Event-triggered integral reinforcement learning for nonlinear continuous-time systems

Qichao Zhang
The state Key Laboratory of Management
and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences,
University of Chinese Academy of Sciences, Beijing, China,
Email: zhangqichao2013@163.com

Dongbin Zhao
The state Key Laboratory of Management
and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences,
University of Chinese Academy of Sciences, Beijing, China,
Email: dongbin.zhao@ia.ac.cn

***Abstract*—In this paper, the optimal control problem for the continuous-time nonlinear systems with partially unknown dynamics is investigated. The event-triggered internal reinforcement learning (IRL) is proposed to approach the solution of the Hamilton-Jacobi-Bellman (HJB) equation. Note that the knowledge of internal dynamics is relaxed, and the event-triggered control scheme is adopted to reduce the computational burden and communication resources. For the online implementation purpose, a single-critic neural network (NN) structure is constructed to approach the optimal value function and the optimal policy with convergence analysis. Finally, a simulation example is provided to demonstrate the effectiveness of the proposed algorithm.**

***Index Terms*—Internal reinforcement learning, event-triggered, neural network, online learning**

I. Introduction

As an important branch of machine learning, reinforcement learning (RL) [1] has received spreading attention in various fields such as economics [2], optimal control [3], and so on. In general, we can divide RL in the optimal control filed into two categories: model-based RL and model-free RL. For the model-based RL, the knowledge of system dynamics is required to design the optimal controller. For the model-free RL, the system data is required and utilized to obtain the optimal policy rather than the knowledge of system dynamics. It should be mentioned that integral RL (IRL) proposed in [4] is a main technique to relax the knowledge of the internal dynamics for continuous-time (CT) nonlinear systems, which means that only partially knowledge of system dynamics is required.

IRL technique has been widely used to solve the optimal tracking problem, zero-sum game [5], nonzero-sum game [6], and so on. Recently, combining the off-policy scheme and IRL technique, an offline iterative learning algorithm is proposed for partially unknown ZS game [7]. Furthermore, a data-driven RL method without the knowledge of system dynamics is proposed based on IRL for uncertain system [8], H_∞ [9], fully cooperative games [10], and so on. The NN identifier with an approximation error [11], which is usually time-consuming, is not required in the data-driven RL approach based on off-policy scheme and IRL technique. Note that only the collected system data is required instead of the knowledge of system dynamics, which means that the data-driven RL is a model-free RL method.

Event-triggered control (ETC) scheme, which has been widely investigated in the communication resources-limited wireless sensor network, can save communication resources and reduce computational burdens effectively. Recently, the event-triggered scheme has been integrated with the RL to design an event-triggered optimal controller for the nonlinear system [12]. Sahoo *et al.* proposed an event-triggered NN controller for the unknown nonlinear CT system [13]. Zhang *et al.* proposed an event-triggered adaptive dynamic programming (ADP) method for zero-sum games [14] and uncertain nonlinear systems [15], respectively. The event-triggered RL method was applied to the load frequency control of power systems in [16]. In [17], the event-triggered ADP approach was developed for the nonlinear systems without requiring exact knowledge of internal system dynamics. Then, the event-triggered optimal control for partially unknown nonlinear systems, where the control input was constrained, was proposed in [18]. It should be mentioned that the NN-identifier is required to identify the unknown internal dynamics in [17, 18].

To the best of our knowledge, there is still no event-triggered IRL algorithms for the optimal control of CT nonlinear systems with partially unknown dynamics. It motivates our research. In this paper, a novel event-triggered IRL algorithm is proposed to design the optimal controller by approximating the solution of the HJB equation. Note that the identification process in [17, 18] is not required any more. The event-triggered tuning law is given with the single critic NN structure. Then, a triggering condition is designed to guarantee the UUB of the critic weights. Simulation results prove the effectiveness of the proposed scheme.

II. Preliminary

A. Problem Statement

Consider the continuous-time nonlinear system given by

$$\dot{x}(t) = f(x) + g(x)u(t), \tag{1}$$

with the system state $x(t) \in \mathbf{R}^n$, the nonlinear functions $f(x) \in \mathbf{R}^n$, $g(x) \in \mathbf{R}^{n\times m}$, and the control input $u(t) \in \mathbf{R}^m$. Let $f(0) = 0$, $f(x) + g(x)u$ is Lipschitz continuous on a set

978-1-5386-2726-6/17/$31.00 ©2017 IEEE

$\Omega \in \mathbf{R}^n$ that contains the origin. Assume that the system is stabilizable on Ω.

