18

Online Gaming: Real Time Solution of Nonlinear Two-Player Zero-Sum Games Using Synchronous Policy Iteration

Kyriakos G. Vamvoudakis and Frank L. Lewis
Automation and Robotics Research Institute, The University of Texas at Arlington, USA

1. Introduction

Games provide an ideal environment in which to study computational intelligence, offering a range of challenging and engaging problems. Game theory (Tijs, 2003) captures the behavior in which a player's success in selecting strategies depends on the choices of other players. One goal of game theory techniques is to find (saddle point) equilibria, in which each player has an outcome that cannot be improved by unilaterally changing his strategy (e.g. Nash equilibrium). The H_∞ control problem is a *minimax* optimization problem, and hence a zero-sum game where the controller is a minimizing player and the disturbance a maximizing one. Since the work of George Zames in the early 1980s, H_∞ techniques have been used in control systems, for sensitivity reduction and disturbance rejection. This chapter is concerned with 2-player zero-sum games that are related to the H_∞ control problem, as formulated by (Basar & Olsder, 1999; Basar & Bernard, 1995; Van Der Shaft, 1992).

Game theory 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 and generate fixed control policies that are then implemented in online controllers in real time.

In this chapter we provide methods for online gaming, that is for solution of 2-player zero-sum infinite horizon games *online*, through learning the saddle point strategies in real-time. The dynamics may be nonlinear in continuous-time and are assumed known. A novel neural network 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 (Doya,

2001; Doya et al 2001; Howard, 1960; Barto et al 2004; Sutton & Barto, 1998). 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 terms, RL refers to the learning approach of an actor or agent which modifies its actions, or control policies, based on stimuli received in response to its interaction with its environment. This learning can be extended along two dimensions: i) nature of interaction (competitive or collaborative) and ii) the number of decision makers (single or multi agent).

In view of the advantages offered by the RL methods, 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) (Howard, 1960; Sutton & Barto, 1998; D. Vrabie et al, 2009), 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 (Bertsekas & Tsitsiklis, 1996) and (Werbos, 1974; Werbos 1992). Recently, online policy iteration methods for continuous-time systems have been developed by (D. Vrabie et al, 2009).

In recent work (Vamvoudakis & Lewis, 2010), we developed an online approximate solution method based on PI for the (1-player) infinite horizon optimal control problem for continuous-time nonlinear systems with known dynamics. This is an optimal adaptive controller that uses two adaptive structures, one for the value (cost) function and one for the control policy. The two structures are tuned simultaneously online to learn the solution of the HJ equation and the optimal policy.

This chapter presents an optimal adaptive control method that converges online to the solution to the 2-player differential game (and hence the solution of the bounded L_2 gain problem). Three approximator structures are used. Parameter update laws are given to tune critic, actor, and disturbance neural networks simultaneously online to converge to the solution to the HJ equation and the saddle point policies, while also guaranteeing closed-loop stability. Rigorous proofs of performance and convergence are given.

The chapter 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 synchronous zero-sum game PI algorithm. 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 in section 4, the main result is presented in Theorem 2, which shows how to tune all three approximators simultaneously by using measurements along the system trajectories in real

time and Theorem 3, which proves exponential convergence to the critic neural network and convergence to the approximate Nash solution. Proofs using Lyapunov techniques guarantee convergence and closed-loop stability. Section 5 presents simulation examples that show the effectiveness of the online synchronous zero-sum game CT PI algorithm in learning the optimal value, control and disturbance for both linear and nonlinear systems. Interestingly, a simulation example shows that the two-player online game converges faster than an equivalent online 1-player (optimal control) problem when all the neural networks are tuned simultaneously in real time. Therefore, it is indicated that one learns faster if one has an opponent and uses synchronous policy iteration techniques.

2. Background: Two player differential game, and policy iteration

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 $x(t) \in \mathbb{R}^n$, control $u(x) \in \mathbb{R}^m$, and disturbance $d(x) \in \mathbb{R}^q$, Assume that $f(x)$ is locally Lipschitz, $\|f(x)\| < b_f \|x\|$, and $f(0) = 0$ so that $x = 0$ is an equilibrium point of the system. Furthermore take $g(x), k(x)$ as continuous.
Define the performance index (Lewis & Syrmos, 1995)

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

for $Q(x) \geq 0$, $R = R^T > 0$, $r(x, u, d) = Q(x) + u^T R u - \gamma^2 \|d\|^2$ and $\gamma \geq \gamma^* \geq 0$, where γ^* is the smallest γ for which the system is stabilized (Van Der Shaft, 1992). For feedback policies $u(x)$ and disturbance policies $d(x)$, define the value or cost of the policies as

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

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

$$0 = r(x, u, d) + (\nabla V)^T (f(x) + g(x)u(x) + k(x)d(x)), \; V(0) = 0 \tag{4}$$

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

$$H(x, \nabla V, u, d) = r(x, u, d) + (\nabla V)^T (f(x) + g(x)u(x) + k(x)d) \tag{5}$$

For feedback policies (Basar & Bernard, 1995), a solution $V(x) \geq 0$ to (4) is the value (5) 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 (Basar & Bernard, 1995; Basar & Olsder, 1999)

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

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

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

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

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

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

and write

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

Note that global solutions to the HJI (11) may not exist. Moreover, if they do, they may not be smooth. For a discussion on viscosity solutions to the HJI, see (Ball & Helton, 1996; Bardi & Capuzzo-Dolcetta, 1997; Basar & Bernard, 1995). 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 $V_a(x) \geq V(x) \geq 0$. Linearize the system (1) about the origin to obtain the Generalized ARE (See Section IV.A). 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 (Van Der Shaft, 1992).

It is shown in (Basar & Bernard, 1995) that if $V^*(x)$ is the minimum non-negative 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 solution of the HJI equation

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 (Feng et al. 2009) which uses an inner loop with iterations on the control, and (Abu-Khalaf, Lewis, 2008; Abu-Khalaf et al. , 2006; Van Der Shaft, 1992) which uses an inner loop with iterations on the disturbance. These are in effect extensions of Kleinman's algorithm (Kleinman, 1968) to nonlinear 2-player games. The complementarity of these algorithms is shown in (Vrabie, 2009). Here, we shall use the latter algorithm (e.g. (Abu-Khalaf, Lewis, 2008; Abu-Khalaf et al., 2006; Van Der Shaft, 1992)).

Policy Iteration (PI) Algorithm for 2-Player Zero-Sum Differential Games (Van Der Shaft, 1992)

Initialization: Start with a stabilizing feedback 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

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

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

On convergence, set $V_{j+1}(x) = V_j^i(x)$

3. Update the control policy using

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

Go to 1.

■

Nota Bene: In practice, the iterations in i and j are continued until some convergence criterion is met, e.g. $\left\| V_j^{i+1} - V_j^i \right\|$ or, respectively $\left\| V_{j+1} - V_j \right\|$ is small enough in some suitable norm.

Given a feedback policy $u(x)$, write the Hamilton-Jacobi (HJ) equation

$$0 = Q(x) + \nabla V^T(x)\big(f(x) + g(x)u(x)\big) + u^T(x) R u(x) + \frac{1}{4\gamma^2} \nabla V^T(x) k k^T \nabla V(x), \qquad V(0) = 0 \tag{15}$$

for fixed $u(x)$. The minimal non negative solution $V(x)$ to this equation is the so-called available storage for the given $u(x)$ (Van Der Shaft, 1992). Note that the inner loop of this algorithm finds the available storage for u_j, where it exists.

Assuming that the available storage at each index j is smooth on a local domain of validity, the convergence of this algorithm to the minimal nonnegative solution to the HJI equation is shown in (Abu-Khalaf & Lewis, 2008; Van Der Shaft, 1992). Under these assumptions, the existence of smooth solutions at each step to the Lyapunov-like equation (12) was further shown in (Abu-Khalaf et al., 2006). Also shown was the asymptotic stability of $(f + gu_j + kd^i)$ at each step. In fact, the inner loop yields $V_j^{i+1}(x) \geq V_j^i(x), \forall x$ while the outer loop yields $V_{j+1}(x) \leq V_j(x), \forall x$ until convergence to V^*.

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 policy/disturbance policy pairs.

3. Approximator structure and solution of the Lyapunov equation

The PI Algorithm is a *sequential* algorithm that solves the HJI equation (11) and finds the Nash solution (u^*, d^*) based on sequential solutions of the nonlinear Lyapunov equation (12). That is, while the disturbance policy is being updated, the feedback policy is held constant. In this section, we use PI to lay a rigorous foundation for the NN approximator structure required *on-line solution of the 2-player zero-sum differential game in real time.* In the next section, this structure will be used to develop an adaptive control algorithm of novel form that converges to the ZS game solution. It is important to define the neural network structures and the NN estimation errors properly or such an adaptive algorithm cannot be developed.

The PI algorithm itself is not implemented in this chapter. Instead, here one implements both loops, the outer feedback control update loop and the inner disturbance update loop, *simultaneously* using neural network learning implemented as differential equations for tuning the weights, while simultaneously keeping track of and learning the value $V(x(t),u,d)$ (3) of the current control and disturbance by solution of the Lyapunov equation (4)/(12). We call this *synchronous PI* for zero-sum games.

3.1 Value function approximation: Critic Neural Network Structure

This chapter uses nonlinear approximator structures (e.g. neural networks) for Value Function Approximation (VFA) (Bertsekas & Tsitsiklis, 1996; Werbos, 1974; Werbos, 1992), therefore sacrificing some representational accuracy in order to make the representation manageable in practice. Sacrificing accuracy in the representation of the value function is not so critical, since the ultimate goal is to find a good policy and not necessarily an accurate value function. Based on the structure of the PI algorithm in Section IIB, VFA for online 2-player games requires three approximators, which are taken as neural networks (NN), one for the value function, one for the feedback control policy, and one for the disturbance policy. These are motivated respectively by the need to solve equations (12), (14), and (13).

To solve equation (12), we use VFA, which here requires approximation in Sobolev norm (Adams & Fournier, 2003), that is, approximation of the value $V(x)$ as well as its gradient $\nabla V(x)$. The following definition describes uniform convergence that is needed later.

Definition 2. (uniform convergence). A sequence of functions $\{p_n\}$ converges uniformly to p if $\forall \varepsilon > 0, \exists N(\varepsilon) : \sup \left\| p_n(x) - p(x) \right\| < \varepsilon, n > N(\varepsilon)$.

Assumption 1. For each feedback control and disturbance policy the nonlinear Lyapunov equation (12) has a smooth local solution $V(x) \geq 0$.

