Neural Networks 71 (2015) 150–158

Contents lists available at ScienceDirect

Neural Networks

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

Reinforcement learning solution for HJB equation arising in constrained optimal control problem

Biao Luo [a], Huai-Ning Wu [b], Tingwen Huang [c], Derong Liu [d,*]

[a] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[b] Science and Technology on Aircraft Control Laboratory, Beihang University (Beijing University of Aeronautics and Astronautics), Beijing 100191, China
[c] Texas A&M University at Qatar, PO Box 23874, Doha, Qatar
[d] School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China

ARTICLE INFO

Article history:
Received 4 January 2015
Received in revised form 15 August 2015
Accepted 16 August 2015
Available online 24 August 2015

Keywords:
Constrained optimal control
Data-based
Off-policy reinforcement learning
Hamilton–Jacobi–Bellman equation
The method of weighted residuals

ABSTRACT

The constrained optimal control problem depends on the solution of the complicated Hamilton–Jacobi–Bellman equation (HJBE). In this paper, a data-based off-policy reinforcement learning (RL) method is proposed, which learns the solution of the HJBE and the optimal control policy from real system data. One important feature of the off-policy RL is that its policy evaluation can be realized with data generated by other behavior policies, not necessarily the target policy, which solves the insufficient exploration problem. The convergence of the off-policy RL is proved by demonstrating its equivalence to the successive approximation approach. Its implementation procedure is based on the actor–critic neural networks structure, where the function approximation is conducted with linearly independent basis functions. Subsequently, the convergence of the implementation procedure with function approximation is also proved. Finally, its effectiveness is verified through computer simulations.

1. Introduction

Optimal control is an important part of control theory, which has been widely investigated over the past several decades. The bottleneck of its applications to nonlinear systems is that it depends on the solution of the Hamilton–Jacobi–Bellman equation (HJBE) (Bertsekas, 2005; Hull, 2003; Lewis, Vrabie, & Syrmos, 2013), which is extremely difficult to obtain analytically. Over the past few years, reinforcement learning (RL) (Lendaris, 2009; Powell, 2007; Precup, Sutton, & Dasgupta, 2001; Sutton & Barto, 1998), has appeared as an efficient tool to solve the HJBE and many meaningful results (Faust, Ruymgaart, Salman, Fierro, & Tapia, 2014; Jiang & Jiang, 2012; Lee, Park, & Choi, 2012; Liu, Wang, & Li, 2014; Liu & Wei, 2014; Luo, Wu, Huang, & Liu, 2014; Modares & Lewis, 2014; Murray, Cox, Lendaris, & Saeks, 2002; Vamvoudakis & Lewis, 2010; Vrabie & Lewis, 2009; Vrabie, Pastravanu, Abu-Khalaf, & Lewis, 2009; Wang, Liu, & Li, 2014; Wei & Liu, 2012; Yang, Liu, & Wang, 2014; Yang, Liu, Wang, & Wei, 2014; Zhao, Xu, & Jagannathan, 2014) have been reported. For example, appropriate estimators were employed for approximating value function such that the temporal difference error is minimized (Doya, 2000). Murray et al. (2002) suggested two policy iteration algorithms that avoid the necessity of knowing the internal system dynamics. Vrabie et al. (2009) extended their result and proposed a new policy iteration algorithm to solve the linear quadratic regulation problem online along a single state trajectory. A nonlinear version of this algorithm was presented in Vrabie and Lewis (2009) by using neural network (NN) approximator. Vamvoudakis and Lewis (2010) also gave a so-called synchronous policy iteration algorithm which tunes synchronously the weight parameters of both NNs in the actor–critic structure. An integral reinforcement learning (IRL) method (Modares & Lewis, 2014) was introduced to solve the linear quadratic tracking problem of partially-unknown continuous-time systems. Online adaptive optimal control (Jiang & Jiang, 2012) and Q-learning (Lee et al., 2012) algorithms were developed for linear quadratic regulator problem. Off-policy RL approaches were proposed to solve the nonlinear data-based optimal control problem (Luo et al., 2014) and partially model-free H_∞ control problem (Luo, Wu, & Huang, 2015). However, it is noted that control constraints are not involved in these results.