Define an infinite-horizon integral cost function with control policy $u(t)$ as

$$\begin{aligned} V(x_0) &= \int_0^\infty \big(x^T Q x + u^T R u\big) d\tau \\ &\triangleq \int_t^\infty U(x(\tau), u(\tau)) d\tau \end{aligned} \tag{2}$$

where the utility $U(x,u) = x^T Q x + u^T(t) R u(t)$. Here, Q and R are positive definite symmetric matrices with $R = rr^T$. r is an appropriate lower triangular matrix.

The value function with any admissible feedback control policy $u(x) \in \Psi(\Omega)$ is

$$V(x(t)) = \int_t^\infty U\big(x(\tau), u(x(\tau))\big) d\tau \tag{3}$$

If the value function (3) is continuously differentiable, the CT Bellman equation is

$$U\big(x, u(x)\big) + \big(\nabla V(x)\big)^T \big(f(x) + g(x)u(x)\big) = 0, V(0) = 0 \tag{4}$$

where $\nabla V(x) = \partial V(x)/\partial x$.

It is desired to find an optimal control policy such that the value function is minimized. Define the Hamiltonian function as

$$H(x, u, \nabla V) = U\big(x(t), u(x)\big) + \big(\nabla V(x)\big)^T \big(f(x) + g(x)u(x)\big) \tag{5}$$

The optimal value function $V^*(x)$ satisfies the HJB equation

$$0 = \min_{u \in \Psi(\Omega)} H(x, u, \nabla V^*) \tag{6}$$

Accordingly, the optimal control policy is

$$u^*(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V^*(x) \tag{7}$$

With the control policy (7), the HJB equation can be rewritten as

$$\begin{aligned} 0 =& x^T Q x + \big(\nabla V(x)\big)^T f(x) \\ &- \frac{1}{4} \big(\nabla V(x)\big)^T g(x) R^{-1} g^T(x) \nabla V(x) \end{aligned} \tag{8}$$

Note that solving the HJB equation is difficult due to its inherent nonlinear property. The policy iteration (PI) algorithm is usually used to solve the HJB equation iteratively.

Explicitly the complete knowledge of the system dynamics $f(x)$ and $g(x)$ is required. According to [4], given an admissible policy and an integration time interval T, the IRL algorithm is presented without the knowledge of the internal dynamics $f(x)$. During the policy evaluation, the internal dynamics $f(x)$ is relaxed by the integral operation for (9) on the time interval $[t, t+T]$ and only the dynamics $g(x)$ is required for IRL.

Remark 1: The convergence of the IRL algorithm is proven in [4]. In fact, solving for $V^i(x)$ in (11) is equivalent to finding the solution of (9). Based on the convergence results of PI algorithm, the IRL algorithm also converges to the solution of the HJB equation on trajectories originating in Ω [6]. Note that an initial admissible policy is required for both PI and IRL algorithms, which is usually chosen based on experimental experience.

Algorithm 1 (Policy Iteration)

1: Select an initial admissible policy $u^0(x) \in \Psi(\Omega)$.
2: **Policy Evaluation.** Solve for $V^k(x)$ using

$$0 = U\big(x, u^k(x)\big) + \big(\nabla V^k(x)\big)^T \big(f(x) + g(x)u^k(x)\big) \tag{9}$$

with $V^k(0) = 0$.
3: **Policy Improvement.** Update the control policy

$$u^{k+1}(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V^k(x) \tag{10}$$

Algorithm 2 (Integral Reinforcement Learning)

1: Select an initial admissible policy $u^0(x) \in \Psi(\Omega)$.
2: **Policy Evaluation.** Solve for $V^k(x)$ using

$$V^k\big(x(t)\big) = \int_t^{t+T} U\big(x, u^k(x)\big) d\tau + V^k\big(x(t+T)\big) \tag{11}$$

with $V^k(0) = 0$.
3: **Policy Improvement.** Update the control policy

$$u^{k+1}(x) = -\frac{1}{2} R^{-1} g^T(x) \nabla V^k(x)$$

III. Event-triggered IRL

To propose the ETC mechanism, we first define a monotonically increasing sequence of triggering instants $\{\tau_j\}_{j=0}^\infty$, where τ_j is the j^{th} consecutive sampling instant with $\tau_j < \tau_{j+1}, j \in \mathbf{N}$ with $\mathbf{N} = \{0, 1, 2, \cdots\}$. Then a sampled-data system characterized by the triggering instants is introduced, where the controller is updated based on the sampled state $\hat{x}_j = x(\tau_j)$ for all $t \in [\tau_j, \tau_{j+1})$. Define the event-based error as

$$e_j(t) = \hat{x}_j - x(t), \forall t \in [\tau_j, \tau_{j+1}), j \in \mathbf{N}, \tag{12}$$

where $x(t)$ and $\hat{x}_j$ denote the current state and the sampled state, respectively.

