Automatica 95 (2018) 213–221

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Output feedback Q-learning for discrete-time linear zero-sum games with application to the H-infinity control☆

Syed Ali Asad Rizvi, Zongli Lin *

Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, P.O. Box 400743, Charlottesville, VA 22904-4743, USA

ARTICLE INFO

Article history:
Received 20 April 2017
Received in revised form 15 January 2018
Accepted 21 April 2018

Keywords:
Reinforcement learning
ADP
Q-learning
Zero-sum games
H-infinity control
Output feedback

ABSTRACT

Approximate dynamic programming techniques usually rely on the feedback of the measurement of the complete state, which is generally not available in practical situations. In this paper, we present an output feedback Q-learning algorithm towards finding the optimal strategies for the discrete-time linear quadratic zero-sum game, which encompasses the H-infinity optimal control problem. A new representation of the Q-function in the output feedback form is derived for the zero-sum game problem and the optimal output feedback policies are presented. Then, a Q-learning algorithm is developed that learns the optimal control strategies online without needing any information about the system dynamics, which makes the control design completely model-free. It is shown that the proposed algorithm converges to the optimal solution obtained by solving the game algebraic Riccati equation (GARE). Unlike the value function based approach used for output feedback, the proposed Q-learning scheme does not require a discounting factor that is generally adopted to mitigate the effect of excitation noise bias. It is known that this discounting factor may compromise the closed-loop stability. The proposed method overcomes the excitation noise bias problem without resorting to the discounting factor, and therefore, converges to the nominal GARE solution. As a result, the closed-loop stability is preserved. An application to the H-infinity autopilot controller for the F-16 aircraft is demonstrated by simulation.

1. Introduction

H-infinity control offers robust performance and stabilization guarantee (Chen, 2013), which make it a good candidate in control problems involving external and cross-coupling disturbances. It finds many important applications such as rotorcrafts (Postlethwaite, Smerlas, Walker, Gubbels, Baillie, Strange, & Howitt, 1999), VSTOL aircrafts (Hyde, Glover, & Shanks, 1995), guided projectiles (Strub, Theodoulis, Gassmann, Dobre, & Basset, 2015), satellites (Frapard & Champetier, 1997) and power systems (Al-Tamimi, Lewis, & Wang, 2007). It has been shown that the H-infinity problem is strongly related with the zero-sum game problem in game theory (Başar & Bernhard, 2005). The Bellman optimality principle plays a key role in solving the optimal control problems using the well-known Bellman or Hamilton–Jacobi–Bellman (HJB) equations. The main difficulty comes from finding the analytical solutions to the HJB equation as it is a partial differential equation. For the special case of linear systems, finding the solution of an optimal control problem leads to solving the algebraic Riccati equations (AREs) associated with the control methods such as linear quadratic regulator (LQR) and the H-infinity optimal control (Başar & Bernhard, 2005; Lancaster & Rodman, 1995; Lewis & Syrmos, 1995). These AREs are nonlinear in the unknown parameters and solving these equations requires complete knowledge of the system dynamics. Computational algorithms have been used to iteratively solve the AREs owing to the difficulty in solving these equations (Hewer, 1971; Lancaster & Rodman, 1995). These algorithms still require full knowledge of the system dynamics.

Reinforcement learning (RL) is a method of solving dynamic optimization problems in which an actor or agent interacts with its environment (system) and modifies its actions, or control policies, based on some stimuli received in response to its actions. In control theory, such techniques are often referred to as adaptive dynamic programming (ADP) or heuristic dynamic programming (HDP). These learning based control methods have been successfully applied towards finding optimal feedback controllers for dynamical systems represented by ordinary differential or difference equations without requiring full knowledge of the system dynamics (Si, 2004). RL-ADP methods often employ function approximation techniques to learn the solution of the Bellman optimality equation. Two techniques that have been successfully applied in these learning control schemes are value function approximation and

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Raul Ordóñez under the direction of Editor Miroslav Krstic.

* Corresponding author.
E-mail addresses: sr9gs@virginia.edu (S.A.A. Rizvi), zl5y@virginia.edu (Z. Lin).

https://doi.org/10.1016/j.automatica.2018.05.027
0005-1098/

Q-learning. Q-learning performs learning over the complete state and action space while value function learning focuses on the state space only. Q-learning offers a completely model-free solution (Watkins & Dayan, 1992; Werbos, 1992). A background on these techniques is given in Lewis and Liu (2013) and Lewis and Vrabie (2009).

Popular optimal control problems such as the linear quadratic regulator (LQR) have been successfully solved using the Q-learning approach (Bradtke, Ydstie, & Barto, 1994; Landelius, 1997). However, in these works, access to the complete state vector is needed. To overcome the requirement of full-state feedback, output feedback techniques have been successfully used that make the whole control design more practical as the number of sensors required is considerably reduced. However, these classical output feedback techniques rely on a dynamic model of the system to estimate the system state, which, in the case of reinforcement learning, is assumed to be unavailable.

