01 Jun 2019

Data-Driven Integral Reinforcement Learning for Continuous-Time Non-Zero-Sum Games

Yongliang Yang

Liming Wang

Hamidreza Modares
Missouri University of Science and Technology, modaresh@mst.edu

Dawei Ding

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/ele_comeng_facwork/3837

Recommended Citation

Y. Yang et al., "Data-Driven Integral Reinforcement Learning for Continuous-Time Non-Zero-Sum Games," *IEEE Access*, vol. 7, pp. 82901-82912, Institute of Electrical and Electronics Engineers (IEEE), Jun 2019. The definitive version is available at https://doi.org/10.1109/ACCESS.2019.2923845

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 18, 2019, accepted June 12, 2019, date of publication June 19, 2019, date of current version July 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2923845

Data-Driven Integral Reinforcement Learning for Continuous-Time Non-Zero-Sum Games

YONGLIANG YANG[1,2], (Member, IEEE), LIMING WANG[1,2], HAMIDREZA MODARES[3], (Member, IEEE), DAWEI DING[1,2], YIXIN YIN[1,2], (Member, IEEE), AND DONALD WUNSCH[4], (Fellow, IEEE)

[1]School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2]Key Laboratory of Knowledge Automation for Industrial Processes, Ministry of Education, University of Science and Technology Beijing, Beijing 100083, China
[3]Department of Mechanical Engineering, Michigan State University, East Lansing, MI 48824, USA
[4]Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, MO 65401, USA

Corresponding author: Dawei Ding (ddaweiauto@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61873028 and 61333002, in part by the China Postdoctoral Science Foundation under Grant 2018M641197, in part by the Fundamental Research Funds for the China Central Universities of USTB under Grant FRF-TP-18-031A1, FRF-BD-17-002A, and FRF-GF-17-B48, 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 the DARPA/Microsystems Technology Office, and in part by the Army Research Laboratory through a Cooperative Agreement under Grant W911NF-18-2-0260.

ABSTRACT This paper develops an integral value iteration (VI) method to efficiently find online the Nash equilibrium solution of two-player non-zero-sum (NZS) differential games for linear systems with partially unknown dynamics. To guarantee the closed-loop stability about the Nash equilibrium, the explicit upper bound for the discounted factor is given. To show the efficacy of the presented online model-free solution, the integral VI method is compared with the model-based off-line policy iteration method. Moreover, the theoretical analysis of the integral VI algorithm in terms of three aspects, i.e., positive definiteness properties of the updated cost functions, the stability of the closed-loop systems, and the conditions that guarantee the monotone convergence, is provided in detail. Finally, the simulation results demonstrate the efficacy of the presented algorithms.

INDEX TERMS Coupled Riccati equations, integral reinforcement learning, non-zero-sum games, optimal control.

I. INTRODUCTION

Game theory is a powerful and natural framework to represent the interactions among multiple players, where each player seeks to maximize its own interest. Game theory has been widely and successfully used in variety of engineering sectors, including, power systems [1], transportation [2], and control systems [3]. In zero-sum (ZS) games, which are strictly competitive games, each player's gain or loss is exactly balanced by others. In contrast, non-zero-sum (NZS) games can take into account both individual self-interests, as well as global group interest, such as mixed H_2/H_∞ control [4], etc. In this paper, NZS games with two players for continuous-time linear systems are investigated.

Differential games, for which the states of agents evolve based on differential dynamic equations, have been originally introduced in [5]. In ZS differential game theory, the Nash equilibrium seeking results in solving coupled Hamilton-Jacobi equations (HJEs) [6]–[8]. For the linear systems, the HJEs reduce to coupled algebraic Riccati equations (CAREs) [9], [10]. For NZS differential games, on the other hand, the Nash equilibrium solution is found by solving coupled HJEs for nonlinear systems and CAREs for linear systems [11], [12]. It is difficult or even impossible to obtain an analytical solution to coupled HJEs or CAREs. Many approaches are presented to approximate the solution to the CAREs, such as Newton's method, Lyapunov iteration [13], Riccati iteration [14], parallel synchronous algorithm [15] etc. However, these numerical methods are essentially off-line and require the complete knowledge of systems dynamics. In reality, however, this knowledge might not be available. Therefore, it is desired to develop an online method and obviate the complete model requirement.

Adaptive dynamic programming (ADP)/ reinforcement learning (RL) is a bio-inspired learning method trying to find the optimal policy that optimizes the cumulative reward [16]. RL has been widely used in the dynamic optimization

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaoli Luan.

applications, such as ZS games with two players [17], NZS games with multiple players [18], optimal regulation/tracking problem with only single-agent systems [19], [20], consensus problem of multi-agent systems [21]–[24], etc. Value iteration (VI) and policy iteration (PI) algorithms are typical ADP/RL methods to approximate the optimal value policy [25], [26]. In PI algorithm, the initial policy has to be admissible in order to guarantee the closed-loop stability in the iterative learning process [27]–[29]. In contrast, VI algorithm does not require an admissible initial policy [30]. However, the closed-loop stability of the VI algorithm in each iteration can not be guaranteed. In this paper, a novel integral VI method is developed to obviate the requirement of initial admissible policy while guaranteeing the closed-loop stability during the learning process.

Typical model-free ADP/RL methods are Q-learning [31]–[33] and off-policy RL [34]–[36]. In the Q-learning algorithm, the action-dependent value function representation is used to evaluate the action given the state. However, only convergence is considered for Q-learning algorithm in [32]. On the other hand, the off-policy RL method is equivalent to the model-based PI algorithm, which still requires the initial policy to be admissible. Therefore, it is desired to obviate the admissibility of initial policy while guaranteeing the closed-loop stability for model-free ADP/RL methods. The main contributions of this paper are summarized as follows:

1) A novel data-driven value iteration algorithm is developed for solving the NZS games for linear dynamical systems.
2) An explicit upper-bound for the discounted factor is given to ensure the asymptotic stability of closed-loop system with the Nash equilibrium. Moreover, it is shown that the undiscounted NZS games can be viewed as a special case of the discounted NZS games.
3) For the presented data-driven value iteration algorithm, theoretical analysis is discussed in terms of the positive-definiteness of the iterative value function, the closed-loop stability and the convergence to the optimal case.

The rest of this paper is organized as follows. In Section II, the problem formulations with preliminaries are presented. It is shown that the coupled AREs are sufficient and necessary to the Nash equilibrium. In Section III, an integral VI algorithm and its equivalent form are considered. In Section IV, the positive definiteness of the updated cost functions, the stability discussions concerning the closed-loop systems and the conditions that guarantee the monotone convergence are proven. In Section V, examples are given to demonstrate the effectiveness of the proposed algorithm. Finally, the conclusion is made in Section VI.

II. PROBLEM FORMULATION

We consider the continuous-time linear dynamical systems

$$\dot{x}(t) = Ax + B_1u_1 + B_2u_2, x(0) = x_0, \tag{1}$$

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

Assumption 1: The matrix pair $\left(A, \begin{bmatrix} B_1 & B_2 \end{bmatrix}\right)$ *is stabilizable.* □

For each player, the NZS differential games on an infinite time horizon aim to minimize the following discounted cost function defined as

$$\begin{aligned} V_1(x_0) &= \int_0^\infty e^{-2\alpha_1\tau}\left(x^TQ_1x + u_1^TR_{11}u_1 + u_2^TR_{12}u_2\right)d\tau \\ V_2(x_0) &= \int_0^\infty e^{-2\alpha_2\tau}\left(x^TQ_2x + u_1^TR_{21}u_1 + u_2^TR_{22}u_2\right)d\tau \end{aligned} \tag{2}$$

where $Q_1 \geq 0, Q_2 \geq 0, R_{11} > 0, R_{12} \geq 0, R_{21} \geq 0, R_{22} > 0$ are penalty functions for players one and two, and $\alpha_1 > 0, \alpha_2 > 0$ are discount factors for players one and two, respectively. As shown later, the non-zero discounted factor is given to ensure the asymptotic stability of closed-loop system with the Nash equilibrium.

The following definitions are required for subsequent discussions.

Definition 1 (Admissible Control): Feedback control policy pair $\mu = \{u_1, u_2\}$ *is said to be admissible with respect to the performance* (2) *on a compact set* $\Omega \in R^n$, *denoted as* $\mu \in \psi(\Omega)$, *if* $\mu = \{u_1, u_2\}$ *is continuous on* Ω, μ *stabilizes (1) on* Ω, $u_i(0) = 0$ *for* $i = 1, 2$, *and the performance functions* $V_i(x_0)$ *in* (2) *take finite values for* $\forall x_0 \in \Omega$. □

Definition 2 (Nash Equilibrium Strategies): A two-tuple of strategies $\mu^* = \left\{u_1^*, u_2^*\right\}$ *with* $\mu^* \in \psi(\Omega)$, $i = 1, 2$ *is said to constitute a Nash equilibrium solution for a two-player finite games in extensive form, if the following two inequalities are satisfied for* $i = 1, 2$*:*

$$\begin{aligned} V_1\left(x(0); u_1^*, u_2^*\right) &\leq V_1\left(x(0); u_1, u_2^*\right), \ \forall u_1 \\ V_2\left(x(0); u_1^*, u_2^*\right) &\leq V_2\left(x(0); u_1^*, u_2\right), \ \forall u_2. \end{aligned} \tag{3}$$

The two-tuple of quantities $\left\{u_1^*, u_2^*\right\}$ *is known as a Nash equilibrium outcome of the two-player games.* □

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

Problem 1 (Discounted Two-Player NZS Games): For the two players in system (1), *find the Nash equilibrium strategies* (u_1^*, u_2^*) *with respect to the cost functions defined in* (2). □

A. COUPLED ALGEBRAIC RICCATI EQUATIONS FOR DISCOUNTED NZS GAMES

In this subsection, the sufficient and necessary conditions of the Nash equilibrium of Problem 1, named the coupled algebraic Riccati equations, are introduced.

