IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Model-Free Temporal Difference Learning for Non-Zero-Sum Games

Liming Wang, Yongliang Yang,
Dawei Ding, Yixin Yin
School of Automation and
Electrical Engineering
University of Science
and Technology Beijing
Beijing, China
limingwangustb@163.com;
yangyongliang@ustb.edu.cn;
dingdawei@ustb.edu.cn;
yyx@ies.ustb.edu.cn

Zhishan Guo
Department of Electric and
Computer Engineering
University of Central Florida
Orlando, Florida, USA
Zhishan.Guo@ucf.edu

Donald C. Wunsch
Department of Electrical and
Computer Engineering
Missouri University of
Science and Technology
Rolla, Missouri, USA
dwunsch@mst.edu

***Abstract*—In this paper, we consider the two-player non-zero-sum games problem for continuous-time linear dynamic systems. It is shown that the non-zero-sum games problem results in solving the coupled algebraic Riccati equations, which are nonlinear algebraic matrix equations. Compared with the algebraic Riccati equation of the linear dynamic systems with only one player, the coupled algebraic Riccati equations of non-zero-sum games with multi-player are more difficult to be solved directly. First, the policy iteration algorithm is introduced to find the Nash equilibrium of the non-zero-sum games, which is the sufficient and necessary condition to solve the coupled algebraic Riccati equations. However, the policy iteration algorithm is offline and requires complete knowledge of the system dynamics. To overcome the above issues, a novel online iterative algorithm, named integral temporal difference learning algorithm, is developed. Moreover, an equivalent compact form of the integral temporal difference learning algorithm is also presented. It is shown that the integral temporal difference learning algorithm can be implemented in an online fashion and requires only partial knowledge of the system dynamics. In addition, in each iteration step, the closed-loop stability using the integral temporal difference learning algorithm is analyzed. Finally, the simulation study shows the effectiveness of the presented algorithm.**

***Index Terms*—integral temporal difference learning, value iteration, non-zero-sum games, Nash equilibrium**

I. Introduction

In game theory, multiple decision makers or players interact with each other and try to maximize their own interests [1]. It can be divided into two categories: zero-sum (ZS) games and non-zero-sum (NZS) games [2]. In ZS games, the sum of the interests of all players remains to be zero, so the increase in the interests of one player will lead to the reduction of the remaining players, which is essentially competitive game; In contrast, for NZS games, the relationship among players can be either competitive or cooperative. Recently, game theory has been widely used in control application [3], [4], economics management [5], [6], power systems [7] and wireless sensor networks [8].

For the systems with only one player, the optimal control problem requires solving the algebraic Riccati equation (ARE) for linear systems or the Hamilton-Jacobi-Bellman (HJB) equation for nonlinear systems, which is difficult to be solved analytically [9]. The NZS games with multiple players usually result in solving the coupled AREs for linear systems [10] and the coupled Hamilton-Jacobi equations (HJEs) for nonlinear systems [11]–[13]. The coupled HJEs/AREs are more challenge to be solved. Dynamic programming is a well-known method for solving the dynamic optimization problem. However, 'the curse of dimensionality' exists due to the essence of backward-in-time [2]. In order to overcome this issue, the forward-in-time method is desired [9], [14].

As a powerful and effective tool, adaptive dynamic programming (ADP) plays an important role in finding the optimal control policies of various problems, such as multi-agent consensus problems [15], [16], input constrained problems [17], [18], tracking problem [19], robust control problem [20], [21] and intermittent feedback design [22] for system with uncertainty [23]. The offline ADP method generates a sequence of value functions satisfying the Lyapunov equations [17], which requires a well-defined region to apply the least-squares

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61333002, in part by the Fundamental Research Funds for the China Central Universities of USTB under Grant FRF-TP-18-031A1 and FRF-BD-17-002A, in part by the China Post-Doctoral Science Foundation under Grant 2018M641197, in part by the National Science Foundation under Grant CNS-1850851, in part by the Mary K. Finley Endowment, in part by the Missouri S&T Intelligent Systems Center, in part by the Lifelong Learning Machines program from DARPA/Microsystems Technology Office, and in part by the Army Research Laboratory under Cooperative Agreement No. W911NF-18-2-0260. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

978-1-7281-2009-6/$31.00 ©2019 IEEE

(LS) method. In addition, two iterative ADP algorithms, policy iteration (PI) algorithm and value iteration (VI) algorithm have been extensively studied. [24] proposes an online PI algorithm with two iterative steps to solve the optimal control policies. On the other hand, VI algorithms are presented in [25] to solve the coupled AREs for continuous-time linear systems. In this paper, we present the integral temporal difference (TD) learning method to approximate the solution to the coupled AREs in the NZS games for continuous-time linear systems with two-player.

The remainders of this paper are arranged as follows. Section II provides the problem statement and gives the coupled AREs of the NZS games. Section III gives an offline policy iteration algorithm. Section IV presents an online integral TD learning algorithm and an equivalent form of this algorithm, the stability proof of closed-loop systems is completed. The simulation with forth-order systems in Section V supports the theory. Finally, a conclusion is given in Section VI.

