. John Wiley, 1995.

[26] Lewis F.L., Jagannathan S., Yesildirek A. Neural Network Control of Robot Manipulators and Nonlinear Systems. Taylor & Francis 1999.

[27] Khalil H. K. Nonlinear Systems. Prentice-Hall, 1996.

[28] Stevens B., Lewis F. L., *Aircract Control and Simulation,* 2nd edition, John Willey, New Jersey, 2003.

[29] Nevistic V. , Primbs J. A. Constrained nonlinear optimal control: a converse HJB approach. Technical Report 96-021, California Institute of Technology, 1996.