Model-Free Reinforcement Learning for Nonlinear Zero-Sum Games with Simultaneous Explorations

Qichao Zhang, Dongbin Zhao, Yuanheng Zhu
The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
Email: zhangqichao2013@163.com, dongbin.zhao@ia.ac.cn, zyh7716155@163.com

Xi Chen
Mail Box 150
Baoji, China
Email: zhangqichao2013@163.com,

***Abstract*—In this paper, the continuous-time unknown nonlinear zero-sum game is investigated using a model-free online learning method. First, motivated by model-based policy iteration, an iterative equation without any knowledge of system dynamics is derived by introducing simultaneous explorations. Then, the model-free reinforcement learning based on the derived iterative equation is developed to approach the solution of the Hamilton-Jacobi-Isaacs equation. For the online implementation purpose, three neural networks are constructed to approach the value function, control and disturbance policies, respectively. Finally, a simulation example is provided to demonstrate the effectiveness of the proposed scheme.**

I. Introduction

As an important part of game theory, zero-sum game has been investigated in various fields such as economics and management sciences [1], behavioral ecology [2] and control theory [3], etc. The two-player zero-sum game is regarded as a kind of typical completely competitive game, where one player is to minimize the performance index while the other is a maximizing one. Since it can provide an effective method to study the H_∞ control problem which relies on solving the Hamilton-Jacobi-Isaacs (HJI) equation for the dynamical systems with disturbances, many existing works on the two-player zero-sum game have been studied like [4], [5]. To find the Nash equalibrium of the zero-sum game, the game algebraic Riccati equation should be solved for linear systems while the HJI equation for nonlinear systems. Unfortunately, the HJI equation is intractable to be solved due to the inherent nonlinearity.

Recently, adaptive/approximate dynamic programming (ADP) and reinforcement learning (RL) have been widely used to approximately solve the HJI equation [6], [7]. For the zero-sum game with completely known system dynamics, Abu-Khalaf et al. propose an off-line inter-outer-loop policy iteration (PI) to solve the constrained HJI equation in [8]. Wu and Luo put forward an online simultaneous policy update algorithm (SPUA) with only one iterative loop in [9]. For the zero-sum game with partially unknown dynamics, an integral reinforcement learning (IRL) technique is used to relieve the dependence of the internal dynamics for linear systems in [10]. Luo et al. develop an off-policy RL for nonlinear systems with unknown internal dynamics using real system data in [11]. This off-policy RL is proved to be equivalent to the SPUA.

To the best of our knowledge, there are two kinds of common methods to solve the zero-sum game with completely unknown system dynamics. The first one is to identify or reconstruct the system model using neural networks (NNs) like [12], [13], then model-based algorithms can be used for the identification system. However, this method is detrimental to find the optimal solution due to the existence of the identifier error. Another one are model-free algorithms like Q-learning and some data-driven algorithms. Zhu and Zhao in [14] develop a model-free iterative ADP with the critic-actor-disturbance structure for unknown nonlinear systems using real system data, which is proved to be a Gauss-Newton method.

It should be noted that this kind of data-driven algorithm in [14] is an off-policy learning method. In [15], a model-free online IRL algorithm with simultaneous explorations is developed for the input-affine nonlinear systems, where the system dynamics and identification process are both not required. This online learning scheme is extended to the linear zero-sum game with completely unknown dynamics in [16]. And the proposed IRL and integral Q-learning algorithms for the linear zero-sum game are proved to be equivalent to the Newton's method.

In this paper, we extend the work in [16] to the continuous-time nonlinear zero-sum game without the exact system dynamics. An online model-free RL algorithm with simultaneous explorations under critic-actor-disturbance structure is developed to learn the Nash equilibrium solution of the HJI equation. The critic-actor-disturbance NNs are updated simultaneously in the developed algorithm and the unknown parameters can be determined based on the least-squares (LS) principle.

The rest of the paper is mainly organized as follows. Section II introduces the zero-sum game and the model-based SPUA algorithm to the corresponding HJI equation. A model-free RL algorithm to the unknown zero-sum game is given with only one iterative equation and the convergence analysis of each iteration in the model-free RL algorithm with the neural network approximations is proposed in Section III. Section IV demonstrates the performance of the developed algorithm by a simulation example, and the conclusion is given in section V.

This work is supported by National Natural Science Foundation of China (NSFC) under Grants No. 61273136, No. 61573353 and No. 61533017.

978-1-5090-0620-5/16/$31.00 ©2016 IEEE

II. PRELIMINARY

A. Continuous-time nonlinear zero-sum game

Consider the continuous-time nonlinear dynamical systems described by

$$\begin{aligned} \dot{x} &= f(x) + g(x)u + k(x)w \\ z &= h(x) \end{aligned} \tag{1}$$

where $x \in \Omega \in \Re^n$ is the state vector, $u(t) \in \Re^m$ is the control input, $w(t) \in \Re^q$ is the external disturbance with $w(t) \in L_2(0,\infty)$, $z \in \Re^p$ is the fictitious output, $f(x) \in \Re^n$, $g(x) \in \Re^{n\times m}$, $k(x) \in \Re^{n\times q}$ and $h(x) \in \Re^p$ are the continuous smooth system dynamics. Assume $f(x)+g(x)u+k(x)w$ is Lipschitz continuous with $f(0) = 0$ and $w(0) = 0$, thus $x = 0$ is an equilibrium of system (1).

