Solving Nonlinear Zero-Sum Game with Completely Unknown Dynamics via Iterative Adaptive Dynamic Programming

Yuanheng Zhu, *Student Member, IEEE*
Supervisor: Dongbin Zhao and Haibo He

***Abstract*—H_∞ control is a powerful method to attenuate disturbance that presents in the control system. The design of such controllers relies on solving the zero-sum game. In the practical applications, however, the exact dynamics is mostly unknown and the identification of the system causes approximation error which is detrimental to the control performance. To overcome this problem, an iterative adaptive dynamic programming algorithm is presented in this paper to solve the continuous-time unknown nonlinear zero-sum game with only online measurement. To this end, a model-free approach to the Hamilton-Jacobi-Isaacs equation is proposed based on policy iteration, and the value, control and disturbance policies are approximated by neural networks under the critic-actor-disturbance structure. Our algorithm is proved to be a Gauss-Newton method solving an optimization problem and uniformly converge to the optimal solution. Simulation results verify the superior in eliminating dynamics knowledge and saving learning time than other ones.**

I. Introduction

OPTIMAL control is a topic of intensive study in control theory. It aims to find a policy that minimizes certain performance index. In various control applications, there are numerous situations where disturbance exists in the system and plays a negative role on the control effect. In these cases, H_∞ control [1], [2] provides a powerful method that attenuates the disturbance effect. According to the game theory [3], the H_∞ control is equivalent to the solvability of a two-player zero-sum game (ZSG) where the controller is to minimize the performance index in the worst-case disturbance. When consider a system with the continuous-time (CT) nonlinear dynamics, the ZSG can be solved by the Hamilton-Jacobi-Isaacs (HJI) equation. However due to the inherent nonlinearity, it is intractable to give an analytic solution to the HJI equation.

Recently a new technique called adaptive/approximate dynamic programming (ADP) [4], [5] has been widely studied in control area, also including ZSG problem. For example, Abu-Khalaf et al. put forward an offline inner-outer-loop policy iteration (PI) to solve the control-saturated HJI equation in [6]. Zhang et al. [7] study the specific situation for which the saddle point may not exist. In [8], Dierks and Jagannathan use a single online approximator to address the ZSG online. Vamvoudakis and Lewis [9] also give an online ADP algorithm which extends their synchronous policy iteration algorithm (SPIA) [10] to the ZSG based on the critic-actor-disturbance neural-networks (NNs) structure. Unfortunately, these two works require the complete system dynamics. Motivated by that, Wu and Luo [11] resort to the idea of integral reinforcement learning (IRL), which relieves the dependence of the internal dynamics.

ADP is developed from reinforcement learning (RL) [12] which more concerns of discrete-time (DT) systems. In practical applications, the exact mathematical dynamics is frequently unknown. Some researchers tune NNs to identify the unknown dynamics; then apply ADP on the modeled systems to solve the optimal control problems [13], [14]. Unfortunately, the approximation errors in the identifier NNs are detrimental to the optimality of the results. The training of the identifier NNs also increases the computational cost and learning time. Hence a totally model-free approach is more direct and efficient to the unknown systems. In [15], Jiang and Jiang perform a robust ADP to the nonlinear optimal control problem without any system dynamics. As for linear quadratic ZSG where the dynamics is linear and the performance index is quadratic, the HJI equation reduces to the generalized algebraic Riccati equation (GARE) and Vrabie and Lewis [16] use a model-free method to obtain its solution. Another research that studies the same problem is presented in [17]. But when considering nonlinear systems, the unknown nonlinear ZSG is rarely studied in the literature, except [18] where the authors study the optimal tracking problem with H_∞ control method.

In this paper, we consider the optimal control of a continuous-time unknown nonlinear zero-sum game. An iterative ADP algorithm is designed to approximately solve the problem with online measurement. The original model-based policy iteration to the HJI equation is converted to a model-free iteration with the additional control inputs. To approach the value, control and disturbance policies, neural network approximators and the critic-actor-disturbance structure are used. The HJI equation is approximately solved by iterating the NNs parameters. It is further proved that the iteration is equivalent to a Gauss-Newton method and uniformly converges to the optimal value and saddle point. Simulated experiment demonstrates the performance of our algorithm.

Y. Zhu and D. Zhao are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yuanheng.zhu@gmail.com, dongbin.zhao@ia.ac.cn).

H. He is with the Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island Kingston, RI 02881, USA (e-mail: he@ele.uri.edu

This research is funded by the IEEE Computational Intelligence Society Graduate Student Research Grant 2015.

II. Preliminary

Consider the following continuous-time nonlinear system

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

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ is the control signal, $\omega(t) \in \mathbb{R}^q$ is the external disturbance satisfying $\omega(t) \in L_2[0,\infty)$, $z \in \mathbb{R}^p$ is the fictitious output, and $f(x)$, $g(x)$, $k(x)$ are the system dynamics vector and matrices with appropriate dimensions. Assume $f(x)$, $g(x)$, and $k(x)$ are Lipschitz continuous and $f(0) = 0$.

