Automatica 71 (2016) 348–360

Contents lists available at ScienceDirect

Automatica

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

Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design☆

Tao Bian [1], Zhong-Ping Jiang

Control and Networks Lab, Department of Electrical and Computer Engineering, Tandon School of Engineering, New York University, 5 Metrotech Center, Brooklyn, NY 11201, USA

ARTICLE INFO

Article history:
Received 16 March 2015
Received in revised form
29 February 2016
Accepted 19 April 2016
Available online 22 June 2016

Keywords:
Value iteration
Adaptive dynamic programming
Optimal control
Adaptive control
Stochastic approximation

ABSTRACT

This paper presents a novel non-model-based, data-driven adaptive optimal controller design for linear continuous-time systems with completely unknown dynamics. Inspired by the stochastic approximation theory, a continuous-time version of the traditional value iteration (VI) algorithm is presented with rigorous convergence analysis. This VI method is crucial for developing new adaptive dynamic programming methods to solve the adaptive optimal control problem and the stochastic robust optimal control problem for linear continuous-time systems. Fundamentally different from existing results, the *a priori* knowledge of an initial admissible control policy is no longer required. The efficacy of the proposed methodology is illustrated by two examples and a brief comparative study between VI and earlier policy-iteration methods.

1. Introduction

Dynamic programming (DP) (Bellman, 1957) is an approach to solving optimal control problems for dynamic systems using Bellman's principle of optimality. However, the implementation of traditional DP methods in real-world applications is prohibited due to the "curse of dimensionality" (Bellman, 1961) and the "curse of modeling" (Bertsekas & Tsitsiklis, 1996). Approximate DP and neuro-DP were introduced to conquer these two shortcomings by approximating the value function and control policy via on-line learning. In the past few decades, numerous approximate DP methods have been developed to solve the optimal control problem for Markov decision processes (Barto, Sutton, & Anderson, 1983; Bertsekas, 2011; Sutton, 1988; Tsitsiklis, 1994; Watkins & Dayan, 1992). The interested reader can consult the nice tutorials by Bertsekas and Tsitsiklis (1996), Si, Barto, Powell, and Wunsch (2004) and Sutton and Barto (1998). Despite their popularity in real-world applications, the above-mentioned methods usually overlooked the stability issue of the system. Moreover, the underlying state space is assumed either finite or countable, which is not always applicable in many applications.

Different from the early approximate DP methods, a new method known as adaptive dynamic programming (ADP), or heuristic dynamic programming in some literature, is introduced recently to find stabilizing optimal controllers for continuous-state space control systems via online learning. Over the past decade, ADP methods for discrete-time systems have attracted considerable attention of many researchers; see, Heydari (2014), Lewis and Vamvoudakis (2011), Ni, He, Zhong, and Prokhorov (2015), Prokhorov and Wunsch (1997), Wang, Jin, Liu, and Wei (2011), Wang, Liu, Wei, Zhao, and Jin (2012), Zhang, Luo, and Liu (2009), and references therein. In parallel, the research in ADP for continuous-time systems is developed in Jiang and Jiang (2012, 2013), Lewis and Vrabie (2009), Lewis, Vrabie, and Vamvoudakis (2012), Murray, Cox, Lendaris, and Saeks (2002), Vrabie, Pastravanu, Abu-Khalaf, and Lewis (2009), Vrabie, Vamvoudakis, and Lewis (2013), Xu, Jagannathan, and Lewis (2012), and numerous references therein. For more recent work on continuous-time ADP, the interested reader may find more detailed references in Bian, Jiang, and Jiang (2014, 2015, in press), Jiang and Jiang (2014b), Li, Liu, and Wang (2014), Modares and Lewis (2014), Song et al. (2015) and Zargarzadeh, Dierks, and Jagannathan (2015).

Since existing continuous-time ADP methods are established based on policy iteration (PI) techniques (Beard, Saridis, & Wen, 1997; Kleinman, 1968; Leake & Liu, 1967), a common assumption is that a stabilizing control policy is known to start the learning

☆ This work was partially supported by the National Science Foundation under Grants ECCS-1230040 and ECCS-1501044. The material in this paper was not presented at any IFAC conference. This paper was recommended for publication in revised form by Associate Editor Raul Ordóñez under the direction of Editor Miroslav Krstic.

