Policy Iteration Based Optimal Control for Partially Unknown Nonlinear Systems via Semi-parametric Regression Model

Jingliang Sun, Chunsheng Liu, and Nian Liu

***Abstract*—This paper presents a novel online learning algorithm, in an actor-critic structure, to find state-feedback optimal controllers for partially unknown nonlinear systems. The algorithm converges online to the optimal solution under the condition of initial stabilizing controller. It is derived from integral reinforcement learning (IRL) technique, and makes use of semi-parametric regression model (SPRM) to approximate the optimal controller and the optimal cost function of a control dynamic system. The convergence to the optimal controller is proven, and the stability of the closed-loop nonlinear system is also guaranteed. The feasibility of the proposed learning algorithm is demonstrated in simulation on two example systems.**

***Index Terms*—policy iteration, optimal control, actor-critic structure, SPRM, nonlinear systems.**

I. Introduction

During the past several years, the optimal methods have been the mainstay design technique for feedback control systems and been developed widely to deal with complex design problems in aerospace control, process control, vehicles, communications system, robotics and numerous other applications [1]. A core challenge of obtaining the solution of nonlinear optimal control problems is that it often falls to solve the Hamilton-Jacobi-Bellman (HJB) equation which is required the full knowledge of the system dynamics.

However, the HJB equation is intractable or impossible to solve analytically for nonlinear systems. In addition, from the perspective of real-world control applications, it is desired to design online optimal controller under the absence of the knowledge of dynamics, especially for the system without the knowledge of internal dynamics.

To address the above issues, reinforcement learning (RL) is introduced. RL provides online direct adaptive schemes that converge to optimal control solutions for unknown systems. A computational intelligent learning technique known as policy iteration (PI) based on RL [2] has been widely used to approximate an optimal controller and optimal cost function for both linear [3]-[8] and nonlinear [9]-[12] partially unknown or completely unknown systems. The PI technique includes two-step iteration: policy evaluation and policy improvement, which starts with a given initial admissible control policy, and evaluates the performance of current policy and then obtains improved policy sequentially or simultaneously until the policy improvement no longer changes. During the development of adaptive optimal control based on PI algorithm, the offline iteration algorithm and the online update algorithm have been presented to approximate optimal control policy [13]-[17]. Naturally, the value function approximation (VFA) that is essential to implement the PI algorithm has played a vital role in an actor-critic structure [13]. Obviously, the approximate accuracy will impact directly the performance of system and the convergence of the optimal controller, which depends on the choice of the function approximator.

To the best of our knowledge, however, almost all of the existed works have chosen BP neural network (NN) or RBF neural network as the functional approximator to implement the PI algorithm. Nevertheless, the neural network approximator belongs to parametric function approximator, which has some drawbacks such as over learning, local minima and poor convergence problems. The learning result is relative to an initial value and it is difficult to converge to a unique optimal policy [18]. In addition, as a typical non-parametric kernel method, support vector machine (SVM), which is based on Vapnik's structural risk minimization (SRM) [19], has perfect generalization property and can be able to overcome the existing weakness in parametric function approximator. But it will decrease the interpretative capability of the model, when the large amount of information is provided by experience. Therefore, in this paper, we presented an approach to approximate value function, named semi-parametric regression model (SPRM) that elegantly combined the advantages of parametric with non-parametric approaches. Compared with pure parametric or non-parametric models, the SPRM has better adaptability and stronger interpretative capability [20].

Based on the above analysis, a novel adaptive optimal control by using PI algorithm based on SPRM is proposed and the SPRM is used to online approximate value function during the process of learning. Compared with the method presented in [18], the novel approach in this paper regarded the SPRM as a function approximator to get the value function, thus accomplishing the adaptive optimal control based on the idea of IRL to improve the generalization ability and the robustness, which is a promising and practical method to overcome the problems of NN approximator.

Manuscript received December 30, 2015; revised June 2, 2016. This work was supported by National Natural Science Foundation of China under Grant 61473147.

The authors are with the Automation Department, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016 China (e-mail: sunjingliangac@163.com. liuchsh@nuaa.edu.cn, liunian@nuaa.edu.cn).

II. Preliminaries of Optimal Control Problems

Consider a nonlinear time-invariant affine in the input dynamical system given by

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t); \quad x(0) = x_0 \tag{1}$$

where state $x(t) \in R^n$, internal dynamics $f(x(t)) \in R^n$, input coupling matrix $g(x(t)) \in R^{n\times m}$ and control input $u(t) \in R^m$. Assume that $f(0) = 0$ and $f(x) + g(x)u$ is Lipschitz continuous on a compact set Ω that contains the origin, and that the system is stabilizable on Ω. That is, there exists a control policy $u(t)$ such that the given system is asymptotically stable on Ω.

For optimal control problem, the control objective is to design an optimal control law for the system (1) that ensures all the signals involved in the closed-loop system are uniformly ultimately bounded (UUB), while minimizing the infinite horizon performance cost function:

$$V(x_0) = \int_0^\infty r(x(\tau), u(\tau))d\tau; \quad V(0) = 0 \tag{2}$$

where $r(x,u) = Q(x) + u^T Ru$ with $Q(x)$ positive definite, i.e. $\forall x \neq 0, Q(x) \neq 0$ and $x = 0 \Rightarrow Q(x) = 0$ and the weighted matrix $R = R^T > 0 \in R^{m\times m}$.