In the event-based control method, the triggering instants are determined by a triggering condition. Generally, the triggering condition is determined by the event-based error and a designed state-dependent threshold. When the event-based error exceeds the state-dependent threshold, an event is triggered. Then, the system states are sampled, which resets the event-based error $e_j(t)$ to zero. Accordingly, the designed event-based controller $u(\hat{x}_j) \triangleq \mu(\hat{x}_j)$ is updated. Note that the system states are held until the next triggering instant. Clearly, the control signal $\mu(\hat{x}_j)$ is a function of the event-based state vector, which is executed based on the latest sampled state $\hat{x}_j$ instead of the current value $x(t)$. That is,

the event-based controller is only updated at the triggering instant sequence $\{\tau_j\}_{j=0}^{\infty}$, and remains unchanged in each time interval $t \in [\tau_j, \tau_{j+1})$. Hence, this control signal $\mu(\hat{x}_j)$ with $j \in \mathbf{N}$ is a piecewise constant function on each segment $[\tau_j, \tau_{j+1})$.

With the event-based control input $\mu(\hat{x}_j)$, the sampled-data version of the system (1) can be rewritten as

$$\dot{x}(t) = f(x) + g(x)\mu\left(x(t) + e_j(t)\right) \tag{13}$$

Considering the event-based sampling rule, the optimal control policy (7) becomes

$$\mu^*(\hat{x}_j) = -\frac{1}{2}R^{-1}g^T(\hat{x}_j)\nabla V^*(\hat{x}_j) \tag{14}$$

for all $t \in [\tau_j, \tau_{j+1})$, where $\nabla V^*(\hat{x}_j) = \partial V^*(x)/\partial x|_{x=\hat{x}_j}$.

Using the event-triggered control policy (14), the Hamiltonian function becomes

$$\begin{aligned}
&H(x, \mu^*(\hat{x}_j), \nabla V^*) \\
=&x^T Q x + \big(\mu^*(\hat{x}_j)\big)^T R\mu^*(\hat{x}_j) + \big(\nabla V^*(x)\big)^T\big(f + g\mu^*(\hat{x}_j)\big) \\
=&\big(\mu^*(\hat{x}_j)\big)^T R\mu^*(\hat{x}_j) - 2\big(u^*(x)\big)^T R\mu^*(\hat{x}_j) + \big(u^*(x)\big)^T Ru^*(x) \\
=&\|r^T\big(u^*(x) - \mu^*(\hat{x}_j)\big)\|^2
\end{aligned} \tag{15}$$

The integral form of equation (15) along the time interval $[t, t+T]$ can be described as

$$\begin{aligned}
&\int_t^{t+T} U\big(x(\tau), \mu^*(\hat{x}_j)\big)d\tau + V^*\big(x(t+T)\big) - V^*(x(t)) \\
=&\int_t^{t+T} \|r^T\big(u^*(x) - \mu^*(\hat{x}_j)\big)\|^2 d\tau
\end{aligned} \tag{16}$$

For the convenience of analysis, the following assumption is introduced.

Assumption 1: The optimal controller $u^*(x)$ is Lipschitz continuous with respect to the event-based error,

$$\|u^*(x) - u^*(\hat{x}_j)\| = \|u^*(x) - u^*(x + e_j)\| \le l\|e_j\|,$$

where l is a positive real constant.

Furthermore, we present the following event-triggered IRL algorithm.

Algorithm 3 (Event-triggered IRL)

1: Select an initial admissible policy $u^0(x) \in \Psi(\Omega)$.

2: **Policy Evaluation.** Solve for $V^k(x(t))$ using

$$\int_t^{t+T} U\big(x(\tau), \mu^*(\hat{x}_j)\big)d\tau + V^*\big(x(t+T)\big) - V^*(x(t)) = 0, \tag{17}$$

with $t = \tau_j, j \in \mathbf{N}$ and $V^k(0) = 0$.

