Available online at www.sciencedirect.com

ScienceDirect

Journal of the Franklin Institute 355 (2018) 2610–2630

www.elsevier.com/locate/jfranklin

General value iteration based single network approach for constrained optimal controller design of partially-unknown continuous-time nonlinear systems

Geyang Xiao[a,b], Huaguang Zhang[a,b,*], Qiuxia Qu[a,b], He Jiang[a,b]

[a] *College of Information Science and Engineering, Northeastern University, Box 134, Shenyang 110819, PR China*

[b] *The Key Laboratory of Integrated Automation of Process Industry (Northeastern University) of the National Education Ministry, Shenyang, 110004, PR China*

Received 5 January 2017; received in revised form 5 November 2017; accepted 6 February 2018

Available online 10 February 2018

Abstract

In this paper, a novel iterative approximate dynamic programming scheme is proposed by introducing the learning mechanism of value iteration (VI) to solve the constrained optimal control problem for CT affine nonlinear systems with utilizing only one neural network. The idea is to show the feasibility of introducing the VI learning mechanism to solve for the constrained optimal control problem from a theoretical point of view, and thus the initial admissible control can be avoided compared with most existing works based on policy iteration (PI). Meanwhile, the initial condition of the proposed VI based method can be more general than the traditional VI method which requires the initial value function to be a zero function. A general analytical method is proposed to demonstrate the convergence property. To simplify the architecture, only one critic neural network is adopted to approximate the iterative value function while implementing the proposed method. At last, two simulation examples are proposed to validate the theoretical results.

* Corresponding author at: College of Information Science and Engineering, Northeastern University, Box 134, Shenyang 110819, PR China.

E-mail addresses: xgyalan@outlook.com (G. Xiao), hgzhang@ieee.org (H. Zhang), quqiuxia2010@163.com (Q. Qu), jianghescholar@163.com (H. Jiang).

https://doi.org/10.1016/j.jfranklin.2018.02.001

1. Introduction

Over the last several decades, optimal control has been widely investigated and becomes an important part of control theory. The goal of optimal control is to find a control law for a given system and a predefined optimality long-term performance criterion. The bottleneck of the optimal control arising in nonlinear systems is that it requires to solve the Hamilton–Jacobi–Bellman (HJB) equation, which is a nonlinear partial differential equation (PDE) and difficult to get the analytical solution [1,2]. On the other hand, in industrial applications, the control signals (such as currents, flows, voltages and torques) [3] are always subject to some constraint conditions. Ignoring the limitation of the actuators may lead to deterioration of the system performance or even cause system instability. Thus, the control constraints should be taken into consideration while designing the optimal controller.

Similar to optimal control, reinforcement learning (RL) [3–6], which attempts to imitate the learning mechanisms observed in mammals, is concerned with how the actor to modify its action based on the response from the interacted environment so as to learn an optimized performance criterion. The similarity and connection between optimal control and RL have prompted a major effort towards introducing and developing RL-based techniques to seek the optimal control for nonlinear systems [7–14]. [15–17] proposed an actor-critic structure based on the idea of RL to solve the HJB equation approximatively. Two neural networks (NN) are involved in the structure where one critic NN is trained to approximate the performance function and one actor NN is trained to approximate the optimized control policy. The corresponding methodology is called approximate dynamic programming (ADP), or adaptive dynamic programming as well, and attracted lots of attention [18–25].

Generally, there are two classical types in the ADP framework [15]: policy iteration (PI) and value iteration (VI). One major difference between PI and VI methods is that PI requires an initial admissible control policy which stabilizes the system states. If an admissible control policy is available, the PI based methods can be utilized to obtain the optimal solution in both online manner and offline manner, because the admissibility of the initial condition ensures the derived iterative control policies to be admissible [26] and thus the system states will not diverge from the origin. However, from a mathematical point of view, the initial admissible control can be regarded as a suboptimal control which requires to solve the nonlinear PDEs analytically and intractable to derive. Accordingly, the admissibility of the initial control policy is a restrictive condition actually [27–29]. To overcome the defect, the VI-based methods can be utilized free from the initial admissible assumption [30].

Since the ADP is a feasible way to seek for optimal solution for nonlinear systems, some researchers introduced this RL based technique for the optimal control problem with control constraints [26,29,31–41]. For the continuous-time (CT) nonlinear systems, most works were proposed based on PI while introducing the ADP framework. In [26], the authors introduced the PI method to design a constrained optimal controller for CT nonlinear systems in an off-line manner, and developed the work for H_∞ control problem of constrained control CT affine nonlinear systems in [31]. In [32], the H_∞ control problem was studied by initializing with an admissible control for constrained inputs CT nonlinear systems with finite-horizon. A PI based approach called IRL was proposed in [33] to solve the optimal control for constrained control CT nonlinear systems with partial unknown dynamics, and the method was developed for optimal tracking control problem in [34]. In [35], a PI based approach called off-policy was proposed to solve the optimal control for constrained control CT nonlinear systems, which can be trained by data generated by arbitrary behavior policies. The