Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009

Online Actor Critic Algorithm to Solve the Continuous-Time Infinite Horizon Optimal Control Problem

Kyriakos G. Vamvoudakis, Frank L. Lewis, *Fellow, IEEE*

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. 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 terms in the actor tuning law being required to guarantee closed-loop dynamical stability. The convergence to the optimal controller is proven, and stability of the system is also guaranteed. Simulation examples show the effectiveness of the new algorithm.**

I. Introduction

COST-EFFECTIVE solutions are regularly required in the control loops which are placed at the high levels of the hierarchy of a complex integrated control applications; the so called "money-making" loops. Such a controller is given by the solution of an optimal control problem. Optimal control policies satisfy the specified system performances while minimizing a structured cost index which describes the balance between desired performances 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). For this reason considerable effort has been devoted to developing algorithms which approximately solve this equation (e.g. [1], [3], [13]). Far more results are available for the solution of the discrete-time HJB equation. Good overviews are given in [4], [15].

Some of the methods involve a computational intelligence technique known as Policy Iteration (PI) [8]. 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 the sense defined herein) control policy and then uses this information to obtain a new improved (i.e. which will have a lower associated cost) control policy. 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 cost can be evaluated only in the case of admissible control policies, admissibility being a condition for the control policy which is used to initialize the algorithm.

Actor/critic structures based on Value Iteration have been introduced and further developed by Werbos [19], [20], [21] with the purpose of solving the optimal control problem online in real-time. Werbos defined four types of actor-critic algorithms based on value iteration, subsumed under the concept of Approximate or Adaptive Dynamic Programming (ADP) algorithms. Adaptive critics have been described in [14] for discrete-time systems and [2], [7], [17], [18] for continuous-time systems.

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 [9] which requires iterative solutions of Lyapunov equations. In the nonlinear systems case, successful application of the PI method was limited until [3], 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.

The key to solving practically the CT nonlinear Lyapunov equations was in the use of neural networks (NN) [1] which can be trained to become approximate solutions of these equations. In fact the PI algorithm for CT systems can be built on an actor/critic structure which involves two neural networks: one, the critic NN, is trained to become an approximation of the Lyapunov equation solution at the policy evaluation step, while the second one is trained to approximate an improving policy at the policy improving step.

In [17], [18] 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

This work was supported by the National Science Foundation ECS-0501451 and the Army Research Office W91NF-05-1-0314.

K. G. Vamvoudakis and F. L. Lewis are with the Automation and Robotics Research Institute, University of Texas at Arlington, 7300 Jack Newell Blvd. S., Fort Worth, TX 76118 USA (phone/fax: +817-272-5938; e-mail: {kyriakos, lewis@arri.uta.edu).

978-1-4244-3553-1/09/$25.00 ©2009 IEEE

algorithm was based on *sequential updates* of the critic (policy evaluation) and actor (policy improvement) neural networks (i.e. while one is tuned the other one remains constant).

This paper is concerned with developing approximate online solutions, based on PI, for the infinite horizon optimal control problem for continuous-time (CT) 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 'synchronous' policy iteration. This approach is a version of Generalized Policy Iteration (GPI), as introduced in [16]. An "almost synchronous" version of PI has been described in [7], without providing explicit training laws for either actor or critic networks, nor convergence analysis.

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 certain persistence of excitation condition. The second involves adding nonstandard extra terms to the actor neural network tuning algorithm that are required to guarantee close loop stability, along with stability and convergence proofs.

The paper is organized as follows. Section II provides the formulation of the optimal control problem followed by the general description of neural network value function approximation (VFA). Section III introduces the synchronous online adaptive algorithms for the actor and critic networks based on PI. Results for convergence and stability are given. Section IV presents simulation examples that show the effectiveness of the online synchronous CT PI algorithm.