II. Problem Formulation

We consider the continuous-time linear dynamical system

$$\dot{x}\left(t\right) = Ax + B_1 u_1 + B_2 u_2, \tag{1}$$

where $x \in \mathbb{R}^n$ is the system state with initial state x_0. $u_1 \in \mathbb{R}^{m_1}$ is the player one and $u_2 \in \mathbb{R}^{m_2}$ is the player two.

For each player, the NZS differential game on an infinite time horizon is to minimize the following performance functions defined as

$$V_1\left(x_0\right) = \int_0^{\infty} \left(x^T Q_1 x + u_1^T R_{11} u_1 + u_2^T R_{12} u_2\right) d\tau \tag{2}$$

$$V_2\left(x_0\right) = \int_0^{\infty} \left(x^T Q_2 x + u_1^T R_{21} u_1 + u_2^T R_{22} u_2\right) d\tau \tag{3}$$

where $Q_i, i = 1,2$ is positive definiteness matrix, $R_{ij},\ i,j = 1,2$ is also positive definiteness matrix.

The following assumption and definitions are required for the subsequent discussions.

Definition 1: ***(Admissible Control)*** Feedback control policies $u_i = \mu_i\left(x\right)$ are called as admissible with respect to (2) and (3) on a set $\Omega \in \mathbb{R}^n$, denoted by $\mu_i \in \psi\left(\Omega\right)$, if $\mu_i\left(x\right)$ is continuous on Ω, $\mu_i\left(x\right) = 0$, $\mu_i\left(x\right)$ stabilizes (1) on Ω, and (2) and (3) are finite for $\forall x_0 \in \Omega$.

Definition 2: ***(Nash Equilibrium)*** For NZS games with two players, $\left(\mu_1^*, \mu_2^*\right)$, is a Nash equilibrium solution, if the following inequality is satisfied for $\forall \mu_i^* \in \Omega_i$, $i = 1,2$

$$V_1^* \triangleq V_1\left(\mu_1^*, \mu_2^*\right) \leqslant V_1\left(\mu_1, \mu_2^*\right) \tag{4}$$

$$V_2^* \triangleq V_2\left(\mu_1^*, \mu_2^*\right) \leqslant V_2\left(\mu_1^*, \mu_2\right) \tag{5}$$

Assumption 1: The matrix pair $\left(A, \left[\begin{array}{cc} B_1 & B_2 \end{array}\right]\right)$ is stabilizable.

In this paper, the problem of interest can be formulated as follows.

Problem 1: ***(Two-Player NZS games)*** Consider the system (1), find a Nash equilibrium solution defined by Definition 2, $\left(\mu_1^*, \mu_2^*\right)$, such that the performance functions described by (2) and (3) are minimized.

In the next, we will give an equivalent condition to solve Problem 1, named the coupled algebraic Riccati equation.

Lemma 1: [10] Under Assumption 1, consider the system (1) with the performance functions defined by (2) and (3). Then, $\left(K_1^*, K_2^*\right)$, defined as $K_i^* = R_{ii}^{-1} B_i^T P_i^*$, $i = 1,2$, is a feedback Nash equilibrium if and only if $\left(P_1^*, P_2^*\right)$ is a symmetric stabilizing solution of the coupled AREs (6) and (7).

$$\begin{aligned} 0 = {} & A^T P_1^* + P_1^* A + Q_1 - P_2^* B_2 R_{22}^{-1} B_2^T P_1^* \\ & - P_1^* B_2 R_{22}^{-1} B_2^T P_2^* - P_1^* B_1 R_{11}^{-1} B_1^T P_1^* \\ & + P_2^* B_2 R_{22}^{-1} R_{12} R_{22}^{-1} B_2^T P_2^* \qquad (6) \\ 0 = {} & A^T P_2^* + P_2^* A + Q_2 - P_1^* B_1 R_{11}^{-1} B_1^T P_2^* \\ & - P_2^* B_1 R_{11}^{-1} B_1^T P_2^* - P_2^* B_2 R_{22}^{-1} B_2^T P_2^* \\ & + P_1^* B_1 R_{11}^{-1} R_{21} R_{11}^{-1} B_1^T P_1^* \qquad (7) \end{aligned}$$

Note that the coupled AREs is quadratic in P_1^* and P_2^*, where P_1^* and P_2^* are also coupled. This makes the coupled AREs difficult to solve. Therefore, in the next section, iterative methods are presented to solve the coupled AREs.

III. Offline Policy Iteration for Solving coupled AREs

In this section, the offline algorithm based on policy iteration will be given to solve the NZS games with two-player.