H_∞ control is to find a controller that renders the performance index

$$J(x(0), u, \omega) = \int_0^\infty \left(h^T(x)h(x) + u^T R u - \gamma^2 \omega^T \omega\right) d\tau$$

nonpositive for all $\omega \in L_2[0,\infty)$ with $x(0) = 0$, where $R > 0$, $\gamma \geq \gamma^* \geq 0$. If such controller exists, it is said that the system has L_2-gain $\leq \gamma$. γ^* represents the smallest value for which the problem is still solvable.

Given a control policy $u(t) \equiv u(x(t))$ and a disturbance policy $\omega(t) \equiv \omega(x(t))$, their value function is defined as

$$\begin{aligned} V(x(0)) &= \int_0^\infty \left(h^T h + u^T R u - \gamma^2 \omega^T \omega\right) d\tau \\ &\equiv \int_0^\infty r(x(t), u, \omega) d\tau \end{aligned}$$

A differential equivalent to the above is a Lyapunov equation

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

where ∇ denotes the partial derivative operator, i.e. $\nabla V = \partial V / \partial x$. Define the Hamiltonian function

$$H(x, \nabla V, u, \omega) \equiv r(x, u, \omega) + \nabla V^T (f + gu + k\omega)$$

Based on game theory, H_∞ problem can be solved by a two-player zero-sum game, where the control signal aims to maximize the performance index while the disturbance signal acts as an opponent and tries to minimize it. The continuous-time nonlinear ZSG is to find the feedback control and disturbance policies which gain the optimal value

$$V^*(x) \equiv \min_u \max_\omega J(x, u, \omega)$$

If the following Nash condition is satisfied

$$\min_u \max_\omega J(x, u, \omega) = \max_\omega \min_u J(x, u, \omega)$$

ZSG has a unique solution, i.e. the saddle point (u^*, ω^*), and u^* is an H_∞ controller for (1). According to the stationary conditions, the formulations of u^* and ω^* are derived as

$$\begin{aligned} \frac{\partial H(x, \nabla V^*, u, \omega)}{\partial u} = 0 \Rightarrow u^* = -\frac{1}{2} R^{-1} g^T \nabla V^* \\ \frac{\partial H(x, \nabla V^*, u, \omega)}{\partial \omega} = 0 \Rightarrow \omega^* = \frac{1}{2} \gamma^{-2} k^T \nabla V^* \end{aligned}$$

After substituting u^* and ω^* into the Lyapunov equation (2), we get the Hamilton-Jacobi-Isaacs equation

$$\begin{aligned} \nabla V^{*T} f + h^T h - \frac{1}{4} \nabla V^{*T} g R^{-1} g^T \nabla V^* \\ + \frac{1}{4} \gamma^{-2} \nabla V^{*T} k k^T \nabla V^* = 0, V^*(0) = 0 \end{aligned} \tag{3}$$

Assumption 1: [9] Select $\gamma > 0$. Assume system (1) is zero-state observable, and there exists a control policy $u(x)$ for which the system has L_2-gain $\leq \gamma$ on a set $\Omega \in \mathbb{R}^n$ and is asymptotically stable. Assume the smooth solution of the HJI equation (3) also exists on Ω.

From the above statement, it is crucial to solve the HJI equation in the ZSG. A common approach is based on policy iteration. Given an appropriate initial value function V_0, a sequence of the values $\{V_i\}_{i=0}^\infty$ are produced by solving the following Lyapunov equation with $V_i(0) = 0$

$$\nabla V_i^T (f + g u_i + k \omega_i) + h^T h + u_i^T R u_i - \gamma^2 \omega_i^T \omega_i = 0 \tag{4}$$

and using the policies updating law

$$u_{i+1} = -\frac{1}{2} R^{-1} g^T \nabla V_i, \quad \omega_{i+1} = \frac{1}{2} \gamma^{-2} k^T \nabla V_i \tag{5}$$

It is easy to prove that following the iteration, the sequence $\{V_i\}_{i=0}^\infty$ converge to the optimal value V^* as $i \to \infty$. But when reviewing (4) and (5), it is observed that the iteration requires the complete knowledge of the dynamics, limiting its application when such information is unknown.

III. A Neural-Network Based Approach to HJI Equation Without System Dynamics

Suppose two arbitrary policies u and ω are implemented and we assume they stabilize the system (1) in a compact region. Denote u_i and ω_i as the results of the i-th iteration in (4) and (5), which are further used to compute V_i, u_{i+1}, ω_{i+1} at the $(i+1)$-th iteration. Along the solutions of the system, the time derivative of V_i equals $\dot{V}_i = \nabla V_i^T (f + gu + k\omega)$. Subtracting 0 from (4) and utilizing (5), we have

$$\dot{V}_i = -2u_{i+1}^T R(u - u_i) + 2\gamma^2 \omega_{i+1}^T (\omega - \omega_i) - r(x, u_i, \omega_i)$$

Based on the idea of integral reinforcement learning, the following equation is formulated

$$\begin{aligned} 0 = V_i(x(t')) - V_i(x(t)) + \int_t^{t'} 2u_{i+1}^T R(u - u_i) d\tau \\ - \int_t^{t'} 2\gamma^2 \omega_{i+1}^T (\omega - \omega_i) d\tau + \int_t^{t'} r(x, u_i, \omega_i) d\tau. \end{aligned} \tag{6}$$

