

Neural-network-based robust optimal control of uncertain nonlinear systems using model-free policy iteration algorithm

Chao Li, Ding Wang, and Derong Liu

Abstract— In this paper, we establish a robust optimal control law for a class of continuous-time uncertain nonlinear systems by using a neural-network-based model-free policy iteration approach. The robust control law of the original uncertain nonlinear system is derived by adding a feedback gain to the optimal control law of the nominal system. It is proven that this robust control law can achieve optimality under a specified cost function. Then, the neural-network-based model-free policy iteration algorithm is developed to solve the Hamilton-Jacobi-Bellman equation corresponding to the nominal system without system dynamics. The actor-critic technique and the least squares implementation method are used to obtain the optimal control policy of the nominal system. A numerical simulation is given to verify the applicability of the present robust optimal control scheme.

I. INTRODUCTION

Many practical control systems such as transportation systems, chemical reactors, electrical networks, and power systems suffer from model uncertainties. This may degrade the closed-loop system performance severely. Hence, the problem of designing robust control policy for uncertain nonlinear systems has drawn considerable attention in recent literature [1], [2], [3], [4]. Lin et al. [3] showed that the robust control problem can be solved by studying the optimal control problem of the corresponding nominal system. Wang et al. [4] developed an iterative learning algorithm for designing the robust control policy of a class of uncertain nonlinear systems. Wang et al. [5] established an online policy iteration approach to design the robust control law for a class of nonlinear systems with uncertainties. However, the optimality of the robust control policy is not discussed and the complete system dynamics is difficult to obtain, which motivate our research. The novelty of this paper is that we establish the robust optimal control policy for uncertain nonlinear systems using model-free policy iteration algorithm with completely unknown system dynamics.

The starting point of the strategy of this paper is the nonlinear optimal control problem. The optimal control law can be obtained by solving the Hamilton-Jacobi-Bellman (HJB)

This work was supported in part by the National Natural Science Foundation of China under Grants 61233001, 61273140, 61304086, 61533017, and U1501251, in part by Beijing Natural Science Foundation under Grant 4132078, and in part by the Early Career Development Award of SKLMCCS.

C. Li and D. Wang are with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (lichao2012@ia.ac.cn; ding.wang@ia.ac.cn). D. Liu is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (derong@ustb.edu.cn).

equation which is a partial differential equation. Due to the “curse of dimensionality” [6], dynamic programming which provides a conventional method in solving optimization and optimal control problems is often computationally untenable. To avoid this difficulty, based on function approximators, such as neural networks (NNs), adaptive dynamic programming (ADP) was proposed by Werbos [7] and Prokhorov and Wunsch [8] as a method to solve the optimal control problem. Reinforcement learning (RL) is an effective computational method and it can find the optimal policy interactively. Lewis and Vrabie [9], and Lewis and Liu [10] stated that the ADP technique is closely related to the field of RL. Recently, the researches on ADP and RL have gained much attention from various scholars [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21].

In the existing literature of of ADP-based and RL-based optimal control, either policy iteration (PI) or value iteration is utilized to solve the HJB equation. The system dynamics are necessary when employing the traditional PI algorithms. However, in many situations, it is difficult to obtain the exact model of control plant. The ADP and RL schemes, which have the learning and optimal capabilities, can relax the need for a complete and accurate model by using compact parameterized function representations whose parameters are adjusted through adaption. Jiang and Jiang [22] presented a novel PI approach for continuous-time linear systems with completely unknown dynamics. Vrabie and Lewis [23] derived an integral RL method to obtain direct adaptive optimal control for nonlinear input-affine continuous-time systems with partially unknown dynamics. Lee et al. [24], [25] presented an integral RL algorithm for continuous-time systems without the exact knowledge of the system dynamics. Luo et al. [26] addressed the model-free nonlinear optimal control problem based on data by using the RL technique. Bian et al. [27] proposed a novel optimal control design scheme for continuous-time nonaffine nonlinear dynamic systems with unknown dynamics by ADP.

To begin with, the problem formulation is provided. The robust control law of the original uncertain system is derived by adding a feedback gain to the optimal control law of the nominal system. It can be proved that the robust control law also achieves optimality under the definition of a specified cost function. Then, the optimal control policy of the nominal system is obtained by the neural-network-based model-free PI algorithm without system dynamics. At last, the effectiveness of the robust optimal control scheme established in this paper is demonstrated by a simulation example.