Definition 3: ***(Riccati Operator)*** For each player, define the Riccati operator $Ric_i\left(X_1, X_2\right)$ as

$$\begin{aligned} & Ric_1\left(X_1, X_2\right) = A^T X_1 + X_1 A + Q_1 - X_2 B_2 R_{22}^{-1} B_2^T X_1 \\ & - X_1 B_2 R_{22}^{-1} B_2^T X_2 - X_1 B_1 R_{11}^{-1} B_1^T X_1 \\ & + X_2 B_2 R_{22}^{-1} R_{12} R_{22}^{-1} B_2^T X_2, \qquad (8) \\ & Ric_2\left(X_1, X_2\right) = A^T X_2 + X_2 A + Q_2 - X_1 B_1 R_{11}^{-1} B_1^T X_2 \\ & - X_2 B_1 R_{11}^{-1} B_1^T X_1 - X_2 B_2 R_{22}^{-1} B_2^T X_2 \\ & + X_1 B_1 R_{11}^{-1} R_{21} R_{11}^{-1} B_1^T X_1. \qquad (9) \end{aligned}$$

Note that the operator Ric_i has an important role in evaluating the performance defined by (2) and (3). $Ric_i\left(X_1, X_2\right) = 0$ means that the performance functions (2) and (3) are minimized and system (1) has reached optimal. If $0 < Ric_i\left(X_1^{(k+1)}, X_2^{(k+1)}\right) < Ric_i\left(X_1^{(k)}, X_2^{(k)}\right)$ holds, it indicates that the performance of step $k+1$ is closer to the optimal solution than that of step k.

The coupled AREs (6) and (7) can be solved iteratively by using the policy iteration algorithm, as shown in Algorithm 1. Also, the corresponding Bellman equations can be obtained as:

$$\dot{V}_1^{(k)}\left(x_t\right) + r_1\left(x_t, u_1^{(k)}, u_2^{(k)}\right) = 0 \tag{10}$$

$$\dot{V}_2^{(k)}\left(x_t\right) + r_2\left(x_t, u_1^{(k)}, u_2^{(k)}\right) = 0 \tag{11}$$

where $V_i^{(k)}(x_t) = x_t^T P_i^{(k)} x_t$, $r_i\left(x_t, u_1^{(k)}, u_2^{(k)}\right) = x^T Q_i x + (u_1^{(k)})^T R_{i1} u_1^{(k)} + (u_2^{(k)})^T R_{i2} u_2^{(k)}$. The above Bellman equations can be further expressed as Lyapunov equations as

$$\begin{aligned} 0 = &\left(\bar{A}^{(k)}\right)^T P_1^{(k)} + P_1^{(k)} \bar{A}^{(k)} + Q_1 \\ &+\left(K_1^{(k)}\right)^T R_{11} K_1^{(k)} + \left(K_2^{(k)}\right)^T R_{12} K_2^{(k)} \end{aligned} \tag{12}$$

$$\begin{aligned} 0 = &\left(\bar{A}^{(k)}\right)^T P_2^{(k)} + P_2^{(k)} \bar{A}^{(k)} + Q_2 \\ &+\left(K_1^{(k)}\right)^T R_{21} K_1^{(k)} + \left(K_2^{(k)}\right)^T R_{22} K_2^{(k)} \end{aligned} \tag{13}$$

where $\bar{A}^{(k)} = A - B_1 K_1^{(k)} - B_2 K_2^{(k)}$.

Algorithm 1 Offline Policy Iteration Algorithm

```
1: Given a pair of initial admissible control $(u_1^{(0)}, u_2^{(0)})$, such
   that the system (1) is a stable closed loop system.
2: Policy Evaluation: solve (12) and (13) for $P_1^{(k)}, P_2^{(k)}$.
3: Policy Improvement:
       $K_1^{(k+1)} = R_{11}^{-1} B_1^T P_1^{(k)}$
       $K_2^{(k+1)} = R_{22}^{-1} B_2^T P_2^{(k)}$
4: Stop at convergence, otherwise, set $k = k + 1$ and go to
   Step 2
```

The offline algorithm 1 needs to know the complete system model in advance, i.e., both A and B_1, B_2. Moreover, the algorithm will be invalid when the system changes, or disturbance exists. In the next section, a novel online method will be developed.

IV. Main Results

A. Integral TD learning

This section presents the main algorithm, i.e., the integral TD learning, to solve the NZS games with two players in an online fashion. In addition, instead of the complete system knowledge, only partial knowledge of the system dynamics is required.

Consider the system (1) and its value function (2) and (3), the value function can be rewritten as

$$\begin{aligned} V_i(x_t) &= \int_t^{t+T} x^T \bar{Q}_i x d\tau + \int_{t+T}^{\infty} x^T \bar{Q}_i x d\tau \\ &= \int_t^{t+T} x^T \bar{Q}_i x d\tau + V_i(x_{t+T}), \end{aligned} \tag{14}$$

where

$$\bar{Q}_i = Q_i + (K_1)^T R_{i1} K_1 + (K_2)^T R_{i2} K_2$$

and (u_1, u_2) guarantee the stability of the closed-loop system.