The optimal control problem can be formulated: given the continuous-time system (1), a admissible control set U and an infinite horizon cost function (2), then find a control policy $u(t)$ that minimizes the cost function [9].

Now we define the Hamiltonian function as

$$H(x,u,\nabla V_x) = r(x(t),u(t)) + \nabla V_x^T (f(x) + g(x)u(t)) \tag{3}$$

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

$$\min_{u\in U}[H(x,u,\nabla V_x^*)] = 0 \tag{4}$$

where ∇V_x denotes the gradient of the cost function $V(x)$ with respect to x, which is a column vector.

A necessary condition for optimization of $H(x,u,\nabla V_x)$ with respect to $u(t)$ is that

$$\frac{\partial H}{\partial u} = 0 \Rightarrow 2Ru(t) + g^T(x)\nabla V_x^* = 0 \tag{5}$$

Which results in

$$u^*(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla V_x^* \tag{6}$$

Substituting (6) into Eq. (3), and the HJB equation in terms of ∇V_x^* is obtained:

$$\begin{aligned} 0 &= Q(x) + (\nabla V_x^*)^T f(x) - \frac{1}{4}(\nabla V_x^*)^T g(x)R^{-1}g^T(x)\nabla V_x^* \\ V^*(0) &= 0 \end{aligned} \tag{7}$$

Remark 1: The HJB equation (7) relates to solving the partial differential equations, which is extremely difficult to solve. From the perspective of real-time applications, the requirement of the full knowledge of the control system dynamics is intractable to satisfy.

III. Semi-parametric Regression Model (SPRM)

It is noted that the support vector machine (SVM) learning formulation is based on the principle of structural risk minimization (SRM), which is a non-parametric regression model. Unlike parametric regression that minimizes an objective function based on training data, SVM attempts to minimize a bound on the generalization error, which overcomes some disadvantages of parametric regression such as over learning, local minima and poor convergence problems. Hence, we combine a ε-SVM with parametric regression model (NN model) so as to construct a SPRM, and further use it to approximate the value function in optimal control. Here, we will present the principle of SPRM that is a key step to accomplish our control objective.

The value function $V(x)$ can be formulated as

$$V(x) = \langle w, \phi(x)\rangle + W_l' \varphi_l(x) \tag{8}$$

where x is the input data of the SPRM, $\phi(x) \in R^n$ is a nonlinear mapping from the input space to a high-dimension feature space, w and W_l' are weight vector of the non-parametric and parametric model, respectively, $\varphi_l(x) \in R^n$ is a basis function vector, $\langle\cdot,\cdot\rangle$ denotes inner products.

The optimization problem can be formulated in the primal space [21]. As the following

$$\text{minimize} \frac{1}{2}\|w\| + \frac{C}{L}\sum_{i=t-L}^{t-1}(\xi_i + \xi_i^*) \tag{9}$$

$$\text{subject to} \begin{cases} \langle w, \phi(x_i)\rangle + W_l'\varphi_l(x_i) - V(x_i) \le \varepsilon + \xi_i^* \\ V(x_i) - \langle w, \phi(x_i)\rangle - W_l'\varphi_l(x_i) \le \varepsilon + \xi_i \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$

in which ε is the maximum value of tolerable error, ξ_i and ξ_i^* are slack variables, $\|\cdot\|$ is the Euclidean norm, and C is a punishment factor that denotes the trade-off between the model complexity and the tolerance to the error large than ε.

Construct the Lagrange function

$$\begin{aligned} L(w,\xi,\xi^*,\alpha,\alpha^*,\gamma,\gamma^*) &= \frac{1}{2}\|w\|^2 + \frac{C}{L}\sum_{i=t-L}^{t-1}(\xi_i + \xi_i^*) \\ &- \sum_{i=t-L}^{t-1}\alpha_i^*\left[\xi_i^* + \varepsilon + V_i(x_i) - \langle w, \phi(x)\rangle - b\right] \\ &- \sum_{i=t-L}^{t-1}\alpha_i\left[\xi_i + \varepsilon - V_i(x_i) + \langle w, \phi(x)\rangle + b\right] \\ &- \sum_{i=t-L}^{t-1}\xi_i^*\gamma_i^* - \sum_{i=t-L}^{t-1}\xi_i\gamma_i \end{aligned} \tag{10}$$

Take the partial derivative of the Lagrange function with respect to w, b, ξ_i^* and ξ_i, respectively. We get the following equations:

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=t-L}^{t-1}(\alpha_i - \alpha_i^*)\phi(x_i) \\ \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=t-L}^{t-1}(\alpha_i^* - \alpha_i) = 0 \\ \frac{\partial L}{\partial \xi_i^*} = 0 \Rightarrow \frac{C}{L} - \alpha_i^* - \gamma_i^* = 0 \\ \frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \frac{C}{L} - \alpha_i - \gamma_i = 0 \end{cases} \tag{11}$$