II. PROBLEM FORMULATION

Consider the continuous-time uncertain nonlinear system

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

where $x(t) \in \mathbb{R}^n$ is the system state, $\bar{u}(x) \in \mathbb{R}^m$ is the control input, $f(\cdot)$ and $g(\cdot)$ are differentiable in their arguments with appropriate dimensionalities, $f(0) = 0$, and $\bar{d}(x) \in \mathbb{R}^m$ is the uncertain nonlinear perturbation. Let $x(0) = x_0$ be the initial state. We assume that $\bar{d}(0) = 0$, so that $x = 0$ is an equilibrium of the system (1). For the nominal system

$$\dot{x}(t) = f(x) + g(x)u(x), \quad (2)$$

we assume that $f(x) + g(x)u(x)$ is Lipschitz continuous on a set $\Omega \subset \mathbb{R}^n$ containing the origin, and there exists a continuous control policy that asymptotically stabilizes the nominal system (2).

For the original uncertain system (1), in order to deal with the robust control problem, we should find a feedback control policy $\bar{u}(x)$, such that the closed-loop system is globally asymptotically stable for all uncertainties $\bar{d}(x)$. This problem can be transformed into designing an optimal control policy for the corresponding nominal system (2) with an appropriate cost function. We denote $d(x) = R^{1/2}\bar{d}(x)$, where $R \in \mathbb{R}^{m \times m}$ is a symmetric positive definite matrix, and $d(x)$ is bounded by a known function, i.e., $\|d(x)\| \leq d_M(x)$ with $d_M(0) = 0$. In order to deal with the optimal control problem for the nominal system (2), we have to find the feedback control policy $u(x)$, which minimizes the cost function given by

$$\begin{aligned} J(x_0) &= \int_0^\infty \{d_M^2(x(\tau)) + u^\top(x(\tau))Ru(x(\tau))\} d\tau \\ &= \int_0^\infty r(x(\tau), u(x(\tau))) d\tau, \end{aligned} \quad (3)$$

where $r(x(t), u(x(t)))$ is the utility function.

Based on the optimal control theory [28], the designed feedback control law must not only stabilize the system on Ω , but also guarantee that the cost function $J(x_0)$ is finite. In other words, the control policy must be admissible [13]. Let $\Psi(\Omega)$ be the set of all the admissible control laws on Ω . The optimal cost function of the system (2) can be formulated as

$$J^*(x_0) = \min_{u \in \Psi(\Omega)} \int_0^\infty r(x(\tau), u(x(\tau))) d\tau \quad (4)$$

with $J^*(0) = 0$, and $J^*(x)$ satisfies the HJB equation

$$0 = \min_{u \in \Psi(\Omega)} H(x, u, \nabla J^*(x)), \quad (5)$$

where the term $\nabla J(x) = \partial J(x)/\partial x$ denotes the partial derivative of the cost function with respect to the state, and the Hamiltonian function is defined as

$$\begin{aligned} H(x, u, \nabla J(x)) &= d_M^2(x) + u^\top(x)Ru(x) \\ &\quad + (\nabla J(x))^\top(f(x) + g(x)u(x)). \end{aligned} \quad (6)$$

Assume that the minimum on the right hand side of (5) exists and is unique. Then, the optimal feedback control policy can be obtained as

$$\begin{aligned} u^*(x) &= \arg \min_{u \in \Psi(\Omega)} H(x, u, \nabla J^*(x)) \\ &= -\frac{1}{2}R^{-1}g^\top(x)\nabla J^*(x). \end{aligned} \quad (7)$$

Based on (6) and (7), the HJB equation (5) becomes

$$\begin{aligned} 0 &= d_M^2(x) + (\nabla J^*(x))^\top f(x) \\ &\quad - \frac{1}{4}(\nabla J^*(x))^\top g(x)R^{-1}g^\top(x)\nabla J^*(x) \end{aligned} \quad (8)$$

with $J^*(0) = 0$.

III. COMPUTATIONAL ROBUST CONTROLLER DESIGN USING MODEL-FREE PI ALGORITHM

In this section, we present the robust optimal controller design. By adding a feedback gain to the optimal control law of the nominal system, the robust optimal control law is established. Then we investigate a model-free PI algorithm to solve the optimal control problem for the nominal system (2) with completely unknown system dynamics. To implement this algorithm, the NN-based actor-critic technique and the least squares method are used.

A. Robust Optimal Control of Uncertain Nonlinear Systems

To establish the robust optimal control policy for the original system (1), we modify the optimal control policy (7) by proportionally adding a feedback gain,

$$\bar{u}(x) = \pi u^*(x) = -\frac{1}{2}\pi R^{-1}g^\top(x)\nabla J^*(x). \quad (9)$$

Lemma 1: For the nominal system (2), the feedback control policy given by (9) ensures that the closed-loop system is asymptotically stable for all $\pi \geq 1/2$.