Model-free state estimation techniques have gained attention recently. In particular, neural network observers have been developed to provide model-free state estimation (Liu, Huang, Wang, & Wei, 2013; Zhong & He, 2017). Model-free output feedback control can also be achieved by designing a controller directly in the input–output feedback form. These methods have the advantage that the need of a separate state observer is eliminated. A model-free input–output data based scheme was first presented in Lewis and Vamvoudakis (2011) to learn an output feedback LQR controller by using the value function approach. Following the same line, the authors in Kiumarsi, Lewis, Naghibi-Sistani, and Karimpour (2015) solved the model-free optimal tracking problem. Similarly, the continuous-time linear quadratic output feedback problem was addressed in Modares, Lewis, and Jiang (2016). It should be noted that the value function approach is affected by the excitation noise bias problem because the Bellman equation associated with the value function does not include excitation noise. Whereas, an exploratory noise signal is necessary in RL to guarantee parameter convergence. To address this difficulty, a discounting factor in the cost function was introduced in all these output feedback model-free designs, which leads to a sub-optimal solution to the optimal control problem. Studies like Postoyan, Busoniu, Nesic, and Daafouz (2017) have shown that this discounting factor may result in closed-loop instability. This issue of excitation noise bias in the value functions has also been discussed in Kiumarsi, Lewis, and Jiang (2017) for the model-free H-infinity problem. A Q-learning scheme was recently developed in Rizvi and Lin (2017) to solve the output feedback LQR problem without resorting to the discounting factor.

Recently, RL-ADP methods have been successfully applied to provide a model-free solution to the zero-sum game problem. A Q-learning solution to the discrete-time linear quadratic zero-sum game was first developed in Al-Tamimi, Lewis, and Abu-Khalaf (2007), where its application to the H-infinity control problem was shown. Later, the continuous-time zero-sum game problem was solved using partially model-free (Vrabie & Lewis, 2011) and completely model-free (Li, Liu, & Wang, 2014) integral reinforcement learning methods. For nonlinear systems, interested readers can refer to Luo, Wu, and Huang (2015). Recently, an off-policy ADP algorithm was proposed in Kiumarsi et al. (2017) to solve the discrete-time linear quadratic zero-sum game problem, where the issue of excitation noise bias was also addressed. However, in all these works, the measurement of the complete state vector is required. Although output feedback RL algorithms based on value functions (Lewis & Vamvoudakis, 2011) can be used to solve the zero-sum game and the H-infinity problems, these methods are prone to the excitation noise bias. Consequently, a discounting factor is adopted to ensure convergence to a sub-optimal solution. Such a sub-optimal solution, however, does not ensure the closed-loop stability as discussed earlier. Although there exists a lower bound on the discounting factor above which the closed-loop stability may be ensured, the computation of this bound requires knowledge of the system dynamics, which is assumed to be unknown in this problem (Postoyan et al., 2017).

The contributions of this paper are as follows. Compared to the state feedback based model-free approaches (Al-Tamimi, Lewis, & Abu-Khalaf, 2007; Kiumarsi et al., 2017), we have proposed an output feedback method, which is more practical in real-world applications. That is, we seek an output feedback model-free solution towards solving the discrete-time linear quadratic zero-sum games and the associated H-infinity control problem. We have developed an output feedback Q-function description which is more comprehensive than the value function description (Lewis & Vamvoudakis, 2011) due to the explicit dependence of the Q-function on the control inputs and disturbances. In contrast to Lewis and Vamvoudakis (2011), the issue of excitation noise bias is not present in our work due to the inclusion of the input terms in the cost function, which results in the cancellation of noise dependent terms in the Bellman equation. A proof of excitation noise immunity of the proposed scheme is provided. The proposed algorithm does not require a discounting factor which is used in output feedback value function learning. It has been shown that the proposed method guarantees closed-loop stability and that the learned output feedback controller is the optimal controller corresponding to the solution of the original Riccati equation. To the best of our knowledge, this is the first work that performs output feedback with Q-learning for the H-infinity problem. Also, our approach is different from the recently proposed off-policy technique used in Kiumarsi et al. (2017), which also addresses the excitation noise issue but requires an initially stabilizing policy and full-state feedback. Both of these requirements are not present in this work. We note that the output feedback law we developed here is completely model-free. While other output feedback control schemes exist in the literature, they require certain knowledge of the system dynamics and employ a separate state estimator (see, for example, He, Ge, Li, Chew, & Ng, 2015; He, He, & Ge, 2016).

The remainder of this paper is organized as follows. Section 2 provides a description of the problem. In Section 3, we present an output feedback representation of the Q-function for the discrete-time linear quadratic zero-sum game problem, using which the optimal output feedback policies are developed. The main result of this paper is presented in Section 4, where the Q-learning iterative algorithm is proposed that learns the optimal output feedback policies online without requiring any knowledge of the system dynamics. Finally, Section 5 includes simulation results on the proposed scheme. Some concluding remarks are made in Section 6.

2. Problem description

Consider a discrete-time linear time-invariant system in the state-space form,

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Ew_k, \\ y_k &= Cx_k, \end{aligned} \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the system state vector, $u_k \in \mathbb{R}^{m_1}$ is the control input vector, $w_k \in \mathbb{R}^{m_2}$ is the disturbance input vector, and $y_k \in \mathbb{R}^p$ is the output vector. The zero-sum problem game can be formulated as a minimax problem with the cost function of the form (Al-Tamimi, Lewis, & Abu-Khalaf, 2007; Başar & Bernhard, 2005; Kiumarsi et al., 2017),

$$V^*(x_k) = \min_{u_i} \max_{w_i} \sum_{i=k}^{\infty} r(x_i, u_i, w_i), \tag{2}$$