For this zero-sum game, define the performance index

$$J(x(0),u,w) = \int_0^\infty (h^{\mathrm{T}}(x)h(x) + u^{\mathrm{T}}Ru - \gamma w^{\mathrm{T}}w)\mathrm{d}\tau$$

where $R = R^{\mathrm{T}} > 0$, $\gamma \geq \gamma^* \geq 0$ and $x(0) = 0$. Here, γ^* represents the smallest γ for which the system (1) is stabilized. And we aim to find the controller which renders the performance index nonpositive for all $w \in L_2(0,\infty)$. That is, the L_2 gain of system (1) is less than or equal to γ.

For a fixed control policy $u(t) = u(x(t))$ and a disturbance policy $w(t) = w(x(t))$, the value function is defined as

$$V(x) = \int_0^\infty (h^{\mathrm{T}}h + u^{\mathrm{T}}Ru - \gamma^2 w^{\mathrm{T}}w)\mathrm{d}\tau = \int_0^\infty r(x,u,w)\mathrm{d}\tau$$

The Bellman equation for the zero-sum game is

$$r(x,u,w)+\nabla V^{\mathrm{T}}(f(x)+g(x)u+k(x)w) = 0, V(0) = 0 \tag{2}$$

where $\nabla V = \partial V/\partial x \in \Re^n$ denotes the gradient. Then the Nash equilibrium (u^*, w^*) which is a unique solution exists if the Nash condition holds

$$\min_u \max_w J(x,u,w) = \max_w \min_u J(x,u,w)$$

with the optimal value

$$V^*(x) = \min_u \max_w J(x,u,w)$$

Define the Hamiltonian function

$$H(x,\nabla V,u,w) = r(x,u,w) + \nabla V^{\mathrm{T}}(f(x) + g(x)u + k(x)w)$$

Then, the HJI equation can be given by

$$\min_u \max_w H(x,\nabla V^*,u,w) = 0 \tag{3}$$

Applying the stationary conditions, we can obtain the optimal control and disturbance policies

$$\begin{aligned} u^*(x) &= -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla V^* \\ w^*(x) &= \frac{1}{2}\gamma^{-2}k^{\mathrm{T}}(x)\nabla V^* \end{aligned}$$

After substituting $u^*(x)$ and $w^*(x)$ into (3), the HJI equation can be rewritten as

$$\begin{aligned} &\nabla V^{*\mathrm{T}} f(x) + h^{\mathrm{T}}h - \frac{1}{4}\nabla V^{*\mathrm{T}} g(x)R^{-1}g^{\mathrm{T}}(x)\nabla V^* \\ &+ \frac{1}{4}\nabla V^{*\mathrm{T}}\gamma^{-2}k(x)k^{\mathrm{T}}(x)\nabla V^* = 0, V^*(0) = 0 \end{aligned} \tag{4}$$

Then we have to solve the HJI equation (4) for the optimal value function V^* and the optimal policies u^* and w^*. However, it is difficult to obtain an analytic solution of (4) directly for nonlinear systems.

B. Model-based method to the HJI equation

To obtain the optimal value function and optimal policies, it is essential to approach the solution to the HJI equation. In order to facilitate comparison, we first give the model-based SPUA with modification in [9].

Algorithm 1 SPUA

1: Start with two initial stabilizing policies u_0, w_0. Let $i = 0$.
2: Solve the following Lyapunov equation with $V_i(0) = 0$

$$\nabla V_i^{\mathrm{T}}(f(x)+g(x)u_i+k(x)w_i)+h^{\mathrm{T}}h+u_i^{\mathrm{T}}Ru_i-\gamma^2 w_i^{\mathrm{T}}w_i = 0 \tag{5}$$

3: Update the control and disturbance policies

$$\begin{aligned} u_{i+1}(x) &= -\frac{1}{2}R^{-1}g^{\mathrm{T}}(x)\nabla V_i \\ w_{i+1}(x) &= \frac{1}{2}\gamma^{-2}k^{\mathrm{T}}(x)\nabla V_i \end{aligned} \tag{6}$$

4: Set $i = i + 1$ and go to step 2.

In [9], SPUA is proved to be equivalent to the Newton's iteration method, which means that this algorithm can converge to the solution of the HJI equation as the iteration index i goes to infinity. It should be mentioned that this algorithm depends on the complete knowledge of the system dynamics. Even though the off-policy learning scheme in [11] can relieve the dependence of the internal dynamics $f(x)$, the drift dynamics $g(x)$ and $k(x)$ are still required.

III. AN ONLINE NEURAL-NETWORK BASED APPROACH TO THE HJI EQUATION WITH SIMULTANEOUS EXPLORATIONS

A. Model-free RL with simultaneous explorations

To solve the HJI equation for zero-sum game with completely unknown dynamics, a model-free RL algorithm is developed in this subsection. Consider the nonlinear zero-sum game with two external simultaneous explorations as follows

$$\begin{aligned} \dot{x} &= f(x) + g(x)(u + e_u) + k(x)(w + e_w) \\ z &= h(x) \end{aligned} \tag{7}$$

where $e_u \in \Re^m$ and $e_w \in \Re^q$ denote the exploration signals added to the control and disturbance policies, respectively. According to [15], the following definition about the exploration signals is given.