Next, neural network approximation is introduced to solve (6) approximately. According to the Weirstrass high-order approximation theorem, a smooth function can be uniformly approximated on a compact set by neural networks. On the compact set Ω, we define[1]

$$\begin{aligned} V_i(x) &= c_{1,i+1}^T \phi_1(x) + \varepsilon_{1,i+1}(x) \\ u_{i+1}(x) &= c_{2,i+1}^T \phi_2(x) + \varepsilon_{2,i+1}(x) \\ \omega_{i+1}(x) &= c_{3,i+1}^T \phi_3(x) + \varepsilon_{3,i+1}(x) \end{aligned}$$

and

$$\begin{aligned} u_i(x) &= c_{2,i}^T \phi_2(x) + \varepsilon_{2,i}(x) \\ \omega_i(x) &= c_{3,i}^T \phi_3(x) + \varepsilon_{3,i}(x) \end{aligned}$$

[1]To denote uniformly, the NN coefficients of V_i use subscript $(i+1)$ and the other value functions below follow the same rule.

where $\phi_1 : \mathbb{R}^n \to \mathbb{R}^{K_1}$, $\phi_2 : \mathbb{R}^n \to \mathbb{R}^{K_2}$, $\phi_3 : \mathbb{R}^n \to \mathbb{R}^{K_3}$ are linearly independent basis function vectors, $c_{1,\bullet} \in \mathbb{R}^{K_1}$, $c_{2,\bullet} \in \mathbb{R}^{K_2 \times m}$, $c_{3,\bullet} \in \mathbb{R}^{K_3 \times q}$ are the coefficient vector and matrices, $\varepsilon_{1,\bullet}$, $\varepsilon_{2,\bullet}$, $\varepsilon_{3,\bullet}$ are the reconstruction errors with appropriate dimensions. K_1, K_2, K_3 are the numbers of neurons in the hidden layers. Assume basis functions, coefficients and reconstruction errors are all bounded over Ω and when $K_1 \to \infty, K_2 \to \infty, K_3 \to \infty$, we have $\varepsilon_{1,\bullet} \to 0, \varepsilon_{2,\bullet} \to 0, \varepsilon_{3,\bullet} \to 0$.

After substituting the above NNs into (6), we yield

$$\begin{aligned}\varepsilon_L =&\big(\phi_1(x(t')) - \phi_1(x(t))\big)^T c_{1,i+1} \\ &+ \int_t^{t'} 2\phi_2^T c_{2,i+1} R(u - c_{2,i}^T \phi_2) d\tau \\ &- \int_t^{t'} 2\gamma^2 \phi_3^T c_{3,i+1} (\omega - c_{3,i}^T \phi_3) d\tau \\ &+ \int_t^{t'} r(x, c_{2,i}^T \phi_2, c_{3,i}^T \phi_3) d\tau\end{aligned}$$

where ε_L is the Lyapunov equation error due to the NNs reconstruction errors, defined by

$$\begin{aligned}\varepsilon_L \equiv& -\varepsilon_{1,i+1}(x(t')) + \varepsilon_{1,i+1}(x(t)) \\ &- \int_t^{t'} 2\left((u - c_{2,i}^T \phi_2)^T R \varepsilon_{2,i+1} - \phi_2^T c_{2,i+1} R \varepsilon_{2,i}\right. \\ &\qquad \left. - \varepsilon_{2,i+1}^T R \varepsilon_{2,i}\right) d\tau \\ &+ \int_t^{t'} 2\gamma^2 \left((\omega - c_{3,i}^T \phi_3)^T \varepsilon_{3,i+1} - \phi_3^T c_{3,i+1} \varepsilon_{3,i}\right. \\ &\qquad \left. - \varepsilon_{3,i+1}^T \varepsilon_{3,i}\right) d\tau \\ &- \int_t^{t'} (2\phi_2^T c_{2,i} R \varepsilon_{2,i} + \varepsilon_{2,i}^T R \varepsilon_{2,i}) d\tau \\ &+ \int_t^{t'} \gamma^2 (2\phi_3^T c_{3,i} \varepsilon_{3,i} + \varepsilon_{3,i}^T \varepsilon_{3,i}) d\tau\end{aligned}$$

Now three NNs structure is utilized, i.e. the critic, actor, and disturbance NN approximators for the value, control and disturbance policies respectively. As the ideal coefficients are unknown, a group of estimations, $W_{1,i+1}, W_{2,i+1}, W_{3,i+1}$, replace $c_{1,i+1}, c_{2,i+1}, c_{3,i+1}$ and parameterize the NN approximators as

$$\begin{aligned}\hat{V}_i(x) &= W_{1,i+1}^T \phi_1(x) \\ \hat{u}_{i+1}(x) &= W_{2,i+1}^T \phi_2(x) \\ \hat{\omega}_{i+1}(x) &= W_{3,i+1}^T \phi_3(x)\end{aligned}$$

