Automatica 46 (2010) 878–888

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem[☆]

Kyriakos G. Vamvoudakis[*], Frank L. Lewis

Automation and Robotics Research Institute, The University of Texas at Arlington, 7300 Jack Newell Blvd. S., Ft. Worth, TX 76118, USA

ARTICLE INFO

Article history:
Received 3 May 2009
Received in revised form
19 November 2009
Accepted 13 February 2010
Available online 17 March 2010

Keywords:
Synchronous policy iteration
Adaptive critics
Adaptive control
Persistence of excitation
LQR

ABSTRACT

In this paper we discuss an online algorithm based on policy iteration for learning the continuous-time (CT) optimal control solution with infinite horizon cost for nonlinear systems with known dynamics. That is, the algorithm learns online in real-time the solution to the optimal control design HJ equation. This method finds in real-time suitable approximations of both the optimal cost and the optimal control policy, while also guaranteeing closed-loop stability. We present an online adaptive algorithm implemented as an actor/critic structure which involves simultaneous continuous-time adaptation of both actor and critic neural networks. We call this 'synchronous' policy iteration. A persistence of excitation condition is shown to guarantee convergence of the critic to the actual optimal value function. Novel tuning algorithms are given for both critic and actor networks, with extra nonstandard terms in the actor tuning law being required to guarantee closed-loop dynamical stability. The convergence to the optimal controller is proven, and the stability of the system is also guaranteed. Simulation examples show the effectiveness of the new algorithm.

1. Introduction

Optimal control (Lewis & Syrmos, 1995) has emerged as one of the fundamental design philosophies of modern control systems design. Optimal control policies satisfy the specified system performance while minimizing a structured cost index which describes the balance between desired performance and available control resources.

From a mathematical point of view the solution of the optimal control problem is based on the solution of the underlying Hamilton–Jacobi–Bellman (HJB) equation. Until recently, due to the intractability of this nonlinear differential equation for continuous-time (CT) systems, which form the object of interest in this paper, only particular solutions were available (e.g. for the linear time-invariant case, the HJB becomes the Riccati equation). For this reason considerable effort has been devoted to developing algorithms which approximately solve this equation (Abu-Khalaf & Lewis, 2005; Beard, Saridis, & Wen, 1997; Murray, Cox, Lendaris, & Saeks, 2002). Far more results are available for the solution of the discrete-time HJB equation. Good overviews are given in Bertsekas and Tsitsiklis (1996), Si, Barto, Powel, and Wunsch (2004) and Werbos (1974, 1989, 1992).

Some of the methods involve a computational intelligence technique known as Policy Iteration (PI) (Howard, 1960; Sutton & Barto, 1998). PI refers to a class of algorithms built as a two-step iteration: *policy evaluation* and *policy improvement*. Instead of trying a direct approach to solving the HJB equation, the PI algorithm starts by evaluating the cost of a given initial admissible (in a sense to be defined herein) control policy. This is often accomplished by solving a nonlinear Lyapunov equation. This new cost is then used to obtain a new improved (i.e. which will have a lower associated cost) control policy. This is often accomplished by minimizing a Hamiltonian function with respect to the new cost. (This is the so-called 'greedy policy' with respect to the new cost.) These two steps of policy evaluation and policy improvement are repeated until the policy improvement step no longer changes the actual policy, thus convergence to the optimal controller is achieved. One must note that the infinite horizon cost can be evaluated only in the case of admissible control policies, which requires that the policy be stabilizing. Admissibility is in fact a condition for the control policy which is used to initialize the algorithm.

Werbos defined actor–critic online learning algorithms to solve the optimal control problem based on so-called Value Iteration (VI), which does not require an initial stabilizing control policy (Werbos, 1974, 1989, 1992). He defined a family of VI

[☆] This work was supported by the National Science Foundation ECS-0501451 and the Army Research Office W91NF-05-1-0314. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Derong Liu under the direction of Editor Miroslav Krstic.

[*] Corresponding author. Tel.: +1 817 272 5938; fax: +1 817 272 5989.
E-mail addresses: kyriakos@arri.uta.edu (K.G. Vamvoudakis), lewis@uta.edu (F.L. Lewis).

0005-1098/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2010.02.018

algorithms which he termed Adaptive Dynamic Programming (ADP) algorithms. He used a critic neural network (NN) for value function approximation (VFA) and an actor NN for approximation of the control policy. Adaptive critics have been described in Prokhorov and Wunsch (1997) for discrete-time systems and Baird (1994), Hanselmann, Noakes, and Zaknich (2007), Vrabie and Lewis (2008) and Vrabie, Vamvoudakis, and Lewis (2009) for continuous-time systems.

Generalized Policy Iteration has been discussed in Sutton and Barto (1998). This is a family of optimal learning techniques which has PI at one extreme. In generalized PI, at each step one does not completely evaluate the cost of a given control, but only updates the current cost estimate *towards* that value. Likewise, one does not fully update the control policy to the greedy policy for the new cost estimate, but only updates the policy *towards* the greedy policy. Value Iteration in fact belongs to the family of generalized PI techniques.

In the linear CT system case, when quadratic indices are considered for the optimal stabilization problem, the HJB equation becomes the well known Riccati equation and the policy iteration method is in fact Newton's method proposed by Kleinman (1968), which requires iterative solutions of Lyapunov equations. In the case of nonlinear systems, successful application of the PI method was limited until Beard et al. (1997), where Galerkin spectral approximation methods were used to solve the nonlinear Lyapunov equations describing the policy evaluation step in the PI algorithm. Such methods are known to be computationally intensive. These are all offline methods for PI.

The key to solving practically the CT nonlinear Lyapunov equations was in the use of neural networks (NN) (Abu-Khalaf & Lewis, 2005) which can be trained to become approximate solutions of these equations. In fact the PI algorithm for CT systems can be built on Werbos' actor/critic structure which involves two neural networks: the critic NN, is trained to approximate the solution of the nonlinear Lyapunov equation at the policy evaluation step, while the actor neural network is trained to approximate an improving policy at the policy improving step. The method of Abu-Khalaf and Lewis (2005) is also an offline method.

In Vrabie and Lewis (2008) and Vrabie, Pastravanu, Lewis, and Abu-Khalaf (2009) was developed an online PI algorithm for continuous-time systems which converges to the optimal control solution without making explicit use of any knowledge on the internal dynamics of the system. The algorithm was based on *sequential updates* of the critic (policy evaluation) and actor (policy improvement) neural networks. That is, while one NN is tuned the other one remains constant.

This paper is concerned with developing an online approximate solution, based on PI, for the infinite horizon optimal control problem for continuous-time nonlinear systems with known dynamics. We present an online adaptive algorithm which involves *simultaneous tuning* of both actor and critic neural networks (i.e. both neural networks are tuned at the same time). We term this algorithm 'synchronous' policy iteration. This approach is an extremal version of the generalized Policy Iteration introduced in Sutton and Barto (1998).

This approach to policy iteration is motivated by work in adaptive control (Ioannou & Fidan, 2006; Tao, 2003). Adaptive control is a powerful tool that uses online tuning of parameters to provide effective controllers for nonlinear or linear systems with modeling uncertainties and disturbances. Closed-loop stability while learning the parameters is guaranteed, often by using Lyapunov design techniques. Parameter convergence, however, often requires that the measured signals carry sufficient information about the unknown parameters (persistence of excitation condition).

There are two main contributions in this paper. The first involves introduction of a nonstandard 'normalized' critic neural network tuning algorithm, along with guarantees for its convergence based on a persistence of excitation condition regularly required in adaptive control. The second involves adding nonstandard extra terms to the actor neural network tuning algorithm that are required to guarantee closed-loop stability, along with stability and convergence proofs.

The paper is organized as follows. Section 2 provides the formulation of the optimal control problem, followed by the general description of policy iteration and neural network value function approximation. Section 3 discusses tuning of the critic NN, in effect designing an observer for the unknown value function. Section 4 presents the online synchronous PI method, and shows how to simultaneously tune the critic and actor NNs to guarantee convergence and closed-loop stability. Results for convergence and stability are developed using a Lyapunov technique. Section 5 presents simulation examples that show the effectiveness of the online synchronous CT PI algorithm in learning the optimal value and control for both linear systems and nonlinear systems.

2. The optimal control problem and value function approximation

2.1. Optimal control and the continuous-time HJB equation

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

$$\dot{x}(t) = f(x(t)) + g(x(t))\, u(x(t)); \quad x(0) = x_0 \tag{1}$$

with state $x(t) \in R^n$, $f(x(t)) \in R^n$, $g(x(t)) \in R^{n\times m}$ and control input $u(t) \in R^m$. We assume that, $f(0) = 0$, $f(x) + g(x)u$ is Lipschitz continuous on a set $\Omega \subseteq R^n$ that contains the origin, and that the system is stabilizable on Ω, *i.e.* there exists a continuous control function $u(t) \in U$ such that the system is asymptotically stable on Ω. The system dynamics $f(x)$, $g(x)$ are assumed known.

Define the infinite horizon integral cost

$$V(x_0) = \int_0^\infty r(x(\tau), u(\tau))\mathrm{d}\tau \tag{2}$$

where $r(x, u) = Q(x) + u^T Ru$ with $Q(x)$ positive definite, *i.e.* $\forall x \neq 0, Q(x) > 0$ and $x = 0 \Rightarrow Q(x) = 0$, and $R \in R^{m\times m}$ a symmetric positive definite matrix.

Definition 1 (*Abu-Khalaf & Lewis, 2005, Admissible Policy*). A control policy $\mu(x)$ is defined as admissible with respect to (2) on Ω, denoted by $\mu \in \Psi(\Omega)$, if $\mu(x)$ is continuous on Ω, $\mu(0) = 0$, $u(x) = \mu(x)$ stabilizes (1) on Ω, and $V(x_0)$ is finite $\forall x_0 \in \Omega$.