Thus, the dual form of (9) can be rewrite a quadratic programming (QP) problem as follows [18]

$$\begin{aligned}\text{minimize } & \frac{1}{2}\sum_{i,j=t-L}^{t-1}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i,x_j) \\ & +\varepsilon\sum_{i=t-L}^{t-1}(\alpha_i+\alpha_i^*)-\sum_{i=t-L}^{t-1}V_i(x_i)(\alpha_i-\alpha_i^*)\end{aligned} \tag{12}$$

$$\text{subject to } \begin{cases}\sum_{i=t-L}^{t-1}(\alpha_i-\alpha_i^*)\varphi_l(x_i)=0 \quad \text{for all} 1\le l\le n \\ \alpha_i,\alpha_i^*\in[0,C/L]\end{cases}$$

where $K(x_i,x_j)$ is a kernel function given by

$$K(x_i,x_j)=\langle\phi(x_i),\phi(x_j)\rangle=\phi(x_i)^T\phi(x_j).$$

Solving the above dual problem, the regression model of value function $V(x)$ can be obtained

$$V(x)=\sum_{i=t-L}^{t-1}(\alpha_i-\alpha_i^*)K(x_i,x)+W_l'\varphi_l(x) \tag{13}$$

Comparing with (8), we can get

$$w=\sum_{i=t-L}^{t-1}(\alpha_i-\alpha_i^*)\phi(x_i) \tag{14}$$

By applying the Karush-Kuhn-Tucker (KKT) condition, we obtain

$$\begin{bmatrix}\vec{\varphi}(x_1)\\ \vec{\varphi}(x_2)\\ \cdots\\ \vec{\varphi}(x_n)\end{bmatrix}\begin{bmatrix}W_1' & & & \\ & W_2' & & \\ & & \cdots & \\ & & & W_n'\end{bmatrix}=\begin{bmatrix}V_1'(x_1)\\ V_2'(x_2)\\ \cdots\\ V_n'(x_n)\end{bmatrix} \tag{15}$$

where

$W_l^T=[W_1',W_2',\cdots,W_n']^T$, $\vec{\varphi}(x_l)=[\vec{\varphi}_1(x_l),\vec{\varphi}_2(x_l),\cdots,\vec{\varphi}_n(x_l)]^T$ and $V_l'(x_l)=V(x)-\sum_{i=t-L}^{t-1}(\alpha_i-\alpha_i^*)K(x_i,x_l)$. Hence, the weight vector W_l' can be obtained by solving the linear equation (15) easily.

IV. Policy Iteration Algorithm for Solving the HJB Equation

In this part, a novel online iteration algorithm, named policy iteration (PI), was adopted to solve infinite horizon optimal control problem by solving the HJB equation (7) without requiring the knowledge of the internal dynamics $f(x)$.

Given an admission policy $\mu^{(0)}(x(t))$ and an integration time interval $T>0$, according to the idea of the dynamic programming, the value function (2) can be revised as the following form:

$$V^{\mu}(x(t))=\int_t^{t+T}r(x(\tau),\mu(x(\tau)))d\tau+V^{\mu}(x(t+T)) \tag{16}$$

in which the integrand

$$\rho(x(t),t,t+T)=\int_t^{t+T}r(x(\tau),\mu(x(\tau)))d\tau$$

is known as the integral reinforcement (IRL) learning form on the time interval $[t,t+T]$.

It should be noted that the formula (7) named as the IRL Bellman equation does not contain system dynamics $(f(\cdot),g(\cdot))$, which makes the PI algorithm possible without giving $f(x)$.

Let $\mu^{(0)}(x(t))\in\Psi(\Omega)$ be an admission policy, then there exists a time interval $T>0$, such that, if $x(t)\in\Omega$, then also $x(t+T)\in\Omega$. Thus the PI algorithm can be implemented between the following two steps: policy evaluation and policy improvement.

1) (Policy evaluation step) Solve for the value function $V^{\mu^{(i)}}(x(t))$ using the equation (7)

$$\begin{aligned}V^{\mu^{(i)}}(x(t))=&\int_t^{t+T}r(x(\tau),\mu^{(i)}(x(\tau)))d\tau \\ &+V^{\mu^{(i)}}(x(t+T)),\ V^{\mu^{(i)}}(0)=0\end{aligned} \tag{17}$$

2) (Policy improvement step) Update the control policy using

$$\mu^{(i+1)}(x)=-\frac{1}{2}R^{-1}g^T(x)\nabla V_x^{\mu^{(i)}} \tag{18}$$

The proposed PI algorithm is derived from the success of the online adaptive critic techniques proposed by computational intelligence researchers [22], which is a contraction map guaranteeing the convergence of the PI algorithm.

It's easy to prove the solution $V^{\mu^{(i)}}$ of the formula (17) is equivalent to solving the solution of

$$\begin{aligned}0=&r(x,\mu^{(i)}(x))+(\nabla V_x^{\mu^{(i)}})^T(f(x)+g(x)\mu^{(i)}(x)) \\ &V^{\mu^{(i)}}(0)=0\end{aligned} \tag{19}$$

by integrating (19) over the time interval $[t,t+T]$ (see [11] and references therein).

Remark 2: Note that although the formulas (17) and (19) have the same solution for optimal control problems, the formula (17) does not contain the knowledge of the internal dynamics $f(x)$, which needs to be known explicitly in (19). Thus, the PI algorithm is necessary for us to deal with the partially unknown systems presented in this paper.

