Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Adaptive Optimal Control Algorithm for Continuous-Time Nonlinear Systems Based on Policy Iteration

D. Vrabie and F.L. Lewis, *Fellow IEEE*

Abstract— **In this paper we develop a new online adaptive control scheme, for partially unknown nonlinear systems, which converges to the optimal state feedback control solution for affine in the inputs nonlinear systems. The derivation of the optimal adaptive control algorithm is presented in a continuous-time framework. The optimal control solution will be obtained in a direct fashion, without system identification.**

The algorithm is an online approach to policy iterations based on an adaptive critic structure to find an approximate solution to the state feedback, infinite-horizon, optimal control problem.

I. Introduction

Adaptive control is an on-line design approach which has the objective of maintaining consistent performance of systems which have known structure but unknown constant or slowly time-varying parameter values. An indirect adaptive control strategy has as first step the online estimation of the system parameters followed by model based controller design, whereas the parameters of a direct adaptive controller are directly identified, the plant being described in terms of the controller parameters [10], [20].

Direct adaptive control techniques modify the parameters of the controller in the sense of minimizing the error between the desired output, *i.e.* the output of a reference model, and the output of the closed loop system. Thus, these techniques are not optimal in the sense of minimizing a formal performance function of the sort specified for optimal control.

Optimal adaptive controllers can be obtained using the indirect approach. For linear systems with quadratic cost this requires the solution of the algebraic Riccati equation (ARE) associated with the optimal control. In the nonlinear case the solution of the well known Hamilton Jacobi Bellman (HJB) equation [16] needs to be found. However the HJB equation is generally difficult to solve. Techniques for obtaining approximate solutions for the HJB equation have been developed in [9], [2], [15]. All these methods are offline approaches which require prior knowledge of the system dynamics, thus the approximate optimal controllers derived using these techniques are not sensitive to changes in the system dynamics and thus are not adaptive. Stabilizing adaptive controllers that are inverse optimal, with respect to some relevant cost not specified by the designer, have been derived and analyzed in [17], [14].

This paper proposes an adaptive strategy to determine online an approximate optimal controller for partially unknown, affine in the input, nonlinear systems, without prior identification of the nonlinear system's internal dynamics. The online strategy is constructed using a policy iteration technique, first formulated in [8], who alternates between *policy evaluation* and *policy improvement* steps. Various results using offline policy iterations for solving the optimal control problem have been presented and discussed, including convergence guarantees, in [12], [2], [1], [6].

The main contribution of this paper is given by the online quality of the policy iteration algorithm, which at the same time does not require knowledge of the system internal dynamics, and thus can be viewed as a direct optimal adaptive control technique. Unlike the regular adaptive controllers which rely on online identification of the system dynamics followed by model based controller design, the policy iteration method we are proposing here relies on identification of the cost function associated with a given control policy followed by policy improvement in the sense of minimizing the identified cost.

The convergence guarantees of the continuous time policy iteration technique to the optimal controller was given for linear systems in [12], as well as for nonlinear systems in both unconstrained and constrained control case in [2] and [1]. Implementation of these algorithms requires complete knowledge of the system dynamics. Online adaptive partially model free algorithms based on policy iteration algorithms for approximate optimal control have been developed in [18] and [21] for linear continuous time systems. Both approaches avoid the necessity of knowing the internal system dynamics.

We now propose a new policy iteration technique that will solve in an online fashion, along a single state trajectory, the optimal control problem for continuous-time nonlinear systems using only partial knowledge about the system dynamics (*i.e.* the internal dynamics of the system need not be known). This is in fact a direct adaptive control scheme for partially unknown systems that converges to the optimal control solution without any knowledge on the plant internal description (*i.e.* the internal dynamics of the plant need not be in a specific parametric form).

The following section of the paper includes a short

Manuscript received February 29, 2008. This work was supported by the National Science Foundation ECS-0501451 and the Army Research Office W91NF-05-1-0314.