Lemma 1 [6]: Under Assumption 1, consider the system (1) *with the performance functions defined by* (2). *Then,* (K_1^*, K_2^*), *defined as*

$$K_i^* = R_{ii}^{-1}B_i^TP_i^*, i = 1, 2,$$

is a feedback Nash equilibrium if and only if (P_1^*, P_2^*) *is a symmetric stabilizing solution of the CAREs* (4) *and* (5)*, as shown at the bottom of this page, with*

$$H_1 = B_1 R_{11}^{-1} B_1^T, H_2 = B_2 R_{22}^{-1} B_2^T$$

$$H_3 = B_2 R_{22}^{-1} R_{12} R_{22}^{-1} B_2^T, H_4 = B_1 R_{11}^{-1} R_{21} R_{11}^{-1} B_1^T. \square$$

Definition 3 (Riccati Operator): For each player, Riccati operator $Ric_{\alpha_i}(X_1, X_2)$ *is defined as in* (6) *and* (7)*, as shown at the bottom of this page.* □

Remark 1: The Riccati operator Ric has an important role in evaluating policy for each player with respect to the performance defined by (2)*.*

1) *If* $Ric_{\alpha_i}\left(P_1^*, P_2^*\right) = 0$*, then the performance indices* (2) *are minimized and both players in system* (1) *has reached the Nash equilibrium.*
2) *If* $0 < Ric_{\alpha_i}\left(P_1^{(k+1)}, P_2^{(k+1)}\right) < Ric_{\alpha_i}\left(P_1^{(k)}, P_2^{(k)}\right)$ *holds, then the performance of step* $k+1$ *is closer to the optimal solution than step* k*.* □

B. OFFLINE POLICY ITERATION ALGORITHM

For iterative ADP algorithm, the optimal feedback gain is obtained by successive approximation. In k-*th* iteration, we denote the admissible policy for the player i as $u_i^{(k)} = -K_i^{(k)} x$, $i = 1, 2$. Then, the corresponding discounted Bellman equation can be written as [37]:

$$\dot{V}_1^{(k)}(x_t) - 2\alpha_1 V_1^{(k)} + r_1\left(x_t, u_1^{(k)}, u_2^{(k)}\right) = 0, \tag{8}$$

$$\dot{V}_2^{(k)}(x_t) - 2\alpha_2 V_2^{(k)} + r_2\left(x_t, u_1^{(k)}, u_2^{(k)}\right) = 0, \tag{9}$$

where

$$V_i^{(k)}(x_t) = x_t^T P_i^{(k)} x_t,$$
$$r_i\left(x_t, u_1^{(k)}, u_2^{(k)}\right) = \left(u_1^{(k)}\right)^T R_{i1} u_1^{(k)} + \left(u_2^{(k)}\right)^T R_{i2} u_2^{(k)} + x^T Q_i x \tag{10}$$

Denote

$$\bar{A}^{(k)} = A - B_1 K_1^{(k)} - B_2 K_2^{(k)}$$
$$\bar{A}_{\alpha_i}^{(k)} = \bar{A}^{(k)} - \alpha_i I$$

Then, the Bellman equations (8) and(9) can be equivalently written as the following Lyapunov equations,

$$\left(\bar{A}_{\alpha_1}^{(k)}\right)^T P_1^{(k)} + P_1^{(k)} \bar{A}_{\alpha_1}^{(k)} + Q_1 + \left(K_1^{(k)}\right)^T R_{11} K_1^{(k)} + \left(K_2^{(k)}\right)^T R_{12} K_2^{(k)} = 0, \tag{11}$$

$$\left(\bar{A}_{\alpha_2}^{(k)}\right)^T P_2^{(k)} + P_2^{(k)} \bar{A}_{\alpha_2}^{(k)} + Q_2 + \left(K_1^{(k)}\right)^T R_{21} K_1^{(k)} + \left(K_2^{(k)}\right)^T R_{22} K_2^{(k)} = 0. \tag{12}$$

The PI algorithm has been successfully used to solve the HJE and ARE in optimal control theory [20], [37]. Here, the PI algorithm is extended to approximate the solution to the CAREs iteratively. The closed-loop dynamics with $K_i^{(k)} = R_{ii}^{-1} B_i^T P_i^{(k)}$ can be written as $\dot{x}(t) = \bar{A}^{(k)} x(t)$. Then, the following offline PI algorithm can be presented to find the solution to the CAREs (4) and (5).

Algorithm 1 Offline Policy Iteration Algorithm

1: Given initial admissible control gain $K_1^{(0)}, K_2^{(0)}$, such that the system (1) is a stable closed-loop system.
2: Policy Evaluation: solve (11) and (12) for $P_1^{(k)}, P_2^{(k)}$.
3: Policy Improvement: update the control policy gain as,

$$K_1^{(k+1)} = R_{11}^{-1} B_1^T P_1^{(k)} \tag{13}$$
$$K_2^{(k+1)} = R_{22}^{-1} B_2^T P_2^{(k)} \tag{14}$$

4: Stop at convergence, otherwise set $k = k + 1$ and go to step 2

Remark 2: As shown in [8], the convergence of Algorithm 1 to the solution of the CAREs (4) *and* (5)*, and the closed-loop stability of the iterative control policy for each player can be guaranteed.* □

Note that in Algorithm 1, the solution to the CAREs (4) and (5) is obtained offline, and it requires complete knowledge of the system dynamics (1). In the subsequent sections, an online integral VI algorithm is developed to solve the CAREs (4) and (5) with only partial knowledge of the system dynamics. In addition, the initial policy has to be admissible in order to guarantee the closed-loop stability in each iteration. In the following, this requirement can also be relaxed.

III. INTEGRAL VI ALGORITHM

In this section, a novel integral VI algorithm is developed to solve Problem 1.

A. INTEGRAL VI ALGORITHM

Consider the system (1) with the performance functions (2), a novel equivalent representation with a stabilizing policy u_i

$$0 = -2\alpha_1 P_1^* + A^T P_1^* + P_1^* A + Q_1 - P_2^* H_2 P_1^* - P_1^* H_2 P_2^* - P_1^* H_1 P_1^* + P_2^* H_3 P_2^*, \tag{4}$$
$$0 = -2\alpha_2 P_2^* + A^T P_2^* + P_2^* A + Q_2 - P_1^* H_1 P_2^* - P_2^* H_1 P_1^* - P_2^* H_2 P_2^* + P_1^* H_4 P_1^*. \tag{5}$$
$$Ric_{\alpha_1}(X_1, X_2) = A^T X_1 + X_1 A + Q_1 - 2\alpha_1 X_1 - X_2 H_2 X_1 - X_1 H_2 X_2 - X_1 H_1 X_1 + X_2 H_3 X_2, \tag{6}$$
$$Ric_{\alpha_2}(X_1, X_2) = A^T X_2 + X_2 A + Q_2 - 2\alpha_2 X_2 - X_1 H_1 X_2 - X_2 H_1 X_1 - X_2 H_2 X_2 + X_1 H_4 X_1. \tag{7}$$

can be described as

$$V_i(x_t) = e^{-2\alpha_i T} V_i(x_{t+T}) + \int_t^{t+T} e^{-2\alpha_i(\tau-t)} \times \left(x^T Q_i x + u_1^T R_{i1} u_1 + u_2^T R_{i2} u_2\right) d\tau. \quad (15)$$

From (15), the integral temporal difference error for a given policy $u_i^{(k)}$ can be defined as

$$\delta_t\left(V_i^{(k)}, u_1^{(k)}, u_2^{(k)}, T\right) = \int_t^{t+T} e^{-2\alpha_i(\tau-t)} x^T \bar{Q}_i^{(k)} x d\tau + e^{-2\alpha_i T} V_i^{(k)}(x_{t+T}) - V_i^{(k)}(x_t), \quad (16)$$

where $\bar{Q}_i^{(k)} = Q_i + \left(K_1^{(k)}\right)^T R_{i1} K_1^{(k)} + \left(K_2^{(k)}\right)^T R_{i2} K_2^{(k)}$. To design the TD(0) algorithm, the value function update can be represented with the learning rates η_1 and η_2 as

$$V_1^{(k+1)}(x_t) = V_1^{(k)}(x_t) + \eta_1 \delta_t\left(V_1^{(k)}, u_1^{(k)}, u_2^{(k)}, T\right),$$
$$V_2^{(k+1)}(x_t) = V_2^{(k)}(x_t) + \eta_2 \delta_t\left(V_2^{(k)}, u_1^{(k)}, u_2^{(k)}, T\right). \quad (17)$$

The learning rates η_1 and η_2 should be properly designed to guarantee the closed-loop stability and the convergence for the learning process, as discussed later in Section IV.

