21st International Symposium on
Mathematical Theory of Networks and Systems
July 7-11, 2014. Groningen, The Netherlands

Data-based Reinforcement Learning algorithm with Experience Replay for Solving Constrained Nonzero-sum Differential Games

S. Yasini, *Student Member*, *IEEE*, A. Karimpour, and M. B. Naghibi Sistani

***Abstract*—In this paper a partially model-free reinforcement learning (RL) algorithm based on experience replay is developed for finding online the Nash equilibrium solution of the multi-player nonzero-sum (NZS) differential games. In order to avoid the performance degradation or even system instability, the amplitude limitation on the control inputs is considered in the design procedure. The proposed algorithm is implemented on actor-critic structure for every player in the game, where both actor and critic networks are tuned at the same time. The game players learn online the solution of the constrained coupled Hamilton-Jacobi (HJ) equations, without using any knowledge on the internal system dynamics. The idea of experience replay is used to relax the requirement for checking the restrictive persistence of excitation (PE) condition which is difficult to verify or implement online. The closed-loop stability is analyzed and the convergence to the Nash equilibrium of the game is shown. A simulation study example is provided showing the effectiveness of the proposed approach.**

I. Introduction

DIFFRENTIAL games adapt characteristics from game theory and optimal control theory, having many potential applications in control engineering, economics, military [1], etc. Nonzero-sum (NZS) differential games are generalization of optimal control theory in situations where more than one player or controller make decision to control a single nonlinear system, each tries to minimize his own performance criterion [2].

For nonlinear dynamical systems, finding Nash equilibrium solution to the NZS games is equivalent to solving the coupled Hamilton-Jacobi (HJ) equations. However, solving the coupled HJ equations is a very difficult problem and an alternative approach is to find the approximate solution. For linear systems with quadratic cost functions, solving NZS game requires obtaining the solution to the coupled algebraic Riccati equations (AREs) [3-5]. In [5], [6], solution of coupled AREs was considered by means of offline iterative procedures.

Policy iteration (PI) algorithm [7] is particular class of reinforcement learning (RL) [8] techniques which provides effective means of learning solutions to the coupled HJ equations in an online manner. The PI algorithm is based on two steps of policy evaluation to determine the value function and policy improvement to obtain improved policies.

S. Yasini, A. Karimpour and M. B. Naghibi Sistani are with the Electrical Engineering Department, Ferdowsi University of Mashhad, Azadi Sq. 9177948974, Mashhad, Iran (phone: +98-511-8803000;

E-mail addresses: sh.yasini@ stu-mail.um.ac.ir, karimpor@um.ac.ir, mb-naghib@um.ac.ir).

Although great progress has been done in developing RL algorithms that provide online solution to the one player game [9-11], and two-player zero-sum (ZS) game [12-14]; however, fewer results are available considering NZS games of continuous-time (CT) systems.

An algorithm based on PI was proposed in [15] that provides online solution to the linear quadratic Nash NZS games for partially-unknown CT systems. Vamvoudakis and Lewis [16] introduced an online model-based synchronous PI algorithm on actor-critic neural network (NN) structure for solving the multi-player NZS games which involves synchronous adaptation of both actor and critic NNs associated with each player in the game. This work was subsequently extended in [17], for partially-unknown nonlinear CT systems. In [18], approximate dynamic programming algorithm was developed to solve the two-player NZS games of nonlinear CT systems with known dynamics, where only a critic network was used for every player instead of dual actor-critic network. Although efficient; however, the aforementioned work in [15-18], did not take into account the saturation nonlinearity which is unavoidable in most actuators. In fact, neglecting of actuator saturation can be source of performance degradation or even instability of the closed-loop system.

Another important issue in existing RL-based adaptive optimal control methods for solving differential games is the need to ensure the persistence of excitation (PE) condition for parameter convergence. This condition is very difficult to implement in practical control algorithms. The idea of experience replay (ER) was employed in [19] to guarantee parameter convergence in direct adaptive control without relying on PE condition. However, their work does not consider any optimality guarantees on the closed-loop system. Modares et. al [20, 21], proposed an online algorithm which uses the idea of ER for solving the optimal control of constrained-input CT systems (Solution of Hamilton-Jacobi-Bellman equation).

In this paper we aim to extend the effort in [20, 21] and develop an online algorithm based on PI to solve *N*-player NZS differential games for partially-unknown dynamical systems with amplitude limitation on control inputs and relaxed PE. It should be noted that when NZS is considered the problem become more difficult due to the existing of coupling terms in the coupled HJ equations which results in a difficulty challenge in obtaining tuning laws for the critic and actor NN weights to guarantee the stability of the system while converging to the Nash solution of the game. The nonquadratic cost functional is employed to encode the input

ISBN: 978-90-367-6321-9

constraints into the NZS differential games and derive the corresponding coupled HJ equations in contrast to [15-18], where no input saturation is considered. Moreover, the need to check for the restrictive PE condition is relaxed in comparison to other online adaptive optimal procedures [12-18] for solving differential games, where the system states need to be PE for parameter convergence. *N*-critic and *N*-actor structures are used to approximate the optimal control laws and the optimal value functions, where all NNs are tuned at the same time. The stability of the system is analyzed in the sense of Lyapunov and the convergence to the Nash equilibrium is shown.

The paper is organized as follows. Formulation of constrained *N*-player NZS games is derived in the next section. In section III, an offline PI algorithm for solving the coupled HJ equations is introduced. Main results of this paper are presented in section IV where an online ER RL algorithm based on PI is introduced for finding the Nash equilibrium solution to the NZS game. Sections V, VI show simulation study and discussion of the results, respectively.