Proof. We show that the optimal cost function $J^*(x)$ is a Lyapunov function. In light of (4), we can easily find that $J^*(x)$ is positive definite for $x \neq 0$. Considering (8) and (9), the derivative of $J^*(x)$ along the trajectory of the closed-loop system is

$$\begin{aligned} \dot{J}^*(x) &= (\nabla J^*(x))^\top(f(x) + g(x)\bar{u}(x)) \\ &= -d_M^2(x) - \frac{1}{2}\left(\pi - \frac{1}{2}\right)\|R^{-1/2}g^\top(x)\nabla J^*(x)\|^2. \end{aligned}$$

Hence, $\dot{J}^*(x) < 0$ for all $\pi \geq 1/2$ and $x \neq 0$. The conditions for the Lyapunov local stability theory are satisfied and the closed-loop system is asymptotically stable. \square

Theorem 1: For the original system (1), there exists a positive number $\pi_1 \geq 1$, such that for any feedback gain $\pi > \pi_1$, the robust control law developed by (9) ensures that the closed-loop system is asymptotically stable.

Proof. We select $V(x) = J^*(x)$ as the Lyapunov function candidate. Taking the time derivative of $V(x)$ along the trajectory of the closed-loop system (1), we obtain

$$\dot{V}(x) = (\nabla J^*(x))^\top(f(x) + g(x)(\bar{u}(x) + \bar{d}(x))).$$

Based on the results in Lemma 1 and the relationship $\|R^{1/2}\bar{d}(x)\| \leq d_M(x)$, we find that

$$\begin{aligned}\dot{V}(x) &\leq -\left\{d_M^2(x) + \frac{1}{2}\left(\pi - \frac{1}{2}\right)\|R^{-1/2}g^\top(x)\nabla J^*(x)\|^2\right. \\ &\quad \left.- \|R^{-1/2}g^\top(x)\nabla J^*(x)\|d_M(x)\right\}.\end{aligned}$$

Let $\xi = [d_M(x), \|R^{-1/2}g^\top(x)\nabla J^*(x)\|]^\top$. Then, we have $\dot{V}(x) \leq -\xi^\top \Gamma \xi$, where

$$\Gamma = \begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2}(\pi - \frac{1}{2}) \end{bmatrix}.$$

From the above equation, we observe that there exists a positive number $\pi_1 \geq 1$ such that any $\pi > \pi_1$ can guarantee the positive definiteness of Γ . Then, we have $\dot{V}(x) < 0$, which implies that the closed-loop system is asymptotically stable. \square

According to Theorem 1, $\bar{u}(x)$ is the robust stabilizing control policy of the original system (1) for any gain $\pi > \pi_1$. We will show that the robust control policy also possesses the optimality with appropriate feedback gain. For the original system (1), we define the following cost function

$$\bar{J}(x_0) = \int_0^\infty \left\{ Q(x(\tau)) + \frac{1}{\pi} \bar{u}^\top(x(\tau)) R \bar{u}(x(\tau)) \right\} d\tau, \quad (10)$$

where

$$\begin{aligned}Q(x) &= \frac{1}{4}(\pi - 1)(\nabla J^*(x))^\top g(x) R^{-1} g^\top(x) \nabla J^*(x) \\ &\quad + d_M^2(x) - (\nabla J^*(x))^\top g(x) \bar{d}(x).\end{aligned} \quad (11)$$

Lemma 2: There exists a positive number $\pi_2 \geq 2$ such that for all $\pi > \pi_2$, the function $Q(x)$ is positive definite.

Proof. Adding and subtracting $(1/(\pi-1))d^\top(x)d(x)$ to (11), we find that

$$\begin{aligned}Q(x) &= \frac{1}{4(\pi-1)} \left((\pi-1)(\nabla J^*)^\top g(x) R^{-1/2} - 2d^\top(x) \right) \\ &\quad \left((\pi-1)(\nabla J^*)^\top g(x) R^{-1/2} - 2d^\top(x) \right)^\top \\ &\quad - \frac{1}{\pi-1} d^\top(x) d(x) + d_M^2(x).\end{aligned}$$

Based on the relationship $\|d(x)\| \leq d_M(x)$ and $\pi > 2$, we can obtain

$$Q(x) \geq d_M^2(x) - \frac{1}{\pi-1} d^\top(x) d(x) \geq \frac{\pi-2}{\pi-1} d_M^2(x).$$

This proves that $Q(x)$ is a positive definite function. \square

Theorem 2: Consider the original system (1) with cost function (10). There exists a positive number π^* such that for any feedback gain $\pi > \pi^*$, the feedback control policy obtained by (9) is the robust optimal control of the original uncertain nonlinear system.