For any admissible control policy $\mu \in \Psi(\Omega)$, if the associated cost function

$$V^\mu(x_0) = \int_0^\infty r(x(\tau), \mu(x(\tau)))\mathrm{d}\tau \tag{3}$$

is C^1, then an infinitesimal version of (3) is the so-called nonlinear Lyapunov equation

$$0 = r(x, \mu(x)) + \left(V_x^\mu\right)^T (f(x) + g(x)\mu(x)), \quad V^\mu(0) = 0 \tag{4}$$

where V_x^μ denotes the partial derivative of the value function V^μ with respect to x. (Note that the value function does not depend explicitly on time.)

We define the gradient here as a column vector, and use at times the alternative operator notation $\nabla \equiv \partial/\partial x$.

Eq. (4) is a Lyapunov equation for nonlinear systems which, given a controller $\mu(x) \in \Psi(\Omega)$, can be solved for the value function $V^\mu(x)$ associated with it. Given that $\mu(x)$ is an admissible control policy, if $V^\mu(x)$ satisfies (4), with, then $V^\mu(x)$ is a Lyapunov function for the system (1) with control policy $\mu(x)$.

The optimal control problem can now be formulated: Given the continuous-time system (1), the set $\mu \in \Psi(\Omega)$ of admissible control policies and the infinite horizon cost functional (2), find an admissible control policy such that the cost index (2) associated with the system (1) is minimized.

Defining the Hamiltonian of the problem

$$H(x,u,V_x)=r(x(t),u(t))+V_x^T(f(x(t))+g(x(t))\mu(t)),\tag{5}$$

the optimal cost function $V^*(x)$ defined by

$$V^*(x_0)=\min_{\mu\in\Psi(\Omega)}\left(\int_0^\infty r(x(\tau),\mu(x(\tau)))\mathrm{d}\tau\right)$$

with $x_0=x$ is known as the *value function*, and satisfies the HJB equation

$$0=\min_{\mu\in\Psi(\Omega)}[H(x,\mu,V_x^*)].\tag{6}$$

Assuming that the minimum on the right hand side of (6) exists and is unique then the optimal control function for the given problem is

$$\mu^*(x)=-\frac{1}{2}R^{-1}g^T(x)V_x^*(x).\tag{7}$$

Inserting this optimal control policy in the nonlinear Lyapunov equation we obtain the formulation of the HJB equation in terms of V_x^*

$$\begin{aligned}&0=Q(x)+V_x^{*T}(x)f(x)-\frac{1}{4}V_x^{*T}(x)g(x)R^{-1}g^T(x)V_x^*(x)\\&V^*(0)=0.\end{aligned}\tag{8}$$

For the linear system case, considering a quadratic cost functional, the equivalent of this HJB equation is the well known Riccati equation.

In order to find the optimal control solution for the problem one only needs to solve the HJB equation (8) for the value function and then substitute the solution in (7) to obtain the optimal control. However, due to the nonlinear nature of the HJB equation finding its solution is generally difficult or impossible.

2.2. Policy iteration

The approach of synchronous policy iteration used in this paper is motivated by Policy iteration (PI) (Sutton & Barto, 1998). Therefore in this section we describe PI.

Policy iteration (PI) (Sutton & Barto, 1998) is an iterative method of reinforcement learning (Doya, 2000) for solving optimal control problems, and consists of policy evaluation based on (4) and policy improvement based on (7). Specifically, the PI algorithm consists in solving iteratively the following two equations:

Policy iteration algorithm:

1. given $\mu^{(i)}(x)$, solve for the value $V^{\mu^{(i)}}(x(t))$ using

$$\begin{aligned}&0=r(x,\mu^{(i)}(x))+(\nabla V^{\mu^{(i)}})^T(f(x)+g(x)\mu^{(i)}(x))\\&V^{\mu^{(i)}}(0)=0\end{aligned}\tag{9}$$

2. update the control policy using

$$\mu^{(i+1)}=\arg\min_{u\in\Psi(\Omega)}[H(x,u,\nabla V_x^{(i)})],\tag{10}$$

which explicitly is

$$\mu^{(i+1)}(x)=-\frac{1}{2}R^{-1}g^T(x)\nabla V_x^{(i)}.\tag{11}$$

To ensure convergence of the PI algorithm an initial admissible policy $\mu^{(0)}(x(t))\in\Psi(\Omega)$ is required. It is in fact required by the desired completion of the first step in the policy iteration: i.e. finding a value associated with that initial policy (which needs to be admissible to have a finite value and for the nonlinear Lyapunov equation to have a solution). The algorithm then converges to the optimal control policy $\mu^*\in\Psi(\Omega)$ with corresponding cost $V^*(x)$. Proofs of convergence of the PI algorithm have been given in several references. See Abu-Khalaf and Lewis (2005), Baird (1994), Beard et al. (1997), Hanselmann et al. (2007), Howard (1960), Murray et al. (2002) and Vrabie and Lewis (2008).

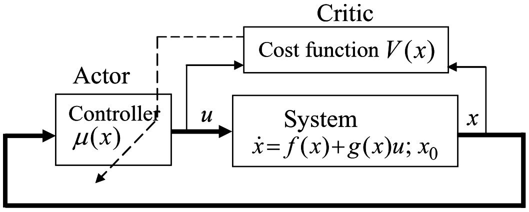


Fig. 1. Actor/critic structure.

Policy iteration is a Newton method. In the linear time-invariant case, it reduces to the Kleinman algorithm (Kleinman, 1968) for solution of the Riccati equation, a familiar algorithm in control systems. Then, (9) become a Lyapunov equation.

2.3. Value function approximation (VFA)

The standard PI Algorithm just discussed proceeds by alternately updating the critic value and the actor policy by solving respectively the Eqs. (9) and (11). In this paper, the fundamental update equations in PI namely (9) for the value and (11) for the policy are used to design two neural networks. Then, by contrast to standard PI, it is shown how to tune these critic and actor neural networks *simultaneously* in real-time to guarantee convergence to the control policy as well as stability during the training process.

The policy iteration algorithm, as other reinforcement learning algorithms, can be implemented on an actor/critic structure which consists of two neural network structures to approximate the solutions of the two Eqs. (9) and (10) at each step of the iteration. The structure is presented in Fig. 1.

In the actor/critic structure (Werbos, 1974, 1989, 1992) the cost $V^{\mu^{(i)}}(x(t))$ and the control $\mu^{(i+1)}(x)$ are approximated at each step of the PI algorithm by neural networks, called respectively the critic Neural Network (NN) and the actor NN. Then, the PI algorithm consists in tuning alternatively each of the two neural networks. The critic NN is tuned to solve (9) (in a least-squares sense Finlayson, 1990), and the actor NN to solve (11). Thus, while one NN is being tuned, the other is held constant. Note that, at each step in the iteration, the critic neural network is tuned to evaluate the performance of the current control policy.

The critic NN is based on value function approximation (VFA). In the following, it is desired to determine a rigorously justifiable form for the critic NN. Since one desires approximation in Sobolev norm, that is, approximation of the value $V(x)$ as well as its gradient, some discussion is given that relates normal NN approximation usage to the Weierstrass higher-order approximation theorem.

The solutions to the nonlinear Lyapunov equations (4), (9) may not be smooth for general nonlinear systems, except in a generalized sense (Sontag & Sussmann, 1995). However, in keeping with other work in the literature (Van der Schaft, 1992) we make the following assumptions.

Assumption 1. The solution to (4) is smooth, i.e. $V(x)\in C^1(\Omega)$.

Assumption 2. The solution to (4) is positive definite. This is guaranteed for stabilizable dynamics if the performance functional satisfies zero-state observability (Van der Schaft, 1992), which is guaranteed by the condition that $Q(x)>0, x\in\Omega-\{0\}; Q(0)=0$ be positive definite.

Assumption 1 allows us to bring in informal style of the Weierstrass higher-order approximation Theorem (Abu-Khalaf & Lewis, 2005; Finlayson, 1990) and the results of Hornik, Stinchcombe, and White (1990), which state that then 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

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

$$V(x) \equiv C_1^T\phi_1(x) + \sum_{i=N+1}^{\infty} c_i\varphi_i(x) \tag{12}$$

$$\frac{\partial V(x)}{\partial x} = \sum_{i=1}^{\infty} c_i\frac{\partial\varphi_i(x)}{\partial x} = \sum_{i=1}^{N} c_i\frac{\partial\varphi_i(x)}{\partial x} + \sum_{i=N+1}^{\infty} c_i\frac{\partial\varphi_i(x)}{\partial x} \tag{13}$$

where $\phi_1(x) = [\varphi_1(x)\ \varphi_2(x)\cdots\varphi_N(x)]^T : R^n \to R^N$ and the last terms in these equations converge uniformly to zero as $N \to \infty$. (Specifically, the basis set is dense in the Sobolev norm $W^{1,\infty}$, Adams & Fournier, 2003.) 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, 1998).

Thus, it is justified to assume there exist weights W_1 such that the value function $V(x)$ is approximated as

$$V(x) = W_1^T\phi_1(x) + \varepsilon(x). \tag{14}$$

Then $\phi_1(x) : R^n \to R^N$ is called the NN activation function vector, N the number of neurons in the hidden layer, and $\varepsilon(x)$ the NN approximation error. As per the above, the NN activation functions $\{\varphi_i(x) : i = 1, N\}$ are selected so that $\{\varphi_i(x) : i = 1, \infty\}$ provides a complete independent basis set such that $V(x)$ and its derivative

$$\frac{\partial V}{\partial x} = \left(\frac{\partial\phi_1(x)}{\partial x}\right)^T W_1 + \frac{\partial\varepsilon}{\partial x} = \nabla\phi_1^T W_1 + \nabla\varepsilon \tag{15}$$

are uniformly approximated. Then, as the number of hidden layer neurons $N \to \infty$, the approximation errors $\varepsilon \to 0, \nabla\varepsilon \to 0$ uniformly (Finlayson, 1990). In addition, for fixed N, the NN approximation errors $\varepsilon(x)$, and $\nabla\varepsilon$ are bounded by constants on a compact set (Hornik et al., 1990).