Definition 1 [15]: For given stabilizing control and disturbance policies (u, w), the exploration singles (e_u, e_w) is said

to be invariantly admissible on Ω, if they can stabilize the system (7) in the compact region Ω.

We call the exploration singles in Definition 1 Ω-invariant exploration signals. Given two initial stabilizing control policies (u_0, w_0) and the Ω-invariant exploration signals (e_u, e_w) for system (7), the derivative of $V_i(x)$ with respect to time for the i-th iteration is calculated as

$$\dot{V}_i(x) = \nabla V_i^{\mathrm{T}} \left(f(x) + g(x)(u_i + e_u) + k(x)(w_i + e_w) \right)$$

Based on (5) and (6), we have

$$\begin{aligned} \dot{V}_i(x) =& \nabla V_i^{\mathrm{T}} g(x) e_u + \nabla V_i^{\mathrm{T}} k(x) e_w - r(x, u_i, w_i) \\ =& -2u_{i+1}^{\mathrm{T}} R e_u + 2\gamma^2 w_{i+1}^{\mathrm{T}} e_w - r(x, u_i, w_i) \end{aligned} \tag{8}$$

where $V_i(x)$, u_{i+1} and w_{i+1} are the unknown function and vectors be solved for, and $V_i(0) = 0$. Integrating (8) on the interval $[t, t + \Delta t]$ with $\Delta t > 0$, we have

$$\begin{aligned} & V_i(x(t)) - V_i(x(t + \Delta t)) - \int_t^{t+\Delta t} 2u_{i+1}^{\mathrm{T}} R e_u \mathrm{d}\tau \\ & + \int_t^{t+\Delta t} 2\gamma^2 w_{i+1}^{\mathrm{T}} e_w \mathrm{d}\tau = \int_t^{t+\Delta t} r(x, u_i, w_i) \mathrm{d}\tau \end{aligned} \tag{9}$$

Then, the solution of the HJI equation can be obtained iteratively following (9) with given initial policies and exploration signals. Until now, the model-free RL with simultaneous explorations can be developed by solving the iterative equation (9) instead of the iterative equations (5) and (6).

Remark 1: Evidently, the real system data u_i and w_i are required in the equation (9) rather than the system dynamics $f(x), g(x)$ and $k(x)$, which means that the proposed algorithm is a model-free approach. In fact, the knowledge of system dynamics is embedded in the real system data in (9).

Remark 2: To solve the HJI equation, two Ω-invariant nonzero exploration signals are added to the control and disturbance policies in the developed algorithm. So the system trajectories are produced with the policies $u_i + e_u$ and $w_i + e_w$. In fact, adding probing noises to the control inputs is also a common and effective approach to guarantee the persistence of excitation (PE) condition. In some degree, the equation (9) is equal to the equations (5) and (6). Compared with the probing noises in the SPUA which are used to guarantee the PE condition, the simultaneous explorations in our algorithm can not only guarantee the PE condition but also help to relieve the dependence of the system dynamics. According to [16], the most common types of exploration signals have exponentially decreasing probing noises, sinusoidal signals with different frequencies and random noises.

B. NN-based implementation

For implementation purposes, neural networks are constructed to approach the solution of (9) with the critic-actor-disturbance structure. According to the Weirstrass high-order approximation theorem [17], the solution (V_i, u_{i+1}, w_{i+1}) of (9) can be uniformly approximated on the compact set Ω by

$$\begin{aligned} V_i(x) &= W_{c,i+1}^{\mathrm{T}} \phi_c(x) + \varepsilon_{c,i+1} \\ u_{i+1}(x) &= W_{u,i+1}^{\mathrm{T}} \phi_u(x) + \varepsilon_{u,i+1} \\ w_{i+1}(x) &= W_{w,i+1}^{\mathrm{T}} \phi_w(x) + \varepsilon_{w,i+1} \end{aligned} \tag{10}$$

where $\phi_c : \Re^n \to \Re^{K_c}$, $\phi_u : \Re^n \to \Re^{K_u}$ and $\phi_w : \Re^n \to \Re^{K_w}$ are linearly independent basis function vectors, $W_{c,i+1} \in \Re^{K_c}$, $W_{u,i+1} \in \Re^{K_u \times m}$ and $W_{w,i+1} \in \Re^{K_w \times q}$ are the unknown coefficient vector and matrices with K_c, K_u and K_w the numbers of hidden neurons, $\varepsilon_{c,i+1}$, $\varepsilon_{u,i+1}$ and $\varepsilon_{w,i+1}$ are the reconstruction errors with appropriate dimensions. It is shown in [18] that as $K_c \to \infty, K_u \to \infty$, $K_w \to \infty$, the reconstruction errors $\varepsilon_{c,i+1}$, $\varepsilon_{u,i+1}$, $\varepsilon_{w,i+1}$ converge to zero.