Proof. The Hamiltonian function of the system (1) with cost function (10) is

$$\begin{aligned}\bar{H}(\nabla \bar{J}(x)) &= Q(x) + \frac{1}{\pi} \bar{u}^\top(x) R \bar{u}(x) \\ &\quad + (\nabla \bar{J}(x))^\top (f(x) + g(x)(\bar{u}(x) + \bar{d}(x))),\end{aligned}$$

where $\pi > \pi_2 \geq 2$. Replacing $\bar{J}(x)$ with $J^*(x)$ and observing (11), the Hamiltonian function $\bar{H}(x)$ becomes

$$\begin{aligned}\bar{H}(x) &= d_M^2(x) + (\nabla J^*(x))^\top f(x) \\ &\quad + \frac{1}{4}(\pi-1)(\nabla J^*(x))^\top g(x) R^{-1} g^\top(x) \nabla J^*(x) \\ &\quad + \frac{1}{\pi} \bar{u}^\top(x) R \bar{u}(x) + (\nabla J^*(x))^\top g(x) \bar{u}(x).\end{aligned}$$

Using (8) and (9), we can obtain that $\bar{H}(\nabla J^*(x)) = 0$, which shows that $J^*(x)$ is a solution of the HJB equation of the system (1). Then, we say that the control law (9) achieves optimality with cost function (10). Based on Theorem 1, there exists a positive number $\pi^* \triangleq \max\{\pi_1, \pi_2\}$ such that for any $\pi > \pi^*$, the control law (9) can not only stabilize system (1), but also achieve optimality with the specified cost function. This completes the proof. \square

B. Online Model-free PI Algorithm

The formulation developed in (7) displays an array of closed-form expression, which obviates the need to search for the optimal control policy via optimization process. To obtain the optimal control policy, the existence of $J^*(x)$ satisfying (8) is the necessary and sufficient condition. Instead of directly solving (8), we can successively solve the nonlinear Lyapunov equation

$$0 = r(x, u_i(x)) + (\nabla J_i(x))^\top (f(x) + g(x)u_i(x)),$$

and update the control policy based on

$$u_{i+1}(x) = -\frac{1}{2} R^{-1} g^\top(x) \nabla J_i(x) \quad (12)$$

to obtain the solution $J^*(x)$. This successive approximation is known as the PI algorithm [29], and it is the fundamental for the model-free PI algorithm. In [30], it was shown that on the domain Ω , the cost function $J_i(x)$ uniformly converges to $J^*(x)$ with monotonicity $J_{i+1}(x) < J_i(x)$, and the control policy $u_i(x)$ is admissible and converges to $u^*(x)$.

To deal with the optimal control problem with completely unknown system dynamics, we develop an online model-free PI algorithm. We consider the following nonlinear system explored by a known bounded piecewise continuous probing signal $e(t)$

$$\dot{x}(t) = f(x(t)) + g(x(t))(u(x(t)) + e(t)). \quad (13)$$

The derivative of the cost function (3) with respect to time along the trajectory of the explored nominal system (13) can be calculated as

$$\begin{aligned}\dot{J}(x) &= (\nabla J(x))^\top (f(x) + g(x)(u(x) + e)) \\ &= -r(x, u(x)) + (\nabla J(x))^\top g(x)e.\end{aligned} \quad (14)$$

Integrating (14) from t to $t+T$ along the trajectory generated by the explored nominal system (13), we obtain the integral equation

$$\begin{aligned}J(x(t+T)) - J(x(t)) &= \int_t^{t+T} (\nabla J(x))^\top g(x) e d\tau \\ &\quad - \int_t^{t+T} r(x, u(x)) d\tau,\end{aligned} \quad (15)$$

where the integral is well-defined since $J(x)$ and the interval $[t, t+T]$ are finite. This means that $J(x)$ as the unique solution of (14), also satisfies (15). Under the admissible control policy $u(x)$, if the state x is generated by the system (13), solving for $J(x)$ from the integral equation (15) is equivalent to finding the solution of (14).

Using the representation $J_i(x)$ and $u_i(x)$, and considering the policy improvement (12), the formulation (15) can be rewritten as

$$\begin{aligned} J_i(x(t+T)) - J_i(x(t)) = & -2 \int_t^{t+T} u_{i+1}^\top(x) R e d\tau \\ & - \int_t^{t+T} r(x, u_i(x)) d\tau. \end{aligned} \quad (16)$$

Since $f(x)$ and $g(x)$ do not appear in the integral equation (16), which is derived from (12) and (15), the PI algorithm can be done without knowing the system dynamics. Thus, we obtain the online model-free PI algorithm.

Algorithm 1. Model-free PI Algorithm

1. Give a small positive real number ϵ . Let $i = 0$ and start with an initial admissible control policy $u_0(x)$.
2. **Policy Evaluation and Improvement:** Based on the control policy $u_i(x)$, solve $J_i(x)$ and $u_{i+1}(x)$ from the integral equation (16).
3. If $\|u_{i+1}(x) - u_i(x)\| \leq \epsilon$, stop and obtain the approximate optimal control policy for the nominal system; else, set $i = i + 1$ and go to Step 2.