II. Problem Formulation

A. Constrained multi-player nonzero-sum games

Consider the following *N*-player nonlinear differential game

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

where $x \in \Re^n$ is the state vector, $u_j \in \Re^{m_j}$ are players. We assume that $f(.) \in \Re^n$, $g_j(.) \in \Re^{n \times m_j}$ are locally Lipschitz, and system (1) is zero-state observable on the compact set Ω [22].

Definition 1. A set of feedback control policies $\{u_i(x); i \in N\}$ is admissible on Ω, if the dynamics of closed-loop system (1) are stable and the cost function for the given set have finite value.

For admissible control policies, define the cost function as

$$V_i(x(0), u_1, ..., u_N) = \int_t^{\infty} (Q_i(x) + \sum_{j=1}^{N} M_i(u_j)) d\tau, \quad i \in N \tag{2}$$

where $Q_i(.) \geq 0$, and $M_i(.) \in \Re$ is positive definite. To consider the input constraints, a nonquadratic functional associated with player i, is employed as follows [23]

$$M_i(u_j) = 2 \int_0^{u_j} (\bar{u}_j \phi^{-1}(v_j / \bar{u}_j))^T R_{ij} dv_j, \tag{3}$$

where $v_j \in \Re^{m_j}$, $\bar{u}_j$ is saturating bound for actuators, $R_{ii} > 0$, $R_{ij} \geq 0$ are considered to be diagonal matrices for simplicity of analysis, and $\phi(.) = \tanh(.)$.

Definition 2 [2]. The set of control strategies $\{u_1^*, u_2^*, ..., u_N^*\}$ with $u_i^* \in \Omega_i, i \in N$, constructs a Nash equilibrium solution if for all $u_i^* \in \Omega_i$, $i \in N$

$$V_i^* \equiv V_i(u_1^*, u_2^*, ..., u_i^*, ..., u_N^*) \leq V_i(u_1^*, u_2^*, ..., u_i, ..., u_N^*). \tag{5}$$

A differential equivalent to each value function is given by the following nonlinear Lyapunov equations

$$\begin{aligned} &LE_i(x, \nabla V_i, u_1, u_2, ..., u_N) = \\ &Q_i(x) + 2 \sum_{j=1}^{N} \int_0^{u_j} (\bar{u}_j \tanh^{-1}(v_j / \bar{u}_j))^T R_{ij} dv_j \\ &+ \nabla V_i^T (f(x) + \sum_{j=1}^{N} g_j(x) u_j) = 0, \; V_i(0) = 0, i \in N \end{aligned} \tag{6}$$

where $\nabla V_i = \partial V_i / \partial x \in \Re^{n_i}$ is the partial derivative of value function. According to the Bellman's optimality principle, the optimal control policies are obtained as

$$u_i^*(x) = -\bar{u}_i \tanh(\frac{1}{2\bar{u}_i} R_{ii}^{-1} g_i^T \nabla V_i^*), \quad i \in N \tag{7}$$

where $V_i^*, i \in N$ are solutions to the *N*-coupled Hamilton-Jacobi (HJ) equations

$$\begin{aligned} &Q_i(x) + \nabla V_i^{*T} f - \nabla V_i^{*T} \sum_{j=1}^{N} \bar{u}_j g_j(x) \tanh(D_j^*) \\ &+ \sum_{j=1}^{N} \nabla V_j^{*T} \bar{u}_j g_j R_{jj}^{-T} R_{ij} \tanh(D_j^*) \\ &+ \sum_{j=1}^{N} \bar{u}_j^2 \bar{R}_{ij} \ln(\underline{\mathbf{1}} - \tanh^2(D_j^*)) = 0; \quad V_i^*(0) = 0, i \in N \end{aligned} \tag{8}$$

where $D_j^* = (1/2\bar{u}_j) R_{jj}^{-1} g_j^T \nabla V_j^* \in \Re^{m_j}$, $\underline{\mathbf{1}}$ is a column vector with all its elements equal to 1, $\bar{R}_{ij} \in \Re^{1 \times m_j}$. Since the coupled HJ equations in (8) are difficult to solve, an approximate solution is sought.

B. Offline PI algorithm to solve coupled HJ equations

An offline PI algorithm is now introduced that solves the coupled HJ equations by iterating on the Lyapunov equations in (6). Note that for any time $t > T$, and any time interval $T > 0$, the value function (2) satisfies [9]

$$\begin{aligned} &V_i(x(t-T)) = \\ &\int_{t-T}^{t} (Q_i(x(\tau) + 2 \sum_{j=1}^{N} \int_0^{u_j} (\bar{u}_j \tanh^{-1}(v_j / \bar{u}_j))^T R_{ij} dv_j) d\tau + V_i(x(t)) \end{aligned} \tag{9}$$

where $i \in N$. It is shown in [9] that (6) and (9) have the same solution for the one player game. Using (9), the following PI algorithm can be used to solve the coupled HJ (8) using only partial knowledge of the system dynamics.