According to the Weierstrass higher-order approximation Theorem (Abu-Khalaf & Lewis, 2005; Finlayson, 1990; Hornik et al., 1990), there exists a complete independent basis set $\{\varphi_i(x)\}$ such that the solution $V(x)$ to (4) and its gradient are uniformly approximated, that is, there exist coefficients c_i such that

$$V(x) = \sum_{i=1}^{\infty} c_i \varphi_i(x) = \sum_{i=1}^{N} c_i \varphi_i(x) + \sum_{i=N+1}^{\infty} c_i \varphi_i(x)$$

$$V(x) \equiv C_1^T \phi_1(x) + \sum_{i=N+1}^{\infty} c_i \varphi_i(x) \tag{16}$$

$$\frac{\partial V(x)}{\partial x} = \sum_{i=1}^{\infty} c_i \frac{\partial \varphi_i(x)}{\partial x} = \sum_{i=1}^{N} c_i \frac{\partial \varphi_i(x)}{\partial x} + \sum_{i=N+1}^{\infty} c_i \frac{\partial \varphi_i(x)}{\partial x} \tag{17}$$

where $\phi_1(x) = [\varphi_1(x)\ \varphi_2(x) \cdots \varphi_N(x)]^T : \mathbb{R}^n \to \mathbb{R}^N$, and the second terms in these equations converge uniformly to zero as $N \to \infty$. Specifically, the linear subspace generated by the basis set is dense in the Sobolev norm $W^{1,\infty}$ (Adams, Fournier, 2003).
Therefore, assume there exist NN weights W_1 such that the value function $V(x)$ is approximated as

$$V(x) = W_1^T \phi_1(x) + \varepsilon(x) \tag{18}$$

with $\phi_1(x) : \mathbb{R}^n \to \mathbb{R}^N$ the NN activation function vector, N the number of neurons in the hidden layer, and $\varepsilon(x)$ the NN approximation error. For approximation in Sobolev space, the NN activation functions $\{\varphi_i(x) : i = 1, N\}$ should be selected so that $\{\varphi_i(x) : i = 1, \infty\}$ provides a complete independent basis set such that $V(x)$ and its derivative are uniformly approximated, e.g., additionally

$$\frac{\partial V}{\partial x} = \left(\frac{\partial \phi_1(x)}{\partial x} \right)^T W_1 + \frac{\partial \varepsilon}{\partial x} = \nabla \phi_1^T W_1 + \nabla \varepsilon \tag{19}$$

Then, as the number of hidden-layer neurons $N \to \infty$, the approximation errors $\varepsilon \to 0$, $\nabla \varepsilon \to 0$ uniformly (Abu-Khalaf & Lewis, 2005; Finlayson, 1990). In addition, for fixed N, the NN approximation errors $\varepsilon(x)$, and $\nabla \varepsilon$ are bounded by constants locally (Hornik et al., 1990).
We refer to the NN with weights W_1 that performs VFA as the *critic* NN.
Standard usage of the Weierstrass high-order approximation Theorem uses polynomial approximation. However, non-polynomial basis sets have been considered in the literature (e.g. (Hornik et al., 1990; Sandberg, 1997)). The NN approximation literature has considered a variety of activation functions including sigmoids, tanh, radial basis functions, etc.
Using the NN VFA, considering fixed feedback and disturbance policies $u(x(t)), d(x(t))$, equation (4) becomes

$$H(x, W_1, u, d) = Q(x) + u^T R u - \gamma^2 \|d\|^2 + W_1^T \nabla \phi_1 (f(x) + g(x)u(x) + k(x)d(x)) = \varepsilon_H \tag{20}$$

where the residual error is

$$\varepsilon_H = -\left(\nabla \varepsilon\right)^T (f + gu + kd)$$

$$= -(C_1 - W_1)^T \nabla \phi_1 (f + gu + kd) - \sum_{i=N+1}^{\infty} c_i \nabla \varphi_i(x)(f + gu + kd) \tag{21}$$

Under the Lipschitz assumption on the dynamics, this residual error is bounded locally.

The following Proposition has been shown in (Abu-Khalaf & Lewis, 2005; Abu-Khalaf & Lewis, 2008).

Define $|v|$ as the magnitude of a scalar v, $\|x\|$ as the vector norm of a vector x, and $\|\ \|_2$ as the induced matrix 2-norm.

Proposition 1. For any policies $u(x(t)), d(x(t))$ the least-squares solution to (20) exists and is unique for each N. Denote this solution as W_1 and define

$$V_1(x) = W_1^T \phi_1(x) \tag{22}$$

Then, as $N \to \infty$:

a. $\sup|\varepsilon_H| \to 0$

b. $\sup\|W_1 - C_1\|_2 \to 0$

c. $\sup|V_1 - V| \to 0$

d. $\sup\|\nabla V_1 - \nabla V\| \to 0$

■

This result shows that $V_1(x)$ converges uniformly in Sobolev norm $W^{1,\infty}$ (Adams & Fournier, 2003) to the exact solution $V(x)$ to (4) as $N \to \infty$, and the weights W_1 converge to the first N of the weights, C_1, which exactly solve (4).

The effect of the approximation error on the HJI equation (8) is

$$Q(x) + W_1^T \nabla\varphi_1(x) f(x) - \frac{1}{4} W_1^T \nabla\varphi_1(x) g(x) R^{-1} g^T(x) \nabla\varphi_1^T W_1 + \frac{1}{4\gamma^2} W_1^T \nabla\varphi_1(x) k k^T \nabla\varphi_1^T W_1 = \varepsilon_{HJI} \tag{23}$$

where the residual error due to the function approximation error is

$$\varepsilon_{HJI} \equiv -\nabla\varepsilon^T f + \tfrac{1}{2} W_1^T \nabla\varphi_1 g R^{-1} g^T \nabla\varepsilon + \tfrac{1}{4} \nabla\varepsilon^T g R^{-1} g^T \nabla\varepsilon - \tfrac{1}{2\gamma^2} W_1^T \nabla\varphi_1 k k^T \nabla\varepsilon - \tfrac{1}{4\gamma^2} \nabla\varepsilon^T k k^T \nabla\varepsilon \tag{24}$$

It was also shown in (Abu-Khalaf & Lewis, 2005; Abu-Khalaf & Lewis, 2008) that this error converges uniformly to zero as the number of hidden layer units N increases. That is, $\forall \varepsilon > 0, \exists N(\varepsilon) : \sup\|\varepsilon_{HJI}\| < \varepsilon, N > N(\varepsilon)$.

3.2 Tuning and convergence of the critic neural network

In this section are addressed the tuning and convergence of the critic NN weights when fixed feedback control and disturbance policies are prescribed. Therefore, the focus is on solving the nonlinear Lyapunov-like equation (4) (e.g. (12)) for a fixed feedback policy u and fixed disturbance policy d.

In fact, this amounts to the *design of an observer for the value function*. Therefore, this algorithm is consistent with adaptive control approaches which first design an observer for the system state and unknown dynamics, and then use this observer in the design of a feedback control.

The ideal weights of the critic NN, W_1 which provide the best approximate solution for (20) are unknown. Therefore, the output of the critic neural network is

$$\hat{V}(x) = \hat{W}_1^T \phi_1(x) \tag{25}$$

where $\hat{W}_1$ are the current estimated values of W_1. The approximate nonlinear Lyapunov-like equation is then

$$H(x,\hat{W}_1,u,d) = \hat{W}_1{}^T\nabla\phi_1(f+gu+kd) + Q(x) + u^T Ru - \gamma^2\|d\|^2 = e_1 \tag{26}$$

with e_1 a residual equation error. In view of Proposition 1, define the critic weight estimation error

$$\tilde{W}_1 = W_1 - \hat{W}_1 .$$

Then,

$$e_1 = -\tilde{W}_1{}^T\nabla\phi_1(f+gu) + \varepsilon_H .$$

Given any feedback control policy u, it is desired to select $\hat{W}_1$ to minimize the squared residual error

$$E_1 = \tfrac{1}{2}e_1{}^T e_1 .$$

Then $\hat{W}_1(t)\to W_1$ and $e_1\to\varepsilon_H$. Select the tuning law for the critic weights as the normalized gradient descent algorithm

$$\dot{\hat{W}}_1 = -a_1\frac{\partial E_1}{\partial\hat{W}_1} = -a_1\frac{\sigma_1}{(1+\sigma_1{}^T\sigma_1)^2}[\sigma_1{}^T\hat{W}_1 + h^T h + u^T Ru - \gamma^2\|d\|^2] \tag{27}$$

where $\sigma_1 = \nabla\phi_1(f+gu+kd)$. This is a nonstandard modified Levenberg-Marquardt algorithm where $(\sigma_1^T\sigma_1+1)^2$ is used for normalization instead of $(\sigma_1^T\sigma_1+1)$. This is required in the theorem proofs, where one needs both appearances of $\sigma_1/(1+\sigma_1^T\sigma_1)$ in (27) to be bounded (Ioannou & Fidan, 2006; Tao, 2003).
Note that, from (20),

$$Q(x) + u^T Ru - \gamma^2\|d\|^2 = -W_1{}^T\nabla\varphi_1(f+gu+kd) + \varepsilon_H . \tag{28}$$

Substituting (28) in (27) and, with the notation

$$\bar{\sigma}_1 = \sigma_1/(\sigma_1^T\sigma_1+1)\ ,\ m_s = 1+\sigma_1^T\sigma_1 \tag{29}$$

we obtain the dynamics of the critic weight estimation error as

$$\dot{\tilde{W}}_1 = -a_1\bar{\sigma}_1\bar{\sigma}_1{}^T\tilde{W}_1 + a_1\bar{\sigma}_1\frac{\varepsilon_H}{m_s}. \tag{30}$$

To guarantee convergence of $\hat{W}_1$ to W_1, the next Persistence of Excitation (PE) assumption and associated technical lemmas are required.
Persistence of Excitation (PE) Assumption. Let the signal $\bar{\sigma}_1$ be persistently exciting over the interval $[t,t+T]$, *i.e.* there exist constants $\beta_1>0$, $\beta_2>0$, $T>0$ such that, for all *t*,

$$\beta_1 \mathrm{I} \le S_0 \equiv \int_t^{t+T} \bar{\sigma}_1(\tau)\bar{\sigma}_1^{\mathrm{T}}(\tau)d\tau \le \beta_2 \mathrm{I}. \tag{31}$$

with I the identity matrix of appropriate dimensions.
The PE assumption is needed in adaptive control if one desires to perform system identification using e.g. RLS (Ioannou & Fidan, 2006; Tao, 2003). It is needed here because one effectively desires to identify the critic parameters to approximate $V(x)$.
The properties of tuning algorithm (27) are given in the subsequent results. They are proven in (Vamvoudakis & Lewis, 2010).
Technical Lemma 1. Consider the error dynamics system with output defined as

$$\dot{\tilde{W}}_1 = -a_1\bar{\sigma}_1\bar{\sigma}_1^{\mathrm{T}}\tilde{W}_1 + a_1\bar{\sigma}_1\frac{\varepsilon_H}{m_s}$$

$$y = \bar{\sigma}_1^{T}\tilde{W}_1 . \tag{32}$$

The PE condition (31) is equivalent to the uniform complete observability (UCO) (Lewis, Jagannathan, Yesildirek, 1999) of this system, that is there exist constants $\beta_3 > 0$, $\beta_4 > 0$, $\mathrm{T} > 0$ such that, for all t,

$$\beta_3 \mathrm{I} \le S_1 \equiv \int_t^{t+T} \Phi^{T}(\tau,t)\bar{\sigma}_1(\tau)\bar{\sigma}_1^{\mathrm{T}}(\tau)\Phi(\tau,t)d\tau \le \beta_4 \mathrm{I}. \tag{33}$$

with $\Phi(t_1,t_0),\ t_0 \le t_1$ the state transition matrix of (32) and I the identity matrix of appropriate dimensions.

■

Technical Lemma 2. Consider the error dynamics system (32). Let the signal $\bar{\sigma}_1$ be persistently exciting. Then:

a. The system (32) is exponentially stable. In fact if $\varepsilon_H = 0$ then $\left\|\tilde{W}(k\mathrm{T})\right\| \le e^{-\alpha k T}\left\|\tilde{W}(0)\right\|$ with

$$\alpha = -\frac{1}{\mathrm{T}}\ln(\sqrt{1-2a_1\beta_3}) . \tag{34}$$

b. Let $\left\|\varepsilon_H\right\| \le \varepsilon_{\max}$ and $\left\|y\right\| \le y_{\max}$. Then $\left\|\tilde{W}_1\right\|$ converges exponentially to the residual set

$$\tilde{W}_1(t) \le \frac{\sqrt{\beta_2 \mathrm{T}}}{\beta_1}\left\{\left[y_{\max} + \delta\beta_2 a_1\left(\varepsilon_{\max} + y_{\max}\right)\right]\right\} . \tag{35}$$

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

■

The next result shows that the tuning algorithm (27) is effective under the PE condition, in that the weights $\hat{W}_1$ converge to the actual unknown weights W_1 which solve the nonlinear Lyapunov-like equation (20) in a least-squares sense for the given feedback and disturbance policies $u(x(t)),\ d(x(t))$. That is, (25) converges close to the actual value function of the current policies. The proof is in (Vamvoudakis & Lewis, 2010).