In practice, constraints widely exist in real control systems and have damaging effects on the system performance, and thus should

* Corresponding author.
E-mail addresses: biao.luo@hotmail.com (B. Luo), whn@buaa.edu.cn (H.-N. Wu), tingwen.huang@qatar.tamu.edu (T. Huang), derong@ustb.edu.cn (D. Liu).

http://dx.doi.org/10.1016/j.neunet.2015.08.007
0893-6080/

be accounted for during the controller design process. For the constrained optimal control problem, several results (Abu-Khalaf & Lewis, 2005; He & Jagannathan, 2005, 2007; Heydari & Balakrishnan, 2013; Liu, Wang, & Yang, 2013; Lyshevski, 1998; Modares, Lewis, & Naghibi-Sistani, 2013; Zhang, Luo, & Liu, 2009) have been reported recently. A nonquadratic cost functional was introduced by Lyshevski (1998) to confront input constraints, and then the associated HJBE was reformulated accordingly. As the extensions of the method in Saridis and Lee (1979) and Beard, Saridis, and Wen (1997) to handle constrained optimal control problem, model-based successive approximation method was used for solving the HJBE of continuous-time systems (Abu-Khalaf & Lewis, 2005) and discrete-time systems (Chen & Jagannathan, 2008). Modares, Lewis, and Naghibi-Sistani (2014) developed an experience-replay based IRL algorithm for nonlinear partially unknown constrained-input systems. A heuristic dynamic programming was used to solve the constrained optimal control problem of nonlinear discrete-time systems (Zhang et al., 2009). The single network based adaptive critics method was proposed for finite-horizon nonlinear constrained optimal control design (Heydari & Balakrishnan, 2013). However, the data-based constrained nonlinear optimal control problem is rarely studied with off-policy RL and still remains an open issue.

In this paper, a data-based off-policy RL method is proposed for learning the constrained optimal control policy form real system data instead of using mathematical model. The rest of the paper is arranged as follows. Section 2 gives the problem description and Section 3 presents a model-based successive approximation method. The data-based off-policy RL method is developed in Section 4. Section 5 shows the simulation results and Section 6 gives the conclusions.

Notation: $\mathbb{R}$ and $\mathbb{R}^n$ are the set of real numbers and the n-dimensional Euclidean space, respectively. $\|\cdot\|$ denotes the vector norm or matrix norm in $\mathbb{R}^n$. The superscript T is used for the transpose and I denotes the identify matrix of appropriate dimension. $\bigtriangledown \triangleq \partial/\partial x$ denotes a gradient operator. $C^1(\mathcal{X})$ is a function space on $\mathcal{X}$ with continuous first derivatives. Let $\mathcal{X}$ and $\mathcal{U}$ be compact sets, denote $\mathcal{D} \triangleq \{(x,u)|x \in \mathcal{X}, u \in \mathcal{U}\}$. For column vector functions $s_1(x,u)$ and $s_2(x,u)$, where $(x,u) \in \mathcal{D}$, define inner product $\langle s_1(x,u), s_2(x,u)\rangle_{\mathcal{D}} \triangleq \int_{\mathcal{D}} s_1^{\mathrm{T}}(x,u)s_2(x,u)\mathrm{d}(x,u)$ and norm $\|s_1(x,u)\|_{\mathcal{D}} \triangleq \langle s_1(x,u), s_2(x,u)\rangle_{\mathcal{D}}^{1/2}$.

2. Problem description

Let us consider the following continuous-time nonlinear system:

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

where $x = [x_1, \ldots, x_n]^{\mathrm{T}} \in \mathbb{R}^n$ is the state, x_0 is the initial state and $u = [u_1, \ldots, u_m]^{\mathrm{T}} \in \mathbb{R}^m$ is the control input constrained by $|u_i| \leqslant \beta$, $\beta > 0$. Assume that $f(x) + g(x)u(x)$ is Lipschitz continuous on $\mathcal{X}$ that contains the origin, $f(0) = 0$, and the system is stabilizable on $\mathcal{X}$, i.e., there exists a continuous control function $u(x)$ such that the system is asymptotically stable. $f(x)$ and $g(x)$ are continuous *unknown* vector or matrix functions of appropriate dimensions.

