Error Bounds of Adaptive Dynamic Programming Algorithms for Solving Undiscounted Optimal Control Problems

Derong Liu, *Fellow, IEEE*, Hongliang Li, and Ding Wang

***Abstract*— In this paper, we establish error bounds of adaptive dynamic programming algorithms for solving undiscounted infinite-horizon optimal control problems of discrete-time deterministic nonlinear systems. We consider approximation errors in the update equations of both value function and control policy. We utilize a new assumption instead of the contraction assumption in discounted optimal control problems. We establish the error bounds for approximate value iteration based on a new error condition. Furthermore, we also establish the error bounds for approximate policy iteration and approximate optimistic policy iteration algorithms. It is shown that the iterative approximate value function can converge to a finite neighborhood of the optimal value function under some conditions. To implement the developed algorithms, critic and action neural networks are used to approximate the value function and control policy, respectively. Finally, a simulation example is given to demonstrate the effectiveness of the developed algorithms.**

***Index Terms*— Adaptive critic designs, adaptive dynamic programming (ADP), approximate dynamic programming, neural networks, neurodynamic programming, nonlinear systems, optimal control.**

I. Introduction

DYNAMIC programming [1] is a very effective tool in solving the optimal control problem of nonlinear systems, which relies on solving the Hamilton–Jacobi–Bellman (HJB) equation. However, it is often computationally untenable to run dynamic programming to obtain optimal solutions due to the curse of dimensionality [2]. Adaptive dynamic programming (ADP) [3], also known as approximate dynamic programming [4]–[6] or neurodynamic programming [7], has received significantly increasing attention as a learning method for optimizing the policy when interacting with the environment. The ADP techniques have been applied in many practical areas, such as call admission control [8], engine control [9], energy system control [10], temperature control of water gas shift reaction [11], differential games [12]–[14], interconnected systems [15], and multiagent systems [16]. Some comprehensive surveys are given in [3], [17], and [18]. Value iteration and policy iteration are two classes of ADP algorithms to solve optimal control problems of nonlinear systems with continuous state and action spaces. Optimistic policy iteration represents a spectrum of iterative algorithms, which contains the value iteration and policy iteration, and it is also known as generalized policy iteration [19] or modified policy iteration [20].

Value iteration algorithms can solve the optimal control problem of the nonlinear systems without requiring an initial stabilizing control policy. It iterates between value function update and policy improvement until the iterative value function converges to the optimal one. Al-Tamimi *et al.* [21] proved the convergence of the value-iteration-based heuristic dynamic programming algorithm for solving the discrete-time HJB equation. Dierks *et al.* [22] relaxed the need of partial knowledge of the system dynamics by online system identification, and demonstrated the convergence of neural network implementation using Lyapunov theory. In [23] and [24], an iterative ADP algorithm was derived to solve the near-optimal control for discrete-time affine nonlinear systems with control constraints. Wang *et al.* [25] solved the finite-horizon optimal control problem for discrete-time nonlinear systems with unspecified terminal time. Heydari and Balakrishnan [26] derived a value-iteration-based ADP algorithm to solve the fixed-final-time finite-horizon optimal control problem. In [27] and [28], a greedy heuristic dynamic programming algorithm was presented to solve the optimal tracking control problem for a class of discrete-time nonlinear systems. Zhang *et al.* [29] proposed an iterative heuristic dynamic programming algorithm to solve the optimal tracking control problem for a class of nonlinear discrete-time systems with time delays. Liu *et al.* [30] and Wang *et al.* [31] presented an iterative ADP algorithm to solve the optimal control for unknown nonaffine nonlinear discrete-time systems with discount factor in the cost function. The book written by Zhang *et al.* [32] gave a good summary on value-iteration-based iterative ADP methods for solving the optimal control of nonlinear systems. For all the value iteration algorithms mentioned above, it is assumed that the value function and control policy update equations can be exactly solved at each iteration.

In contrast to value iteration, the policy iteration algorithm [33] requires an initial stabilizing control policy.

Manuscript received September 12, 2014; revised December 3, 2014 and February 3, 2015; accepted February 5, 2015. Date of publication March 3, 2015; date of current version May 15, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61034002, Grant 61233001, Grant 61273140, Grant 61304086, and Grant 61374105, and in part by the Beijing Natural Science Foundation under Grant 4132078. The acting Editor-in-Chief who handled the review of this paper was Professor Haibo He.

The authors are with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: derong.liu@ia.ac.cn; hongliang.li@ia.ac.cn; ding.wang@ia.ac.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/TNNLS.2015.2402203

The policy iteration is built to iterate between policy evaluation and policy improvement until it converges to the optimal control. The convergence of policy iteration algorithm for continuous-time nonlinear systems was given in [34]. In [35], the policy iteration algorithm was applied to optimal robust guaranteed cost control of continuous-time uncertain nonlinear systems. Chen and Jagannathan [36] analyzed the convergence of policy iteration algorithm for solving the generalized HJB equation of discrete-time affine nonlinear systems offline. In [37], a policy iteration ADP approach was developed to obtain the optimal control of discrete-time nonlinear systems. Optimistic policy iteration generalizes the value iteration and policy iteration algorithms. Tsitsiklis [38] established the convergence of optimistic policy iteration for discounted finite-state Markov decision problems. Bertsekas [39] gave the convergence of optimistic policy iteration within weighted sup-norm contraction framework. However, it is still an open problem if the convergence results of optimistic policy iteration can be extended to undiscounted optimal control of discrete-time nonlinear systems.