D. Vrabie and F. 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: dvrabie@uta.edu).

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

overview of nonlinear optimal control and the HJB equation, the derivation of the proposed algorithm with convergence analysis, the description of the adaptive controller structure, and the online implementation approach. The third section presents the adaptive optimal control results obtained in simulation while considering a linear and a nonlinear system.

II. Continuous-Time Adaptive Critic Solution for the Infinite Horizon Optimal Control Problem

A. Optimal control and the continuous-time HJB equation

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

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

with $x(t)\in R^n$, $f(x(t))\in R^n$, $g(x(t))\in R^{n\times m}$ and the input $u(t)\in U\subset\mathbb{R}^m$. We assume that $f(x)+g(x)u$ is Lipschitz continuous on a set $\Omega\subseteq\mathbb{R}^n$ that contains the origin and that the dynamical 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.

Define the infinite horizon integral cost

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

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

Definition 1 [2] (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

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

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

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

where V_x^μ denotes the partial derivative of the value function V^μ with respect to x, as the value function does not depend explicitly on time. Equation (4) is a Lyapunov equation for nonlinear systems which, given the 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 $u\in\Psi(\Omega)$ 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

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

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

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

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

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

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

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

This is a necessary and sufficient condition for the optimal value function [11]. 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, solving the HJB equation is generally difficult. It also requires complete knowledge of the system dynamics (*i.e.* the functions $f(x), g(x)$ need to be known).

B. Adaptive optimal control algorithm based on policy iterations

In the following we propose a new online iterative algorithm which will adapt to solve the infinite horizon optimal control problem without using knowledge regarding the system internal dynamics (*i.e.* the system function $f(x)$).

Let $\mu(x)$ be an admissible policy for (1), such that the closed loop system is asymptotically stable on Ω. Then the infinite horizon cost for any $x(t)\in\Omega$ is given by (3) and $V^\mu(x(t))$ serves as a Lyapunov function for (1). The cost function (3) can be written as

$$V^\mu(x(t))=\int_t^{t+T} r(x(\tau),\mu(x(\tau)))d\tau+V^\mu(x(t+T)). \qquad (9)$$

Based on (9) and (6), considering an initial admissible control policy $\mu^{(0)}(x)$, the following policy iteration scheme can be derived

1. solve for $V^{\mu^{(i)}}(x)$ using

$$\begin{aligned}&V^{\mu^{(i)}}(x(t))=\int_t^{t+T} r(x(\tau),\mu^{(i)}(x(\tau)))d\tau+V^{\mu^{(i)}}(x(t+T)),\\&V^{\mu^{(i)}}(0)=0\end{aligned} \qquad (10)$$

2. update the control policy using

$$\mu^{(i+1)}(x)=\arg\min_{\mu}\{H(x,\mu,V_x^{\mu^{(i)}})\} \quad (11)$$

which is

$$\mu^{(i+1)}(x)=-R^{-1}g^T(x)V_x^{\mu^{(i)}}(x)\,. \quad (12)$$

Equations (10) and (12) formulate a new policy iteration algorithm to solve for the optimal control without making use of any knowledge of the system internal dynamics $f(x)$. The implementation of the algorithm is straightforward and will be discussed in section II-*D*. This algorithm is an online version of the offline algorithms proposed in [1], [2] motivated by the success of the online adaptive critic techniques proposed by computational intelligence researchers [22], [19], [4].

C. Convergence analysis

In this section we prove the convergence of the online optimal adaptive control algorithm.

Lemma 1 Solving for $V^{\mu^{(i)}}$ in equation (10) is equivalent with finding the solution of the Lyapunov equation

$$0=r(x,\mu^{(i)}(x))+V_x^{\mu^{(i)}T}(f(x)+g(x)\mu^{(i)}(x)),\ V^{\mu^{(i)}}(0)=0\,. \quad (13)$$