The next control policy gain is designed by

$$K_i^{(k+1)} = R_{ii}^{-1} B_i^T P_i^{(k+1)} \quad i = 1, 2 \quad (18)$$

Note that the value function $V_i^{(k+1)}(x_t)$ is quadratic in it argument x_t. Then, $V_i^{(k+1)}(x_t)$ can be parameterized as

$$V_i^{(k+1)}(x_t) = x_t^T P_i^{(k+1)} x_t = \left(\bar{P}_i^{(k+1)}\right)^T \bar{x}_t,$$

where $\bar{x}_t \in \mathbb{R}^{n(n+1)/2}$ represents a column vector

$$\bar{x}_t = \begin{bmatrix} x_1 x_1 & x_1 x_2 & x_1 x_3 & \dots & x_{n-1} x_n & x_n x_n \end{bmatrix}^T,$$

and $\bar{P}_i^{(k+1)} \in \mathbb{R}^{n(n+1)/2}$ represents a column vector

$$\bar{P}_i^{(k+1)} = \begin{bmatrix} P_i^{(k+1)}(1,1) & 2P_i^{(k+1)}(1,2) & 2P_i^{(k+1)}(1,3) & \cdots & 2P_i^{(k+1)}(n,n-1) & P_i^{(k+1)}(n,n) \end{bmatrix}^T.$$

Then, update rule (17) can be expressed as:

$$\bar{x}_t^T \bar{P}_i^{(k+1)} = V_i^{(k)}(x_t) + \eta_i \delta_t\left(V_i^{(k)}, u_1^{(k)}, u_2^{(k)}, T\right) \triangleq d_i^k \quad (19)$$

From the definition of δ_t in (16), the term d_i^k in (19) contains an integral term. Therefore, to solve d_i^k, the following additional dynamics (20) is introduced:

$$\dot{W}_i = 2\alpha_i W_i + x^T \bar{Q}_i^{(k)} x, \quad W_i(0) = 0, \quad (20)$$

during the simulation, note that initial state for (20) is reset to zero at each interval $(t, t+T)$. Then the integral term in (16) can be calculated as

$$e^{-2\alpha_i T} W_i(x_{t+T}) - W_i(x_t) = \int_t^{t+T} e^{-2\alpha_i(\tau-t)} x^T \bar{Q}_i^{(k)} x d\tau. \quad (21)$$

Inserting (21) into (16), one has an equivalent form,

$$\delta_t\left(V_i^{(k)}, u_1^{(k)}, u_2^{(k)}, T\right) = -\left(V_i^{(k)}(x_t) + W_i(x_t)\right) e^{-2\alpha_i T}\left(V_i^{(k)}(x_{t+T}) + W_i(x_{t+T})\right). \quad (22)$$

Then, the term d_i^k in (19) can be equivalently expressed as

$$d_i^k = V_i^{(k)}(x_t) + \eta_i e^{-2\alpha_i T}\left(V_i^{(k)}(x_{t+T}) + W_i(x_{t+T})\right) - \eta_i\left(V_i^{(k)}(x_t) + W_i(x_t)\right).$$

Therefore, the update rule (19) can be rewritten as:

$$\bar{x}_t^T \bar{P}_i^{(k+1)} = d_i^k, \quad (23)$$

Note that in *k-th* iteration, only the term $\bar{P}_i^{(k+1)}$ is unknown. Therefore, the least squares method is employed to solve $\bar{P}_i^{(k+1)}$ by collecting $N\left(\geq n(n+1)/2\right)$ sample points is to ensure the number of equations is greater than the number of unknowns for (23).

Remark 3: It is worthwhile to highlight the role of the discounted factor, α_i, in the discounted NZS games. In Algorithm 1, it is required that the matrix

$$\bar{A}^{(0)} = A - B_1 K_1^{(0)} - B_2 K_2^{(0)} \quad (24)$$

is Hurwitz. In contrast, in the integral VI algorithm, the initial policy does not need to be admissible. As given in Step 1 in Algorithm 2, it is only required that the matrix

$$\bar{A}_{\alpha_i}^{(0)} = A - H_1 P_1^{(0)} - H_2 P_2^{(0)} - \alpha_i I \quad (25)$$

is Hurwitz. A comparison between (24) and (25) indicates that the requirement of admissibility about the initial policy is no longer needed for the integral VI algorithm. □

Remark 4: Existing results on the NZS games are usually without discounted factors, such as the cases in [7]–[10]. In this paper, the discounted factor is allowed to be zero, i.e., the NZS games without discounted factors can be viewed as special cases in our formulation. □

In the optimal control theory, the discount factor in the performance has effects on the closed-loop stability, which is required to be within some certain range, as discussed in [19]. To guarantee the closed-loop stability, the bound of the discount factor α_i for the NZS games is discussed in the next theorem.

Theorem 1 (Upper Bound for the Discount Factor α_i): Consider the system (1), then the origin of system (1) is asymptotically stable if (26) or (27) holds.

$$\alpha_1 \leq \left\| (H_1 Q_1)^{1/2} \right\| \quad (26)$$

$$\alpha_2 \leq \left\| (H_2 Q_2)^{1/2} \right\| \quad (27)$$

Proof: Denote $\bar{A} = A - H_1 P_1^* - H_2 P_2^*$. Then, the CAREs (4) and (5) can be rewritten as:

$$\bar{A}^T P_1^* + P_1^* \bar{A} - 2\alpha_1 P_1^* + Q_1 + \left(P_1^*\right)^T H_1 P_1^* + \left(P_2^*\right)^T H_3 P_2^* = 0 \quad (28)$$

$$\bar{A}^T P_2^* + P_2^* \bar{A} - 2\alpha_2 P_2^* + Q_2 + \left(P_1^*\right)^T H_4 P_1^* + \left(P_2^*\right)^T H_2 P_2^* = 0 \quad (29)$$

Assume that λ is an eigenvalue of the closed-loop system $\bar{A}$, one has $\bar{A}x = \lambda x$. First, for the player one, multiplying both sides of (28) by nonzero vector x^T and x with $x \in R^n$, one can obtain [19]

$$2\left(Re\left(\lambda\right) - \alpha_1\right) x^T P_1^* x = -x^T \left(Q_1 + \left(P_1^*\right)^T H_1 P_1^* + \left(P_2^*\right)^T H_3 P_2^*\right) x. \quad (30)$$

Using the inequality $a^2 + b^2 \geq 2ab$ and since $P_1^* > 0$, (30) becomes

$$\begin{aligned} Re\left(\lambda\right) - \alpha_1 &\leq -\left\| \left(Q_1 \left(P_1^*\right)^{-1}\right)^{1/2} \right\| \\ &\quad \times \left\| \left(\left(P_1^*\right)^T H_1 P_1^* + \left(P_2^*\right)^T H_3 P_2^*\right)^{1/2} \left(P_1^*\right)^{-1/2} \right\|. \end{aligned} \quad (31)$$

To guarantee the stability of the closed-loop system, it is required that $Re(\lambda) < 0$, i.e.,

$$\begin{aligned} \alpha_1 &\leq \left\| \left(Q_1 \left(P_1^*\right)^{-1}\right)^{1/2} \right\| \\ &\quad \times \left\| \left(\left(P_1^*\right)^T H_1 P_1^* + \left(P_2^*\right)^T H_3 P_2^*\right)^{1/2} \left(P_1^*\right)^{-1/2} \right\|. \end{aligned} \quad (32)$$

Since $\left(P_2^*\right)^T H_3 P_2^* \geq 0$, (32) holds if the following inequality is satisfied,

$$\begin{aligned} \alpha_1 &\leq \left\| \left(Q_1 \left(P_1^*\right)^{-1}\right)^{1/2} \right\| \left\| \left[\left(P_1^*\right)^T H_1 P_1^*\right]^{1/2} \left(P_1^*\right)^{-1/2} \right\| \\ &\leq \left\| \left(Q_1 \left(P_1^*\right)^{-1}\right)^{1/2} \right\| \left\| \left(H_1 P_1^*\right)^{1/2} \right\| \\ &\leq \left\| {Q_1}^{1/2} \right\| \left\| {H_1}^{1/2} \right\| \end{aligned} \quad (33)$$

Using the fact that $\|A\|\,\|B\| \geq \|AB\|$, one can obtain the sufficient condition to (33) as given in (26). From the above analysis, it is shown that the condition (26) guarantees the asymptotic stability of the closed-loop system. Similarly, for the discounted factor α_2, one can obtain (27). This completes the proof. ■

B. EQUIVALENT INTEGRAL VI WITH DISCOUNT FACTOR

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

Consider the system (1) and feedback control $u_i^{(k)} = -k_i^{(k)} x$, we can obtain $x_\tau = e^{\bar{A}^{(k)}(\tau - t)} x_t$, substituting the equation into (35), the following equation can be obtained

$$\begin{aligned} P_i^{(k+1)} &= \left(1 - \eta_i\right) P_i^{(k)} + \eta_i \left(\int_0^T e^{\left(\bar{A}_{\alpha_i}^{(k)}\right)^T t} \bar{Q}_i^{(k)} e^{\bar{A}_{\alpha_i}^{(k)} t} dt \right. \\ &\quad \left. + e^{\left(\bar{A}_{\alpha_i}^{(k)}\right)^T T} P_i^{(k)} e^{\bar{A}_{\alpha_i}^{(k)} T} \right) \\ &= P_i^{(k)} + \eta_i \left(\int_0^T e^{\left(\bar{A}_{\alpha_i}^{(k)}\right)^T t} \bar{Q}_i^{(k)} e^{\bar{A}_{\alpha_i}^{(k)} t} dt \right. \\ &\quad \left. + \int_0^T \frac{d}{dt} \left(e^{\left(\bar{A}_{\alpha_i}^{(k)}\right)^T t} P_i^{(k)} e^{\bar{A}_{\alpha_i}^{(k)} t} \right) dt \right) \\ &= P_i^{(k)} \\ &\quad + \eta_i \int_0^T e^{\left(\bar{A}_{\alpha_i}^{(k)}\right)^T t} Ric_{\alpha_i}\left(P_1^{(k)}, P_2^{(k)}\right) e^{\bar{A}_{\alpha_i}^{(k)} t} dt \end{aligned} \quad (34)$$

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

Algorithm 2 Online integral VI Algorithm With Discount Factor

1: Let $k = 0$. Start with a pair of initial matrices $\left(P_1^{(0)}, P_2^{(0)}\right)$ such that $\bar{A}_{\alpha_i}^{(0)}$ is Hurwitz for $i = 1, 2$.
2: Update the control policy for player i such that
$$u_i^{(k)}(x) = -K_i^k x = -R_{ii}^{-1} B_i^T P_i^{(k)} x \quad i = 1, 2.$$
3: For $k \geq 0$, $k \in \mathbb{N}$, first, collect N sample state data, then use the LS method to solve (23) for $P_1^{(k+1)}$, $P_2^{(k+1)}$.
4: Stop the online algorithm when the following criterion is satisfied for a specified value of ε:
$$max\left(\left\|P_1^{(k+1)} - P_1^{(k)}\right\|, \left\|P_2^{(k+1)} - P_2^{(k)}\right\|\right) \leq \varepsilon.$$
Otherwise, set $k = k + 1$ and go to step 2.