As the ideal coefficients are unknown, the actual output of the three NNs with a group of estimations can be given by

$$\begin{aligned} \hat{V}_i(x) &= \hat{W}_{c,i+1}^{\mathrm{T}} \phi_c(x) \\ \hat{u}_{i+1}(x) &= \hat{W}_{u,i+1}^{\mathrm{T}} \phi_u(x) \\ \hat{w}_{i+1}(x) &= \hat{W}_{w,i+1}^{\mathrm{T}} \phi_w(x) \end{aligned} \tag{11}$$

Given a strictly increasing time sequence $\{t_k\}_{k=0}^{l}$ for each interval with $l > 0$, the residual error by substituting (11) into (9) is defined as

$$\begin{aligned} e_k =& \left(\phi_c\left(x(t_{k+1})\right) - \phi_c\left(x(t_k)\right) \right)^{\mathrm{T}} \hat{W}_{c,i+1} \\ & + \int_{t_k}^{t_{k+1}} 2\phi_u^{\mathrm{T}} \hat{W}_{u,i+1} R e_u \mathrm{d}\tau - \int_{t_k}^{t_{k+1}} 2\gamma^2 \phi_w^{\mathrm{T}} \hat{W}_{w,i+1} e_w \mathrm{d}\tau \\ & + \int_{t_k}^{t_{k+1}} r(x, \hat{W}_{u,i}^{\mathrm{T}} \phi_u, \hat{W}_{w,i}^{\mathrm{T}} \phi_w) \mathrm{d}\tau \end{aligned} \tag{12}$$

By Kronecker product $\otimes$, we have

$$\begin{aligned} \phi_u^{\mathrm{T}} \hat{W}_{u,i+1} R e_u &= (e_u^{\mathrm{T}} R \otimes \phi_u^{\mathrm{T}}) \mathrm{vec}(\hat{W}_{u,i+1}) \\ \phi_w^{\mathrm{T}} \hat{W}_{w,i+1} e_w &= (e_w^{\mathrm{T}} \otimes \phi_w^{\mathrm{T}}) \mathrm{vec}(\hat{W}_{w,i+1}) \end{aligned}$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization of a matrix formed by stacking the columns of the matrix into a single column vector.

Define $\tilde{W}_i = [\hat{W}_{c,i}^{\mathrm{T}}, \mathrm{vec}(\hat{W}_{u,i})^{\mathrm{T}}, \mathrm{vec}(\hat{W}_{w,i})^{\mathrm{T}}]^{\mathrm{T}}$, a compact form of (12) can be written as

$$e_k = \theta_k^{\mathrm{T}}(x) \tilde{W}_{i+1} + \xi_k(\tilde{W}_i) \tag{13}$$

where $\tilde{W}_i \in \Re^{\tilde{K}}$ with $\tilde{K} = K_c + mK_u + pK_w$, $i \in 0, 1, ...$ and $k \in 0, ..., l$ denote the iterative index and time sequence index, respectively, and θ_k, ξ_k are defined as

$$\theta_k(x) = \begin{bmatrix} \phi_c(x(t_{k+1})) - \phi_c(x(t_k)) \\ \int_{t_k}^{t_{k+1}} 2Re_u \otimes \phi_u \mathrm{d}\tau \\ -\int_{t_k}^{t_{k+1}} 2\gamma^2 e_w \otimes \phi_w \mathrm{d}\tau \end{bmatrix} \in \Re^{\tilde{K}}$$

$$\xi_k(\tilde{W}_i) = \int_{t_k}^{t_{k+1}} r(x, \hat{W}_{u,i}^{\mathrm{T}} \phi_u, \hat{W}_{w,i}^{\mathrm{T}} \phi_w) \mathrm{d}\tau \in \Re$$

It is desired to determine the estimated weights $\tilde{W}_{i+1}$ by minimizing the sum of squared residual error $\min_{\tilde{W}_{i+1}} \sum_{k=0}^{l-1} e_k^2$, which can be seen as a least-squares (LS) problem. First, the PE assumption to guarantee the convergence of $\tilde{W}_{i+1}$ is given.

Assumption 1: There exist $l_0 > 0$ and $\delta > 0$ for each $i \geq 0$ such that for all $l \geq l_0$, we have

$$\frac{1}{l} \sum_{k=0}^{l-1} \theta_k(x) \theta_k^{\mathrm{T}}(x) \geq \delta I_{\tilde{K}}$$

where $I_{\tilde{K}}$ is the identity matrix with appropriate dimensions.

According to the PE condition and the equation (13), the solution of this LS problem yields

$$\tilde{W}_{i+1} = -(\Theta_i^{\mathrm{T}}(x)\Theta_i(x))^{-1}\Theta_i^{\mathrm{T}}(x)\Xi_i(\tilde{W}_i) \tag{14}$$

where

$$\Theta_i(x) = [\theta_0(x), ..., \theta_{l-1}(x)]^{\mathrm{T}}$$
$$\Xi_i(\tilde{W}_i) = [\xi_0(\tilde{W}_i), ..., \xi_{l-1}(\tilde{W}_i)]^{\mathrm{T}}$$

Remark 3: To update the estimated weights by (14), we have to guarantee $\Theta_i^{\mathrm{T}}(x)\Theta_i(x))^{-1}$ exists. Therefore, the collected data in $\Theta_i(x)$ should satisfy that $rank(\Theta_i) = \tilde{K}$, which means that Θ_i^{T} has full column rank. Since the number of collected data is determined by the length of the time sequence $\{t_k\}_{k=0}^{l}$, the length index of time sequence for each interval l should be larger than $\tilde{K}$ to guarantee the condition of full column rank.

Now the model-free RL with simultaneous explorations for the nonlinear zero-sum game are proposed. Similar to most PI algorithms, the initial stabilizing policies are required. The flowchart of the model-free RL is given in Fig. 1.