Proof

Since $\mu^{(i)}$ is an admissible control policy over Ω then the function $V^{\mu^{(i)}}$, defined as in (3), satisfies equation (13) with $r(x(t),\mu^{(i)}(x(t)))>0; x(t)\neq 0$ and is a Lyapunov function of the system. Integrating (13) over the interval [*t*,*t*+*T*] one obtains

$$-\int_t^{t+T}\frac{dV^{\mu^{(i)}}(x(t))}{dt}dt=\int_t^{t+T}r(x(\tau),\mu^{(i)}(x(\tau)))d\tau\,,$$

which is equation (10)

$$V^{\mu^{(i)}}(x(t))=\int_t^{t+T}r(x(\tau),\mu^{(i)}(x(\tau)))d\tau+V^{\mu^{(i)}}(x(t+T))\,.$$

This means that the solution of the Lyapunov equation (13) $V^{\mu^{(i)}}$ satisfies also equation (10). To complete de proof we will now show that equation (10) has a unique solution.

Assume that there exists another cost function V, continuously differentiable, such that

$$V(x(t))=\int_t^{t+T}r(x(\tau),\mu^{(i)}(x(\tau)))d\tau+V(x(t+T)),\ V(0)=0\,. \quad (14)$$

This cost function also satisfies

$$\dot V(x(t))=-r(x(t),\mu^{(i)}(x(t)))\,. \quad (15)$$

Subtracting (15) from (13) we obtain

$$\left(\frac{d[V(x)-V^{\mu^{(i)}}(x)]}{dx}\right)^T[f(x)+g(x)\mu^{(i)}(x)]=0 \quad (16)$$

which must hold for any $x(t)$ on the system trajectories generated by the admissible policy $\mu^{(i)}$. Thus $V(x)=V^{\mu^{(i)}}(x)+c, \forall x\in\Omega$. The relation must hold for $x=0$ which implies that $c=0$ and thus $V(x)=V^{\mu^{(i)}}(x), \forall x\in\Omega$, *i.e.* equation (10) has a unique solution. ∎

Remark 1 Although the same solution is obtained whether solving the equation (10) or (13), solving equation (10) does not require any knowledge on the system dynamics $f(x)$.

From Lemma 1 it follows that the algorithm (10) and (12) is equivalent to iterating between (13) and (12), without using knowledge of the system internal dynamics.

Theorem 1 *(convergence)* The policy iteration (10) and (12) converges to the optimal control solution on the trajectories having initial state $x_0\in\Omega$.

Proof: In [2], [1] it was shown that using policy iteration conditioned by an initial admissible policy $\mu^{(0)}(x)$, all the subsequent control policies will be admissible and the iteration (13) and (12) will converge to the solution of the HJB equation.

Based on the proven equivalence between the equations (10) and (13) we can conclude that the proposed online adaptive optimal control algorithm will converge to the solution of the optimal control problem (2), on any subset $\Omega_{x_0}^{\mu^{(i)}}\subset\Omega$, without using knowledge on the internal dynamics of the controlled system (1). ∎

D. Online implementation of the algorithm without using knowledge of the system internal dynamics

For the implementation of the iteration scheme given by (10) and (12) one only needs to have knowledge of the input to state dynamics, *i.e.* the function $g(x)$, which is required for the policy update in equation (12). One can see that knowledge on the internal state dynamics, described by $f(x)$, is not required. The information regarding the system $f(x)$ matrix is embedded in the states $x(t)$ and $x(t+T)$ which are sampled online.

In order to solve for the cost function $V^{\mu^{(i)}}(x)$ in equation (10) we will use a neural network to obtain an approximation of the value function for any given initial state $x\in\Omega$. Due to the universal approximation property [7], a neural network is a natural choice for this application. The cost function $V^{\mu^{(i)}}(x(t))$ will be approximated by

$$V^{\mu^{(i)}}(x)=\sum_{j=1}^{L}w_j^{\mu^{(i)}}\phi_j(x)=(\mathbf{w}_L^{\mu^{(i)}})^T\boldsymbol{\varphi}_L(x) \quad (17)$$

which is a neural network with *L* neurons on the hidden layer and activation functions $\phi_j(x)\in C^1(\Omega), \phi_j(0)=0$. $w_j^{\mu^{(i)}}$ denote the weights of the neural network, $\boldsymbol{\varphi}_L(x)$ is the vector of activation functions and $\mathbf{w}_L^{\mu^{(i)}}$ is the weight vector. Note that there exists an approximation error between the neural network and the true value of the cost function. This issue will be addressed in a future paper while we continue the following derivations assuming that the neural network is an exact description of the cost function.