Theorem 1. Let $u(x(t)), d(x(t))$ be any bounded policies. Let tuning for the critic NN be provided by (27) and assume that $\bar{\sigma}_1$ is persistently exciting. Let the residual error in (20) be bounded $\|\varepsilon_H\| < \varepsilon_{\max}$. Then the critic parameter error converges exponentially with decay factor given by (34) to the residual set

$$\tilde{W}_1(t) \le \frac{\sqrt{\beta_2 \mathrm{T}}}{\beta_1}\left\{\left[1 + 2\delta\beta_2 a_1\right]\varepsilon_{\max}\right\}. \tag{36}$$

■

Remark 1. Note that, as $N \to \infty$, $\varepsilon_H \to 0$ uniformly (Abu-Khalaf & Lewis, 2005; Abu-Khalaf & Lewis, 2008). This means that $\varepsilon_{\max}$ decreases as the number of hidden layer neurons in (25) increases.

Remark 2. This theorem requires the assumption that the feedback policy $u(x(t))$ and the disturbance policy $d(x(t))$ are bounded, since the policies appear in (21). In the upcoming Theorems 2 and 3 this restriction is removed.

3.3 Action and disturbance neural network

It is important to define the neural network structure and the NN estimation errors properly for the control and disturbance or an adaptive algorithm cannot be developed. To determine a rigorously justified form for the actor and the disturbance NN, consider one step of the Policy Iteration algorithm (12)-(14). Suppose that the solution $V(x)$ to the nonlinear Lyapunov equation (12) for given control and disturbance policies is smooth and given by (16). Then, according to (17) and (13), (14) one has for the policy and the disturbance updates:

$$u = -\frac{1}{2}R^{-1}g^T(x)\sum_{i=1}^{\infty} c_i \nabla\varphi_i(x) \tag{37}$$

$$d = \frac{1}{2\gamma^2}k^T(x)\sum_{i=1}^{\infty} c_i \nabla\varphi_i(x) \tag{38}$$

for some unknown coefficients c_i. Then one has the following result.

The following proposition is proved in (Abu-Khalaf & Lewis, 2008) for constrained inputs. Non-constrained inputs are easier to prove.

Proposition 2. Let the least-squares solution to (20) be W_1 and define

$$u_1 = -\frac{1}{2}R^{-1}g^T(x)\nabla V_1(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla\phi_1{}^T(x)W_1 \tag{39}$$

$$d_1 = \frac{1}{2\gamma^2}k^T(x)\nabla V_1(x) = \frac{1}{2\gamma^2}k^T(x)\nabla\phi_1{}^T(x)W_1 \tag{40}$$

with V_1 defined in (22). Then, as $N \to \infty$:

a. $\sup\|u_1 - u\| \to 0$

b. $\sup\|d_1 - d\| \to 0$

c. There exists a number of NN hidden layer neurons N_0 such that u_1 and d_1 stabilize the system (1) for $N>N_0$.

■

In light of this result, the ideal feedback and disturbance policy updates are taken as (39), (40) with W_1 unknown. Therefore, define the feedback policy in the form of an action neural network which computes the control input in the structured form

$$\hat{u}(x) = -\tfrac{1}{2}R^{-1}g^T(x)\nabla\phi_1{}^T\hat{W}_2, \tag{41}$$

where $\hat{W}_2$ denotes the current estimated values of the ideal NN weights W_1. Define the actor NN estimation error as

$$\tilde{W}_2 = W_1 - \hat{W}_2 \tag{42}$$

Likewise, define the disturbance in the form of a disturbance neural network which computes the disturbance input in the structured form

$$\hat{d}(x) = \tfrac{1}{2\gamma^2}k^T(x)\nabla\phi_1{}^T\hat{W}_3, \tag{43}$$

where $\hat{W}_3$ denotes the current estimated values of the ideal NN weights W_1. Define the disturbance NN estimation error as

$$\tilde{W}_3 = W_1 - \hat{W}_3 \tag{44}$$

4. Online solution of 2-player zero-sum games using neural networks

This section presents our main results. An online adaptive PI algorithm is given for online solution of the zero-sum game problem which involves simultaneous, or synchronous, tuning of critic, actor, and disturbance neural networks. That is, the weights of all three neural networks are tuned at the same time. This approach is a version of Generalized Policy Iteration (GPI), as introduced in (Sutton & Barto, 1998). In the standard Policy Iteration algorithm (12)-(14), the critic and actor NNs are tuned sequentially, e.g. one at a time, with the weights of the other NNs being held constant. By contrast, we *tune all NN simultaneously in real-time*.

The next definition and facts complete the machinery required for the main results.

Definition 3. (Lewis, Jagannathan, Yesildirek, 1999) (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 $T(B,\zeta(0))$ such that $\|\zeta(t)\| \le B$ for all $t \ge t_0 + T$.

Facts 1.

a. $g(.), k(.)$ are bounded by constants:

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

b. The NN approximation error and its gradient are bounded locally so that

$$\|\varepsilon\| < b_\varepsilon, \ \|\nabla\varepsilon\| < b_{\varepsilon_x}$$

c. The NN activation functions and their gradients are bounded locally so that

$$\|\phi_1(x)\| < b_\phi \,,\ \|\nabla\phi_1(x)\| < b_{\phi_x}$$

■

The main Theorems are now given, which provide the tuning laws for the actor, critic and disturbance neural networks that guarantee convergence of the synchronous online zero-sum game PI algorithm in real-time to the game saddle point solution, while guaranteeing closed-loop stability.

Theorem 2. System stability and convergence of NN weights. Let the dynamics be given by (1), the critic NN be given by (25), the control input be given by actor NN (41) and the disturbance input be given by disturbance NN (43). Let tuning for the critic NN be provided by

$$\dot{\hat{W}}_1 = -a_1 \frac{\sigma_2}{(\sigma_2^T\sigma_2+1)^2}[\sigma_2{}^T\hat{W}_1 + Q(x) - \gamma^2\left\|\hat{d}\right\|^2 + \hat{u}^T R\hat{u}] \tag{45}$$

where $\sigma_2 = \nabla\phi_1(f + g\hat{u} + k\hat{d})$. Let the actor NN be tuned as

$$\dot{\hat{W}}_2 = -\alpha_2\left\{(F_2\hat{W}_2 - F_1\bar{\sigma}_2^T\hat{W}_1) - \frac{1}{4}\bar{D}_1(x)\hat{W}_2 m^T(x)\hat{W}_1\right\} \tag{46}$$

and the disturbance NN be tuned as

$$\dot{\hat{W}}_3 = -\alpha_3\left\{\left(F_4\hat{W}_3 - F_3\bar{\sigma}_2^T\hat{W}_1\right) + \frac{1}{4\gamma^2}\bar{\mathrm{E}}_1(x)\hat{W}_3 m^T\hat{W}_1\right\} \tag{47}$$

where $\bar{D}_1(x) \equiv \nabla\phi_1(x)g(x)R^{-1}g^T(x)\nabla\phi_1{}^T(x)$, $\bar{E}_1(x) \equiv \nabla\phi_1(x)kk^T\nabla\phi_1{}^T(x)$, $m \equiv \frac{\sigma_2}{(\sigma_2{}^T\sigma_2+1)^2}$, and $F_1 > 0, F_2 > 0, F_3 > 0, F_4 > 0$ are tuning parameters. Let Facts 1 hold and let $Q(x) > 0$. Suppose that $\bar{\sigma}_2 = \sigma_2/(\sigma_2^T\sigma_2+1)$ is persistently exciting. Let the tuning parameters be selected as detailed in the proof. Then there exists an N_0 such that, for the number of hidden layer units $N > N_0$ 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 3. See the comments following equation (24). Let $\varepsilon > 0$ and let N_0 be the number of hidden layer units above which $\sup\left\|\varepsilon_{HJI}\right\| < \varepsilon$. In the proof it is seen that the theorem holds for $N > N_0$.

Remark 4. The theorem shows that PE is needed for proper identification of the value function by the critic NN, and that nonstandard tuning algorithms are required for the actor and the disturbance NN to guarantee stability.

Remark 5. The assumption $Q(x) > 0$ is sufficient but not necessary for this result. If this condition is replaced by zero state observability, the proof still goes through, however it is tedious and does not add insight. The method used would be the technique used in the

proof of technical Lemma 2 Part a in (Vamvoudakis & Lewis), or the standard methods of (Ioannou & Fidan, 2006; Tao, 2003).

Remark 6. The tuning parameters F_1, F_2, F_3, F_4 in (46), and (47) must be selected to make the matrix M in (A.10) positive definite.

Theorem 3. Exponential Convergence and Nash equilibrium. Suppose the hypotheses of Theorem 1 and Theorem 2,. Then Theorem 1 holds with

$$\varepsilon_{\max} > \frac{1}{4}\left\|\tilde{W}_2\right\|^2\left\|\frac{\bar{D}_1}{m_s}\right\| - \frac{1}{4\gamma^2}\left\|\tilde{W}_3\right\|^2\left\|\frac{\bar{E}_1}{m_s}\right\| + \varepsilon\left\|\frac{1}{m_s}\right\|$$

where $m_s = {\sigma_2}^T\sigma_2 + 1$, so that exponential convergence of $\hat{W}_1$ to the approximate optimal critic value W_1 is obtained. Then:

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

$$\hat{u}_1 = -\frac{1}{2}R^{-1}g^T(x)\nabla\phi_1{}^T(x)\hat{W}_1 \tag{48}$$

$$\hat{d}_1 = \frac{1}{2\gamma^2}k^T(x)\nabla\phi_1{}^T(x)\hat{W}_1 \tag{49}$$

b. $\hat{u}(x),\hat{d}(x)$ (see (41) and (43)) converges to the approximate Nash equilibrium solution of the ZS game.

Proof. Consider the UUB weights $\tilde{W}_1$, $\tilde{W}_2$ and $\tilde{W}_3$ as proved in Theorem 2.

a. The approximate HJI equation is

$$H(x,\hat{W}_1) = Q(x) + \hat{W}_1{}^T\nabla\varphi_1(x)f(x) - \frac{1}{4}\hat{W}_1{}^TD_1\hat{W}_1 + \frac{1}{4\gamma^2}\hat{W}_1{}^TE_1\hat{W}_1 - \hat{\varepsilon}_{HJI} \tag{50}$$

After adding zero we have

$$H(x,\hat{W}_1) = \tilde{W}_1{}^T\nabla\varphi_1(x)f(x) - \frac{1}{4}\tilde{W}_1{}^TD_1\tilde{W}_1 - \frac{1}{2}\tilde{W}_1{}^TD_1\hat{W}_1 + \frac{1}{4\gamma^2}\tilde{W}_1{}^TE_1\tilde{W}_1 + \frac{1}{2\gamma^2}\tilde{W}_1{}^TE_1\hat{W}_1 - \varepsilon_{HJI} \tag{51}$$

But

$$\hat{W}_1 = -\tilde{W}_1 + W_1 \tag{52}$$

After taking norms in (52) and letting $\|W_1\| < W_{1\max}$ one has

$$\left\|\hat{W}_1\right\| = \left\|-\tilde{W}_1 + W_1\right\| \le \left\|\tilde{W}_1\right\| + \left\|W_1\right\| \le \left\|\tilde{W}_1\right\| + W_{\max} \tag{53}$$