Lemma 1: The PI algorithm proposed converges uniformly to the optimal control solution on trajectories originating in Ω, that is

$$\begin{aligned}&\forall\varepsilon>0,\exists i_0:\forall i\ge i_0 \\ &\sup_{x\in\Omega}\left|V^{\mu^{(i)}}(x)-V^*(x)\right|<\varepsilon \\ &\sup_{x\in\Omega}\left|\mu^{\mu^{(i)}}(x)-u^*(x)\right|<\varepsilon\end{aligned} \tag{20}$$

Proof: the proof is presented in the paper [23] for details.

Remark 3: Lemma 1 presents the fact that the optimal control based on PI algorithm guarantees the control policy converges uniformly to the optimal control solution under an initial stabilizing controller.

V. The Convergence and Implementation of PI Algorithm by Using SPRM Approximation

This section combines the SPRM with PI algorithm in an

actor-critic structure to implement the proposed algorithm and to improve the accuracy of approximation, and the convergence of the proposed algorithm is presented.

The value function $V(x)$ always can be described as follows:

$$V(x)=\langle w,\phi(x)\rangle+W_l'\varphi(x)=W^T\sigma(x) \tag{21}$$

where $W=[w\ W_l'^{\,T}]^T$ denotes the augmented regularization weighted vector of SPRM and $\sigma(x)=[\phi(x)\ \ \varphi(x)]$ is the vector of augmented function.

Considering the approximate error of the SPRM, the output of the SPRM can be noted

$$\hat{V}(x)=\langle w,\phi(x)\rangle+W_l'\varphi(x)=\hat{W}^T\sigma(x) \tag{22}$$

Using the SPRM description for the cost function, equation (17) can be written as

$$\begin{aligned}\hat{W}^{\mu^{(i)}T}\sigma(x(t))=&\int_t^{t+T}r(x(\tau),\mu^{(i)}(x(\tau)))d\tau\\&+\hat{W}^{\mu^{(i)}T}\sigma(x(t+T))\end{aligned} \tag{23}$$

The policy improvement step (18) then can be written as

$$\hat{\mu}^{(i+1)}(x)=-\frac{1}{2}R^{-1}g^T(x)\nabla\sigma^T(x)\hat{W}^{\mu^{(i)}} \tag{24}$$

Define the residual error produced by the SPRM approximation as ε_{HJB}. Then, applying the formula (10) and (11), the residual error can be obtained:

$$\hat{W}^T\nabla\sigma(x)f(x)-\frac{1}{4}\hat{W}^TD(x)W+Q(x)=\varepsilon_{HJB} \tag{25}$$

with $D(x)=\nabla\sigma(x)g(x)R^{-1}g^T(x)\nabla\sigma^T(x)$.

Remark 4: The objective of the SPRM is to drive the residual error to converge to zero. Thus, the parameters $W^{\mu^{(i)}}$ of the cost functional approximation $V^{\mu^{(i)}}$ converge to the optimal weights and the control policy converges to the optimal control policy.

Now, we will discuss the convergence of the proposed PI algorithm and the optimal solution by using SPRM to approximate the exact cost function. The next notion of practical stability is needed.

Definition 1:[9] (UUB) A time signal $\zeta(t)$ is said to the uniformly ultimately bounded (UUB) if there exists a compact set $S\subset R^n$ so that for all $\zeta(0)\in S$ there exists a bound B and a time $T(B,\zeta(0))$ such that $\|\zeta(t)\|\le B$ for all $t\ge t_0+T$.

Theorem 1: Based on the convergence of the PI algorithm, then, the Hamiltonian function $H(\hat{u},\hat{W},x)=\int_{t-T}^{t}(Q(x)+\hat{u}^TR\hat{u}-\varepsilon_{HJB})d\tau+\hat{W}^T\sigma(x(t))-\hat{W}^T\sigma(x(t-T))$ is UUB with the control policy $\hat{u}=-\frac{1}{2}R^{-1}g^T(x)\nabla\sigma^T(x)\hat{W}$. That is to say, $\hat{W}^T\sigma(x(t))$ converges to the approximate HJB solution and $\hat{u}$ converges to the optimal solution.

Proof: According to the principle of the SPRM, we know that the weights $\tilde{W}=W-\hat{W}$ is UUB, thus

$$\begin{aligned}H(\hat{u},\hat{W},x)=&\int_{t-T}^{t}(Q(x)+\frac{1}{4}\hat{W}^TD(x)\hat{W}-\varepsilon_{HJB})d\tau\\&+\hat{W}^T\sigma(x(t))-\hat{W}^T\sigma(x(t-T))\end{aligned} \tag{26}$$