II. The Optimal Control Problem and the Policy Iteration Algorithm

A. Optimal control and the continuous-time HJB equation

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

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

with state $x(t)\in R^n$, $f(x(t))\in R^n$, $g(x(t))\in R^{n\times m}$ and control input $u(t)\in U\subset\mathbf{R}^m$. We assume that $f(0)=0,\ g(0)=0$, $f(x)+g(x)u$ is Lipschitz continuous on a set $\Omega\subseteq\mathbf{R}^n$ 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)* is assumed to be known.

Define the infinite horizon integral cost

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

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 $R\in\mathbf{R}^{m\times m}$ a positive definite matrix.

Definition 1. [1] (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, $\mu(0)=0$, $\mu(x)$ 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

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

is C^1, then an infinitesimal version of (3) is

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

where $V_x{}^{\mu}$ 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). Equation (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 $r(x,\mu(x))\geq 0$, 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

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

the optimal cost function $V^*(x)$ satisfies the HJB equation

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

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

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

Inserting this optimal control policy in the Hamiltonian we obtain the formulation of the HJB equation in terms of $V_x{}^*$

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

This is a necessary and sufficient condition for the optimal value function [12]. 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.

B. *Neural network approximation of the value function*

Policy iteration is an iterative method of reinforcement learning [16] for solving (8), and consists of policy improvement based on (7) and policy evaluation based on (4). See [1], [2], [3], [7], [8], [13] and [17].

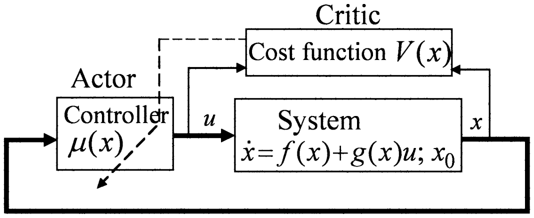

Fig. 1. Actor/Critic Structure.

In the actor/critic structure the Critic and the Actor functions are approximated by neural networks, and the PI algorithm consists in tuning alternatively each of the two neural networks. 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). Thus, assume there exist weights W_1 such that the value $V(x)$ is approximated by a neural network as

$$V(x)=W_1^T\phi_1(x)+\varepsilon(x) \qquad (9)$$

where $\phi_1(x):R^n\to R^N$ is the activation functions vector, N the number of neurons in the hidden layer, and $\varepsilon(x)$ the NN approximation error. It is known that $\varepsilon(x)$ is bounded by a constant on a compact set. Select the activation functions to provide a *complete* basis set such that $V(x)$ and its derivative

$$\frac{\partial V}{\partial x}=\nabla\phi_1^T W_1+\frac{\partial\varepsilon}{\partial x} \qquad (10)$$

are uniformly approximated. According to the Weierstrass higher-order approximation theorem [1], such a basis exists if $V(x)$ is sufficiently smooth. This means that, as the number of hidden-layer neurons $N\to\infty$, the approximation error $\varepsilon\to 0$ uniformly.

Using the NN value function approximation, considering a fixed control policy $u(t)$, the Hamiltonian (5) becomes

$$H(x,u,W_1)=W_1^T\nabla\phi_1(f+gu)+Q(x)+u^T Ru=\varepsilon_H \qquad (11)$$

where the residual error due to the function approximation error is

$$\varepsilon_H=\left(\nabla\varepsilon_x\right)^T(f+gu). \qquad (12)$$

Under the Lipschitz assumption on the dynamics, this residual error is bounded on a compact set. Moreover, in [1] it has been shown that, under certain assumptions, as the number of hidden layer neurons $N\to\infty$, one has $\varepsilon_H\to 0$.

III. ONLINE GENERALIZED PI ALGORITHM WITH SYNCHRONOUS TUNING OF ACTOR AND CRITIC NEURAL NETWORKS

Standard PI algorithms for CT systems are offline methods. We develop an online adaptive learning method for policy iteration that uses simultaneous continuous-time tuning for the actor and critic neural networks. That is, both actor NN and critic NN are tuned at the same time. We term this 'synchronous' online PI for CT systems. The dynamics are assumed to be known.