Algorithm 3 Equivalent Integral VI Algorithm With Discount Factor

1: Start with initial matrices $\left(P_1^{(0)}, P_2^{(0)}\right)$ and select a suitable T.
2: Value Update: solve (34) for $P_1^{(k+1)}$, $P_2^{(k+1)}$.
3: Stop until the following criterion is satisfied for a specified threshold ε:
$$max\left(\left\|P_1^{(k+1)} - P_1^{(k)}\right\|, \left\|P_2^{(k+1)} - P_2^{(k)}\right\|\right) \leq \varepsilon.$$
Otherwise, set $k = k + 1$ and go to step 2.

Remark 5: Algorithms 2 and 3 are equivalent to each other. However, Algorithms 2 and 3 are different for implementation purpose. As shown (34)*, $A_{\alpha_i}^{(k)}$, which contains the model knowledge A, is required to calculate $P_i^{(k+1)}$. Therefore, Algorithm 3 is a model-based algorithm. In contrast, as shown in* (23)*, the value function parameter $P_i^{(k+1)}$ is determined by collecting the online data instead of model knowledge. Therefore, Algorithm 2 is a data-driven algorithm.* □

Remark 6: As shown in Figure 1-(a) and Figure 1-(b), in classical VI and PI, the iterative algorithm is implemented between the value function and the control policy. The value function update in PI or VI depends on the policy in the previous iteration and includes two steps in each iteration. However, in algorithm 3, one can observe that $P_i^{(k+1)}$ can be determined directly based on $P_i^{(k)}$ using equation (34)*, as shown in Figure 1-(c). That is, the value function*

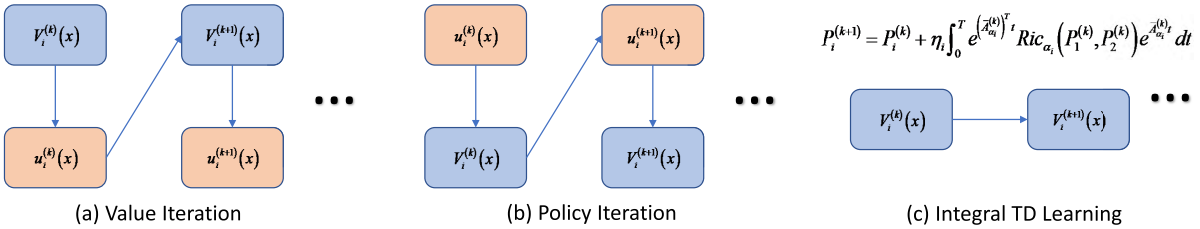

FIGURE 1. Diagram of VI, PI, and presented integral VI (IVI) algorithms.

update only depends on the value function itself. Therefore, the integral VI algorithm can be viewed as a simple one-step iteration. □

IV. MAIN RESULTS

In this section, we give the theoretical analysis of the integral VI algorithm in terms of three aspects, i.e., positive definiteness properties of the updated cost functions, the stability discussions concerning the closed-loop systems, and the conditions that guarantee the monotone convergence.

A. POSITIVE DEFINITENESS OF THE INTEGRAL VI ALGORITHM

In this subsection, the positiveness of the iterative value function in the integral VI algorithm is analyzed.

Theorem 2: Suppose that $\eta_i \in (0, 1]$ *for* $i = 1, 2$, *and* $V_i^{(0)}(x)$ *is positive definite, then all the element in the value function sequence* $\left\{V_i^{(k)}(x)\right\}_{k=0}^{\infty}$ *is positive definite.*

Proof: Suppose that $V_i^{(k)}(x)$ is positive definite. Substituting (16) into (17), one can obtain

$$\begin{aligned} x_t^T P_i^{(k+1)} x_t &= (1-\eta_i)\, x_t^T P_i^{(k)} x_t \\ &\quad + \eta_i \int_t^{t+T} e^{-2\alpha_i(\tau - t)} x^T \bar{Q}_i^{(k)} x d\tau \\ &\quad + \eta_i e^{-2\alpha_i T} x_{t+T}^T P_i^{(k)} x_{t+T} \end{aligned} \tag{35}$$

Since $V_i^{(k)}(x_t)$ is a positive definite function, then $P_i^{(k)}$ is a positive definite matrix. Therefore, the first and last terms in (35) are both positive definite functions. In addition, the second term in (35) satisfies $\eta_i \int_t^{t+T} e^{-2\alpha_i(\tau-t)} x_t^T \bar{Q}_i^{(k)} x_t d\tau \geq 0$. Finally, $V_i^{(k+1)}(x_t) = x_t^T P_i^{(k+1)} x_t$ is also positive definite. This completes the proof. ■

B. STABILITY DISCUSSION

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

Before moving on, the following lemma is required.

Lemma 2: For a symmetric matrix $G \in M^{n\times n}$, *and any nonzero matrices* $E_1 \in C^{n\times n}, E_2 \in C^{n\times n}, F_1 \in C^{n\times n}, F_2 \in C^{n\times n}$, *it follows that*

$$G + E_1 E_2 + E_2^T E_1^T + F_1 F_2 + F_2^T F_1^T < 0 \tag{36}$$

if there exists constant $\varepsilon > 0$ *such that*

$$G + \varepsilon E_1 E_1^T + \varepsilon^{-1} E_2^T E_2 + \varepsilon F_1 F_1^T + \varepsilon^{-1} F_2^T F_2 < 0. \tag{37}$$

Proof: Based on the Young's inequality, one has

$$X^T Y^T + YX \leq \varepsilon Y Y^T + \varepsilon^{-1} X^T X \tag{38}$$

for $\forall \varepsilon > 0$. Then, the following two inequalities holds

$$E_1 E_2 + E_2^T E_1^T \leq \varepsilon E_1 E_1^T + \varepsilon^{-1} E_2^T E_2, \tag{39}$$

$$F_1 F_2 + F_2^T F_1^T \leq \varepsilon F_1 F_1^T + \varepsilon^{-1} F_2^T F_2. \tag{40}$$

for any $\varepsilon > 0$. Inserting (39) and (40) into (37), one can obtain (36). This completes the proof. ■

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

Theorem 3: Let $\bar{A}_{\alpha_i}^{(0)}$ *be Hurwitz. Suppose that for each player, there exists a positive definite matrix* Y_i^k *in k-th iteration satisfying the Lyapunov equation*

$$\left(\bar{A}_{\alpha_i}^{(k)}\right)^T Y_i^k + Y_i^k \bar{A}_{\alpha_i}^{(k)} = -I. \tag{41}$$

If η_1 *and* η_2 *satisfy*

$$0 < \eta_{\max} < 1/2 \times \frac{1}{\sqrt{\left(\left\|H_1 Y_i^{(k)}\right\|^2 + \left\|H_2 Y_i^{(k)}\right\|^2\right)\left(\left\|M_1^{(k)}\right\|^2 + \left\|M_2^{(k)}\right\|^2\right)}} \tag{42}$$

for $\forall k \in \mathbb{N}$, *where*

$$\begin{aligned} M_i^{(k)} &= \int_0^T e^{\left(\bar{A}_{\alpha_i}^{(k)}\right)^T t} Ric_{\alpha_i}\left(P_1^{(k)}, P_2^{(k)}\right) e^{\bar{A}_{\alpha_i}^{(k)} t} dt. \\ \eta_{\max} &= \max\{\eta_1, \eta_2\} \end{aligned} \tag{43}$$

Then, $\bar{A}_{\alpha_i}^{(k)}$ *is Hurwitz for all* $k \in \mathbb{N}$.

Proof: We will prove this theorem by deduction.

First, from the assumption, $\bar{A}_{\alpha_i}^{(0)}$ is Hurwitz. Suppose that $\bar{A}_{\alpha_i}^{(k)}$ is Hurwitz. In *k-th* iteration, there exists a positive definite matrix denoted by $Y_i^k \in \mathbb{R}_p^{n\times n}$ such that (41) is satisfied.