Offline PI Algorithm

Given initial admissible policies $u_1^0,...,u_N^0$, determine the computation accuracy ε

Do for $l = 0,1,2,...$

Solve for the constrained *N*-tuple of cost $V_1^l(x),...,V_N^l(x)$ using

$$V_i^l(x(t-T)) = V_i^l(x(t)) + \int_{t-T}^{t} (Q_i(x(\tau)) + 2\sum_{j=1}^{N} \int_0^{u_j^l} (\bar{u}_j \tanh^{-1}(v_j/\bar{u}_j))^T R_{ij} dv_j) d\tau;\ i \in N \tag{10}$$

Update the *N*-tuple of feedback control policies

$$u_i^{l+1} = -\bar{u}_i \tanh(\frac{1}{2\bar{u}_i} R_{ii}^{-1} g_i^T \nabla V_i^l);\ i \in N \tag{11}$$

While $\left\| V_i^{l-1} - V_i^{\infty} \right\| > \varepsilon$

Similar to [17], we solve the two-player NZS game. The method can be directly extended to more than two players. Consider the nonlinear affine in the input dynamical system

$$\dot{x} = f(x) + g_1(x)u_1(x) + g_2(x)u_2(x) \tag{12}$$

According to the Weirstrass high-order approximation theorem [24], there exist neural networks such that the solutions $V_i(x), i = 1,2$, and their gradients are uniformly approximated on a compact set Ω [25]

$$V_i(x) = W_i^T \sigma_i(x) + \varepsilon_i(x), i = 1,2 \tag{13}$$

$$\nabla V_i = \nabla \sigma_i^T(x) W_i + \nabla \varepsilon_i(x), i = 1,2 \tag{14}$$

where $\sigma_i(x) : \Re^n \to \Re^P$ are linearly independent activation function basis set vectors for the cost functions, $\varepsilon_i(x)$ are NN approximation errors, $W_i \in R^P$ are constant parameter vectors, and P is the number of neurons in the hidden layer.

Assumption 1 [17].

a. The NN activation functions and their gradients are bounded as $\left\|\sigma_i(x)\right\| \le b_{\sigma i}, \left\|\nabla \sigma_i(x)\right\| \le b_{\sigma ix}, i = 1,2$. The NN approximation errors and their gradients are bounded as $\left\|\varepsilon_i(x)\right\| \le b_{\varepsilon i}, \left\|\nabla \varepsilon_i(x)\right\| \le b_{\varepsilon ix}, \ i = 1,2$ over the compact set Ω.

b. The functions f, $g_i(.)$, are uniformly bounded on the compact set Ω, i.e., $\left\|f(x)\right\| \le b_f \left\|x\right\|$, $\left\|g_i(x)\right\| \le b_{gi}, i = 1,2$.

By using the value function approximation (VFA) NN in (13), the Bellman errors based on (9) becomes

$$\int_{t-T}^{t} \left(Q_i(x(\tau)) + M_i(u_1(\tau)) + M_i(u_2(\tau))\right) d\tau + W_i^T \Delta\sigma_i \equiv \varepsilon_{Bi} \tag{15}$$

where $\varepsilon_{Bi}\ i = 1,2$, are NZS Bellman equation errors due to the NN approximation error and are obtained as

$$\varepsilon_{B_i} = -\int_{t-T}^{t} \nabla \varepsilon_i^T \left(f + g_1 u_1 + g_2 u_2\right) d\tau,\ i = 1,2 \tag{16}$$

Note that

$$\Delta\sigma_i = \Delta\sigma_i(x(t)) = \int_{t-T}^{t} \nabla\sigma_i(x)(f + g_1 u_1 + g_2 u_2) d\tau,\ i = 1,2 \tag{17}$$

Next, using (15)-(17), the constrained coupled HJ equations associated with first and second player are

$$\int_{t-T}^{t} \Big(Q_1 + W_1^T \nabla\sigma_1 f + \bar{u}_1^2 \bar{R}_{11} \ln(\underline{1} - \tanh^2(D_1)) + \bar{u}_2^2 \bar{R}_{12} \ln(\underline{1} - \tanh^2(D_2)) + W_2^T \nabla\sigma_2 g_2 \bar{u}_2 R_{22}^{-T} R_{12} \tanh(D_2) - W_1^T \nabla\sigma_1 g_2 \bar{u}_2 \tanh(D_2) + \varepsilon_{CHJ1}\Big) d\tau = 0 \tag{18}$$

$$\int_{t-T}^{t} \Big(Q_2 + W_2^T \nabla\sigma_2 f + \bar{u}_1^2 \bar{R}_{21} \ln(\underline{1} - \tanh^2(D_1)) + \bar{u}_2^2 \bar{R}_{22} \ln(\underline{1} - \tanh^2(D_2)) + W_1^T \nabla\sigma_1 g_1 \bar{u}_1 R_{11}^{-T} R_{21} \tanh(D_1) - W_2^T \nabla\sigma_2 g_1 \bar{u}_1 \tanh(D_1) + \varepsilon_{CHJ2}\Big) d\tau \tag{19}$$

where $D_i = (1/2\bar{u}_i) R_{ii}^{-1} g_i^T \nabla\sigma_i^T W_i$, and $\varepsilon_{CHJi}, i = 1,2$, are the coupled HJ approximation errors due to the function approximation error. The coupled HJ approximation errors are bounded by constants, i.e., $\sup_{x\in\Omega} \left\|\varepsilon_{CHJi}\right\| \le \varepsilon_{him}$ [17].

Similarly, according to (7) the ideal optimal feedback control policies can be approximated as

$$u_i(x) = -\bar{u}_i \tanh(\frac{1}{2\bar{u}_i} R_{ii}^{-1} g_i^T \nabla\sigma_i^T W_i),\ i = 1,2 \tag{20}$$

III. Online ER RL Algorithm for NZS Game

An online ER-based algorithm based on the PI algorithm in the previous section is now given that solves the NZS games of constrained-input systems without using any knowledge on the internal system dynamics. An actor-critic structure is considered for each player, where both actor and critic are tuned online at the same time.