Theorem 3: An initial admissible control policy $u_0(x)$ is given for the nominal system (2). Using the model-free PI algorithm established in Algorithm 1, the cost function and the control policy converge to the optimal ones as $i \rightarrow \infty$, i.e.,

$$J_i(x) \rightarrow J^*(x), \quad u_i(x) \rightarrow u^*(x).$$

Proof. If the initial policy $u_0(x)$ is admissible, during the iteration process on (12) and (14), all the subsequent control policies will be admissible [23]. Moreover, the iteration result will converge to the solution of the HJB equation. Based on the formation process of (16), we can conclude that the proposed model-free PI algorithm will converge to the solution of the optimal control problem for the nominal system (2) without using the knowledge of system dynamics. The proof is completed. \square

C. Neural-network-based Implementation

In the following part of this section, we discuss the NN-based implementation method of the established model-free PI algorithm. A critic network and an actor network are used to approximate the cost function and the control policy of the nominal system, respectively. We assume that for the nominal system, $J_i(x)$ and $u_{i+1}(x)$ are represented on a compact set Ω by single-layer NNs as

$$\begin{aligned} J_i(x) &= \sum_{j=1}^{N_c} \omega_{ij} \phi_j(x) + \varepsilon_c(x), \\ u_{i+1}(x) &= \sum_{j=1}^{N_a} \nu_{ij} \psi_j(x) + \varepsilon_a(x), \end{aligned}$$

where $\omega_{ij} \in \mathbb{R}$ and $\nu_{ij} \in \mathbb{R}^m$ are unknown bounded ideal weights which will be determined by the established model-free PI algorithm, $\phi_j(x) \in \mathbb{R}$ and $\psi_j(x) \in \mathbb{R}$, $\{\phi_j\}_{j=1}^{N_c}$ and $\{\psi_j\}_{j=1}^{N_a}$ are the sequences of activation functions that are linearly independent and complete, and $\varepsilon_c(x) \in \mathbb{R}$ and $\varepsilon_a(x) \in \mathbb{R}^m$ are the bounded NN approximation errors. Since the ideal weights are unknown, the outputs of the critic network and the actor network are

$$\hat{J}_i(x) = \sum_{j=1}^{N_c} \hat{\omega}_{ij} \phi_j(x) = \hat{\omega}_i^\top \phi(x), \quad (17)$$

$$\hat{u}_{i+1}(x) = \sum_{j=1}^{N_a} \hat{\nu}_{ij} \psi_j(x) = \hat{\nu}_i^\top \psi(x), \quad (18)$$

where $\hat{\omega}_i$ and $\hat{\nu}_i$ are the current estimated weights, and

$$\phi(x) = [\phi_1(x), \phi_2(x), \dots, \phi_{N_c}(x)]^\top \in \mathbb{R}^{N_c}$$

$$\psi(x) = [\psi_1(x), \psi_2(x), \dots, \psi_{N_a}(x)]^\top \in \mathbb{R}^{N_a}$$

$$\hat{\omega}_i = [\hat{\omega}_{i1}, \hat{\omega}_{i2}, \dots, \hat{\omega}_{iN_c}]^\top \in \mathbb{R}^{N_c}$$

$$\hat{\nu}_i = [\hat{\nu}_{i1}, \hat{\nu}_{i2}, \dots, \hat{\nu}_{iN_a}]^\top \in \mathbb{R}^{N_a \times m}.$$

Define $\text{col}\{\hat{\nu}_i^\top\} = [\hat{\nu}_{i1}^\top, \hat{\nu}_{i2}^\top, \dots, \hat{\nu}_{iN_a}^\top]^\top \in \mathbb{R}^{mN_a}$, then

$$\begin{aligned} \hat{u}_{i+1}^\top(x) R e &= (\hat{\nu}_i^\top \psi(x))^\top R e \\ &= (\psi(x) \otimes (R e))^\top \text{col}\{\hat{\nu}_i^\top\}, \end{aligned}$$

where \otimes represents the Kronecker product. Substituting the expressions (17) and (18) into the integral equation (16), we obtain the following general form

$$\lambda_k^\top \left[\begin{array}{c} \hat{\omega}_i \\ \text{col}\{\hat{\nu}_i^\top\} \end{array} \right] = \theta_k \quad (19)$$

with

$$\theta_k = \int_{t+(k-1)T}^{t+kT} r(x, \hat{u}_i(x)) d\tau$$

$$\begin{aligned} \lambda_k &= \left[\left(\phi(x(t+(k-1)T)) - \phi(x(t+kT)) \right)^\top, \right. \\ &\quad \left. - 2 \int_{t+(k-1)T}^{t+kT} (\psi(x) \otimes (R e))^\top d\tau \right]^\top, \end{aligned}$$

where the measurement time is from $t + (k-1)T$ to $t + kT$. Since (19) is only a 1-dimensional equation, we cannot guarantee the uniqueness of the solution. Similar to [24], we use the least squares sense method to solve the parameter vector over the compact set Ω . For any positive integral K , we denote $\Lambda = [\lambda_1, \lambda_2, \dots, \lambda_K]$ and $\Theta = [\theta_1, \theta_2, \dots, \theta_K]^\top$. Then, we have the following K -dimensional equation

$$\Lambda^\top \left[\begin{array}{c} \hat{\omega}_i \\ \text{col}\{\hat{\nu}_i^\top\} \end{array} \right] = \Theta.$$