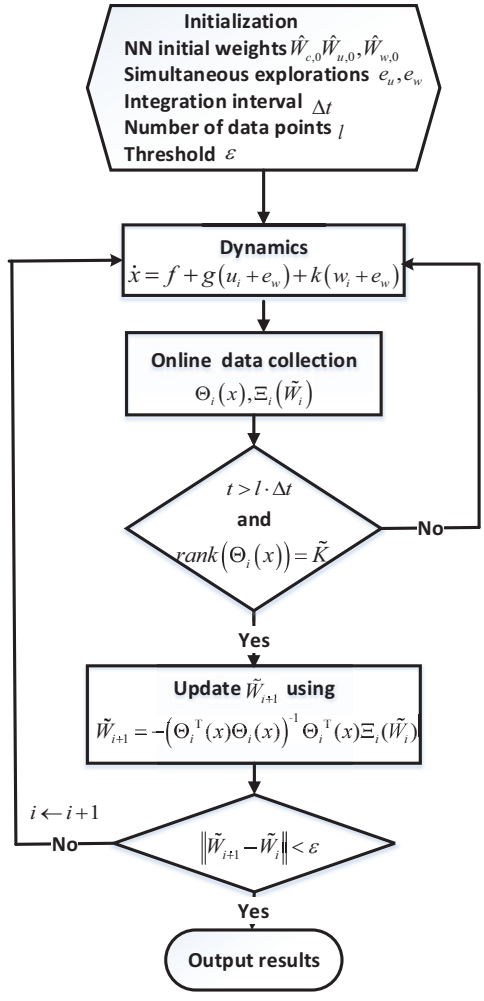

Fig. 1. The flowchart of the model-free RL with simultaneous explorations

C. Uniformly convergence analysis

In this subsection, we will prove $\left(\hat{V}_i, \hat{u}_{i+1}, \hat{w}_{i+1}\right)$ under the NN approximations (11) can uniformly approach the solution (V_i, u_{i+1}, w_{i+1}) in (10) .

Theorem 1: Suppose that Assumption 1 holds, there exist $K_c^*, K_u^*, K_w^* > 0$ such that if $K_c \geq K_c^*, K_u \geq K_u^*, K_w \geq K_w^*$, then

$$|\hat{V}_i(x) - V_i(x)| \leq \varepsilon, |\hat{u}_{i+1}(x) - u_{i+1}| \leq \varepsilon, |\hat{w}_{i+1} - w_{i+1}| \leq \varepsilon$$

Proof: Assume that there exists a solution $\bar{V}_i$ of the following equation with $\bar{V}_i(0) = 0$

$$\nabla\bar{V}_i^{\mathrm{T}}(f + g\hat{u}_i + k\hat{w}_i) + h^{\mathrm{T}}h + \hat{u}_i^{\mathrm{T}}R\hat{u}_i - \gamma^2\hat{w}_i^{\mathrm{T}}\hat{w}_i = 0 \tag{15}$$

Define

$$\bar{u}_{i+1} = -\tfrac{1}{2}R^{-1}g^{\mathrm{T}}\nabla\bar{V}_i$$
$$\bar{w}_{i+1} = \tfrac{1}{2}\gamma^{-2}k^{\mathrm{T}}\nabla\bar{V}_i$$

According to (9) along the time sequence $\{t_k\}_{k=0}^{l}$, we have

$$\begin{aligned}\bar{V}_i(x(t_k)) - \bar{V}_i(x(t_{k+1})) - \int_{t_k}^{t_{k+1}} 2\bar{u}_{i+1}^{\mathrm{T}}Re_u\mathrm{d}\tau \\ + \int_{t_k}^{t_{k+1}} 2\gamma^2\bar{w}_{i+1}^{\mathrm{T}}e_w\mathrm{d}\tau = \int_{t_k}^{t_{k+1}} r(x, \hat{u}_i, \hat{w}_i)\mathrm{d}\tau\end{aligned} \tag{16}$$

Similar with (10), the solution $\left(\bar{V}_i(x), \bar{u}_{i+1}, \bar{w}_{i+1}\right)$ can be presented as

$$\begin{aligned}\bar{V}_i(x) &= \bar{W}_{c,i+1}^{\mathrm{T}}\phi_c(x) + \bar{\varepsilon}_{c,i+1} \\ \bar{u}_{i+1}(x) &= \bar{W}_{u,i+1}^{\mathrm{T}}\phi_u(x) + \bar{\varepsilon}_{u,i+1} \\ \bar{w}_{i+1}(x) &= \bar{W}_{w,i+1}^{\mathrm{T}}\phi_w(x) + \bar{\varepsilon}_{w,i+1}\end{aligned} \tag{17}$$

Substitute (17) in (16) and based on (12), we have

$$e_k = \theta_k^{\mathrm{T}}(x)\Gamma_{i+1} + \pi_k$$

where

$$\Gamma_{i+1} = \begin{bmatrix} \hat{W}_{c,i+1} - \bar{W}_{c,i+1} \\ \mathrm{vec}(\hat{W}_{u,i+1} - \bar{W}_{u,i+1}) \\ \mathrm{vec}(\hat{W}_{w,i+1} - \bar{W}_{w,i+1}) \end{bmatrix}$$

$$\begin{aligned}\pi_k = \bar{\varepsilon}_{c,i+1}(x(t_k)) - \bar{\varepsilon}_{c,i+1}(x(t_{k+1})) \\ - \int_{t_k}^{t_{k+1}} 2e_u^{\mathrm{T}}R\bar{\varepsilon}_{u,i+1}\mathrm{d}\tau + \int_{t_k}^{t_{k+1}} 2\gamma^2 e_w^{\mathrm{T}}\bar{\varepsilon}_{w,i+1}\mathrm{d}\tau\end{aligned}$$