Substituting the weight estimation error $\tilde{W}=W-\hat{W}$ into the $H(\hat{u},\hat{W},x)$, we get

$$\begin{aligned}&H(\hat{u},\hat{W},x)=\\&\int_{t-T}^{t}(Q(x)+\frac{1}{4}(W-\tilde{W})^TD(x)(W-\tilde{W})-\varepsilon_{HJB})d\tau\\&+(W-\tilde{W})^T\sigma(x(t))-(W-\tilde{W})^T\sigma(x(t-T))\\&=\int_{t-T}^{t}\Big(Q(x)+\frac{1}{4}W^TD(x)W-\frac{1}{2}\tilde{W}^TD(x)\\&+\frac{1}{4}\tilde{W}^TD(x)\tilde{W}-\varepsilon_{HJB}\Big)d\tau+W^T\sigma(x(t))\\&-\tilde{W}^T\sigma(x(t))-W^T\sigma(x(t-T))+\tilde{W}^T\sigma(x(t-T))\\&=\int_{t-T}^{t}(-\frac{1}{2}\tilde{W}^TD(x)W+\frac{1}{4}\tilde{W}^TD(x)\tilde{W}-\varepsilon_{HJB})d\tau\\&-\tilde{W}^T\sigma(x(t))+\tilde{W}^T\sigma(x(t-T))\\&+\int_{t-T}^{t}(Q(x)+\frac{1}{4}W^TD(x)W)d\tau\\&+W^T\sigma(x(t))-W^T\sigma(x(t-T))\end{aligned} \tag{27}$$

Since the term

$$\begin{aligned}&\int_{t-T}^{t}(Q(x)+\frac{1}{4}W^TD(x)W)d\tau\\&+W^T\sigma(x(t))-W^T\sigma(x(t-T))=0\end{aligned} \tag{28}$$

Then

$$\begin{aligned}&H(\hat{u},\hat{W},x)=\\&\int_{t-T}^{t}(-\frac{1}{2}\tilde{W}^TD(x)W+\frac{1}{4}\tilde{W}^TD(x)\tilde{W}-\varepsilon_{HJB})d\tau\\&-\tilde{W}^T\sigma(x(t))+\tilde{W}^T\sigma(x(t-T))\end{aligned} \tag{29}$$

By taking norms on both sides and taking into account the $\sup_{x\in\Omega}\|\varepsilon_{HJB}\|<\varepsilon$ and letting

$$\begin{aligned}&\|H(\hat{u},\hat{W},x)\|\le\\&\int_{t-T}^{t}\Big(\frac{1}{4}\|\tilde{W}\|^2\|D(x)\|+\frac{1}{2}\|\tilde{W}\|^2W_{\max}\|D(x)\|+\varepsilon\Big)\\&+\|\tilde{W}\|\|\sigma(x(t))\|+\|\tilde{W}\|\|\sigma(x(t-T))\|\end{aligned} \tag{30}$$

All the signals on the right-hand side of (30) are UUB. So $H(\hat{u},\hat{W},x)$ is UUB and the convergence to the approximate HJB solution is obtained. Because of the convergence of PI algorithm and equation (24), we get $\hat{u}$ converges to the optimal solution.

Next, we will introduce the structure of the whole algorithm bridged with the PI algorithm and the machine learning, which denotes the basic idea of this paper.

The structure of the IRL adaptive optimal controller is presented in Fig. 1. The PI technique in this section has been implemented through applying the proposed SPRM to converging in real time to an optimal control solution by measuring data along the system trajectories and without knowing the internal dynamics of the system.

In this structure, a series of state-action pairs (x_t,μ_t) comprised of each action μ_t computed by (24)

and the current system state x_t can be constituted. The state-action pair (x_t, μ_t) and the estimated cost function V_t are regarded as the input and the output of the SPRM, respectively. In order to reduce the number of adjustable parameters in (21) improving the learning speed, a model interval is established with fixed length and a sliding time window is introduced into the learning system. In addition, two sample sets are designed, i.e., a data buffer memory DB and a working sample set WD. Define the size of WD and DB are L and l, respectively. Then, the samples in the WD can be described as $\{(x_i, \mu_i), V_{i-1} \mid i = t-L, t-L+1, \cdots, t-1\}$.

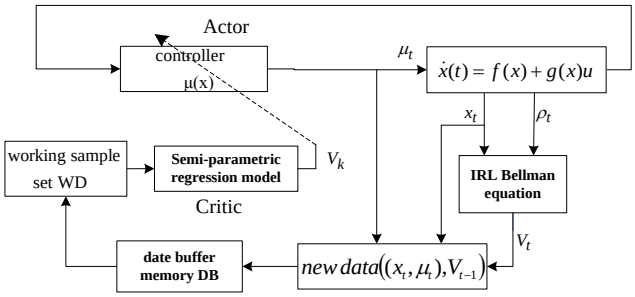


Fig. 1. The actor-critic structure based on semi-parametric model.

According to the sliding time window theory, the working sample data will be updated during the learning process. When the number of the samples in DB exceeds a predefined threshold, the samples in DB will be sent to the working sample set WD while the same amount samples in WD will be removed following the rule that first-in-first-out to keep a fixed length sample window.

Remark 5: The structure proposed is a hybrid continuous-time/discrete-time adaptive control structure, which is different from a traditional control structure. Obviously, it has a discrete-time sampled portion for policy evaluation and policy updates and continuous-time controllers.

VI. Simulations

To support the new online PI algorithm for continuous-time systems, two simulation examples are offered: one is linear system and one is nonlinear system. Both the two cases validate the proposed online adaptive optimal algorithm and guarantee the system and the cost function converge to actual optimal values.