3: **Policy Improvement.** Update the event-triggered control policy

$$\mu^{k+1}(\hat{x}_j) = -\frac{1}{2}R^{-1}g^T(\hat{x}_j)\nabla V^k(\hat{x}_j), t = \tau_j, j \in \mathbf{N}.$$

Compared with the IRL algorithm, the policy evaluation and policy improvement for the event-triggered IRL are only occurred at the triggering instants $t = \tau_j$ with the sampled state $\hat{x}_j$. Note that at the triggering instants, we have $u^*(x) = \mu^*(\hat{x}_j)$, which means the right side of the equation (15) is equal to zero. The Hamilton function $H(x, \mu^*(\hat{x}_j), \nabla V^*)$ is not equal to zero during the triggering intersample time, as a transformation error is introduced because of the event-based transformation from (7) to (14). A suitable triggering condition should be designed to attain a tradeoff between the stabilization and resource utilization. In the next section, we will propose the NN-based event-triggered IRL with a designed triggering condition which can guarantee the convergence of the critic weights and the closed-loop system.

IV. Approximate Optimal Controller Design

In this section, a critic network is constructed to approximate the optimal value function. The NN-based optimal value function can be formulated as

$$V^*(x) = w_c^T\phi(x) + \varepsilon(x) \tag{18}$$

where $w_c \in \mathbf{R}^N$ are the ideal weights of critic NN which in turn corresponds to the *stabilizing* solution of the underlying HJB equation, $\phi(x) \in \mathbf{R}^N$ is the activation function vector, N is the number of hidden neurons, and $\varepsilon \in \mathbf{R}$ is the approximation error of critic NN.

Assumption 2:

(a) The NN approximation error and its gradient are bounded over the compact set $\mathbf{R}$, i.e., $\|\varepsilon(x)\| \le b_\varepsilon$ and $\|\nabla\varepsilon\| \le b_{\nabla\varepsilon}$.

(b) The NN activation function and its gradient are bounded, i.e., $\|\phi(x)\| \le b_\phi$ and $\|\nabla\phi(x)\| \le b_{\nabla\phi}$.

Using the value function approximation (18) in event-triggered Bellman equation (17), it becomes

$$w_c^T\left(\phi(x(\tau_j + T)) - \phi(x(\tau_j))\right) + \int_{\tau_j}^{\tau_j+T} U(x, \mu(\hat{x}_j)d\tau = \Delta\varepsilon_{\tau_j} \tag{19}$$

where $\Delta\varepsilon_{\tau_j} = \int_{\tau_j}^{\tau_j+T} \nabla\varepsilon^T(f + g\mu(\hat{x}_j))d\tau$. Under Assumption 1, we can deduce that $\Delta\varepsilon_{\tau_j}$ is also bounded on the compact set $\mathbf{R}$, i.e., $\Delta\varepsilon_{\tau_j} \le \Delta\varepsilon_M$.

Since the ideal weight matrix is unknown, the actual output of critic NN can be presented as

$$\hat{V}(x) = \hat{w}_c^T\phi(x), \tag{20}$$

where $\hat{w}_c$ represents the estimation of the unknown weight matrix w_c.

Based on the value function approximation, the control policy is approximated by

$$\hat{\mu}(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla\phi(x)\hat{w}_c \tag{21}$$

Define the estimation errors of critic NN as

$$\tilde{w}_c = w_c - \hat{w}_c \tag{22}$$

Therefore, the approximate Bellman equation becomes

$$\int_t^{t+T} U(x,\hat{\mu}(x))d\tau+\hat{w}_c^T(t)\left(\phi\big(x(t+T)\big)-\phi\big(x(t)\big)\right)=e(t) \tag{23}$$

Denote that $\Delta\phi\big(x(t)\big) \triangleq \phi\big(x(t+T)\big)-\phi\big(x(t)\big)$. The Hamiltonian error $e(t)$ provides useful information to adapt the estimation weights. For the event-triggered mechanism, transmitted measurements are available only at triggering instants. Therefore, we define the Hamiltonian error at the triggering instant t_j as

$$\begin{aligned}&\int_{\tau_j}^{\tau_j+T} U(\hat{x}_j,\hat{\mu}(\hat{x}_j))d\tau+\hat{w}_c^T(\tau_j)\left(\phi\big(x(\tau_j+T)\big)-\phi\big(x(\tau_j)\big)\right)\\&=e(\tau_j)\end{aligned} \tag{24}$$

where $\mu(\hat{x}_j)=-\frac{1}{2}R^{-1}g^T(\hat{x}_j)\nabla\phi(\hat{x}_j)\hat{w}_c$ denotes the event-triggered control policy. The error can be reduced by a gradient-descent method with the partial derivative toward $\hat{w}_c$

$$\frac{\partial e(\tau_j)}{\partial\hat{w}_c}=\phi\big(x(\tau_j+T)\big)-\phi\big(x(\tau_j)\big)\triangleq\Delta\phi_{\tau_j}$$