Now (51) becomes by taking into account (53),

$$\left\|H(x,\hat{W}_1)\right\| \le \left\|\tilde{W}_1\right\|\left\|\nabla\varphi_1(x)\right\|\left\|f(x)\right\| - \frac{1}{4}\left\|\tilde{W}_1\right\|^2\left\|\bar{D}_1\right\| - \frac{1}{2}\left\|\tilde{W}_1\right\|\left\|\bar{D}_1\right\|\left(\left\|\tilde{W}_1\right\| + W_{\max}\right)$$

$$+\frac{1}{4\gamma^2}\left\|\tilde{W}_1\right\|^2\left\|\bar{E}_1\right\|+\frac{1}{2\gamma^2}\left\|\tilde{W}_1\right\|\left\|\bar{E}_1\right\|\left(\left\|\tilde{W}_1\right\|+W_{\max}\right)+\left\|\varepsilon_{HJI}\right\| \quad (54)$$

Let Facts 1 hold and also $\sup\left\|\varepsilon_{HJI}\right\|<\varepsilon$ then (54) becomes

$$\begin{aligned}\left\|H(x,\hat{W}_1)\right\| &\le b_{\phi_x}b_f\left\|\tilde{W}_1\right\|\|x\|+\frac{1}{4}\left\|\tilde{W}_1\right\|^2\left\|\bar{D}_1\right\|+\frac{1}{2}\left\|\tilde{W}_1\right\|\left\|\bar{D}_1\right\|\left(\left\|\tilde{W}_1\right\|+W_{\max}\right)\\ &+\frac{1}{4\gamma^2}\left\|\tilde{W}_1\right\|^2\left\|\bar{E}_1\right\|+\frac{1}{2\gamma^2}\left\|\tilde{W}_1\right\|\left\|\bar{E}_1\right\|\left(\left\|\tilde{W}_1\right\|+W_{\max}\right)+\varepsilon\end{aligned} \quad (55)$$

All the signals on the right hand side of (55) are UUB. So $\left\|H(x,\hat{W}_1)\right\|$ is UUB and convergence to the approximate HJI solution is obtained.

b. According to Theorem 1 and equations (39), (40) and (41), (43), $\left\|\hat{u}-u_1\right\|$ and $\left\|\hat{d}-d_1\right\|$ are UUB because $\left\|\hat{W}_2-W_1\right\|$ and $\left\|\hat{W}_3-W_1\right\|$ are UUB

So the pair $\hat{u}(x),\hat{d}(x)$ gives the Nash equilibrium solution of the zero-sum game.
This completes the proof.

■

Remark 7. The theorems make no mention of finding the minimum nonnegative solution to the HJI. However they do guarantee convergence to a solution $(u(x),d(x))$ such that $(f(x)+g(x)u(x)+k(x)d(x))$ is stable. This is only accomplished by the minimal nonnegative HJI solution. Practical implementation, in view of the Policy Iteration Algorithm, would start with initial weights of zero in the disturbance NN (43). NN usage suggests starting with the initial control NN weights in (41) randomly selected and nonzero.
Note that the dynamics is required to be known to implement this algorithm in that $\sigma_2=\nabla\phi_1(f+g\hat{u}+k\hat{d})$, $\bar{D}_1(x)$, $\bar{E}_1(x)$ and (41), (43) depend on *f(x), g(x), k(x).*

5. Simulations

Here we present simulations of a linear and a nonlinear system to show that the game can be solved ONLINE by learning in real time, using the method of this chapter. We also present Simulation B to show that that one learns FASTER if one has an opponent. *That is, the two-player online game converges faster than an equivalent online 1-player (optimal control) problem when all the NNs are tuned online in real time.*

5.1 Linear system

Consider the continuous-time F16 aircraft plant with quadratic cost function used in (Stevens & Lewis, 2003). 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 the elevator actuator voltage and the disturbance is wind gusts on angle of attack. One has the dynamics $\dot{x}=Ax+Bu+Kd$,

$$\dot{x}=\begin{bmatrix}-1.01887 & 0.90506 & -0.00215\\ 0.82225 & -1.07741 & -0.17555\\ 0 & 0 & -1\end{bmatrix}x+\begin{bmatrix}0\\0\\1\end{bmatrix}u+\begin{bmatrix}1\\0\\0\end{bmatrix}d$$

where Q and R in the cost function are identity matrices of appropriate dimensions and $\gamma = 5$. In this linear case the solution of the HJI equation is given by the solution of the game algebraic Riccati equation (GARE)

$$A^T P + PA + Q - PBR^{-1}B^T P + \frac{1}{\gamma^2} PKK^T P = 0$$

Since the value is quadratic in the LQR case, the critic NN basis set $\phi_1(x)$ was selected as the quadratic vector in the state components $x \otimes x$ with $\otimes$ the Kronecker product. Redundant terms were removed to leave $n(n+1)/2 = 6$ components. Solving the GARE gives the parameters of the optimal critic as $W_1^* = [1.6573 \quad 1.3954 \quad -0.1661 \quad 1.6573 \quad -0.1804 \; 0.4371]^T$ which are the components of the Riccati solution matrix *P*.

The synchronous zero-sum game PI algorithm is implemented as in Theorem 2. PE was ensured by adding a small probing noise to the control and the disturbance input. Figure 1 shows the critic parameters, denoted by $\hat{W}_1 = [W_{c1} \quad W_{c2} \quad W_{c3} \; W_{c4} \quad W_{c5} \quad W_{c6}]^T$ converging to the optimal values. In fact after 600s the critic parameters converged to $\hat{W}_1(t_f) = [1.7090 \quad 1.3303 \quad -0.1629 \quad 1.7354 \quad -0.1730 \; 0.4468]^T$. The actor parameters after 600s converge to the values of $\hat{W}_2(t_f) = [1.7090 \quad 1.3303 \quad -0.1629 \quad 1.7354 \quad -0.1730 \; 0.4468]^T$. The disturbance parameters after 600s converge to the values of $\hat{W}_3(t_f) = [1.7090 \quad 1.3303 \quad -0.1629 \quad 1.7354 \quad -0.1730 \; 0.4468]^T$.

Then, the actor NN is given as

$$\hat{u}_2(x) = -\tfrac{1}{2} R^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} 2x_1 & 0 & 0 \\ x_2 & x_1 & 0 \\ x_3 & 0 & x_1 \\ 0 & 2x_2 & 0 \\ 0 & x_3 & x_2 \\ 0 & 0 & 2x_3 \end{bmatrix}^T \begin{bmatrix} 1.7090 \\ 1.3303 \\ -0.1629 \\ 1.7354 \\ -0.1730 \\ 0.4468 \end{bmatrix}.$$

Then, the disturbance NN is given as

$$\hat{d}(x) = \tfrac{1}{2\gamma^2} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} 2x_1 & 0 & 0 \\ x_2 & x_1 & 0 \\ x_3 & 0 & x_1 \\ 0 & 2x_2 & 0 \\ 0 & x_3 & x_2 \\ 0 & 0 & 2x_3 \end{bmatrix}^T \begin{bmatrix} 1.7090 \\ 1.3303 \\ -0.1629 \\ 1.7354 \\ -0.1730 \\ 0.4468 \end{bmatrix}.$$

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.

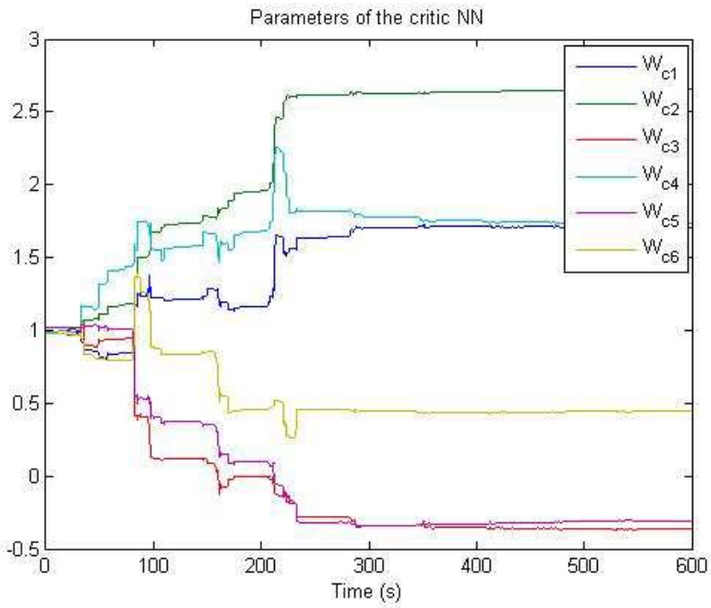


Fig. 1. Convergence of the critic parameters to the parameters of the optimal critic.

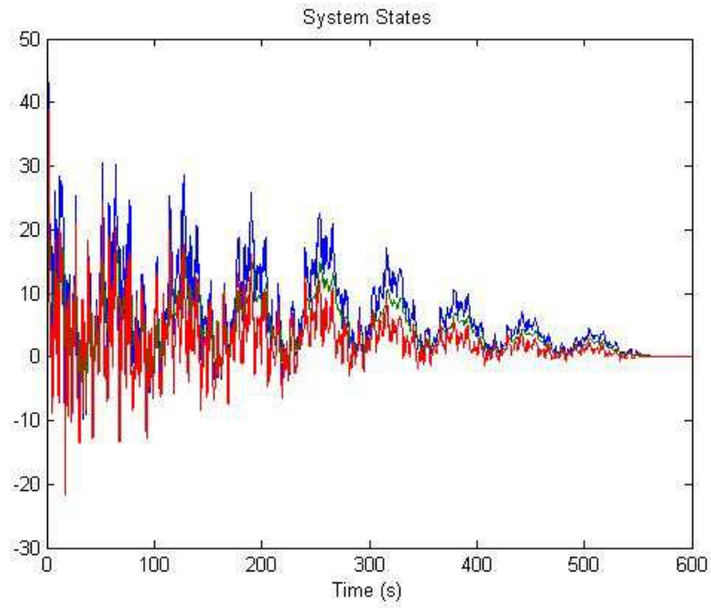


Fig. 2. Evolution of the system states for the duration of the experinment.

5.2 Single player linear system

The purpose of this example is to show that one learns FASTER if one has an opponent. That is, the online two-player game converges faster than an equivalent online 1-player (optimal control) problem. In this example, we use the method for online solution of the optimal control problem presented in (Vamvoudakis & Lewis, 2010). That is, Theorem 2 without the disturbance NN (47).

Consider the continuous-time F16 aircraft plant described before but with $d = 0$. Solving the ARE with Q and R identity matrices of appropriate dimensions, gives the parameters of the optimal critic as

$$W_1^* = [1.4245 \quad 1.1682 \quad -0.1352 \quad 1.4361 \quad -0.1516 \quad 0.4329]^T .$$

Figure 3 shows the critic parameters, denoted by $\hat{W}_1 = [W_{c1} \quad W_{c2} \quad W_{c3} \; W_{c4} \quad W_{c5} \quad W_{c6}]^T$ converging to the optimal values. In fact after 800s the critic parameters converged to $\hat{W}_1(t_f) = [1.4270 \quad 1.1654 \quad -0.1367 \quad 1.4387 \quad -0.1496 \; 0.4323]^T$. The actor parameters after 800s converge to the values of $\hat{W}_2(t_f) = [1.4270 \quad 1.1654 \quad -0.1367 \quad 1.4387 \quad -0.1496 \; 0.4323]^T$.
In comparison with part A, it is very clear that the two-player zero-sum game algorithm has faster convergence skills than the single-player game (e.g. optimal control problem) by a factor of two. As a conclusion the critic NN learns faster when there is an oponent for the control input, namely a disturbance.