A. Linear System Example

Consider the continuous-time F16 aircraft longitudinal plant with quadratic cost function used in [10], which has the dynamics $\dot{x} = Ax + Bu$ given by

$$\dot{x} = \begin{bmatrix} -1.01887 & 0.90506 & -0.00215 \\ 0.8225 & -1.07741 & -0.17555 \\ 0 & 0 & -1 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u \qquad (31)$$

where the system state vector is $x = [\alpha \; q \; \delta_e]^T$, where α denotes the angle of attack, q is the pitch rate and δ_e is the elevator deflection angle. The control input is the elevator actuator voltage. The cost function $J = \int_0^\infty (x^T Q x + u^T R u)\, dt$, with $Q = Q^T \geq 0$ and $R = R^T > 0$.

In this linear system, in order to validate the proposed algorithm, the design parameters are chosen as $x_0 = [0.1 \; 0 \; 0.3]^T$, $Q = diag(10, 10, 10)$, $R = 1$. The length of train data is $L = 50$ and the length of data buffer memory is $l = 0.4 * L = 20$. In the SPRM, the Gaussian kernel $\phi(x) = K(x, y) = \exp\left(\frac{-(x-y)^2}{2\sigma^2}\right)$ is chosen for SVM regression model with kernel parametric $\rho = 50$, the penalty factor $C = 8$ and tolerance error $\varepsilon = 0.0002$. While in the parametric part of the SPRM, the symmetric sigmoidal function $\varphi(x) = \frac{1-e^{-x}}{1+e^{-x}}$ is chosen to be used as basis function. The initial input train sample is randomly chosen in the subset $[0, 1]$ and the initial output train sample is set to zero. The sampling period is T=0.1s, in this way, every 2s, the cost function was solved for and a policy update was performed. The results of applying the algorithm is presented in Fig. and Fig. .

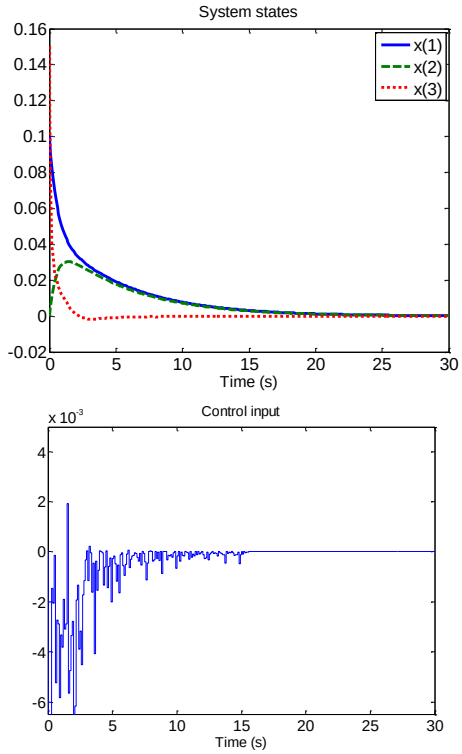


Fig. 2. The trajectories of system states and control input.

The Fig. shows that the system states converge to the steady state $x = [0 \; 0 \; 0]^T$ after about 20s by using appropriate control input. In other words, the proposed algorithm guarantees the F16 aircraft longitudinal plant to converge to zero under reasonable control input. One can see that the control input is a discrete signal because of the actor-critic structure that has a hybrid continuous-time/discrete-time adaptive optimal structure. At every 2s period, the data buffer memory collects the train data set at the sample times $t, t+T, t+2T, \cdots t+lT$. The cost function in critic part was evaluated while the control input in actor part held until a new policy updated.

Furthermore, compared with the adaptive optimal control based on parametric regression model by using Neural Network (NN) presented in [10], the adaptive optimal control

based on SPRM in this paper performs better robustness and less experiment time that was about 750s in [10].

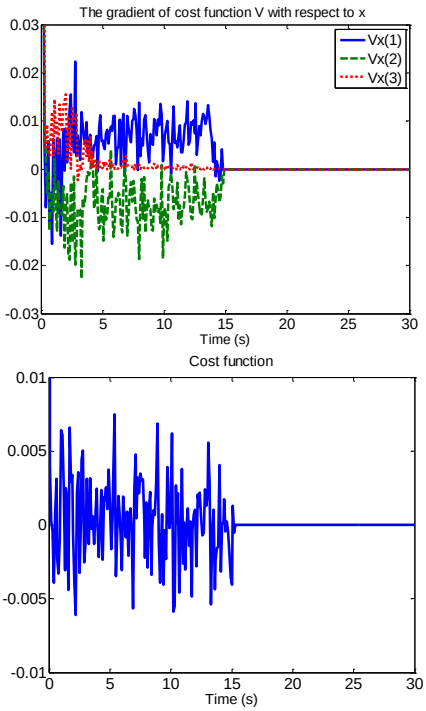


Fig. 3. Converge of the cost function.

The Fig. 3 presents the convergence of the cost function and the gradient of cost function $V(x)$ with respected to x which are continuous in critic part. After about 15s, both the cost function and the gradient of cost function $V(x)$ with respected to x no longer change. It means that the control input $\mu(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla\sigma^T(x)W$ closes to optimal value and then drives the system to converge to the steady states.