Therefore, for a pair of given policy $(u_1^{(k)}, u_2^{(k)})$, the integral TD error $\delta_t\left(V_i^{(k)}, u_1^{(k)}, u_2^{(k)}, T\right)$ is defined as

$$\begin{aligned} \delta_t\left(V_i^{(k)}, u_1^{(k)}, u_2^{(k)}, T\right) = &\int_t^{t+T} x^T \bar{Q}_i^{(k)} x d\tau + V_i^{(k)}(x_{t+T}) \\ &-V_i^{(k)}(x_t), \end{aligned} \tag{15}$$

where $\bar{Q}_i^{(k)} = Q_i + \left(K_1^{(k)}\right)^T R_{i1} K_1^{(k)} + \left(K_2^{(k)}\right)^T R_{i2} K_2^{(k)}$. Then update method of value function can be expressed as

$$V_i^{(k+1)}(x_t) = V_i^{(k)}(x_t) + \eta_i \delta_t\left(V_i^{(k)}, u_1^{(k)}, u_2^{(k)}, T\right). \tag{16}$$

The policy update can be further determined as

$$K_i^{(k+1)} = R_{ii}^{-1} B_i^T P_i^{(k+1)} \qquad i = 1, 2. \tag{17}$$

In the next, the least squares (LS) method is employed to implement the integral TD algorithm. To describe the LS method, we introduce the concept of the Kronecker product [26] as follows

$$V_i^{(k)}(x_t) = x_t^T P_i^{(k)} x_t = \left(x_t^T \otimes x_t^T\right) vec\left(P_i^{(k)}\right)$$

where

$$\begin{aligned} x_t^T \otimes x_t^T = [\ &x_1 x_1 \quad x_1 x_2 \quad \dots \quad x_2 x_1 \\ &\dots \quad x_n x_{n-1} \quad x_n x_n\ ], \end{aligned}$$

$$\begin{aligned} vec\left(P_i^{(k)}\right) = \Big[\ &P_i^{(k)}(1,1) \quad P_i^{(k)}(1,2) \quad \dots \quad P_i^{(k)}(2,1) \\ &\dots \quad P_i^{(k)}(n, n-1) \quad P_i^{(k)}(n,n)\ \Big]^T. \end{aligned}$$

Then, update rule (16) can be expressed as:

$$\begin{aligned} &\left(x_t^T \otimes x_t^T\right) vec\left(P_i^{(k+1)}\right) = V_i^{(k)}(x_t) \\ &+\eta_i \delta_t\left(V_i^{(k)}, u_1^{(k)}, u_2^{(k)}, T\right) \end{aligned} \tag{18}$$

Therefore, the update rule (16) can be rewritten as

$$\left(x_t^T \otimes x_t^T\right) vec\left(P_i^{(k+1)}\right) = d_i \tag{19}$$

with

$$\begin{aligned} d_i = &V_i^{(k)}(x_t) + \eta_i \Bigg(V_i^{(k)}(x_{t+T}) - V_i^{(k)}(x_t) \\ &+ \int_t^{t+T} x^T \bar{Q}_i^{(k)} x d\tau \Bigg) \end{aligned}$$

To ensure the uniqueness and existence of solutions in (19), the condition that $N \geqslant n^2$ is satisfied during the LS method.

B. Equivalent integral TD learning

In this subsection, we give an equivalent formulation with a compact form of the integral TD learning algorithm developed in the previous subsection.

Before moving on, inserting (15) into (16), one has

$$\begin{aligned} x_t^T P_i^{(k+1)} x_t = &\eta_i \left[\int_t^{t+T} x^T \bar{Q}_i^{(k)} x d\tau + x_{t+T}^T P_i^{(k)} x_{t+T}\right] \\ &+ (1 - \eta_i) x_t^T P_i^{(k)} x_t \end{aligned} \tag{20}$$

Algorithm 2 Online Integral TD Learning Algorithm

```
1: Let $k = 0$. Start with a pair of initial matrices $\left(P_1^{(0)}, P_2^{(0)}\right)$ such that (1) is stable and select a suitable T.
2: For $k \geqslant 0$, at first, collect $N$ sample state data, then use the LS method to solve matrices $P_1^{(k+1)}$, $P_2^{(k+1)}$ that satisfy (19).
3: Update the control polices such that
   $u_i^{(k+1)}(x) = -K_i^{k+1}x = -R_{ii}^{-1}{B_i}^T P_i^{(k+1)}x \quad i = 1, 2.$
4: Stop the value function update if the following criterion is satisfied for a specified value of $\varepsilon$:
   $max\left(\left\|P_1^{(k+1)} - P_1^{(k)}\right\|, \left\|P_2^{(k+1)} - P_2^{(k)}\right\|\right) \leqslant \varepsilon$
   otherwise, set $k = k + 1$ and go to step 2.
```

