ISA Transactions 52 (2013) 611–621

Contents lists available at ScienceDirect

ISA Transactions

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

A policy iteration approach to online optimal control of continuous-time constrained-input systems

Hamidreza Modares [a,*], Mohammad-Bagher Naghibi Sistani [a], Frank L. Lewis [b]

[a] Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad 91775-1111, Iran
[b] University of Texas at Arlington Research Institute, 7300 Jack Newell Blvd. S., Ft. Worth, TX 76118, USA

ARTICLE INFO

Article history:
Received 15 March 2012
Received in revised form
23 January 2013
Accepted 6 April 2013
Available online 24 May 2013

Keywords:
Optimal control
Reinforcement learning
Policy iteration
Neural networks
Input constraints

ABSTRACT

This paper is an effort towards developing an online learning algorithm to find the optimal control solution for continuous-time (CT) systems subject to input constraints. The proposed method is based on the policy iteration (PI) technique which has recently evolved as a major technique for solving optimal control problems. Although a number of online PI algorithms have been developed for CT systems, none of them take into account the input constraints caused by actuator saturation. In practice, however, ignoring these constraints leads to performance degradation or even system instability. In this paper, to deal with the input constraints, a suitable nonquadratic functional is employed to encode the constraints into the optimization formulation. Then, the proposed PI algorithm is implemented on an actor–critic structure to solve the Hamilton–Jacobi–Bellman (HJB) equation associated with this nonquadratic cost functional in an online fashion. That is, two coupled neural network (NN) approximators, namely an actor and a critic are tuned online and simultaneously for approximating the associated HJB solution and computing the optimal control policy. The critic is used to evaluate the cost associated with the current policy, while the actor is used to find an improved policy based on information provided by the critic. Convergence to a close approximation of the HJB solution as well as stability of the proposed feedback control law are shown. Simulation results of the proposed method on a nonlinear CT system illustrate the effectiveness of the proposed approach.

1. Introduction

The optimal control of continuous-time (CT) nonlinear systems is a challenging subject in control engineering. Solving such a problem requires solving the Hamilton–Jacobi–Bellman (HJB) equation [1], which has remained intractable in all but very special problems. This has inspired researchers to present various approaches for obtaining approximate solutions to the HJB equation.

One of the existing approximation approaches for solving the HJB equation is a power-series based method [2,3]. This approach separates the system nonlinearities into a power-series and then, to avoid prohibitive computational effort, computes local estimates by using only a few terms of the series. The second approach to approximate the HJB solution is the state-dependent Riccati equation (SDRE) [4,5]. This approach is the extension of the well-known Riccati equation to nonlinear systems. But solving the SDRE is much more difficult than solving the Riccati equation, because the coefficients in the SDRE are functions of the states instead of being constant-valued as in the Riccati equation.

Another elegant approach to approximate the HJB solution is policy iteration (PI) [6], where an iterative process is used to find a sequence of approximations converging to the solution of the HJB equation. PI is a class of reinforcement learning (RL) [7,8] methods that have two-step iterations: policy evaluation and policy improvement. In the policy evaluation step, the cost associated with a control policy is evaluated by solving a nonlinear Lyapunov equation (LE). In the policy improvement step, the algorithm finds an improved policy under which the system performs better. These two steps are repeated until the policy converges to a near-optimal policy. Considerable research has been conducted for approximating the HJB solution of discrete-time systems using PI algorithms [9–30]. However, due to the complex nature of the HJB equation for nonlinear CT systems, only few results are available [31–38].

The first practical PI algorithm developed for nonlinear CT systems was proposed by Beard [31]. He utilized the Galerkin approximation method to find approximate solutions to the LE in the policy evaluation step of the PI algorithm. However, Galerkin approximation method requires the evaluation of numerous integrals, which is computationally intensive [32]. A computationally effective algorithm to find near-optimal control laws was presented

* Corresponding author. Tel.: +98 9155624453.
E-mail addresses: Ha.modarres@stu-mail.um.ac.ir, reza_modares@yahoo.com (H. Modares).

0019-0578/$ - see front matter © 2013 ISA. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.isatra.2013.04.004

by Abu-Khalaf and Lewis [32]. They used neural network (NN) approximators to approximate solutions to the LE. Their results showed the suitability of NN approximators for PI methods. Although efficient, both methods presented in [31,32] are offline techniques. Developing online learning algorithms for solving optimal control problems is of great interest in the control systems society, since in this manner additional approaches such as adaptive control can be integrated with the optimal control to develop adaptive optimal control algorithms for systems with parametric uncertainties or even unknown dynamics.

An online PI algorithm was first presented by Doya [33] for optimal control of CT systems. Nevertheless, this algorithm was not shown to guarantee the stability of the control system. Murray et al. [34], proposed a PI algorithm which converges to the optimal control solution without using an explicit, a priori obtained, model of the drift dynamics of the system. However, it requires measurements of the state derivatives. Vrabie and Lewis [35] presented an online PI algorithm which solves the optimal problem, using only partial knowledge about the system dynamics and without requiring measurements of the state derivatives. However, the inherently discrete nature of their controller prevents the development of stability proof of the closed-loop system. Vamvoudakis and Lewis [36] proposed an online algorithm based on PI algorithm with guaranteed closed-loop stability for CT systems with completely known dynamics. Inspired by the work in [36], Dierks and Jagannathan [38] presented a single online approximator-based optimal scheme with guaranteed stability. Moreover, motivated by the work of [36], Bhasin et al. [37] presented an online PI algorithm where the requirement of knowing the system drift dynamics was eliminated by employing a NN to identify the drift dynamics. Although efficient, none of these online PI algorithms takes into account the input constraints caused by actuator saturation.