Assuming the estimated weights $W_{2,i}, W_{3,i}$, w.r.t. $c_{2,i}, c_{3,i}$, are already known, then $W_{1,i+1}, W_{2,i+1}, W_{3,i+1}$ can be determined in the least-squares (LS) principle. Given a strictly increasing time sequence $\{t_k\}_{k=0}^l$, for each interval define the residual error e_k as

$$\begin{aligned}e_k =&\big(\phi_1(x(t_{k+1})) - \phi_1(x(t_k))\big)^T W_{1,i+1} \\ &+ \int_{t_k}^{t_{k+1}} 2\phi_2^T W_{2,i+1} R(u - W_{2,i}^T \phi_2) d\tau \\ &- \int_{t_k}^{t_{k+1}} 2\gamma^2 \phi_3^T W_{3,i+1} (\omega - W_{3,i}^T \phi_3) d\tau \\ &+ \int_{t_k}^{t_{k+1}} r(x, W_{2,i}^T \phi_2, W_{3,i}^T \phi_3) d\tau \end{aligned} \tag{7}$$

By Kronecker product $\otimes$, we have

$$\begin{aligned}&\phi_2^T W_{2,i+1} R(u - W_{2,i}^T \phi_2) = \\ &\qquad \big((u - W_{2,i}^T \phi_2)^T R \otimes \phi_2^T\big) \mathbf{v}(W_{2,i+1}) \\ &\phi_3^T W_{3,i+1} (\omega - W_{3,i}^T \phi_3) = \big((\omega - W_{3,i}^T \phi_3)^T \otimes \phi_3^T\big) \mathbf{v}(W_{3,i+1})\end{aligned}$$

where $\mathbf{v}(\cdot)$ is a vector function which transforms a matrix into a vector by stacking columns. After that (7) can be rewritten into a linear form

$$e_k = \theta_k^T(\bar{W}_i) \bar{W}_{i+1} + \xi(\bar{W}_i)$$

where $W_{1,i+1}, W_{2,i+1}, W_{3,i+1}$ are integrated into the vector $\bar{W}_{i+1} = [W_{1,i+1}^T, \mathbf{v}(W_{2,i+1})^T, \mathbf{v}(W_{3,i+1})^T]^T \in \mathbb{R}^{\bar{K}}$ and $\bar{K} = K_1 + mK_2 + qK_3$. $\bar{W}_i$ is defined in the same way by $W_{1,i}, W_{2,i}, W_{3,i}$, and θ_k, ξ_k are defined as

$$\begin{aligned}\theta_k(\bar{W}_i) &= \begin{bmatrix} \phi_1(x(t_{k+1})) - \phi_1(x(t_k)) \\ \int_{t_k}^{t_{k+1}} 2R(u - W_{2,i}^T \phi_2) \otimes \phi_2 d\tau \\ -\int_{t_k}^{t_{k+1}} 2\gamma^2 (\omega - W_{3,i}^T \phi_3) \otimes \phi_3 d\tau \end{bmatrix} \in \mathbb{R}^{\bar{K}} \\ \xi_k(\bar{W}_i) &= \int_{t_k}^{t_{k+1}} r(x, W_{2,i}^T \phi_2, W_{3,i}^T \phi_3) d\tau \in \mathbb{R}\end{aligned}$$

The estimated weights $\bar{W}_{i+1}$ are determined by solving the LS problem

$$\min_{\bar{W}_{i+1}} \sum_{k=0}^{l-1} e_k^2$$

Assumption 2 (Persistency of excitation (PE)): For each $i \geq 0$, there exist $l_0 > 0$ and $\delta > 0$ such that for all $l \geq l_0$, we have

$$\frac{1}{l} \sum_{k=0}^{l-1} \theta_k(\bar{W}_i) \theta_k^T(\bar{W}_i) \geq \delta I_{\bar{K}}$$

where $I_{\bar{K}}$ is the identity matrix with the given size.

Based on the PE condition, the solution to the LS problem is directly calculated by

$$\bar{W}_{i+1} = -\big(\Theta^T(\bar{W}_i)\Theta(\bar{W}_i)\big)^{-1} \Theta^T(\bar{W}_i) \Xi(\bar{W}_i) \tag{8}$$

where

$$\Theta(\bar{W}_i) = [\theta_0(\bar{W}_i), \ldots, \theta_{l-1}(\bar{W}_i)]^T \tag{9}$$

$$\Xi(\bar{W}_i) = [\xi_0(\bar{W}_i), \ldots, \xi_{l-1}(\bar{W}_i)]^T \tag{10}$$

Now the iterative ADP algorithm solving the unknown nonlinear ZSG is proposed. Given a set of initial policies weights, $W_{2,0}$ and $W_{3,0}$, the parameters for the critic-actor-disturbance NNs structure are iterated following (8). If a sequence of the system trajectories are given and the PE condition holds continually, no dynamics is needed in our algorithm.