A. Critic NNs using the technique of ER

Since the ideal weights W_i are unknown, one must consider the current weight estimates. Hence, the output of the critic NNs can be written as

$$\hat{V}_i(x) = \hat{W}_i^T \sigma_i(x), i = 1,2 \tag{21}$$

The approximate Bellman equations become

$$e_i = \hat{W}_i^T \Delta\sigma_i(x(t)) + \int_{t-T}^{t} (Q_i + 2\int_0^{u_1} (\bar{u}_i \tanh^{-1}(v_1/\bar{u}_1))^T R_{i1} dv_1 + 2\int_0^{u_2} (\bar{u}_2 \tanh^{-1}(v_2/\bar{u}_2))^T R_{i2} dv_2) d\tau \tag{22}$$

where the errors e_i, $i=1,2$, are Temporal Difference (TD) errors and the objective is to bring them to their minimum values. In order to obtain a convergence of $\hat{W}_i$, to their optimal values and eliminate the need to ensure the PE condition, the idea of ER is employed [19-21]. By this, the critic NN weights for each player are adapted online using recorded past data, concurrently with current data. We define the TD errors corresponding to the previous data as

$$e_i(t_k)=\hat{W}_i^T\Delta\sigma_{ik}+\int_{t_k-T}^{t_k}\left(Q_i+M_i(u_1)+M_i(u_2)\right)d\tau \tag{23}$$

where $\Delta\sigma_{ik}=\Delta\sigma_i(t_k)=\sigma_i(x(t_k))-\sigma_i(x(t_k-T))$, and the integral term in (23) are evaluated at time history t_k, $k=1,...,l$, and are stored in history stack. The aim is to select $\hat{W}_i$, to minimize the following performance error

$$\begin{aligned}E=&\frac{1}{2}e_1^T(t)e_1(t)+\frac{1}{2}\sum_{k=1}^{l}e_1^T(t_k)e_1(t_k)\\&+\frac{1}{2}e_2^T(t)e_2(t)+\frac{1}{2}\sum_{k=1}^{l}e_2^T(t_k)e_2(t_k)\end{aligned} \tag{24}$$

where $l\in Z^+$ is the size of the history stack and Z^+ is the set of only positive integer numbers. The tuning for the critic NNs can be obtained by using a gradient-descent rule as

$$\begin{aligned}\dot{\hat{W}}_i=&-a_i\frac{\partial E}{\partial\hat{W}_i}=-a_i\frac{\Delta\sigma_i}{(\Delta\sigma_i^T\Delta\sigma_i+1)^2}e_i(t)\\&-a_2\sum_{k=1}^{l}\frac{\Delta\sigma_{ik}}{(\Delta\sigma_{ik}^T\Delta\sigma_{ik}+1)^2}e_i(t_k),\ i=1,2,\ t>t_k\ge 0\end{aligned} \tag{25}$$

where $a_i>0$, $i=1,2$ are the learning rates. The error dynamics for the critic NN weights can be written as

$$\begin{aligned}\dot{\tilde{W}}_i=&-a_i\left(\frac{\Delta\sigma_i\Delta\sigma_i^T}{(\Delta\sigma_i^T\Delta\sigma_i+1)^2}+\sum_{k=1}^{l}\frac{\Delta\sigma_{ik}\Delta\sigma_{ik}^T}{(\Delta\sigma_{ik}^T\Delta\sigma_{ik}+1)^2}\right)\tilde{W}_i(t)+\\&a_i\left(\frac{\Delta\sigma_i}{(\Delta\sigma_i^T\Delta\sigma_i+1)^2}\varepsilon_{Bi}(t)+\sum_{k=1}^{l}\frac{\Delta\sigma_{ik}}{(\Delta\sigma_{ik}^T\Delta\sigma_{ik}+1)^2}\varepsilon_{Bi}(t_k)\right),i=1,2\end{aligned} \tag{26}$$

where $\tilde{W}_i=W_i-\hat{W}_i$. Note that in (25) and (26) the time *t* is used for the current time and the index *k* is the *k*-th sample data stored in history stack.

Condition 1. By using the ER-based tuning laws (25), the only condition should be checked for parameter convergence is that the number of linearly independent data in $\mathrm{Y}_i=\left[\Delta\bar{\sigma}_i(t_1),...,\Delta\bar{\sigma}_i(t_k)\right]$, with $\Delta\bar{\sigma}_i=\Delta\sigma_i/(\Delta\sigma_i^T\Delta\sigma_i+1)$, corresponding to the first and second NN weights, should be equal to the dimension of the basis function in (13), that is, $rank(Y_i)=P,\ i=1,2$.

Technical Lemma 1. Let the critic NNs (13) with the ER-based update laws in (25) are used to evaluate the given admissible control policies u_i. Then, for $\varepsilon_{Bi}(t)=0$, $\hat{W}_i$, converge exponentially to the unknown weights W_i. Moreover, for bounded $\varepsilon_{Bi}(t)$, $\tilde{W}_i$ converge exponentially to the residual sets $R_{si}=\{\tilde{W}_i\,|\,\tilde{W}_i\le c_i\varepsilon_{mi}\}$, where ε_{mi} are bounds for $\varepsilon_{Bi}(t)$, and $c_i, i=1,2$ are constants.

Proof. Follows as in Modares et al. [21]. ■

B. Tuning of actor NNs

The objective of this section is to find the control feedback policies which minimize the approximated cost functions in (25). The ideal control policies according to (20) are unknown due to unknown weights W_i. Therefore, we define the actor NNs which compute the constrained control inputs as

$$\hat{u}_1(x)=-\bar{u}_1\tanh\left((1/2\bar{u}_1)R_{11}^{-1}g_1^T\nabla\sigma_1^T\hat{W}_3\right) \tag{27}$$

$$\hat{u}_2(x)=-\bar{u}_2\tanh\left((1/2\bar{u}_2)R_{22}^{-1}g_2^T\nabla\sigma_2^T\hat{W}_4\right) \tag{28}$$

where $\hat{W}_3$, $\hat{W}_4$ are current estimated values of the ideal weights W_1, W_2, and $\tilde{W}_3=W_1-\hat{W}_3$, $\tilde{W}_4=W_2-\hat{W}_4$ are actor NN estimation errors. Note that in order to guarantee the closed-loop stability, the control policies in (20) are defined in the form of nonstandard actor NNs in (27), (28).