Next, we need to show the Hurwitzness of the matrix $\bar{A}_{\alpha_i}^{(k+1)}$. As the following discussions, this will be done by finding the sufficient condition $\left(\bar{A}_{\alpha_i}^{(k+1)}\right)^T Y_i^{(k)} + Y_i^{(k)} \bar{A}_{\alpha_i}^{(k+1)} < 0$ that guarantees the Hurwitzness of the

matrix $\bar{A}_{\alpha_i}^{(k+1)}$. Rewriting $\bar{A}_{\alpha_i}^{(k+1)}$ using the fact that $P_i^{(k+1)} = P_i^{(k)} + \eta_i M_i^{(k)}$ in (34) yields

$$\begin{aligned}\bar{A}_{\alpha_i}^{(k+1)} &= A - H_1 P_1^{(k+1)} - H_2 P_2^{(k+1)} - \alpha_i I \\ &= \bar{A}_{\alpha_i}^{(k)} - \eta_1 H_1 M_1^{(k)} - \eta_2 H_2 M_2^{(k)}, \quad i = 1, 2. \end{aligned} \tag{44}$$

Based on (41) and (44), one has

$$\begin{aligned}&\left(\bar{A}_{\alpha_i}^{(k+1)}\right)^T Y_i^{(k)} + Y_i^{(k)} \bar{A}_{\alpha_i}^{(k+1)} \\ &\quad = -I - \left(\eta_1 H_1 M_1^{(k)} + \eta_2 H_2 M_2^{(k)}\right)^T Y_i^{(k)} \\ &\qquad - Y_i^{(k)} \left(\eta_1 H_1 M_1^{(k)} + \eta_2 H_2 M_2^{(k)}\right) \end{aligned} \tag{45}$$

Based on (45) and Lemma 2, $\left(\bar{A}_{\alpha_i}^{(k+1)}\right)^T Y_i^{(k)} + Y_i^{(k)} \bar{A}_{\alpha_i}^{(k+1)} < 0$ holds if

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

The above inequality can be guaranteed by

$$\begin{aligned}&\varepsilon_i^2 \left(\left(H_1 Y_i^{(k)}\right)^T H_1 Y_i^{(k)} + \left(H_2 Y_i^{(k)}\right)^T H_2 Y_i^{(k)}\right) - \varepsilon_i I \\ &\quad + \left(\eta_1^2 \left(M_1^{(k)}\right)^T M_1^{(k)} + \eta_2^2 \left(M_2^{(k)}\right)^T M_2^{(k)}\right) < 0 \end{aligned} \tag{46}$$

Note that (46) is a matrix inequality quadratic in the variable ε_i. To transform this into a scalar inequality, multiplying both sides of (46) by nonzero vector x^T and x with $x \in R^n$, one can obtain

$$\begin{aligned}&\varepsilon_i^2 \left(\left\|H_1 Y_i^{(k)} x\right\|^2 + \left\|H_2 Y_i^{(k)} x\right\|^2\right) - \varepsilon_i \|x\|^2 \\ &\quad + \left(\eta_1^2 \left\|M_1^{(k)} x\right\|^2 + \eta_2^2 \left\|M_2^{(k)} x\right\|^2\right) < 0 \end{aligned} \tag{47}$$

Since H_1 and H_2 are positive definiteness matrices, $Y_i^{(k)}$ is also positive definiteness matrix and $x \neq 0$, then $\left\|H_1 Y_i^{(k)} x\right\|^2 + \left\|H_2 Y_i^{(k)} x\right\|^2 > 0$. Therefore, (47) is a scalar inequality quadratic in ε_i. In this case, the existence condition for $\varepsilon_i \in \mathbb{R}_+$ can be determined as

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

That is,

$$\begin{aligned}&0 < \eta_{\max} < 1/2 \\ &\times \frac{1}{\sqrt{\left(\left\|H_1 Y_i^{(k)}\right\|^2 + \left\|H_2 Y_i^{(k)}\right\|^2\right)\left(\left\|M_1^{(k)}\right\|^2 + \left\|M_2^{(k)}\right\|^2\right)}} \end{aligned} \tag{48}$$

which ensures the existence of $\varepsilon_i > 0$ in (47). Finally, the requirement of η_1, η_2 to guarantee that $\bar{A}_{\alpha_i}^{(k+1)}$ is Hurwitz can be summarized as in (48). This completes the proof. ■

C. CONVERGENCE ANALYSIS

In this subsection, the effect of the parameters, the learning rate η_i, on the convergence of the integral VI algorithm is discussed.

Theorem 4: Define the following parameters

$$\begin{aligned} M_i^{(k)} &= \int_0^T e^{\left(\bar{A}_{\alpha_i}^{(k)}\right)^T t} Ric_{\alpha_i}\left(P_1^{(k)}, P_2^{(k)}\right) e^{\bar{A}_{\alpha_i}^{(k)} t} dt \\ S_1^{(k)} &= e^{\left(\bar{A}_{\alpha_1}^{(k)}\right)^T T} Ric_{\alpha_1}\left(P_1^{(k)}, P_2^{(k)}\right) e^{\bar{A}_{\alpha_1}^{(k)} T} \\ &\quad - \eta_2 M_2^{(k)} H_2 M_1^{(k)} - \eta_2 M_1^{(k)} H_2 M_2^{(k)} \\ \beta_1^{(k)} &= \eta_2 \left(M_2^{(k)} H_3 P_2^{(k)} + P_2^{(k)} H_3 M_2^{(k)} \right. \\ &\quad \left. - M_2^{(k)} H_2 P_1^{(k)} - P_1^{(k)} H_2 M_2^{(k)}\right) \\ &\quad + \eta_2^2 M_2^{(k)} H_3 M_2^{(k)} \\ \rho_1^{(k)-} &= \left\|Ric_{\alpha_1}\left(P_1^{(k)}, P_2^{(k)}\right)\right\| - \left\|S_1^{(k)}\right\| \\ \sigma_1^{(k)} &= \left\|M_1^{(k)} H_1 M_1^{(k)}\right\| \\ S_2^{(k)} &= e^{\left(\bar{A}_{\alpha_2}^{(k)}\right)^T T} Ric_{\alpha_2}\left(P_1^{(k)}, P_2^{(k)}\right) e^{\bar{A}_{\alpha_2}^{(k)} T} \\ &\quad - \eta_1 M_1^{(k)} H_1 M_2^{(k)} - \eta_1 M_2^{(k)} H_1 M_1^{(k)} \\ \rho_2^{(k)-} &= \left\|Ric_{\alpha_2}\left(P_1^{(k)}, P_2^{(k)}\right)\right\| - \left\|S_2^{(k)}\right\| \\ \sigma_2^{(k)} &= \left\|M_2^{(k)} H_2 M_2^{(k)}\right\| \\ \beta_2^{(k)} &= \eta_1 \left(M_1^{(k)} H_4 P_1^{(k)} + P_1^{(k)} H_4 M_1^{(k)} \right. \\ &\quad \left. - M_1^{(k)} H_1 P_2^{(k)} - P_2^{(k)} H_1 M_1^{(k)}\right) \\ &\quad + \eta_1^2 M_1^{(k)} H_4 M_1^{(k)} \end{aligned}$$

Then, the following propositions hold.

a) *If we consider* (6) *and* (34)*, then the following matrix recursive equation holds for* $P_1^{(k+1)}$ *and* $P_2^{(k+1)}$*:*

$$\begin{aligned}&Ric_{\alpha_1}\left(P_1^{(k+1)}, P_2^{(k+1)}\right) \\ &\quad = (1 - \eta_1) Ric_{\alpha_1}\left(P_1^{(k)}, P_2^{(k)}\right) \\ &\qquad + \eta_1 S_1^{(k)} - \eta_1^2 M_1^{(k)} H_1 M_1^{(k)} + \beta_1^{(k)}, \end{aligned} \tag{49}$$

$$\begin{aligned}&Ric_{\alpha_2}\left(P_1^{(k+1)}, P_2^{(k+1)}\right) \\ &\quad = (1 - \eta_2) Ric_{\alpha_2}\left(P_1^{(k)}, P_2^{(k)}\right) \\ &\qquad + \eta_2 S_2^{(k)} - \eta_2^2 M_2^{(k)} H_2 M_2^{(k)} + \beta_2^{(k)}. \end{aligned} \tag{50}$$

b) *For player one, when the learning rate* η_1 *satisfies* $\eta_1 \in (0, 1]$*, if* (48) *and the following conditions hold*:

– **Condition 1***:*

$$\Delta_{11} = \left(\rho_1^{(k)-}\right)^2 - 4\sigma_1^{(k)} \left\|\beta_1^{(k)}\right\| > 0, \tag{51}$$

$$\frac{\rho_1^{(k)-}-\sqrt{\Delta_{11}}}{2\sigma_1^{(k)}} \le \eta_1 \le \frac{\rho_1^{(k)-}+\sqrt{\Delta_{11}}}{2\sigma_1^{(k)}}, \tag{52}$$

$$0 < \frac{\rho_1^{(k)-}-\sqrt{\Delta_{11}}}{2\sigma_1^{(k)}} < 1, \tag{53}$$

Then,

$$\left\|Ric_{\alpha_1}\left(P_1^{(k+1)},P_2^{(k+1)}\right)\right\| \le \left\|Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right)\right\|$$

holds for every $k \in \mathbb{N}$.