Using the NN value function approximation, considering a fixed control policy $u(t)$, the nonlinear Lyapunov equation (4) becomes

$$H(x, u, W_1) = W_1^T\nabla\phi_1(f + gu) + Q(x) + u^TRu = \varepsilon_H \tag{16}$$

where the residual error due to the function approximation error is

$$\begin{aligned}\varepsilon_H &= -(\nabla\varepsilon)^T(f + gu)\\ &= -(C_1 - W_1)^T\nabla\phi_1(f + gu) - \sum_{i=N+1}^{\infty} c_i\nabla\varphi_i(x)(f + gu).\end{aligned} \tag{17}$$

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

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.

Definition 2 (*Uniform Convergence*). A sequence of functions $\{f_k\}$ converges uniformly to f on a set Ω if $\forall\varepsilon > 0, \exists N(\varepsilon) : \sup_{x\in\Omega} \|f_n(x) - f(x)\| < \varepsilon$.

The following Lemma has been shown in Abu-Khalaf and Lewis (2005).

Lemma 1. *For any admissible policy $u(t)$, the least-squares solution to (16) exists and is unique for each N. Denote this solution as W_1 and define*

$$V_1(x) = W_1^T\phi_1(x). \tag{18}$$

Then, as $N \to \infty$:

a. $\sup_{x\in\Omega} |\varepsilon_H| \to 0$
b. $\|W_1 - C_1\|_2 \to 0$
c. $\sup_{x\in\Omega} |V_1 - V| \to 0$
d. $\sup_{x\in\Omega} \|\nabla V_1 - \nabla V\| \to 0$. □

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).

Since the object of interest in this paper is finding the solution of the HJB using the above introduced function approximator, it is interesting now to look at the effect of the approximation error on the HJB equation (8)

$$W_1^T\nabla\varphi_1 f - \frac{1}{4}W_1^T\nabla\varphi_1 gR^{-1}g^T\nabla\varphi_1^T W_1 + Q(x) = \varepsilon_{HJB} \tag{19}$$

where the residual error due to the function approximation error is

$$\varepsilon_{HJB} = -\nabla\varepsilon^T f + \frac{1}{2}W_1^T\nabla\varphi_1 gR^{-1}g^T\nabla\varepsilon + \frac{1}{4}\nabla\varepsilon^T gR^{-1}g^T\nabla\varepsilon. \tag{20}$$

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

3. Tuning and convergence of critic NN

In this section we address the issue of tuning and convergence of the critic NN weights when a fixed admissible control policy is prescribed. Therefore, the focus is on the nonlinear Lyapunov equation (4) for a fixed policy u.

In fact, this amounts to the *design of an observer for the value function* which is known as 'cost function' in the optimal control literature. 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 weights of the critic NN, W_1 which provide the best approximate solution for (16) are unknown. Therefore, the output of the critic neural network is

$$\hat{V}(x) = \hat{W}_1^T\phi_1(x) \tag{21}$$

where $\hat{W}_1$ are the current estimated values of the ideal critic NN weights W_1. Recall that $\phi_1(x) : R^n \to R^N$ is the vector of activation functions, with N the number of neurons in the hidden layer. The approximate nonlinear Lyapunov equation is then

$$H(x, \hat{W}_1, u) = \hat{W}_1^T\nabla\phi_1(f + gu) + Q(x) + u^TRu = e_1. \tag{22}$$

In view of Lemma 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 admissible control policy u, it is desired to select $\hat{W}_1$ to minimize the squared residual error

$$E_1 = \frac{1}{2}e_1^Te_1.$$

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

$$\dot{\hat{W}}_1 = -a_1\frac{\partial E_1}{\partial\hat{W}_1} = -a_1\frac{\sigma_1}{(\sigma_1^T\sigma_1 + 1)^2}[\sigma_1^T\hat{W}_1 + Q(x) + u^TRu] \tag{23}$$

where $\sigma_1 = \nabla\phi_1(f + gu)$. This is a 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 proofs, where one needs both appearances of $\sigma_1/(1 + \sigma_1^T\sigma_1)$ in (23) to be bounded (Ioannou & Fidan, 2006; Tao, 2003).

Note that, from (16),

$$Q(x) + u^T R u = -W_1^T \nabla \varphi_1 (f + gu) + \varepsilon_H. \tag{24}$$

Substituting (24) in (23) and, with the notation

$$\bar{\sigma}_1 = \sigma_1 / (\sigma_1^T \sigma_1 + 1), \qquad m_s = 1 + \sigma_1^T \sigma_1 \tag{25}$$

we obtain the dynamics of the critic weight estimation error as

$$\dot{\tilde{W}}_1 = -a_1 \bar{\sigma}_1 \bar{\sigma}_1^{\mathrm{T}} \tilde{W}_1 + a_1 \bar{\sigma}_1 \frac{\varepsilon_H}{m_s}. \tag{26}$$

Though it is traditional to use critic tuning algorithms of the form (23), 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. This development is motivated by adaptive control techniques that appear in Ioannou and Fidan (2006) and Tao (2003).

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,

$$\beta_1 I \le S_0 \equiv \int_t^{t+T} \bar{\sigma}_1(\tau) \bar{\sigma}_1^T(\tau) \mathrm{d}\tau \le \beta_2 I. \tag{27}$$

The PE assumption is needed in adaptive control if one desires to perform system identification using e.g. RLS (Ioannou & Fidan, 2006). It is needed here because one effectively desires to identify the critic parameters to approximate $V(x)$.

Technical Lemma 1. *Consider the error dynamics system with output defined as*

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

$$y = \bar{\sigma}_1^T \tilde{W}_1. \tag{28}$$

The PE condition (27) is equivalent to the uniform complete observability (UCO) (Lewis, Liu, & Yesildirek, 1995) of this system, that is there exist constants $\beta_3 > 0$, $\beta_4 > 0$, $T > 0$ such that, for all t,

$$\beta_3 I \le S_1 \equiv \int_t^{t+T} \Phi^T(\tau, t) \bar{\sigma}_1(\tau) \bar{\sigma}_1^T(\tau) \Phi(\tau, t) \mathrm{d}\tau \le \beta_4 I \tag{29}$$

with $\Phi(t_1, t_0)$, $t_0 \le t_1$ the state transition matrix of (28).

Proof. System (28) and the system defined by $\dot{\tilde{W}}_1 = a_1 \bar{\sigma}_1 u$, $y = \bar{\sigma}_1^T \tilde{W}_1$ are equivalent under the output feedback $u = -y + \varepsilon_H / m_s$. Note that (27) is the observability gramian of this last system. □

The importance of UCO is that bounded input and bounded output implies that the state $\tilde{W}_1(t)$ is bounded. In Theorem 1 we shall see that the critic tuning law (23) indeed guarantees boundedness of the output in (28).

Technical Lemma 2. *Consider the error dynamics system (28). Let the signal $\bar{\sigma}_1$ be persistently exciting. Then:*

(a) *The system (28) is exponentially stable. In fact if $\varepsilon_H = 0$ then $\|\tilde{W}(kT)\| \le \mathrm{e}^{-\alpha kT} \|\tilde{W}(0)\|$ with*

$$\alpha = -\frac{1}{T} \ln(\sqrt{1 - 2a_1 \beta_3}). \tag{30}$$

(b) *Let $\|\varepsilon_H\| \le \varepsilon_{\max}$ and $\|y\| \le y_{\max}$ then $\|\tilde{W}_1\|$ converges exponentially to the residual set*

$$\tilde{W}_1(t) \le \frac{\sqrt{\beta_2 T}}{\beta_1} \left\{ [y_{\max} + \delta \beta_2 a_1 (\varepsilon_{\max} + y_{\max})] \right\} \tag{31}$$

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

Proof. See Appendix. □

The next result shows that the tuning algorithm (23) 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 equation (16) for the given control policy $u(t)$. That is, (21) converges close to the actual value function of the current control policy.

Theorem 1. *Let $u(t)$ be any admissible bounded control policy. Let tuning for the critic NN be provided by (23) and assume that $\bar{\sigma}_1$ is persistently exciting. Let the residual error in (16) be bounded $\|\varepsilon_H\| < \varepsilon_{\max}$. Then the critic parameter error converges exponentially with decay factor given by (30) to the residual set*

$$\tilde{W}_1(t) \le \frac{\sqrt{\beta_2 T}}{\beta_1} \left\{ [1 + 2\delta \beta_2 a_1] \varepsilon_{\max} \right\}. \tag{32}$$

Proof. Consider the following Lyapunov function candidate

$$L(t) = \frac{1}{2} tr\{\tilde{W}_1^T a_1^{-1} \tilde{W}_1\}. \tag{33}$$

The derivative of L is given by

$$\dot{L} = -tr\left\{\tilde{W}_1^T \frac{\sigma_1}{m_s^2} [\sigma_1^T \tilde{W}_1 - \varepsilon_H]\right\}$$

$$\dot{L} = -tr\left\{\tilde{W}_1^T \frac{\sigma_1 \sigma_1^T}{m_s^2} \tilde{W}_1\right\} + tr\left\{\tilde{W}_1^T \frac{\sigma_1}{m_s} \frac{\varepsilon_H}{m_s}\right\}$$

$$\dot{L} \le -\left\|\frac{\sigma_1^T}{m_s} \tilde{W}_1\right\|^2 + \left\|\frac{\sigma_1^T}{m_s} \tilde{W}_1\right\| \left\|\frac{\varepsilon_H}{m_s}\right\|$$

$$\dot{L} \le -\left\|\frac{\sigma_1^T}{m_s} \tilde{W}_1\right\| \left[\left\|\frac{\sigma_1^T}{m_s} \tilde{W}_1\right\| - \left\|\frac{\varepsilon_H}{m_s}\right\|\right]. \tag{34}$$

Therefore $\dot{L} \le 0$ if

$$\left\|\frac{\sigma_1^T}{m_s} \tilde{W}_1\right\| > \varepsilon_{\max} > \left\|\frac{\varepsilon_H}{m_s}\right\|, \tag{35}$$

since $\|m_s\| \ge 1$.

