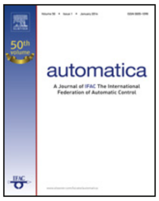
Automatica 50 (2014) 193–202

Contents lists available at ScienceDirect

Automatica

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



Brief paper

Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems[☆]

Hamidreza Modares[a,1], Frank L. Lewis[b], Mohammad-Bagher Naghibi-Sistani[a]

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

ARTICLE INFO

Article history:
Received 21 November 2012
Received in revised form
18 September 2013
Accepted 22 September 2013
Available online 24 October 2013

Keywords:
Integral reinforcement learning
Experience replay
Optimal control
Neural networks
Input constraints

ABSTRACT

In this paper, an integral reinforcement learning (IRL) algorithm on an actor–critic structure is developed to learn online the solution to the Hamilton–Jacobi–Bellman equation for partially-unknown constrained-input systems. The technique of experience replay is used to update the critic weights to solve an IRL Bellman equation. This means, unlike existing reinforcement learning algorithms, recorded past experiences are used concurrently with current data for adaptation of the critic weights. It is shown that using this technique, instead of the traditional persistence of excitation condition which is often difficult or impossible to verify online, an easy-to-check condition on the richness of the recorded data is sufficient to guarantee convergence to a near-optimal control law. Stability of the proposed feedback control law is shown and the effectiveness of the proposed method is illustrated with simulation examples.

1. Introduction

In this paper, an integral reinforcement learning (IRL) algorithm is developed to find optimal control solutions online for partially-unknown continuous-time systems subject to input constraints. Moreover, the idea of the experience replay is used to learn optimal solutions more efficiently by using past experiences during learning. It is well known that optimal control solutions can be derived by solving the Hamilton–Jacobi–Bellman (HJB) equation. The HJB equation is nonlinear and extremely intractable to solve by analytical approaches (Lewis, Vrabie, & Syrmos, 2012); thus several approximate methods have been presented in the literature to address the optimal control solutions. Traditional approaches for approximating the HJB solution are normally offline and require complete knowledge of the system dynamics. In practical applications, however, it is often desirable to design controllers conducive to real-time implementation and able to handle modeling uncertainties.

Over the last few decades, reinforcement learning (RL) (Bertsekas & Tsitsiklis, 1996; Powell, 2007; Sutton & Barto, 1998) has been effectively used to design learning-based adaptive optimal controllers. Considerable research has been conducted for approximating the HJB solution for discrete-time systems using RL algorithms. However, few results are available for continuous-time systems (Abu-Khalaf & Lewis, 2005; Beard, 1995; Bhasin et al., 2012; Doya, 2000; Murray, Cox, Lendaris, & Saeks, 2002; Vamvoudakis & Lewis, 2010; Vamvoudakis, Vrabie, & Lewis, 2013; Vrabie & Lewis, 2009). A survey of RL-based feedback control designs is found in Lewis and Vrabie (2009), Lewis, Vrabie, and Vamvoudakis (2012).

Existing online RL-based algorithms for solving the HJB equation for continuous-time (CT) nonlinear systems either require complete knowledge of the system dynamics (Doya, 2000; Vamvoudakis & Lewis, 2010) or lack a rigorous stability analysis (Doya, 2000; Murray et al., 2002), except for Bhasin et al. (2012); Vamvoudakis et al. (2013). In Bhasin et al. (2012), a system identification procedure was used along with the RL to find the optimal control solution for partially-unknown systems. Vamvoudakis et al. (2013) presented an IRL-based algorithm for partially-unknown systems which does not require a system identification

[☆] This work was supported by NSF grant ECCS-1128050, NSF grant IIS-1208623, AFOSR EOARD Grant #13-3055, China NNSF grant 61120106011, and China Education Ministry Project 111 (No. B08015). The material in this paper was partially presented at AIAA Guidance, Navigation, and Control Conference (GNC), August 19–22, 2013, Boston, Massachusetts, USA. This paper was recommended for publication in revised form by Associate Editor Kok Lay Teo under the direction of Editor Ian R. Petersen.