Using the neural network description for the value function, equation (17), equation (10) can be written as

$$\mathbf{w}_L^{\mu^{(i)}T}\boldsymbol{\varphi}_L(x(t))=\int_t^{t+T} r(x,\mu^{(i)}(x))d\tau+\mathbf{w}_L^{\mu^{(i)}T}\boldsymbol{\varphi}_L(x(t+T)). \tag{18}$$

As the cost function was replaced with the neural network approximation, equation (18) will have the residual error

$$\delta_L^i(x(t))=\int_t^{t+T} r(x,\mu^{(i)}(x))d\tau+\mathbf{w}_L^{\mu^{(i)}T}[\boldsymbol{\varphi}_L(x(t+T))-\boldsymbol{\varphi}_L(x(t))]. \tag{19}$$

From the perspective of temporal difference learning methods [5], [3] this error can be viewed as temporal difference residual error.

Denote with $\Omega_{x_0}^{\mu^{(i)}}\subset\Omega$ the trajectory generated by the control policy $\mu^{(i)}(x)$, starting form the initial state x_0. Note that, as the proposed method is in the class of adaptive control techniques all the computations must be performed based on the information that can be acquired form the system along a trajectory generated by a given admissible control policy.

To determine the parameters of the neural network approximating the cost function, in the least-squares sense, we use the method of weighted residuals. Thus we seek to minimize the objective

$$S=\int_{\Omega_{x_0}^{\mu^{(i)}}}\delta_L^i(x)\delta_L^i(x)dx\,. \tag{20}$$

This amounts to $\int_{\Omega_{x_0}^{\mu^{(i)}}}\dfrac{d\delta_L^i(x)}{d\mathbf{w}_L^{\mu^{(i)}}}\delta_L^i(x)dx=0$.

Using the inner product notation for the Lebesgue integral one can write

$$\left\langle \frac{d\delta_L^i(x)}{d\mathbf{w}_L^{\mu^{(i)}}},\delta_L^i(x)\right\rangle_{\Omega_{x_0}^{\mu^{(i)}}}=0$$

which is

$$\left\langle[\boldsymbol{\varphi}_L(x(t+T))-\boldsymbol{\varphi}_L(x(t))],[\boldsymbol{\varphi}_L(x(t+T))-\boldsymbol{\varphi}_L(x(t))]^T\right\rangle\mathbf{w}_L^{\mu^{(i)}}+$$
$$+\left\langle[\boldsymbol{\varphi}_L(x(t+T))-\boldsymbol{\varphi}_L(x(t))],\int_t^{t+T} r(x(s),\mu^{(i)}(x(s)))ds\right\rangle=0$$

Conditioned by

$\Phi=\left\langle[\boldsymbol{\varphi}_L(x(t+T))-\boldsymbol{\varphi}_L(x(t))],[\boldsymbol{\varphi}_L(x(t+T))-\boldsymbol{\varphi}_L(x(t))]^T\right\rangle$ being invertible, then we obtain

$$\mathbf{w}_L^{\mu^{(i)}}=-\Phi^{-1}\left\langle[\boldsymbol{\varphi}_L(x(t+T))-\boldsymbol{\varphi}_L(x(t))],\int_t^{t+T} r(x(s),\mu^{(i)}(x(s)))ds\right\rangle \tag{21}$$

To show that matrix Φ is invertible the following technical results are needed.

Definition 2 [13] (linearly independent set of functions) A set of functions $\{\phi_j\}_1^N$ is said to be linearly independent if $\sum_{j=1}^N c_j\phi_j(x)=0$ a.e. on $\mathbb{R}$ implies that $c_1=\cdots=c_N=0$.

Lemma 2 If the set $\{\phi_j\}_1^N$ is linearly independent and $u\in\Psi(\Omega)$ then the set $\left\{\nabla\phi_j^T(f+gu)\right\}_1^N$ is also linearly independent.