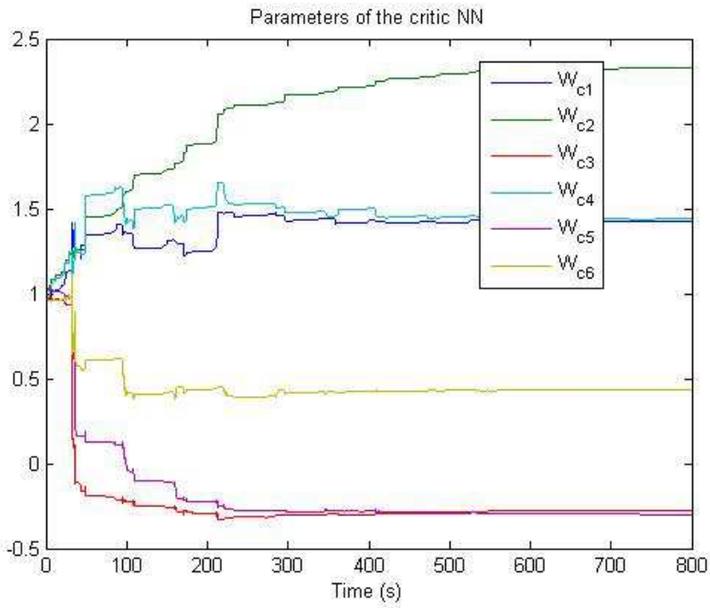


Fig. 3. Convergence of the critic parameters to the parameters of the optimal critic.

5.3 Nonlinear system

Consider the following affine in control input nonlinear system, with a quadratic cost constructed as in (Nevistic & Primbs, 1996; D. Vrabie, Vamvoudakis & Lewis, 2009)

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

where

$$f(x) = \begin{bmatrix} -x_1 + x_2 \\ -{x_1}^3 - {x_2}^3 + 0.25x_2(\cos(2x_1) + 2)^2 - 0.25x_2 \frac{1}{\gamma^2}(\sin(4x_1) + 2)^2 \end{bmatrix}$$

$$g(x) = \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix}, \quad k(x) = \begin{bmatrix} 0 \\ (\sin(4x_1) + 2) \end{bmatrix}.$$

One selects $Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $R = 1$, $\gamma = 8$.

The optimal value function is $V^*(x) = \frac{1}{4}{x_1}^4 + \frac{1}{2}{x_2}^2$ the optimal control signal is $u^*(x) = -(\cos(2x_1) + 2)x_2$ and $d^*(x) = \frac{1}{\gamma^2}(\sin(4x_1) + 2)x_2$.

One selects the critic NN vector activation function as

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

Figure 4 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 80s have

$$\hat{W}_1(t_f) = [0.0008 \quad 0.4999 \quad 0.2429 \quad 0.0032]^T$$

The actor parameters after 80s converge to the values of

$$\hat{W}_2(t_f) = [0.0008 \quad 0.4999 \quad 0.2429 \quad 0.0032]^T,$$

and the disturbance parameters after 300s converge to the values of

$$\hat{W}_3(t_f) = [0.0008 \quad 0.4999 \quad 0.2429 \quad 0.0032]^T.$$

So that the actor NN

$$\hat{u}_2(x) = -\tfrac{1}{2}R^{-1}\begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix}^T \begin{bmatrix} 2x_1 & 0 \\ 0 & 2x_2 \\ 4x_1^3 & 0 \\ 0 & 4x_2^3 \end{bmatrix}^T \begin{bmatrix} 0.0008 \\ 0.4999 \\ 0.2429 \\ 0.0032 \end{bmatrix}$$

also converged to the optimal control, and the disturbance NN

$$\hat{d}(x) = \tfrac{1}{2\gamma^2}\begin{bmatrix} 0 \\ \sin(4x_1) + 2 \end{bmatrix}^T \begin{bmatrix} 2x_1 & 0 \\ 0 & 2x_2 \\ 4x_1^3 & 0 \\ 0 & 4x_2^3 \end{bmatrix}^T \begin{bmatrix} 0.0008 \\ 0.4999 \\ 0.2429 \\ 0.0032 \end{bmatrix}$$

also converged to the optimal disturbance.

The evolution of the system states is presented in Figure 5. Figure 6 shows the optimal value function. The identified value function given by $\hat{V}_1(x) = \hat{W}_1^T \phi_1(x)$ is virtually indistinguishable from the exact solution and so is not plotted. In fact, Figure 7 shows the 3-

D plot of the difference between the approximated value function and the optimal one. This error is close to zero. Good approximation of the actual value function is being evolved. Figure 8 shows the 3-D plot of the difference between the approximated control, by using the online algorithm, and the optimal one. This error is close to zero.
Finally Figure 9 shows the 3-D plot of the difference between the approximated disturbance, by using the online algorithm, and the optimal one. This error is close to zero.

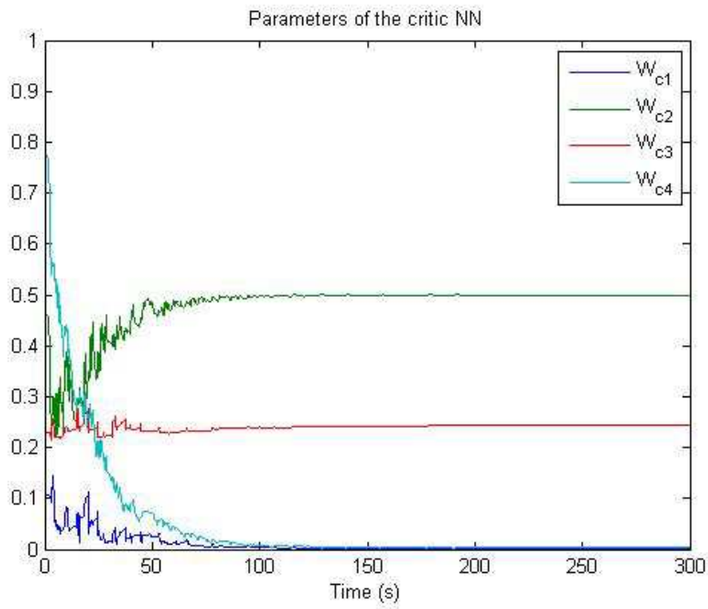


Fig. 4. Convergence of the critic parameters.

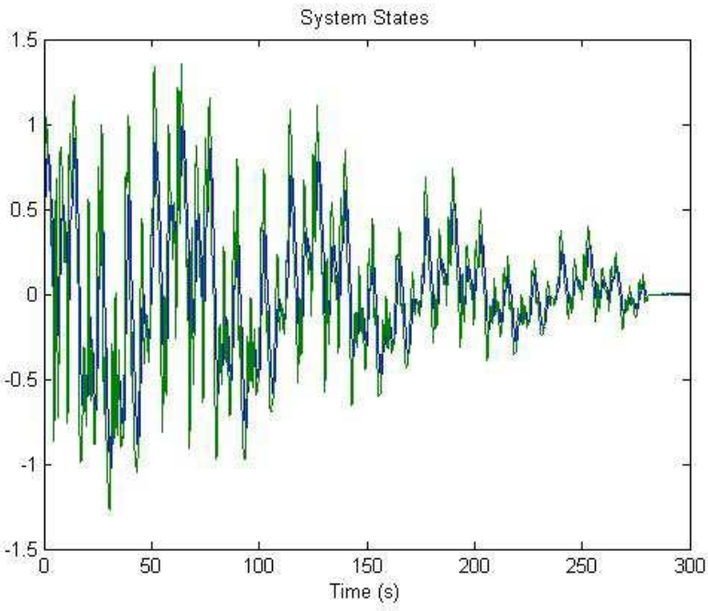


Fig. 5. Evolution of the system states.

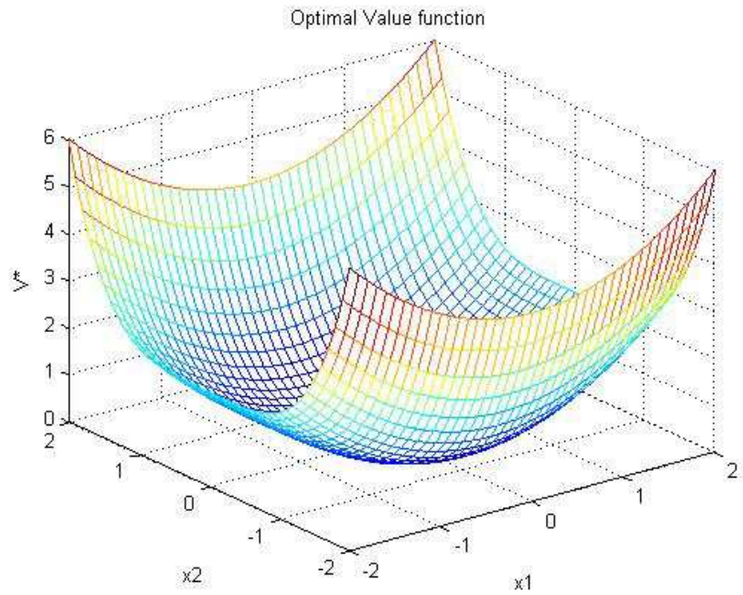


Fig. 6. Optimal Value function.

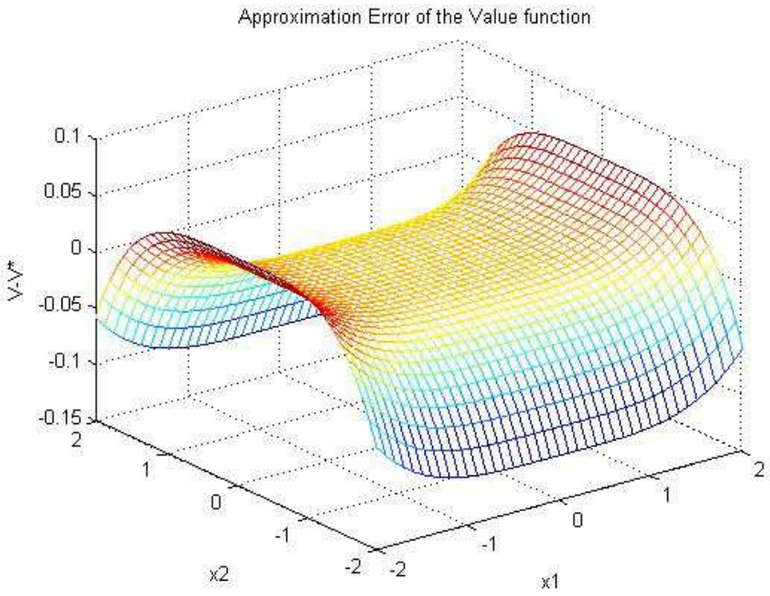


Fig. 7. 3D plot of the approximation error for the value function.

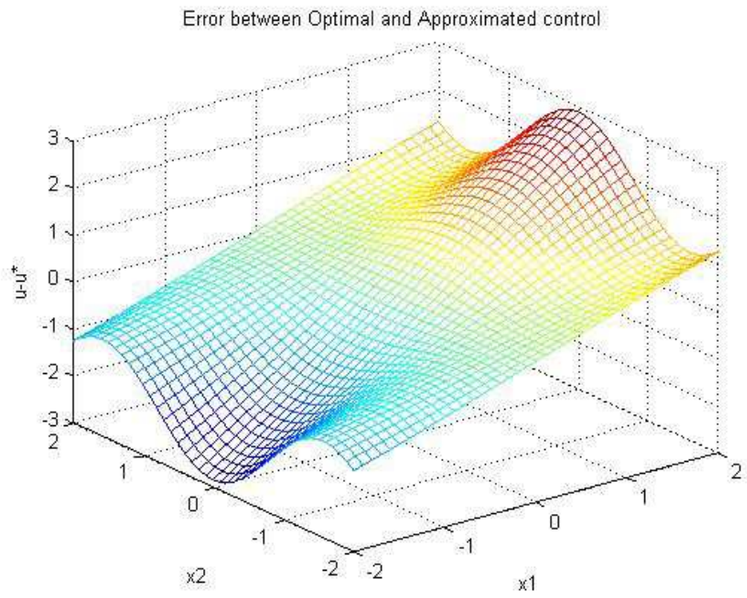


Fig. 8. 3D plot of the approximation error for the control.