IV. Convergence Theoretical Analysis

A. Convergence of the iteration

We first demonstrate the convergence of our algorithm by proving its equivalence to a Gauss-Newton iteration. As we assume there exists a solution of the HJI equation, let the optimal value and saddle point policies be represented by NNs in the following form

$$V^*(x) = c_{1,*}^T \phi_1(x) + \varepsilon_{1,*}(x) \tag{11}$$
$$u^*(x) = c_{2,*}^T \phi_2(x) + \varepsilon_{2,*}(x) \tag{12}$$
$$\omega^*(x) = c_{3,*}^T \phi_3(x) + \varepsilon_{3,*}(x) \tag{13}$$

Now consider V^*, u^*, ω^* and use the IRL method. A similar equation is derived as (6) after some manipulation

$$\begin{aligned} 0 = V^*(x(t')) - V^*(x(t)) + \int_t^{t'} u^{*T} R(2u - u^*) d\tau \\ - \int_t^{t'} \gamma^2 \omega^{*T}(2\omega - \omega^*) d\tau + \int_t^{t'} h^T h d\tau \end{aligned}$$

After substituting NNs approximation (11), (12), (13), it becomes

$$\begin{aligned} \varepsilon_{HJI} =& \big(\phi_1(x(t')) - \phi_1(x(t))\big)^T c_{1,*} \\ &+ \int_t^{t'} \phi_2^T c_{2,*} R(2u - c_{2,*}^T \phi_2) d\tau \\ &- \int_t^{t'} \gamma^2 \phi_3^T c_{3,*}(2\omega - c_{3,*}^T \phi_3) d\tau + \int_t^{t'} h^T h d\tau \end{aligned}$$

where ε_{HJI} is the HJI equation error with

$$\begin{aligned} \varepsilon_{HJI} \equiv& -\varepsilon_{1,*}(x(t')) + \varepsilon_{1,*}(x(t)) \\ &- \int_t^{t'} \big(2(u - c_{2,*}^T \phi_2)^T R \varepsilon_{2,*} - \varepsilon_{2,*}^T R \varepsilon_{2,*}\big) d\tau \\ &+ \int_t^{t'} \gamma^2 \big(2(\omega - c_{3,*}^T \phi_3)^T \varepsilon_{3,*} - \varepsilon_{3,*}^T \varepsilon_{3,*}\big) d\tau \end{aligned}$$

When the ideal values of $c_{1,*}, c_{2,*}, c_{3,*}$ are determined, the approximated optimal value and saddle point to the HJI equation are acquired. As $c_{1,*}, c_{2,*}, c_{3,*}$ are unknown, we use $W_{1,*}, W_{2,*}, W_{3,*}$ as their estimations. With $\{t_k\}_{k=0}^l$, the estimated weights formulate a set of residual errors

$$\begin{aligned} d_k =& \big(\phi_1(x(t_{k+1})) - \phi_1(x(t_k))\big)^T W_{1,*} \\ &+ \int_{t_k}^{t_{k+1}} \phi_2^T W_{2,*} R(2u - W_{2,*}^T \phi_2) d\tau \\ &- \int_{t_k}^{t_{k+1}} \gamma^2 \phi_3^T W_{3,*}(2\omega - W_{3,*}^T \phi_3) d\tau + \int_{t_k}^{t_{k+1}} h^T h d\tau \end{aligned}$$

The problem becomes a nonlinear least-squares problem (NLSP) to find the parameters that minimize the square error

$$\min_{\bar{W}_*} D^T(\bar{W}_*) D(\bar{W}_*) \tag{14}$$

where $\bar{W}_*$ denotes $\bar{W}_* = [W_{1,*}^T, \mathbf{v}(W_{2,*})^T, \mathbf{v}(W_{3,*})^T]^T \in \mathbb{R}^{\bar{K}}$ and $D(\bar{W}_*)$ is the residual error vector $D(\bar{W}_*) = [d_0, \ldots, d_{l-1}]^T \in \mathbb{R}^l$. It has been demonstrated that Gauss-Newton method is a feasible approach to this optimization problem. Next lemmas reveal the connection between Gauss-Newton method and our iterative ADP algorithm.

Lemma 1: The Jacobian matrix $\mathbf{J} \in \mathbb{R}^{l * \bar{K}}$ of NLSP w.r.t. (14) is defined as

$$(\mathbf{J}(\bar{W}_*))_{ij} = \frac{\partial (D(\bar{W}_*))_i}{\partial (\bar{W}_*)_j}$$

When substituting $\bar{W}_*$ into (9), we have $\mathbf{J}(\bar{W}_*) = \Theta(\bar{W}_*)$.

Proof: The partial derivatives of d_k to the NN weights are

$$\begin{aligned} \frac{\partial d_k}{\partial W_{1,*}} &= \phi_1(x(t_{k+1})) - \phi_1(x(t_k)) \\ \frac{\partial d_k}{\partial W_{2,*}} &= \int_{t_k}^{t_{k+1}} 2(\phi_2 u^T R - \phi_2 \phi_2^T W_{2,*} R) d\tau \\ \frac{\partial d_k}{\partial W_{3,*}} &= -\int_{t_k}^{t_{k+1}} 2\gamma^2(\phi_3 \omega^T - \phi_3 \phi_3^T W_{3,*}) d\tau \end{aligned}$$

From Kronecker product representation, it is straightforward to infer $\partial d_k / \partial \bar{W}_* = \theta_k(\bar{W}_*)$. Hence $\mathbf{J}(\bar{W}_*) = \Theta(\bar{W}_*)$. ■