For the proof see [2]. ■

We now introduce a lemma proving that Φ can be inverted.

Lemma 3 Let $\mu(x)\in\Psi(\Omega)$ such that $f(x)+g(x)\mu(x)$ is asymptotically stable. If the set $\{\phi_j\}_1^N$ is linearly independent then $\exists T>0$ such that $\forall x(t)\in\Omega$ the set $\left\{\bar{\phi}_j(x(t),T)=\phi_j(x(t+T))-\phi_j(x(t))\right\}_1^N$ is also linearly independent.

Proof

If the vector field $\dot{x}=f(x)+g(x)\mu(x)$ is asymptotically stable then along the system trajectories $\varphi(\tau;x(t),\mu),x(t)\in\Omega$, we have that

$$\begin{aligned}\phi(x(t))&=-\int_t^\infty\phi_x(f+g\mu)(\varphi(\tau;x(t),\mu))d\tau\\&=-\int_t^{t+T}\phi_x(f+g\mu)(\varphi(\tau;x(t),\mu))d\tau+\phi(x(t+T))\end{aligned} \tag{22}$$

$$\phi(x(t+T))-\phi(x(t))=\int_t^{t+T}\frac{\partial\phi}{\partial x}(f+g\mu)(\varphi(\tau;x(t),\mu))d\tau \tag{23}$$

Suppose that the lemma is not true. Then for all $T>0$ there exists a nonzero constant vector $c\in\mathbb{R}^N$ such that $\forall x_0\in\Omega\quad c^T[\phi(x(t+T))-\phi(x(t))]\equiv0$. This implies that $\forall T>0$, $c^T\int_t^{t+T}\frac{\partial\phi}{\partial x}(f+g\mu)(\varphi(\tau;x(t),\mu))d\tau\equiv0$ and thus, $\forall x(t)\in\Omega$, $c^T\frac{\partial\phi}{\partial x}(f+g\mu)(\varphi(\tau;x(t),\mu))\equiv0$. This means that $\left\{\nabla\phi_j^T(f+gu)\right\}_1^N$ is not linearly independent contradicting Lemma 2. Thus $\exists T>0$ such that $\forall x(t_0)\in\Omega$ the set $\left\{\bar{\phi}_j(x(t_0),T)\right\}_1^N$ is also linearly independent. ■

Based on the result of Lemma 3, there exist values of T such that the matrix Φ is invertible and the parameters W_i of the cost function can be calculated. The selection of T for the online implementation is related to the excitation condition requirement and will be addressed in a future paper.

After the parameters of the parameters $\mathbf{w}_L^{\mu^{(i)}}$ of the neural network approximating the cost function $V^{\mu^{(i)}}(x(t))$ have been determined, the new improved control policy can be simply calculated as

$$\mu^{(i+1)}(x) = -R^{-1} g^T(x) \left(\frac{\partial \varphi_L(x)}{\partial x} \right)^T \mathbf{w}_L^{\mu^{(i)}} . \tag{24}$$

The flowchart of the online algorithm is presented in Fig. 1.

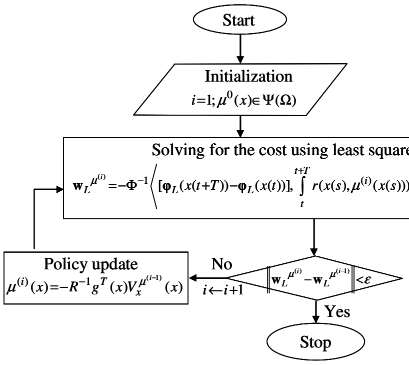
Start
Initialization
$i=1; \mu^0(x) \in \Psi(\Omega)$
Solving for the cost using least squares
$\mathbf{w}_L^{\mu^{(i)}} = -\Phi^{-1} \left\langle [\boldsymbol{\varphi}_L(x(t+T)) - \boldsymbol{\varphi}_L(x(t))], \int_t^{t+T} r(x(s), \mu^{(i)}(x(s)))ds \right\rangle$
$\left\| \mathbf{w}_L^{\mu^{(i)}} - \mathbf{w}_L^{\mu^{(i-1)}} \right\| < \varepsilon$
No
$i \leftarrow i+1$
Yes
Stop
Policy update
$\mu^{(i)}(x) = -R^{-1} g^T(x) V_x^{\mu^{(i-1)}}(x)$