Since most realistic systems have a large or continuous state space, the value function needs to be approximated. Bertsekas [40] and Busoniu *et al.* [41] provided good surveys on approximate value iteration and approximate policy iteration methods with value function approximation. Error bounds of some ADP methods have been established for discounted infinite-horizon optimal control problems formalized by Markov decision processes. Roy [42] established performance loss bounds for approximate value iteration with state aggregation. Munos [43], [44] gave bounds on the error between the performance of policies induced by the approximate value iteration algorithm and the optimal policy as a function of weighted L_p-norms of the approximation errors. Munos [45] also provided error bounds for approximate policy iteration using quadratic norm. Bertsekas [39] established error bounds for approximate policy iteration based on weighted sup-norm contractions. Perkins and Precup [46] studied a model-free form of approximate policy iteration, and proved that the approximate policy iteration algorithm can converge to a unique solution from any initial policy. Li and Si [47] proposed an approximate robust policy iteration using a multilayer perceptron neural network and analyzed the error bounds for approximate value function. Thiery and Scherrer [48] and Scherrer *et al.* [49] proposed three implementations of approximate modified policy iteration and provided error propagation analysis. Bertsekas [39] established error bounds for approximate optimistic policy iteration and extended the result of Thiery and Scherrer [48] and Scherrer *et al.* [49]. In this paper, we will show what will happen for the undiscounted optimal control problem with continuous state and action spaces in the presence of function approximation errors.

In [50], an inequality version of the HJB equation was used to derive bounds on the optimal cost function. For undiscounted discrete-time nonlinear systems, Rantzer [51] introduced a relaxed value iteration scheme to simplify computation based on upper and lower bounds of the optimal cost function, where the distance from optimal values can be kept within prespecified bounds. In [52], the relaxed value iteration scheme was used to solve the optimal switching between linear systems, the optimal control of a linear system with piecewise linear cost, and a partially observable Markov decision problem. In [53], the relaxed value iteration scheme was applied to receding horizon control schemes for discrete-time nonlinear systems. Compared with [51], Liu and Wei [54] and Wei *et al.* [55] presented a convergence analysis for the approximate value iteration algorithm using a new expression of approximation errors at each iteration.

For the optimal control problems with continuous state and action spaces, ADP methods use a critic neural network to approximate the value function and an action neural network to approximate the control policy. Iterating on these approximate models will inevitably give rise to approximation errors. However, the research on ADP methods considering the approximation errors of neural networks is quite sparse for undiscounted infinite-horizon optimal control problems. The main topic of this paper is to understand how the approximation errors at each iteration influence the ADP algorithms for solving undiscounted infinite-horizon optimal control problems of discrete-time deterministic nonlinear systems. Discrete-time deterministic optimal control problem is a major part in the field of optimal control [1], and covers a large class of systems [21]–[32]. We consider approximation errors in both value function and control policy update equations. First, we utilize a new assumption instead of the contraction assumption in discounted optimal control problems. We establish the error bounds for the approximate value iteration algorithm based on a new error condition, which extends a result of Liu and Wei [54]. Then, we establish the error bounds for approximate policy iteration. Furthermore, we prove the convergence of exact optimistic policy iteration by a novel method, and establish the error bounds for approximate optimistic policy iteration. It is shown that the iterative approximate value function can converge to a finite neighborhood of the optimal value function under some conditions. To implement the developed algorithms, two multilayer feedforward neural networks are used to approximate the value function and control policy. Finally, a simulation example is given to demonstrate the effectiveness of the developed algorithms.

The remainder of this paper is organized as follows. Section II provides the problem formulation of undiscounted infinite-horizon optimal control problems of discrete-time nonlinear systems. We establish the error bounds for approximate value iteration, approximate policy iteration, and approximate optimistic policy iteration in Sections III–V, respectively. In Section VI, we propose the neural network implementation of the developed approach. Section VII presents a simulation example to demonstrate the effectiveness of the developed algorithms. Finally, the conclusions are drawn in Section VIII.

The following notations will be used throughout this paper. $\mathbb{R}$ denotes the set of real numbers and $\mathbb{R}^n$ is the n-dimensional Euclidean space. A function $V: \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ is positive definite if: 1) it is continuous and 2) $V(0) = 0$ and $V(x) > 0$ $\forall x \in \Omega - \{0\}$. The notation $\kappa_1 \geq \kappa_2$ means $\kappa_1(s) \geq \kappa_2(s)$, $\forall s \in \mathbb{R}^n$.

II. Problem Formulation

Consider a discrete-time deterministic nonlinear dynamical system described by

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

where $x_k \in \mathbb{R}^n$ is the system state at time k and $u_k \in \mathbb{R}^m$ is the control input. Let x_0 be the initial state. The system function $f(x_k, u_k)$ is Lipschitz continuous on a compact set $\Omega \subseteq \mathbb{R}^n$ containing the origin, and $f(0, 0) = 0$. Hence, $x = 0$ is an equilibrium state of (1) under the control $u = 0$. Assume that (1) is stabilizable on the compact set Ω [21].

Definition 1: A nonlinear dynamical system is defined to be stabilizable on a compact set $\Omega \subseteq \mathbb{R}^n$ if there exists a control input $u \in \mathbb{R}^m$ such that, for all initial conditions $x_0 \in \Omega$, the state $x_k \to 0$ as $k \to \infty$.

Define the undiscounted infinite-horizon cost function as

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

where U is a positive definite utility function. In this paper, the utility function is chosen as the quadratic form $U(x_k, u_k) = x_k^T Q x_k + u_k^T R u_k$, where Q and R are positive definite matrices with suitable dimensions. Our goal is to find a control policy $u_k = u(x_k)$, which can minimize the cost function (2) for every initial state $x_0 \in \Omega$. For the optimal control problem, the designed feedback control must not only stabilize (1) on Ω but also guarantee that the cost function (2) is finite, i.e., the control must be admissible [21].