For player two, when the learning rate η_2 *satisfies* $\eta_2 \in (0, 1]$, *and if* (48) *and the following conditions hold,*

– **Condition 2***:*

$$\Delta_{21} = \left(\rho_2^{(k)-}\right)^2 - 4\sigma_2^{(k)}\left\|\beta_2^{(k)}\right\| > 0, \tag{54}$$

$$\frac{\rho_2^{(k)-}-\sqrt{\Delta_{21}}}{2\sigma_2^{(k)}} \le \eta_2 \le \frac{\rho_2^{(k)-}+\sqrt{\Delta_{21}}}{2\sigma_2^{(k)}},$$

$$0 < \frac{\rho_2^{(k)-}-\sqrt{\Delta_{21}}}{2\sigma_2^{(k)}} < 1. \tag{55}$$

Then,

$$\left\|Ric_{\alpha_2}\left(P_1^{(k+1)},P_2^{(k+1)}\right)\right\| \le \left\|Ric_{\alpha_2}\left(P_1^{(k)},P_2^{(k)}\right)\right\|$$

holds for every $k \in \mathbb{N}$.

c) *If learning rate* η_1 *and* η_2 *does not vanish at* $k = \infty$, *the pair* $(P_1^{(k)}, P_2^{(k)})$ *will monotonically converge to the* (P_1^*, P_2^*), *i.e.,* $Ric_{\alpha_1}(P_1^*, P_2^*) = \lim_{k\to\infty}\left\|Ric_{\alpha_i}\left(P_1^{(k)},P_2^{(k)}\right)\right\| = 0$ *for* $i = 1, 2$.

Proof: a) First, for player one, (34) can be equivalently rewritten as

$$P_1^{(k+1)} = P_1^{(k)} + \eta_1 M_1^{(k)}. \tag{56}$$

Then, applying (56) to the Riccati operator representation (6) yields,

$$\begin{aligned}
&Ric_{\alpha_1}\left(P_1^{(k+1)},P_2^{(k+1)}\right)\\
&= -2\alpha_1 P_1^{(k+1)} + A^T P_1^{(k+1)} + P_1^{(k+1)}A + Q_1\\
&\quad - P_1^{(k+1)}H_1P_1^{(k+1)} - P_2^{(k+1)}H_2P_1^{(k+1)}\\
&\quad - P_1^{(k+1)}H_2P_2^{(k+1)} + P_2^{(k+1)}H_3P_2^{(k+1)}\\
&= -2\alpha_1\left(P_1^{(k)}+\eta_1M_1^{(k)}\right) + A^T\left(P_1^{(k)}+\eta_1M_1^{(k)}\right)\\
&\quad + \left(P_1^{(k)}+\eta_1M_1^{(k)}\right)A + Q_1\\
&\quad - \left(P_1^{(k)}+\eta_1M_1^{(k)}\right)H_1\left(P_1^{(k)}+\eta_1M_1^{(k)}\right)\\
&\quad - \left(P_2^{(k)}+\eta_2M_2^{(k)}\right)H_2\left(P_1^{(k)}+\eta_1M_1^{(k)}\right)\\
&\quad - \left(P_1^{(k)}+\eta_1M_1^{(k)}\right)H_2\left(P_2^{(k)}+\eta_2M_2^{(k)}\right)\\
&\quad + \left(P_2^{(k)}+\eta_2M_2^{(k)}\right)H_3\left(P_2^{(k)}+\eta_2M_2^{(k)}\right)\\
&= Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right) - \eta_1^2M_1^{(k)}H_1M_1^{(k)} + \beta_1\\
&\quad + \eta_1\left[\left(\bar{A}_{\alpha_1}^{(k)}\right)^T M_1^{(k)} + M_1^{(k)}\bar{A}_{\alpha_1}^{(k)}\right.\\
&\quad \left. - \eta_2M_1^{(k)}H_2M_2^{(k)} - \eta_2M_2^{(k)}H_2M_1^{(k)}\right].
\end{aligned} \tag{57}$$

The terms $\left(\bar{A}_{\alpha_1}^{(k)}\right)^T M_1^{(k)} + M_1^{(k)}\bar{A}_{\alpha_1}^{(k)}$ in (57) can be written as

$$\begin{aligned}
&\left(\bar{A}_{\alpha_1}^{(k)}\right)^T M_1^{(k)} + M_1^{(k)}\bar{A}_{\alpha_1}^{(k)}\\
&= \int_0^T \frac{d}{dt}\left(e^{\left(\bar{A}_{\alpha_1}^{(k)}\right)^T t} Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right) e^{\bar{A}_{\alpha_1}^{(k)}t}\right)dt\\
&= e^{\left(\bar{A}_{\alpha_1}^{(k)}\right)^T T} Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right) e^{\bar{A}_{\alpha_1}^{(k)}T}\\
&\quad - Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right)
\end{aligned} \tag{58}$$

Finally, inserting (58) into (57) yields (49). Similarly, for player two, one can obtain (50).

b) When the learning rate η_1 satisfies condition 1 and (48), combining the fact (49) in proposition a) and the properties of the matrix norm yields

$$\begin{aligned}
&\left\|Ric_{\alpha_1}\left(P_1^{(k+1)},P_2^{(k+1)}\right)\right\|\\
&\le (1-\eta_1)\left\|Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right)\right\| + \eta_1\|S_1\| + \eta_1^2\sigma_1^k + \|\beta_1\|\\
&= \psi_{11}\left\|Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right)\right\|
\end{aligned} \tag{59}$$

where

$$\begin{aligned}
\psi_{11} = 1 &- \eta_1\frac{\rho_1^{(k)-}}{\left\|Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right)\right\|}\\
&+ \eta_1^2\frac{\sigma_1^k}{\left\|Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right)\right\|} + \frac{\|\beta_1\|}{\left\|Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right)\right\|}
\end{aligned}$$

In order to satisfy

$$\left\|Ric_{\alpha_1}\left(P_1^{(k+1)},P_2^{(k+1)}\right)\right\| \le \left\|Ric_{\alpha_1}\left(P_1^{(k)},P_2^{(k)}\right)\right\|$$

for each $k \in \mathbb{N}$, a sufficient condition can be selected as $\psi_{11} \le 1$ by (59), i.e.,

$$f_1(\eta_1) = \sigma_1^k\eta_1^2 - \rho_1^{(k)-}\eta_1 + \|\beta_1\| \le 0, \tag{60}$$

To guarantee the existence of the solution to the above quadratic inequality, the equation $f_1(\eta_1) = 0$ should have distinct solutions, i.e., (51) should be satisfied. Then, the solution to the equation $f_1(\eta_1) = 0$ can be denoted as $\frac{\rho_1^{(k)-}-\sqrt{\Delta_{11}}}{2\sigma_1^{(k)}}$ and $\frac{\rho_1^{(k)-}+\sqrt{\Delta_{11}}}{2\sigma_1^{(k)}}$. Therefore, (52) yields (60). On the other hand, $\eta_1 \in (0, 1)$ is assumed in condition 1). Then, to avoid the contradiction of the requirements $\eta_1 \in (0, 1)$ and (52), (53) is needed. This completes the proof of condition 1. The proof of condition 2 for $P_2^{(k)}$ is similar to the case for condition 1.

c) First, note that $\left\|Ric_{\alpha_i}\left(P_1^{(k)}, P_2^{(k)}\right)\right\| \geq 0$ for $\forall k \in \mathbb{N}$, i.e., it is lower-bound by zero. In addition, from proposition b, the sequence $\left\{\left\|Ric_{\alpha_i}\left(P_1^{(k)}, P_2^{(k)}\right)\right\|\right\}_{k=0}^{\infty}$ is monotonically decreasing. Therefore, $P_1^{(k+1)} = P_1^{(k)} = P_1^{(\infty)}$ as $k \to \infty$. In addition, from (34), one has

$$\eta_1 \int_0^T e^{\left(\bar{A}_{\alpha_i}^{(\infty)}\right)^T t} Ric_{\alpha_i}\left(P_1^{(\infty)}, P_2^{(\infty)}\right) e^{\bar{A}_{\alpha_i}^{(\infty)} t} dt = 0 \quad (61)$$

Note that $\eta_1 \neq 0$, then (61) is equivalent to

$$\int_0^T e^{\left(\bar{A}_{\alpha_i}^{(\infty)}\right)^T t} Ric_{\alpha_i}\left(P_1^{(\infty)}, P_2^{(\infty)}\right) e^{\bar{A}_{\alpha_i}^{(\infty)} t} dt = 0$$

Since the exponential function can not be zero, then $Ric_{\alpha_i}\left(P_1^{(\infty)}, P_2^{(\infty)}\right) = 0$. Therefore, $P_1^{(\infty)} = P_1^*$ and $P_2^{(\infty)} = P_2^*$. That is, (P_1^*, P_2^*) converges to the solution to the CAREs. This completes the proof. ■

Remark 7: As the learning rate η_1 and η_2 increase, the convergence speed of the integral VI algorithm will be faster, when the learning rate of the integral VI algorithm is sufficiently large, the integral VI algorithm outperforms the PI algorithm. To guarantee the positive definiteness in Theorem 2, the max value of η_1 and η_2 can not exceed 1. □

Remark 8: For player one, the learning rate η_1 need to satisfy (51), (52) *and* (53) *in condition* 1, *which contain $\rho_1^{(k)-}$. Note that $\rho_1^{(k)-}$ is affected by η_2. Then, the learning rate η_1 is not independent of η_2. Similarly, for player two, the learning rate η_2 also depends on η_1.* □

V. SIMULATION STUDY

Here we present simulations of NZS differential games for linear systems, the games can be solved by the integral VI method and another method, Lyapunov iteration method that is used as a reference to verify the effectiveness of the proposed method for the NZS differential games.

Consider the following two-player continuous-time linear systems with [38]

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

$$B_1 = \begin{bmatrix} 0.4422 \\ 3.0447 \\ -5.52 \\ 0 \end{bmatrix}, B_2 = \begin{bmatrix} 0.1761 \\ -7.5922 \\ 4.99 \\ 0 \end{bmatrix}.$$

where $Q_1 = diag\,([3.5, 2, 4, 5])$, $R_{11} = 1$, $R_{12} = 0.25$ and $Q_2 = diag\,([1.5, 6, 3, 1])$, $R_{21} = 0.6$, $R_{22} = 2$, $\alpha_1 = 5$, $\alpha_2 = 10$, The initial state is slected as $x(0) = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^T$.