If Λ^\top has full column rank, the parameters can be solved by

$$\left[\begin{array}{c} \hat{\omega}_i \\ \text{col}\{\hat{\nu}_i^\top\} \end{array} \right] = (\Lambda \Lambda^\top)^{-1} \Lambda \Theta. \quad (20)$$

Therefore, we need to guarantee that the number of collected points K satisfies $K \geq \text{rank}(\Lambda) = N_c + mN_a$, which will

make $(\Lambda\Lambda^T)^{-1}$ exist. The least squares problem in (20) can be solved in real time by collecting enough data points generated from the system (13).

Based on the model-free PI algorithm, we obtain the approximation solution of the optimal control problem. We can conclude that the approximate optimal control policy $\hat{u}_i(x)$ can be obtained. According to (9), we have the robust control law

$$\bar{u}(x) = \pi \hat{u}_i(x).$$

Therefore, the robust optimal control law of the uncertain nonlinear system is derived.

IV. NUMERICAL SIMULATION

In this section, we consider the classical multi-machine power system with governor controllers [31]

$$\begin{aligned}\dot{\delta}_i(t) &= \omega_i(t), \\ \dot{\omega}_i(t) &= -\frac{D_i}{2H_i}\omega_i(t) + \frac{\omega_0}{2H_i}[P_{mi}(t) - P_{ei}(t)], \\ \dot{P}_{mi}(t) &= \frac{1}{T_i}[-P_{mi}(t) + u_{gi}(t)], \\ P_{ei}(t) &= E'_{qi} \sum_{j=1}^N E'_{qj} [B_{ij} \sin \delta_{ij}(t) + G_{ij} \cos \delta_{ij}(t)],\end{aligned}$$

where, for $1 \leq i, j \leq N$, N is the number of the generators. The meanings and values of these parameters are same as those in [31].

The second generator of the power system is selected to validate the applicability of the developed algorithm in this numerical simulation. Similarly, as in [31], we rewrite the second generator as the following form

$$\dot{x} = \begin{bmatrix} x_2 \\ -\frac{D}{2H}x_2 + \frac{\omega_0}{2H}x_3 \\ -\frac{1}{T}x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{T} \end{bmatrix} (\bar{u}(x) + \bar{d}(x)). \quad (21)$$

We define the states as $x_1 = \Delta\delta(t) = \delta(t) - \delta_0$, $x_2 = \Delta\omega(t) = \omega(t) - \omega_0$, $x_3 = \Delta P_m(t) = P_m(t) - P_e(t)$, and the system control is $\bar{u}(x(t)) = u_g(t) - P_e(t)$. The term $\bar{d}(t) = -E'_q[\delta_1 \cos(x_1 - \delta_3) - \delta_2 \sin(x_1 - \delta_3)] \times (x_2 - \delta_4)$ reflects the uncertainty caused by the other generators of the multi-machine power system, where $\delta_1, \delta_2, \delta_3$, and δ_4 are unknown parameters with $\delta_1 \in [0, 0.9]$, $\delta_2 \in [-0.45, 0.45]$, $\delta_3 \in [-60, 60]$, and $\delta_4 \in [-2, 2]$. We set $R = I$ and choose $d_M(x) = 10\sqrt{10}\|x\|$ as the bound of $d(x)$. According to the aforementioned results, the cost function can be represented as

$$J(x_0) = \int_0^\infty \{1000\|x(\tau)\|^2 + u^T(x(\tau))Ru(x(\tau))\} d\tau.$$

Assume that the exact knowledge of the dynamics (21) is fully unknown. We adopt the model-free PI algorithm to tackle the optimal control problem. In this simulation study, the activation functions are chosen as

$$\begin{aligned}\phi(x) &= [x_1^2, x_1x_2, x_1x_3, x_2^2, x_2x_3, x_3^2]^T, \\ \psi(x) &= [x_1, x_2, x_3]^T.\end{aligned}$$

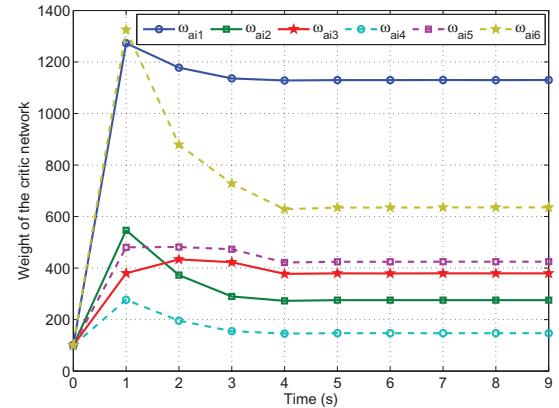


Fig. 1. Evolutions of the weight of the critic network (ω_{aij} represents $\hat{\omega}_{ij}$, $j = 1, 2, \dots, 6$)