This provides an effective practical bound for $\|\bar{\sigma}_1^T \tilde{W}_1\|$, since $L(t)$ decreases if (35) holds.

Consider the estimation error dynamics (28) with the output bounded effectively by $\|y\| < \varepsilon_{\max}$, as just shown. Now Technical Lemma 2 shows exponential convergence to the residual set

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

This completes the proof. □

Remark 1. Note that, as $N \to \infty$, $\varepsilon_H \to 0$ uniformly (Abu-Khalaf & Lewis, 2005). This means that $\varepsilon_{\max}$ decreases as the number of hidden layer neurons in (21) increases.

Remark 2. This theorem requires the assumption that the control policy $u(t)$ is bounded, since $u(t)$ appears in ε_H. In the upcoming Theorem 2 this restriction is removed.

4. Action neural network and online synchronous policy iteration

We will now present an online adaptive PI algorithm which involves simultaneous, or synchronous, tuning of both the actor and critic neural networks. That is, the weights of both neural networks are tuned at the same time. This approach is a version of Generalized Policy Iteration (GPI), as introduced in Sutton and Barto (1998). In standard policy iteration, the critic and actor NN are tuned sequentially, with the weights of the other NN being held constant. By contrast, we tune both NN simultaneously in real-time.

It is desired to determine a rigorously justified form for the actor NN. To this end, let us consider one step of the Policy Iteration algorithm (9)–(11). Suppose that the solution $V(x) \in C^1(\Omega)$ to the nonlinear Lyapunov equation (9) for a given admissible policy $u(t)$ is given by (12). Then, according to (13) and (11) one has for the policy update

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

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

Lemma 2. *Let the least-squares solution to (16) be W_1 and define*

$$u_1(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla V_1(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla\phi_1^T(x)W_1 \tag{38}$$

with V_1 defined in (18).
Then, as $N \to \infty$:

(a) $\sup_{x\in\Omega}\|u_1 - u\| \to 0$
(b) *There exists an N_0 such that $u_1(x)$ is admissible for $N > N_0$.*

Proof. See Abu-Khalaf and Lewis (2005).
In light of this result, the ideal control policy update is taken as (38), with W_1 unknown. Therefore, define the control policy in the form of an action neural network which computes the control input in the structured form

$$u_2(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla\phi_1^T\hat{W}_2, \tag{39}$$

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

$$\tilde{W}_2 = W_1 - \hat{W}_2. \tag{40}$$

The next definition and assumptions complete the machinery required for our main result.

Definition 3 (*Lewis, Jagannathan, & Yesildirek, 1999, UUB*). The equilibrium point $x_e = 0$ of (1) is said to be uniformly ultimately bounded (UUB) if there exists a compact set $S \subset R^n$ so that for all $x_0 \in S$ there exists a bound B and a time $T(B, x_0)$ such that $\|x(t) - x_e\| \le B$ for all $t \ge t_0 + T$.

We make the following assumptions.

Assumption 3. a. $f(.)$, is Lipschitz, and $g(.)$ is bounded by a constant

$$\|f(x)\| < b_f\|x\|, \qquad \|g(x)\| < b_g.$$

b. The NN approx error and its gradient are bounded on a compact set containing Ω so that

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

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

$$\|\phi_1(x)\| < b_\phi$$
$$\|\nabla\phi_1(x)\| < b_{\phi_x}. \quad \square$$

These are standard assumptions, except for the rather strong assumption on $g(x)$ in *a*. Assumption 3c is satisfied, e.g. by sigmoids, tanh, and other standard NN activation functions.

We now present the main Theorem, which provides the tuning laws for the actor and critic neural networks that guarantee convergence of the synchronous online PI algorithm to the optimal controller, while guaranteeing closed-loop stability.

Theorem 2. *Let the dynamics be given by (1), the critic NN be given by (21) and the control input be given by actor NN (39). Let tuning for the critic NN be provided by*

$$\dot{\hat{W}}_1 = -a_1\frac{\sigma_2}{(\sigma_2^T\sigma_2+1)^2}\,[\sigma_2^T\hat{W}_1 + Q(x) + u_2^TRu_2] \tag{41}$$

where $\sigma_2 = \nabla\phi_1(f + gu_2)$, and assume that $\bar{\sigma}_2 = \sigma_2/(\sigma_2^T\sigma_2 + 1)$ is persistently exciting. Let the actor NN be tuned as

$$\dot{\hat{W}}_2 = -\alpha_2\left\{(F_2\hat{W}_2 - F_1\bar{\sigma}_2^T\hat{W}_1) - \frac{1}{4}\bar{D}_1(x)\hat{W}_2m^T(x)\hat{W}_1\right\} \tag{42}$$

where

$$\bar{D}_1(x) \equiv \nabla\phi_1(x)g(x)R^{-1}g^T(x)\nabla\phi_1^T(x),$$
$$m \equiv \frac{\sigma_2}{(\sigma_2^T\sigma_2+1)^2},$$

and $F_1 > 0$ and $F_2 > 0$ are tuning parameters. Let Assumptions 1–3 hold, and 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 $\tilde{W}_1$, and the actor NN error $\tilde{W}_2$ are UUB. Moreover, Theorem 1 holds with $\varepsilon_{\max}$ defined in the proof, so that exponential convergence of $\hat{W}_1$ to the approximate optimal critic value W_1 is obtained.

Proof. See Appendix. □

Remark 3. Let $\varepsilon > 0$ and let N_0 be the number of hidden layer units above which $\sup_{x\in\Omega}\|\varepsilon_{HJB}\| < \varepsilon$. In the proof it is seen that the theorem holds for$N > N_0$. Additionally, ε provides an effective bound on the critic weight residual set in Theorem 1. That is, $\varepsilon_{\max}$ in (32) is effectively replaced by ε.

Remark 4. The theorem shows that PE is needed for proper identification of the value function by the critic NN, and that a nonstandard tuning algorithm is required for the actor NN to guarantee stability. The second term in (42) is a cross-product term that involves both the critic weights and the actor weights. It is needed to guarantee good behavior of the Lyapunov function, i.e. that the energy decreases to a bounded compact region.

Remark 5. The tuning parameters F_1 and F_2 in (42) must be selected to make the matrix M in (A.22) positive definite.

Note that the dynamics is required to implement this algorithm in that $\sigma_2 = \nabla\phi_1(f + gu_2)$, $\bar{D}_1(x)$, and (39) depend on $f(x)$, $g(x)$.

5. Simulation results

To support the new synchronous online PI algorithm for CT systems, we offer two simulation examples, one linear and one nonlinear. In both cases we observe convergence to the actual optimal value function and control.

5.1. Linear system example

Consider the continuous-time F16 aircraft plant with quadratic cost function used in Stevens and Lewis (2003)

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

where Q and R in the cost function are identity matrices of appropriate dimensions. In this linear case the solution of the HJB equation is given by the solution of the algebraic Riccati equation (ARE). 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. Solving the ARE gives the parameters of the optimal critic as

$$W_1^* = [\,1.4245 \;\; 1.1682 \;\; -0.1352 \;\; 1.4349 \;\; -0.1501 \;\; 0.4329\,]^T$$

which are the components of the Riccati solution matrix P.

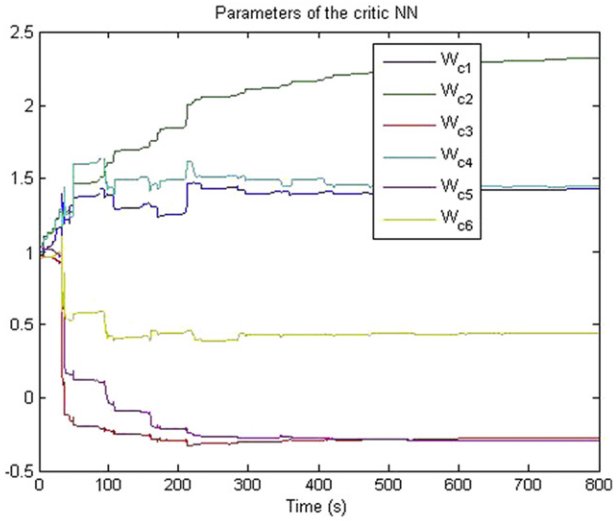


Fig. 2. Convergence of the critic parameters to the parameters of the optimal critic.

The synchronous PI algorithm is implemented as in Theorem 2. PE was ensured by adding a small probing noise to the control input. Fig. 2 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

$$\hat{W}_1(t_f) = [\, 1.4279 \;\; 1.1612 \;\; -0.1366 \;\; 1.4462 \;\; -0.1480 \;\; 0.4317 \,]^T.$$

The actor parameters after 800s converge to the values of

$$\hat{W}_2(t_f) = [\, 1.4279 \;\; 1.1612 \;\; -0.1366 \;\; 1.4462 \;\; -0.1480 \;\; 0.4317 \,]^T.$$

Then, the actor NN is given by (39) as

$$\hat{u}_2(x) = -\frac{1}{2} R^{-1} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} 2x_1 & 0 & 0 \\ x_2 & x_1 & 0 \\ x_3 & 0 & x_1 \\ 0 & 2x_2 & 0 \\ 0 & x_3 & x_2 \\ 0 & 0 & 2x_3 \end{bmatrix}^T \begin{bmatrix} 1.4279 \\ 1.1612 \\ -0.1366 \\ 1.4462 \\ -0.1480 \\ 0.4317 \end{bmatrix}$$

i.e. approximately the correct optimal control solution $u = -R^{-1}B^T Px$.