E-mail addresses: tbian@nyu.edu (T. Bian), zjiang@nyu.edu (Z.-P. Jiang).

[1] Tel.: +1 718 260 3779; fax: +1 718 260 3906.

http://dx.doi.org/10.1016/j.automatica.2016.05.003
0005-1098/

process. However, this assumption is quite strong. When the system model is not fully accessible, finding an initial stabilizing control policy usually involves solving either a linear matrix inequality (LMI) (Jeung, Oh, Kim, & Park, 1996) or some matrix Riccati equations (Khargonekar, Petersen, & Zhou, 1990; Zhou & Khargonekar, 1988), which is computationally expensive. Besides, since the bounds on the system parameters are often assumed known in these methods (which may be exploited to find the initial stabilizing control policy), the system model is not necessarily fully unknown.

In this paper, we depart from the commonly used PI scheme, and propose a new continuous-time VI algorithm that leads to the development of two ADP methods for linear continuous-time, continuous-state space systems. Both the optimal control and stochastic robust optimal control problems will be studied. Employing the VI method has at least two significant advantages: (1) an initial stabilizing control policy is not required; and (2) there is no need to solve matrix equation per-iteration. Due to these two advantages, VI has become the most widely used and best understood algorithm for solving discounted Markov decision problems (Puterman, 1994). Furthermore, VI method for discrete-time, continuous-state space systems can also be found in Bertsekas (2005, Proposition 4.4.1) and Lancaster and Rodman (1995, Section 17.5), for the setting of linear systems; and in Liu, Wang, Zhao, Wei, and Jin (2012), for a nonlinear extension. Unfortunately, VI methods for continuous-time, continuous-state space systems are still not well established. In Vrabie et al. (2013), some efforts have been made to derive a continuous-time VI algorithm, but the convergence of Vrabie's VI has not been proved.

Different from the past results, the continuous-time VI method given in this paper is inspired by the asymptotic stability property of the differential matrix Riccati equation (DMRE). It has been pointed out in Kučera (1973), Shayman (1986) and Willems (1971), that under observability and stabilizability assumptions, the unique symmetric positive definite solution to the algebraic Riccati equation (ARE) is locally asymptotically stable (LAS) for the DMRE, backward in time. Since it is not easy to implement DMRE in ADP design, we borrow the idea of stochastic approximation to construct an iterative updating scheme based on the DMRE, and then use stochastic approximation method to show the convergence. While the stochastic approximation method used in this paper is inspired by the methods in Abounadi, Bertsekas, and Borkar (2002), Andrieu, Moulines, and Priouret (2005) and Chen and Zhu (1986), it is slightly different from these results in the sense that the solution to the proposed algorithm stays in a subset of a level set of the Lyapunov function (see Lemma 3.4).

Since VI is employed in our ADP design, instead of starting from an initial stabilizing control policy, the proposed algorithms start from an arbitrary real symmetric and positive definite matrix representing the initial value function. Moreover, since the data matrices (Θ in (19) and Π in (24)) in the ADP algorithms are independent of the number of learning iterations, there is no need to recalculate the matrix inverse in each iteration. This implies that our methods, for some systems, may be more computationally efficient in each iteration than the methods in Jiang and Jiang (2012) and Vrabie et al. (2009) (see Remark 4.1 and Section 6.3). The obtained results are first tested by a single machine-infinite bus power system. Then, we test our stochastic robust ADP method with a human arm movement task (Wolpert, Diedrichsen, & Flanagan, 2011). A comparison between our continuous-time VI and Kleinman's algorithm is also given. These examples show that our method serves as a powerful tool to solve non-model-based adaptive optimal control problems.

The remainder of this paper is organized as follows. In Section 2, some preliminaries regarding the optimal control problem for linear continuous-time systems are introduced. In Section 3, a new continuous-time VI method is presented with rigorous convergence analysis. Moreover, a detailed comparison between Vrabie's VI and our method is given. Based on the obtained result, two ADP methods for deterministic linear continuous-time systems are developed in Section 4. In Section 5, the obtained ADP algorithm is extended to solve the robust optimal control problem for linear continuous-time stochastic systems with input-dependent noise. Two examples are presented in Section 6. Finally, the conclusion is drawn in Section 7.