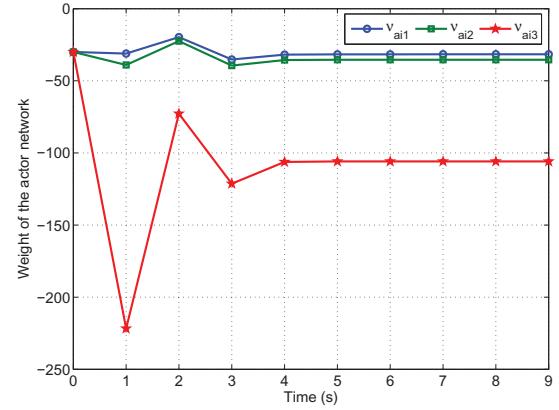


Fig. 2. Evolutions of the weight of the actor network (ν_{aij} represents $\hat{\nu}_{ij}$, $j = 1, 2, 3$)

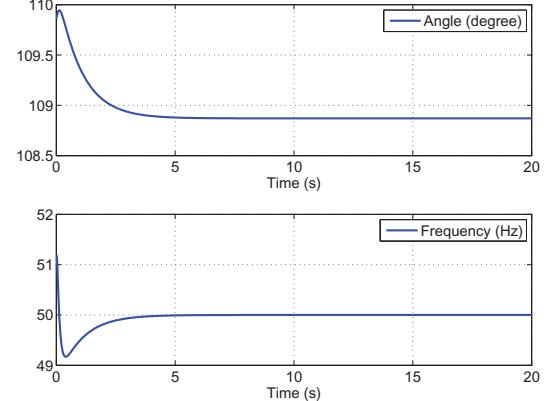


Fig. 3. The angle and frequency trajectories of the controlled generator when setting $\pi = 3$, $\delta_1 = 0.5$, $\delta_2 = 0.3$, $\delta_3 = 50$, and $\delta_4 = 2$

From these parameters, we know that $N_c = 6$ and $N_a = 3$. Then, we conduct the simulation with $K = 10$. During the simulation, the initial weights of the critic network and the actor network are chosen as

$$\begin{aligned}\hat{\omega}_0 &= 100 \times [1, 1, 1, 1, 1, 1]^T, \\ \hat{\nu}_0 &= -30 \times [1, 1, 1]^T.\end{aligned}$$

Let the initial state be $x_0 = [1, 1, 1]^\top$. The time interval $T = 0.1\text{s}$ and the probing signal $e(t) = 0.01 \sin(2\pi t) + 0.01 \cos(2\pi t)$ are used in the learning process. The least squares problem is solved after 10 samples are acquired, and thus the weights of the neural networks are updated every 1s. Figs. 1 and 2 illustrate the evolutions of the weights of the critic network and the actor network, respectively. The precision $\epsilon = 10^{-4}$ is achieved after nine iterations. At $t = 9\text{s}$,

$$\begin{aligned}\hat{\omega}_9 &= [1129.8863, 275.4226, 379.4565, 147.0610, \\ &\quad 424.6685, 635.3140]^\top, \\ \hat{\nu}_9 &= [-31.6230, -35.3899, -105.8822]^\top.\end{aligned}$$

To evaluate the robust control performance, the scalar parameters are chosen as $\pi = 3$, $\delta_1 = 0.5$, $\delta_2 = 0.3$, $\delta_3 = 50$, and $\delta_4 = 2$, respectively. Under the action of the robust control strategy, the angle and frequency trajectories of the generator during the first 20s is shown in Fig. 3. In light of Theorem 2, it also achieves optimality with cost function defined as (10). These results authenticate the availability of the robust optimal control scheme developed in this paper.

V. CONCLUSION

A robust optimal control policy for a class of uncertain nonlinear systems is developed in this paper, under the framework of the model-free PI algorithm. It is proved that the robust control law of the original uncertain system achieves optimality under a specified cost function. Then, the robust optimal control problem is transformed into an optimal control problem. The optimal control policy of the nominal system is developed without the system dynamics. The simulation study verifies the good control performance.

