Neurocomputing 000 (2017) 1–10

Contents lists available at ScienceDirect

Neurocomputing

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

Data-driven adaptive dynamic programming schemes for non-zero-sum games of unknown discrete-time nonlinear systems

He Jiang, Huaguang Zhang*, Kun Zhang, Xiaohong Cui

College of Information Science and Engineering, Northeastern University, Box 134, 110819 Shenyang, PR China

ARTICLE INFO

Article history:
Received 7 November 2016
Revised 1 March 2017
Accepted 4 September 2017
Available online xxx

Communicated by Zidong Wang

Keywords:
Reinforcement learning
Adaptive dynamic programming
Data-driven
Non-zero-sum games
Neural networks

ABSTRACT

This paper integrates game theory, optimal control theory and reinforcement learning to deal with the discrete-time (DT) multi-player non-zero-sum game issue. As is known, the solutions to non-zero-sum game problems are the outcomes of coupled Riccati equations or coupled Hamilton–Jacobi ones, which are generally difficult to solve analytically and require the knowledge of accurate system mathematical models. However, for most practical industrial systems, the system dynamics cannot be obtained accurately or even unavailable, and the conventional model-based methods will be invalid. To overcome this deficiency, we develop data-based adaptive dynamic programming (ADP) algorithms for completely unknown multi-player systems. Firstly, the Nash equilibrium and stationarity conditions are used to formulate the DT multi-player non-zero-sum game, and then policy iteration algorithm is applied to approximate optimal solutions successively. Secondly, a novel online ADP algorithm combined with a neural-network-based identification scheme is designed and only requires the system data instead of the real system models. Subsequently, a data-driven action-dependent heuristic dynamic programming approach is presented and circumvents the estimation errors caused by the identification learning procedure. Finally, two simulation examples are provided to illustrate the feasibility of our schemes.

1. Introduction

Dynamic programming is a conventional approach in dealing with optimal control problems and computes the optimal solutions backward-in-time [1]. However, it is computationally untenable to achieve the results due to the well-known "curse of dimensionality" [2]. As a brain-like intelligent method, adaptive dynamic programming (ADP) [3–10] with the abilities of adaptivity and self-learning is an important branch of reinforcement learning (RL) [11–14] and able to compute the solution forward-in-time. So far, ADP has successfully addressed various optimal control issues for both continuous-time (CT) and discrete-time (DT) systems, such as optimal tracking control [15–19], finite-horizon optimal control [20–24], robust optimal control [25–28], optimal control with time-delays [29–31] and constrained inputs [32–34], and optimal control applied on nonaffine systems [35,36], Markov jump systems [37,38] and power systems [39,40].

It is worth mentioning that most traditional ADP methods require accurate mathematical models of system dynamics. However, in the practical industrial systems, system models are difficult to attain accurately, and the model-based ADP methods can not be directly applied in the real world control processes. Fortunately, input–output system data can be measured and utilized, which motivates the development of data-driven control techniques [41,42]. There are two mainstream data-driven schemes within the scope of ADP. The first approach employs accessed system data to design optimal controllers directly [43,44], and the second one is to reconstruct data-driven system structures through neural networks (NNs) to approximate the real system ones [45,46].

Recently, researchers have attempted to bring together ADP approaches and the game theory to investigate multi-player game problems, such as graphical games, zero-sum and non-zero-sum games. As is known, the solutions to the general optimal control problems are Riccati and Hamilton–Jacobi–Bellman (HJB) equations for linear and nonlinear systems, respectively. Two-player zero-sum games are usually employed to describe robust H_∞ control issues, and due to the disturbance input term, the solutions, i.e., Hamilton–Jacobi–Isaacs (HJI) equations are more difficult than HJB ones [47–49]. In addition, the game theory and reinforcement learning have been applied on multi-agent systems to formulate and solve graphical games. In [50], the online adaptive learning solutions to differential graphical games were obtained by em-

* Corresponding author.
E-mail addresses: jianghescholar@163.com (H. Jiang), hgzhang@ieee.org, zhg516516@gmail.com (H. Zhang), nukgnahz@163.com (K. Zhang), xiaohong19821206@126.com (X. Cui).

http://dx.doi.org/10.1016/j.neucom.2017.09.020
0925-2312/

ploying reinforcement learning methods. Afterwards, the idea of Vamvoudakis et al. [50] was applied and extended to different cases of heterogeneous linear multi-agent systems [51], nonlinear multi-agent systems [52], disturbance rejection [53] and DT graphical games [54].