Let the tuning of critic NN be provided by

$$\begin{cases}\hat{w}_c(\tau_j)=\hat{w}_c(\tau_{j-1})-\alpha_c\frac{\Delta\phi_{\tau_j}}{\left(1+\Delta\phi_{\tau_j}^T\Delta\phi_{\tau_j}\right)^2}e(\tau_j),\ t=\tau_j\\ \dot{\hat{w}}_c(t)=0,\quad \tau_{j-1}<t<\tau_j\end{cases} \tag{25}$$

where $\alpha_c>0$ denotes the learning rate.

Now, we will give the triggering condition to guarantee the convergence of the critic weights and the closed-loop system under the event-triggered updating law (25).

Theorem 1: Consider the nonlinear system with (1) and the critic NN with (20). The critic NN is updated by (25). The transmission of system states following the triggering condition

$$\|e_j(t)\|^2\le\frac{(1-\beta^2)\lambda_{\min}(Q)\|x\|^2+\|r^T\mu^*(\hat{x}_j)\|^2}{l^2\|r^T\|^2} \tag{26}$$

where $\lambda_{\min}$ is the minimal eigenvalue of Q, and $\beta\in(0,1)$ is a designed sample frequency parameter. Then, the closed-loop system state and critic estimation errors are Uniformly Ultimately Bounded (UUB).

Proof: Consider the following Lyapunov function candidate

$$\begin{aligned}L=&L_v+L_c\\=&\int_{t-T}^{t}V(x)d\tau+\frac{1}{2}\tilde{w}_c^T\alpha_c^{-1}\tilde{w}_c\end{aligned} \tag{27}$$

1) *(In interevent intervals, $t\in(\tau_j,\tau_{j+1})$)* During any triggering intervals, the critic NN weights remain unchanged. So the differential of L only includes $\dot{L}_v$. Along the evolution of $(f+g\mu(\hat{x}_j))$, the time derivative of L_v is

$$\begin{aligned}\dot{L}_v=&\int_{t-T}^{t}(\nabla V(x))^T\big(f+g\mu(\hat{x}_j)\big)d\tau\\=&\int_{t-T}^{t}-x^TQx-u^TRu+(\nabla V(x))^Tg\big(\mu(\hat{x}_j)-u(x)\big)d\tau\\=&\int_{t-T}^{t}-x^TQx+u^T(x)Ru(x)-2u^T(x)R\mu(\hat{x}_j)d\tau\\=&\int_{t-T}^{t}-x^TQx+\|r^Tu(x)-r^T\mu(\hat{x}_j)\|^2-\|r^T\mu(\hat{x}_j)\|^2d\tau\end{aligned} \tag{28}$$

Based on Assumption 1, we have

$$\dot{L}_v\le\int_{t-T}^{t}-\lambda_{\min}(Q)\|x\|^2+l^2\|r^T\|^2\|e_k(t)\|^2-\|r^T\mu(\hat{x}_j)\|^2d\tau \tag{29}$$

If the triggering condition is designed as (26), we can conclude that

$$\dot{L}=\dot{L}_v\le-\beta^2\lambda_{\min}(Q)\|x\|^2 \tag{30}$$

2) *(At triggering instants, $t=\tau_{j+1}$)* Based on (30), we have $\dot{L}_v\le0$ during the interevent intervals. Note that the system state x is continuous for the sample-data system. Hence, for $\forall t=\tau_{j+1}$, we have $L_v(\hat{x}_{j+1})\le L_v((\hat{x}_{j+1}^-))$. According to (27), the difference of L_c is written as

$$\begin{aligned}\Delta L_c=&L_c(\tau_{j+1})-L_c(\tau_{j+1}^-)\\=&\frac{1}{2}\tilde{w}_c^T(\tau_{j+1})\alpha_c^{-1}\tilde{w}_c(\tau_{j+1})-\frac{1}{2}\tilde{w}_c^T(\tau_j)\alpha_c^{-1}\tilde{w}_c(\tau_j)\end{aligned} \tag{31}$$