Definition 2: A control $\mu(x)$ is said to be admissible with respect to the cost function (2) on Ω if $\mu(x)$ is continuous on a compact set $\Omega \subseteq \mathbb{R}^n$, $\mu(0) = 0$, $\mu(x)$ stabilizes (1) on Ω, and, $\forall x_0 \in \Omega$, $J(x_0, \mu)$ is finite.

For any admissible control policy $\mu(x)$, the map from any state x to the value of (2) is called a value function $V^{\mu}(x)$. Then, we define the optimal value function as

$$V^*(x) = \inf_{\mu}\{V^{\mu}(x)\}. \tag{3}$$

According to Bellman's principle of optimality [2], the optimal value function $V^*(x)$ satisfies the discrete-time HJB equation

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

If it can be solved for V^*, the optimal control policy $\mu^*(x)$ can be obtained by

$$\mu^*(x) = \arg\min_{\mu}\{U(x, \mu) + V^*(f(x, \mu))\}. \tag{5}$$

Similar to [56], we define the Hamiltonian function as

$$H(x, u, V) = U(x, u) + V(f(x, u)). \tag{6}$$

Then, we define the mapping $\mathcal{T}_{\mu}$ as

$$(\mathcal{T}_{\mu} V)(x) = H(x, \mu(x), V) \tag{7}$$

and define the mapping $\mathcal{T}$ as

$$(\mathcal{T} V)(x) = \min_{\mu} H(x, \mu(x), V). \tag{8}$$

For convenience, $\mathcal{T}_{\mu}^k$ denotes the composition of mapping $\mathcal{T}_{\mu}$ k times

$$\left(\mathcal{T}_{\mu}^k V\right)(x) = \left(\mathcal{T}_{\mu}\left(\mathcal{T}_{\mu}^{k-1} V\right)\right)(x). \tag{9}$$

Similarly, the mapping $\mathcal{T}^k$ is defined by

$$(\mathcal{T}^k V)(x) = \left(\mathcal{T}\left(\mathcal{T}^{k-1} V\right)\right)(x). \tag{10}$$

Therefore, the discrete-time HJB equation (4) can be written compactly as

$$V^* = \mathcal{T} V^* \tag{11}$$

i.e., V^* is the fixed point of $\mathcal{T}$.

In this paper, we assume the following monotonicity property holds, which was used in [56].

Assumption 1: If $V \leq V'$, then $H(x, u, V) \leq H(x, u, V')$, $\forall x, u$.

Besides the above monotonicity assumption, the contraction assumption in [39] is often required for the discounted optimal control problem. However, for the undiscounted optimal control problem, we utilize the following assumption in [52] instead of the contraction assumption.

Assumption 2: Suppose the condition $0 \leq V^*(f(x, u)) \leq \lambda U(x, u)$ holds uniformly for some $0 < \lambda < \infty$.

The positive constant λ gives a measure on how contractive the optimally controlled system is, i.e., how close the total value function is to the cost of a single step [52].

Equation (4) reduces to the Riccati equation in the linear quadratic regulator case, which can be efficiently solved. In the general nonlinear case, the HJB equation cannot be solved exactly. Some ADP methods using function approximation structures are derived to learn the near-optimal control policy and value function associated with the HJB equation. Because of the approximation errors, the control policy and value function are generally impossible to obtain accurately at each iteration. Therefore, it is necessary to analyze the convergence and to establish the error bounds for ADP algorithms considering function approximation errors.

III. Approximate Value Iteration

Section III-A presents the exact value iteration and is followed by an approximate value iteration algorithm with an analysis of error bounds in Section III-B.

A. Value Iteration

For the value iteration algorithm, it starts with any initial positive definite value function $V_0(x)$ or $V_0(\cdot) = 0$. Then, the control policy $\pi_1(x)$ can be obtained by

$$\pi_1(x) = \arg\min_{u}\{U(x, u) + V_0(f(x, u))\}. \tag{12}$$

For $i = 1, 2, \ldots$, the value iteration algorithm iterates between the value function update

$$\begin{aligned} V_i(x) &= \mathcal{T} V_{i-1}(x) \\ &= \min_{u}\{U(x, u) + V_{i-1}(f(x, u))\} \\ &= U(x, \pi_i(x)) + V_{i-1}(f(x, \pi_i(x))) \end{aligned} \tag{13}$$

and the policy improvement

$$\pi_{i+1}(x) = \arg\min_u \{U(x,u) + V_i(f(x,u))\}. \tag{14}$$

It should satisfy that $V_i(0) = 0$ and $\pi_i(0) = 0$, $\forall i \geq 1$. The value iteration algorithm does not require an initial stabilizing control policy.

According to Assumption 1, it is easy to give the following lemma.

Lemma 1: If $V_0 \geq \mathcal{T}V_0$, the value function sequence $\{V_i\}$ is a monotonically nonincreasing sequence, i.e., $V_i \geq V_{i+1}$, $\forall i \geq 0$. If $V_0 \leq \mathcal{T}V_0$, the value function sequence $\{V_i\}$ is a monotonically nondecreasing sequence, i.e., $V_i \leq V_{i+1}$, $\forall i \geq 0$.

For undiscounted optimal control problems, the convergence of value iteration algorithm has been given in the following theorem [51].

Theorem 1: Let Assumptions 1 and 2 hold. Suppose that $0 \leq \alpha V^* \leq V_0 \leq \beta V^*$, $0 \leq \alpha \leq 1$, and $1 \leq \beta < \infty$. The value function V_i and the control policy π_{i+1} are iteratively updated by (13) and (14). Then, the value function sequence $\{V_i\}$ approaches V^* according to the inequalities

$$\left[1 - \frac{1-\alpha}{(1+\lambda^{-1})^i}\right] V^* \leq V_i \leq \left[1 + \frac{\beta - 1}{(1+\lambda^{-1})^i}\right] V^* \quad i \geq 1. \tag{15}$$

Moreover, the value function V_i converges to V^* uniformly on Ω as $i \to \infty$.