The evolution of the system states is presented in Fig. 3. One can see that after 750 s convergence of the NN weights in both critic and actor has occurred. This shows that the probing noise effectively guaranteed the PE condition. On convergence, the PE condition of the control signal is no longer needed, and the probing signal was turned off. After that, the states remain very close to zero, as required.

5.2. Nonlinear system example

Consider the following affine in control input nonlinear system, with a quadratic cost derived as in Nevistic and Primbs (1996) and Vrabie et al. (2009)

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

where

$$f(x) = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2(1 - (\cos(2x_1) + 2)^2) \end{bmatrix}$$

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

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

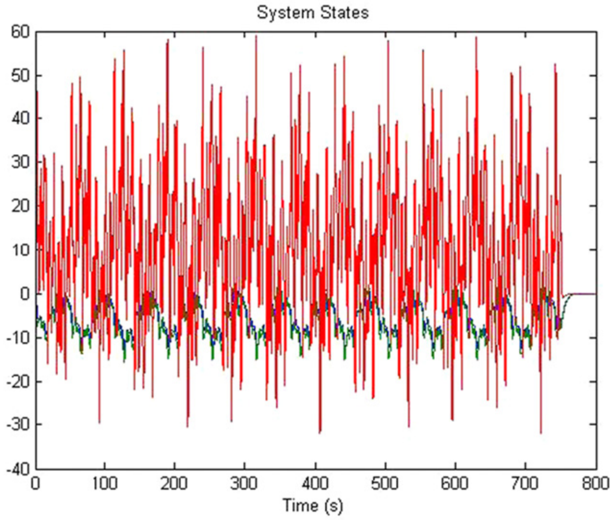


Fig. 3. Evolution of the system states for the duration of the experiment.

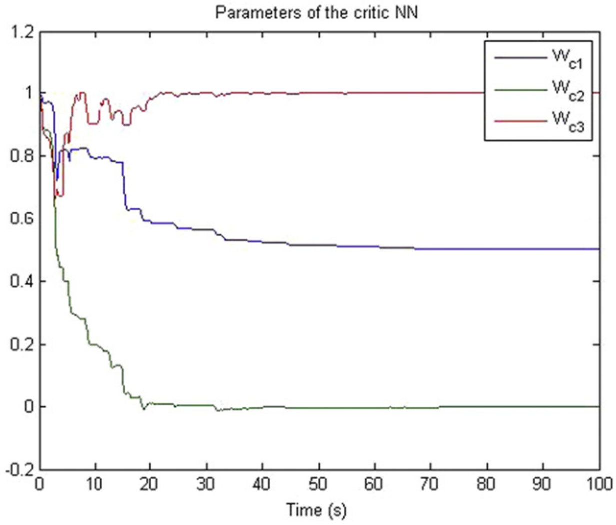


Fig. 4. Convergence of the critic parameters.

Using the procedure in Nevistic and Primbs (1996) the optimal value function is

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

and the optimal control signal is

$$u^*(x) = -(\cos(2x_1) + 2)x_2.$$

One selects the critic NN vector activation function as

$$\phi_1(x) = [\, x_1^2 \quad x_1 x_2 \quad x_2^2 \,]^T.$$

Fig. 4 shows the critic parameters, denoted by

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

These converge after about 80s to the correct values of

$$\hat{W}_1(t_f) = [\, 0.5017 \quad -0.0020 \quad 1.0008 \,]^T.$$

The actor parameters after 80s converge to the values of

$$\hat{W}_2(t_f) = [\, 0.5017 \quad -0.0020 \quad 1.0008 \,]^T.$$

So that the actor NN (39)

$$\hat{u}_2(x) = -\frac{1}{2} R^{-1} \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix}^T \begin{bmatrix} 2x_1 & 0 \\ x_2 & x_1 \\ 0 & 2x_2 \end{bmatrix}^T \begin{bmatrix} 0.5017 \\ -0.0020 \\ 1.0008 \end{bmatrix}$$

also converged to the optimal control.

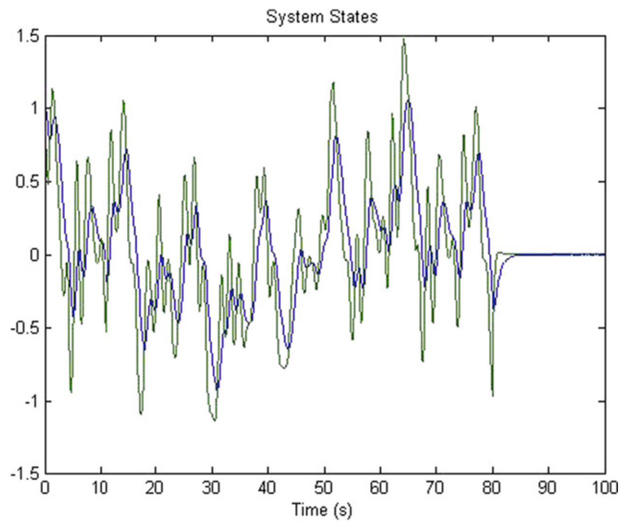


Fig. 5. Evolution of the system states for the duration of the experiment.

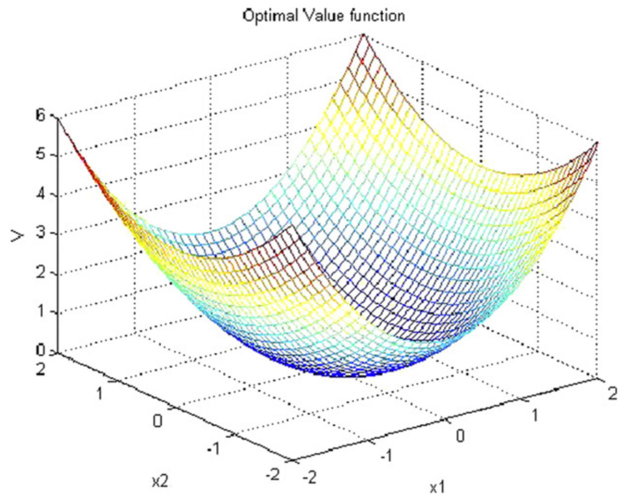


Fig. 6. Optimal value function.

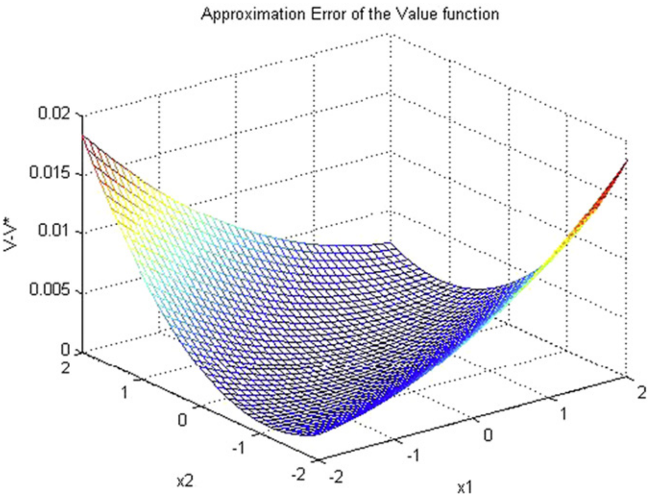


Fig. 7. 3D plot of the approximation error for the value function.

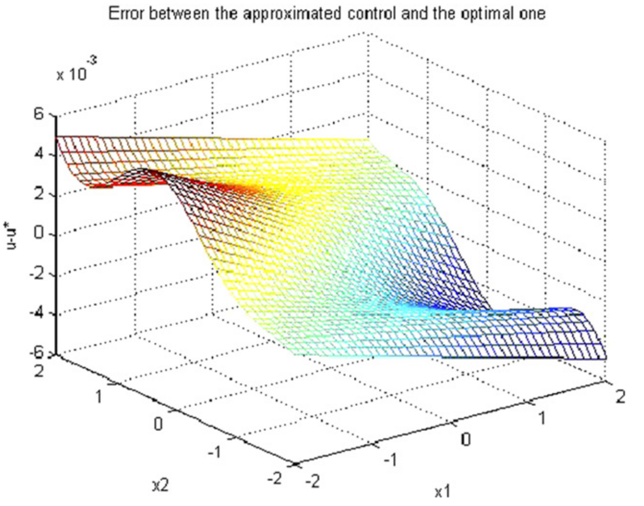


Fig. 8. 3D plot of the approximation error for the control.

The evolution of the system states is presented in Fig. 5. One can see that after 80s convergence of the NN weights in both critic and actor has occurred. This shows that the probing noise effectively guaranteed the PE condition. On convergence, the PE condition of the control signal is no longer needed, and the probing signal was turned off. After that, the states remain very close to zero, as required.

Fig. 6 show the optimal value function. The identified value function given by $\hat{V}_1(x) = \hat{W}_1^T \phi_1(x)$ is virtually indistinguishable. In fact, Fig. 7 shows the 3D plot of the difference between the approximated value function, by using the online algorithm, and the optimal one. This error is close to zero. Good approximation of the actual value function is being evolved.

Finally Fig. 8 shows the 3D plot of the difference between the approximated control, by using the online algorithm, and the optimal one. This error is close to zero.

6. Conclusions

In this paper we have proposed a new adaptive algorithm which solves the continuous-time optimal control problem for affine in the inputs nonlinear systems. We call this algorithm synchronous online PI for CT systems. The algorithm requires complete knowledge of the system model. For this reason our research efforts will now be directed towards integrating a third neural network with the actor/critic structure with the purpose of approximating in an online fashion the system dynamics, as suggested by Werbos (1974, 1989, 1992).