B. Nonlinear System Example

Consider the following affine in control input nonlinear system with quadratic cost function applied in [11], which has the dynamics $\dot{x} = f(x) + g(x)u, x \in R^2$ given by

$$\dot{x} = \begin{bmatrix} -x_1 + x_2 \\ -\frac{1}{2}(x_1 + x_2) + \frac{1}{2}x_2\sin^2(x_1) \end{bmatrix} + \begin{bmatrix} 0 \\ \sin(x_1) \end{bmatrix} u \qquad (32)$$

with the cost function

$$J = \int_0^\infty (Q(x) + u^T Ru)dt, Q(x) = x_1^2 + x_2^2, R = 1.$$

In this nonlinear affine system, the design parameters are chosen as follows: the initial states $x_0 = [0.1\ \ 0]^T$. The length of train data is $L = 50$ and the length of data buffer memory is $l = 0.4 * L = 20$. In the semi-parametric regression model, the Gaussian kernel $\phi(x) = K(x, y) = \exp\left(\frac{-(x-y)^2}{2\sigma^2}\right)$ is chosen for SVM regression model with kernel parametric $\rho = 0.6$, the penalty factor $C = 300$ and tolerance error $\varepsilon = 0.0002$. While in the parametric part of the SPRM, the symmetric sigmoidal function $\varphi(x) = \frac{1-e^{-x}}{1+e^{-x}}$ is chosen to be used as basis function. The initial input train sample is randomly chosen in the subset $[0, 1]$ and the initial output train sample is set to zero. The sample time is T=0.1s, in this way, every 2s, the cost function was solved for and a policy update was performed. The results of applying the algorithm is presented in Fig. 4 and Fig. 5.

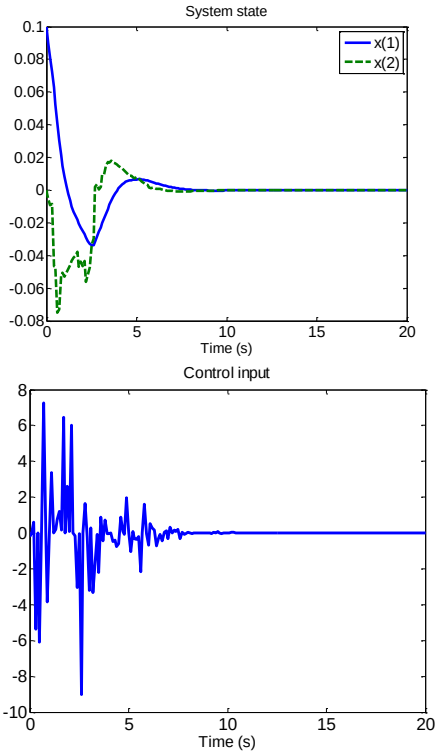


Fig. 4. The trajectories of system states and control input.

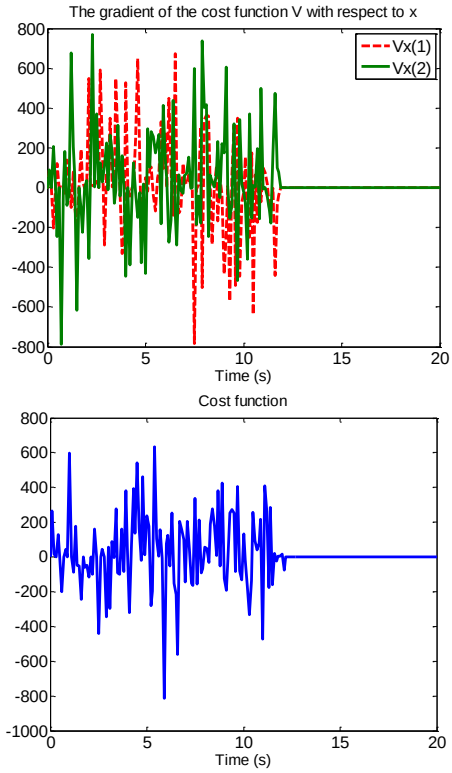


Fig. 5. Converge of the cost function.

The Fig. 4 and Fig. 5 point out that the adaptive optimal algorithm based on SPRM proposed in this paper performs desired virtues and promising potential. One can see that the system states converge to the steady state quickly under admissible control input and the cost function and the gradient

of cost function with respect to x converge to zero at about 12s, the input $\mu(x) = -\frac{1}{2}R^{-1}g^T(x)\nabla\sigma^T(x)W$ closes to optimal value since the objective of our controller is stabilizing control rather than tracking control which is the same as linear system example.

The two examples validate the proposed adaptive optimal algorithm based on SPRM in which the internal dynamics $f(x)$ is not required. It shows that the SPRM exerts its advantages such as simplicity, generalization performance and robustness compared with parametric model such as NN regression model. Besides, in order to guarantee the initial output finite in SPRM, additional small noise signal is added to guarantee the difference of the output of the parametric model part, thus resulting the fluctuating of the cost function and the gradient of cost function with respect to x while it is no needed when the system converges to the optimal values.