Theorem 1 (Closed-loop stability). Consider the dynamical system given by (12), the feedback control policies given by (27), (28). Let Assumptions 1 holds and condition 1 be satisfied. Let the NN weight tuning laws for the critic NNs provided by

$$\begin{aligned}\dot{\hat{W}}_i=&-a_i\frac{\Delta\sigma_i}{\left(\Delta\sigma_i^T\Delta\sigma_i+1\right)^2}(\Delta\sigma_i^T\hat{W}_i(t)\\&+\int_{t-T}^{t}\left(Q_i+M_i(\hat{u}_1)+M_i(\hat{u}_2)\right)d\tau)\\&-a_i\sum_{k=1}^{l}\frac{\Delta\sigma_{ik}}{\left(\Delta\sigma_{ik}^T\Delta\sigma_{ik}+1\right)^2}(\Delta\sigma_{ik}^T\hat{W}_1(t)\\&+\int_{t_k-T}^{t_k}\left(Q_i+M_i(\hat{u}_1)+M_i(\hat{u}_2)\right)d\tau);\ i=1,2\end{aligned} \tag{29}$$

Let the actor NN tuning laws for the first and second player be provided by

$$\begin{aligned}\dot{\hat{W}}_3(t)=&-a_3\Bigg(B_1\hat{W}_3+\Pi_1\frac{\Delta\bar{\sigma}_1^T}{s_1}\hat{W}_1+\Pi_1'\frac{\Delta\bar{\sigma}_2^T}{s_2}\hat{W}_2\\&+\frac{h_1}{s_1^2}\Pi_1\int_{t-T}^{t}\left(\Pi_1^T\hat{W}_3\right)d\tau+\frac{h_2}{s_2^2}\Pi_1'\int_{t-T}^{t}\left(\Pi_1'^T\hat{W}_3\right)d\tau\Bigg)\end{aligned} \tag{30}$$

$$\begin{aligned}\dot{\hat{W}}_4(t)=&-a_4\Bigg(B_2\hat{W}_4+\Pi_2\frac{\Delta\bar{\sigma}_1^T}{s_1}\hat{W}_1+\Pi_2'\frac{\Delta\bar{\sigma}_2^T}{s_2}\hat{W}_2\\&+\frac{h_1}{s_1^2}\Pi_2\int_{t-T}^{t}\left(\Pi_2^T\hat{W}_4\right)d\tau+\frac{h_2}{s_2^2}\Pi_2'\int_{t-T}^{t}\left(\Pi_2'^T\hat{W}_4\right)d\tau\Bigg)\end{aligned} \tag{31}$$

where $h_i = a_i T^3/3$, $s_i = \Delta\sigma_i^T \Delta\sigma_i + 1$, $\Delta\bar{\sigma}_i = 1/(\Delta\sigma_i^T \Delta\sigma_i + 1)$, for $i = 1, 2$, and $B_1 > 0$, $B_2 > 0$ are tuning parameters, and

$$\Pi_1 = \nabla\sigma_1 g_1 \bar{u}_1 \left(\tanh(\delta\hat{D}_1) - \tanh(\hat{D}_1)\right) \quad (32)$$

$$\Pi_2 = \nabla\sigma_2 g_2 \bar{u}_2 R_{22}^{-T} R_{12} \left(\tanh(\delta\hat{D}_2) - \tanh(\hat{D}_2)\right) \quad (33)$$

$$\Pi_1' = \nabla\sigma_1 g_1 \bar{u}_1 R_{11}^{-T} R_{21} \left(\tanh(\delta\hat{D}_1) - \tanh(\hat{D}_1)\right) \quad (34)$$

$$\Pi_2' = \nabla\sigma_2 g_2 \bar{u}_2 \left(\tanh(\delta\hat{D}_2) - \tanh(\hat{D}_2)\right) \quad (35)$$

where

$\hat{D}_1 = (1/2\bar{u}_1) R_{11}^{-1} g_1^T \nabla\sigma_1^T \hat{W}_3$, $\hat{D}_2 = (1/2\bar{u}_2) R_{22}^{-1} g_2^T \nabla\sigma_2^T \hat{W}_4$.

Then, the closed-loop system states *x*, the critic NN errors $\tilde{W}_1$, $\tilde{W}_2$, and the actor NN errors $\tilde{W}_3$, $\tilde{W}_4$, are all UUB for sufficiently large number of NN neurons, provided that

$$a_i < 3/T^3 \quad (36)$$

$$B_i > (l+4)\left((a_1/2s_1^2)\Pi_i \Pi_i^T + (a_2/2s_2^2)\Pi_i' \Pi_i'^T\right), i = 1, 2 \quad (37)$$

Proof. Proof is not provided due to page limitation. ■

Remark 1. In the proposed algorithm the players in the game learn online the optimal policies corresponding to the Nash equilibrium of the game without using the knowledge on the internal dynamics of the system.

Remark 2. Note that for the ER-based PI algorithm the restrictive PE condition is relaxed to the condition 1 that can easily be checked online.

Theorem 2 (Convergence). Given that the assumptions in Theorem 1 hold, then the actor and critic NNs converge to the approximate coupled HJ solution.