Lemma 2: Given a parametric vector $\bar{W}_i \in \mathbb{R}^{\bar{K}}$, if Assumption 2 keeps satisfied, computing $\bar{W}_{i+1}$ based on (8) is equivalent to the Gauss-Newton iteration

$$\bar{W}_{i+1} = \bar{W}_i - \big(\mathbf{J}^T(\bar{W}_i)\mathbf{J}(\bar{W}_i)\big)^{-1} \mathbf{J}^T(\bar{W}_i) D(\bar{W}_i) \tag{15}$$

Proof: From Lemma 1 and based on the definitions of Θ and Ξ, we have

$$\mathbf{J}(\bar{W}_i)\bar{W}_i - D(\bar{W}_i) = -\Xi(\bar{W}_i)$$

Substitute into (8)

$$\begin{aligned} \bar{W}_{i+1} =& \big(\mathbf{J}^T(\bar{W}_i)\mathbf{J}(\bar{W}_i)\big)^{-1} \mathbf{J}^T(\bar{W}_i) \big(\mathbf{J}(\bar{W}_i)\bar{W}_i - D(\bar{W}_i)\big) \\ =& \bar{W}_i - \big(\mathbf{J}^T(\bar{W}_i)\mathbf{J}(\bar{W}_i)\big)^{-1} \mathbf{J}^T(\bar{W}_i) D(\bar{W}_i) \end{aligned}$$

■

According to the above analysis, it is revealed that our algorithm is actually a Gauss-Newton method to the optimization problem (14), and the next theorem is concluded whose proof follows the convergence results given in [19].

Theorem 1: Suppose the following conditions remain satisfied

1) Assumption 2 holds continually;
2) there exists $\bar{W}_* \in \mathbb{R}^{\bar{K}}$ such that $\mathbf{J}^T(\bar{W}_*) D(\bar{W}_*) = 0$;
3) the Jacobian matrix $\mathbf{J}(\bar{W}_*)$ at $\bar{W}_*$ has full rank $\bar{K}$;
4) $\rho\Big(\big(\mathbf{J}^T(\bar{W}_*)\mathbf{J}(\bar{W}_*)\big)^{-1}\big(\sum_{i=1}^l D_i(\bar{W}_*)\nabla^2 D_i(\bar{W}_*)\big)\Big) < 1$, where $\rho(A)$ indicates the spectral radius of a square matrix A and ∇^2 is the Hessian matrix.

Under the above conditions, there exists $\varepsilon > 0$ such that the sequence $\{\bar{W}_i\}$ generated by the iterative ADP (8) converges to $\bar{W}_*$ for all $\bar{W}_0 \in \mathbb{D} \equiv \{\bar{W} | \|\bar{W} - \bar{W}_*\| < \varepsilon\}$ [2].

The first requirement in Theorem 2 demonstrates the PE condition is necessary, which is commonly premised in adaptive control algorithms. A common approach to guarantee the PE condition is adding probing noise to the control inputs, in our case, i.e. $u = u' + e_u$ and $\omega = \omega' + e_\omega$, where u' and ω' are two policies for which $(f + gu' + k\omega')$ is stable and

[2]Throughout this paper, we use $|\cdot|$ as the magnitude of a scalar, $\|\cdot\|$ as the vector norm of a vector, and $\|\cdot\|_2$ as the induced matrix 2-norm.

e_u and e_ω are the probing noise. The most common probing noise is composed of several sinusoidal signals with different frequencies.

B. Uniformly converge to the HJI solution

In this part, it is to be proved that the NN approximators, $\hat{V}_i, \hat{u}_i, \hat{\omega}_i$, uniformly approximate V_i, u_i, ω_i defined by (4) and (5). The convergence of the algorithm to the HJI solution is concluded afterwards.

Lemma 3: Under Assumption 2, for each $i \geq 0$,

$$\lim_{K_1,K_2,K_3\to\infty} \hat{V}_i(x) = V_i(x)$$
$$\lim_{K_1,K_2,K_3\to\infty} \hat{u}_{i+1}(x) = u_{i+1}(x)$$
$$\lim_{K_1,K_2,K_3\to\infty} \hat{\omega}_{i+1}(x) = \omega_{i+1}(x)$$

The proving process is similar to Theorem 3.1 in [15], so we omit it here.

Theorem 2: Under Assumptions 1 and 2, for any arbitrary $\epsilon > 0$, there exist $i^* > 0$, $K_1^* > 0$, $K_2^* > 0$, $K_3^* > 0$, such that

$$\left|\hat{V}_i(x) - V^*(x)\right| \leq \epsilon$$
$$\left\|\hat{u}_{i+1}(x) - u^*(x)\right\| \leq \epsilon$$
$$\left\|\hat{\omega}_{i+1}(x) - \omega^*(x)\right\| \leq \epsilon$$

hold for all $x \in \Omega$, if $i > i^*$, $K_1 > K_1^*$, $K_2 > K_2^*$, $K_3 > K_3^*$.