B. Error Bounds for Approximate Value Iteration

For the approximate value iteration algorithm, function approximation structures like neural networks are usually used to approximate the value function V_i and the control policy π_i. In general, the value function and control policy update equations (13) and (14) cannot be solved accurately, because we only have some samples from the state space and there exist approximation errors for function approximation structures. Here, we use $\hat{V}_i$ and $\hat{\pi}_i$ to stand for the approximate expressions of V_i and π_i, respectively. We assume that there exist finite positive constants $\underline{\delta} \leq 1$ and $\overline{\delta} \geq 1$ that make

$$\underline{\delta}\mathcal{T}_{\hat{\pi}_i}\hat{V}_{i-1} \leq \hat{V}_i \leq \overline{\delta}\mathcal{T}_{\hat{\pi}_i}\hat{V}_{i-1} \tag{16}$$

hold uniformly, $\forall i = 1, 2, \ldots$ Similarly, we also assume that there exist finite positive constants $\underline{\sigma} \leq 1$ and $\overline{\sigma} \geq 1$ that make

$$\underline{\sigma}\mathcal{T}\hat{V}_{i-1} \leq \mathcal{T}_{\hat{\pi}_i}\hat{V}_{i-1} \leq \overline{\sigma}\mathcal{T}\hat{V}_{i-1} \tag{17}$$

hold uniformly, $\forall i = 1, 2, \ldots$ Combining (16) and (17), we obtain

$$\underline{\sigma}\ \underline{\delta}\mathcal{T}\hat{V}_{i-1} \leq \hat{V}_i \leq \overline{\sigma}\ \overline{\delta}\mathcal{T}\hat{V}_{i-1}. \tag{18}$$

For simplicity, (18) can be written as

$$\underline{\epsilon}\mathcal{T}\hat{V}_{i-1} \leq \hat{V}_i \leq \overline{\epsilon}\mathcal{T}\hat{V}_{i-1} \tag{19}$$

by denoting

$$\underline{\sigma}\ \underline{\delta} \triangleq \underline{\epsilon}, \quad \overline{\sigma}\ \overline{\delta} \triangleq \overline{\epsilon}. \tag{20}$$

Based on Assumptions 1 and 2, we can establish the error bounds for the approximate value iteration by the following theorem.

Theorem 2: Let Assumptions 1 and 2 hold. Suppose that $0 \leq \alpha V^* \leq V_0 \leq \beta V^*$, $0 \leq \alpha \leq 1$ and $1 \leq \beta < \infty$. The approximate value function $\hat{V}_i$ satisfy the iterative error condition (19). Then, the value function sequence $\{\hat{V}_i\}$ approaches V^* according to the following inequalities:

$$\underline{\epsilon}\left[1 - \sum_{j=1}^{i} \frac{\lambda^j \underline{\epsilon}^{j-1}(1-\underline{\epsilon})}{(\lambda+1)^j} - \frac{\lambda^i \underline{\epsilon}^i (1-\alpha)}{(\lambda+1)^i}\right] V^* \leq \hat{V}_{i+1}$$
$$\leq \overline{\epsilon}\left[1 + \sum_{j=1}^{i} \frac{\lambda^j \overline{\epsilon}^{j-1}(\overline{\epsilon}-1)}{(\lambda+1)^j} + \frac{\lambda^i \overline{\epsilon}^i (\beta-1)}{(\lambda+1)^i}\right] V^* \quad \forall i \geq 0. \tag{21}$$

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

$$\frac{\underline{\epsilon}}{1+\lambda-\underline{\epsilon}\lambda} V^* \leq \lim_{i\to\infty} \hat{V}_i \leq \frac{\overline{\epsilon}}{1+\lambda-\overline{\epsilon}\lambda} V^* \tag{22}$$

under the condition $\overline{\epsilon} < 1/\lambda + 1$.

Proof: First, we prove the lower bound of the approximate value function $\hat{V}_{i+1}$ by mathematical induction. Letting $i = 1$ in (19), we obtain

$$\hat{V}_1 \geq \underline{\epsilon}\mathcal{T}\hat{V}_0 = \underline{\epsilon}\mathcal{T}V_0. \tag{23}$$

Considering $\alpha V^* \leq V_0$ and Assumption 1, we obtain

$$\hat{V}_1 \geq \underline{\epsilon}\mathcal{T}V_0 \geq \alpha\underline{\epsilon}\mathcal{T}V^* = \alpha\underline{\epsilon}V^*. \tag{24}$$

Thus, the lower bound of $\hat{V}_{i+1}$ holds for $i = 0$. According to (19), Assumptions 1 and 2, we obtain

$$\begin{aligned}
\hat{V}_2 &\geq \underline{\epsilon}\mathcal{T}\hat{V}_1 = \underline{\epsilon}\min_u\{U(x,u) + \hat{V}_1(f(x,u))\} \\
&\geq \underline{\epsilon}\min_u\{U(x,u) + \alpha\underline{\epsilon}V^*(f(x,u))\} \\
&\geq \underline{\epsilon}\min_u\left\{\left(1 + \lambda\frac{\alpha\underline{\epsilon}-1}{\lambda+1}\right)U(x,u) \right. \\
&\qquad \left. + \left(\alpha\underline{\epsilon} - \frac{\alpha\underline{\epsilon}-1}{\lambda+1}\right)V^*(f(x,u))\right\} \\
&= \underline{\epsilon}\left(1 + \lambda\frac{\alpha\underline{\epsilon}-1}{\lambda+1}\right)\min_u\{U(x,u) + V^*(f(x,u))\} \\
&= \underline{\epsilon}\left(1 - \frac{\lambda(1-\underline{\epsilon})}{\lambda+1} - \frac{\lambda\underline{\epsilon}(1-\alpha)}{\lambda+1}\right)V^*.
\end{aligned} \tag{25}$$

Hence, the lower bound of $\hat{V}_{i+1}$ holds for $i = 1$. The lower bound of $\hat{V}_{i+1}$ in (21) can be proved by repeating the argument $i+1$ times.