Consider the system (1) with the feedback control $u_i = -k_i^{(k)}x$, one can obtain $x_\tau = e^{\bar{A}^{(k)}(\tau - t)}x_t$. Then, inserting x_τ into (20) yields

$$
\begin{aligned}
P_i^{(k+1)} &= (1 - \eta_i) P_i^{(k)} + \eta_i \int_0^T e^{\left(\bar{A}^{(k)}\right)^T t} \bar{Q}_i^{(k)} e^{\bar{A}^{(k)}t} dt \\
&\quad + \eta_i e^{\left(\bar{A}^{(k)}\right)^T T} P_i^{(k)} e^{\bar{A}^{(k)}T} \\
&= P_i^{(k)} + \eta_i \int_0^T e^{\left(\bar{A}^{(k)}\right)^T t} \bar{Q}_i^{(k)} e^{\bar{A}^{(k)}t} dt \\
&\quad + \eta_i \int_0^T \frac{d}{dt}\left(e^{\left(\bar{A}^{(k)}\right)^T t} P_i^{(k)} e^{\bar{A}^{(k)}t}\right) dt \\
&= P_i^{(k)} + \eta_i \int_0^T e^{\left(\bar{A}^{(k)}\right)^T t} Ric_i\left(P_1^{(k)}, P_2^{(k)}\right) e^{\bar{A}^{(k)}t} dt
\end{aligned}
\tag{21}
$$

where $Ric_i\left(P_1^{(k)}, P_2^{(k)}\right) = \left(\bar{A}^{(k)}\right)^T P_i^{(k)} + P_i^{(k)}\bar{A}^{(k)} + \bar{Q}_i^{(k)}$.

Algorithm 3 Compact Form of Integral TD Algorithm

```
1: Start with initial matrices $\left(P_1^{(0)}, P_2^{(0)}\right)$ such that the closed-loop system is stable and select a suitable T.
2: Value Update: solve (21) for $P_1^{(k+1)}$, $P_2^{(k+1)}$.
3: Stop the equivalent algorithm when the following criterion is satisfied for a specified value of $\varepsilon$:
   $max\left(\left\|P_1^{(k+1)} - P_1^{(k)}\right\|, \left\|P_2^{(k+1)} - P_2^{(k)}\right\|\right) \leqslant \varepsilon.$
   Otherwise, set $k = k + 1$ and go to step 2.
```

In algorithm 3, one can know that the update of $P_i^{(k+1)}$ only depends on $P_i^{(k)}$ from (21). That is, the compact form of integral TD algorithm is essentially a one-step update iteration algorithm.

C. Stability Discussion

In this subsection, the stability analysis of the closed loop system (1) will be given.

Lemma 2: For a symmetric matrix $G \in M^{n\times n}$, and any nonzero matrices $\mathcal{N}_1 \in C^{n\times n}, \mathcal{N}_2 \in C^{n\times n}, \mathcal{M}_1 \in C^{n\times n}, \mathcal{M}_2 \in C^{n\times n}$, it follows that $G + \mathcal{N}_1\mathcal{N}_2 + \mathcal{N}_2^T\mathcal{N}_1^T + \mathcal{M}_1\mathcal{M}_2 + \mathcal{M}_2^T\mathcal{M}_1^T < 0$ if there exists a constant $\varepsilon > 0$ such that $G + \varepsilon\mathcal{N}_1\mathcal{N}_1^T + \varepsilon^{-1}\mathcal{N}_2^T\mathcal{N}_2 + \varepsilon\mathcal{M}_1\mathcal{M}_1^T + \varepsilon^{-1}\mathcal{M}_2^T\mathcal{M}_2 < 0$ holds.

Proof. The proof of the this lemma follows from that of [27, Lemma 2.4] and is omitted here. ■

The next theorem discusses the stability of the closed-loop system when applying the integral TD learning algorithm.

Theorem 1: Assume that $\bar{A}^{(0)}$ is Hurwitz. Let $Y_i^k, i = 1, 2$ be the solution of Lyapunov equation $\left(\bar{A}^{(0)}\right)^T Y_i^k + Y_i^k \bar{A}^{(0)} = -I$. If η_i satisfies (22) for each $k \in \mathbb{Z}_+$, then $\bar{A}^{(k)}$ is Hurwitz for all $k \in \mathbb{Z}_+$.

$$
\begin{aligned}
&0 < \eta_{\max} < 1/2 \times \\
&\frac{1}{\sqrt{\left(\left\|H_1 Y_i^{(k)}\right\|^2 + \left\|H_2 Y_i^{(k)}\right\|^2\right)\left(\left\|M_1^{(k)}\right\|^2 + \left\|M_2^{(k)}\right\|^2\right)}}
\end{aligned}
\tag{22}
$$

where $M_i^{(k)} = \int_0^T e^{\left(\bar{A}^{(k)}\right)^T t} Ric_i\left(P_1^{(k)}, P_2^{(k)}\right) e^{\bar{A}^{(k)}t} dt$, $Ric_i\left(P_1^{(k)}, P_2^{(k)}\right) = \left(\left(\bar{A}^{(k)}\right)^T P_i^{(k)} + P_i^{(k)}\bar{A}^{(k)} + \bar{Q}_i^{(k)}\right)$, $\eta_{\max} = \max\{\eta_1, \eta_2\}$ and $H_i = B_i R_{ii}^{-1} B_i^T$, $i = 1, 2$.