Acknowledgements

The authors would like to acknowledge the insight and guidance received from Draguna Vrabie, currently a Ph.D. student at UTA's Automation & Robotics Research Institute.

Appendix. Proofs

Proof for Technical Lemma 2 Part a. This is a more complete version of results in Ioannou and Fidan (2006) and Tao (2003).

Set $\varepsilon_H = 0$ in (26). Take the Lyapunov function

$$L = \frac{1}{2}\tilde{W}_1^T a_1^{-1} \tilde{W}_1. \tag{A.1}$$

The derivative is

$$\dot{L} = -\tilde{W}_1^T \bar{\sigma}_1 \bar{\sigma}_1^T \tilde{W}_1.$$

Integrating both sides

$$L(t+T)-L(t) = -\int_t^{t+T} \tilde{W}_1^T\sigma_1(\tau)\sigma_1^T(\tau)\tilde{W}_1 \mathrm{d}\tau$$

$$\begin{aligned} L(t+T) &= L(t) - \tilde{W}_1^T(t)\int_t^{t+T}\Phi^T(\tau,t)\sigma_1(\tau)\sigma_1^T(\tau) \\ &\quad \times \Phi(\tau,t)\mathrm{d}\tau\tilde{W}_1(t) \\ &= L(t)-\tilde{W}_1^T(t)S_1\tilde{W}_1(t) \le (1-2a_1\beta_3)L(t). \end{aligned}$$

So

$$L(t+T) \le (1-2a_1\beta_3)L(t). \tag{A.2}$$

Define $\gamma = (1-2a_1\beta_3)$. By using norms we write (A.2) in terms of $\tilde{W}_1$ as

$$\frac{1}{2a_1}\left\|\tilde{W}(t+T)\right\|^2 \le \sqrt{(1-2a_1\beta_3)}\frac{1}{2a_1}\left\|\tilde{W}(t)\right\|^2$$

$$\left\|\tilde{W}(t+T)\right\| \le \sqrt{(1-2a_1\beta_3)}\left\|\tilde{W}(t)\right\|$$

$$\left\|\tilde{W}(t+T)\right\| \le \gamma\left\|\tilde{W}(t)\right\|.$$

Therefore

$$\|\tilde{W}(kT)\| \le \gamma^k\|\tilde{W}(0)\| \tag{A.3}$$

i.e. $\tilde{W}(t)$ decays exponentially. To determine the decay time constant in continuous time, note that

$$\|\tilde{W}(kT)\| \le \mathrm{e}^{-\alpha kT}\|\tilde{W}(0)\| \tag{A.4}$$

where $\mathrm{e}^{-\alpha kT} = \gamma^k$. Therefore the decay constant is

$$\alpha = -\frac{1}{T}\ln(\gamma) \Leftrightarrow \alpha = -\frac{1}{T}\ln(\sqrt{1-2a_1\beta_3}). \tag{A.5}$$

This completes the proof. □

Proof for Technical Lemma 2 Part b. Consider the system

$$\begin{cases}\dot{x}(t) = B(t)u(t) \\ y(t) = C^T(t)x(t).\end{cases} \tag{A.6}$$

The state and the output are

$$\begin{cases}x(t+T) = x(t) + \displaystyle\int_t^{t+T} B(\tau)u(\tau)\mathrm{d}\tau \\ y(t+T) = C^T(t+T)x(t+T).\end{cases} \tag{A.7}$$

Let $C(t)$ be PE, so that

$$\beta_1 I \le S_C \equiv \int_t^{t+T} C(\lambda)C^T(\lambda)\mathrm{d}\lambda \le \beta_2 I. \tag{A.8}$$

Then,

$$y(t+T) = C^T(t+T)x(t) + \int_t^{t+T} C^T(t+T)B(\tau)u(\tau)\mathrm{d}\tau$$

$$\int_t^{t+T} C(\lambda)\left(y(\lambda) - \int_t^{\lambda} C^T(\lambda)B(\tau)u(\tau)\mathrm{d}\tau\right)\mathrm{d}\lambda$$

$$= \int_t^{t+T} C(\lambda)C^T(\lambda)x(t)\mathrm{d}\lambda$$

$$\int_t^{t+T} C(\lambda)\left(y(\lambda) - \int_t^{\lambda} C^T(\lambda)B(\tau)u(\tau)\mathrm{d}\tau\right)\mathrm{d}\lambda = S_C x(t)$$

$$x(t) = S_C^{-1}\left\{\int_t^{t+T} C(\lambda)\left(y(\lambda) - \int_t^{\lambda} C^T(\lambda)B(\tau)u(\tau)\mathrm{d}\tau\right)\mathrm{d}\lambda\right\}.$$

Taking the norms in both sides yields

$$\begin{aligned} \|x(t)\| &\le \left\|S_C^{-1}\int_t^{t+T} C(\lambda)y(\lambda)\mathrm{d}\lambda\right\| \\ &\quad + \left\|S_C^{-1}\left\{\int_t^{t+T} C(\lambda)\left(\int_t^{\lambda} C^T(\lambda)B(\tau)u(\tau)\mathrm{d}\tau\right)\mathrm{d}\lambda\right\}\right\| \end{aligned}$$

$$\begin{aligned} \|x(t)\| &\le (\beta_1 I)^{-1}\left(\int_t^{t+T} C(\lambda)C^T(\lambda)\mathrm{d}\lambda\right)^{\frac{1}{2}} \\ &\quad \times \left(\int_t^{t+T} y(\lambda)^T y(\lambda)\mathrm{d}\lambda\right)^{\frac{1}{2}} \\ &\quad + \left\|S_C^{-1}\right\|\left\{\int_t^{t+T}\left\|C(\lambda)C^T(\lambda)\right\|\mathrm{d}\lambda\int_t^{t+T}\|B(\tau)u(\tau)\|\mathrm{d}\tau\right\} \end{aligned}$$

$$\|x(t)\| \le \frac{\sqrt{\beta_2 T}}{\beta_1}y_{\max} + \frac{\delta\beta_2}{\beta_1}\int_t^{t+T}\|B(\tau)\|\cdot\|u(\tau)\|\mathrm{d}\tau \tag{A.9}$$

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

$$\dot{\tilde{W}}_1(t) = a_1\bar{\sigma}_1 u. \tag{A.10}$$

Note that setting $u = -y + \frac{\varepsilon_H}{m_s}$ with output given $y = \bar{\sigma}_1^T\tilde{W}_1$ turns (A.10) into (26). Set $B = a_1\bar{\sigma}_1, C = \bar{\sigma}_1, x(t) = \tilde{W}_1$ so that (A.6) yields (A.10). Then,

$$\|u\| \le \|y\| + \left\|\frac{\varepsilon_H}{m_s}\right\| \le y_{\max} + \varepsilon_{\max} \tag{A.11}$$

since $\|m_s\| \ge 1$. Then,

$$\begin{aligned} N &\equiv \int_t^{t+T}\|B(\tau)\|\cdot\|u(\tau)\|\mathrm{d}\tau = \int_t^{t+T}\|a_1\bar{\sigma}_1(\tau)\|\cdot\|u(\tau)\|\mathrm{d}\tau \\ &\le a_1(y_{\max}+\varepsilon_{\max})\int_t^{t+T}\|\bar{\sigma}_1(\tau)\|\mathrm{d}\tau \\ &\le a_1(y_{\max}+\varepsilon_{\max})\left[\int_t^{t+T}\|\bar{\sigma}_1(\tau)\|^2\mathrm{d}\tau\right]^{1/2}\left[\int_t^{t+T}1\mathrm{d}\tau\right]^{1/2}. \end{aligned}$$

By using (A.8),

$$N \le a_1(y_{\max}+\varepsilon_{\max})\sqrt{\beta_2 T}. \tag{A.12}$$

Finally (A.9) and (A.12) yield,

$$\tilde{W}_1(t) \le \frac{\sqrt{\beta_2 T}}{\beta_1}\left\{[y_{\max} + \delta\beta_2 a_1(\varepsilon_{\max}+y_{\max})]\right\}. \tag{A.13}$$

This completes the proof. □

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

$$L(t) = V(x) + \frac{1}{2}tr(\tilde{W}_1^T a_1^{-1}\tilde{W}_1) + \frac{1}{2}tr(\tilde{W}_2^T a_2^{-1}\tilde{W}_2). \tag{A.14}$$

With the chosen tuning laws one can then show that the errors $\tilde{W}_1$ and $\tilde{W}_2$ are UUB and convergence is obtained.