In addition, the upper bound can be proved similarly. Therefore, the lower and upper bounds of $\hat{V}_{i+1}$ in (21) have been proved.

Finally, we prove that the value function sequence $\{\hat{V}_i\}$ converges to a finite neighborhood of V^* uniformly on Ω as $i \to \infty$. Since the sequence $\{\lambda^j \underline{\epsilon}^{j-1}(1-\underline{\epsilon})/(\lambda+1)^j\}$ is a geometric series, we have

$$\sum_{j=1}^{i} \frac{\lambda^j \underline{\epsilon}^{j-1}(1-\underline{\epsilon})}{(\lambda+1)^j} = \frac{\frac{\lambda(1-\underline{\epsilon})}{\lambda+1}\left(1 - \left(\frac{\lambda\underline{\epsilon}}{\lambda+1}\right)^i\right)}{1 - \frac{\lambda\underline{\epsilon}}{\lambda+1}}. \tag{26}$$

Considering $\lambda\underline{\epsilon}/(\lambda+1) < 1$, we have

$$\lim_{i\to\infty} \hat{V}_i \geq \frac{\underline{\epsilon}}{1+\lambda-\underline{\epsilon}\lambda} V^*. \tag{27}$$

For the other part, if $\lambda\overline{\epsilon}/(\lambda+1) < 1$, i.e., $\overline{\epsilon} < 1/\lambda + 1$, we can show that

$$\lim_{i\to\infty} \hat{V}_i \leq \frac{\overline{\epsilon}}{1+\lambda-\overline{\epsilon}\lambda} V^*. \tag{28}$$

Thus, we complete the proof. ■

Remark 1: We can find that the lower and upper bounds in (22) are both monotonically increasing functions of $\underline{\epsilon}$ and $\overline{\epsilon}$, respectively. The condition $\overline{\epsilon} < 1/\lambda + 1$ should satisfy to make the upper bound in (22) be finite and positive. Since $\underline{\epsilon} \leq 1$, the lower bound in (22) is always positive. The values of $\underline{\epsilon}$ and $\overline{\epsilon}$ may gradually refine during the iterative process similar to [57], in which a crude initial operator approximation is gradually refined with new iterations. We can also derive that a larger λ will lead to a slower convergence rate and a larger error bound. In addition, a larger λ also requires more accurate iteration to converge. When $\underline{\epsilon} = \overline{\epsilon} = 1$, the value function sequence $\{\hat{V}_i\}$ converges to V^* uniformly on Ω as $i \to \infty$.

IV. Approximate Policy Iteration

In Section IV-A, we present and analyze the exact policy iteration and establish the error bounds for approximate policy iteration in Section IV-B.

A. Policy Iteration

For the policy iteration algorithm, an initial stabilizing control policy is usually required. In this paper, we start the policy iteration from an initial value function V_0, which satisfies $V_0 \geq \mathcal{T}V_0$. We can see that the obtained control policy $\mu_1(x)$ by

$$\mu_1(x) = \arg\min_u \{U(x,u) + V_0(f(x,u))\} \tag{29}$$

is asymptotically stable for (1), because we have

$$\begin{aligned} V_0(f(x,\mu_1(x))) - V_0(x) &\leq V_0(f(x,\mu_1(x))) - (\mathcal{T}V_0)(x) \\ &= -U(x,\mu_1(x)) \leq 0. \end{aligned}$$

For $i = 1, 2, \ldots$, the policy iteration algorithm iterates between policy evaluation

$$V_{\mu_i}(x) = U(x,\mu_i(x)) + V_{\mu_i}(f(x,\mu_i(x))) \tag{30}$$

and policy improvement

$$\mu_{i+1}(x) = \arg\min_u \{U(x,u) + V_{\mu_i}(f(x,u))\}. \tag{31}$$

When the policy evaluation equation (30) cannot be solved directly, the following iterative process can be used to solve the value function at the policy evaluation step:

$$V_{\mu_i}^{j+1}(x) = U(x,\mu_i(x)) + V_{\mu_i}^{j}(f(x,\mu_i(x))), \quad j \geq 0 \tag{32}$$

with $V_{\mu_i}^0 = V_{\mu_{i-1}}$, $\forall i \geq 1$ and $V_{\mu_1}^0 = V_{\mu_0} = V_0$. Then, we can obtain the following lemma.

Lemma 2: Let Assumption 1 hold. Suppose that $V_0 \geq \mathcal{T}V_0$. Let μ_{i+1} and $V_{\mu_i}^j$ be updated by (31) and (32). Then, the sequence $\{V_{\mu_i}^j\}$ is a monotonically nonincreasing sequence, i.e., $V_{\mu_i}^j \geq V_{\mu_i}^{j+1}$, $\forall i \geq 1$. Moreover, as $j \to \infty$, the limit of $V_{\mu_i}^j$ denoted by $V_{\mu_i}^\infty$ exists, and it is equal to V_{μ_i}, $\forall i \geq 1$.