Proof. Consider all the UUB weight errors in Theorem 1. The approximate constrained coupled HJ equations associated with first player is

$$H_1\left(x, \hat{W}_1, \hat{u}_1, \hat{u}_2\right) = \int_{t-T}^{t} \Big(Q_1 + \hat{W}_1^T \nabla\sigma_1 f - \hat{W}_1^T \nabla\sigma_1 g_1 \bar{u}_1 \tanh(\hat{D}_1) - \hat{W}_1^T \nabla\sigma_1 g_2 \bar{u}_2 \tanh(\hat{D}_2) + \hat{W}_3^T \nabla\sigma_1 g_1 \bar{u}_1 \tanh(\hat{D}_1) - \hat{W}_3^T \nabla\sigma_1 g_1 \bar{u}_1 \tanh(\delta\hat{D}_1) + \hat{W}_4^T \nabla\sigma_2 g_2 \bar{u}_2 R_{22}^{-T} R_{12} \tanh(\hat{D}_2) - - \hat{W}_4^T \nabla\sigma_2 g_2 \bar{u}_2 R_{22}^{-T} \tanh R_{12}(\delta\hat{D}_2) + \bar{u}_1^2 R_{11} \varepsilon_{\hat{D}_1} + \bar{u}_2^2 R_{12} \varepsilon_{\hat{D}_2} \Big) d\tau \quad (38)$$

and likewise is derived for the second player. After adding zero from HJ equations in (18) and (19) and using the fact that $\tilde{W}_1 = W_1 - \hat{W}_1$, $\tilde{W}_3 = W_1 - \hat{W}_3$, $\tilde{W}_4 = W_2 - \hat{W}_4$, one has

$$\begin{aligned} H_1\left(x, \hat{W}_1, \hat{u}_1, \hat{u}_2\right) = \int_{t-T}^{t} \Big(& - \tilde{W}_1^T \nabla\sigma_1 f + \tilde{W}_1^T \nabla\sigma_1 g_1 \bar{u}_1 \tanh(\hat{D}_1) \\ & + \tilde{W}_1^T \nabla\sigma_1 g_2 \bar{u}_2 \tanh(\hat{D}_2) + \tilde{W}_3^T \nabla\sigma_1 g_1 \bar{u}_1 \left(\tanh(\delta\hat{D}_1) - \tanh(\hat{D}_1)\right) \\ & + \tilde{W}_4^T \nabla\sigma_2 g_2 \bar{u}_2 R_{22}^{-T} R_{12} \left(\tanh(\delta\hat{D}_2) - \tanh(\hat{D}_2)\right) \\ & + W_1^T \nabla\sigma_1 g_2 \bar{u}_2 \left(\tanh(D_2) - \tanh(\hat{D}_2)\right) \\ & + W_1^T \nabla\sigma_1 g_1 \bar{u}_1 \left(\tanh(\delta D_1) - \tanh(\delta\hat{D}_1)\right) \\ & + W_2^T \nabla\sigma_2 g_2 \bar{u}_2 R_{22}^{-T} R_{12} \left(\tanh(\delta D_2) - \tanh(\delta\hat{D}_2)\right) \\ & + W_2^T \nabla\sigma_2 g_2 \bar{u}_2 R_{22}^{-T} R_{12} \left(\tanh(\hat{D}_2 - \tanh(D_2)\right) \\ & + \bar{u}_1^2 R_{11} \varepsilon_{\hat{D}_1} + \bar{u}_2^2 R_{12} \varepsilon_{\hat{D}_2} - \bar{u}_1^2 R_{11} \varepsilon_{D_1} - \bar{u}_2^2 R_{12} \varepsilon_{D_2} + \varepsilon_{CHJ1} \Big) d\tau \end{aligned} \quad (39)$$

and similarly is obtained for the second player. Now using Assumption 1, taking norms in (39) reveals

$$\begin{aligned} \left\| H_1\left(x, \hat{W}_1, \hat{u}_1, \hat{u}_2\right) \right\| \le \int_{t-T}^{t} \Big(& b_f b_{\sigma 1x} \|x\| \left\|\tilde{W}_1^T\right\| + \\ & b_{\sigma 1x} \left(b_{g1}\bar{u}_1 + b_{g2}\bar{u}_2\right) \left\|\tilde{W}_1^T\right\| + b_{\sigma 1x} b_{g1} \bar{u}_1 \left\|\tilde{W}_3^T\right\| + \\ & b_{\sigma 2x} b_{g2} \bar{u}_2 \lambda_{\max}(R_{22}^{-T}) \lambda_{\max}(R_{12}) \left\|\tilde{W}_4^T\right\| + b_{\sigma 1x} b_{g2} \bar{u}_2 \left\|W_1^T\right\| + \\ & b_{\sigma 1x} b_{g1} \bar{u}_1 \left\|W_1^T\right\| + 2 b_{\sigma 2x} b_{g2} \bar{u}_2 \lambda_{\max}(R_{22}^{-T}) \lambda_{\max}(R_{12}) \left\|W_2^T\right\| + \|\varepsilon_1\| \Big) d\tau \end{aligned} \quad (40)$$

where

$\|\varepsilon_1\| \le \bar{u}_1^2 \lambda_{\max}(R_{11}) \left(\varepsilon_{\hat{D}_1} - \varepsilon_{D_1}\right) + \bar{u}_2^2 \lambda_{\max}(R_{12}) \left(\varepsilon_{\hat{D}_2} - \varepsilon_{D_2}\right) + \varepsilon_{h1m}$

, and ε_{h1m}, is bounds for ε_{CHJ1}. One can likewise derive for the second player. All the signals on the right hand side of (40) is UUB. Therefore, $\left\|H_1(x, \hat{W}_1, \hat{u}_1, \hat{u}_2)\right\|$ and $\left\|H_2(x, \hat{W}_2, \hat{u}_1, \hat{u}_2)\right\|$ are UUB and convergence to the approximate coupled HJ solutions is obtained. This completes the proof. ■

Theorem 3. Given that the assumptions in Theorem 1 hold, then $\hat{u}_1$ and $\hat{u}_2$ converge to the approximate Nash equilibrium solution of the NZS game.