Proof. We will prove this theorem by deduction. First, suppose that $\bar{A}^{(0)}$ is Hurwitz. Suppose also that $\bar{A}^{(k)}$ is Hurwitz. Then, there exists a positive definite matrix denoted by $Y_i^k \in C_p^{n\times n}$ such that $\left(\bar{A}^{(k)}\right)^T Y_i^{(k)} + Y_i^{(k)}\bar{A}^{(k)} = -I$. Next, we need to show the Hurwitzness of the matrix $\bar{A}^{(k+1)}$. In the follows, we will find the sufficient condition $\left(\bar{A}^{(k+1)}\right)^T Y_i^{(k)} + Y_i^{(k)}\bar{A}^{(k+1)} < 0$ that guarantees the Hurwitzness of the matrix $\bar{A}^{(k+1)}$.

Rewriting $\bar{A}^{(k+1)}$ using the fact that $P_i^{(k+1)} = P_i^{(k)} + \eta_i M_i^{(k)}$ in (21) yields

$$
\begin{aligned}
\bar{A}^{(k+1)} &= A - B_1 K_1^{(k+1)} - B_2 K_2^{(k+1)} \\
&= \bar{A}^{(k)} - \eta_1 H_1 M_1^{(k)} - \eta_2 H_2 M_2^{(k)}
\end{aligned}
\tag{23}
$$

Based on (23), $\left(\bar{A}^{(k+1)}\right)^T Y_i^{(k)} + Y_i^{(k)}\bar{A}^{(k+1)} < 0$ can be rearranged as:

$$
\begin{aligned}
&-I - \left(\eta_1 H_1 M_1^{(k)} + \eta_2 H_2 M_2^{(k)}\right)^T Y_i^{(k)} \\
&\quad - Y_i^{(k)}\left(\eta_1 H_1 M_1^{(k)} + \eta_2 H_2 M_2^{(k)}\right) < 0
\end{aligned}
\tag{24}
$$

Based on Lemma 2, the next inequality holds for any nonzero vector $x \in \mathbb{R}^n$

$$
\begin{aligned}
&\varepsilon_i^2\left(\left\|H_1 Y_i^{(k)} x\right\|^2 + \left\|H_2 Y_i^{(k)} x\right\|^2\right) - \varepsilon_i\|x\|^2 \\
&+ \left(\eta_1^2\left\|M_1^{(k)} x\right\|^2 + \eta_2^2\left\|M_2^{(k)} x\right\|^2\right) < 0
\end{aligned}
\tag{25}
$$

Because $\left\| \left(H_1 + H_2 \right) Y_i^{(k)} x \right\| > 0$, (25) is obviously quadratic in ε_i. In this case, the existence condition for $\varepsilon_i \in \mathbb{R}$ can be obtained by solving $D_i > 0$

$$
\begin{aligned}
D_i = &\|x\|^4 - 4\left(\eta_1^2 \left\| H_1 Y_i^{(k)} x \right\|^2 + \eta_2^2 \left\| H_2 Y_i^{(k)} x \right\|^2 \right) \times \\
&\left(\left\| M_1^{(k)} x \right\|^2 + \left\| M_2^{(k)} x \right\|^2 \right) \\
\geqslant &\|x\|^4 - 4\eta_{\max}^2 \left(\left\| H_1 Y_i^{(k)} x \right\|^2 + \left\| H_2 Y_i^{(k)} x \right\|^2 \right) \times \\
&\left(\left\| M_1^{(k)} x \right\|^2 + \left\| M_2^{(k)} x \right\|^2 \right) > 0
\end{aligned}
$$

that is, (22) holds. Note that, (22) ensures the existence of $\varepsilon_i > 0$ in (25). Thus, the above analysis guarantees the fact that $\bar{A}^{(k+1)}$ is Hurwitz. This completes the proof. ■

V. Simulation Study

In this section, we show the efficacy of the proposed integral TD method using the simulation with forth-order systems. Consider the following two-player continuous-time linear system [25]:

$$
A = \begin{bmatrix} -0.0366 & 0.0271 & 0.0188 & -0.4555 \\ 0.0482 & -1.0100 & 0.0024 & -4.0208 \\ 0.1002 & 0.2855 & -0.7070 & 1.3229 \\ 0 & 0 & 1 & 0 \end{bmatrix},
$$

$$
B_1 = \begin{bmatrix} 0.4422 & 3.0447 & -5.52 & 0 \end{bmatrix}^T,
$$

$$
B_2 = \begin{bmatrix} 0.1761 & -7.5922 & 4.99 & 0 \end{bmatrix}^T.
$$

Define

$$
\begin{aligned}
V_1 &= \int_0^\infty \left(x^T Q_1 x + u_1^T R_{11} u_1 + u_2^T R_{12} u_2 \right) dt \\
V_2 &= \int_0^\infty \left(x^T Q_2 x + u_1^T R_{21} u_1 + u_2^T R_{22} u_2 \right) dt
\end{aligned}
$$

where $Q_1 = diag\left([3.5, 2, 4, 5]\right)$, $R_{11} = 1$, $R_{12} = 0.25$ and $Q_2 = diag\left([1.5, 6, 3, 1]\right)$, $R_{21} = 0.6$, $R_{22} = 2$, $\eta_1 = 0.7$, $\eta_2 = 0.9$.