Proof: We prove the lemma by mathematical induction. Letting $i = 1$ and $j = 0$ in (32), we obtain

$$\begin{aligned} V_{\mu_1}^1(x) &= U(x,\mu_1(x)) + V_{\mu_1}^0(f(x,\mu_1(x))) \\ &= U(x,\mu_1(x)) + V_0(f(x,\mu_1(x))) \\ &= \mathcal{T}V_0 \leq V_0 = V_{\mu_1}^0(x). \end{aligned} \tag{33}$$

According to (32) and Assumption 1, we have

$$V_{\mu_1}^2(x) = \mathcal{T}_{\mu_1} V_{\mu_1}^1(x) \leq \mathcal{T}_{\mu_1} V_{\mu_1}^0(x) = V_{\mu_1}^1(x). \tag{34}$$

Similarly, we can obtain that $V_{\mu_i}^j \geq V_{\mu_i}^{j+1}$ holds for $i = 1$ by induction. Since the sequence $\{V_{\mu_1}^j\}$ is a monotonically nonincreasing sequence and $V_{\mu_1}^j \geq 0$, the limit of $V_{\mu_1}^j$ exists, which is denoted by $V_{\mu_1}^\infty$, and $V_{\mu_1}^j \geq V_{\mu_1}^\infty$. Considering

$$V_{\mu_1}^{j+1}(x) = U(x,\mu_1(x)) + V_{\mu_1}^j(f(x,\mu_1(x))), \quad j \geq 0 \tag{35}$$

we have

$$V_{\mu_1}^{j+1}(x) \geq U(x,\mu_1(x)) + V_{\mu_1}^\infty(f(x,\mu_1(x))), \quad j \geq 0. \tag{36}$$

Letting $j \to \infty$ in (36), we have

$$V_{\mu_1}^\infty(x) \geq U(x,\mu_1(x)) + V_{\mu_1}^\infty(f(x,\mu_1(x))). \tag{37}$$

Similarly, we obtain

$$V_{\mu_1}^\infty(x) \leq U(x,\mu_1(x)) + V_{\mu_1}^j(f(x,\mu_1(x))), \quad j \geq 0. \tag{38}$$

Letting $j \to \infty$ in (38), we can obtain

$$V_{\mu_1}^\infty(x) \leq U(x,\mu_1(x)) + V_{\mu_1}^\infty(f(x,\mu_1(x))). \tag{39}$$

Combining (37) and (39), we have

$$V_{\mu_1}^\infty(x) = U(x,\mu_1(x)) + V_{\mu_1}^\infty(f(x,\mu_1(x))). \tag{40}$$

Considering (30), we obtain that $V_{\mu_i}^\infty(x) = V_{\mu_i}(x)$ holds for $i = 1$.

We assume that it holds for $V_{\mu_i}^j \geq V_{\mu_i}^{j+1}$ and $V_{\mu_i}^\infty(x) = V_{\mu_i}(x)$, $\forall i \geq 1$. Then, considering (31) and (32), we obtain

$$V_{\mu_{i+1}}^1 = \mathcal{T}_{\mu_{i+1}} V_{\mu_{i+1}}^0 = \mathcal{T}_{\mu_{i+1}} V_{\mu_i} \leq V_{\mu_i} = V_{\mu_{i+1}}^0. \tag{41}$$

According to (32) and Assumption 1, we have

$$V_{\mu_{i+1}}^2 = \mathcal{T}_{\mu_{i+1}} V_{\mu_{i+1}}^1 \leq \mathcal{T}_{\mu_{i+1}} V_{\mu_{i+1}}^0 = V_{\mu_{i+1}}^1. \tag{42}$$

Similarly, we can obtain that $V_{\mu_{i+1}}^j \geq V_{\mu_{i+1}}^{j+1}$ holds for $i+1$ by induction, and $V_{\mu_{i+1}}^\infty(x) = V_{\mu_{i+1}}(x)$. Therefore, the proof is completed. ■

Lemma 3: Let Assumption 1 hold. Suppose that $V_0 \geq \mathcal{T}V_0$. Let μ_{i+1} and $V_{\mu_i}^j$ be updated by (31) and (32). Then, the sequence $\{V_{\mu_i}\}$ is a monotonically nonincreasing sequence, i.e., $V_{u_i} \geq V_{u_{i+1}}$, $\forall i \geq 0$.

Proof: According to Lemma 2, we obtain

$$V_{u_{i+1}}^0 \geq V_{u_{i+1}}^\infty = V_{u_{i+1}}. \tag{43}$$

Then, considering

$$V_{u_i} \geq \mathcal{T}V_{u_i} = \mathcal{T}_{\mu_{i+1}} V_{u_i} = \mathcal{T}_{\mu_{i+1}} V_{u_{i+1}}^0 \tag{44}$$

and Assumption 1, we obtain

$$V_{u_i} \geq \mathcal{T}_{\mu_{i+1}} V_{u_{i+1}} = V_{u_{i+1}}. \tag{45}$$

Therefore, the sequence $\{V_{\mu_i}\}$ is a monotonically nonincreasing sequence, $\forall i \geq 0$. ∎

Based on the lemmas above, an extended result in [51, Proposition 5] is given in the following theorem.

Theorem 3: Let Assumptions 1 and 2 hold. Suppose that $V^* \leq V_0 \leq \beta V^*$, $1 \leq \beta < \infty$, and that $V_0 \geq \mathcal{T} V_0$. Let μ_{i+1} and $V_{\mu_i}^j$ be updated by (31) and (32). Then, the value function sequence $\{V_{\mu_i}\}$ approaches V^* according to the inequalities

$$V^* \leq V_{\mu_i} \leq \left[1 + \frac{\beta - 1}{(1+\lambda^{-1})^i}\right] V^* \quad \forall i \geq 1. \tag{46}$$

Proof: First, we prove that $V_{\mu_i} \leq V_i$ holds by induction, $\forall i \geq 1$, where V_i is defined in Section III. According to Lemma 2, we have $V_{\mu_1} \leq \mathcal{T} V_0 = V_1$. Assume it holds for $V_{\mu_i} \leq V_i$, $\forall i \geq 1$. Considering Assumption 1 and Lemma 2, we have

$$V_{u_{i+1}} \leq \mathcal{T} V_{u_{i+1}}^0 = \mathcal{T} V_{u_i} \leq \mathcal{T} V_i = V_{i+1}. \tag{47}$$