Notations: Throughout this paper, I_n denotes the identity matrix of dimension n. $\mathbb{R}$ and $\mathbb{R}_+$ denote the set of real numbers and the set of nonnegative real numbers, respectively. $\mathbb{Z}_+$ denotes the set of nonnegative integers. $|\cdot|$ denotes the Euclidean norm for vectors, or the induced matrix norm for matrices. $\mathscr{S}^n$ denotes the normed space of all n-by-n real symmetric matrices, equipped with the induced matrix norm. $\mathscr{S}^n_+ = \{P \in \mathscr{S}^n : P \geq 0\}$. For a matrix $A \in \mathbb{R}^{n\times m}$, $A^\dagger$ denotes the pseudoinverse of A; $\text{vec}(A) = [a_1^T, a_2^T, \ldots, a_m^T]^T$, where $a_i \in \mathbb{R}^n$ is the ith column of A. For any $A \in \mathscr{S}^n$, denote $\text{vecs}(A) = [a_{11}, a_{12}, \ldots, a_{1n}, a_{22}, a_{23}, \ldots, a_{n-1n}, a_{nn}]^T$, where $a_{ij} \in \mathbb{R}$ is the (i, j)th element of matrix A. $\otimes$ indicates the Kronecker product. A function $f : Q \to \mathbb{R}_+$, where $Q \subseteq \mathbb{R}^n$ and $0 \in Q$, is called positive definite, if $f(x) > 0$ for all $x \in Q \setminus \{0\}$, and $f(0) = 0$. A function f is of class $\mathscr{C}^0(Q)$, where $Q \subseteq \mathbb{R}^n$, if f is continuous on Q. For a continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}$, $\nabla V(x) \in \mathbb{R}^{1\times n}$ denotes the gradient of V at x. For any $0 < T < \infty$, $\mathscr{D}([0, T], \mathscr{S}^n)$ denotes the space of functions from $[0, T]$ to $\mathscr{S}^n$, that are right-continuous with left-hand limits, equipped with the Skorokhod topology (Skorokhod, 1956).

2. Mathematical preliminaries

2.1. Review of stochastic approximation

Stochastic approximation, also known as stochastic gradient descent in some literature, serves as an important tool to solve stochastic optimization problems. Over the past several decades, various stochastic approximation methods have been developed (Abounadi et al., 2002; Andrieu et al., 2005; Borkar, 1998; Chen, 2002; Chen & Zhu, 1986; Kushner & Yin, 2003; Ljung, 1977; Robbins & Monro, 1951). Consider the following stochastic approximation algorithm:

$$\theta_{k+1} = \theta_k + \epsilon_k g(\theta_k, w_k) + Z_k,$$

where g is a nonlinear function, $\{w_k\}_{k=1}^\infty$ is a sequence of i.i.d. zero-mean random noise, ϵ_k is the step size, and Z_k is the projection term. Assume

$$\epsilon_k > 0, \qquad \sum_{k=0}^{\infty} \epsilon_k = \infty, \qquad \sum_{k=0}^{\infty} \epsilon_k^2 < \infty, \tag{1}$$

and the following dynamical system

$$\dot\theta = \mathcal{E}_w g(\theta, w),$$

where $\mathcal{E}$ represents the expectation operator, is asymptotically stable at a fixed-point θ^*, then under some mild conditions on θ_0, w_k and Z_k, one can show (Kushner & Yin, 2003) that $\lim_{k\to\infty} \theta_k = \theta^*$ with probability one.

In this paper, we develop a continuous-time VI algorithm for linear systems based on the stochastic approximation theory.

2.2. System description

This paper considers the following linear system:

$$\dot x = Ax + Bu, \quad x(0) = \xi, \tag{2}$$

where $x \in \mathbb{R}^n$ is the system state, $u \in \mathbb{R}^m$ is the input, and $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are unknown matrices. Assume (A, B) is stabilizable.