According to the tuning law (25) of critic NN, we have its discrete-time dynamical form

$$\hat{w}_c(\tau_{j+1})=\hat{w}_c(\tau_j)-\alpha_c\frac{\Delta\phi_{\tau_j}}{\left(1+\Delta\phi_{\tau_j}^T\Delta\phi_{\tau_j}\right)^2}e(\tau_j) \tag{32}$$

Denote that $m_s=\left(1+\Delta\phi_{\tau_j}^T\Delta\phi_{\tau_j}\right)$. Then, we have

$$\begin{aligned}\tilde{w}_c(\tau_{j+1})=&\tilde{w}_c(\tau_j)+\alpha_c\frac{\Delta\phi_{\tau_j}}{m_s^2}\\&\times\left(\int_{\tau_j}^{\tau_j+T}U(\hat{x}_j,\hat{\mu}(\hat{x}_j))d\tau+\hat{w}_c^T(\tau_j)\Delta\phi_{\tau_j}\right)\\=&\tilde{w}_c(\tau_j)-\alpha_c\frac{\Delta\phi_{\tau_j}\Delta\phi_{\tau_j}^T}{m_s^2}\tilde{w}_c(\tau_j)\\&+\alpha_c\frac{\Delta\phi_{\tau_j}}{m_s^2}\big(w_c^T\Delta\phi_{\tau_j}+\int_{\tau_j}^{\tau_j+T}U(\hat{x}_j,\hat{\mu}(\hat{x}_j))d\tau\big)\\=&\tilde{w}_c(\tau_j)-\alpha_c\frac{\Delta\phi_{\tau_j}\Delta\phi_{\tau_j}^T}{m_s^2}\tilde{w}_c(\tau_j)+\alpha_c\frac{\Delta\phi_{\tau_j}}{m_s^2}\Delta\varepsilon_{\tau_j}\end{aligned} \tag{33}$$

where $\Delta\varepsilon_{\tau_j}=\int_{\tau_j}^{\tau_j+T}\nabla\varepsilon^T\dot{x}d\tau$.

So ΔL_c has

$$\begin{aligned}\Delta L_c =&\frac{1}{2}\tilde{w}_c^T(\tau_{j+1})\alpha_c^{-1}\tilde{w}_c(\tau_{j+1}) - \frac{1}{2}\tilde{w}_c^T(\tau_j)\alpha_c^{-1}\tilde{w}_c(\tau_j)\\ =&-\tilde{w}_c^T(\tau_j)\frac{\Delta\phi_{\tau_j}\Delta\phi_{\tau_j}^T}{m_s^2}\tilde{w}_c(\tau_j)+\tilde{w}_c^T(\tau_j)\frac{\Delta\phi_{\tau_j}}{m_s^2}\Delta\varepsilon_{\tau_j}\\ &+\frac{\alpha_c}{2}\left(-\frac{\Delta\phi_{\tau_j}\Delta\phi_{\tau_j}^T}{m_s^2}\tilde{w}_c(\tau_j)+\frac{\Delta\phi_{\tau_j}}{m_s^2}\Delta\varepsilon_{\tau_j}\right)^2\\ \leq&-\frac{1}{2}\left\|\frac{\Delta\phi_{\tau_j}\Delta\phi_{\tau_j}^T}{m_s^2}\right\|\|\tilde{w}_c(\tau_j)\|^2+\frac{\Delta\varepsilon_{\tau_j}^2}{2\|m_s^2\|}\\ &+\alpha_c\left\|\frac{\Delta\phi_{\tau_j}}{m_s^2}\right\|^2\Delta\varepsilon_{\tau_j}^2+\alpha_c\left\|\frac{\Delta\phi_{\tau_j}\Delta\phi_{\tau_j}^T}{m_s^2}\right\|^2\|\tilde{w}_c(\tau_j)\|^2\end{aligned} \tag{34}$$

Note that $\left\|\frac{\Delta\phi_{\tau_j}\Delta\phi_{\tau_j}^T}{m_s^2}\right\|<1$, we have

$$\begin{aligned}\Delta L\leq\Delta L_c\leq&-(\frac{1}{2}-\alpha_c)\left\|\frac{\Delta\phi_{\tau_j}\Delta\phi_{\tau_j}^T}{m_s^2}\right\|\|\tilde{w}_c(\tau_j)\|^2\\ &+\frac{\|m_s^2\|+2\alpha_c\|\Delta\phi_{\tau_j}\|^2}{2\|m_s^2\|^2}\Delta\varepsilon_M^2\end{aligned} \tag{35}$$