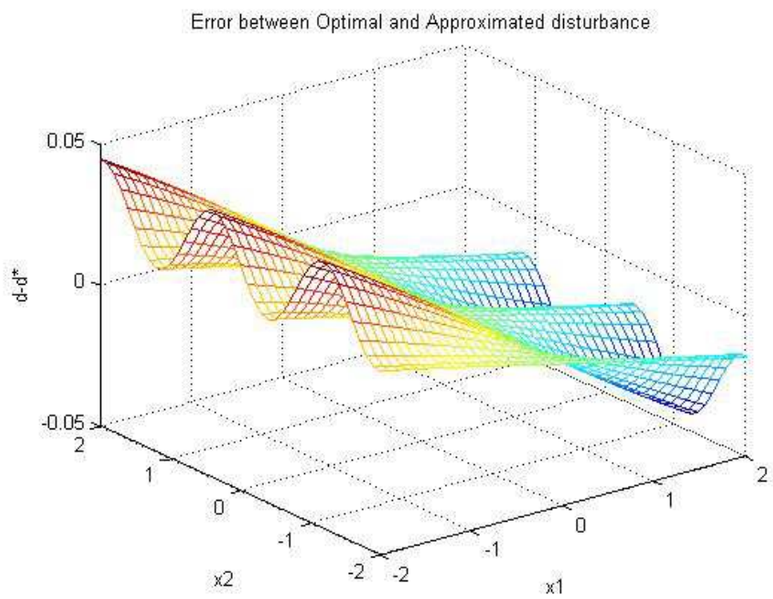


Fig. 9. 3D plot of the approximation error for the disturbance.

6. Appendix

Proof for Theorem 2: The convergence proof is based on Lyapunov analysis.
We consider the Lyapunov function

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

The derivative of the Lyapunov function is given by

$$\dot{L}(x) = \dot{V}(x) + \tilde{W}_1{}^T \alpha_1^{-1}\dot{\tilde{W}}_1 + \tilde{W}_2{}^T \alpha_2^{-1}\dot{\tilde{W}}_2 + \tilde{W}_3{}^T \alpha_3^{-1}\dot{\tilde{W}}_3 \tag{A.2}$$

First term is,

$$\dot{V}(x) = W_1{}^T\left(\nabla\phi_1 f(x) - \frac{1}{2}\bar{D}_1(x)\hat{W}_2 + \frac{1}{2\gamma^2}\bar{E}_1(x)\hat{W}_3\right) + \nabla\varepsilon^T(x)\left(f(x) - \frac{1}{2}g(x)R^{-1}g^T(x)\nabla\phi_1{}^T\hat{W}_2 + \frac{1}{2\gamma^2}kk^T\nabla\phi_1{}^T\hat{W}_3\right).$$

Then

$$\begin{aligned}\dot{V}(x) &= W_1{}^T\left(\nabla\phi_1 f(x) - \frac{1}{2}\bar{D}_1(x)\hat{W}_2 + \frac{1}{2\gamma^2}\bar{E}_1(x)\hat{W}_3\right) + \varepsilon_1(x)\\ &= W_1{}^T\nabla\phi_1 f(x) + \frac{1}{2}W_1{}^T\bar{D}_1(x)\left(W_1 - \hat{W}_2\right) - \frac{1}{2}W_1{}^T\bar{D}_1(x)W_1\\ &\quad - \frac{1}{2\gamma^2}W_1{}^T\bar{E}_1(x)\left(W_1 - \hat{W}_3\right) + \frac{1}{2\gamma^2}W_1{}^T\bar{\mathrm{E}}_1(x)W_1 + \varepsilon_1(x)\end{aligned}$$

$$= W_1{}^T\nabla\phi_1 f(x) + \frac{1}{2}W_1{}^T\bar{D}_1(x)\tilde{W}_2 - \frac{1}{2}W_1{}^T\bar{D}_1(x)W_1 - \frac{1}{2\gamma^2}W_1{}^T\bar{E}_1(x)\tilde{W}_3 + \frac{1}{2\gamma^2}W_1{}^T\bar{\mathrm{E}}_1(x)W_1 + \varepsilon_1(x)$$

$$= W_1{}^T\sigma_1 + \frac{1}{2}W_1{}^T\bar{D}_1(x)\tilde{W}_2 - \frac{1}{2\gamma^2}W_1{}^T\bar{E}_1(x)\tilde{W}_3 + \varepsilon_1(x)$$

where $\varepsilon_1(x) \equiv \dot{\varepsilon}(x) = \nabla\varepsilon^T(x)(f(x) - \frac{1}{2}g(x)R^{-1}g^T(x)\nabla\phi_1{}^T\hat{W}_2 + \frac{1}{2\gamma^2}kk^T\nabla\phi_1{}^T\hat{W}_3)$.

From the HJI equation $W_1{}^T\sigma_1 = -h^Th - \frac{1}{4}W_1{}^T\bar{D}_1(x)W_1 + \frac{1}{4\gamma^2}W_1{}^T\bar{E}_1(x)W_1 + \varepsilon_{HJI}(x)$.

Then

$$\dot{L}_V(x) = -h^Th - \frac{1}{4}W_1{}^T\bar{D}_1(x)W_1 + \frac{1}{4\gamma^2}W_1{}^T\bar{D}_1(x)W_1 + \frac{1}{2}W_1{}^T\bar{E}_1(x)\tilde{W}_2 - \frac{1}{2\gamma^2}W_1{}^T\bar{E}_1(x)\tilde{W}_3 + \varepsilon_{HJI}(x) + \varepsilon_1(x)$$

$$\equiv \dot{\bar{L}}_V(x) + \frac{1}{2}W_1{}^T\bar{D}_1(x)\tilde{W}_2 - \frac{1}{2\gamma^2}W_1{}^T\bar{E}_1(x)\tilde{W}_3 + \varepsilon_1(x). \tag{A.3}$$

where $\dot{\bar{L}}_V(x) = -h^T h - \frac{1}{4} W_1^T \bar{D}_1(x) W_1 + \frac{1}{4\gamma^2} W_1^T \bar{D}_1(x) W_1 + \varepsilon_{HJI}(x) + \varepsilon_1(x)$

Second term is,

$$\dot{L}_1 = \tilde{W}_1^T \alpha_1^{-1} \dot{\tilde{W}}_1 = \tilde{W}_1^T \alpha_1^{-1} \alpha_1 \frac{\sigma_2}{\left(\sigma_2^T \sigma_2 + 1\right)^2} (\sigma_2^T \hat{W}_1 + Q(x) + \frac{1}{4} \hat{W}_2^T \bar{D}_1 \hat{W}_2 - \frac{1}{4\gamma^2} \hat{W}_3^T \bar{E}_1 \hat{W}_3)$$

$$\dot{L}_1 = \tilde{W}_1^T \frac{\sigma_2}{\left(\sigma_2^T \sigma_2 + 1\right)^2} (-\sigma_2^T \tilde{W}_1 + \frac{1}{4} \tilde{W}_2^T \bar{D}_1(x) \tilde{W}_2 - \frac{1}{4\gamma^2} \tilde{W}_3^T \bar{E}_1 \tilde{W}_3 + \varepsilon_{HJI}(x))$$

$$= \dot{\bar{L}}_1 + \frac{1}{4} \tilde{W}_1^T \frac{\sigma_2}{\left(\sigma_2^T \sigma_2 + 1\right)^2} \left(\tilde{W}_2^T \bar{D}_1(x) \tilde{W}_2 - \frac{1}{\gamma^2} \tilde{W}_3^T \bar{E}_1 \tilde{W}_3 \right) \qquad \text{(A.4)}$$

where $\dot{\bar{L}}_1 = \tilde{W}_1^T \frac{\sigma_2}{\left(\sigma_2^T \sigma_2 + 1\right)^2} \left(-\sigma_2^T \tilde{W}_1 + \varepsilon_{HJI}(x) \right) = \tilde{W}_1^T \bar{\sigma}_2 \left(-\sigma_2^T \tilde{W}_1 + \frac{\varepsilon_{HJI}(x)}{m_s} \right)$.

By adding the terms of (A.3) and (A.4) we have

$$\dot{L}(x) = -Q(x) - \frac{1}{4} W_1^T \bar{D}_1(x) W_1 + \frac{1}{4\gamma^2} W_1^T \bar{E}_1(x) W_1 + \frac{1}{2} W_1^T \bar{D}_1(x) \tilde{W}_2 - \frac{1}{2\gamma^2} W_1^T \bar{E}_1(x) \tilde{W}_3 + \varepsilon_{HJI}(x) + \varepsilon_1(x)$$
$$+ \tilde{W}_1^T \frac{\sigma_2}{\left(\sigma_2^T \sigma_2 + 1\right)^2} (-\sigma_2^T \tilde{W}_1 + \frac{1}{4} \tilde{W}_2^T \bar{D}_1(x) \tilde{W}_2 - \frac{1}{4\gamma^2} \tilde{W}_3^T \bar{E}_1 \tilde{W}_3 + \varepsilon_{HJI}(x)) + \tilde{W}_2^T \alpha_2^{-1} \dot{\tilde{W}}_2 + \tilde{W}_3^T \alpha_3^{-1} \dot{\tilde{W}}_3$$

$$\dot{L}(x) = \dot{\bar{L}}_1 + \dot{\bar{L}}_V(x) + \varepsilon_1(x) - \frac{1}{4} \tilde{W}_2^T \bar{D}_1(x) W_1 \frac{\bar{\sigma}_2^T}{m_s} W_1 + \frac{1}{4} \tilde{W}_2^T \bar{D}_1(x) W_1 \frac{\bar{\sigma}_2^T}{m_s} \tilde{W}_1$$
$$+ \frac{1}{4} \tilde{W}_2^T \bar{D}_1(x) \tilde{W}_2 \frac{\bar{\sigma}_2^T}{m_s} W_1 + \frac{1}{4} \tilde{W}_2^T \bar{D}_1(x) \hat{W}_2 \frac{\bar{\sigma}_2^T}{m_s} \hat{W}_1 - \frac{1}{4\gamma^2} \tilde{W}_3^T \bar{E}_1(x) W_1 \frac{\bar{\sigma}_2^T}{m_s} \tilde{W}_1$$
$$- \frac{1}{4\gamma^2} \tilde{W}_3^T \bar{E}_1(x) \tilde{W}_3 \frac{\bar{\sigma}_2^T}{m_s} W_1 - \frac{1}{4\gamma^2} \tilde{W}_3^T \bar{E}_1(x) \hat{W}_3 \frac{\bar{\sigma}_2^T}{m_s} \hat{W}_1 + \frac{1}{2} \tilde{W}_2^T \bar{D}_1(x) W_1 - \frac{1}{2\gamma^2} \tilde{W}_3^T \bar{E}_1(x) W_1$$
$$+ \frac{1}{4\gamma^2} \tilde{W}_3^T \bar{E}_1(x) W_1 \frac{\bar{\sigma}_2^T}{m_s} W_1 - \tilde{W}_2^T \alpha_2^{-1} \dot{\hat{W}}_2 - \tilde{W}_3^T \alpha_3^{-1} \dot{\hat{W}}_3 \qquad \text{(A.5)}$$

where $\bar{\sigma}_2 = \frac{\sigma_2}{\sigma_2^T \sigma_2 + 1}$ and $m_s = \sigma_2^T \sigma_2 + 1$.