Hence the derivative of the Lyapunov function is given by

$$\begin{aligned} \dot{L}(x) &= \dot{V}(x) + \tilde{W}_1^T\alpha_1^{-1}\dot{\tilde{W}}_1 + \tilde{W}_2^T\alpha_2^{-1}\dot{\tilde{W}}_2 \\ &= \dot{L}_V(x) + \dot{L}_1(x) + \dot{L}_2(x). \end{aligned} \tag{A.15}$$

First term is,

$$\begin{aligned} \dot{V}(x) &= W_1^T\left(\nabla\phi_1 f(x) - \frac{1}{2}\overline{D}_1(x)\hat{W}_2\right) \\ &\quad + \nabla\varepsilon^T(x)\left(f(x) - \frac{1}{2}g(x)R^{-1}g^T(x)\nabla\phi_1^T\hat{W}_2\right). \end{aligned}$$

Then

$$\begin{aligned}\dot{V}(x) &= W_1^T\left(\nabla\phi_1 f(x) - \frac{1}{2}\bar{D}_1(x)\hat{W}_2\right) + \varepsilon_1(x)\\ &= W_1^T\nabla\phi_1 f(x) + \frac{1}{2}W_1^T\bar{D}_1(x)\left(W_1 - \hat{W}_2\right)\\ &\quad - \frac{1}{2}W_1^T\bar{D}_1(x)W_1 + \varepsilon_1(x)\\ &= W_1^T\nabla\phi_1 f(x) + \frac{1}{2}W_1^T\bar{D}_1(x)\tilde{W}_2 - \frac{1}{2}W_1^T\bar{D}_1(x)W_1 + \varepsilon_1(x)\\ &= W_1^T\sigma_1 + \frac{1}{2}W_1^T\bar{D}_1(x)\tilde{W}_2 + \varepsilon_1(x)\end{aligned}$$

where

$$\varepsilon_1(x) \equiv \dot{\varepsilon}(x) = \nabla\varepsilon^T(x)\left(f(x) - \frac{1}{2}g(x)R^{-1}g^T(x)\nabla\phi_1^T(x)\hat{W}_2\right).$$

From the HJB equation

$$W_1^T\sigma_1 = -Q(x) - \frac{1}{4}W_1^T\bar{D}_1(x)W_1 + \varepsilon_{HJB}(x).$$

Then

$$\begin{aligned}\dot{L}_V(x) &= -Q(x) - \frac{1}{4}W_1^T\bar{D}_1(x)W_1\\ &\quad + \frac{1}{2}W_1^T\bar{D}_1(x)\tilde{W}_2 + \varepsilon_{HJB}(x) + \varepsilon_1(x)\\ &\equiv \dot{\bar{L}}_V(x) + \frac{1}{2}W_1^T\bar{D}_1(x)\tilde{W}_2 + \varepsilon_1(x).\end{aligned} \tag{A.16}$$

Second term is,

$$\begin{aligned}\dot{L}_1 &= \tilde{W}_1^T\alpha_1^{-1}\dot{\tilde{W}}_1\\ &= \tilde{W}_1^T\alpha_1^{-1}\alpha_1\frac{\sigma_2}{\left(\sigma_2^T\sigma_2+1\right)^2}\left(\sigma_2^T\hat{W}_1 + Q(x) + \frac{1}{4}\hat{W}_2^T\bar{D}_1\hat{W}_2\right)\\ &= \tilde{W}_1^T\frac{\sigma_2}{\left(\sigma_2^T\sigma_2+1\right)^2}\left(\sigma_2^T\hat{W}_1 + Q(x) + \frac{1}{4}\hat{W}_2^T\bar{D}_1(x)\hat{W}_2\right.\\ &\quad \left. -\ Q(x) - \sigma_1^T W_1 - \frac{1}{4}W_1^T\bar{D}_1(x)W_1 + \varepsilon_{HJB}(x)\right)\\ &= \tilde{W}_1^T\frac{\sigma_2}{\left(\sigma_2^T\sigma_2+1\right)^2}\left(\sigma_2^T(x)\hat{W}_1 - \sigma_1^T(x)W_1\right.\\ &\quad \left. +\ \frac{1}{4}\hat{W}_2^T\bar{D}_1(x)\hat{W}_2 - \frac{1}{4}W_1^T\bar{D}_1(x)W_1 + \varepsilon_{HJB}(x)\right)\\ &= \tilde{W}_1^T\frac{\sigma_2}{\left(\sigma_2^T\sigma_2+1\right)^2}\left(-\ \tilde{W}_1^T\left(\nabla\phi_1^T(x)\right)^T f(x)\right.\\ &\quad - \frac{1}{2}\hat{W}_2^T\bar{D}_1(x)\hat{W}_1 + \frac{1}{2}W_1^T\bar{D}_1(x)W_1 + \frac{1}{4}\hat{W}_2^T\bar{D}_1(x)\hat{W}_2\\ &\quad \left. -\ \frac{1}{4}W_1^T\bar{D}_1(x)W_1 + \varepsilon_{HJB}(x)\right)\\ &= \tilde{W}_1^T\frac{\sigma_2}{\left(\sigma_2^T\sigma_2+1\right)^2}\left(-f(x)^T\nabla\phi_1^T(x)\tilde{W}_1 + \frac{1}{2}\hat{W}_2^T\bar{D}_1(x)\tilde{W}_1\right.\\ &\quad \left. + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)\tilde{W}_2 + \varepsilon_{HJB}(x)\right).\end{aligned}$$

$$\begin{aligned}\dot{L}_1 &= \tilde{W}_1^T\frac{\sigma_2}{\left(\sigma_2^T\sigma_2+1\right)^2}\left(-\sigma_2^T\tilde{W}_1 + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)\tilde{W}_2 + \varepsilon_{HJB}(x)\right)\\ &= \dot{\bar{L}}_1 + \frac{1}{4}\tilde{W}_1^T\frac{\sigma_2}{\left(\sigma_2^T\sigma_2+1\right)^2}\tilde{W}_2^T\bar{D}_1(x)\tilde{W}_2\end{aligned} \tag{A.17}$$

where

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

Finally by adding the terms (A.16) and (A.17)

$$\begin{aligned}\dot{L}(x) &= -Q(x) - \frac{1}{4}W_1^T\bar{D}_1(x)W_1 + \frac{1}{2}W_1^T\bar{D}_1(x)\tilde{W}_2\\ &\quad + \varepsilon_{HJB}(x) + \varepsilon_1(x) + \tilde{W}_1^T\frac{\sigma_2}{\left(\sigma_2^T\sigma_2+1\right)^2}\\ &\quad \times\left(-\sigma_2^T\tilde{W}_1 + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)\tilde{W}_2 + \varepsilon_{HJB}(x)\right) + \tilde{W}_2^T\alpha_2^{-1}\dot{\tilde{W}}_2\end{aligned}$$

$$\begin{aligned}\dot{L}(x) &= \dot{\bar{L}}_V + \dot{\bar{L}}_1 + \varepsilon_1(x) - \tilde{W}_2^T\alpha_2^{-1}\dot{\hat{W}}_2 + \frac{1}{2}\tilde{W}_2^T\bar{D}_1(x)W_1\\ &\quad + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2^T}{m_s}\tilde{W}_1 - \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2^T}{m_s}W_1\\ &\quad + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)\tilde{W}_2\frac{\bar{\sigma}_2^T}{m_s}W_1 + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)\hat{W}_2\frac{\bar{\sigma}_2^T}{m_s}\hat{W}_1\end{aligned} \tag{A.18}$$

where

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

In order to select the update law for the action neural network, write (A.18) as

$$\begin{aligned}\dot{L}(x) &= \dot{\bar{L}}_V + \dot{\bar{L}}_1 + \varepsilon_1(x) - \tilde{W}_2^T\left[\alpha_2^{-1}\dot{\hat{W}}_2 - \frac{1}{4}\bar{D}_1(x)\hat{W}_2\frac{\bar{\sigma}_2^T}{m_s}\hat{W}_1\right]\\ &\quad + \frac{1}{2}\tilde{W}_2^T\bar{D}_1(x)W_1 + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2^T}{m_s}\tilde{W}_1\\ &\quad - \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2^T}{m_s}W_1 + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2}{m_s}\tilde{W}_2\end{aligned}$$

and we define the actor tuning law as

$$\dot{\hat{W}}_2 = -\alpha_2\left\{\left(F_2\hat{W}_2 - F_1\bar{\sigma}_2^T\hat{W}_1\right) - \frac{1}{4}\bar{D}_1(x)\hat{W}_2 m^T\hat{W}_1\right\}. \tag{A.19}$$

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

$$\begin{aligned}&\tilde{W}_2^T F_2\hat{W}_2 - \tilde{W}_2^T F_1\bar{\sigma}_2^T\hat{W}_1\\ &\quad = \tilde{W}_2^T F_2(W_1 - \tilde{W}_2) - \tilde{W}_2^T F_1\bar{\sigma}_2^T(W_1 - \tilde{W}_1)\\ &\quad = \tilde{W}_2^T F_2 W_1 - \tilde{W}_2^T F_2\tilde{W}_2 - \tilde{W}_2^T F_1\bar{\sigma}_2^T W_1 + \tilde{W}_2^T F_1\bar{\sigma}_2^T\tilde{W}_1.\end{aligned}$$

Overall

$$\begin{aligned}\dot{L}(x) &= -Q(x) - \frac{1}{4}W_1^T\bar{D}_1(x)W_1 + \varepsilon_{HJB}(x)\\ &\quad + \tilde{W}_1^T\bar{\sigma}_2\left(-\bar{\sigma}_2^T\tilde{W}_1 + \frac{\varepsilon_{HJB}(x)}{m_s}\right) + \varepsilon_1(x)\\ &\quad + \frac{1}{2}\tilde{W}_2^T\bar{D}_1(x)W_1 + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2^T}{m_s}\tilde{W}_1\\ &\quad - \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2^T}{m_s}W_1 + \frac{1}{4}\tilde{W}_2^T\bar{D}_1(x)W_1\frac{\bar{\sigma}_2}{m_s}\tilde{W}_2\\ &\quad + \tilde{W}_2^T F_2 W_1 - \tilde{W}_2^T F_2\tilde{W}_2 - \tilde{W}_2^T F_1\bar{\sigma}_2^T W_1 + \tilde{W}_2^T F_1\bar{\sigma}_2^T\tilde{W}_1.\end{aligned} \tag{A.20}$$

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

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

Also, since $Q(x) > 0$ there exists q such that $x^T q x < Q(x)$ for $x \in \Omega$. It is shown in Abu-Khalaf and Lewis (2005) that ε_{HJB} converges to zero uniformly as N increases.

