Automatica 78 (2017) 79–87

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Data-driven approximate value iteration with optimality error bound analysis☆

Yongqiang Li [a], Zhongsheng Hou [b], Yuanjing Feng [a], Ronghu Chi [c]

[a] College of Information Engineering, Zhejiang University of Technology, Hangzhou, China
[b] Advanced Control Systems Lab, School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China
[c] School Automation and Electronic Engineering, Qingdao University Science and Technology, Qingdao, China

ARTICLE INFO

Article history:
Received 21 September 2015
Received in revised form
30 October 2016
Accepted 16 November 2016

Keywords:
Data-driven control
Approximate dynamic programming
Domain of attraction
Asymptotic stabilization

ABSTRACT

Features of the data-driven approximate value iteration (AVI) algorithm, proposed in Li et al. (2014) for dealing with the optimal stabilization problem, include that only process data is required and that the estimate of the domain of attraction for the closed-loop is enlarged. However, the controller generated by the data-driven AVI algorithm is an approximate solution for the optimal control problem. In this work, a quantitative analysis result on the error bound between the optimal cost and the cost under the designed controller is given. This error bound is determined by the approximation error of the estimation for the optimal cost and the approximation error of the controller function estimator. The first one is concretely determined by the approximation error of the data-driven dynamic programming (DP) operator to the DP operator and the approximation error of the value function estimator. These three approximation errors are zeros when the data set of the plant is sufficient and infinitely complete, and the number of samples in the interested state space is infinite. This means that the cost under the designed controller equals to the optimal cost when the number of iterations is infinite.

1. Introduction

Optimal control is a method for finding a controller for a dynamic system such that a given cost is as optimal as possible. DP, based on optimality principle (Bellman, 1957), is a useful tool of solving the optimal control problem. However, DP suffers from the accurate modeling and the curse of dimensionality. To overcome this problem, the idea of approximate dynamic programming (ADP) was proposed by Werbos in 1968 firstly (Werbos, 1968) and received more and more attentions in recent decades (Balakrishnan, Ding, & Lewis, 2008; Jiang & Jiang, 2013; Liu, Li, & Wang, 2015; Wang, Zhang, & Liu, 2009).

Recursive DP algorithms can be divided into two classes: policy iteration (PI) and value iteration (VI) (Bertsekas, 2001). This paper focuses on VI. VI derived directly from the property of DP operator, that is, the optimal cost function is obtained after successively using the DP operator on a function over an infinite number of times. However, for plants with infinite state space and control input space, VI may be implementable only through approximations, which lead the development of approximate VI (AVI) (Al-Tamimi, Lewis, & Abu-Khalaf, 2007; Liu, Wang, & Yang, 2013; Liu & Wei, 2013). Instead of updating a value function for all states, it can be done only for some states and estimate the updated value function for the remaining states by a function estimator.

In practice, numerous plants are difficult to be modeled accurately. Hence, the data-driven control approaches in which controllers are designed directly from data and the modeling step is bypassed, has shown great promise and thus recently undergone extensive research (Hou & Jin, 2011a,b; Hou & Wang, 2013). However, most ADP algorithms require prior plant knowledge/model. Thus, the wide applicability of ADP necessitates the development of data-driven ADP algorithms, which only require the plant data instead of the plant knowledge/model. For discrete-time systems, several data-driven ADP algorithms, based on Q value function (*i.e.*, action-dependent value function), have been proposed (Al-Tamimi et al., 2007; He & Jagannathan, 2007; Xu, Hou, Lian, & He, 2013).

☆ This work is partially supported by National Natural Science Foundation of China (61433002, 61120106009, and 61379020), Natural Science Foundation of Zhejiang Province (LQ16F030009) and Taishan Scholar program of Shandong Province of China. The material in this paper was presented at the 53rd IEEE Conference on Decision and Control, December 15–17, 2014, Los Angeles, CA, USA. This paper was recommended for publication in revised form by Associate Editor Andrey V. Savkin under the direction of Editor Ian R. Petersen.

E-mail addresses: yqli@zjut.edu.cn (Y. Li), zhshhou@bjtu.edu.cn (Z. Hou), fyjing@zjut.edu.cn (Y. Feng), ronghu_chi@hotmail.com (R. Chi).

http://dx.doi.org/10.1016/j.automatica.2016.12.019
0005-1098/

ADP methods are widely applied for two classes of plants: Markov decision process (MDP) and dynamical systems described by ordinary differential/difference equations. Given that the state spaces considered in the MDP are finite or countable, the stability issue is usually overlooked. For dynamical systems, stability must be considered in the context of ADP while feedback control problems are studied (Lewis & Vrabie, 2009). As discussed in Balakrishnan et al. (2008), controllers derived from model-based ADP algorithms assure stability of closed-loops because they are optimal controllers essentially. However, the analysis of closed-loop stability for data-driven ADP methods is quite different from that of the model-based framework and considerably difficult. A few results are published (Al-Tamimi et al., 2007; He & Jagannathan, 2007; Liu, Sun, Si, Guo, & Mei, 2012; Sokolov, Kozma, Werbos, & Werbos, 2015). For linear systems or affine nonlinear systems, stability results are available in Al-Tamimi et al. (2007) and He and Jagannathan (2007). For general nonlinear systems, stability results for action-dependent heuristic DP are provided in Liu et al. (2012) and Sokolov et al. (2015). However, such results do not describe the domain of attraction (DOA) for closed-loops.