Among the practical engineering applications, a large class of real world systems have more than one controller, such as networked communication systems, coupled large-scale systems and microgrid systems. The optimal control issue of these systems with multiple controllers can be formulated by the non-zero-sum game theory, which was originally investigated in [55] and has a great number of promising applications in military, economics and control engineering. The non-zero-sum game aims to produce a set of optimal control policies to form the Nash equilibrium, which can not only minimize the individual performance index function for each player but also guarantee the stability of the systems [56]. For the linear non-zero-sum game systems, one needs to solve a set of coupled Riccati equations to obtain the Nash equilibrium. For the nonlinear cases, more complicated coupled Hamilton–Jacobi (HJ) equations are required to be solved to achieve the results. Unfortunately, it is generally difficult to get global solutions [57]. In order to overcome this difficulty, many ADP-based approaches have been proposed to approximate the optimal solutions of non-zero-sum games [57–61]. In [57], a policy iteration (PI) based online adaptive learning algorithm implemented by critic and actor networks was proposed to solve coupled HJ equations of CT non-zero-sum games. Based on the theoretical framework formulated by the work [57], a single network ADP algorithm was designed instead of the traditional dual network architecture and successfully reduced the computational burden in [58]. Identification schemes with the online synchronous approximate optimal learning approach and the experience replay technique were presented for unknown CT systems in [59] and [60], respectively. In [61], the data-driven off-policy integral reinforcement learning method was extended to the model-free non-zero-sum games without any identification processes. The aforementioned Refs. [57–61] all considered the CT non-zero-sum game issues. However, there are still few works concerning DT version.

This paper combines ADP algorithms and data-driven techniques to address model-free DT non-zero-sum game problems. Firstly, the Nash equilibrium and stationarity conditions of optimization are employed to derive and formulate the DT non-zero-sum game, and develop the PI scheme to learn and approximate the optimal solutions iteration by iteration in Section 2. Secondly, two ADP algorithms, an identification-based one and a direct data-driven one, are proposed in Section 3 and Section 4, respectively. Thirdly, Section 5 provides two numerical simulation examples to demonstrate the feasibility of our proposed approaches. Finally, a brief conclusion is drawn in Section 6.

2. Problem statement

Let us consider the discrete-time N-player system as follows:

$$x(k+1) = f(x(k)) + \sum_{j=1}^{N} g_j(x(k))u_j(k) \tag{1}$$

where $x(k) \in \Omega \subset \mathbb{R}^n$ denotes the system state and $u_j(k) \in \mathbb{R}^{m_j}$ represents the control input; $f(x) \in \mathbb{R}^n$ and $g_j(x) \in \mathbb{R}^{n\times m_j}$ are the system functions.

The performance index function associated with each player is described by

$$J_i(x(0), u_1, u_2, \ldots, u_N) = \sum_{p=0}^{\infty} r_i(x(p), u_i(p), u_{(-i)}(p)) \tag{2}$$

where $u_{(-i)} = \{u_j : j = 1, 2, \ldots, N, \ \ j \neq i\}$; $r_i(x(p), u_i(p), u_{(-i)}(p)) = x^T(p)Q_i x(p) + u_i^T(p)R_{ii}u_i(p) + \sum_{j=1, j\neq i}^{N} u_j^T(p)R_{ij}u_j(p)$ with the symmetric positive definite weight matrices $Q_i > 0$, $R_{ii} > 0$ and $R_{ij} > 0$.