Select $\varepsilon > 0$ and $N_0(\varepsilon)$ such that $\sup_{x\in\Omega}\left\|\varepsilon_{HJB}\right\| < \varepsilon$. Then, assuming $N > N_0$ and writing in terms of $\tilde{Z} = \begin{bmatrix} x \\ \bar{\sigma}_2^T\tilde{W}_1 \\ \tilde{W}_2 \end{bmatrix}$ (A.20) becomes

$$\begin{aligned}
\dot{L} &< \frac{1}{4}\left\|W_1\right\|^2\left\|\bar{D}_1(x)\right\| + \varepsilon + \frac{1}{2}\left\|W_1\right\| b_{\varepsilon_x} b_{\varphi_x} b_g^2 \sigma_{\min}(R) \\
&\quad - \tilde{Z}^T \begin{bmatrix} qI & 0 & 0 \\ 0 & I & \left(-\frac{1}{2}F_1 - \frac{1}{8m_s}\bar{D}_1 W_1\right)^T \\ 0 & -\frac{1}{2}F_1 - \left(\frac{1}{8m_s}\bar{D}_1 W_1\right) & F_2 - \frac{1}{8}\left(\bar{D}_1 W_1 m^T + m W_1^T \bar{D}_1\right) \end{bmatrix} \tilde{Z} \\
&\quad + \tilde{Z}^T \begin{bmatrix} \dfrac{\underset{\varepsilon}{b_{\varepsilon_x} b_f}}{m_s} \\ \left(\frac{1}{2}\bar{D}_1 + F_2 - F_1\bar{\sigma}_2^T - \frac{1}{4}\bar{D}_1 W_1 m^T\right) W_1 + \frac{1}{2} b_{\varepsilon_x} b_g^2 b_{\phi_x} \sigma_{\min}(R) \end{bmatrix}. \qquad (A.21)
\end{aligned}$$

Define

$$M = \begin{bmatrix} qI & 0 & 0 \\ 0 & I & \left(-\frac{1}{2}F_1 - \frac{1}{8m_s}\bar{D}_1 W_1\right)^T \\ 0 & -\frac{1}{2}F_1 - \left(\frac{1}{8m_s}\bar{D}_1 W_1\right) & F_2 - \frac{1}{8}\left(\bar{D}_1 W_1 m^T + m W_1^T \bar{D}_1\right) \end{bmatrix} \qquad (A.22)$$

$$d = \begin{bmatrix} \dfrac{\underset{\varepsilon}{b_{\varepsilon_x} b_f}}{m_s} \\ \left(\frac{1}{2}\bar{D}_1 + F_2 - F_1\bar{\sigma}_2^T - \frac{1}{4}\bar{D}_1 W_1 m^T\right) W_1 + \frac{1}{2} b_{\varepsilon_x} b_g^2 b_{\phi_x} \sigma_{\min}(R) \end{bmatrix}$$

$$c = \frac{1}{4}\left\|W_1\right\|^2\left\|\bar{D}_1(x)\right\| + \varepsilon + \frac{1}{2}\left\|W_1\right\| b_{\varepsilon_x} b_{\varphi_x} b_g^2 \sigma_{\min}(R).$$

Let the parameters be chosen such that $M > 0$. Now (A.21) becomes

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

Completing the squares, the Lyapunov derivative is negative if

$$\left\|\tilde{Z}\right\| > \frac{\|d\|}{2\sigma_{\min}(M)} + \sqrt{\frac{d^2}{4\sigma_{\min}^2(M)} + \frac{c+\varepsilon}{\sigma_{\min}(M)}} \equiv B_Z. \qquad (A.23)$$

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 et al., 1999) the analysis above demonstrates that the state and the weights are UUB.

This completes the proof. □

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*, *41*(5), 779–791.

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

Baird, L. C. III (1994). Reinforcement learning in continuous time: advantage updating. In *Proc. of ICNN*. Orlando FL.

Beard, R., Saridis, G, & Wen, J. (1997). Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation. *Automatica*, *33*(12), 2159–2177.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. MA: Athena Scientific.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, *12*(1), 219–245.

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

Hanselmann, T., Noakes, L., & Zaknich, A. (2007). Continuous-time adaptive critics. *IEEE Transactions on Neural Networks*, *18*(3), 631–647.

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

Howard, R. A. (1960). *Dynamic programming and markov processes*. Cambridge, MA: MIT Press.

Ioannou, P., & Fidan, B. (2006). *Advances in design and control*, *Adaptive control tutorial*. PA: SIAM.

Kleinman, D. (1968). On an iterative technique for riccati equation computations. *IEEE Transactions on Automatic Control*, *13*(1), 114–115.

Lewis, F. L., Jagannathan, S., & Yesildirek, A. (1999). *Neural network control of robot manipulators and nonlinear systems*. Taylor & Francis.

Lewis, F. L., Liu, K., & Yesildirek, A. (1995). Neural net controller with guaranteed tracking performance. *IEEE Transactions on Neural Networks*, *6*(3), 703–715.

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

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

Nevistic, V., & Primbs, J. A. (1996). Constrained nonlinear optimal control: a converse HJB approach, California Institute of Technology, Pasadena, CA 91125, *Tech rep. CIT-CDS 96-021*.

Prokhorov, D. Prokhorov, & Wunsch, D. (1997). Adaptive critic designs. *IEEE Transactions on Neural Networks*, *8*(5), 997–1007.

Sandberg, E. W. (1998). Notes on uniform approximation of time-varying systems on finite time intervals. *IEEE Transactions on Circuits and Systems—1: Fundamental Theory and Applications*, *45*(8), 863–865.

Si, J., Barto, A., Powel, W., & Wunsch, D. (2004). *Handbook of learning and approximate dynamic programming*. New Jersey: John Wiley.

Sontag, E. D., & Sussmann, H. J. (1995). Nonsmooth control-Lyapunov functions. In *IEEE proc. CDC95* (pp. 2799–2805).

Stevens, B., & Lewis, F. L. (2003). *Aircraft control and simulation* (2nd ed.). New Jersey: John Willey.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning — an introduction*. Cambridge, MA: MIT Press.

Tao, G. (2003). *Adaptive and learning systems for signal processing, communications and control series*, *Adaptive control design and analysis*. Hoboken, NJ: Wiley-Interscience.

Van der Schaft, A. J. (1992). *L*2-gain analysis of nonlinear systems and nonlinear state feedback *H*∞ control. *IEEE Transactions on Automatic Control*, *37*(6), 770–784.

Vrabie, D., & Lewis, F. (2008). Adaptive optimal control algorithm for continuous-time nonlinear systems based on policy iteration. In *IEEE proc. CDC08* (pp. 73–79).

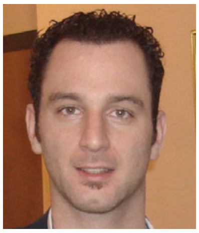
Vrabie, D., Pastravanu, O., Lewis, F. L., & Abu-Khalaf, M. (2009). Adaptive optimal control for continuous-time linear systems based on policy iteration. *Automatica*, *45*(2), 477–484. doi:10.1016/j.automatica.2008.08.017.

Vrabie, D., Vamvoudakis, K. G., & Lewis, F. (2009). Adaptive optimal controllers based on generalized policy iteration in a continuous-time framework, In *IEEE mediterranean conference on control and automation, MED'09* (pp. 1402–1409).

Werbos, P. J. (1974). Beyond regression: new tools for prediction and analysis in the behavior sciences. *Ph.D. thesis*.

Werbos, P.J. (1989). Neural networks for control and system identification. In *IEEE proc. CDC89, Vol. 1* (pp. 260–265).

Werbos, P. J. (1992). Approximate dynamic programming for real-time control and neural modeling. In D. A. White, & D. A. Sofge (Eds.), *Handbook of intelligent control*. New York: Van Nostrand Reinhold.



Kyriakos G. Vamvoudakis was born in Athens, Greece. He received the Diploma in Electronic and Computer Engineering from the Technical University of Crete, Greece in 2006 with highest honors and the M.Sc. Degree in Electrical Engineering from The University of Texas at Arlington in 2008. He is currently working toward the Ph.D. degree and working as a research assistant at the Automation and Robotics Research Institute, The University of Texas at Arlington. His current research interests include approximate dynamic programming, neural network feedback control, optimal control, adaptive control and systems biology. He is a member of Tau Beta Pi, Eta Kappa Nu and Golden Key honor societies and is listed in *Who's Who in the world*.

Mr. Vamvoudakis is a registered Electrical/Computer engineer (PE) and member of Technical Chamber of Greece.

Frank L. Lewis, Fellow IEEE, Fellow IFAC, Fellow UK Institute of Measurement & Control, PE Texas, UK Chartered Engineer, is Distinguished Scholar Professor and Moncrief-O'Donnell Chair at University of Texas at Arlington's Automation and Robotics Research Institute. He obtained the Bachelor's Degree in Physics/EE and the MSEE at Rice University, the MS in Aeronautical Engineering from Univ. W. Florida, and the Ph.D. at Ga. Tech. He works in feedback control, intelligent systems, distributed control systems, and sensor networks. He is author of 6 US patents, 216 journal papers, 330 conference papers, 14 books, 44 chapters, and 11 journal special issues. He received the Fulbright Research Award, NSF Research Initiation Grant, ASEE *Terman Award*, Int. Neural Network Soc. *Gabor Award* 2009, UK Inst Measurement & Control *Honeywell Field Engineering Medal* 2009. He received Outstanding Service Award from Dallas IEEE Section, and was selected as Engineer of the year by Ft. Worth IEEE Section. He is listed in Ft. Worth Business Press Top 200 Leaders in Manufacturing. He served on the NAE Committee on Space Station in 1995. He is an elected Guest Consulting Professor at South China University of Technology and Shanghai Jiao Tong University. He is a Founding Member of the Board of Governors of the Mediterranean Control Association. He helped to win the IEEE Control Systems Society Best Chapter Award (as Founding Chairman of DFW Chapter), the National Sigma Xi Award for Outstanding Chapter (as President of UTA Chapter), and the US SBA Tibbets Award in 1996 (as Director of ARRI's SBIR Program).