E-mail addresses: modares@uta.edu (H. Modares), lewis@uta.edu (F.L. Lewis), naghig@yahoo.com (M.-B. Naghibi-Sistani).

[1] Tel.: +98 9155624456; fax: +98 5118172233.

0005-1098/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.automatica.2013.09.043

procedure. In fact, the IRL algorithm (Vrabie, Pastravanu, Abu-Khalaf, & Lewis, 2009; Vrabie & Lewis, 2009) allows development of a Bellman equation in such a way that does not contain the system dynamics.

The existing mentioned RL-based control design methods did not take into account the input constraints caused by actuator saturation. However, failure to account for actuator saturation often severely destroys the system performance, or may even lead to instability. In our recent work (Modares, Lewis, & Naghibi-Sistani, 2012), we presented an online algorithm to solve the H_∞ control problem for constrained-input systems. Nevertheless, it requires complete knowledge of the system dynamics.

Another problem related to the existing online RL-based control design methods is that to guarantee convergence to a near-optimal control solution, a persistence of excitation (PE) condition is required to be satisfied. However, traditional PE conditions are often difficult or impossible to check online. Also, due to the requirement for the PE condition, the existing RL-based algorithms for CT systems are sample inefficient, that is, they require many samples from the real world in order to learn the optimal policy.

In order to reduce sample complexity and use available data more effectively, the experience replay technique has been proposed in the context of RL for discrete-time systems (Adam, Busoniu, & Babuska, 2012; Dung, Komeda, & Takagi, 2008; Kalyanakrishnan & Stone, 2007; Lin, 1992; Wawrzynski, 2009; Xu, Jagannathan, & Lewis, 2012), without providing a proof of convergence and stability. In this technique, a number of recent samples are stored in a database and they are presented repeatedly to the underlying RL algorithm. A related idea called concurrent learning was introduced in Chowdhary (2010) and Chowdhary and Johnson (2010) for adaptive control of uncertain systems. They showed that the concurrent use of recorded and current data can lead to the stability of a model reference adaptive controller as long as the recorded data is sufficiently rich. However, their results were focused on direct adaptive control, and in particular, that work did not establish any optimality guarantees on the closed-loop system.

The contributions of this paper are introducing for the first time the use of the experience replay to the IRL (Vrabie & Lewis, 2009) algorithms and incorporating the actuator limitations into control design. Specifically, in a first contribution, our proposed IRL algorithm takes into account the input constraints caused by actuator saturation, in contrast to the existing IRL algorithms (Vamvoudakis et al., 2013; Vrabie & Lewis, 2009) and other online RL-based algorithms for uncertain CT systems. Second, it is shown that the experience replay provides simplified conditions to check for PE-like requirements in real time by more efficient use of current and past data. The closed-loop stability of the overall system is ensured by using the Lyapunov theory. Simulations show that using the experience replay in the critic weights' tuning law significantly speeds up the convergence.

This paper is organized as follows. The next section provides an overview of the optimal control for CT systems with input constraints. Section 3 presents an offline IRL algorithm for solving the optimal control problem. The proposed online IRL algorithm with experience replay is presented in Section 4. Sections 5 and 6 present simulation results and conclusion, respectively.

2. Optimal control problem for systems with input constraints

In this section, the optimal control problem for CT systems with input constraints is formulated.

Let the system dynamics be described by

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

where $x \in \mathbb{R}^n$ is the system state vector, $f(x) \in \mathbb{R}^n$ is the drift dynamics of the system, $g(x) \in \mathbb{R}^{n\times m}$ is the input dynamics of the system and $u(t) \in \mathbb{R}^m$ is the control input. We denote $\Omega_u = \{u | u \in \mathbb{R}^m,\ |u_i(t)| \leqslant \lambda,\ i = 1,\dots,m\}$ as the set of all inputs satisfying the input constraints, where λ is the saturating bound. It is assumed that $f(x) + g(x)\,u$ is Lipschitz and the system (1) is stabilizable.