Definition 1. [57] (Admissible control) For $\forall i$, the control policy $u_i(x) \in \psi(\Omega)$ is admissible with respect to (2) on the compact set $\Omega \subset \mathbb{R}^n$, if $u_i(x)$ can not only make (1) stable but also guarantee (2) finite.

Assumption 1. The N-player system (1) is controllable and there exists at least one set of admissible control policies.

Given admissible state feedback control policies $u_i(x)$ for $\forall i$, the value function is given by

$$V_i(x(k)) = \sum_{p=k}^{\infty} r_i(x(p), u_i(p), u_{(-i)}(p)). \tag{3}$$

Define the optimal value function as

$$V_i^*(x(k)) = \min_{u_i} \sum_{p=k}^{\infty} r_i(x(p), u_i(p), u_{(-i)}(p)). \tag{4}$$

Definition 2. [60] (Nash equilibrium) A set of control policies $\{u_1^*, u_2^*, \ldots, u_N^*\}$ is regarded as the solution to a Nash equilibrium of the N-player game, if, for $\forall i$, the inequality holds as below

$$J_i^* \triangleq J_i(u_1^*, u_2^*, \ldots, u_i^*, \ldots, u_N^*) \leq J_i(u_1^*, u_2^*, \ldots, u_i, \ldots, u_N^*). \tag{5}$$

Lemma 1. *Let us consider the N-player system (1) along with the value function (3). Based on the stationarity conditions of optimization, for ∀i, one can get optimal control policy*

$$u_i^*(k) = -\frac{1}{2}R_{ii}^{-1}g_i^T(x(k))\nabla V_i^*(x(k+1)) \tag{6}$$

where $\nabla V_i^*(x(k+1)) = \partial V_i^*(x(k+1))/\partial x(k+1)$ *and, for* $\forall i$, $V_i^*(x)$ *satisfies the coupled equation as below*

$$V_i^*(x(k)) = V_i^*(x(k+1)) + x^T(k)Q_i x(k) + u_i^{*T}(k)R_{ii}u_i^*(k) + \sum_{j=1, j\neq i}^{N} u_j^{*T}(k)R_{ij}u_j^*(k). \tag{7}$$

Proof. In light of Definition 2 and related works [54,56,57], Lemma 1 can be easily derived. □

Theorem 1. *Let Assumption 1 hold. Let* $V_i^*(x)$ *satisfy the coupled Eq. (7) and each control policy* $u_i^*(x)$ *use the form of (6). Then, one has*

1. *The dynamics of the system (1) are asymptotically stable;*
2. *The game value of each player is expressed as* $J_i^*(x(k), u_i^*(k), u_{(-i)}^*(k)) = V_i^*(x(k))$;
3. *The control policies* u_i^* *and* $u_{(-i)}^*$ *constitute a Nash equilibrium.*

Proof. 1. Since $V_i^*(x(k))$ satisfies (7), it can be acquired that

$$V_i^*(x(k+1)) - V_i^*(x(k)) = -x^T(k)Q_i x(k) - u_i^{*T}(k)R_{ii}u_i^*(k) - \sum_{j=1, j\neq i}^{N} u_j^{*T}(k)R_{ij}u_j^*(k) \leq 0. \tag{8}$$

which implies $V_i^*(x(k))$ can serve as Lyapunov function, and the dynamics of the system (1) are asymptotically stable.

2. Rewrite the performance index function (2) by adding and subtracting $V_i(x(k))$ as

$$J_i(x(k), u_1, u_2, \ldots, u_N) = \sum_{p=k}^{\infty} r_i(x(p), u_i(p), u_{(-i)}(p)) - V_i(x(k)) + V_i(x(k)). \tag{9}$$

For $\forall i$, let $u_i = u_i^*$ for both $J_i(x(k), u_1, u_2, \ldots, u_N)$ and $V_i(x(k))$. According to (2) and (3), one can obtain