The control of systems subject to input constraints is of increasing importance, since almost all actuators in real-world applications are subject to saturation. In fact, control design methods that ignore the constraints on the magnitude of the control inputs may lead to performance degradation and even system instability. Hence, during the control development, due attention must be paid to the constraints which the control signals must comply with. This issue is of more importance when one designs an online learning control method, because instability may easily occur as a result of continuing online adaptation and learning during input saturation. This motivates our research into incorporating the actuator saturation limits when designing a PI algorithm for optimal control of CT systems.

This paper is concerned with developing an online optimal control method for CT systems in the presence of constraints on the input amplitude. To deal with actuator saturation, a suitable nonquadratic functional is used to encode the constraints into the optimization formulation. Then, a PI algorithm on an actor–critic structure is developed to solve the associated HJB equation online. That is, the optimal control law and the optimal value function are approximated as the output of two NNs, namely an actor NN and a critic NN. The problem of solving the HJB equation is then converted to simultaneously adjusting the weights of these two NNs. Given an arbitrarily nonoptimal control policy by the action network, the critic network guides the action network toward the optimal solution by successive adaptation. The closed-loop stability of the overall system and boundedness of the actor and critic NNs weights are assured by using Lyapunov theory. To our knowledge, this is the first treatment in which the input constraints are considered during the design of an online PI learning algorithm for solving the optimal control problem. Note that, although in [39] the authors presented an actor–critic algorithm for control of discrete-time systems with input constraints, their method does not converge to the optimal feedback control solution for a user-defined cost function, as it only minimizes a norm of the output error.

This paper is organized as follows. In the next section, some notations and definitions are given. An overview of optimal control for CT systems with input constraints is given in Section 3. This requires preliminary offline design. The development and implementation of the proposed online PI algorithm is presented in Section 4. Sections 5 and 6 present simulation results and conclusion, respectively.

2. Preliminaries

2.1. Notations and definitions

Throughout the paper, $\Re$ denotes the real numbers, $\Re^n$ denotes the real n vectors, $\Re^{m\times n}$ denotes the real $m\times n$ matrices, I denotes the identity matrix with appropriate dimension, for a scalar v, $|v|$ denotes the absolute value of v, for a vector x, $\|x\|$ indicates the Euclidean norm of x, for a matrix M, $\|M\|$ indicates the induced 2-norm of M, and $tr\{M\}$ denotes the trace of the matrix M. With $sgn\ (z)$ we denote the sign function defined as follows

$$sgn\ (z)=\begin{cases}1 & z\geq 0\\ -1 & z<0\end{cases}$$

Finally, we write $(.)^T$ to denote transpose and $\lambda_{\min}(.)$ to denote the minimum eigenvalue of a Hermitian matrix.

Lemma 1. (Young's inequality) *[40]*: *For any two vectors x and y, it holds that*

$$x^T y\leq \frac{\|x\|^2}{2}+\frac{\|y\|^2}{2}. \tag{1}$$

Definition 1. (Uniformly ultimately bounded (UUB) stability) [41]: Consider the nonlinear system (2)

$$\dot{x}=f(x,t) \tag{2}$$

with state $x(t)\in\Re^n$. The equilibrium point x_e is said to be UUB if there exists a compact set $\Omega\subset\Re^n$ so that for all $x_0\subset\Omega$, there exists a bound B and a time $T(B,x_0)$ such that $\|x(t)-x_0\|\leq B$ for all $t\geq t_0+T$. That is, after a transition period T, the state remains within the ball of radius B around x_0.

Definition 2. (Exponential stability) [42]: The equilibrium state x_e of the system (2) is exponentially stable if there exists an $\eta>0$, and for every $\varepsilon>0$ there exists a $\delta(\varepsilon)>0$ such that $|x(t;t_0,x_0)-x_e|\leq\varepsilon\, e^{-\eta(t-t_0)}$ for all $t>t_0$, whenever $|x_0-x_e|<\delta(\varepsilon)$.

Definition 3. (Zero-state observability) [41]: System (2) with measured output $y=h(x)$ is zero-state observable if $y(t)\equiv 0\ \forall t\geq 0$ implies that $x(t)\equiv 0\ \forall t\geq 0$.

Definition 4. (Persistently exciting (PE) signal) [42]: The bounded vector signal $z(t)$ is PE over the interval $[t,\ t+T_1]$ if there exists $T_1>0$, $\gamma_1>0$, and $\gamma_2>0$, such that for all t.

$$\gamma_1 I\leq \int_t^{t+T_1} z(\tau)\, z^T(\tau)\, d\tau\leq\gamma_2\, I \tag{3}$$

Definition 5. (Lipschitz) [42]: A function $f:[a,b]\to R$ is Lipschitz on $[a,b]$ if $|f(x_1)-f(x_2)|\leq k|x_1-x_2|$ for all $x_1,x_2\in[a,b]$, where $k\geq 0$ is a constant.

2.2. Function approximation by neural networks

The NN universal approximation property indicates that any continuous function $f(x)$ can be approximated arbitrary closely using a two-layer NN with appropriate weights on a compact set.