Based on the LS principle, we have $\sum_{k=0}^{l-1} e_k^2 \leq \sum_{k=0}^{l-1} \pi_k^2$. Then

$$\begin{aligned}\sum_{k=0}^{l-1} \Gamma_{i+1}^{\mathrm{T}}\theta_k(x)\theta_k^{\mathrm{T}}(x)\Gamma_{i+1} = \sum_{k=0}^{l-1}(e_k - \pi_k)^2 \\ \leq \sum_{k=0}^{l-1} 2(e_k^2 + \pi_k^2) \leq 4l \max_{0\leq k\leq l-1} \pi_k^2\end{aligned}$$

From Assumption 1, we yield $\|\Gamma_{i+1}\|^2 \leq \frac{4}{\delta} \max_{0\leq k\leq l-1} \pi_k^2$. Note that as $K_c, K_u, K_w \to \infty$, $\bar{\varepsilon}_{c,i+1}, \bar{\varepsilon}_{u,i+1}, \bar{\varepsilon}_{w,i+1} \to 0$ and $\pi_k \to 0$. Then we have $\lim_{K_c, K_u, K_w \to \infty} \|\Gamma_{i+1}\| \to 0$ on Ω. Considering

$$\hat{V}_i - \bar{V}_i = (\hat{W}_{c,i+1} - \bar{W}_{c,i+1})\phi_c(x) - \bar{\varepsilon}_{c,i+1}$$

We have

$$\lim_{K_c,K_u,K_w\to\infty} \hat{V}_i(x) = \bar{V}_i(x) \tag{18}$$

Similarly,

$$\begin{aligned} \lim_{K_c,K_u,K_w\to\infty} \hat{u}_{i+1}(x) &= \bar{u}_{i+1}(x) \\ \lim_{K_c,K_u,K_w\to\infty} \hat{w}_{i+1}(x) &= \bar{w}_{i+1}(x) \end{aligned} \tag{19}$$

Then, we will prove that as $K_c, K_u, K_w \to \infty$, $\bar{V}_i \to V_i, \bar{u}_{i+1} \to u_{i+1}$ and $\bar{w}_{i+1} \to w_{i+1}$ for all $i = 0, 1, ...$

(a) For $i = 0$, we have $\bar{V}_0 = V_0, \bar{u}_1 = u_1, \bar{w}_1 = w_1$ according to the definitions of $\bar{V}_i, \bar{u}_{i+1}$ and $\bar{w}_{i+1}$.

(b) For some $i > 0$, assume that as $K_c, K_u, K_w \to \infty$, $\bar{V}_{i-1} \to V_{i-1}, \bar{u}_i \to u_i$ and $\bar{w}_i \to w_i$. From (8) and (15), we can get

$$\begin{aligned} \bar{V}_i(x) - V_i(x) = &\int_t^\infty -2(\bar{u}_{i+1} - u_{i+1})^{\mathrm{T}} R e_u \mathrm{d}\tau \\ &+ \int_t^\infty 2\gamma^2 (\bar{w}_{i+1} - w_{i+1})^{\mathrm{T}} e_w \mathrm{d}\tau \end{aligned}$$

With the NN approximation structure, we yield

$$\begin{aligned} &\phi_c^{\mathrm{T}}(\bar{W}_{c,i+1} - W_{c,i+1}) + \int_t^\infty 2\phi_u^{\mathrm{T}}(\bar{W}_{u,i+1} - W_{u,i+1}) \\ &\times R e_u \mathrm{d}\tau - \int_t^\infty 2\gamma^2 \phi_w^{\mathrm{T}}(\bar{W}_{u,i+1} - W_{u,i+1}) e_w \mathrm{d}\tau = \varepsilon_\Gamma \end{aligned}$$

According to the approximation property of NNs, the equation (19) and the hypotheses (b), we have $\lim_{K_c,K_u,K_w\to\infty} \varepsilon_\Gamma \to 0$ as $K_c, K_u, K_w \to \infty$. It implies that

$$\begin{aligned} \lim_{K_c,K_u,K_w\to\infty} \hat{V}_i(x) &= V_i(x) \\ \lim_{K_c,K_u,K_w\to\infty} \hat{u}_{i+1}(x) &= u_{i+1}(x) \\ \lim_{K_c,K_u,K_w\to\infty} \hat{w}_{i+1} &= w_{i+1}(x) \end{aligned} \tag{20}$$

Based on (18), (19) and (20), there exist $K_c^*, K_u^*, K_w^* > 0$ such that if $K_c \geq K_c^*, K_u \geq K_u^*, K_w \geq K_w^*$, then

$$\begin{aligned} |\hat{V}_i(x) - V_i(x)| &\leq |\hat{V}_i(x) - \bar{V}_i(x)| + |\bar{V}_i(x) - V_i(x)| \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

Similarly, we have $|\hat{u}_{i+1}(x) - u_{i+1}| \leq \varepsilon, |\hat{w}_{i+1} - w_{i+1}| \leq \varepsilon$. This completes the proof.

■

According to Theorem 1, the NN approximations (11) can approach the solution of (9) at each iteration. As the iteration index $i \to \infty$, the model-free RL can approach the solution of the HJI equation.