Therefore, considering (3) and Theorem 1, we can obtain the conclusion. ∎

B. Error Bounds for Approximate Policy Iteration

Similar to Section III, we use function approximation structures to approximate the value function and control policy. Here, we use $\hat{V}_{\hat{\mu}_i}$ and $\hat{\mu}_i$ to stand for the approximate expressions of V_{μ_i} and μ_i, respectively. We assume that there exist finite positive constants $\underline{\delta} \leq 1$ and $\overline{\delta} \geq 1$ that make

$$\underline{\delta} V_{\hat{\mu}_i} \leq \hat{V}_{\hat{\mu}_i} \leq \overline{\delta} V_{\hat{\mu}_i} \tag{48}$$

hold uniformly, $\forall i = 1, 2, \ldots$, where $V_{\hat{\mu}_i}$ is the exact value function associated with $\hat{\mu}_i$ Considering Lemma 2, we have

$$\hat{V}_{\hat{\mu}_i} \leq \overline{\delta} \mathcal{T}_{\hat{\mu}_i} \hat{V}_{\hat{\mu}_{i-1}}. \tag{49}$$

Similarly, we assume that there exist finite positive constants $\underline{\sigma} \leq 1$ and $\overline{\sigma} \geq 1$ that make

$$\underline{\sigma} \mathcal{T} \hat{V}_{\hat{\mu}_{i-1}} \leq \mathcal{T}_{\hat{\mu}_i} \hat{V}_{\hat{\mu}_{i-1}} \leq \overline{\sigma} \mathcal{T} \hat{V}_{\hat{\mu}_{i-1}} \tag{50}$$

hold uniformly, $\forall i = 1, 2, \ldots$ Combining (49) and (50), we obtain

$$\hat{V}_{\hat{\mu}_i} \leq \overline{\sigma}\ \overline{\delta} \mathcal{T} \hat{V}_{\hat{\mu}_{i-1}}. \tag{51}$$

On the other hand, considering (48), (50), and Assumption 1, we obtain

$$\hat{V}_{\hat{\mu}_i} \geq \underline{\delta} V_{\hat{\mu}_i} \geq \underline{\sigma}\ \underline{\delta} V^*. \tag{52}$$

Therefore, the whole approximation errors in the value function and control policy update equations can be expressed by

$$\underline{\sigma}\ \underline{\delta} V^* \leq \hat{V}_{\hat{\mu}_i} \leq \overline{\sigma}\ \overline{\delta} \mathcal{T} \hat{V}_{\hat{\mu}_{i-1}}. \tag{53}$$

Using the notation in (20), (53) can be written as

$$\underline{\epsilon} V^* \leq \hat{V}_{\hat{\mu}_i} \leq \overline{\epsilon} \mathcal{T} \hat{V}_{\hat{\mu}_{i-1}}. \tag{54}$$

Similar to Section III, we can establish the error bounds for approximate policy iteration by the following theorem.

Theorem 4: Let Assumptions 1 and 2 hold. Suppose that $V^* \leq V_0 \leq \beta V^*$, $1 \leq \beta < \infty$, and that $V_0 \geq \mathcal{T} V_0$. The approximate value function $\hat{V}_{\hat{\mu}_i}$ satisfies the iterative error condition (54). Then, the value function sequence $\{\hat{V}_{\hat{\mu}_i}\}$ approaches V^* according to the following inequalities:

$$\underline{\epsilon} V^* \leq \hat{V}_{\hat{\mu}_{i+1}} \leq \overline{\epsilon} \left[1 + \sum_{j=1}^{i} \frac{\lambda^j \overline{\epsilon}^{j-1}(\overline{\epsilon} - 1)}{(\lambda+1)^j} + \frac{\lambda^i \overline{\epsilon}^i (\beta - 1)}{(\lambda+1)^i}\right] V^*, \quad i \geq 0.$$

Moreover, the approximate value function sequence $\{\hat{V}_{\mu_i}\}$ converges to a finite neighborhood of V^* uniformly on Ω as $i \to \infty$

$$\underline{\epsilon} V^* \leq \lim_{i \to \infty} \hat{V}_{\hat{\mu}_i} \leq \frac{\overline{\epsilon}}{1 + \lambda - \overline{\epsilon}\lambda} V^* \tag{55}$$

under the condition $\overline{\epsilon} < 1/\lambda + 1$.

V. Approximate Optimistic Policy Iteration

In Section V-A, we prove the convergence of exact optimistic policy iteration, and establish the error bounds for approximate optimistic policy iteration in Section V-B.

A. Optimistic Policy Iteration

According to Lemma 2, we can see that the policy evaluation can be obtained as $j \to \infty$ in (32). However, this process will usually take a long time to converge. To avoid this problem, the optimistic policy iteration algorithm only makes finite iterations in (32).

