Automatica 61 (2015) 274–281

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Non-zero sum Nash Q-learning for unknown deterministic continuous-time linear systems☆

Kyriakos G. Vamvoudakis

Center for Control, Dynamical-systems and Computation (CCDC), University of California, Santa Barbara, CA 93106-9560, USA

ARTICLE INFO

Article history:
Received 5 May 2014
Received in revised form
8 July 2015
Accepted 8 August 2015
Available online 5 September 2015

Keywords:
Q-learning
Nash-games
Uncertain systems
Model-free formulation

ABSTRACT

This work proposes a novel Q-learning algorithm to solve the problem of non-zero sum Nash games of linear time invariant systems with N-players (control inputs) and centralized uncertain/unknown dynamics. We first formulate the Q-function of each player as a parametrization of the state and all other the control inputs or players. An integral reinforcement learning approach is used to develop a model-free structure of N-actors/N-critics to estimate the parameters of the N-coupled Q-functions online while also guaranteeing closed-loop stability and convergence of the control policies to a Nash equilibrium. A 4th order, simulation example with five players is presented to show the efficacy of the proposed approach.
© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Most game-based (Basar & Olsder, 1999) control system design techniques rely on complete knowledge of the systems to be controlled. This is not the case when the entire system is not modeled exactly or some parameters are uncertain or completely unknown and for that reason one has to find more intelligent techniques that can still guarantee optimal performance and closed-loop stability. Reinforcement learning (RL) (Sutton & Barto, 1998) is a machine learning approach that is developed primarily for systems with discrete dynamics and actions. In the control literature, the field where RL methods are studied is called approximate dynamic programming (ADP). There is a substantial work on ADP for discrete-time systems, that solve mostly optimal regulation and tracking problems with policy iteration algorithms (Liu & Wei, 2014), value iteration algorithms (Wei, Wang, Liu, & Yang, 2014) and other iterative algorithms (Wei & Liu, 2014b). The efficiency of such algorithms has been shown extensively in several practical problems, e.g. Wei and Liu (2014a,c). There are several algorithms for continuous-time systems based on ADP and optimal adaptive control (Bertsekas & Tsitsiklis, 1996; Busoniu, Babuska, deSchutter, & Ernst, 2010; Lewis, Vrabie, & Vamvoudakis, 2012; Powell, 2007; Vrabie, Vamvoudakis, & Lewis, 2012; Zhang, Liu, Luo, & Wang, 2012) that can achieve the desired controller performance but rely on complete or partial knowledge of the dynamics or use identifiers to approximate unknown functions. One other obvious solution to get from continuous state space to discrete, and apply well known RL techniques (Sutton & Barto, 1998), is to quantize the state space, in a form of state aggregation. But one of the pitfalls of such a quantization approach is that the solution found is probably suboptimal since due to quantization the set of actions is reduced and hence the optimal policy for the continuous-time problem may not be in the obtained set of actions. Moreover discretization of a continuous state space system may cause the forfeit of the Markovian properties of the process and as a result the convergence proofs may no longer be valid. Game theory (Basar & Olsder, 1999; Tijs, 2003; Vrabie et al., 2012), provides an ideal environment to study multi-player decision and control problems (e.g. coupled large-scale systems), and offers a wide range of challenging and engaging problems. Since each controller wants to minimize its own cost function, Nash strategy offers a nice framework to study control robustness. In continuous-time linear systems with multiple decision makers and quadratic costs, one has to rely on solving complicated coupled matrix Riccati equations that require complete knowledge of the system matrices and need to be solved offline and then implemented online in the controller. In the era of complex and big data systems, modeling the processes exactly is most of the time infeasible and offline solutions make the systems vulnerable to parameter changes (drift) and adversarial attacks. There is a need for intelligent algorithms that self-heal, resist attacks, and allow dynamic optimization. Q-learning is a model-free RL technique developed primarily for discrete-time

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Kok Lay Teo under the direction of Editor Ian R. Petersen.

E-mail address: kyriakos@ece.ucsb.edu.

http://dx.doi.org/10.1016/j.automatica.2015.08.017
0005-1098/© 2015 Elsevier Ltd. All rights reserved.

systems (Watkins & Dayan, 1992). It learns an action-dependent value function that ultimately gives the expected utility of taking a given action in a given state and following the optimal policy thereafter. When such an action-dependent value function is learned, the optimal policy can be computed easily. The *biggest strength* of Q-learning is that it is model free. It has been proven in Watkins and Dayan (1992) that for any finite Markov Decision Process, Q-learning eventually finds an optimal policy. In complex-systems Q-learning needs to store a lot of data, which makes the algorithm infeasible. This problem can be solved effectively by using adaptation techniques. Specifically, Q-learning can be improved by using the universal function approximation property of neural networks (NNs) and especially in the context of ADP (Werbos, 1992) or neuro-dynamic programming (Bertsekas & Tsitsiklis, 1996) that allow us to solve difficult optimization problems online and forward in time. This makes it possible to apply the algorithm to larger problems, even when the state space is continuous, and infinitely large.