REFERENCES

- [1] H. Gao, X. Meng, and T. Chen, “A new design of robust H_2 filters for uncertain systems,” *Systems & Control Letters*, vol. 57, no. 7, pp. 585–593, July 2008.
- [2] Y. H. Lan and Y. Zhou, “Non-fragile observer-based robust control for a class of fractional-order nonlinear systems,” *Systems & Control Letters*, vol. 62, no. 12, pp. 1143–1150, Dec. 2013.
- [3] F. Lin, R. D. Brand, and J. Sun, “Robust control of nonlinear systems: Compensating for uncertainty,” *International Journal of Control*, vol. 56, no. 6, pp. 1453–1459, 1992.
- [4] D. Wang, D. Liu, H. Li, and X. Yang, “A learning optimal control scheme for robust stabilization of a class of uncertain nonlinear systems,” in *Proceedings of the Chinese Control Conference*, Xi'an, China, July 2013, pp. 7834–7839.
- [5] D. Wang, D. Liu, and H. Li, “Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems,” *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 627–632, Apr. 2014.
- [6] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.
- [7] P. J. Werbos, “Approximate dynamic programming for real-time control and neural modeling,” in *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, D. A. White and D. A. Sofge, Eds. New York, NY: Van Nostrand Reinhold, 1992, ch. 13.
- [8] D. V. Prokhorov and D. C. Wunsch, “Adaptive critic designs,” *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 997–1007, Sept. 1997.
- [9] F. L. Lewis and D. Vrabie, “Reinforcement learning and adaptive dynamic programming for feedback control,” *IEEE Circuits and Systems Magazine*, vol. 9, no. 3, pp. 32–50, July 2009.
- [10] F. L. Lewis and D. Liu, *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ: Wiley, 2013.
- [11] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control: Algorithms and Stability*. London, UK: Springer, 2013.
- [12] F. Y. Wang, H. Zhang, and D. Liu, “Adaptive dynamic programming: An introduction,” *IEEE Computational Intelligence Magazine*, vol. 4, no. 2, pp. 39–47, May 2009.
- [13] 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, May 2010.
- [14] J. Fu, H. He, and X. Zhou, “Adaptive learning and control for MIMO system based on adaptive dynamic programming,” *IEEE Transactions on Neural Networks*, vol. 22, no. 7, pp. 1133–1148, July 2011.
- [15] D. Wang, D. Liu, and Q. Wei, “Finite-horizon neuro-optimal tracking control for a class of discrete-time nonlinear systems using adaptive dynamic programming approach,” *Neurocomputing*, vol. 78, no. 1, pp. 14–22, Feb. 2012.
- [16] D. Wang, D. Liu, Q. Wei, D. Zhao, and N. Jin, “Optimal control of unknown nonaffine nonlinear discrete-time systems based on adaptive dynamic programming,” *Automatica*, vol. 48, no. 8, pp. 1825–1832, Aug. 2012.
- [17] D. Liu, H. Li, and D. Wang, “Neural-network-based zero-sum game for discrete-time nonlinear systems via iterative adaptive dynamic programming algorithm,” *Neurocomputing*, vol. 110, pp. 92–100, June 2013.
- [18] T. Dierks and S. Jagannathan, “Optimal control of affine nonlinear continuous-time systems,” in *Proceedings of the American Control Conference*, Baltimore, MD, USA, June 2010, pp. 1568–1573.
- [19] D. Nodland, H. Zargarzadeh, and S. Jagannathan, “Neural network-based optimal adaptive output feedback control of a helicopter UAV,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 7, pp. 1061–1073, July 2013.
- [20] 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, Dec. 2012.
- [21] Z. P. Jiang and Y. Jiang, “Robust adaptive dynamic programming for linear and nonlinear systems: An overview,” *European Journal of Control*, vol. 19, no. 5, pp. 417–425, Sept. 2013.
- [22] 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, Oct. 2012.
- [23] 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, Mar. 2009.
- [24] 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. 11, pp. 2850–2859, Nov. 2012.
- [25] J. Y. Lee, J. B. Park, and Y. H. Choi, “Integral Reinforcement Learning for Continuous-Time Input-Affine Nonlinear Systems With Simultaneous Invariant Explorations,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 916–932, May 2015.
- [26] B. Luo, H. N. Wu, T. Huang, and D. Liu, “Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design” *Automatica*, vol. 50 no. 12, pp. 3281–3290, Dec. 2014.
- [27] T. Bian, Y. Jiang, and Z. P. Jiang, “Adaptive dynamic programming and optimal control of nonlinear nonaffine systems,” *Automatica*, vol. 50 no. 10, pp. 2624–2632, Oct. 2014.
- [28] F. L. Lewis, D. Vrabie, V. Syrmos, *Optimal Control*. Hoboken, NJ: Wiley, 2012.
- [29] R. W. Beard, G. N. Saridis, J. T. Wen, “Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation,” *Automatica*, vol. 33, no. 12, pp. 2159–2177, Dec. 1997.
- [30] G. N. Saridis and C. G. Lee, “An approximation theory of optimal control for trainable manipulators,” *IEEE Transactions on Systems, Man, and Cybernetics—PART B: Cybernetics*, vol. 9, no. 3, pp. 152–159, 1979.
- [31] Y. Jiang, and Z. P. Jiang, “Robust adaptive dynamic programming for large-scale systems with an application to multimachine power systems,” *IEEE Transactions on Circuits and Systems—II: Express Briefs*, vol. 59, no. 10, pp. 693–697, 2012.