A. *Critic NN*

The weights of the critic NN, W_1, which solve (11) are unknown. Then the output of the critic neural network is

$$\hat{V}(x)=\hat{W}_1^T\phi_1(x) \qquad (13)$$

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

$$H(x,\hat{W}_1,u)=\hat{W}_1^T\nabla\phi_1(f+gu)+Q(x)+u^T Ru=e_1 \qquad (14)$$

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

$$\dot{\hat{W}}_1=-a_1\frac{\partial E_1}{\partial\hat{W}_1}=-a_1\frac{\sigma_1}{(\sigma_1^T\sigma_1+1)^2}[\sigma_1^T\hat{W}_1+Q(x)+u^T Ru] \qquad (15)$$

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 (15) to be bounded.

Define the critic weight estimation error $\tilde{W}_1=W_1-\hat{W}_1$ and note that, from (11),

$$Q(x)+u^T Ru=-W_1^T\nabla\varphi_1(f+gu)+\varepsilon_H. \qquad (16)$$

Substitute (16) in (15) and, with the notation $\bar{\sigma}_1=\sigma_1/(\sigma_1^T\sigma_1+1)$ and $m_s=1+\sigma_1^T\sigma_1$, we obtain the dynamics of the critic weight estimation error as

$$\dot{\tilde{W}}_1=-a_1\bar{\sigma}_1\bar{\sigma}_1^{\mathrm{T}}\tilde{W}_1+a_1\bar{\sigma}_1\frac{\varepsilon_H}{m_s}. \qquad (17)$$

Though it is traditional to use critic tuning algorithms of the form (15), it is not generally understood when convergence of the critic weights can be guaranteed. In this paper, we address this issue in a formal manner. To guarantee convergence of $\hat{W}_1$ to W_1, the next Persistence of Excitation (PE) assumption 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 $\beta_1 > 0$, $\beta_2 > 0$, $\mathrm{T} > 0$ such that, for all *t*,

$$\beta_1 \mathrm{I} \le S_0 \equiv \int_t^{t+T} \bar{\sigma}_1(\tau)\bar{\sigma}_1^{\mathrm{T}}(\tau)d\tau \le \beta_2 \mathrm{I}\,. \tag{18}$$

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

$$\begin{aligned}\dot{\tilde{W}}_1 &= -a_1\bar{\sigma}_1\bar{\sigma}_1^{\mathrm{T}}\tilde{W}_1 + a_1\bar{\sigma}_1\frac{\varepsilon_H}{m_s}\\ y &= \bar{\sigma}_1^T\tilde{W}_1\,.\end{aligned} \tag{19}$$

The PE condition (18) is equivalent to the uniform complete observability (UCO) [10] of this system, that is there exist constants $\beta_3 > 0$, $\beta_4 > 0$, $\mathrm{T} > 0$ such that, for all *t*,

$$\beta_3 \mathrm{I} \le S_1 \equiv \int_t^{t+T} \Phi^T(\tau,t)\bar{\sigma}_1(\tau)\bar{\sigma}_1^{\mathrm{T}}(\tau)\Phi(\tau,t)d\tau \le \beta_4 \mathrm{I}\,. \tag{20}$$

with $\Phi(t_1, t_0),\, t_0 \le t_1$ the state transition matrix of (19).

Proof: System (19) 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 (18) 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 (15) indeed guarantees boundedness of the output in (19).

Technical Lemma 2. Consider the error dynamics system (19). Let the signal $\bar{\sigma}_1$ be persistently exciting. Then:

a) The system (19) is exponentially stable. In fact if *ε=0* then $\|\tilde{W}(k\mathrm{T})\| \le e^{-\alpha kT}\|\tilde{W}(0)\|$ with

$$\alpha = -\frac{1}{\mathrm{T}}\ln(\sqrt{1-2a_1\beta_3})\,. \tag{21}$$

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

$$\tilde{W}_1(t) \le \frac{\sqrt{\beta_2 \mathrm{T}}}{\beta_1}\left\{\left[y_{\max} + \delta\beta_2 a_1\left(\varepsilon_{\max} + y_{\max}\right)\right]\right\}. \tag{22}$$

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

Proof: See Appendix.