The optimal control problem under consideration is to learn a state feedback control law $u(t) = u(x(t))$ from real system data, such that the system (1) is closed-loop asymptotically stable, and minimize the following generalized infinite horizon cost functional:

$$V(x_0) \triangleq \int_0^{+\infty} (Q(x(t)) + W(u(t)))\mathrm{d}t, \tag{2}$$

where $Q(x)$ and $W(u)$ are positive definite functions, i.e., for $\forall x \neq 0$, $u \neq 0$, $Q(x) > 0$, $W(u) > 0$, and $Q(x) = 0$, $W(u) = 0$ only when $x = 0$, $u = 0$. Then, the optimal control problem is briefly presented as

$$u(t) = u^*(x) \triangleq \arg\min_u V(x_0). \tag{3}$$

3. Model-based successive approximation method

For the model-based optimal control problem (3), i.e., the mathematical models of $f(x)$ and $g(x)$ are completely known, it can be converted to solving the HJBE. In Abu-Khalaf and Lewis (2005), a model-based successive approximation method was given for solving the HJBE, where the HJBE is successively approximated by a sequence of linear partial differential equations. Before we start, the definition of admissible control (Abu-Khalaf & Lewis, 2005; Beard et al., 1997) is given.

Definition 1 (*Admissible Control*)**.** For the given system (1), $x \in \mathcal{X}$, a control policy $u(x)$ is defined to be admissible with respect to the cost function (2) on $\mathcal{X}$, denoted by $u(x) \in \mathfrak{U}(\mathcal{X})$, if, (1) u is continuous on $\mathcal{X}$, (2) $u(0) = 0$, (3) $u(x)$ stabilizes the system, and (4) $V(x) < \infty, \forall x \in \mathcal{X}$. □

For $\forall u(x) \in \mathfrak{U}(\mathcal{X})$, its value function $V(x)$ of (2) satisfies the following linear partial differential equation (Abu-Khalaf & Lewis, 2005):

$$[\nabla V(x)]^{\mathrm{T}}(f(x) + g(x)u(x)) + Q(x) + W(u) = 0, \tag{4}$$

where $V(x) \in C^1(\mathcal{X})$, $V(x) \geq 0$ and $V(0) = 0$. From the optimal control theory (Anderson & Moore, 2007; Bertsekas, 2005; Lewis et al., 2013), if using the optimal control $u^*(x)$, the Eq. (4) results in the HJBE

$$[\nabla V^*(x)]^{\mathrm{T}}(f(x) + g(x)u^*(x)) + Q(x) + W(u^*) = 0. \tag{5}$$

For the system (1) with input constraints $|u_i| \leqslant \beta$, the following nonquadratic form $W(u)$ for the cost functional (2) can be used (Abu-Khalaf & Lewis, 2005; Lyashevskiy, 1996; Lyshevski, 1998; Modares et al., 2013):

$$W(u) = 2\sum_{l=1}^{m} r_l \int_0^{u_l} \varphi^{-1}(\mu_l)\mathrm{d}\mu_l, \tag{6}$$

where $\mu \in \mathbb{R}^m$, $r_l > 0$ and $\varphi(\cdot)$ is a continuous one-to-one bounded function satisfying $|\varphi(\cdot)| \leqslant \beta$ with $\varphi(0) = 0$. Moreover, $\varphi(\cdot)$ is a monotonic odd function and its derivative is bounded. An example of $\varphi(\cdot)$ is the hyperbolic tangent $\tanh(\cdot)$. Denoting $R = \mathrm{diag}(r_1, \ldots, r_m)$, it follows from Abu-Khalaf and Lewis (2005) and Lyshevski (1998) that the HJBE (5) of the constrained optimal control problem is given by

$$[\nabla V^*]^{\mathrm{T}}\left(f - g\varphi\left(\frac{1}{2}R^{-1}g^{\mathrm{T}}\nabla V^*\right)\right) + Q(x) + W\left(-\varphi\left(\frac{1}{2}R^{-1}g^{\mathrm{T}}\nabla V^*\right)\right) = 0. \tag{7}$$

By solving the HJBE for $V^*(x)$, the optimal control policy is obtained as:

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

For simplicity of description, define

$$\nu^*(x) \triangleq -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla V^*(x). \tag{9}$$