By using the PI algorithm, the solution, $(P_1^{(*)}, P_2^{(*)})$, to the CAREs (4) and (5) can be obtained as

$$P_1^{(*)} = \begin{bmatrix} 0.3463 & 0.0068 & 0.0112 & -0.0156 \\ 0.0068 & 0.1113 & 0.0576 & -0.0139 \\ 0.0112 & 0.0576 & 0.2308 & 0.0250 \\ -0.0156 & -0.0139 & 0.0250 & 0.5112 \end{bmatrix}$$

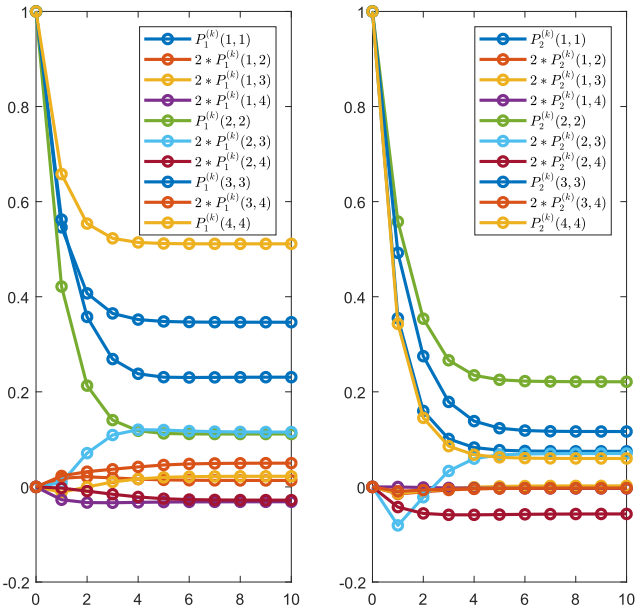

FIGURE 2. The learning process of $(P_1^{(k)}, P_2^{(k)})$ for two-player when $\eta_1 = \eta_2 = 0.7$.

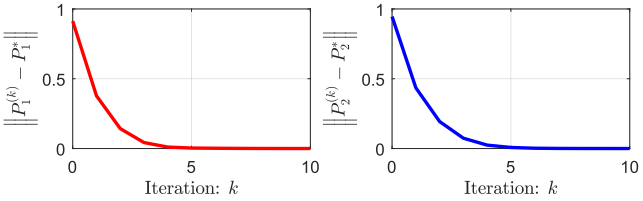

FIGURE 3. Convergence of $P_1^{(k)}$ and $P_2^{(k)}$ to their optimal values P_1^* and P_2^* with $\eta_1 = \eta_2 = 0.7$ during the learning process.

$$P_2^{(*)} = \begin{bmatrix} 0.0748 & -0.0008 & 0.0011 & -0.0016 \\ -0.0008 & 0.2211 & 0.0351 & -0.0284 \\ 0.0011 & 0.0351 & 0.1165 & -0.0013 \\ -0.0016 & -0.0284 & -0.0013 & 0.0598 \end{bmatrix}$$

The integral VI algorithm is implemented using $T = 0.5$. The threshold of the stop criterion is selected as $\epsilon = 10^{-8}$. The initial matrices $P_1^{(0)}$, $P_2^{(0)}$ are selected as identity matrices. In order to solve online for the values of the $P_1^{(k)}$, $P_2^{(k)}$, the LS method is chosen after a set of 15 data samples is collected and thus the policy of the controller is updated every 7.5 sec. Figures (2) and (4) presents the evolution of the parameters of the value function for players one two during the learning process when the learning rate is selected as $\eta_1 = \eta_2 = 0.7$ and $\eta_1 = \eta_2 = 1.0$, respectively. It can be shown that the learning algorithm converges within 5 steps. Moreover, to investigate the convergence of the integral VI algorithm to the solution of the CAREs (4) and (5), The difference between $(P_1^{(k)}, P_2^{(k)})$ and $(P_1^{(*)}, P_2^{(*)})$ is shown in Figures 3 and 5, respectively. One can observe that the value functions for both players converge to $(P_1^{(*)}, P_2^{(*)})$.

A comparison between the integral VI algorithm with different learning rate and the PI algorithm is shown in Figure 6. First, the larger learning rate results in faster convergence

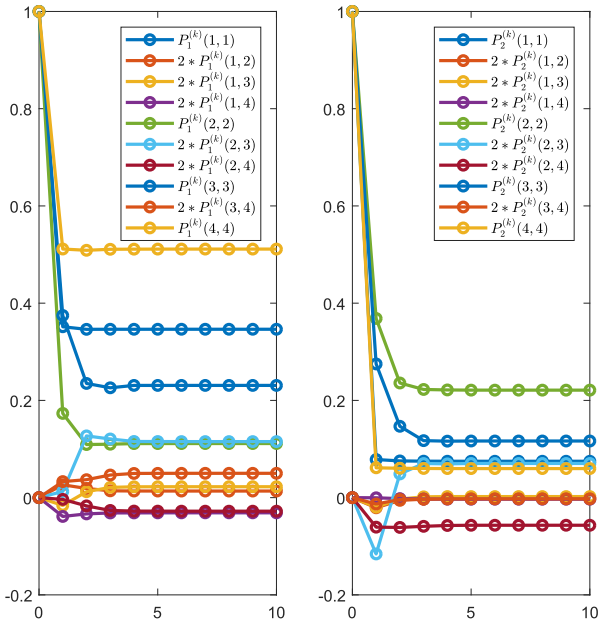

FIGURE 4. The learning process of $(P_1^{(k)}, P_2^{(k)})$ for two-player when $\eta_1 = \eta_2 = 1.0$.

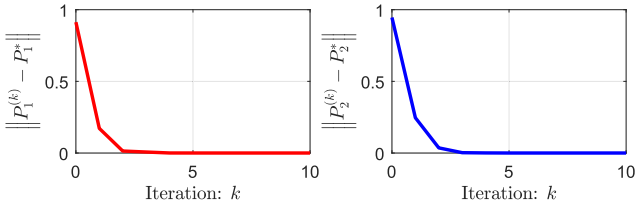

FIGURE 5. Convergence of $P_1^{(k)}$ and $P_2^{(k)}$ to their optimal values P_1^* and P_2^* with $\eta_1 = \eta_2 = 1.0$ during the learning process.

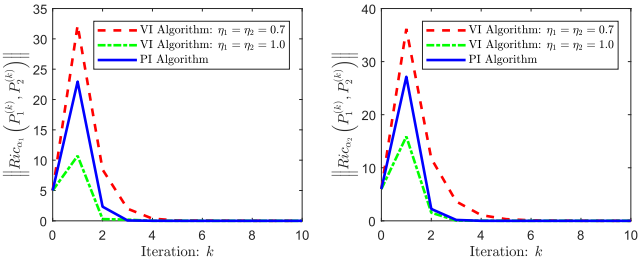

FIGURE 6. The comparison between the integral VI algorithm with different learning rate and the PI algorithm.

speed, i.e., the case of $\eta_i = 1.0$ converges to the optimal case faster than the case of $\eta_i = 0.7$. Second, when the learning rate of the integral VI algorithm is sufficiently large, the integral VI algorithm outperforms the PI algorithm, as shown in Figure 6.

VI. CONCLUSIONS

In this paper, an integral VI algorithm is proposed to find the Nash equilibrium of the NZS games. The presented integral VI algorithm is implemented using the online data to obviate the requirement of the drift dynamics. First, the reward function in the NZS games contains a discounted factor, which is required to be within a given range to guarantee the closed-loop stability. Compared with existing RL algorithms with fixed convergence speed, the convergence of the integral VI method can be tuned by the learning rate. Moreover, as discussed in Section IV, additional conditions on the learning rate are imposed to guarantee the positive definiteness of the iterative value function, closed-loop stability during learning and the convergence of the integral VI algorithm to the solutions of CAREs. Simulation examples demonstrates the effectiveness of the presented integral VI algorithm.

REFERENCES

[1] Q. Xu, N. Zhang, C. Kang, Q. Xia, D. He, C. Liu, Y. Huang, L. Cheng, and J. Bai, "A game theoretical pricing mechanism for multi-area spinning reserve trading considering wind power uncertainty," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1084–1095, Mar. 2016.

[2] S. Phelps, P. McBurney, and S. Parsons, "A novel method for strategy acquisition and its application to a double-auction market game," *IEEE Trans. Syst., Man, Cybern. B. Cybern.*, vol. 40, no. 3, pp. 668–674, Jun. 2010.

[3] F. Seo and M. Sakawa, "A game theoretic approach with risk assessment for international conflict solving," *IEEE Trans. Syst., Man, Cybern.*, vol. 20, no. 1, pp. 141–148, Jan./Feb. 1990.

[4] W. Lin, "Mixed H_2/H_∞ control via state feedback for nonlinear systems," *Int. J. Control*, vol. 64, no. 5, pp. 899–922, 1996.

[5] R. Isaacs, *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. North Chelmsford, MA, USA: Courier Corporation, 1999.

[6] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. Hoboken, NJ, USA: Wiley, 2012.

[7] H. Zhang, Q. Wei, and D. Liu, "An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games," *Automatica*, vol. 47, no. 1, pp. 207–214, Jan. 2011.

[8] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, *Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*. London, U.K.: Institution of Engineering and Technology, 2012.

[9] K. G. Vamvoudakis and F. L. Lewis, "Multi-player non-zero-sum games: Online adaptive learning solution of coupled Hamilton–Jacobi equations," *Automatica*, vol. 47, no. 8, pp. 1556–1569, 2011.