In order to select the update law for the action neural networks, write (A.5) as

$$\dot{L}(x) = \dot{\bar{L}}_V + \dot{\bar{L}}_1 + \varepsilon_1(x) - \tilde{W}_2^T \left[\alpha_2^{-1} \dot{\hat{W}}_2 - \tfrac{1}{4} \bar{D}_1(x) \hat{W}_2 \frac{\bar{\sigma}_2^T}{m_s} \hat{W}_1 \right] - \tilde{W}_3^T \left[\alpha_3^{-1} \dot{\hat{W}}_3 + \tfrac{1}{4\gamma^2} \bar{E}_1(x) \hat{W}_3 \frac{\bar{\sigma}_2^T}{m_s} \hat{W}_1 \right]$$

$$+\frac{1}{2}\tilde{W}_2{}^T\bar{D}_1(x)W_1+\frac{1}{4}\tilde{W}_2{}^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2{}^T}{m_s}\tilde{W}_1-\frac{1}{4}\tilde{W}_2{}^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2{}^T}{m_s}W_1+\frac{1}{4}\tilde{W}_2{}^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2}{m_s}\tilde{W}_2$$

$$-\frac{1}{2\gamma^2}\tilde{W}_3{}^T\bar{\mathrm{E}}_1(x)W_1-\frac{1}{4\gamma^2}\tilde{W}_3{}^T\bar{\mathrm{E}}_1(x)W_1\frac{\bar{\sigma}_2{}^T}{m_s}\tilde{W}_1+\frac{1}{4\gamma^2}\tilde{W}_3{}^T\bar{\mathrm{E}}_1(x)W_1\frac{\bar{\sigma}_2{}^T}{m_s}W_1-\frac{1}{4\gamma^2}\tilde{W}_3{}^T\bar{\mathrm{E}}_1(x)W_1\frac{\bar{\sigma}_2}{m_s}\tilde{W}_3$$

Now define the actor tuning law as

$$\dot{\hat{W}}_2=-\alpha_2\left\{\left(F_2\hat{W}_2-F_1\bar{\sigma}_2^T\hat{W}_1\right)-\frac{1}{4}\bar{D}_1(x)\hat{W}_2m^T\hat{W}_1\right\} \tag{A.6}$$

and the disturbance tuning law as

$$\dot{\hat{W}}_3=-\alpha_3\left\{\left(F_4\hat{W}_3-F_3\bar{\sigma}_2^T\hat{W}_1\right)+\frac{1}{4\gamma^2}\bar{\mathrm{E}}_1(x)\hat{W}_3m^T\hat{W}_1\right\}. \tag{A.7}$$

This adds to $\dot{L}$ the terms

$$\begin{aligned}&\tilde{W}_2^TF_2W_1-\tilde{W}_2^TF_2\tilde{W}_2-\tilde{W}_2^TF_1\bar{\sigma}_2^TW_1+\tilde{W}_2^TF_1\bar{\sigma}_2^T\tilde{W}_1\\&+\tilde{W}_3^TF_4W_1-\tilde{W}_3^TF_4\tilde{W}_3-\tilde{W}_3^TF_3\bar{\sigma}_2^TW_1+\tilde{W}_3^TF_3\bar{\sigma}_2^T\tilde{W}_1\end{aligned}$$

Overall

$$\begin{aligned}\dot{L}(x)=&-Q(x)-\frac{1}{4}W_1{}^T\bar{D}_1(x)W_1+\frac{1}{4\gamma^2}W_1{}^T\bar{E}_1(x)W_1+\varepsilon_{HJI}(x)+\tilde{W}_1{}^T\bar{\sigma}_2\left(-\bar{\sigma}_2{}^T\tilde{W}_1+\frac{\varepsilon_{HJI}(x)}{m_s}\right)+\varepsilon_1(x)\\&+\frac{1}{2}\tilde{W}_2{}^T\bar{D}_1(x)W_1+\frac{1}{4}\tilde{W}_2{}^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2{}^T}{m_s}\tilde{W}_1-\frac{1}{4}\tilde{W}_2{}^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2{}^T}{m_s}W_1\\&+\frac{1}{4}\tilde{W}_2{}^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2}{m_s}\tilde{W}_2-\frac{1}{2\gamma^2}\tilde{W}_3{}^T\bar{\mathrm{E}}_1(x)W_1-\frac{1}{4\gamma^2}\tilde{W}_3{}^T\bar{\mathrm{E}}_1(x)W_1\frac{\bar{\sigma}_2}{m_s}\tilde{W}_3\\&-\frac{1}{4\gamma^2}\tilde{W}_3{}^T\bar{\mathrm{E}}_1(x)W_1\frac{\bar{\sigma}_2{}^T}{m_s}\tilde{W}_1+\frac{1}{4\gamma^2}\tilde{W}_3{}^T\bar{\mathrm{E}}_1(x)W_1\frac{\bar{\sigma}_2{}^T}{m_s}W_1\\&+\tilde{W}_2^TF_2W_1-\tilde{W}_2^TF_2\tilde{W}_2-\tilde{W}_2^TF_1\bar{\sigma}_2^TW_1+\tilde{W}_2^TF_1\bar{\sigma}_2^T\tilde{W}_1\\&+\tilde{W}_3^TF_4W_1-\tilde{W}_3^TF_4\tilde{W}_3-\tilde{W}_3^TF_3\bar{\sigma}_2^TW_1+\tilde{W}_3^TF_3\bar{\sigma}_2^T\tilde{W}_1\end{aligned} \tag{A.8}$$

Now it is desired to introduce norm bounds. It is easy to show that under the Facts 1

$$\left\|\varepsilon_1(x)\right\|<b_{\varepsilon_x}b_f\left\|x\right\|+\tfrac{1}{2}b_{\varepsilon_x}b_g^2b_{\phi_x}\sigma_{\min}(R)\left(\left\|W_1\right\|+\left\|\tilde{W}_2\right\|\right)+\tfrac{1}{2\gamma^2}b_{\varepsilon_x}b_k^2b_{\phi_x}\left(\left\|W_1\right\|+\left\|\tilde{W}_3\right\|\right)$$

Also since $Q(x)>0$ there exists q such that $x^Tqx<Q(x)$ locally. It is shown in (Abu-Khalaf & Lewis, 2008; Abu-Khalaf et al. 2006) that ε_{HJI} converges to zero uniformly as N increases.

Select $\varepsilon > 0$ and $N_0(\varepsilon)$ such that $\sup\left\|\varepsilon_{HJI}\right\| < \varepsilon$ (see comments after (24)). Then assuming $N > N_0$ and writing in terms of $\tilde{Z} = \begin{bmatrix} x \\ \bar{\sigma}_2^{\mathrm{T}} \tilde{W}_1 \\ \tilde{W}_2 \\ \tilde{W}_3 \end{bmatrix}$, (A.8) becomes

$$\dot{L} < \frac{1}{4}\left\|W_1\right\|^2 \left\|\bar{D}_1(x)\right\| + \frac{1}{4\gamma^2}\left\|W_1\right\|^2 \left\|\bar{E}_1(x)\right\| + \varepsilon + \tfrac{1}{2}|\,|W_1|\,|\,b_{\varepsilon_x} b_{\varphi_x} b_g{}^2 \sigma_{\min}(R) + \tfrac{1}{2\gamma^2}|\,|W_1|\,|\,b_{\varepsilon_x} b_k^2 b_{\phi_x}$$

$$-\tilde{Z}^T \begin{bmatrix} qI & 0 & 0 & 0 \\ 0 & I & \left(\tfrac{1}{2}F_1 - \dfrac{1}{8m_s}\bar{D}_1 W_1\right)^T & \tfrac{1}{2}F_3 + \left(\dfrac{1}{8\gamma^2 m_s}\bar{\mathrm{E}}_1 W_1\right) \\ 0 & \tfrac{1}{2}F_1 - \left(\dfrac{1}{8m_s}\bar{D}_1 W_1\right) & F_2 - \tfrac{1}{8}\left(\bar{D}_1 W_1 m^{\mathrm{T}} + m W_1^T \bar{D}_1\right) & 0 \\ 0 & \tfrac{1}{2}F_3 + \left(\dfrac{1}{8\gamma^2 m_s}\bar{\mathrm{E}}_1 W_1\right) & 0 & F_4 + \tfrac{1}{8\gamma^2}\left(\bar{\mathrm{E}}_1 W_1 m^{\mathrm{T}} + m W_1^T \bar{\mathrm{E}}_1\right) \end{bmatrix} \tilde{Z}$$

$$+\tilde{Z}^T \begin{bmatrix} b_{\varepsilon_x} b_f \\ \dfrac{\varepsilon}{m_s} \\ (\tfrac{1}{2}\bar{D}_1 + F_2 - F_1\bar{\sigma}_2^T - \tfrac{1}{4}\bar{D}_1 W_1 m^T)W_1 + \tfrac{1}{2}b_{\varepsilon_x} b_g^2 b_{\phi_x}\sigma_{\min}(R) \\ (-\tfrac{1}{2\gamma^2}\bar{\mathrm{E}}_1 + F_4 - F_3\bar{\sigma}_2^T + \tfrac{1}{4\gamma^2}\bar{\mathrm{E}}_1 W_1 m^T)W_1 + \tfrac{1}{2\gamma^2}b_{\varepsilon_x} b_k^2 b_{\phi_x} \end{bmatrix} \tag{A.9}$$

Define

$$M = \begin{bmatrix} qI & 0 & 0 & 0 \\ 0 & I & \left(\tfrac{1}{2}F_1 - \dfrac{1}{8m_s}\bar{D}_1 W_1\right)^T & \tfrac{1}{2}F_3 + \left(\dfrac{1}{8\gamma^2 m_s}\bar{\mathrm{E}}_1 W_1\right) \\ 0 & \tfrac{1}{2}F_1 - \left(\dfrac{1}{8m_s}\bar{D}_1 W_1\right) & F_2 - \tfrac{1}{8}\left(\bar{D}_1 W_1 m^{\mathrm{T}} + m W_1^T \bar{D}_1\right) & 0 \\ 0 & \tfrac{1}{2}F_3 + \left(\dfrac{1}{8\gamma^2 m_s}\bar{\mathrm{E}}_1 W_1\right) & 0 & F_4 + \tfrac{1}{8\gamma^2}\left(\bar{\mathrm{E}}_1 W_1 m^{\mathrm{T}} + m W_1^T \bar{\mathrm{E}}_1\right) \end{bmatrix} \tag{A.10}$$

$$d = \begin{bmatrix} b_{\varepsilon_x} b_f \\ \dfrac{\varepsilon}{m_s} \\ (\tfrac{1}{2}\bar{D}_1 + F_2 - F_1\bar{\sigma}_2^T - \tfrac{1}{4}\bar{D}_1 W_1 m^T)W_1 + \tfrac{1}{2}b_{\varepsilon_x} b_g^2 b_{\phi_x}\sigma_{\min}(R) \\ (-\tfrac{1}{2\gamma^2}\bar{\mathrm{E}}_1 + F_4 - F_3\bar{\sigma}_2^T + \tfrac{1}{4\gamma^2}\bar{\mathrm{E}}_1 W_1 m^T)W_1 + \tfrac{1}{2\gamma^2}b_{\varepsilon_x} b_k^2 b_{\phi_x} \end{bmatrix}$$

$$c = \frac{1}{4}\|W_1\|^2\|\bar{D}_1(x)\| + \frac{1}{4\gamma^2}\|W_1\|^2\|\bar{E}_1(x)\| + \tfrac{1}{2}||W_1||b_{\varepsilon_x}b_{\varphi_x}b_g{}^2\sigma_{\min}(R) + \tfrac{1}{2\gamma^2}||W_1||b_{\varepsilon_x}b_k^2b_{\phi_x}$$

Let the parameters be chosen such that $M > 0$. Now (A.9) becomes

$$\dot{L} < -\|\tilde{Z}\|^2\sigma_{\min}(M) + \|d\|\|\tilde{Z}\| + c + \varepsilon$$

Completing the squares, the Lyapunov derivative is negative if

$$\|\tilde{Z}\| > \frac{\|d\|}{2\sigma_{\min}(M)} + \sqrt{\frac{d^2}{4\sigma_{\min}^2(M)} + \frac{c+\varepsilon}{\sigma_{\min}(M)}} \equiv B_Z. \tag{A.11}$$