Figure 1. Flowchart of the online algorithm

The solution given by (18) can be obtained in real-time after a sufficient number of data points are collected along a single state trajectory. In practice, the matrix inversion in (18) is not performed, the solution of the equation being obtained using algorithms that involve techniques such as Gaussian elimination, backsubstitution, and Householder reflections. Equation (18) can also be solved by Recursive Least Squares (RLS) in which case a persistence of excitation condition is required. It has to be emphasized that, in order to successfully apply the algorithm, enough excitation must be present in the system. Thus, if the system state reached the equilibrium point (note that the algorithm iterates only on stabilizing controllers), the data measured from the system can no longer be used in the adaptive algorithm; in this case the system must be again excited to the previously considered initial state and a new experiment needs to be conducted having as starting point the last policy obtained in the previous experiment.

The iterations will be stopped (*i.e.* the critic will stop updating the control policy) when the error between the system performance evaluated at two consecutive steps will cross below a designer specified threshold. Also, when this error becomes bigger than the above mentioned threshold the critic will take again the decision to start tuning the actor parameters.

The next section presents the structure of the system with adaptive controller.

E. Control system structure and implementation issues

The proposed optimal adaptive procedure requires only measurements of the states at discrete moments in time, t and $t+T$, as well as knowledge of the observed cost over the time interval $[t, t+T]$. Therefore there is no required knowledge about the system dynamics $f(x)$ for the evaluation of the cost or the update of the control policy. However the $g(x)$ matrix is required for the update of the control policy, using (13), and this makes the online tuning algorithm only partially model free. The control policy is updated at time *t+T*, after observing the state *x*(*t+T*) and it will be used for controlling the system during the time interval [*t+T*, *t+2T*]; thus the algorithm is suitable for online implementation from the control theory point of view.

The structure of the system with the adaptive controller is presented in Fig. 2.

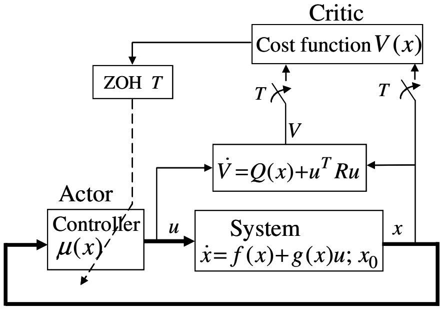

Figure 2. Structure of the system with adaptive controller

Most important is that the system has to be augmented with an extra state $V(t)$, with $\dot{V} = Q(x) + u^T Ru$, in order to extract the information regarding the cost associated with the given policy. It is shown that having little information about the system states, *x*, and the augmented system state, *V*, extracted from the system only at specific time values (*i.e.* $x(t), x(t+T)$ and $V(t+T) - V(t)$), the critic is able to evaluate the performance of the system associated with a given control policy. Then a policy improvement takes place at time *t+T*.

It is observed that the update of both the actor and the critic is performed at discrete moments in time. However, the control action is a full fledged continuous-time control, with its constant gain updated at discrete moments in time. Moreover, the critic update is based on the observations of the continuous-time cost over a finite sample interval. As a result, the algorithm converges to the solution of the continuous-time optimal control problem, as proven in II-*C*.

III. Optimal Adaptive Controller Design for a nonlinear system

In this section we illustrate the results of the adaptive optimal control algorithm considering the nonlinear system in [2] given by the equations

$$\begin{aligned} \dot{x}_1 &= -x_1^3 - x_2 \\ \dot{x}_2 &= x_1 + x_2 + u \end{aligned} . \tag{25}$$

A. Linear case

For a first case we consider a linear version of the system (25), not including the cubic term in the dynamics of the first state, described by the following equations

$$\dot{x} = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u . \tag{26}$$

The simulation was conducted using data obtained from the system at every 0.3s. For the purpose of demonstrating the algorithm the initial state is taken to be different than zero. The initial stabilizing controller was taken to be the truncated version of the initial stabilizing controller derived in [2] for the nonlinear system which is

$$\mu_0(x)=[0.4142 \quad -2.35]x\,. \tag{27}$$

The cost function parameters, namely the Q and R matrices, were chosen to be identity matrices of appropriate dimensions. The following smooth function was used to approximate the cost function of the system

$$\begin{aligned}V(x_1,x_2)=&w_1 x_1^2+w_2 x_1 x_2+w_3 x_2^2+w_4 x_1^4+w_5 x_1^3 x_2+\\+&w_6 x_1^2 x_2^2+w_7 x_1 x_2^3+w_8 x_2^4+w_9 x_1^6+w_{10} x_1^5 x_2+\\+&w_{11} x_1^4 x_2^2+w_{12} x_1^3 x_2^3+w_{13} x_1^2 x_2^4+w_{14} x_1 x_2^5+w_{15} x_2^6\end{aligned} \tag{28}$$

In order to solve online for the neural network weights $w_i, i=\overline{1,15}$ which parameterize the cost function, before each iteration step one needs to setup a least squares problem with the solution given by (13). As the considered neural network has 15 weights we can setup a least squares problem by measuring the cost function associated with a given control policy over 15 time intervals T=0.3s, the initial state and the system state at the end of each time interval. In this way, at every 4.5s, enough data is collected from the system to solve for the cost function and perform a policy update. The result of applying the algorithm is presented in Fig. 3.

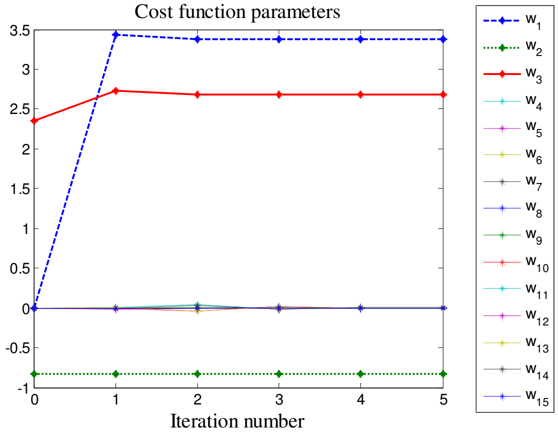

Figure 3. Parameters of the cost function converging to the optimal values

The experiment was performed along the state trajectory having as initial state $x_0=[0.1 \quad 0.1]^T$. The cost function converged to

$$V(x_1,x_2)=3.3784x_1^2\text{-}0.8284x_1x_2+2.6818x_2^2\,, \tag{29}$$

the last 12 parameters being close to zero. The resulting control policy is

$$\mu_5(x)=0.4142x_1\text{-}2.6818x_2\,. \tag{30}$$

This result is consistent with the solution of the Riccati equation underlying the optimal control problem in the linear case.

From Fig. 3 it is clear that the parameters of the cost function, and implicitly the parameters of the control policy, converged after two iteration steps were performed. Thus, after two iteration steps the system will be controlled in an optimal fashion with the controller which was adapted online without using knowledge about the system's internal dynamics.

B. *Nonlinear case*

The proposed adaptive optimal control algorithm is now used with a nonlinear system (25). The required initial stabilizing controller for this system was (27) and the same initial state was chosen $x_0=[0.1 \quad 0.1]^T$. The cost function was approximated as in (28). The evolution of the cost function parameters is presented in Fig. 4.