The initial state is selected as $x(0) = [0; 0; 0; 1]$ and the initial matrices $P_1^{(0)}$ and $P_2^{(0)}$ are selected as zero matrices. The data information of the system is collected at intervals of 0.5 s. After a set of 15 data samples is acquired, that is, 7.5 s, a least squares solution is performed, the iterative method stops until it is satisfied that ϵ less than 10^{-8}. After 20 steps, $P_1^{(k)}$ and $P_2^{(k)}$ for player one and two are stable as shown in figure 1 and 2. As is shown in figure 3 and 4, after 20 steps, both $\left\| P_i^{(k)} - P_i^* \right\|$ and Riccati operator are close to zero. Therefore, the integral TD method converges to the solution of the coupled AREs.

VI. Conclusions

In this paper, an integral temporal difference learning method is proposed to find the Nash equilibrium of two-player non-zero-sum games in an online manner. Only partial knowledge of system dynamics is required for the integral TD learning. The sufficient condition that guarantees the closed-loop stability during the iterative learning phase is discussed. Finally, the simulation study demonstrates the effectiveness of the presented algorithm in this paper.

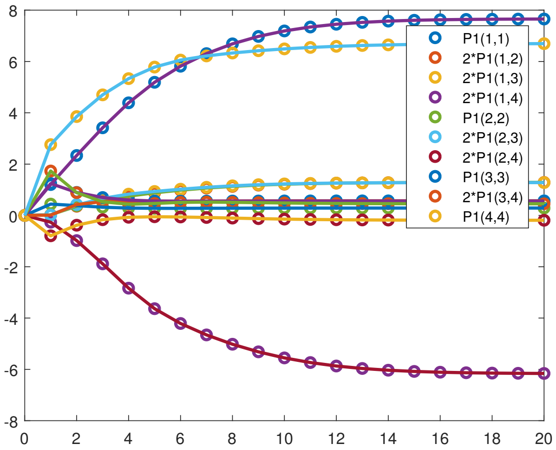

Fig. 1. Learning process of the elements in $P_1^{(k)}$ for player 1

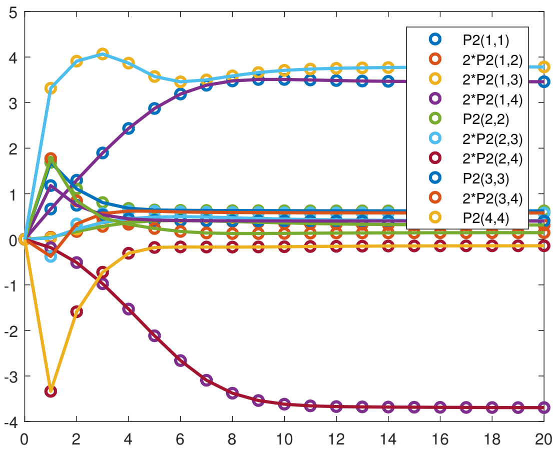

Fig. 2. Learning process of the elements in $P_2^{(k)}$ for player 2

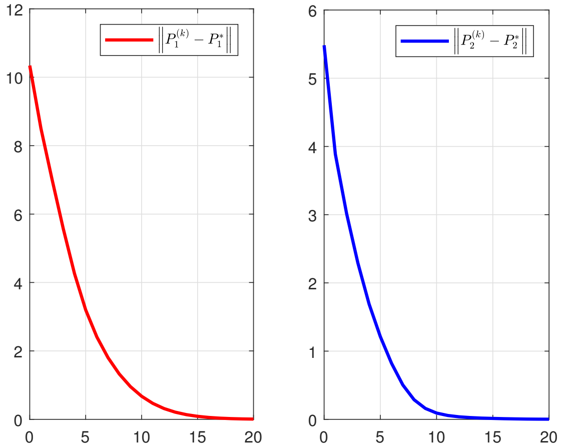

Fig. 3. Convergence of $P_1^{(k)}$ and $P_2^{(k)}$ to their optimal values P_1^* and P_2^* during the learning process