Therefore, if the following conditions are satisfied:

$$\alpha_c<\frac{1}{2}$$

$$\|\tilde{w}_c(\tau_j)\|\leq\sqrt{\frac{2\left(\|m_s^2\|+\alpha_c\|\Delta\phi_{\tau_j}\|^2\right)\Delta\varepsilon_M^2}{(1-2\alpha_c)\|m_s^2\|\|\Delta\phi_{\tau_j}\|^2}}$$

Combining (30) and (35), we can know that the closed-loop system and the critic NN are UUB under the triggering condition (26), which means it will uniformly converge to the optimal solution. This completes the proof. ■

V. Simulation Study

Consider the following nonlinear system:

$$\dot{x}=f(x)+g(x)u \tag{36}$$

where

$$f(x)=\begin{bmatrix}-x_1+x_2\\ -0.5(x_1+x_2)+0.5x_2\sin(x_1)^2\end{bmatrix},$$

$$g(x)=\begin{bmatrix}0\\ \sin(x_1)\end{bmatrix}.$$

Let Q and R be identity matrices of approximate dimensions. The parameter of triggering condition is chosen as $l=5$, $\beta=0.9$. The learning rate is chosen as $\alpha_c=0.2$, and the integral interval $T=0.05s$. We choose the critic NN with structure $3-5-1$. The critic NN activation function is chosen as $\phi(x)=[x_1^2\ x_1x_2\ x_2^2\ x_1^4\ x_2^4]^T$. Let the initial state be $x_0=[1,-1]^T$. We choose the initial weights of critic NN as $\hat{w}_c(0)=[0.327,0.018,0.425,0.467,0.339]^T$. Note that an exponent-form probing noise is added to the control input before 25s to trade off exploration and exploitation in the proposed RL algorithm. The trajectories of system states are shown in Fig. 1. The system states are converged to the equilibrium point rapidly after 23s. The convergence curves of critic weights are shown in Fig. 2. The triggering condition is triggered 118 times, which means that the event-based controller uses 118 samples of the states while the time-triggered controller uses 800 samples. This will reduce the number of controller updates during the learning process. The triggering instants during the learning process for the control policy is illustrated in Fig. 3. The trajectory of the event-triggered control input during the learning process is shown in Fig 4. The simulation results prove the effectiveness of the proposed event-triggered IRL algorithm.

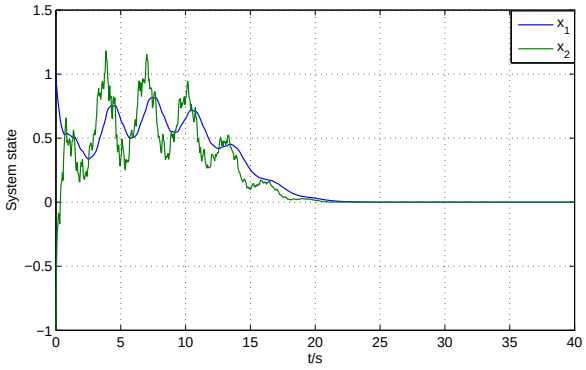

Fig. 1. Trajectories of system state

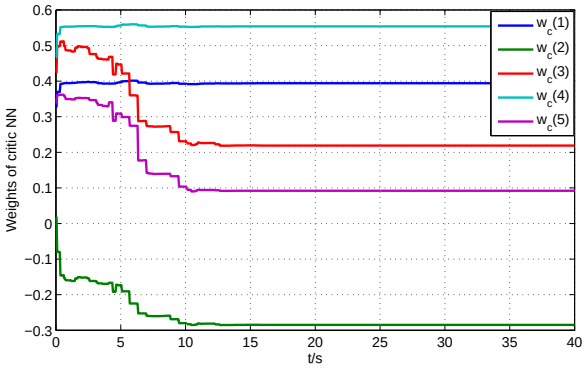

Fig. 2. Convergence of the critic parameters

VI. Conclusion

In this paper, we propose an event-triggered IRL algorithm for the nonlinear continuous-time systems with partially unknown dynamics. Compared with [17, 18], the identification process is not required. For the implementation purpose, a critic NN is constructed to approximate the optimal value function and optimal control policy. Then, the UUB of critic weights are guarantee using Lyapunov method. Simulation results prove the effectiveness of the proposed event-triggered IRL scheme.