Proof. Consider $\hat{u}_1$ and $\hat{u}_2$ in (27) and (28). Then one has

$$\begin{aligned} & \|u_1 - \hat{u}_1\| = \\ & \le \bar{u}_1 \left\| - \tanh(1/(2\bar{u}_1) R_{11}^{-1} g_1^T \nabla\sigma_1^T W_1) + \right. \\ & \left. \tanh(1/(2\bar{u}_1) R_{11}^{-1} g_1^T \nabla\sigma_1^T (W_1 - \tilde{W}_3)) \right\| \end{aligned} \quad (41)$$

$$\begin{aligned} & \|u_2 - \hat{u}_2\| = \\ & \le \bar{u}_1 \left\| - \tanh\left(1/(2\bar{u}_2) R_{22}^{-1} g_2^T \nabla\sigma_2^T W_2\right) + \right. \\ & \left. \tanh\left(1/(2\bar{u}_2) R_{22}^{-1} g_2^T \nabla\sigma_2^T \left(W_2 - \tilde{W}_4\right)\right) \right\| \end{aligned} \quad (42)$$

Hence, $\|u_1 - \hat{u}_1\|$ and $\|u_2 - \hat{u}_2\|$ are UUB. Therefore, the pair $(\hat{u}_1, \hat{u}_2)$ gives the approximate Nash equilibrium solution of the NZS game and this completes the proof. ■

IV. Simulation Study

Let us consider the following affine in input nonlinear system [17]

$$f = \begin{bmatrix} x_2 \\ -x_2 - \frac{1}{2}x_1 + \frac{1}{4}x_2(\cos(2x_1)+2)^2 + \frac{1}{4}x_2\sin((4x_1^2)+2)^2 \end{bmatrix}$$
$$g_1 = [0,\ \cos(2x_1)+2]^T$$
$$g_2 = [0,\ \sin(4x_1^2)+2)]^T \tag{43}$$

We use the online PI algorithm in theorem 2 to solve the two-player NZS game for the system (43). The objective is to control the system (43) with the control constraints of $|u_1| \le 0.59$, $|u_2| \le 0.59$. The weighting matrices are selected as $Q_1 = 2I$, $Q_2 = I$, $R_{11} = 1$, $R_{12} = 2$, $R_{21} = 1$, $R_{22} = 0.5$, where I is the identity matrix of appropriate dimension. The tuning matrices are $B_1 = B_2 = 40I$. The critic NN activation functions are generated from a second order polynomial order as $\sigma_1(x) = \sigma_2(x) = [x_1^2 \quad x_1x_2 \quad x_2^2]^T$. During learning process, a probing noise is added to the both control inputs to excite the system states. The simulation is run for 300 seconds. At the end of the learning process the parameters of the first critic NN weights converge to $\hat{W}_1 = [0.3420\ 0.2647\ 1.147]$, while the parameters of second critic NN converge to $\hat{W}_2 = [0.2955\ 0.3336\ 1.0470]$. Figs 1, 2 show convergence of the first and second critic NN weights. Fig 3 demonstrates the states of the system during online learning process. The probing noise is no longer required and is thus removed after 250s. Then, the system states converge to zero.

It can be observed from figs 1, 2 that the convergence of the critic NN weights has occurred after nearly 150s, and the proposed online algorithm has determined the solution of the game without requirement to ensure the PE condition.

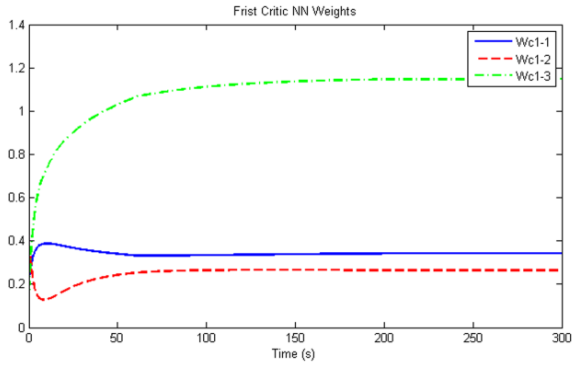

Figure 1. Convergence of the first critic NN parameters.

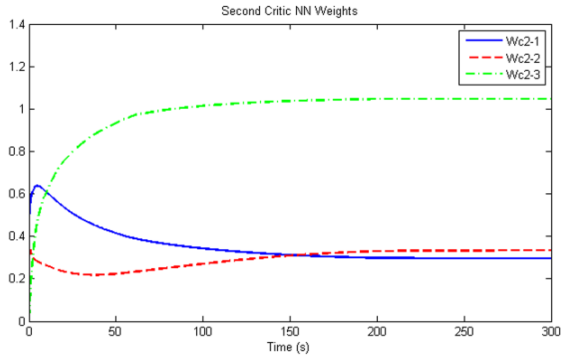

Figure 2. Convergence of the second critic NN parameters.

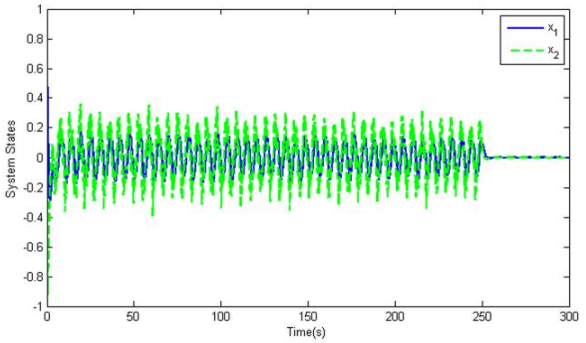

Figure 3. System states during learning process.