Related work

The work of Freiling, Jank, and Abou-kandil (1996) proposes global existence results for the solutions of coupled-Riccati equations in closed-loop Nash games but with known dynamics and without proper convergence and stability proofs. On the other side the authors in Jungers, Castelan, De Pieri, and Abou-Kandil (2008) design robust controllers, inspired by a Nash strategy for systems with polytopic representation of uncertainty. An online method for solving coupled Hamilton–Jacobi equations (coupled-Riccati equations in linear systems) of deterministic nonlinear systems with known dynamics has been proposed in Vamvoudakis and Lewis (2011) along with stability and performance guarantees. In Xu and Xiao (2013) the authors propose an iterative refinement algorithm along with fully mathematical justifications, which sharpens matrix solution upper bounds for the continuous coupled algebraic Riccati equations. The work of Limebeer, Anderson, and Hendel (1994) provides necessary and sufficient conditions for the existence of a solution to the mixed H_2/H_∞ problem in the infinite horizon case in terms of the existence of solutions to a pair of cross-coupled algebraic Riccati equations. The authors in Al-Tamini, Abu-Khalaf, and Lewis (2007), have proposed a sequential update Q-learning approach to solve zero-sum games in systems with discrete dynamics. In the same manner, the work of Kiumarsi, Lewis, Modares, Karimpour, and Naghibi-Sistani (2014) proposes a Q-learning framework to solve the optimal tracking problem of discrete time systems. The author in Littman (2001) has investigated 2-player zero-sum Markov games and proposed minimax Q-learning based methods that do not require explicit knowledge of the environment. Based on the work of Littman (2001), the authors in Hu and Wellman (2003) and Suematsu and Hayashi (2002) propose a Q-learning algorithm for multi-agent systems where the agents choose Nash-equilibrium policies. Specifically for (Hu & Wellman, 2003) the authors have some highly restrictive assumptions on the form of stage games during learning, to guarantee convergence. Whereas in Suematsu and Hayashi (2002) the authors maintain large Q-tables for all agents and provide convergence to a Nash equilibrium when all agents are adaptable, otherwise convergence to an optimal response equilibrium is achieved. A pursuit evasion game between a plane and a missile by using minimax Q-learning has been investigated in Harmon, Baird, and Klopf (1995). Continuous-time systems on the other side lack a proper completely mode-free non-Zero sum Nash game formulation due to dependencies to the system matrices of the optimal controllers and the coupled Riccati equations. For that reason, most of the times one has to rely on discretization of the state and the action space to apply such techniques, and as such lose important information during discretization. Some initial work on Q-learning for continuous time systems has been investigated in Baird (1994) and Doya (2000) but without any convergence and stability guarantees. The authors in Mehta and Meyn (2009) have established connections between Q-learning and nonlinear control of continuous-time models with general state and action space by observing that the Q-function developed in Watkins and Dayan (1992) is an extension of the Hamiltonian that appears in the minimum principle. A non-synchronous ϵ-integral Q-function has been used to propose an ϵ-approximate Q-learning framework for solving the linear quadratic regulator problem of continuous-time systems in Young Lee, Bae Park, and Ho Choi (2012). The authors in Young Lee et al. (2012) require lots of computations due to index iterations, cannot guarantee robustness and only prove uniform ultimate boundedness of the closed-loop system. The work of Xu, Zhao, and Jagannathan (2014) proposes a Q-learning approach to solve the finite-horizon optimal control problem which eventually reduces to solve the differential Riccati equation whereas in the present paper we solve the coupled algebraic Riccati equations. The authors of Palanisamy, Modares, Lewis, and Aurangzeb (2015) propose a Q-learning approach to solve the continuous-time infinite-horizon optimal control problem by writing the Q-function with respect to the state, the control input and the derivatives of the control input. The algorithm proposed in Palanisamy et al. (2015) uses iterations as in Young Lee et al. (2012) on the value function to solve the problem while the proposed algorithm of the current paper will solve the more difficult problem of learning multiple value functions in a synchronous manner. The algorithm in this paper shall be more aligned with the adaptive control setting (Ioannou & Fidan, 2006; Tao, 2003).

Contributions

Since solving Nash game requires complete knowledge of the centralized system dynamics and complicated offline computation, this work proposes a completely mode-free algorithm to solve the coupled Riccati equations arising in multi-player non-zero sum Nash games in deterministic systems. The contributions of the present paper are fourfold. First, a parameterized Q-function is derived for every of the N-players in the game that depends on the state and the control inputs of all the players. Second, we derive mode-free controllers based on the Q-functions parametrization, and we prove that by employing N-coupled minimizations on the derived Q-functions, the controllers form a Nash equilibrium. Third, in order to solve the coupled optimizations problems in an efficient way by overcoming the curse of dimensionality problem, we use $2N$-universal approximators such as NNs to approximate the cost and the control of every player, by using namely a critic and an actor NN. Fourth, we prove asymptotic stability of the closed-loop signals and convergence to a Nash equilibrium by using rigorous Lyapunov-based proofs.

Structure

The remainder of the paper is structured as follows. Section 2 formulates the multi-player non-zero sum game for systems with linear-time invariant dynamics while Section 3 provides a brief background on the existence of a Nash equilibrium. Since Section 3 relies on complete knowledge of the system and input matrices, Section 4 provides a model-free formulation based on a Q-learning approach and a structure based on N-critic and N-actors, to estimate the parameters of each player's Q-function. The effectiveness of the proposed approach is illustrated in Section 5 through a simulation result with unknown dynamics. Finally Section 6 concludes and talks about future work.

Download English Version:

https://daneshyari.com/en/article/695260

Download Persian Version:

https://daneshyari.com/article/695260

Daneshyari.com