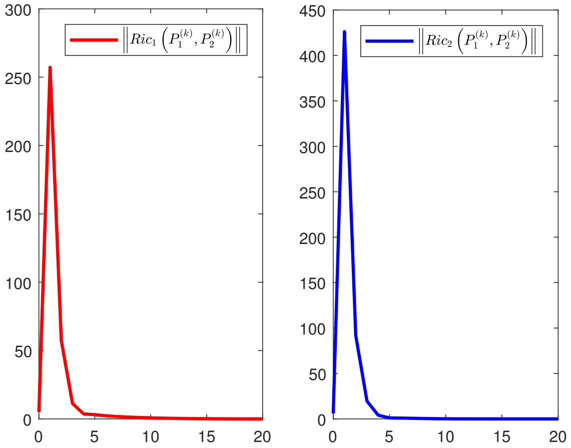

Fig. 4. Convergence of the Riccati operator of player 1 and 2 using the integral TD method for two-player NZS games

References

[1] A. W. Starr and Y. C. Ho, "Nonzero-sum differential games," *Journal of Optimization Theory and Applications*, vol. 3, no. 3, pp. 184–206, Mar 1969.

[2] H. Jiang, H. Zhang, G. Xiao, and X. Cui, "Data-based approximate optimal control for nonzero-sum games of multi-player systems using adaptive dynamic programming," *Neurocomputing*, vol. 275, pp. 192 – 199, 2018.

[3] A. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Transactions on Smart Grid*, vol. 1, no. 3, pp. 320–331, Dec 2010.

[4] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for discrete-time zero-sum games with application to H-infinity control," in *2007 European Control Conference (ECC)*, July 2007, pp. 1668–1675.

[5] L. Zhu and H. Yi, "Differential game analysis of manufacturer enterprise supplier relationship under dynamic market environment," in *2010 2nd IEEE International Conference on Information Management and Engineering*, April 2010, pp. 314–316.

[6] O. Petrosian, M. Nastych, and D. Volf, "Differential game of oil market with moving informational horizon and non-transferable utility," in *2017 Constructive Nonsmooth Analysis and Related Topics (dedicated to the memory of V.F. Demyanov) (CNSA)*, May 2017, pp. 1–4.

[7] L. Wu and J. M. Yang, "Load frequency control of area power system with multi-source power generation units based on differential games tracking control," in *2013 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, Dec 2013, pp. 1–6.

[8] W. Saad, Z. Han, M. Debbah, A. Hjorungnes, and T. Basar, "Coalitional game theory for communication networks," *IEEE Signal Processing Magazine*, vol. 26, no. 5, pp. 77–97, September 2009.

[9] Y. Yang, D. Wunsch, and Y. Yin, "Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 8, pp. 1929–1940, Aug 2017.

[10] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.

[11] R. Song, F. L. Lewis, and Q. Wei, "Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 704–713, March 2017.

[12] D. Zhao, Q. Zhang, D. Wang, and Y. Zhu, "Experience replay for optimal control of nonzero-sum game systems with unknown dynamics," *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 854–865, March 2016.

[13] D. Liu, H. Li, and D. Wang, "Online synchronous approximate optimal learning algorithm for multi-player non-zero-sum games with unknown dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 8, pp. 1015–1027, Aug 2014.

[14] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193 – 202, 2014.

[15] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Leader-follower output synchronization of linear heterogeneous systems with active leader using reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2139–2153, June 2018.

[16] Y. Yang and H. Modares and D. C. Wunsch and Y. Yin, "Optimal containment control of unknown heterogeneous systems with active leaders," *IEEE Transactions on Control Systems Technology*, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8277155

[17] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[18] B. Luo, H.-N. Wu, T. Huang, and D. Liu, "Reinforcement learning solution for HJB equation arising in constrained optimal control problem," *Neural Networks*, vol. 71, pp. 150 – 158, 2015.

[19] H. Modares, F. L. Lewis, and Z. Jiang, "H_∞tracking control of completely unknown continuous-time systems via off-policy reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2550–2562, Oct 2015.

[20] H. Zhang, L. Cui, X. Zhang, and Y. Luo, "Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2226–2236, Dec 2011.

[21] Y. Yang, Z. Guo, H. Xiong, D. Ding, Y. Yin, and D. C. Wunsch, "Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8657370

[22] Y. Yang, K. G. Vamvoudakis, H. Ferraz, and H. Modares, "Dynamic intermittent Q-learning-based model-free suboptimal co-design of $\mathcal{L}_2$-stabilization," *International Journal of Robust and Nonlinear Control*, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rnc.4515

[23] Q. Zhang and D. Zhao, "Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics," *IEEE Transactions on Cybernetics*, pp. 1–12, 2018.

[24] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477 – 484, 2009.

[25] D. Vrabie and F. Lewis, "Integral reinforcement learning for online computation of feedback Nash strategies of nonzero-sum differential games," in *49th IEEE Conference on Decision and Control (CDC)*, Dec 2010, pp. 3066–3071.

[26] T. Y. Chun, J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral temporal difference learning for continuous-time linear quadratic regulations," *International Journal of Control, Automation and Systems*, vol. 15, no. 1, pp. 226–238, Feb 2017.

[27] L. Xie, "Output feedback $H\infty$ control of systems with parameter uncertainty," *International Journal of control*, vol. 63, no. 4, pp. 741–750, 1996.