The next result shows that the tuning algorithm (15) is effective, in that, under the PE assumption, the weights $\hat{W}_1$ converge to the actual unknown weights W_1 which solve the Hamiltonian equation (11) for the given control policy *u(t)*. That is, (13) converges close to the actual value function of the current control policy.

Theorem 1. Let $u(t)$ be any admissible bounded control input. Let tuning for the critic NN be provided by (15) and assume that $\bar{\sigma}_1$ is persistently exciting. Let the residual error in (11) be bounded $\|\varepsilon_H\| < \varepsilon_{\max}$. Then the critic parameter error is practically bounded by

$$\tilde{W}_1(t) \le \frac{\sqrt{\beta_2 \mathrm{T}}}{\beta_1}\left\{\left[1 + 2\delta\beta_2 a_1\right]\varepsilon_{\max}\right\}. \tag{23}$$

Proof:

Consider the following Lyapunov function candidate

$$L(t) = \frac{1}{2}tr\{\tilde{W}_1^T a_1^{-1}\tilde{W}_1\}\,. \tag{24}$$

The derivative of L is given by

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

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

$$\begin{aligned}\dot{L} &\le -\|\frac{\sigma_1^T}{m_s}\tilde{W}_1\|^2 + \|\frac{\sigma_1^T}{m_s}\tilde{W}_1\|\left\|\frac{\varepsilon_H}{m_s}\right\|\\ \dot{L} &\le -\|\frac{\sigma_1^T}{m_s}\tilde{W}_1\|\left[\|\frac{\sigma_1^T}{m_s}\tilde{W}_1\| - \left\|\frac{\varepsilon_H}{m_s}\right\|\right]\end{aligned} \tag{25}$$

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

$$\|\frac{\sigma_1^T}{m_s}\tilde{W}_1\| > \varepsilon_{\max} > \left\|\frac{\varepsilon_H}{m_s}\right\|, \tag{26}$$

since $\|m_s\| \ge 1$. This provides an effective practical bound for $\|\bar{\sigma}_1^T\tilde{W}_1\|$, since $L(t)$ decreases if (26) holds.

Consider the estimation error dynamics (19) with the output bounded effectively by $\|y\| < \varepsilon_{\max}$, as just shown. Now Technical Lemma 2 shows that

$$\tilde{W}_1(t) \le \frac{\sqrt{\beta_2 \mathrm{T}}}{\beta_1}\left\{\left[1 + 2a_1\delta\beta_2\right]\varepsilon_{\max}\right\}. \tag{27}$$

This completes the proof. ■

□

Remark 1. Note that, as $N \to \infty$, $\varepsilon_H \to 0$ uniformly [1]. This means that $\varepsilon_{\max}$ decreases as the number of hidden layer neurons in (13) increases.

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

B. Action NN

The policy improvement step in PI is given by substituting (9) into (7) as

$$u(x) = -\tfrac{1}{2} R^{-1} g^T(x) \nabla\phi_1^T W_1$$

with critic weights 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

$$u_2(x) = -\tfrac{1}{2} R^{-1} g^T(x) \nabla\phi_1^T \hat{W}_2 , \tag{28}$$

where $\hat{W}_2$ denotes the current known values of the actor NN weights.

Based on (8), define the approximate HJB equation

$$\begin{gathered} Q(x) + W_1^T \nabla\phi_1(x) f(x) - \frac{1}{4} W_1^T \bar{D}_1(x) W_1 = \varepsilon_{HJB}(x) \\ W_1^T \phi_1(0) + \varepsilon(0) = 0 \end{gathered} \tag{29}$$

with the notation $\bar{D}_1(x) = \nabla\phi_1(x) g(x) R^{-1} g^T(x) \nabla\phi_1^T(x)$, where W_1 denotes the ideal unknown weights of the critic and actor neural networks which solve the HJB.

We now present the main Theorem, which provides the tuning laws for the actor and critic neural networks that guarantee convergence to the optimal controller along with closed-loop stability. The next notion of practical stability is needed.

Definition 2. [10] (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 $t \ge t_0 + T$.

Theorem 2. Let tuning for the critic NN be provided by

$$\dot{\hat{W}}_1 = -a_1 \frac{\sigma_2}{(\sigma_2^T \sigma_2 + 1)^2} [\sigma_2^T \hat{W}_1 + Q(x) + u_2^T R u_2] \tag{30}$$

where $\sigma_2 = \nabla\phi_1(f + g u_2)$, and assume that $\bar{\sigma}_2 = \sigma_2 / (\sigma_2^T \sigma_2 + 1)$ is persistently exciting. Let the actor NN be tuned as

$$\dot{\hat{W}}_2 = -a_2 (\frac{1}{2} \bar{D}_1(x)(\hat{W}_2 - \hat{W}_1) - \frac{1}{4} \bar{D}_1(x) \hat{W}_2 \frac{\bar{\sigma}_2^T}{m_{s2}} \hat{W}_1) \tag{31}$$

where $m_{s2} = 1 + \sigma_2^T \sigma_2$. Then the closed-loop system state is UUB, the critic parameter error $\tilde{W}_1 = W_1 - \hat{W}_1$ and the actor parameter error $\tilde{W}_2 = W_1 - \hat{W}_2$ are UUB, and (23) holds so that convergence of $\hat{W}_1$ to the approximate optimal critic value W_1 is obtained.

Proof

The convergence proof is based on Lyapunov analysis. We consider the Lyapunov function

$$L(t) = V(x) + \frac{1}{2} tr(\tilde{W}_1^T a_1^{-1} \tilde{W}_1) + \frac{1}{2} tr(\tilde{W}_2^T a_2^{-1} \tilde{W}_2) . \tag{32}$$

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. For space reasons we will present the details of this proof in a future paper.

■

Remark. The theorem shows that PE is neded 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 (31) is a cross-product term that involves both the critic weights and the actor weights. It is needed to yield good behavior of the Lyapunov function, i.e. that the energy decreases to a bounded compact region.

IV. 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.

A. Linear system example

Consider the linear system with quadratic cost function used in [7]

$$\dot{x} = \begin{bmatrix} -1 & -2 \\ 1 & -4 \end{bmatrix} x + \begin{bmatrix} 1 \\ -3 \end{bmatrix} u$$

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. 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. The parameters of the optimal critic are then $W_1^* = [0.3199 \quad -0.1162 \quad 0.1292]^T$.

The synchronous PI algorithm is implemented as in Theorem 2. PE was ensured by adding a small probing noise to the control input. Figure 2 shows the critic parameters, denoted by $\hat{W}_1 = [W_{11} \quad W_{12} \quad W_{13}]^T$, converging to the optimal values. In fact after 200s the critic parameters converged to $\hat{W}_1(t_f) = [0.3192 \quad -0.1174 \quad 0.1287]^T$.

The evolution of the system states is presented in Figure 3. One can see that after 200s convergence of the NN weights in both critic and actor has occurred. Then, 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.

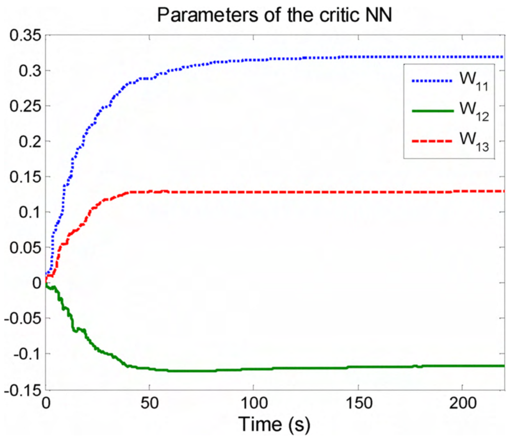

Fig. 2. Convergence of the critic parameters to the parameters of the optimal critic

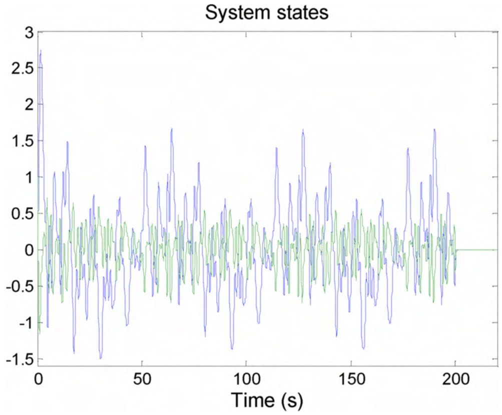

Fig. 3. Evolution of the system states for the duration of the experiment

B. *Nonlinear system example*

Consider the following affine in control input nonlinear system, with a quadratic cost used in [5]

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

where

$$f(x) = \begin{bmatrix} x_2 \\ -x_1(\frac{\pi}{2} + \arctan(5x_1)) - \frac{5x_1^{\ 2}}{2(1+25x_1^{\ 2})} + 4x_2 \end{bmatrix}$$

$$g(x) = \begin{bmatrix} 0 \\ 3 \end{bmatrix}.$$

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

As detailed in [5] the optimal value function is

$$V^*(x) = x_1^{\ 2}(\frac{\pi}{2} + \arctan(5x_1)) + x_2^{\ 2}$$

and the optimal control signal is

$$u^*(x) = -3x_2.$$

One selects the critic NN vector activation function as

$$\phi_1(x) = [x_1 \quad x_2 \quad x_1 x_2 \quad x_1^2 \quad x_2^2 \quad x_1^{\ 2}\arctan(5x_1) \quad x_1^3],$$

in order to have the desired signals to approximate the optimal value function and the optimal control signal.

Figure 4 shows the critic parameters, denoted by

$$\hat{W}_1 = [W_{c1} \quad W_{c2} \quad W_{c3} \quad W_{c4} \quad W_{c5}\ W_{c6} \quad W_{c7}]^T .$$

These converge to the values of

$\hat{W}_1(t_f)$=[0.5674 -0.0239 0.0371 1.7329 0.9741 0.3106 0.4189]^T.

The evolution of the system states is presented in Figure 5. One can see that after 3300s convergence of the NN weights in both critic and actor has occurred. Then, 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.

Figure 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, Figure 7 shows the 3-D 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. The actor NN also converged to the optimal control $u^*(x) = -3x_2$.

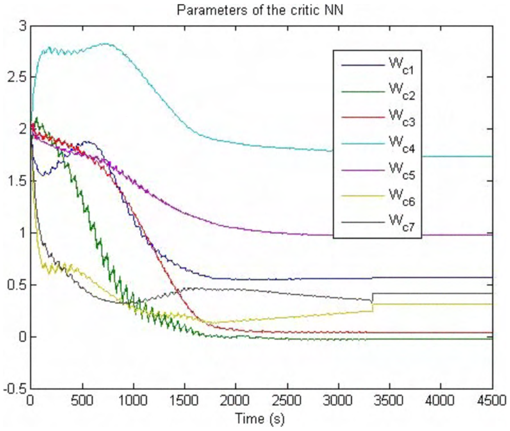

Fig. 4. Convergence of the critic parameters

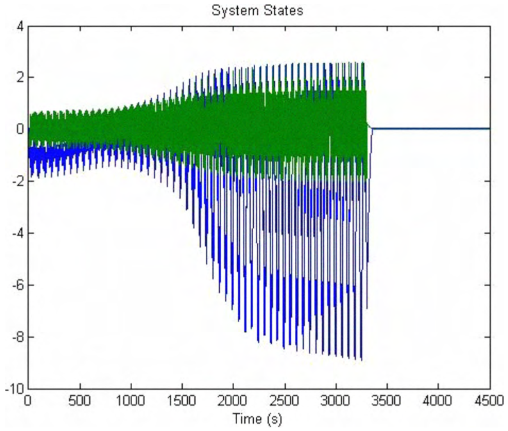

Fig. 5. Evolution of the system states during the first 4500s.