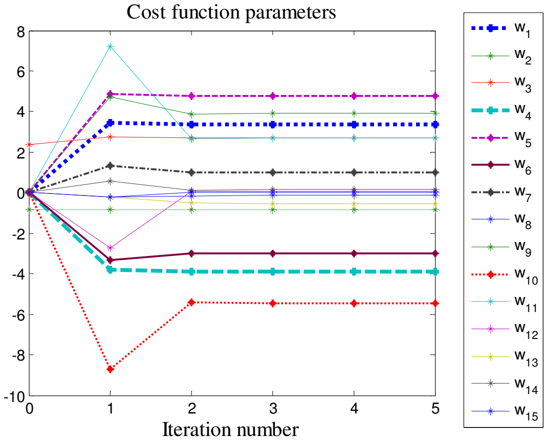

Figure 4. Parameters of the cost function converging to the optimal values

The optimal controller obtained over this trajectory is

$$\begin{aligned}\mu_5(x)=&0.4142x_1\text{-}2.6818x_2-2.3889x_1^3+2.9829x_1^2x_2-\\-&1.5202x_1x_2^2+0.3013x_2^3+2.7257x_1^5-2.6802x_1^4x_2-\\-&0.1213x_1^3x_2^2+1.1147x_1^2x_2^3-0.3917x_1x_2^4-0.016x_2^5\end{aligned} \tag{31}$$

Notice that the first terms are the same with the terms in the controller for the linear system (30).

An experiment for a cost function using terms up to the power 8 was next performed and the result (the weights corresponding to the high order terms were close to zero) indicates that the 6th order polynomial (28) provides a good approximation for the cost function.

IV. Conclusion

In this paper we proposed a new adaptive controller based on policy iteration to solve on-line the continuous time optimal control problem without using knowledge about the system's internal dynamics. Convergence of the proposed algorithm, under the condition of initial stabilizing controller, to the solution of the optimal control problem has been established. Simulation results support the effectiveness of the online adaptive optimal controller.

Issues such as the neural network approximation error, the choice of the sampling time will be addressed in a future extended paper.

Acknowledgements

The authors would like to thank M. Abu-Khalaf and O. Pastravanu for their insightful suggestions which had an influence on this paper.

REFERENCES

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

[2] R. Beard, G. Saridis, and J. Wen, “Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation”, *Automatica*, *33*(12), 2159–2177, 1997.

[3] L. C. Baird III, “Reinforcement Learning in Continuous Time: Advantage Updating”, *Proc. Of ICNN*, Orlando FL, June 1994.

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

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

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

[7] K. Hornik, M. Stinchcombe and H. White, “Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks”, *Neural Networks*, 3, 551–560, 1990.

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

[9] J. Huang, and C. F. Lin, “Numerical approach to computing nonlinear H_∞ control laws”, *Journal of Guidance, Control and Dynamics*, 18(5), 989–994, 1995.

[10] P. Ioannou and B. Fidan, *Adaptive Control Tutorial*, SIAM, Advances in Design and Control, PA, 2006.

[11] D. E. Kirk, *Optimal Control Theory – an introduction*, Dover Pub. Inc., Mineola, New York, 2004.

[12] D. Kleinman, “On an Iterative Technique for Riccati Equation Computations”, *IEEE Trans. on Automatic Control*, 13, 114-115, 1968.

[13] A. N. Kolmogorov and S. V. Fomin, *Elements of the Theory of Functions and Functional Analysis*, Dover Pub. Inc., Mineola, New York, 1999.

[14] M. Krstic and H. Deng, *Stabilization of Nonlinear Uncertain Systems*, Springer, 1998.

[15] H. W. J. Lee, K. L. Teo, W. R. Lee and S. Wang “Construction of suboptimal feedback control for chaotic systems using B-splines with optimally chosen knot points”, *International Journal of Bifurcation and Chaos*, 11(9), 2375–2387, 2001.

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

[17] Z. H. Li and M. Krstic, “Optimal design of adaptive tracking controllers for nonlinear systems”, *Proc. of ACC*, 1191-1197, 1997.

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

[19] D. Prokhorov and D. Wunsch, “Adaptive critic designs,” *IEEE Trans. on Neural Networks*, 8(5), 997-1007, 1997.

[20] J. J. Slotine and W. Li, *Applied Nonlinear Control*, New Jersey: Prentice-Hall, Inc, 1991.

[21] D. Vrabie, O. Pastravanu, F. L. Lewis, “Policy Iteration for Continuous-time Systems with Unknown Internal Dynamics”, *Proceedings of MED*, 2007.

[22] 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.