IV. Simulation

Consider the continuous-time nonlinear zero-sum game in [14] with modification as follows

$$\dot{x} = f(x) + g(x)u + k(x)w$$

where

$$f(x) = \begin{bmatrix} -x_1 + x_2 \\ -x_1 - x_2 + \frac{1}{4}(x_1 + x_2)^2 x_2 - \frac{1}{4\gamma^2} x_2^3 \end{bmatrix}$$

$$g(x) = \begin{bmatrix} 0 \\ x_1 + x_2 \end{bmatrix}, k(x) = \begin{bmatrix} 0 \\ x_2 \end{bmatrix}$$

Note that the system dynamics are assumed to be unknown. Select $h(x) = [x_1^2, x_2^2]$, $R = 1$ and $\gamma = 2$. According to the converse optimal control method in [19], we can obtain the optimal value function and the optimal policies

$$V^*(x) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2,$$

$$u^*(x) = -\frac{1}{2}x_1 x_2 - \frac{1}{2}x_2^2, \quad w^*(x) = \frac{1}{8}x_2^2$$

Then we can construct three NNs to approach the value, control and disturbance policies with the same basis function vectors $\phi_c(x) = \phi_u(x) = \phi_w(x) = [x_1^2, x_1 x_2, x_2^2]^{\mathrm{T}}$. Thus, the ideal coefficients for the NNs should be $W_c^* = [0.5, 0, 0.5]^{\mathrm{T}}$, $W_u^* = [0, -0.5, -0.5]^{\mathrm{T}}$ and $W_w^* = [0, 0, 0.125]^{\mathrm{T}}$. Let the initial system state be $x_0 = [1, -1]^{\mathrm{T}}$, the sampling interval be 0.1s. We choose the time sequence index $l = 20$, which means the estimated weights updating law (14) is tuned every 2s. As the system is self-stable, the initial weights for the critic, actor and disturbance NNs are chosen as $W_{c,0} = W_{u,0} = W_{w,0} = [0, 0, 0]^{\mathrm{T}}$. We experimentally choose two sinusoidal signals with different frequencies as the simultaneous explorations, i.e., $e_u = 0.3(\sin(\pi t) + \sin(5t) + \sin(9.3t) + 0.5)$ and $e_w = 0.3(\sin(2\pi t) + \sin(4t) + \sin(11t))$. The convergence threshold is 10^{-6}. After 6th iterations, the online algorithm is converged and the estimated weights are $\hat{W}_{c,6} = [0.5, 0, 0.5]^{\mathrm{T}}$, $\hat{W}_{u,6} = [0, -0.5, -0.5]^{\mathrm{T}}$ and $\hat{W}_{w,6} = [0, 0, 0.125]^{\mathrm{T}}$. The convergence curves of the estimated weights are shown in Fig. 2-Fig. 4. Compared with the online learning algorithms in [12], [13] that the convergence time is about 100s, the convergence time of 12s in our algorithm is shorter.

Then we apply the obtained optimal controller to the corresponding H_∞ problem with a disturbance signal as

$$w = \begin{cases} 8e^{-(t-t_0)}\cos(t - t_0), & t > t_0 \\ 0, & t < t_0 \end{cases}$$

where $t_0 = 4.5s$. The whole state and action trajectories are shown in Fig. 5. We can see that the closed-loop system under the obtained optimal controller is asymptotically stable, which demonstrates the effectiveness of the proposed algorithm.

V. Conclusion

The continuous-time nonlinear zero-sum game with unknown system dynamics is investigated in this paper. Combining IRL with simultaneous explorations, a model-free RL algorithm is developed to solve the HJI equation approximately. Furthermore, the critic-actor-disturbance NNs are constructed to approach the optimal solution using online learning scheme.

References

[1] A. Kelly, *Decision Making Using Game Theory: An Introduction for Managers*. Cambridge University Press, 2003.

[2] R. Selten, "A note on evolutionarily stable strategies in asymmetric animal conflicts," *Journal of Theoretical Biology*, vol. 84, no. 1, pp. 93–101, 1980.

[3] T. Basar, G. J. Olsder, G. Clsder, T. Basar, T. Baser, and G. J. Olsder, *Dynamic Noncooperative Game Theory*. SIAM, 1995.

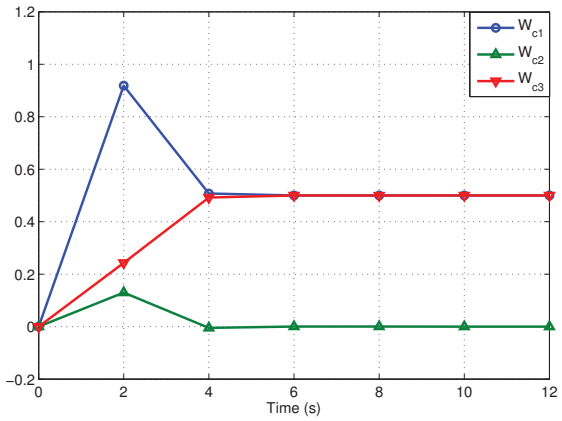

Fig. 2. The convergence curve of $\hat{W}_{c,i+1}$

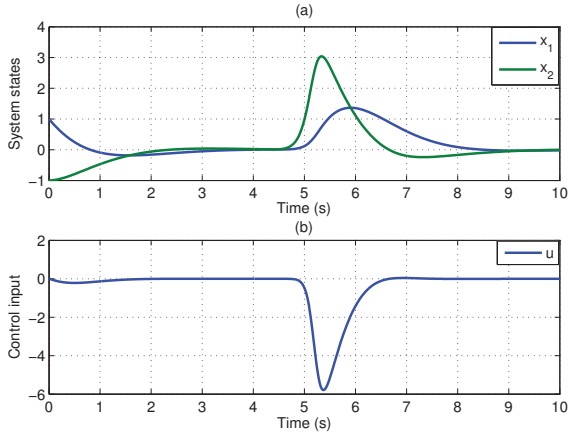

Fig. 5. Close-loop system states (a) and control input (b)