VII. Conclusion

In this paper, a novel online continuous-time optimal control PI algorithm without the knowledge of internal dynamics of the nonlinear system has been proposed and the HJB equation is solved by using SPRM composed by SVM model and NN model which is an intractable problem in traditional optimal control. A data buffer memory is introduced to improve the learning speed and lessen calculative burden. The proof of convergence for the online adaptive optimal algorithm based on SPRM is presented. The two examples verify the effectiveness of the proposed adaptive optimal algorithm, which is similar to the persistent excitation (PE) condition.

References

[1] F. L. Lewis and K. G. Vamvoudakis, "Optimal adaptive control for unknown systems using output feedback by reinforcement learning methods," in *Proc. 8th IEEE International Conference on Control and Automation*, 2010.

[2] A. G. Barto, *Reinforcement Learning: An Introduction*, MIT press, 1998.

[3] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," *Automatica*, vol. 45, no. 2, pp. 477-484, 2009.

[4] B. Kiumarsi, F. L. Lewis *et al*., "Reinforcement -learning for optimal tracking control of linear discrete-time systems with unknown dynamics," *Automatica*, vol. 50, no. 4, pp. 1167-1175, 2014.

[5] M. Palanisamy, H. Modares, F. L. Lewis, and M. Aurangzeb, "Continuous-time q-learning for infinite-horizon discounted cost linear quadratic regulator problems," *IEEE Transactions on Cybernetics*, vol. 45, no. 2, pp. 165-176, 2015.

[6] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems," *Automatica*, vol. 48, no. 1, pp. 2850-2859, 2012.

[7] 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*, 2014.

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

[9] K. G. Vamvoudakis, D. Vrabie, and F. L. Lewis,, "Online adaptive learning of optimal control solutions using integral reinforcement learning," in *Proc. 2011 IEEE Symposium on Adaptive Dynamic Programming And Reinforcement Learning*, 2011.

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

[12] D. Liu, X. Yang, and H. Li, "Adaptive optimal control for a class of continuous-time affine nonlinear systems with unknown internal dynamics," *Neural Computing and Applications*, vol. 23, pp. 1843-1850, 2012.

[13] H.-G. Zhang, X. Zhang, Y.-H. Luo, and J. Yang, "An overview of research on adaptive dynamic programming," *Acta Automatica Sinica*, vol. 39, no. 4, pp. 303-311, 2013.

[14] D.-R. Liu, H.-L. Li, and D. Wang, "Data-based self-learning optimal control: Research progress and prospects," *Acta Automatica Sinica*, vol. 39, no. 11, p. 1858, 2013.

[15] S. G. Khan, G. Herrmann, F. L. Lewis, T. Pipe, and C. Melhuish, "Reinforcement learning and optimal adaptive control: An Overview and implementation examples," *Annual Reviews in Control*, vol. 36, no. 1, pp. 42-59, 2012.

[16] D. Kleinman, "On an iterative technique for riccati equation computations," *IEEE Transactions on Automatic Control*, vol. 13, no. 1, pp. 114-115, 1968.

[17] S. Bhasin, R. Kamalapurkar, M. Johnson, K. G. Vamvoudakis, F. L. Lewis, and W. E. Dixon, "A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82-92, 2013.

[18] Y. H. Cheng., X. S. Wang, and X. L. Tian, "Reinforcement learning method based on semi-parametric regression model," in *Proc. Control and Decision Conference*, 2010.

[19] H. Shah and M. Gopal, "Reinforcement learning with kernel recursive least-squares support vector machine," *International Journal of Machine Learning and Computing*, vol. 2, no. 5, pp. 618-622, 2012.

[20] E. Castillo, A. S. Hadi, B. Lacruz, and R. E. Pruneda, "Semi-parametric nonlinear regression and transformation using functional networks," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2129-2157, 2008.

[21] J. Shin, H. J. Kim, and Y. Kim, "Adaptive support vector regression for Uav flight control," *Neural Netw*, vol. 24, no. 1, pp. 109-120, 2011.

[22] J. J. Murray, S. Member *et al*., "Adaptive dynamic programming," *IEEE Trans. Syst. Man Cyber.*, 2002.

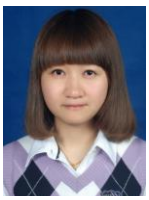
[23] F. L. Lewis *et al*., "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach ☆," *Automatica*, vol. 41, no. 5, pp. 779-791, 2005.

Jingliang Sun was born in 1990. He received his B.S. degree in automation from Tianjin University of Technology, Tianjin, China. He is currently working toward the Ph.D. degree in control theory and control engineering in College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

His current research interests include optimal control, adaptive dynamic programming, reinforcement learning and differential games.

Chunsheng Liu was born in 1955. She received her B.S., M.E. and Ph.D. degrees from Huazhong University of Science & Technology, Xi"an Jiaotong University and Nanjing University of Aeronautics and Astronautics (NUAA), respectively. She is now a professor and Ph.D. supervisor in automation engineering in NUAA.

Her current research interests include adaptive control, fault diagnosis and tolerant control with the application in aircraft.



Nian Liu was born in 1991. She received her B.S. degree in automation from Nanjing University of Aeronautics and Astronautics, Nanjing, China. She is currently working toward the M.S. degree in control theory and control engineering in College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

Her current research interests include adaptive dynamic programming and differential games.