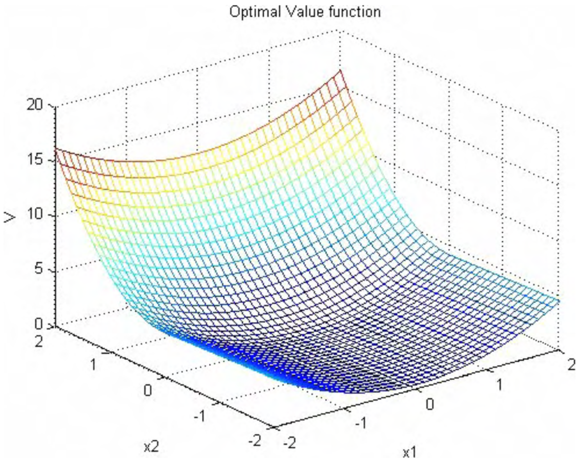

Fig. 6. Optimal Value function.

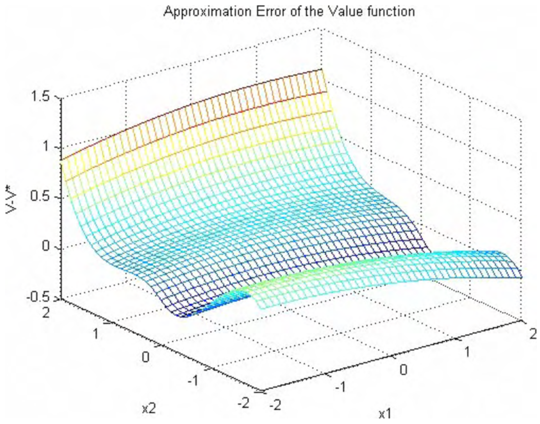

Fig. 7. 3-D plot of the approximation error for the value function.

V. Conclusion

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 [19], [20], [21].

VI. Acknowledgement

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

Appendix

Proof for Technical Lemma 2 Part a: Set $\varepsilon_H = 0$ in (19). Take the Lyapunov function

$$L = \frac{1}{2}\tilde{W}_1^T a_1^{-1}\tilde{W}_1 \qquad \text{(A.1)}$$

The derivative is

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

Integrating both sides

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

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

$$= L(t) - \tilde{W}_1^{\mathrm{T}}(t) S_1 \tilde{W}_1(t) \le (1-2a_1\beta_3)L(t)$$

So

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

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

$$\|\tilde{W}(k\mathrm{T})\| \le \gamma^k \|\tilde{W}(0)\| \qquad \text{(A.3)}$$

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

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

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

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

This completes the proof. ■

Proof for Technical Lemma 2 Part b: Consider the system

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

The state and the output are

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

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

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

Then,

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

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

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

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

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

Taking the norms in both sides yields

$$\|\, x(t)\,\|\le\|\, S_C{}^{-1}\int_t^{t+T} C(\lambda)y(\lambda)d\lambda\,\|$$

$$+\|\, S_C{}^{-1}\left\{\int_t^{t+T} C(\lambda)\left(\int_t^{\lambda} C^T(\lambda)B(\tau)u(\tau)d\tau\right)d\lambda\right\}\|$$

$$\|\, x(t)\,\|\le(\beta_1 \mathrm{I})^{-1}\Big(\int_t^{t+T} C(\lambda)C^T(\lambda)d\lambda\Big)^{\frac{1}{2}}\Big(\int_t^{t+T} y(\lambda)^T y(\lambda)d\lambda\Big)^{\frac{1}{2}}$$

$$+\left\|S_C{}^{-1}\right\|\left\{\int_t^{t+\mathrm{T}}\left\|C(\lambda)C^T(\lambda)\right\|d\lambda\int_t^{t+\mathrm{T}}\|\, B(\tau)u(\tau)\,\|\, d\tau\right\}$$

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

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 $u=-y+\dfrac{\varepsilon_H}{m_s}$ with output given $y=\bar{\sigma}_1{}^T\tilde{W}_1$ turns (A.10) into (19). 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,

$$N\equiv\int_t^{t+T}\|\, B(\tau)\,\|\cdot\|\, u(\tau)\,\|\, d\tau=\int_t^{t+T}\|\, a_1\bar{\sigma}_1(\tau)\,\|\cdot\|\, u(\tau)\,\|\, d\tau$$

$$\le a_1(y_{\max}+\varepsilon_{\max})\int_t^{t+T}\|\,\bar{\sigma}_1(\tau)\,\|\, d\tau$$

$$\le a_1(y_{\max}+\varepsilon_{\max})\left[\int_t^{t+T}\|\,\bar{\sigma}_1(\tau)\,\|^2\, d\tau\right]^{1/2}\left[\int_t^{t+T}1d\tau\right]^{1/2}$$

By using (A.8),

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

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

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

This completes the proof. ■

References

[1] M. Abu-Khalaf, F. L. Lewis, "Nearly Optimal Control Laws for Nonlinear Systems with Saturating Actuators Using a Neural Network HJB Approach", *Automatica*, vol. 41, no. 5, pp. 779-791, 2005.

[2] L. C. Baird III, "Reinforcement Learning in Continuous Time: Advantage Updating", *Proc. Of ICNN*, Orlando FL, vol. 4, pp. 2448-2453, 1994.

[3] R. Beard, G. Saridis, J. Wen, "Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation", *Automatica*, vol. 33, no. 12, pp. 2159–2177, 1997.

[4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, MA, 1996.

[5] J. W. Curtis, R. W. Beard, "Successive Collocation: An Approximation to Optimal Nonlinear Control", *IEEE Proc. ACC01*, vol. 5, pp. 3481-3485, 2001.

[6] K. Doya, "Reinforcement Learning In Continuous Time and Space", *Neural Computation*, 12(1), pp. 219-245, 2000.

[7] T. Hanselmann, L. Noakes, and A. Zaknich, "Continuous-Time Adaptive Critics", *IEEE Transactions on Neural Networks*, 18(3), pp. 631-647, 2007.

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

[9] D. Kleinman, "On an Iterative Technique for Riccati Equation Computations", *IEEE Trans. on Automatic Control*, vol. 13, pp. 114-115, February, 1968.

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

[11] F. L. Lewis, K. Liu, and A.Yesildirek, "Neural Net Controller with Guaranteed Tracking Performance", *IEEE Transactions on Neural Networks*, vol. 6, no. 3, pp. 703-715, 1995.

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

[13] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive Dynamic Programming", *IEEE Trans. on Systems, Man and Cybernetics*, vol. 32, no. 2, pp 140-153, 2002.

[14] D. Prokhorov, D. Wunsch, "Adaptive critic designs," *IEEE Trans. on Neural Networks*, vol. 8, no 5, pp. 997-1007, 1997.

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

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

[17] D. Vrabie, F. Lewis, "Adaptive Optimal Control Algorithm for Continuous-Time Nonlinear Systems Based on Policy Iteration", *IEEE Proc. CDC08*, pp. 73-79, 2008

[18] D. Vrabie, O. Pastravanu, F. Lewis, M. Abu-Khalaf, "Adaptive Optimal Control for Continuous-Time Linear Systems Based on Policy Iteration", *Automatica* (to appear)

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

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

[21] P. Werbos, "Neural networks for control and system identification", *IEEE Proc. CDC89*, vol. 1, pp. 260-265, 1989.