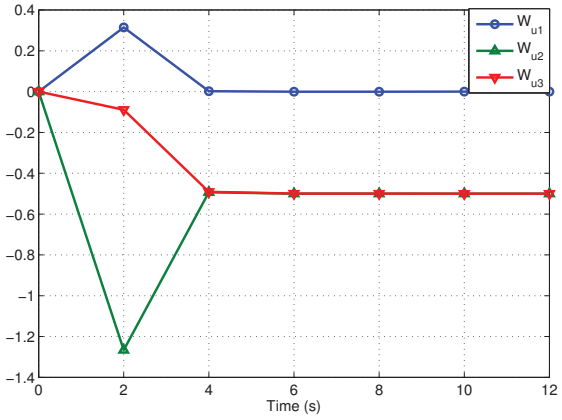

Fig. 3. The convergence curve of $\hat{W}_{u,i+1}$

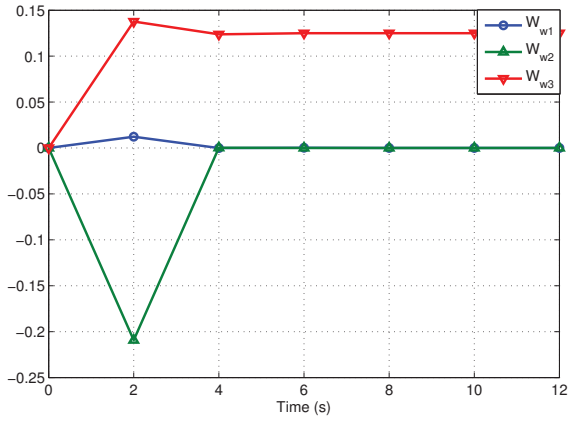

Fig. 4. The convergence curve of $\hat{W}_{w,i+1}$

[4] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control," *Automatica*, vol. 43, no. 3, pp. 473–481, 2007.

[5] J. Huang and C. Lin, "Numerical approach to computing nonlinear H-infinity control laws," *Journal of Guidance, Control, and Dynamics*, vol. 18, no. 5, pp. 989–994, 1995.

[6] D. Liu, H. Li, and D. Wang, "Neural-network-based zero-sum game for discrete-time nonlinear systems via iterative adaptive dynamic programming algorithm," *Neurocomputing*, vol. 110, pp. 92–100, 2013.

[7] D. Zhao, Q. Zhang, X. Li, and L. Kong, "Event-triggered H_∞ control for continuous-time nonlinear system," in *Advances in Neural Networks–ISNN 2015*. Springer, 2015, pp. 62–70.

[8] M. Abu-Khalaf, F. L. Lewis, and J. Huang, "Neurodynamic programming and zero-sum games for constrained control systems," *IEEE Transactions on Neural Networks*, vol. 19, no. 7, pp. 1243–1252, 2008.

[9] H. Wu and B. Luo, "Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear control," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 12, pp. 1884–1895, 2012.

[10] D. Vrabie and F. Lewis, "Adaptive dynamic programming for online solution of a zero-sum differential game," *Journal of Control Theory and Applications*, vol. 9, no. 3, pp. 353–360, 2011.

[11] B. Luo, H.-N. Wu, and T. Huang, "Off-policy reinforcement learning for control design," *IEEE Transactions on Cybernetics*, vol. 45, no. 1, pp. 65–76, 2015.

[12] Q. Wei, R. Song, and P. Yan, "Data-driven zero-sum neuro-optimal control for a class of continuous-time unknown nonlinear systems with disturbance using ADP," *IEEE Transactions on Neural Networks and Learning Systems*, DOI 10.1109/TNNLS.2015.2464080, 2015.

[13] S. Yasini, M. B. N. Sistani, and A. Karimpour, "Approximate dynamic programming for two-player zero-sum game related to H_∞ control of unknown nonlinear continuous-time systems," *International Journal of Control, Automation and Systems*, vol. 13, no. 1, pp. 99–109, 2015.

[14] Y. Zhu and D. Zhao, "Model-free iterative adaptive dynamic programming for solving unknown nonlinear zero-sum game based on online measurement," *Submitted to IEEE Transactions on Neural Networks and Learning Systems*, 2015.

[15] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 916–932, 2015.

[16] H. Li, D. Liu, and D. Wang, "Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 3, pp. 706–714, 2014.

[17] B. Finlayson, "The Method of Weighted Residuals and Variational Principles Academic," *New York*, 1972.

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

[19] V. Nevistić and J. A. Primbs, "Constrained nonlinear optimal control: a converse HJB approach," California Institute of Technology. Tech. Rep. 96-021, 1996.