It is now straightforward to demonstrate that if L exceeds a certain bound, then, $\dot{L}$ is negative. Therefore, according to the standard Lyapunov extension theorem (Lewis, Jagannathan, Yesildirek, 1999) the analysis above demonstrates that the state and the weights are UUB.
To show this from (A.1), one has,

$$\sigma_{\min}(P)\|x\|^2 + \frac{1}{2a_1}\|\tilde{W}_1\|^2 + \frac{1}{2a_2}\|\tilde{W}_2\|^2 + \frac{1}{2a_3}\|\tilde{W}_3\|^2 \le L \le$$

$$\le \sigma_{\max}(P)\|x\|^2 + \frac{1}{2a_1}\|\tilde{W}_1\|^2 + \frac{1}{2a_2}\|\tilde{W}_2\|^2 + \frac{1}{2a_3}\|\tilde{W}_3\|^2 \tag{A.12}$$

$$\tilde{Z}^T\underbrace{\begin{bmatrix}\sigma_{\min}(P) & & \\ & \frac{1}{2\|\bar{\sigma}_2\|^2 a_1} & \\ & & 1/2a_2 \\ & & & 1/2a_3\end{bmatrix}}_{S_1}\tilde{Z} \le L \le \tilde{Z}^T\underbrace{\begin{bmatrix}\sigma_{\max}(P) & & \\ & \frac{1}{2\|\bar{\sigma}_2\|^2 a_1} & \\ & & 1/2a_2 \\ & & & 1/2a_3\end{bmatrix}}_{S_2}\tilde{Z} \tag{A.13}$$

Equation (A.13) is equivalent to

$$\tilde{Z}^T\sigma_{\min}(S_1)\tilde{Z} \le L \le \tilde{Z}^T\sigma_{\max}(S_2)\tilde{Z}$$

Then

$$\sigma_{\min}(S_1)\|\tilde{Z}\|^2 \le L \le \sigma_{\max}(S_2)\|\tilde{Z}\|^2.$$

Therefore,
$$L > \sigma_{\max}(S_2)\left(\frac{\|d\|}{2\sigma_{\min}(M)} + \sqrt{\frac{\|d\|^2}{4\sigma_{\min}^2(M)} + \frac{c+\bar{\varepsilon}_1+\bar{\varepsilon}_2}{\sigma_{\min}(M)}}\right)^2 \tag{A.14}$$

implies (A.11).

Note that condition (A.11) holds if the norm of any component of $\tilde{Z}$ exceeds the bound, i.e. specifically $x > B_Z$ or $\bar{\sigma}_2^{\mathrm{T}}\tilde{W}_1 > B_Z$ or $\tilde{W}_2 > B_Z$ or $\tilde{W}_3 > B_Z$ (Khalil, 1996).
Now consider the error dynamics and the output as in Technical Lemmas 1, 2 and assume $\bar{\sigma}_2$ is persistently exciting

$$\dot{\tilde{W}}_1 = -a_1\bar{\sigma}_2\bar{\sigma}_2^{\mathrm{T}}\tilde{W}_1 + a_1\bar{\sigma}_2\frac{\varepsilon_{HJI}}{m_s} + \frac{a_1}{4m_s^{\,2}}\tilde{W}_2^{\,T}\bar{D}_1(x)\tilde{W}_2 - \frac{a_1}{4\gamma^2 m_s^{\,2}}\tilde{W}_3^{\,T}\bar{\mathrm{E}}_1\tilde{W}_3$$

$$y = \bar{\sigma}_2^{\,T}\tilde{W}_1 . \tag{A.15}$$

Then Theorem 1 is true with

$$\left\|\frac{\sigma_2^{\,T}}{m_s}\tilde{W}_1\right\| > \varepsilon_{\max} > \frac{1}{4}\left\|\tilde{W}_2\right\|^2\left\|\frac{\bar{D}_1}{m_s}\right\| - \frac{1}{4\gamma^2}\left\|\tilde{W}_3\right\|^2\left\|\frac{\bar{\mathrm{E}}_1}{m_s}\right\| + \varepsilon\left\|\frac{1}{m_s}\right\|$$

This provides an effective practical bound for $\left\|\bar{\sigma}_2^{\,T}\tilde{W}_1\right\|$.

This completes the proof.

■

6. References

Abu-Khalaf, M. & Lewis, F. L. (2005). Nearly Optimal Control Laws for Nonlinear Systems with Saturating Actuators Using a Neural Network HJB Approach. *Automatica*, Vol. 41, No. 5, pp. 779-791.

Abu-Khalaf, M. & Lewis, F. L. (2008). Neurodynamic Programming and Zero-Sum Games for Constrained Control Systems. *IEEE Transactions on Neural Networks*, Vol. 19, No. 7, pp. 1243-1252.

Abu-Khalaf, M.; Lewis, F. L. & Huang, J. (2006). Policy Iterations on the Hamilton-Jacobi-Isaacs Equation for H_∞ State Feedback Control With Input Saturation. *IEEE Transactions on Automatic Control*, Vol. 51, No. 12, pp. 1989-1995.

Adams R. & Fournier J. (2003). *Sobolev spaces*, New York: Academic Press.

Ball J.& Helton W. (1996). Viscosity solutions of Hamilton-Jacobi equations arising in nonlinear H_∞-control. *J. Math Syst., Estimat., Control*, Vol. 6, No.1, pp. 1-22.

Bardi M. & Capuzzo-Dolcetta I. (1997). *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*. Boston, MA: Birkhäuser.

Başar T. & Olsder G. J. (1999). *Dynamic Noncooperative Game Theory*, 2nd ed. Philadelphia, PA: SIAM, Vol. 23, SIAM's Classic in Applied Mathematics.

Başar T. & Bernard P. (1995). *H_∞ Optimal Control and Related Minimax Design Problems*. Boston, MA: Birkhäuser.

Bertsekas D. P. & Tsitsiklis J. N. (1996). *Neuro-Dynamic Programming*, Athena Scientific, MA.

Doya K. (2000). Reinforcement Learning In Continuous Time and Space. *Neural Computation*, Vol. 12, No. 1, pp. 219-245.

Doya K., Kimura H. & Kawato M. (2001). Neural Mechanisms of Learning and Control. *IEEE Control Syst. Mag.*, Vol. 21, No. 4, pp. 42-54.

Feng Y., Anderson B. D. & M. Rotkowitz. (2009). A game theoretic algorithm to compute local stabilizing solutions to HJBI equations in nonlinear H∞ control. *Automatica*, Vol. 45, No. 4, pp. 881-888.

Finlayson B. A. (1990). *The method of weighted residuals and variational principles*. New York: Academic Press, 1990.

Hanselmann T., Noakes L. & Zaknich A. (2007). Continuous-Time Adaptive Critics. *IEEE Transactions on Neural Networks*, Vol. 18, No. 3, pp. 631-647.

Hornik K., Stinchcombe M. & White H. (1990). Universal Approximation of an unknown mapping and its derivatives using multilayer feedforward networks, *Neural Networks*, Vol. 3, pp. 551-560.

Howard R. A. (1960). *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, Massachusetts.

Ioannou P. & Fidan B. (2006). *Adaptive Control Tutorial*, SIAM, Advances in Design and Control, PA.

Khalil H. K. (1996). *Nonlinear Systems*. Prentice-Hall.

Kleinman D. (1968). On an Iterative Technique for Riccati Equation Computations. *IEEE Transactions on Automatic Control*, Vol. 13, pp. 114- 115, February.

Lewis F.L., Jagannathan S., Yesildirek A. (1999). *Neural Network Control of Robot Manipulators and Nonlinear Systems*. Taylor & Francis.

Lewis F. L., Syrmos V. L. (1995). *Optimal Control*, John Wiley.

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

Van Der Shaft A. J. (1992). L_2-gain analysis of nonlinear systems and nonlinear state feedback H_∞ control. *IEEE Transactions on Automatic Control*, Vol. 37, No. 6, pp. 770-784.

Sandberg E. W. (1997). Notes on uniform approximation of time varying systems on finite time intervals. *IEEE Transactions on Circuits and Systems-1: Fundamental Theory and Appricatio*ns, Vol. 45, No. 8, pp. 863-865.

Stevens B. & Lewis F. L. (2003). *Aircract Control and Simulation,* 2nd edition, John Willey, New Jersey.

Si J., Barto A., Powel W. & Wunch D. (2004). *Handbook of Learning and Approximate Dynamic Programming*, John Wiley, New Jersey.

Sontag E. D. & Sussman H. J. (1995). Nonsmooth control Lyapunov functions. *IEEE Proc. CDC95*, pp. 2799-2805.

Sutton R. S. & Barto A. G. (1998). *Reinforcement Learning – An Introduction*, MIT Press, Cambridge, Massachusetts.

Tao G. (2003). *Adaptive Control Design and Analysis*, Adaptive and Learning Systems for Signal Processing, Communications and Control Series, Hoboken, NJ: Wiley-Interscience.

Tijs S. (2003). *Introduction to Game Theory*, Hindustan Book Agency, India.

Vamvoudakis K. G. & Lewis F. L. (2010). Online Actor-Critic Algorithm to Solve the Continuous-Time Infinite Horizon Optimal Control Problem. *Automatica*, Vol. 46, No. 5, pp. 878-888.

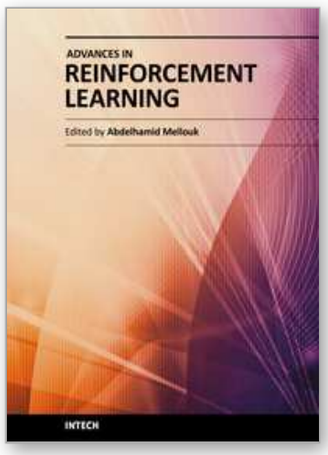
Vrabie D., Pastravanu O., Lewis F. & Abu-Khalaf M. (2009). Adaptive Optimal Control for Continuous-Time Linear Systems Based on Policy Iteration. *Automatica*, Vol. 45, No. 2, pp. 477-484.

Vrabie D., Vamvoudakis K. & Lewis F. (2009). Adaptive optimal controllers based on generalized policy iteration in a continuous-time framework. *Proc. of the IEEE Mediterranean Conf. on Control and Automation*, pp. 1402-1409.

Vrabie D. (2009) *Online Adaptive Optimal Control for Continuous Time Systems*, Ph.D. Thesis, Dept. of Electrical Engineering, Univ. Texas at Arlington, Arlington, TX, USA.

Werbos P.J. (1974). *Beyond Regression: New Tools for Prediction and Analysis in the Behavior Sciences*, Ph.D. Thesis.

Werbos P. J. (1992). 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.



Advances in Reinforcement Learning

Edited by Prof. Abdelhamid Mellouk

ISBN 978-953-307-369-9
Hard cover, 470 pages
Publisher InTech
Published online 14, January, 2011
Published in print edition January, 2011

Reinforcement Learning (RL) is a very dynamic area in terms of theory and application. This book brings together many different aspects of the current research on several fields associated to RL which has been growing rapidly, producing a wide variety of learning algorithms for different applications. Based on 24 Chapters, it covers a very broad variety of topics in RL and their application in autonomous systems. A set of chapters in this book provide a general overview of RL while other chapters focus mostly on the applications of RL paradigms: Game Theory, Multi-Agent Theory, Robotic, Networking Technologies, Vehicular Navigation, Medicine and Industrial Logistic.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Kyriakos G. Vamvoudakis and Frank L. Lewis (2011). Online Gaming: Real Time Solution of Nonlinear Two-Player Zero-Sum Games Using Synchronous Policy Iteration, Advances in Reinforcement Learning, Prof. Abdelhamid Mellouk (Ed.), ISBN: 978-953-307-369-9, InTech, Available from: http://www.intechopen.com/books/advances-in-reinforcement-learning/online-gaming-real-time-solution-of-nonlinear-two-player-zero-sum-games-using-synchronous-policy-ite

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen