Error Bound Analysis of Q-Function for Discounted Optimal Control Problems With Policy Iteration

Pengfei Yan, *Student Member, IEEE*, Ding Wang, *Member, IEEE*, Hongliang Li, and Derong Liu, *Fellow, IEEE*

***Abstract*—In this paper, we present error bound analysis of the Q-function for the action-dependent adaptive dynamic programming for solving discounted optimal control problems of unknown discrete-time nonlinear systems. The convergence of Q-functions derived by a policy iteration algorithm under ideal conditions is given. Considering the approximated errors of the Q-function and control policy in the policy evaluation step and policy improvement step, we establish error bounds of approximate Q-functions in each iteration. With the given boundedness conditions, the approximate Q-function will converge to a finite neighborhood of the optimal Q-function. To implement the presented algorithm, two three-layer neural networks are employed to approximate the Q-function and the control policy, respectively. Finally, a simulation example is utilized to verify the validity of the presented algorithm.**

***Index Terms*—Adaptive dynamic programming (ADP), error analysis, nonlinear systems, policy iteration, Q-function.**

I. Introduction

OPTIMAL control is an indispensable field in modern control theory, and dynamic programming is one of the most effective techniques to solve the optimal control problem by solving the Hamilton–Jacobi–Bellman (HJB) equation [1]. Unfortunately, it is often difficult to obtain the solutions to a group of nonlinear partial difference equations and sometimes the HJB equation even has no analytic solution [2]. Furthermore, it will suffer the "curse of dimensionality" when running dynamic programming to solve the optimal control problem [3]. Adaptive dynamic programming (ADP) [4]–[6], similar to the reinforcement learning, is an improved technique of the dynamic programming and overcomes the problem of the curse of dimensionality. According to the difference of the function approximation architectures, existing ADP algorithms can be classified into several schemes: heuristic dynamic programming (HDP) which approximates the value function [7], dual heuristic programming (DHP) which approximates the derivative of value function [8], [9], global dual heuristic dynamic programming (GDHP) which approximates both the value function and the derivative of value function [10], and their action-dependent versions (ADHDP, ADDHP, and ADGDHP) which employ the state-action value function (also known as Q-function) [11], [12].

ADP has attracted wide-spread attention to obtain the optimal control policy in the interacted environment, and has been applied to practical problems, such as multiagent systems [13], energy system control [14], decentralized control [15], and zero-sum game [16], [17]. The iterative algorithms for ADP to solve the optimal control problem include value iteration, policy iteration, and generalized policy iteration [18] which generalizes value iteration and policy iteration. Value iteration can solve the optimal control problem of linear or nonlinear systems with an initial positive semi-definite function, which will converge to the optimal value function. Al-Tamimi *et al*. [7] proved the convergence of the value iteration algorithm for solving HJB equation of the discrete-time nonlinear systems. Policy iteration can solve the optimal control problems with an initial admissible control. With the iterations between the policy evaluation step and the policy improvement step, the control policy and the value function will converge to the optimal ones. The convergence of the policy iteration algorithm for continuous-time nonlinear systems and discrete-time nonlinear systems have been given in [19] and [20], respectively. Wei *et al*. [21] proved the iterative value function was monotonically nonincreasing and converged to the optimum by properly choosing the parameter. Wei and Liu [22] developed a generalized policy iteration algorithm to solve infinite horizon optimal control algorithm and demonstrated its stability. Wang *et al*. [23] demonstrated the stability of online policy iteration algorithm for continuous-time nonlinear systems. Generalized policy iteration contains value iteration and policy iteration, and its convergence is very difficult to analyze. Some work has been reported in the literature, such as generalized policy iteration for discounted finite-state Markov decision problems [24], and more work needs to be done in the area.

In recent years, data-driven control or model-free control has attracted more and more attention. Many ADP algorithms were designed to derive the optimal control for systems with partially known model [25]–[30] or unknown

Manuscript received October 29, 2015; revised January 28, 2016; accepted February 25, 2016. Date of publication June 1, 2016; date of current version June 22, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61233001, Grant 61273140, Grant 61304086, Grant 61374105, Grant 61533017, and Grant U1501251, in part by Beijing Natural Science Foundation under Grant 4162065, and in part by the Early Career Development Award of SKLMCCS. This paper was recommended by Associate Editor Y. Shen.

P. Yan and D. Wang are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Science, Beijing 100190, China (e-mail: pengfei.yan@ia.ac.cn; ding.wang@ia.ac.cn).

H. Li is with the IBM Research–China, Beijing 100193, China (e-mail: hongliang.li625@foxmail.com).

D. Liu is with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: derong@ustb.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMC.2016.2563982

model [2], [31]–[36]. However, in the above model-free schemes, the dynamic models were still required to be identified using the input-output data beforehand, which would bring additional computational cost on building the model network and introduce the modeling error which decreases the performance of the controller [7]. Fortunately, the action-dependent versions of ADP provide a useful way for deriving the optimal control policy directly by the state-action value function without building a model network [5], [18], [37]–[40]. Hence, the action-dependent versions of ADP are completely model-free, which have attracted many researchers' interest. Si and Wang [37] applied the ADHDP algorithm to solve the cart-pole balancing problem. Song *et al*. [41] developed a self learning optimal control laws for nonlinear systems based on ESN architecture. Yang *et al*. [42] introduced the input constraints to reinforcement learning for unknown continuous-time nonlinear system. Wei and Liu [43] proposed a policy iteration based deterministic Q-learning for discrete-time systems. Li *et al*. [44] developed an integral reinforcement learning for linear continuous-time zero sum games. Zhao *et al*. [45]–[47] improved the convergence rate of the ADHDP algorithm via integrating the prior experience into a supervisor for the driver assistance systems. Wei and Liu [48] introduced iterative ADP for temperature control of water gas shift reaction.

As the function approximation is used in most ADP schemes, the error bound analysis must be established to make sure the ADP methods available. Van Roy [49] established bounds for approximate value iteration with performance loss. Mounos provided the error bounds between approximate policies derived by the approximate value function and the optimal policy using L_p-norms [50], [51] and quadratic norm [52]. Bertsekas [53] gave the error bounds of approximate policy iteration according to weighted sup-norm contractions. Perkins and Precup [54] proved that the approximate policy iteration algorithm with a model-free form can converge to one solution. Liu and Wei [55] presented finite approximation error analysis based optimal control approach for discrete-time nonlinear systems. Rantzer [56] established a relaxed dynamic programming in switching systems which could make the distance from optimal values within preset bounds of the optimal cost function. Grune and Rantzer [57] applied the relaxed value iteration to discrete-time nonlinear systems with a receding horizon control scheme. Liu *et al*. [58] discussed error bounds of ADP algorithms for value iteration, policy iteration and generalized policy iteration. Most of the above works study the error bounds of value function which is just a function of the states. However, as a widely applied scheme, the action-dependent ADP has not been analyzed for error bounds of the Q-function when considering approximation errors, which motivates this paper. To the best of our knowledge, this is the first time to establish the error bounds of the Q-function with policy iteration for discounted optimal control problems considering the approximation errors.

The infinite-horizon optimal control problem of discrete-time deterministic nonlinear systems is an important area of optimal control [1], [59]. In this paper, we consider the approximation errors of the iterative Q-functions in policy evaluation step and policy improvement step. First, we give a data-driven ADP algorithm based on the discounted Q-function with policy iteration. Instead of the contraction assumption, a novel convergence proof of this algorithm under ideal conditions is established. Then, the error bounds of the approximate Q-functions derived by the approximate policy iteration are considered. Furthermore, we prove that the approximate Q-functions will converge to a finite neighborhood of the optimal Q-function. This result provides a theoretical guarantee to introduce an approximator such as neural network to action-dependent ADP, which is important to the development of ADP in both theory and application. To implement the developed algorithm, two three-layer neural networks are used to approximate the Q-function and the controller. Finally, a simulation example of mass-spring system is given to demonstrate the effectiveness of the developed algorithm.

The rest of this paper is organized as follows. Section II presents the infinite-horizon optimal control problem for discrete-time deterministic nonlinear systems. Section III presents the convergence and error bound analysis for the model-free optimal control problem based on policy iteration. Section IV provides the implementation of the developed approach with three-layer neural networks. Section V provides a simulation example to demonstrate the effectiveness of the developed algorithm, and Section VI gives the conclusions.

II. Problem Formulation

In this paper, we consider the following deterministic discrete-time nonlinear dynamical system:

$$x_{k+1} = f(x_k, u_k), \ \ k = 0, 1, 2, \ldots \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the state vector, and $u_k \in \mathbb{R}^m$ is the control input. The system (1) is assumed to be controllable which implies that there exists a continuous control policy on a compact set $\Omega \subseteq \mathbb{R}^n$ that stabilizes the system asymptotically. We assume that $x_k = 0$ is an equilibrium state of the system (1) and $f(0, 0) = 0$. The system function $f(x_k, u_k)$ is Lipschitz continuous with respect to x_k and u_k. The infinite horizon cost function with discount factor for any initial state x_0 is given by

$$J(x_0, u) = \sum_{k=0}^{\infty} \gamma^k U(x_k, u_k) \tag{2}$$

where u is the control sequence defined as $u = \{u_0, u_1, u_2, \ldots\}$, U is the utility function which is continuous and positive definite on x_k and u_k, and γ is the discount factor which satisfies $0 < \gamma \leq 1$. For any feedback control policy $\mu(x)$, the value function $V^{\mu}(x)$, the map from any state x to the value of (2), is defined as

$$V^{\mu}(x) = J(x, \mu(x)). \tag{3}$$

Definition 1: A feedback control policy $\mu(x)$ is admissible with respect to system (1), if $\mu(0) = 0$, $\mu(x)$ is continuous and stabilizes the system, and the value function $V^{\mu}(x)$ is finite.

Our goal is to find an admissible control which can minimize the value function such that

$$V^*(x) = \min_{\mu}\{V^{\mu}(x)\}. \tag{4}$$

According to Bellman's principle of optimality, the optimal value function satisfies the discrete-time HJB equation

$$V^*(x) = \min_{u}\{U(x,u) + \gamma V^*(f(x,u))\} \tag{5}$$

and the optimal control policy $\mu^*(x)$ is given by

$$\mu^*(x) = \arg\min_{u}\{U(x,u) + \gamma V^*(f(x,u))\}. \tag{6}$$

To replace the value function in data-driven ADP schemes, the Q-function which is also known as state-action value function is defined as

$$Q^{\mu}(x,u) = U(x,u) + \gamma Q^{\mu}\left(x^+, \mu\left(x^+\right)\right) \tag{7}$$

where x^+ is the state of the next moment, i.e., $x^+ = f(x,u)$. According to (3), the relationship between value function and Q-function is

$$\begin{aligned} Q^{\mu}(x,\mu(x)) &= U(x,\mu(x)) + \gamma Q^{\mu}\left(x^+, \mu\left(x^+\right)\right) \\ &= J(x,\mu(x)) \\ &= V^{\mu}(x). \end{aligned}$$

Then the optimal Q-function is defined as

$$Q^*(x,u) = \min_{\mu} Q^{\mu}(x,u)$$

and the optimal control policy $\mu^*(x)$ can be obtained by

$$\mu^*(x) = \arg\min_{u} Q^*(x,u). \tag{8}$$

The optimal Q-function satisfies the following Bellman optimality equation:

$$Q^*(x,u) = U(x,u) + \gamma \min_{u^+} Q^*\left(x^+, u^+\right) \tag{9}$$

where u^+ is the action of the next moment. The connection between the optimal value function and the optimal Q-function is

$$V^*(x) = \min_{u} Q^*(x,u).$$

In most situations, the optimal control problem for nonlinear systems has no analytical solution and traditional dynamic programming faces the curse of dimensionality. In this paper, we develop a data-driven iterative ADP algorithm by using Q-function which depends on the state and action to solve the nonlinear optimal control problem. Similar to most of the ADP methods, function approximation structures like neural networks are used to approximate the Q-function (or state-action value function) and the control policy. The approximation errors may increase along with the iteration processes. Therefore, it is necessary to establish the error bounds of Q-function for iteration algorithm considering the function approximation errors.

III. Policy-Iteration-Based Data-Driven ADP

In this section, we develop a data-driven ADP algorithm based on the discounted Q-function with policy iteration algorithm. Convergence analysis of this algorithm under ideal conditions is presented and the error bound for the algorithm considering the approximation errors is established.

A. Policy Iteration Under Ideal Conditions

We assume that the system (1) is unknown, and only an offline data set $\{x_k, u_k, x_{k+1}\}_N$ is available, where x_{k+1} is the next state of x_k and u_k, and N is the number of samples in the data set. x_{k+1} and x_k stand for the dynamic behavior of one-shot data and do not mean that the data set has to take samples from one trajectory. In general, the data set contains a variety of trajectories and scattered data.

For the policy iteration of the data-driven ADP algorithm, it starts with an initial admissible control μ_0. For $i = 0, 1, \ldots,$ the policy iteration algorithm contains policy evaluation phase and policy improvement phase given as follows.

Policy evaluation

$$Q_{j+1}^{\mu_i}(x_k,u_k) = U(x_k,u_k) + \gamma Q_j^{\mu_i}(x_{k+1}, \mu_i(x_{k+1})). \tag{10}$$

Policy improvement

$$\mu_{i+1}(x_k) = \arg\min_{u_k} Q^{\mu_i}(x_k,u_k) \tag{11}$$

where j is the policy evaluation index and i is the policy improvement index. $Q_j^{\mu_i}$ represents the jth evaluation for the ith control policy μ_i, and $Q_0^{\mu_i} = Q_{\infty}^{\mu_{i-1}}$. Let Q^{μ_i} denote the Q-function for μ_i. Next, we will prove that the limit of $Q_j^{\mu_i}$ as $j \to \infty$ exists and $Q_{\infty}^{\mu_i} = Q^{\mu_i}$.

Assumption 1: There exists a finite positive constant λ that makes the condition $\min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}) \le \lambda U(x_k, u_k)$ hold uniformly on Ω.

Remark 1: Assumption 1 is a basic assumption which ensures the convergence of ADP algorithms. For most nonlinear systems, it is easy to find a sufficiently large number λ to satisfy this assumption as $Q^*(\cdot)$ and $U(\cdot)$ are finite.

Lemma 1: Let Assumption 1 hold. Suppose that μ_0 is an admissible control policy and Q^{μ_0} is the Q-function of μ_0. Let $Q_j^{\mu_i}$ and μ_i be updated by (10) and (11). Then we can obtain the following conclusions.

1) The sequence $\{Q_j^{\mu_i}\}$ is monotonically nonincreasing, i.e., $Q_j^{\mu_i} \ge Q_{j+1}^{\mu_i}$, $\forall i \ge 1$. Moreover, as $j \to \infty$, the limit of $Q_j^{\mu_i}$ is denoted by $Q_{\infty}^{\mu_i}$, and equal to Q^{μ_i}, $\forall i \ge 1$.
2) The sequence $\{Q^{\mu_i}\}$ is monotonically nonincreasing, i.e., $Q^{\mu_i} \ge Q^{\mu_{i+1}}$, $\forall i \ge 1$.

Proof: The entire proof includes two parts.

1) μ_0 is an admissible control policy, according to (10) and (11), then we can obtain

$$\begin{aligned} Q_0^{\mu_1}(x_k,u_k) &= Q^{\mu_0}(x_k,u_k) \\ &= U(x_k,u_k) + \gamma Q^{\mu_0}(x_{k+1}, \mu_0(x_{k+1})) \\ &\ge U(x_k,u_k) + \gamma \min_{u_{k+1}} Q^{\mu_0}(x_{k+1}, u_{k+1}) \\ &= U(x_k,u_k) + \gamma Q^{\mu_0}(x_{k+1}, \mu_1(x_{k+1})) \\ &= Q_1^{\mu_1}(x_k,u_k). \end{aligned} \tag{12}$$

So $Q_j^{\mu_1} \geq Q_{j+1}^{\mu_1}$ holds for $j = 0$. If $Q_j^{\mu_1} \geq Q_{j+1}^{\mu_1}$ holds for $j = m-1$, when $j = m$, we can obtain

$$\begin{aligned} Q_m^{\mu_1}(x_k, u_k) &= U(x_k, u_k) + \gamma Q_{m-1}^{\mu_1}(x_{k+1}, \mu_1(x_{k+1})) \\ &\geq U(x_k, u_k) + \gamma Q_m^{\mu_1}(x_{k+1}, \mu_1(x_{k+1})) \\ &= Q_{m+1}^{\mu_1}(x_k, u_k). \end{aligned} \tag{13}$$

According to the mathematical induction, we can obtain that $Q_j^{\mu_i} \geq Q_{j+1}^{\mu_i}$ holds for $i = 1$. Since $\{Q_j^{\mu_1}\}$ is a monotonically nonincreasing sequence and $Q_j^{\mu_1} \geq 0$, the limit of $\{Q_j^{\mu_1}\}$ exists, which is denoted by $Q_\infty^{\mu_1}$, and $Q_j^{\mu_1} \geq Q_\infty^{\mu_1}, \forall j$. According to (10), we have

$$\begin{aligned} &Q_{j+1}^{\mu_1}(x_k, u_k) \\ &\quad \geq U(x_k, u_k) + \gamma Q_\infty^{\mu_1}(x_{k+1}, \mu_1(x_{k+1})), j \geq 0. \end{aligned} \tag{14}$$

Letting $j \to \infty$, we have

$$Q_\infty^{\mu_1}(x_k, u_k) \geq U(x_k, u_k) + \gamma Q_\infty^{\mu_1}(x_{k+1}, \mu_1(x_{k+1})). \tag{15}$$

Similarly, we can obtain

$$\begin{aligned} Q_\infty^{\mu_1}(x_k, u_k) &\leq U(x_k, u_k) \\ &\quad + \gamma Q_j^{\mu_1}(x_{k+1}, \mu_1(x_{k+1})), j \geq 0. \end{aligned} \tag{16}$$

Letting $j \to \infty$, we have

$$Q_\infty^{\mu_1}(x_k, u_k) \leq U(x_k, u_k) + \gamma Q_\infty^{\mu_1}(x_{k+1}, \mu_1(x_{k+1})). \tag{17}$$

Hence, we can obtain

$$Q_\infty^{\mu_1}(x_k, u_k) = U(x_k, u_k) + \gamma Q_\infty^{\mu_1}(x_{k+1}, \mu_1(x_{k+1})). \tag{18}$$

Thus, we can obtain that $Q_\infty^{\mu_i} = Q^{\mu_i}$ holds for $i = 1$. We assume that $Q_j^{\mu_i} \geq Q_{j+1}^{\mu_i}$ and $Q_\infty^{\mu_i} = Q^{\mu_i}$ hold for any $i = l,\ l \geq 1$. According to (10) and (11), we can obtain

$$\begin{aligned} Q_0^{\mu_{l+1}}(x_k, u_k) &= Q^{\mu_l}(x_k, u_k) \\ &= U(x_k, u_k) + \gamma Q^{\mu_l}(x_{k+1}, \mu_l(x_{k+1})) \\ &\geq U(x_k, u_k) + \gamma \min_{u_{k+1}} Q^{\mu_l}(x_{k+1}, u_{k+1}) \\ &= U(x_k, u_k) + \gamma Q^{\mu_l}(x_{k+1}, \mu_{l+1}(x_{k+1})) \\ &= Q_1^{\mu_{l+1}}(x_k, u_k). \end{aligned} \tag{19}$$

Considering (10) and (19), we have

$$\begin{aligned} Q_1^{\mu_{l+1}}(x_k, u_k) &= U(x_k, u_k) + \gamma Q_0^{\mu_{l+1}}(x_{k+1}, \mu_l(x_{k+1})) \\ &\geq U(x_k, u_k) + \gamma Q_1^{\mu_{l+1}}(x_{k+1}, \mu_l(x_{k+1})) \\ &= Q_2^{\mu_{l+1}}(x_k, u_k). \end{aligned} \tag{20}$$

Similarly, we can obtain that $Q_j^{\mu_i} \geq Q_{j+1}^{\mu_i}$ holds for $i = l+1$ by induction and $Q_\infty^{\mu_i} = Q^{\mu_i}$. Therefore, the proof is completed by induction.

2) According to 1), we have

$$Q^{u_i} = Q_0^{u_{i+1}} \geq Q_\infty^{u_{i+1}} = Q^{u_{i+1}}. \tag{21}$$

Therefore, $\{Q^{\mu_i}\}$ is a monotonically nonincreasing sequence and the proof is completed. ■

Theorem 1: Let Assumption 1 hold. Suppose that $Q^* \leq Q^{\mu_0} \leq \beta Q^*$, $1 \leq \beta \leq \infty$. μ_0 is an admissible control policy and Q^{μ_0} is the Q-function of μ_0. Let $Q_j^{\mu_i}$ and μ_i be updated by (10) and (11). Then the Q-function sequence Q^{μ_i} approaches Q^* according to the inequalities

$$Q^* \leq Q^{\mu_i} \leq \left[1 + \frac{\beta - 1}{\left(1 + \lambda^{-1}\right)^i}\right] Q^*, \quad \forall i \geq 1. \tag{22}$$

Proof: According to the definitions of Q^* and Q^{μ_i}, the left-hand side of the inequality (22) always holds for any $i \geq 1$. Next, we prove the right-hand side of (22) by induction. According to Assumption 1, we have

$$\gamma \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}) \leq \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}) \leq \lambda U(x_k, u_k). \tag{23}$$

Considering (10), (11), and (23), we can obtain

$$\begin{aligned} Q_1^{\mu_1}(x_k, u_k) &= U(x_k, u_k) + \gamma \min_{u_{k+1}} Q_0^{\mu_1}(x_{k+1}, u_{k+1}) \\ &= U(x_k, u_k) + \gamma \min_{u_{k+1}} Q^{\mu_0}(x_{k+1}, u_{k+1}) \\ &\leq U(x_k, u_k) + \gamma\beta \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}) \\ &\leq U(x_k, u_k) + \gamma\beta \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}) \\ &\quad + \frac{\beta - 1}{\lambda + 1}\left[\lambda U(x_k, u_k) - \gamma \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1})\right] \\ &= \left(1 + \lambda\frac{\beta - 1}{\lambda + 1}\right) U(x_k, u_k) \\ &\quad + \left(\beta + \frac{\beta - 1}{\lambda + 1}\right)\gamma \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}) \\ &= \left[1 + \frac{\beta - 1}{\left(1 + \lambda^{-1}\right)}\right] Q^*(x_k, u_k). \end{aligned} \tag{24}$$

According to Lemma 1, we have

$$Q^{\mu_1} = Q_\infty^{\mu_1} \leq Q_1^{\mu_1}(x_k, u_k) \leq \left[1 + \frac{\beta - 1}{\left(1 + \lambda^{-1}\right)}\right] Q^*(x_k, u_k) \tag{25}$$

which shows that the right-hand side of (22) holds for $i = 1$. Assume that

$$Q^{\mu_i}(x_k, u_k) \leq \left[1 + \frac{\beta - 1}{\left(1 + \lambda^{-1}\right)^i}\right] Q^*(x_k, u_k)$$

holds for any $i = l,\ l \geq 1$, then we can obtain

$$\begin{aligned} &Q^{\mu_{l+1}}(x_k, u_k) \\ &\quad \leq Q_1^{\mu_{l+1}}(x_k, u_k) \\ &\quad = U(x_k, u_k) + \gamma \min_{u_{k+1}} Q^{\mu_l}(x_{k+1}, u_{k+1}) \\ &\quad \leq U(x_k, u_k) + \gamma\left[1 + \frac{\beta - 1}{\left(1 + \lambda^{-1}\right)^l}\right] \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}) \\ &\qquad + \frac{\beta - 1}{\left(1 + \lambda^{-1}\right)^{l+1}}\left[\lambda U(x_k, u_k) - \gamma \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1})\right] \\ &\quad \leq \left[1 + \frac{\beta - 1}{\left(1 + \lambda^{-1}\right)^{l+1}}\right] Q^*(x_k, u_k) \end{aligned} \tag{26}$$

which shows that the right of (22) holds for $i = l + 1$. According to the mathematical induction, the right of (22) holds. The proof is completed. ■

In policy iteration, an initial admissible control policy is required, which is usually obtained by experience or trial. However, for most nonlinear systems, it is hard to obtain an admissible control policy, especially in data-driven ADP for unknown systems. So we present a novel initial condition for policy iteration.

Lemma 2: Let Assumption 1 hold. Suppose that there is an positive definite function Q_0 satisfying $\gamma Q_0 \geq Q_1^{\mu_1}$ for any x_k, u_k. Let $Q_j^{\mu_1}$ and μ_1 be obtained by (10) and (11). Then $\mu_1(x)$ is an admissible control policy and $Q^{\mu_1} = Q_\infty^{\mu_1}$ is the Q-function of $\mu_1(x)$.

Proof: Considering (10) and (11), we have

$$\mu_1(x_k) = \arg\min_{u_k} Q_0(x_k, u_k) \tag{27}$$

and

$$Q_1^{\mu_1}(x_k, u_k) = U(x_k, u_k) + \gamma Q_0(x_{k+1}, \mu_1(x_{k+1})). \tag{28}$$

Using the assumption of Q_0, we can obtain

$$\begin{aligned}\gamma Q_0(x_k, \mu_1(x_k)) &\geq Q_1^{\mu_1}(x_k, \mu_1(x_k)) \\ &= U(x_k, \mu_1(x_k)) + \gamma Q_0(x_{k+1}, \mu_1(x_{k+1})).\end{aligned} \tag{29}$$

Considering $Q_0(x_k, \mu_1(x_k)) \geq 0$ and

$$Q_0(x_{k+1}, \mu_1(x_{k+1})) - Q_0(x_k, \mu_1(x_k)) \leq -1/\gamma U(x_k, \mu_1(x_k)) \tag{30}$$

we can conclude that the control policy μ_1 is asymptotically stable for the system (1). Then, similar to (12)–(18), we can obtain that $Q^{\mu_1} = Q_\infty^{\mu_1} \leq Q_0$. Thus, the value function of μ_1 is

$$V^{\mu_1}(x_k) = Q^{\mu_1}(x_k, \mu_1(x_k)) \leq Q_0(x_k, \mu_1(x_k)). \tag{31}$$

Therefore, we can conclude that $\mu_1(x_k)$ is an admissible control. The proof is completed. ∎

According to Lemma 2, we can obtain an admissible control by using an initial positive definite function Q_0. Thus, considering Theorem 1, we can obtain the following corollary.

Corollary 1: Let Assumption 1 hold. Suppose that there is an initial positive definite function Q_0 which satisfies $\gamma Q_0 \geq Q_1^{\mu_1}$ for any x_k, u_k. Let $Q_j^{\mu_1}$ and μ_1 be updated by (10) and (11). Then the Q-function sequence $\{Q^{\mu_i}\}$ approaches Q^* according to the following inequalities:

$$Q^* \leq Q^{\mu_i} \leq \left[1 + \frac{\beta - 1}{\left(1 + \lambda^{-1}\right)^i}\right] Q^*, \quad \forall i \geq 1. \tag{32}$$

From Theorem 1 and Corollary 1, we can see that as $i \to \infty$, Q^{μ_i} converges to Q^* in ideal conditions, i.e., the control policy and Q-function in each iteration can be obtained accurately. They also give a convergence rate of Q^{μ_i} with policy iteration. When the discount factor $\gamma = 1$, the discounted optimal control problem turns into an undiscounted optimal control problem, and Theorem 1 and Corollary 1 still hold.

However, in practice, considering that the iteration indices i and j cannot reach infinity as the algorithm must stop in finite steps, there exist convergence errors in the iteration process. In addition, the control policy and Q-function in each iteration are obtained by approximation structures, so there exist approximate errors between approximate and accurate values. Hence, Theorem 1 and Corollary 1 may be invalid and the policy-iteration-based data-driven ADP may even be divergent. To overcome this difficulty, in the following section we establish new error bound analysis results for Q-function considering the convergence and approximation errors.

B. Error Bound for Approximate Policy Iteration

For the approximate policy iteration, function approximation structures are used to approximate the Q-function and the control policy. The approximate expressions of Q^{μ_i} and μ_i are $\hat{Q}^{\hat{\mu}_i}$ and $\hat{\mu}_i$, respectively. We assume that there exist two finite positive constants $\underline{\delta} \leq 1$ and $\bar{\delta} \geq 1$ that make

$$\underline{\delta} Q^{\hat{\mu}_i} \leq \hat{Q}^{\hat{\mu}_i} \leq \bar{\delta} Q^{\hat{\mu}_i} \tag{33}$$

hold uniformly, for any $i \geq 1$, where $Q^{\hat{\mu}_i}$ is the exact Q-function associated with $\hat{\mu}_i$. $\underline{\delta}$ and $\bar{\delta}$ imply the convergence error in j-iteration and the approximation error of $Q^{\hat{\mu}_i}$ in policy evaluation phase. When $\underline{\delta} = \bar{\delta} = 1$, both errors are zero. Considering Lemma 1, we can obtain

$$\hat{Q}^{\hat{\mu}_i} \leq \bar{\delta} Q^{\hat{\mu}_i} \leq \bar{\delta} \hat{Q}_1^{\hat{\mu}_i} \tag{34}$$

where $\hat{Q}_1^{\hat{\mu}_i}(x_k, u_k) = U(x_k, u_k) + \gamma \hat{Q}^{\hat{\mu}_{i-1}}(x_{k+1}, \hat{\mu}_i(x_{k+1}))$. Similarly, we assume that there exist two finite positive constants $\underline{\sigma} \leq 1$ and $\bar{\sigma} \geq 1$ that make

$$\underline{\sigma} Q_1^{\hat{\mu}_i} \leq \hat{Q}_1^{\hat{\mu}_i} \leq \bar{\sigma} Q_1^{\hat{\mu}_i} \tag{35}$$

hold uniformly, $\forall i \geq 1$, where $Q_1^{\hat{\mu}_i}(x_k, u_k) = U(x_k, u_k) + \gamma \hat{Q}^{\hat{\mu}_{i-1}}(x_{k+1}, \mu_i(x_{k+1}))$. $\underline{\sigma}$ and $\bar{\sigma}$ imply the approximation errors of $\hat{\mu}_i$ in the policy improvement phase. If the iterative control policy can be obtained accurately, then $\underline{\sigma} = \bar{\sigma} = 1$. Considering (34) and (35), we can get

$$\hat{Q}^{\hat{\mu}_i} \leq \bar{\sigma} \bar{\delta} Q_1^{\hat{\mu}_i}. \tag{36}$$

On the other hand, considering (33) and (35), we have

$$\hat{Q}^{\hat{\mu}_i} \geq \underline{\delta} Q^{\hat{\mu}_i} \geq \underline{\delta} Q^*. \tag{37}$$

Therefore, the whole approximation errors in the Q-function and control policy update step can be expressed by

$$\underline{\epsilon} Q^* \leq \hat{Q}^{\hat{\mu}_i} \leq \bar{\epsilon} Q_1^{\hat{\mu}_i} \tag{38}$$

where $\underline{\epsilon} = \underline{\delta}$ and $\bar{\epsilon} = \bar{\sigma}\bar{\delta}$. We establish the error bounds for approximate policy iteration by the following theorem.

Theorem 2: Let Assumption 1 hold. Suppose that $Q^* \leq Q^{\mu_0} \leq \beta Q^*$, $1 \leq \beta \leq \infty$. μ_0 is an admissible control policy and Q^{μ_0} is the Q-function of μ_0. The approximate Q-function $\hat{Q}^{\hat{\mu}_i}$ satisfies the iterative error condition (38). Then, the Q-function sequence $\{\hat{Q}^{\hat{\mu}_i}\}$ approaches Q^* according to the following inequalities:

$$\begin{aligned}\underline{\epsilon} Q^* &\leq \hat{Q}^{\hat{\mu}_{i+1}} \\ &\leq \bar{\epsilon}\left[1 + \sum_{j=1}^{i} \frac{\lambda^j \bar{\epsilon}^{j-1}(\bar{\epsilon} - 1)}{(\lambda + 1)^j} + \frac{\lambda^i \bar{\epsilon}^i (\beta - 1)}{(\lambda + 1)^i}\right] Q^*.\end{aligned} \tag{39}$$

Moreover, the approximate value function sequence $\{\hat{Q}^{\hat{\mu}_i}\}$ converges to a finite neighborhood of Q^* uniformly on Ω as $i \to \infty$, that is

$$\underline{\epsilon} Q^* \le \lim_{i\to\infty} \hat{Q}^{\hat{\mu}_i} \le \frac{\bar{\epsilon}}{1+\lambda-\bar{\epsilon}\lambda} Q^* \tag{40}$$

under the condition $\bar{\epsilon} \le 1/\lambda + 1$.

Proof: First, the left-hand side of (39) holds clearly according to (38). Next, we prove the right-hand side of (39) holds for $i \ge 1$. Considering (38) and Assumption 1, we can obtain

$$\begin{aligned}
&\hat{Q}^{\hat{\mu}_2}(x_k, u_k) \\
&\quad \le \bar{\epsilon} Q_1^{\hat{\mu}_2}(x_k, u_k) \\
&\quad = \bar{\epsilon}[U(x_k, u_k) + \gamma \hat{Q}^{\hat{\mu}_1}(x_{k+1}, \hat{\mu}_2(x_{k+1}))] \\
&\quad = \bar{\epsilon}[U(x_k, u_k) + \gamma \min_{u_{k+1}} Q^{\hat{\mu}_1}(x_{k+1}, u_{k+1})] \\
&\quad \le \bar{\epsilon}[U(x_k, u_k) + \gamma\beta\bar{\epsilon} \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1})] \\
&\qquad + \bar{\epsilon}\frac{\bar{\epsilon}\beta - 1}{\lambda+1}[\lambda U(x_k, u_k) - \gamma \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1})] \\
&\quad = \bar{\epsilon}\left(1 + \lambda\frac{\bar{\epsilon}\beta - 1}{\lambda+1}\right) U(x_k, u_k) \\
&\qquad + \bar{\epsilon}\left(\beta\bar{\epsilon} - \frac{\bar{\epsilon}\beta - 1}{\lambda+1}\right)\gamma \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}) \\
&\quad = \bar{\epsilon}\left(1 + \lambda\frac{\bar{\epsilon}\beta - 1}{\lambda+1}\right) Q^*(x_k, u_k) \\
&\quad = \bar{\epsilon}\left(1 + \frac{\lambda(\bar{\epsilon}-1)}{\lambda+1} + \frac{\lambda\bar{\epsilon}(\beta-1)}{\lambda+1}\right) Q^*(x_k, u_k).
\end{aligned} \tag{41}$$

Hence, the upper bound of $\hat{Q}^{\hat{\mu}_{i+1}}$ holds for $i = 1$. Suppose that the upper bound of $\hat{Q}^{\hat{\mu}_i}$ holds for $i \ge 1$. Then, we have

$$\begin{aligned}
&\hat{Q}^{\hat{\mu}_{i+1}}(x_k, u_k) \\
&\quad \le \bar{\epsilon}\left[U(x_k, u_k) + \gamma \min_{u_{k+1}} \hat{Q}^{\hat{\mu}_i}(x_{k+1}, u_{k+1})\right] \\
&\quad \le \bar{\epsilon}\left[U(x_k, u_k) + \gamma\bar{\epsilon}\rho \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1})\right] \\
&\quad \le \bar{\epsilon}\left[U(x_k, u_k) + \gamma\bar{\epsilon}\rho \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1})\right] \\
&\qquad + \bar{\epsilon}\Delta\left[\lambda U(x_k, u_k) - \gamma \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1})\right] \\
&\quad = \bar{\epsilon}(1+\Delta\lambda)\left[U(x_k, u_k) + \gamma \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1})\right] \\
&\quad = \bar{\epsilon}(1+\Delta\lambda) Q^*(x_k, u_k)
\end{aligned} \tag{42}$$

where $\rho = 1 + \sum_{j=1}^{i-1}(\lambda^j\bar{\epsilon}^{j-1}(\bar{\epsilon}-1)/(\lambda+1)^j) + (\lambda^{i-1}\bar{\epsilon}^{i-1}(\beta-1)/(\lambda+1)^{i-1})$, Δ satisfies $\Delta \ge 0$ and $1 + \Delta\lambda = \bar{\epsilon}\rho - \Delta$. Hence, we can calculate

$$\begin{aligned}
\Delta &= \frac{\bar{\epsilon}\rho - 1}{1+\lambda} \\
&= \frac{\bar{\epsilon}-1}{1+\lambda} + \sum_{j=1}^{i-1}\frac{\lambda^j\bar{\epsilon}^j(\bar{\epsilon}-1)}{(\lambda+1)^{j+1}} + \frac{\lambda^{i-1}\bar{\epsilon}^i(\beta-1)}{(\lambda+1)^i} \\
&= \sum_{j=1}^{i}\frac{\lambda^{j-1}\bar{\epsilon}^{j-1}(\bar{\epsilon}-1)}{(\lambda+1)^j} + \frac{\lambda^{i-1}\bar{\epsilon}^i(\beta-1)}{(\lambda+1)^i}.
\end{aligned} \tag{43}$$

Substituting (43) into (42), we can obtain

$$\hat{Q}^{\hat{\mu}_{i+1}} \le \bar{\epsilon}\left[1 + \sum_{j=1}^{i}\frac{\lambda^j\bar{\epsilon}^{j-1}(\bar{\epsilon}-1)}{(\lambda+1)^j} + \frac{\lambda^i\bar{\epsilon}^i(\beta-1)}{(\lambda+1)^i}\right] Q^*. \tag{44}$$

Thus, the upper bound of $\hat{Q}^{\hat{\mu}_i}$ holds for $i+1$. According to the mathematical induction, the right-hand side of (39) holds. Since the sequence $\{\lambda^j\bar{\epsilon}^{j-1}(1-\bar{\epsilon})/(\lambda+1)^j\}$ is a geometric series, we have

$$\sum_{j=1}^{i}\frac{\lambda^j\bar{\epsilon}^{j-1}(\bar{\epsilon}-1)}{(\lambda+1)^j} = \frac{\frac{\lambda(1-\bar{\epsilon})}{\lambda+1}\left(1 - \left(\frac{\lambda\bar{\epsilon}}{\lambda+1}\right)^i\right)}{1 - \frac{\lambda\bar{\epsilon}}{\lambda+1}}. \tag{45}$$

Considering $\bar{\epsilon} \le 1/\lambda + 1$, we have

$$\lim_{i\to\infty} \hat{Q}^{\hat{\mu}_i} \le \frac{\bar{\epsilon}}{1+\lambda-\bar{\epsilon}\lambda} Q^*. \tag{46}$$

The proof is completed. ■

Remark 2: We can find that the upper bound is a monotonically increasing function of $\bar{\epsilon}$. The condition $\bar{\epsilon} \le 1/\lambda + 1$ ensures that the upper bound in (40) is finite and positive. A larger λ will lead to a slower convergence rate and a larger error bound. Besides, a larger λ also requires more accurate iteration to converge. When $\underline{\epsilon} = \bar{\epsilon} = 1$, the approximate Q-function sequence $\hat{Q}^{\hat{\mu}_i}$ converges to Q^* uniformly on Ω as $i \to \infty$.

For the undiscounted optimal control problem, the discount factor $\gamma = 1$, and the Q-function is redefined as

$$Q^{\mu}(x, u) = U(x, u) + Q^{\mu}\left(x^+, \mu\left(x^+\right)\right) \tag{47}$$

and the optimal Q-function satisfies

$$Q^*(x, u) = U(x, u) + \min_{u^+} Q^*\left(x^+, u^+\right). \tag{48}$$

From Theorems 1 and 2, we know that when $\gamma = 1$, all the deductions still hold. So we have the following corollary.

Corollary 2: For the undiscounted optimal control problem with Assumption 1 and the admissible control policy μ_0 satisfying $Q^* \le Q^{\mu_0} \le \beta Q^*$, $1 \le \beta \le \infty$, if the approximate Q-function $\hat{Q}^{\hat{\mu}_i}$ satisfies the iterative error condition (38), the approximate Q-function sequence $\{\hat{Q}^{\hat{\mu}_i}\}$ converges to a finite neighborhood of Q^* uniformly on Ω as $i \to \infty$, that is

$$\underline{\epsilon} Q^* \le \lim_{i\to\infty} \hat{Q}^{\hat{\mu}_i} \le \frac{\bar{\epsilon}}{1+\lambda-\bar{\epsilon}\lambda} Q^* \tag{49}$$

under the condition $\bar{\epsilon} \le 1/\lambda + 1$.

IV. Neural Network Implementation for Approximate Policy Iteration

In the previous section, the approximation Q-function with policy iteration is proven to converge to a finite neighborhood of the optimal one. Hence, it is feasible to approximate the Q-function and the control policy using neural networks. We present the detailed implementation of the proposed algorithm using neural networks in this section.

The structure diagram of the data-driven iterative ADP in this paper is shown in Fig. 1. The outputs of critic network and

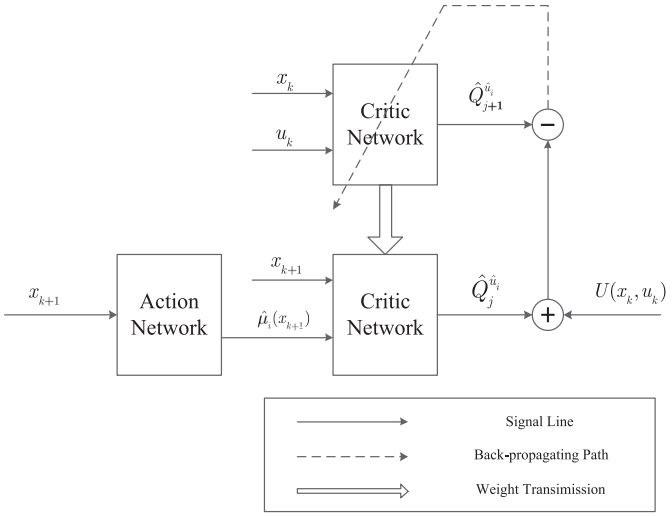

Fig. 1. Structure diagram of data-driven iterative ADP.

the action network are the approximations of the Q-function and the control policy, respectively.

The approximate Q-function $\hat{Q}^{\hat{u}_i}_j(x_k, u_k)$ is expressed by

$$\hat{Q}^{\hat{u}_i}_j(x_k, u_k) = W^{\mathsf{T}}_{c(ij)}\sigma\left(V^{\mathsf{T}}_{c(ij)}\left[x^{\mathsf{T}}_k, u^{\mathsf{T}}_k\right]^{\mathsf{T}}\right) \tag{50}$$

where $\sigma(\cdot)$ is the activation function, which is selected as $\tanh(\cdot)$. The target function of the critic neural network is given by

$$\hat{Q}^*_{i,j+1}(x_k, u_k) = U(x_k, u_k) + \hat{Q}^{\hat{u}_i}_j\left(x_{k+1}, \hat{\mu}_i(x_{k+1})\right) \tag{51}$$

where $x_{k+1} = f(x_k, \hat{\mu}_i(x_k))$. Then, the training function for the critic network is defined by

$$e_{c(i,j+1)}(x_k) = \hat{Q}^{\hat{u}_i}_{j+1}(x_k, u_k) - \hat{Q}^*_{i,j+1}(x_k, u_k) \tag{52}$$

and the performance function to be minimized is defined by

$$E_{c(i,j+1)} = \frac{1}{2}e^{\mathsf{T}}_{c(i,j+1)}e_{c(i,j+1)}. \tag{53}$$

The approximate control policy $\hat{\mu}_i$ is expressed by the action network

$$\hat{\mu}_i(x_k) = W^{\mathsf{T}}_{a(i)}\sigma\left(V^{\mathsf{T}}_{a(i)}x_k\right). \tag{54}$$

The target function of the action network is defined by

$$\hat{\mu}^*_{i+1}(x_k) = \arg\min_{\mu}\hat{Q}^{\hat{\mu}_i}(x_k, u). \tag{55}$$

Then, the error function for training the action network is defined by

$$e_{a(i+1)}(x_k) = \hat{\mu}_{i+1}(x_k) - \hat{\mu}^*_{i+1}(x_k). \tag{56}$$

The performance function of the action network to be minimized is defined by

$$E_{a(i+1)} = \frac{1}{2}e^{\mathsf{T}}_{a(i+1)}e_{a(i+1)}. \tag{57}$$

We use the gradient descent method to update the weights of critic and action networks on a data set. A detailed process of the approximate policy iteration is given in Algorithm 1.

Algorithm 1 Approximate Policy Iteration

1: Initialization:
Initialize critic and action networks randomly;
Select an initial stabilizing control policy μ_0;
Set the approximation errors of policy evaluation step and policy improvement step as ζ and ξ, and maximum iteration numbers of policy evaluation step and policy improvement step as $J_{\max}$ and $I_{\max}$.
2: Set $i = 0$.
3: For $j = 0, 1, \ldots, J_{\max}$, update the Q-function $\hat{Q}^{\hat{\mu}_i}_{j+1}(x_k)$ by minimizing (53) on the training set $\{x_k\}$. When $j = J_{\max}$ or the convergence conditions are met, set $\hat{Q}^{\hat{\mu}_i}(x_k) = \hat{Q}^{\hat{\mu}_i}_{j+1}(x_k)$ and go to Step 4.
4: Update the control policy $\hat{\mu}_{i+1}(x_k)$ by minimizing (57) on the training set $\{x_k\}$.
5: Set $i \leftarrow i + 1$.
6: Repeat Steps 3–5 until the convergence conditions are met.
7: Obtain the approximate optimal control policy $\hat{\mu}_i(x_k)$.

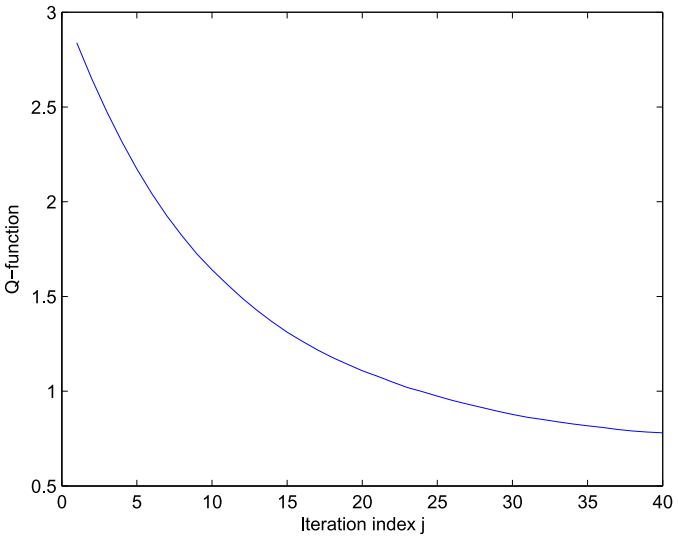

Fig. 2. Convergence of Q-function $\hat{Q}^{\hat{\mu}_i}_j$ on state $[0.5, -0.5]^{\mathsf{T}}$ at $i = 1$.

V. Simulation Study

In this section, we use a simulation example to demonstrate the effectiveness of the developed algorithm. Consider the mass-spring system [34] whose dynamics is

$$\begin{bmatrix} x_{1,k+1} \\ x_{2,k+1} \end{bmatrix} = \begin{bmatrix} 0.05x_{2,k} \\ -0.0005x_{1,k} - 0.0335x^3_{1,k} + x_{2,k} \end{bmatrix} + \begin{bmatrix} 0 \\ 0.05 \end{bmatrix}u_k. \tag{58}$$

Define the Q-function as

$$Q(x_0, u) = \sum_{k=0}^{\infty}\gamma^k\left(x^{\mathsf{T}}_k Q x_k + u^{\mathsf{T}}_k R u_k\right) \tag{59}$$

where $Q = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$, $R = 0.1$, and $\gamma = 0.95$.

In the simulation, we choose two three-layer neural networks as the approximation structures of controller and Q-function. The structures of the critic and action neural networks are chosen as 3–8–1 and 2–8–1, respectively. The

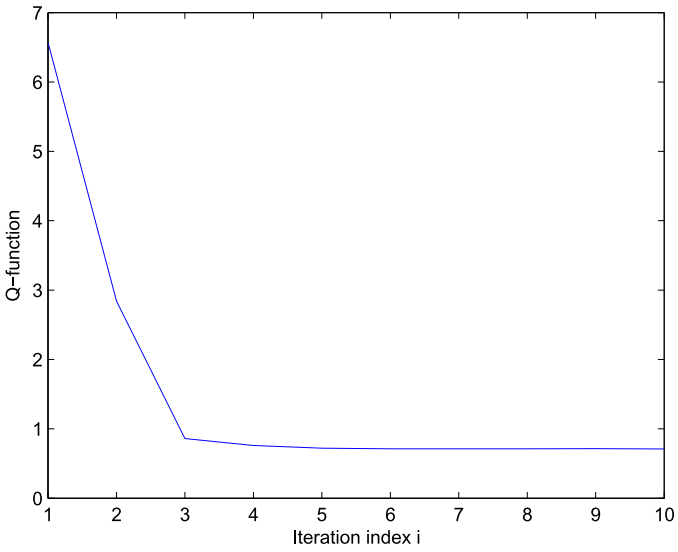

Fig. 3. Convergence of Q-function $\hat{Q}^{\hat{\mu}_i}$ on state $[0.5, -0.5]^{\mathsf{T}}$.

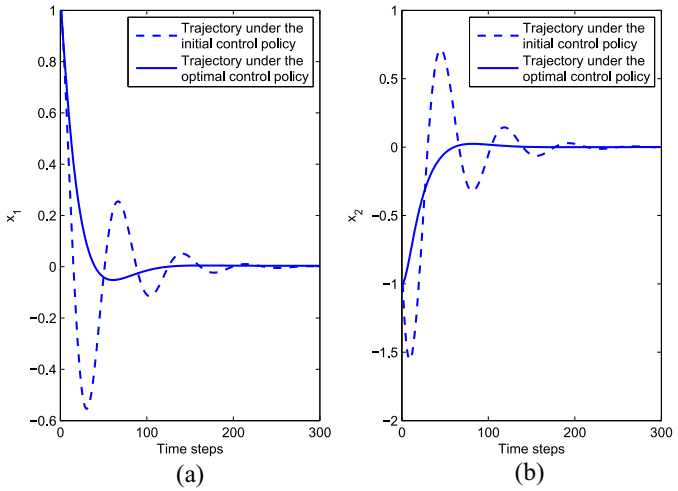

Fig. 4. State trajectories from the state $[1, -1]^{\mathsf{T}}$ for (a) x_1 and (b) x_2.

activation function is selected as $\tanh(\cdot)$. In order to guarantee the convergence of the neural network training, the initial weights of the activation functions are chosen randomly around zero. In this paper, we set the initial weights of both the critic and action networks as random values with uniform distribution of $[-0.01, 0.01]$. The preset approximation errors are $\zeta = \xi = 10^{-8}$, and the maximum iteration steps of policy evaluation and policy improvement are $J_{\max} = 40$ and $I_{\max} = 10$. The compact subset Ω of the state space is chosen as $0 \le x_1 \le 1$ and $0 \le x_2 \le 1$. The training set $\{x_k\}$ contains 1000 samples choosing randomly from the compact set Ω. The initial admissible control policy is chosen as $\mu_0 = [-3, -1]x_k$.

We train the action and critic neural networks offline with Algorithm 1. Fig. 2 illustrates the convergence process of the Q-function $\hat{Q}_j^{\hat{\mu}_i}$ with the iteration index j on state $x = [-0.5, 0.5]^{\mathsf{T}}$ at $i = 1$. Fig. 3 shows the convergence curve of Q-function $\hat{Q}^{\hat{\mu}_i}$ on state $[0.5, 0.5]^{\mathsf{T}}$ with the iteration index i. Fig. 4 shows the state trajectories from the initial state $[1, -1]^{\mathsf{T}}$ to the equilibrium under the initial control policy

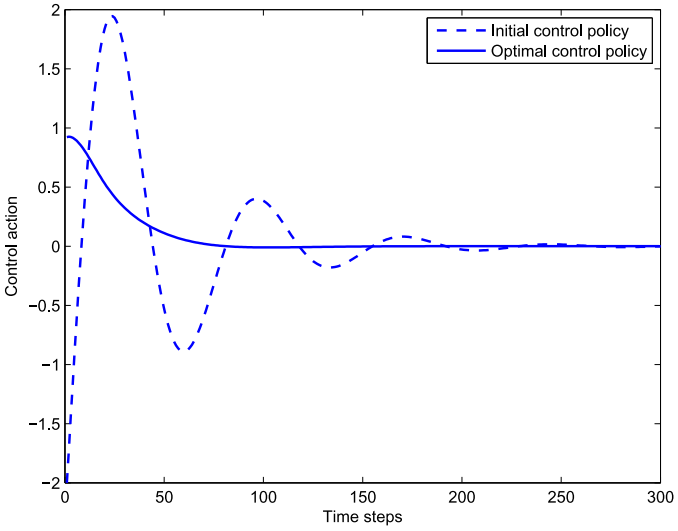

Fig. 5. Action trajectories corresponding to the states from $[1, -1]^{\mathsf{T}}$.

and the approximate optimal control policy obtained by our method, respectively. Fig. 5 shows the action trajectories of the initial control policy and the approximate optimal control policy obtained by our method, respectively.

VI. Conclusion

In this paper, we developed a novel error bound analysis method of Q-function with policy iteration for unknown discounted discrete-time nonlinear systems. A new error condition was given at each iteration, under which the approximate Q-function would converge to a finite neighborhood of the optimal Q-function. Strict mathematical deduction was given to prove the above conclusion. This paper guaranteed that using an approximation structure like neural networks it is possible to solve nonlinear optimal control problems with model-free ADP. Two three-layer neural networks were used to approximate the Q-function and the control policy in the implementation of the developed method. An example was given in the simulation to verify the effectiveness of the developed algorithm.

References

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Belmont, MA, USA: Athena Sci., 2012.

[2] D. Liu, D. Wang, D. Zhao, Q. Wei, and N. Jin, "Neural-network-based optimal control for a class of unknown discrete-time nonlinear systems using globalized dual heuristic programming," *IEEE Trans. Autom. Sci. Eng.*, vol. 9, no. 3, pp. 628–634, Jul. 2012.

[3] R. E. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.

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

[5] 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, USA: Van Nostrand Reinhold, 1992, pp. 493–525.

[6] H. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control: Algorithms and Stability*. London, U.K.: Springer-Verlag, 2012.

[7] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 943–949, Aug. 2008.

[8] D. Zhao, X. Bai, F.-Y. Wang, J. Xu, and W. Yu, "DHP method for ramp metering of freeway traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 990–999, Dec. 2011.

[9] D. Wang and D. Liu, "Neuro-optimal control for a class of unknown nonlinear dynamic systems using SN–DHP technique," *Neurocomputing*, vol. 121, pp. 218–225, Dec. 2013.

[10] G. G. Yen and P. G. DeLima, "Improving the performance of globalized dual heuristic programming for fault tolerant control through an online learning supervisor," *IEEE Trans. Autom. Sci. Eng.*, vol. 2, no. 2, pp. 121–131, Apr. 2005.

[11] C. Liu, X. Xu, and D. Hu, "Multiobjective reinforcement learning: A comprehensive overview," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 45, no. 3, pp. 385–398, Mar. 2015.

[12] D. Liu, X. Yang, D. Wang, and Q. Wei, "Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints," *IEEE Trans. Cybern.*, vol. 45, no. 7, pp. 1372–1385, Jul. 2015.

[13] H. Zhang, J. Zhang, G.-H. Yang, and Y. Luo, "Leader-based optimal coordination control for the consensus problem of multiagent differential games via fuzzy adaptive dynamic programming," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 1, pp. 152–163, Feb. 2015.

[14] T. Huang and D. Liu, "A self-learning scheme for residential energy system control and management," *Neural Comput. Appl.*, vol. 22, no. 2, pp. 259–269, Feb. 2013.

[15] D. Liu, D. Wang, and H. Li, "Decentralized stabilization for a class of continuous-time nonlinear interconnected systems using online learning optimal control approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 2, pp. 418–428, Feb. 2014.

[16] D. Liu, H. Li, and D. Wang, "Online synchronous approximate optimal learning algorithm for multi-player non-zero-sum games with unknown dynamics," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 8, pp. 1015–1027, Aug. 2014.

[17] B. Luo, H.-N. Wu, and T. Huang, "Off-policy reinforcement learning for H_∞ control design," *IEEE Trans. Cybern.*, vol. 45, no. 1, pp. 65–76, Jan. 2015.

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

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

[20] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 621–634, Mar. 2014.

[21] Q. Wei, F.-Y. Wang, D. Liu, and X. Yang, "Finite-approximation-error-based discrete-time iterative adaptive dynamic programming," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2820–2833, Dec. 2014.

[22] Q. Wei and D. Liu, "A novel iterative θ-adaptive dynamic programming for discrete-time nonlinear systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 4, pp. 1176–1190, Oct. 2014.

[23] 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 Trans. Autom. Sci. Eng.*, vol. 11, no. 2, pp. 627–632, Apr. 2014.

[24] J. N. Tsitsiklis, "On the convergence of optimistic policy iteration," *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 59–72, 2002.

[25] T. Dierks and S. Jagannathan, "Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1118–1129, Jul. 2012.

[26] 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.

[27] H. Zhang, L. Cui, and Y. Luo, "Near-optimal control for nonzero-sum differential games of continuous-time nonlinear systems using single-network ADP," *IEEE Trans. Cybern.*, vol. 43, no. 1, pp. 206–216, Feb. 2013.

[28] S. D. Patek, "Partially observed stochastic shortest path problems with approximate solution by neurodynamic programming," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 37, no. 5, pp. 710–720, Sep. 2007.

[29] D. Wang, D. Liu, H. Li, and H. Ma, "Neural-network-based robust optimal control design for a class of uncertain nonlinear systems via adaptive dynamic programming," *Inf. Sci.*, vol. 282, pp. 167–179, Oct. 2014.

[30] B. Luo, H.-N. Wu, and H.-X. Li, "Adaptive optimal control of highly dissipative nonlinear spatially distributed processes with neuro-dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 684–696, Apr. 2015.

[31] 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.

[32] H. Li and D. Liu, "Optimal control for discrete-time affine non-linear systems using general value iteration," *IET Control Theory Appl.*, vol. 6, no. 18, pp. 2725–2736, Dec. 2012.

[33] 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, 2014.

[34] H. Zhang, Y. Luo, and D. Liu, "Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints," *IEEE Trans. Neural Netw.*, vol. 20, no. 9, pp. 1490–1503, Sep. 2009.

[35] X. Yang, D. Liu, and Q. Wei, "Online approximate optimal control for affine non-linear systems with unknown internal dynamics using adaptive dynamic programming," *IET Control Theory Appl.*, vol. 8, no. 16, pp. 1676–1688, Nov. 2014.

[36] B. Luo, H.-N. Wu, T. Huang, and D. Liu, "Reinforcement learning solution for HJB equation arising in constrained optimal control problem," *Neural Netw.*, vol. 71, pp. 150–158, Nov. 2015.

[37] J. Si and Y.-T. Wang, "Online learning control by association and reinforcement," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 264–276, Mar. 2001.

[38] A. Konar, I. G. Chakraborty, S. J. Singh, L. C. Jain, and A. K. Nagar, "A deterministic improved Q-learning for path planning of a mobile robot," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 43, no. 5, pp. 1141–1153, Sep. 2013.

[39] S. Doltsinis, P. Ferreira, and N. Lohse, "An MDP model-based reinforcement learning approach for production station ramp-up optimization: Q-learning analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 9, pp. 1125–1138, Sep. 2014.

[40] B. Luo, T. Huang, H.-N. Wu, and X. Yang, "Data-driven H_∞ control for nonlinear distributed parameter systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 11, pp. 2949–2961, Nov. 2015.

[41] R. Song, W. Xiao, and C. Sun, "A new self-learning optimal control laws for a class of discrete-time nonlinear systems based on ESN architecture," *Sci. China Inf. Sci.*, vol. 57, no. 6, pp. 1–10, 2014.

[42] X. Yang, D. Liu, and D. Wang, "Reinforcement learning for adaptive optimal control of unknown continuous-time nonlinear systems with input constraints," *Int. J. Control*, vol. 87, no. 3, pp. 553–566, Mar. 2014.

[43] Q. Wei and D. Liu, "A novel policy iteration based deterministic Q-learning for discrete-time nonlinear systems," *Sci. China Inf. Sci.*, vol. 58, no. 12, pp. 1–15, 2015.

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

[45] D. Zhao *et al.*, "Full-range adaptive cruise control based on supervised adaptive dynamic programming," *Neurocomputing*, vol. 125, pp. 57–67, Feb. 2014.

[46] D. Zhao, B. Wang, and D. Liu, "A supervised Actor–Critic approach for adaptive cruise control," *Soft Comput.*, vol. 17, no. 11, pp. 2089–2099, 2013.

[47] Z. Dongbin and X. Zhongpu "Adaptive optimal control for the uncertain driving habit problem in adaptive cruise control system," in *Proc. IEEE Int. Conf. Veh. Electron. Safety (ICVES)*, Dongguan, China, 2013, pp. 159–164.

[48] Q. Wei and D. Liu, "Data-driven neuro-optimal temperature control of water–gas shift reaction using stable iterative adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6399–6408, Nov. 2014.

[49] B. Van Roy, "Performance loss bounds for approximate value iteration with state aggregation," *Math. Oper. Res.*, vol. 31, no. 2, pp. 234–244, Feb. 2006.

[50] R. Munos, "Error bounds for approximate value iteration," in *Proc. Nat. Conf. Artif. Intell.*, Pittsburgh, PA, USA, Jul. 2005, pp. 1006–1011.

[51] R. Munos, "Performance bounds in l_p-norm for approximate value iteration," *SIAM J. Control Optim.*, vol. 46, no. 2, pp. 541–561, 2007.

[52] R. Munos, "Error bounds for approximate policy iteration," in *Proc. 20th Int. Conf. Mach. Learn.*, Washington, DC, USA, Aug. 2003, pp. 560–567.

[53] D. P. Bertsekas, "Weighted sup-norm contractions in dynamic programming: A review and some new applications," Dept. Electr. Eng. Comput. Sci., Lab. Inf. Decis. Syst., Tech. Rep. LIDS-P-2884, Cambridge, MA, USA, 2012.
[54] T. J. Perkins and D. Precup, "A convergent form of approximate policy iteration," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2002, pp. 1595–1602.
[55] D. Liu and Q. Wei, "Finite-approximation-error-based optimal control approach for discrete-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 779–789, Apr. 2013.
[56] A. Rantzer, "Relaxed dynamic programming in switching systems," *IEE Proc. Control Theory Appl.*, vol. 153, no. 5, pp. 567–574, Sep. 2006.
[57] L. Grune and A. Rantzer, "On the infinite horizon performance of receding horizon controllers," *IEEE Trans. Autom. Control*, vol. 53, no. 9, pp. 2100–2111, Oct. 2008.
[58] D. Liu, H. Li, and D. Wang, "Error bounds of adaptive dynamic programming algorithms for solving undiscounted optimal control problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1323–1334, Jun. 2015.
[59] D. Liu, D. Wang, F.-Y. Wang, H. Li, and X. Yang, "Neural–network–based online HJB solution for optimal robust guaranteed cost control of continuous-time uncertain nonlinear systems," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2834–2847, Dec. 2014.

Pengfei Yan (S'13) received the B.S. degree in information and computing science from Wuhan University, Wuhan, China, in 2011. He is currently pursuing the Ph.D. degree in the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

He is also with the University of Chinese Academy of Sciences, Beijing. His current research interests include adaptive dynamic programming, data-driven control, adaptive control, and neural networks.

Ding Wang (M'15) received the B.S. degree in mathematics from the Zhengzhou University of Light Industry, Zhengzhou, China, the M.S. degree in operations research and cybernetics from Northeastern University, Shenyang, China, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2007, 2009, and 2012, respectively.

He has been a Visiting Scholar with the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI, USA, since 2015. He is currently an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. He has published over 60 journal and conference papers, and coauthored two monographs. His current research interests include adaptive and learning systems, intelligent control, and neural networks.

Dr. Wang was the Secretariat of the 2014 IEEE World Congress on Computational Intelligence (IEEE WCCI 2014), the Registration Chair of the 5th International Conference on Information Science and Technology (ICIST 2015), and the 4th International Conference on Intelligent Control and Information Processing (ICICIP 2013), and served as the program committee member of several international conferences. He is the Finance Chair of the 12th World Congress on Intelligent Control and Automation (WCICA 2016). He was a recipient of the Excellent Doctoral Dissertation Award of Chinese Academy of Sciences in 2013, and a nomination of the Excellent Doctoral Dissertation Award of Chinese Association of Automation (CAA), in 2014. He serves as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS AND NEUROCOMPUTING. He is a member of Asia-Pacific Neural Network Society, and CAA.

Hongliang Li received the Ph.D. degree in control theory and control engineering from the University of Chinese Academy of Sciences, Beijing, China, in 2015.

He is now with IBM Research—China, Beijing, China. His current research interests include machine learning, deep learning, neural networks, reinforcement learning, and their applications in cognitive commerce.

Derong Liu (S'91–M'94–SM'96–F'05) received the Ph.D. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 1994.

He was a Staff Fellow with General Motors Research and Development Center, Warren, MI, USA, from 1993 to 1995. He was an Assistant Professor with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA, from 1995 to 1999. He joined the University of Illinois at Chicago, Chicago, IL, USA, in 1999, and became a Full Professor of electrical and computer engineering and of computer science, in 2006. He was selected for the "100 Talents Program" by the Chinese Academy of Sciences, in 2008. He has published 14 books (six research monographs and eight edited volumes).

Dr. Liu served as the Editor-in-Chief of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 2010 to 2015. He received the Michael J. Birck Fellowship from the University of Notre Dame, in 1990, the Harvey N. Davis Distinguished Teaching Award from the Stevens Institute of Technology, in 1997, the Faculty Early Career Development CAREER Award from the National Science Foundation, in 1999, the University Scholar Award from University of Illinois from 2006 to 2009, and the Overseas Outstanding Young Scholar Award from the National Natural Science Foundation of China in 2008. He is a Fellow of the International Neural Network Society.