[10] K. G. Vamvoudakis, H. Modares, B. Kiumarsi, and F. L. Lewis, "Game theory-based control system algorithms with real-time reinforcement learning: How to solve multiplayer games online," *IEEE Control Syst.*, vol. 37, no. 1, pp. 33–52, Feb. 2017.

[11] H. Zhang, H. Jiang, C. Luo, and G. Xiao, "Discrete-time nonzero-sum games for multiplayer using policy-iteration-based adaptive dynamic programming algorithms," *IEEE Trans. Cybern.*, vol. 47, no. 10, pp. 3331–3340, Oct. 2017.

[12] Q. Zhang and D. Zhao, "Data-based reinforcement learning for nonzero-sum games with unknown drift dynamics," *IEEE Trans. Cybern.*, vol. 49, no. 8, pp. 2874–2885, Aug. 2019.

[13] T. Damm, V. Dragan, and G. Freiling, "Coupled Riccati differential equations arising in connection with nash differential games," *IFAC Proc. Volumes*, vol. 41, no. 2, pp. 3946–3951, 2008.

[14] L. Cherfi, Y. Chitour, and H. Abou-Kandil, "A new algorithm for solving coupled algebraic Riccati equations," in *Proc. Int. Conf. Comput. Intell. Modelling, Control Automat. Int. Conf. Intell. Agents, Web Technol. Internet Commerce (CIMCA-IAWTIC)*, Nov. 2005, pp. 83–88.

[15] Z. Gajic and X. Shen, *Parallel Algorithms for Optimal Control of Large Scale Linear Systems*. London, U.K.: Springer, 2012.

[16] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

[17] Q. Wei, D. Liu, Q. Lin, and R. Song, "Adaptive dynamic programming for discrete-time zero-sum games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 4, pp. 957–969, Apr. 2018.

[18] R. Song, F. L. Lewis, and Q. Wei, "Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 704–713, Mar. 2017.

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

[20] Y. Yang, D. Wunsch, and Y. Yin, "Hamiltonian-driven adaptive dynamic programming for continuous nonlinear dynamical systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 8, pp. 1929–1940, Aug. 2017.

[21] H. Zhang, J. Zhang, G.-H. Yang, and Y. Luo, "Leader-based optimal coordination control for the consensus problem of multiagent differential games via fuzzy adaptive dynamic programming," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 1, pp. 152–163, Feb. 2015.

[22] 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 Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2139–2153, Jun. 2018.

[23] Y. Yang, H. Modares, D. C. Wunsch, and Y. Yin, "Optimal containment control of unknown heterogeneous systems with active leaders," *IEEE Trans. Control Syst. Technol.*, vol. 27, no. 3, pp. 1228–1236, May 2019.

[24] Y. Yang, S. Cheng, Y. Yin, and D. C. Wunsch, "Containment control of heterogeneous systems with non-autonomous leaders: A distributed optimal model reference approach," *IEEE Access*, vol. 6, pp. 60689–60703, 2018.

[25] T. Y. Chun, J. Y. Lee, J. B. Park, and Y. H. Choi, "Stability and monotone convergence of generalised policy iteration for discrete-time linear quadratic regulations," *Int. J. Control*, vol. 89, no. 3, pp. 437–450, 2016.

[26] T. Y. Chun, J. Y. Lee, J. B. Park, and Y. H. Choi, "Adaptive dynamic programming for discrete-time linear quadratic regulation based on multirate generalised policy iteration," *Int. J. Control*, vol. 91, no. 6, pp. 1223–1240, 2018.

[27] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 621–634, Mar. 2014.

[28] D. Wang, D. Liu, and H. Li, "Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 2, pp. 627–632, Apr. 2014.

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

[30] Q. Wei, D. Liu, and H. Lin, "Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 840–853, Mar. 2016.

[31] K. G. Vamvoudakis, "Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach," *Syst. Control Lett.*, vol. 100, pp. 14–20, Feb. 2017.

[32] B. Luo, Y. Yang, and D. Liu, "Adaptive q-learning for data-based optimal output regulation with experience replay," *IEEE Trans. Cybern.*, vol. 48, no. 12, pp. 3337–3348, Dec. 2018.

[33] 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," *Int. J. Robust Nonlinear Control*, vol. 29, no. 9, pp. 2673–2694, 2019.

[34] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[35] B. Kiumarsi, F. L. Lewis, and Z.-P. Jiang, "H_∞ control of linear discrete-time systems: Off-policy reinforcement learning," *Automatica*, vol. 78, no. 1, pp. 144–152, Apr. 2017.

[36] Y. Yang, Z. Guo, H. Xiong, D.-W. Ding, Y. Yin, and D. C. Wunsch, "Data-driven robust control of discrete-time uncertain linear systems via off-policy reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published. [Online]. Available: https://ieeexplore.ieee.org/document/8657370

[37] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Trans. Autom. Control*, vol. 59, no. 11, pp. 3051–3056, Nov. 2014.

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

YONGLIANG YANG (M'16) received the B.S. degree from Hebei University, Baoding, China, in 2011, and the Ph.D. degree from the University of Science and Technology Beijing (USTB), Beijing, China, in 2017.

He was a Visiting Scholar with the Missouri University of Science and Technology, Rolla, USA, from 2015 to 2017. He is currently an Assistant Professor with USTB. He was supported by the China Scholarship Council. His research interests include adaptive optimal control, distributed optimization and control, and cyber-physical systems.

Dr. Yang was a recipient of the Best Ph.D. Dissertation, the Chancellor's Scholarship, the Scholarship for Outstanding Ph.D. Students in USTB, and the Excellent Graduates Award in Beijing. He also serves as a Reviewer for several international journals and conferences, including *Automatica*, the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, the IEEE CONTROL SYSTEMS LETTERS, the IEEE/CAA JOURNAL OF AUTOMATICA SINICA AND NEUROCOMPUTING.

LIMING WANG received the B.S. degree from Henan University, Kaifeng, China, in 2017. He is currently pursuing the M.S. degree with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China. His research interests include adaptive control and optimal control.

HAMIDREZA MODARES (M'15) received the B.S. degree from the University of Tehran, Tehran, Iran, in 2004, the M.S. degree in electrical engineering from the Shahrood University of Technology, Shahrood, Iran, in 2006, and the Ph.D. degree in electrical engineering from The University of Texas at Arlington, Arlington, TX, USA, in 2015.

He was a Faculty Research Associate with The University of Texas at Arlington, from 2015 to 2016, and an Assistant Professor with the Missouri University of Science and Technology, Rolla, MO, USA, from 2016 to 2018. He is currently an Assistant Professor with the Mechanical Engineering Department, Michigan State University, East Lansing, MI, USA. His current research interests include cyber-physical systems, reinforcement learning, distributed control, robotics, and machine learning.

Dr. Modares was a recipient of the Best Paper Award from the 2015 IEEE International Symposium on Resilient Control Systems. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

DAWEI DING received the B.E. degree from the Ocean University of China, Qingdao, China, in 2003, and the Ph.D. degree in control theory and engineering from Northeastern University, Shenyang, China, in 2010. He is currently a Professor and the Dean of the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China. His research interests include robust control and filtering, fuzzy control, and multidimensional systems.

YIXIN YIN (M'07) received the B.S., M.S., and Ph.D. degrees from the University of Science and Technology Beijing (USTB), Beijing, China, in 1982, 1984, and 2002, respectively.

He has served as the Dean of the School of the Information Engineering, USTB, from 2000 to 2011, where he was the Dean of the School of Automation and Electrical Engineering, from 2011 to 2017. He is a Fellow of the Chinese Society for Artificial Intelligence and a Member of The Chinese Society for Metals and The Chinese Association of Automation. He is currently a Professor with the School of Automation and Electrical Engineering, USTB. His major research interests include the modeling and the control of complex industrial processes, the computer aided design of control systems, intelligent control, and artificial life. He was a recipient of several national awards, including the Outstanding Young Educator, in 1993, the Award of Science and Technology Progress in Education, in 1994, the Award of the National Science and Technology Progress, in 1995, the Special Allowance of the State Council, in 1994, the Best Paper of Japanese Acoustical Society, in 1999, and the Award of Metallurgical Science and Technology, in 2014. He was a Visiting Scholar with several universities in Japan, including The University of Tokyo, the Kyushu Institute of Technology, Kanagawa University, Chiba University, and the Muroran Institute of Technology.

DONALD WUNSCH (F'05) received the B.S. degree in applied mathematics from the University of New Mexico, the M.B.A. degree from Washington University, St. Louis, the M.S. degree in applied mathematics and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, and the Jesuit Core Honors Program from Seattle University.

He was an INNS President. He is currently the Mary K. Finley Missouri Distinguished Professor with the Missouri University of Science and Technology (Missouri S&T). He was with Texas Tech University, Boeing, Rockwell International, and International Laser Systems. He is an INNS Fellow and a Senior Fellow, from 2007 to 2013, the NSF CAREER Award, and the 2015 INNS Gabor Award. He has served as the IJCNN general chair. He has served on several Boards, including the St. Patrick's School Board, the IEEE Neural Networks Council, the International Neural Networks Society, and the University of Missouri Bioinformatics Consortium, the Chair of the Missouri S&T Information Technology and Computing Committee, and the Student Design and Experiential Learning Center Board. His key research contributions are clustering, adaptive resonance, and reinforcement learning architectures, hardware and applications, neurofuzzy regression, traveling salesman problem heuristics, robotic swarms, and bioinformatics.