V. Conclusion

In this paper, an online ER-based RL algorithm was developed to solve coupled HJ equations in NZS games for constrained-input partially-unknown systems. By using the technique of ER the need to assure the PE condition for convergence the algorithm was relaxed to a simplified condition on recorded data. The stability of the closed-loop system in the presence of bounded NN reconstruction errors was guaranteed and convergence of the proposed algorithm to the Nash equilibrium of the NZS game was shown.

References

[1] A. Starr, Y. Ho, "Nonzero-sum differential games," *J. Opt. Theory App*, vol. 3, pp. 184-206, 1969.

[2] T. Basar, G.J. Olsder, *Dynamic Noncooperative Game Theory*. 2nd Edition, Philadelphia, PA: SIAM, SIAM's Classic in Applied Mathematics, 1999.

[3] H. Abu-Kandil, G. Freiling, G. Jank, "Necessary and sufficient conditions for constant solutions of coupled Riccati equations in Nash games," *Systems and Control Letters*, vol. 21, 2003, pp. 295-306.

[4] G. Freiling, G. Jank, H. Abu-Kandil, "On global existence of solutions to coupled Matrix Riccati equations in closed-loop Nash games," *IEEE Trans. On Automatic Control*, vol. 41, pp. 264-269, 1996.

[5] T. Li, Z. Gajic, "Lyapanuv iterations for solving coupled Riccati equations for Nash differential games and algebraic Riccati equations for zero-sum games," in New Trends in Dynamic Games, G. Olsder (ed), Birkhauser, 1995, pp. 333-351.

[6] H. Mukaidani, "Numerical computation of sign-indefinite linear quadratic differential games for weakly coupled large scale systems," *Int. J. of Control*, vol. 80, pp. 75-86, 2007.

[7] R. A. Howard, *Dynamic Programming and Markov Processes*. Cambridge, MA: MIT Press, 1960.

[8] F. L. Lewis, D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits System Magazine*. vol. 9, pp. 32–50, 2009.
[9] D. Vrabie, F. L. Lewis, "Neural network approach to continuous time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*. vol. 22, pp. 237–246, 2009.
[10] K. Vamvoudakis, F. L. Lewis, "Online actor-critic algorithm to solve continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, pp. 787-788, 2010.
[11] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and D. Dixon, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica,* vol. 59, no. 1, pp. 82-92, 2013.
[12] K. Vamvoudakis, F. L. Lewis, "Online neural network solution of nonlinear two-player zero-sum games using synchronous policy iteration," *IEEE Conf. on Decision and Control*, Hilton Atlanta Hotel, Atlanta, GA, USA, 2010, pp. 3040-3047.
[13] H. Modares, F. L. Lewis, M. B. Naghibi-Sistani, "Online solution of nonquadratic two-player zero-sum games arising in H_∞ control of constrained input systems," *Int. J. of Adaptive Control and Signal Processing*, 2012, DOI: 10.1002/acs.2348.
[14] M. Johnson, S. Bhasin, W. E. Dixon, "Nonlinear two-player zero-sum game approximate solution using a policy iteration algorithm," *IEEE Conf. on Decision and Control and European Control Conference,* Orlando, USA, 2011, pp. 142-147.
[15] D. Vrabie, F. L. Lewis, "Integral reinforcement learning for online computation of feedback Nash strategies of nonzero-sum differential games," *49th IEEE Conference on Decision and Control*, Atlanta, GA, USA, 2010.
[16] K. Vamvoudakis, F. L. Lewis, "Non-zero sum games: Online learning solution of coupled Hamilton-Jacobi and coupled Riccati equations," *IEEE Int. Symposium on Intelligent Control*, Denver, Co, USA, pp. 171-178, 2011.
[17] K. Vamvoudakis, F. L. Lewis, "An online integral reinforcement learning algorithm to N-player Nash games," *IEEE Int. Symposium on Intelligent Control* , Dubrovnik, Croatia, pp. 797-702, 2012.
[18] H. Zhang, L. Cui, Y. Luo, "Near-optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network ADP," *IEEE Trans. Cybern*, vol. 45, pp. 206-216, 2013.
[19] G. V. Chowdhary, "Concurrent learning for convergence in adaptive control without persistency of excitation," PhD Thesis at Georgia Institute of Technology, 2010.
[20] H. Modares, F. L. Lewis, M. B. Naghibi-Sistani, G. Chowdhary, T. Yucelen, "Adaptive optimal control for the partially-unknown constrained-input using policy iteration with experience replay," *AIAA Guidance, Navigation, and Control Conference*, Boston, Massachusetts, August 2013.
[21] 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, pp. 193-202, 2014.
[22] F. L. Lewis, D. Vrabie, V. L. Syrmos, *Optimal control*. Wiley, 2012.
[23] S. E. Lyshevski, "Optimal control of nonlinear continuous-time systems: design of bounded controllers via generalized nonquadratic functionals," *In Proc. American Control Conference*, pp. 205-209, 1998.
[24] K. Hornik, M. Stinchcombe, H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks", *Neural Networks*, vol. 3, pp. 551–560, 1990.
[25] P. J. Werbos, "Approximate dynamic programming for real-time control and neural modeling," in D. A. White, D. A. Sofge (Eds.), Handbook of Intelligent Control. New York: Van Nostrand Reinhold. 1992.