Proof: The conclusion is proved from Theorem 1 and Lemma 3. ■

V. Simulation Study

In the experiment we select a nonlinear system from [20] where two online ADP algorithms, CRLA and SPIA, are conducted to solve this nonlinear ZSG. The dynamics is

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ -0.5 * (x_1 + x_2) + 0.5 x_2 \sin(x_1)^2 \end{bmatrix} + \begin{bmatrix} 0 \\ \sin(x_1) \end{bmatrix} u + \begin{bmatrix} 0 \\ \cos(x_1) \end{bmatrix} \omega$$

One selects $h(x) = [x_1, x_2]^T$, $R = 1$, $\gamma = 2$. Note that there exists no analytic solution to the problem. So a fourth order complete polynomial basis function vector is selected for the critic, actor, and disturbance NNs, i.e.

$$\phi_1(x) = \phi_2(x) = \phi_3(x) = [x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, x_1^4, x_1^3x_2, x_1^2x_2^2, x_1x_2^3, x_2^4]^T$$

The total number of parameters is $\bar{K} = 42$. The system starts from $x(0) = [1, -1]^T$ and the integration is conducted at every 0.1s. Initial weights $W_{2,0}$ and $W_{3,0}$ are set to 0. The measurement phase lasts for 20s. Afterwards the learning phase starts training the NN weights. The algorithm needs only 4 iterations to converge. Fig. 1, Fig. 2 and Fig. 3 show the iteration of some parameters in the critic, actor and disturbance NNs. The final actor NN is formulated as

$$\hat{u}_4(x) = [0.0230, 0.0109, 0.1931, -0.8625, 0.0019, -0.0786, -0.0498, 0.0082, -0.0025, -0.0447, 0.0831, -0.0118, 0.0054, -0.0013]\phi_2(x)$$

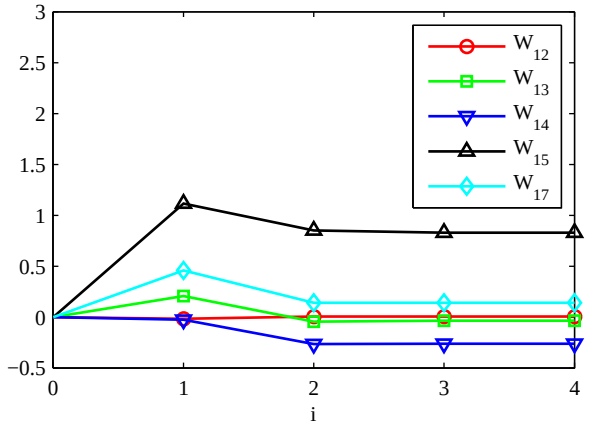

Fig. 1. Iteration of W_1 in the learning phase of Example 2.

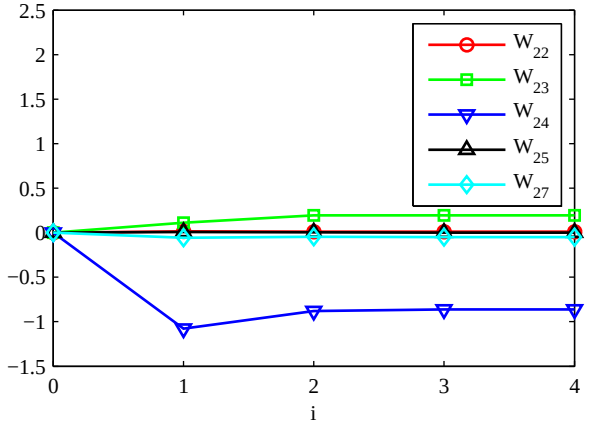

Fig. 2. Iteration of W_2 in the learning phase of Example 2.

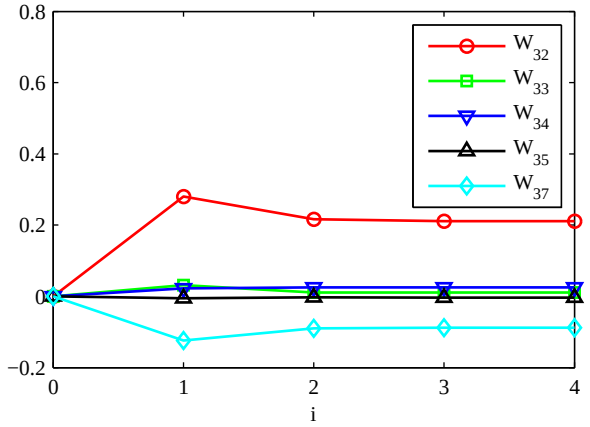

Fig. 3. Iteration of W_3 in the learning phase of Example 2.

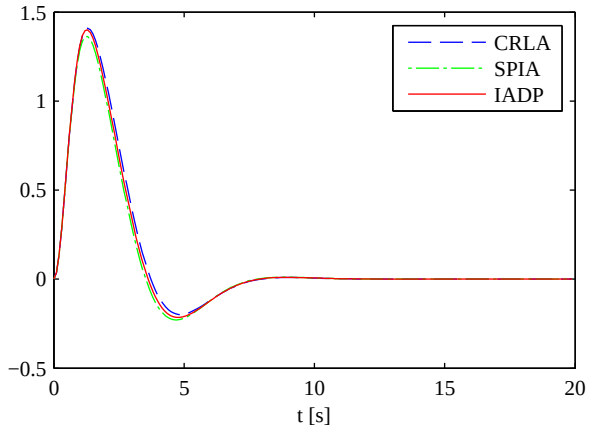

Fig. 4. Trajectories of x_1 for iterative ADP algorithm (IADP), concurrent reinforcement learning algorithm (CRLA), and synchronous policy iteration algorithm (SPIA).