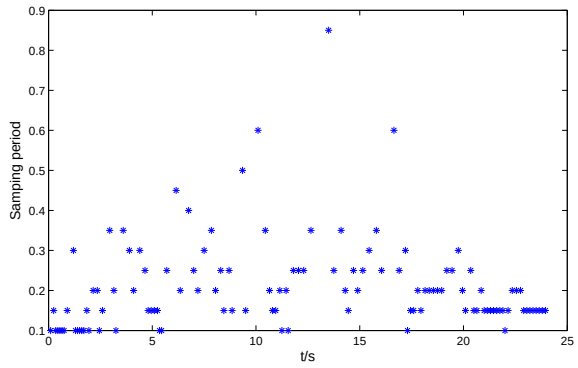

Fig. 3. Triggering instants during the learning process

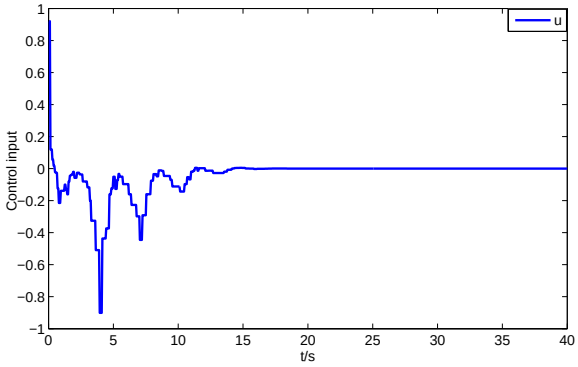

Fig. 4. Trajectory of the event-triggered control input

ACKNOWLEDGMENT

This research is supported by National Natural Science Foundation of China (NSFC) under Grants No. 61573353, No. 61533017, by the National Key Research and Development Plan under Grants 2016YFB0101000.

REFERENCES

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[2] L. Tesfatsion, "Introduction to the special issue on agent-based computational economics," *Journal of Economic Dynamics and Control*, vol. 25, no. 3, pp. 281–293, 2001.

[3] D. Zhao, Z. Xia, and Q. Zhang, "Model-free optimal control based intelligent cruise control with hardware-in-the-loop demonstration," *IEEE Computational Intelligence Magazine*, vol. 12, no. 2, pp. 56–69, 2017.

[4] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237–246, 2009.

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

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

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

[8] Y. Jiang and Z.-P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 5, pp. 882–893, 2014.

[9] Y. Zhu, D. Zhao, and X. Li, "Iterative adaptive dynamic programming for solving unknown nonlinear zero-sum game based on online data," *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2016.2561300, 2016.

[10] Q. Zhang, D. Zhao, and Y. Zhu, "Data-driven adaptive dynamic programming for continuous-time fully cooperative games with partially constrained inputs," *Neurocomputing*, vol. 238, pp. 377–386, 2017.

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

[12] K. G. Vamvoudakis, "Event-triggered optimal adaptive control algorithm for continuous-time nonlinear systems," *IEEE/CAA J. Autom. Sin.*, vol. 1, no. 3, pp. 282–293, 2014.

[13] A. Sahoo, H. Xu, and S. Jagannathan, "Neural network-based event-triggered state feedback control of nonlinear continuous-time systems," *IEEE Transactions on Neural Networks and Learning Systems*, DOI10.1109/TNNLS.2015.2416259, 2015.

[14] Q. Zhang, D. Zhao, and Y. Zhu, "Event-triggered h control for continuous-time nonlinear system via concurrent learning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, DOI: 10.1109/TSMC.2016.2531680, 2017.

[15] Q. Zhang, D. Zhao, and D. Wang, "Event-based robust control for uncertain nonlinear systems using adaptive dynamic programming," *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2016.2614002, 2017.

[16] L. Dong, Y. Tang, H. He, and C. Sun, "An event-triggered approach for load frequency control with supplementary adp," *IEEE Transactions on Power Systems*, vol. 32, no. 1, pp. 581–589, 2017.

[17] X. Zhong and H. He, "An event-triggered adp control approach for continuous-time system with unknown internal states," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 683–694, 2017.

[18] Y. Zhu, D. Zhao, H. He, and J. Ji, "Event-triggered optimal control for partially unknown constrained-input systems via adaptive dynamic programming," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 4101–4109, 2017.