The problem of interest in this paper is to find an optimal constrained policy u^* that drives the state of the system (1) to the origin, by minimizing a performance index as a function of state and control variables. The performance index is defined as

$$V(x(t)) = \int_t^\infty Q(x(\tau)) + U(u(\tau))\,d\tau \tag{2}$$

where $Q(x)$ is a positive definite monotonically increasing function and $U(u)$ is a positive definite integrand function.

Assumption 1. The performance index (2) is zero-state observable (Lewis, Jagannathan, & Yesildirek, 1999).

Definition 1 (*Beard, 1995*)**.** A control policy $u(t) = \mu(x(t)) \equiv \mu(x)$ is said to be admissible with respect to (2) on Ω, defined by $\mu \in \pi(\Omega)$, if $\mu(x)$ is continuous on Ω, $\mu(0) = 0$, $u(t) = \mu(x)$ stabilizes (1) on Ω, and $V(x_0)$ is finite $\forall x_0 \in \Omega$.

To deal with input constraints, the following generalized non-quadratic cost function $U(u)$ is employed in the literature (Abu-Khalaf & Lewis, 2005; Lyshevski, 1998):

$$U(u) = 2\int_0^u (\lambda\,\beta^{-1}(v/\lambda))^T R\,dv \tag{3}$$

where $v \in \mathbb{R}^m$, $\beta(.) = \tanh(.)$, and $R = diag(r_1,\dots,r_m) > 0$ is assumed to be diagonal for simplicity of analysis. Denote $\omega(v) = (\lambda\beta^{-1}(v/\lambda))^T R = [\omega_1(v_1)\dots\omega_m(v_m)]$. Then the integral in (3) is defined as

$$U(u) = 2\int_0^u \omega(v)\,dv = 2\sum_{i=1}^m \int_0^{u_i} \omega_i(v_i)\,dv_i. \tag{4}$$

It is clear that $U(u)$ in (3) is a scalar for $u \in \mathbb{R}^m$. Using (3) in (2), the performance index becomes

$$V(x(t)) = \int_t^\infty \left(Q(x(\tau)) + 2\int_0^u (\lambda\ \tanh^{-1}(v/\lambda))^T R\,dv \right) d\tau. \tag{5}$$

By differentiating V along the system trajectories, the following Bellman equation is given:

$$Q(x) + 2\int_0^u (\lambda\ \tanh^{-1}(v/\lambda))^T R\,dv$$
$$+ \nabla V^T(x)\,(f(x) + g(x)\,u) = 0, \quad V(0) = 0 \tag{6}$$

where $\nabla V(x) = \partial V(x)/\partial x \in \mathbb{R}^n$. Let $V^*(x)$ be the optimal value function. Then, it satisfies the Hamilton–Jacobi–Bellman (HJB) equation (Abu-Khalaf & Lewis, 2005)

$$\min_{\substack{u(\tau)\in\pi(\Omega)\\ t\leqslant\tau<\infty}} \Bigg[Q(x) + 2\int_0^u (\lambda\ \tanh^{-1}(v/\lambda))^T R\,dv$$
$$+ \nabla V^{*T}(x)\,(f(x) + g(x)\,u) \Bigg] = 0. \tag{7}$$

The optimal control input is obtained by differentiating the HJB equation (7) with respect to the control u. The result is

$$u^* = -\lambda\ \tanh\left((1/2\lambda)R^{-1}g^T(x)\ \nabla V^*(x)\right). \tag{8}$$

Using (8) in (3) yields

$$U(u^*) = \lambda\nabla V^{*T}(x)g(x)\tanh(D^*) + \lambda^2\,\bar{R}\ln(\underline{\mathbf{1}} - \tanh^2(D^*)) \tag{9}$$

where $D = (1/2\lambda)\,R^{-1}g^T\nabla V^*(x) \in \mathbb{R}^m$, $\underline{\mathbf{1}}$ is a column vector having all of its elements equal to one, and $\bar{R} = [r_1,\dots,r_m] \in \mathbb{R}^{1\times m}$.