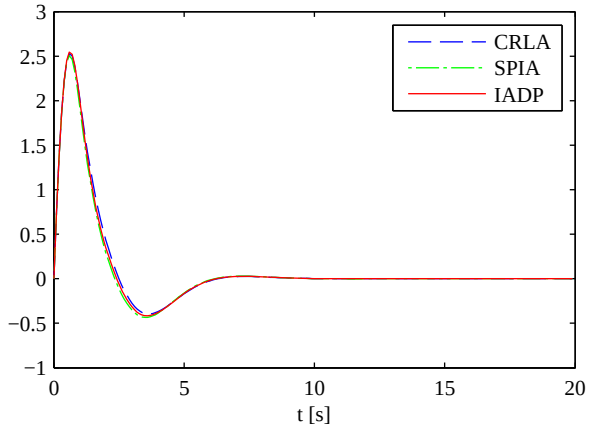

Fig. 5. Trajectories of x_2 for iterative ADP algorithm (IADP), concurrent reinforcement learning algorithm (CRLA), and synchronous policy iteration algorithm (SPIA).

In our algorithm, the time length of the online measurement is 20s, compared to CRLA needs 270s and SPIA needs more than 800s for their online learning [20]. Besides, the implementation of CRLA and SPIA relies on parts of the system dynamics, while no dynamics is needed here.

Next, our converged actor is compared with the results of CRLA and SPIA, provided by [20], in a same finite-energy run. The system is at rest and the disturbance is set to $\omega(t) = 8\cos(t)\exp^{-t}$. Fig. 4 and Fig. 5 illustrate the trajectories of the state variables when executing three controllers separately. From the plots, it is revealed that the difference of performance between the three converged actors is barely noticeable except that CRLA leads the best attenuating effect. Our algorithm performs closely to CRLA and SPIA is the worst.

VI. Conclusion

The continuous-time unknown nonlinear zero-sum game is approximately solved by a model-free iterative ADP algorithm using online measurement in this paper. With additional control inputs, system trajectories contain the complete information of the dynamics, and is properly processed in the algorithm to train the NN approximators of the value, control and disturbance policies. The same online data can be used repeatedly to conduct the iteration and produce the converged solution, which contributes to shorten the learning time.

References

[1] A. Isidori and W. Kang, "H_∞ control via measurement feedback for general nonlinear systems," *IEEE Trans. Autom. Control*, vol. 40, no. 3, pp. 466–472, Mar 1995.

[2] A. J. Van der Schaft, *L2-Gain and Passivity Techniques in Nonlinear Control*, 1st ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1996.

[3] S. Tijs, *Introduction to Game Theory*. India: Hindustan Book Agency, 2003.

[4] F. Lewis and D. Vrabie, "Reinforcement learning and adaptive dynamic programming for feedback control," *IEEE Circuits Syst. Mag.*, vol. 9, no. 3, pp. 32–50, 2009.

[5] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 39–47, May 2009.

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

[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, 2011.

[8] T. Dierks and S. Jagannathan, "Optimal control of affine nonlinear continuous-time systems using an online Hamilton-Jacobi-Isaacs formulation," in *49th IEEE Conf. Decision and Control (CDC),*, Atlanta, GA, USA, 2010, pp. 3048–3053.

[9] K. G. Vamvoudakis and F. Lewis, "Online solution of nonlinear two-player zero-sum games using synchronous policy iteration," *Int. J. Robust and Nonlinear Control*, vol. 22, no. 13, pp. 1460–1483, 2012.

[10] K. G. Vamvoudakis and F. L. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[11] H.-N. Wu and B. Luo, "Neural network based online simultaneous policy update algorithm for solving the HJI equation in nonlinear H_∞ control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 12, pp. 1884–1895, Dec 2012.

[12] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[13] H. Zhang, C. Qin, B. Jiang, and Y. Luo, "Online adaptive policy learning algorithm for H_∞ state feedback control of unknown affine nonlinear discrete-time systems," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2706–2718, Dec 2014.

[14] H. Modares, F. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1513–1525, Oct 2013.

[15] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming for optimal nonlinear control design," 2013, arXiv:1303.2247 [math.DS].

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

[17] H. Li, D. Liu, and D. Wang, "Integral reinforcement learning for linear continuous-time zero-sum games with completely unknown dynamics," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 3, pp. 706–714, 2014.

[18] H. Modares, F. 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.*, 2015, in press.

[19] S. Gratton, A. S. Lawless, and N. K. Nichols, "Approximate Gauss-Newton methods for nonlinear least squares problems," *SIAM J. Optimization*, vol. 18, no. 1, pp. 106–132, 2007.

[20] S. Yasini, A. Karimpour, M.-B. Naghibi Sistani, and H. Modares, "Online concurrent reinforcement learning algorithm to solve two-player zero-sum games for partially unknown nonlinear continuous-time systems," *Int. J. Adaptive Control and Signal Processing*, vol. 29, no. 4, pp. 473–493, 2015.