For general nonlinear systems, the DOA, which is an invariant set that characterizes stabilizable areas around an equilibrium, requires extensive investigation because global stabilization is difficult to achieve. In Li and Hou (2014), under the assumption that plant model is unknown, a set of controllers for stabilizing plants is found directly from data for discrete time systems, and the estimate of the DOA for closed-loops is enlarged by selecting an appropriate Lyapunov function. On the basis of this result, Li, Hou, and Feng (2014) proposed a data-driven AVI algorithm to find a controller from the controller set to minimize the given cost. Because the plant model is unknown, the data-driven DP operator is proposed to replace the DP operator in AVI iterations in order to find an estimation of the optimal cost. Again because the plant model is unknown, a sub-optimal controller is estimated using the relevant data derived from the estimation of the optimal cost. Features of the data-driven AVI algorithm include that only data is required and that the estimate of the DOA for the closed loop is enlarged, but the controller is sub-optimal.

The contribution of this paper is that a quantitative analysis result on the error bound between the optimal cost and the cost under the sub-optimal controller is presented. This error bound is determined by two approximation errors: the first one is caused by using an estimation of the optimal cost and the second one is caused by using an estimation of a sub-optimal controller derived from the estimation of the optimal cost. The first one is concretely determined by the approximation error of the data-driven DP operator to the DP operator and the approximation error of the value function estimator. The approximation error bound of the data-driven DP operator to the DP operator is zero when the data set of the plant is sufficient and infinitely complete. To analyze the approximation error bound of the value function estimator, the value function estimator is selected as the one-output Gaussian processes regression (GPR) with noise-free training data. GPR can provide the standard deviation of the predictions to estimate the approximation error bound and can select the hyperparameters (including the noise level) according to the training data (Rasmussen & Williams, 2006). Because the training data is noise-free, the error bound of the value function estimator is zero when the number of the training data is infinite. In order to analyze the second error bound, the controller function estimator is also selected as the multi-output GPR with noisy training data. When the data set of the plant is sufficient and infinitely complete, the training data for the controller estimator is also noisy-free. Under this condition, the error bound of the controller function estimator is zero when the number of the training data is infinite. As shown by the main result, if these three approximation errors are zeros, the cost under the designed controller equals to the optimal cost when the number of iterations is infinite.

The rest of the paper is organized as follows. In Section 2, the control problem is formulated. In Section 3, the data-driven stabilization method and the data-driven AVI method are briefly introduced. In Section 4, theoretical analysis results about the optimality error bound are presented. In Section 5, simulation results are presented. Finally, in Section 6, the conclusion of the study is drawn.

Notation: $\mathbb{R}$ represents the set of real number. $\mathbb{Z}_+$ represents the set of positive integer number. $\mathbb{R}^n$ represents the set of real vectors with n elements. For a vector $x \in \mathbb{R}^n$, $\|x\|$ represents $\sqrt{x^T x}$ and $x_{(i)}$ represents the ith element of x, $i = 1, 2, \ldots, n$.

2. Problem formulation

Consider the nonlinear discrete-time system

$$x(k+1) = f(x(k), u(k)), \quad x(0) = x_0, \ k \in \overline{\mathbb{Z}}_+, \tag{1}$$

where $x(k) \in \mathbb{R}^n$ is the state, $u(k) \in \mathbb{R}^m$ is the control input and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is an unknown continuous function satisfying $0 = f(0, 0)$.

Although f is unknown, we have a point-wise data set

$$\Pi^d = \left\{ (x^d_{f,i}; u^d_i; x^d_i) \in \mathbb{X}^d_f \times \mathbb{U}^d \times \mathbb{X}^d \;\middle|\; x^d_{f,i} = f\left(x^d_i, u^d_i\right), i = 1, 2, \ldots, N_p \right\} \tag{2}$$

where $(x^d_{f,i}; u^d_i; x^d_i)$ is the ith data point in Π^d, N_p is the number of points in Π^d, $\mathbb{X}^d_f, \mathbb{X}^d \in \mathbb{R}^n$, and $\mathbb{U}^d \in \mathbb{R}^m$. The data set Π^d should contains adequate dynamic information of the plant, that is, Π^d is adequately sufficient and complete. For additional details, see Li and Hou (2014).

Using the data set Π^d, our control objective is to find a nonlinear feedback controller or stationary policy $\mu : \mathbb{R}^n \to \mathbb{R}^m$ such that the closed-loop $x(k+1) = f(x(k), \mu(x(k)))$ is asymptotically stable at $x = 0$ and that the estimate of the DOA for the closed-loop is as large as possible. Meanwhile, for all initial state x_0 in the estimate of the DOA, the infinite horizon discounted cost

$$V_\mu(x_0) = \sum_{k=0}^{\infty} \gamma^k g(x(k), \mu(x(k))) \tag{3}$$

is as small as possible, here the constant $0 < \gamma < 1$ is the discounted factor, and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is the instantaneous cost satisfying uniform boundedness, $g(0, 0) = 0$ and $g(x, u) > 0, \forall x \neq 0, u \neq 0$.

The first part of the above control objective, the asymptotic stabilization and the enlargement of the estimate of the DOA, was realized in our previous work (Li & Hou, 2014). Under the same framework, the second part of the control objective, minimizing the cost function (3) for all initial states in the estimate of the DOA, was partially realized in our another previous work (Li et al., 2014). In this study, we present the most significant theoretical analysis result of the optimal control method proposed in Li et al. (2014).

3. Background materials

3.1. Data-driven stabilization

In our previous work (Li & Hou, 2014), under the assumption that f in (1) is unknown, a state feedback asymptotic stabilization controller is designed directly from available data. By selecting an appropriate Lyapunov function, the estimate of the DOA for the closed-loop is enlarged. The following lemma, which presents sufficient conditions for a feedback controller asymptotically stabilizing plant (1) and an estimate of the DOA for the closed-loop, serves as the theoretical cornerstone of this method.