For the optimistic policy iteration algorithm, we start the iteration from an initial value function V_0, which satisfies $V_0 \geq \mathcal{T} V_0$. The initial control policy $\nu_1 = \mu_1$ can be obtained by (29). For $i = 1, 2, \ldots$, and for any positive integer n_i, the optimistic policy iteration algorithm updates the value function $V_{\nu_i}(x) = V_{\nu_i}^{n_i}(x)$ by

$$V_{\nu_i}^{j+1}(x) = U(x, \nu_i(x)) + V_{\nu_i}^j(f(x, \nu_i(x))), \quad 0 \leq j \leq n_i - 1 \tag{56}$$

where $V_{\nu_i}^0 = V_{\nu_{i-1}}$, $\forall i \geq 1$, and $V_{\nu_1}^0 = V_{\nu_0} = V_0$. Using the definition in (9), the value function $V_{\nu_i}(x)$ can be expressed by $V_{\nu_i}(x) = \mathcal{T}_{\nu_i}^{n_i} V_{\nu_{i-1}}(x)$. The optimistic policy iteration algorithm updates the control policy by

$$\nu_{i+1}(x) = \arg\min_u \{U(x, u) + V_{\nu_i}(f(x, u))\}. \tag{57}$$

The optimistic policy iteration becomes value iteration as $n_i = 1$, and becomes the policy iteration as $n_i \to \infty$. For policy iteration, it solves the value function associated with the current control policy at each iteration, while it takes only one iteration toward that value function for value iteration. However, the value function update in (56) has to stop before $j \to \infty$ in practical implementations. Next, we will show the monotonicity property of value function, which is given in [56], and then establish the convergence property of optimistic policy iteration by a novel method.

Lemma 4: Let Assumption 1 hold. Suppose that $V_0 \geq \mathcal{T} V_0$. Let V_{ν_i} and ν_{i+1} be updated by (56) and (57). Then, the value

function sequence $\{V_{\nu_i}\}$ is a monotonically nonincreasing sequence, i.e., $V_{\nu_i} \geq V_{\nu_{i+1}}$, $\forall i \geq 0$.

Proof: According to Assumption 1 and Lemma 2, we have

$$V_{\nu_0} = V_0 \geq \mathcal{T}V_0 = \mathcal{T}_{\nu_1}V_0 \geq \mathcal{T}_{\nu_1}^{n_1}V_0 = V_{\nu_1}. \tag{58}$$

Thus, it holds for $i = 0$. Similarly, we obtain

$$V_{\nu_1} \geq \mathcal{T}_{\nu_1}^{n_1+1}V_0 = \mathcal{T}_{\nu_1}V_{\nu_1} \geq \mathcal{T}V_{\nu_1} = \mathcal{T}_{\nu_2}V_{\nu_1} \geq V_{\nu_2}. \tag{59}$$

Therefore, the conclusion can be proved by induction. ∎

Theorem 5: Let Assumptions 1 and 2 hold. Suppose that $V^* \leq V_0 \leq \beta V^*$, $1 \leq \beta < \infty$, and that $V_0 \geq \mathcal{T}V_0$. The value function V_{ν_i} and the control policy ν_{i+1} are updated by (56) and (57). Then, the value function sequence $\{V_{\nu_i}\}$ approaches V^* according to the inequalities

$$V^* \leq V_{\nu_i} \leq \left[1 + \frac{\beta - 1}{(1+\lambda^{-1})^i}\right] V^* \quad \forall i \geq 1. \tag{60}$$

Moreover, the value function V_{ν_i} converges to V^* uniformly on Ω.

Proof: First, we prove that $V_{\nu_i} \leq V_i$ holds by mathematical induction, $\forall i \geq 1$, where V_i is defined in Section III. According to Lemma 2, we have

$$V_{\nu_1} = V_{\nu_1}^{n_1} \leq V_{\nu_1}^1 = \mathcal{T}V_0 = V_1. \tag{61}$$

Thus, it holds for $i = 1$. Assume that it holds for $i \geq 1$, i.e., $V_{\nu_i} \leq V_i$. According to Lemma 2, we have

$$V_{\nu_{i+1}} = V_{\nu_{i+1}}^{n_i} \leq V_{\nu_{i+1}}^1 = \mathcal{T}V_{\nu_i}. \tag{62}$$

Considering Assumption 1, we obtain

$$\mathcal{T}V_{\nu_i} \leq \mathcal{T}V_i = V_{i+1}. \tag{63}$$

Thus, we can obtain $V_{\nu_{i+1}} \leq V_{i+1}$. Then, it can also be proved that $V_{\nu_i} \geq V^*$ by mathematical induction, $\forall i \geq 1$. Therefore, considering Theorem 1, we obtain

$$V^* \leq V_{\nu_i} \leq V_i \leq \left[1 + \frac{\beta - 1}{(1+\lambda^{-1})^i}\right] V^*. \tag{64}$$

As $i \to \infty$, the value function V_{ν_i} converges to V^* uniformly on Ω. ∎

Remark 2: Although it only makes finite iterations in (56), the convergence of optimistic policy iteration can still be guaranteed. Moreover, we can find that the convergence rate of optimistic policy iteration is determined by $\lambda/(\lambda+1)$, while for the discounted optimal control problem, the convergence rate is determined by the discount factor [39]. The optimistic policy iteration algorithm has faster convergence rate than the value iteration and requires less computation in value function update than policy iteration.

B. Error Bounds for Approximate Optimistic Policy Iteration

Here, we use $\hat{V}_{\hat{\nu}_i}$ and $\hat{\nu}_i$ to stand for the approximate expressions of V_{ν_i} and ν_i, respectively. Without loss of generality, we let n_i be a constant integer K for all iteration steps.

We assume that there exist finite positive constants $\underline{\delta} \leq 1$ and $\overline{\delta} \geq 1$ that make

$$\underline{\delta}\mathcal{T}_{\hat{\nu}_i}^K \hat{V}_{\hat{\nu}_{i-1}} \leq \hat{V}_{\hat{\nu}_i} \leq \overline{\delta}\mathcal{T}_{\hat{\nu}_i}^K \hat{V}_{\hat{\nu}_{i-1}} \tag{65}$$

hold uniformly, $\forall i = 1, 2, \ldots$ Considering Lemma 2, we have

$$\hat{V}_{\hat{\nu}_i} \leq \overline{\delta}\mathcal{T}_{\hat{\nu}_i}^K \hat{V}_{\hat{\nu}_{i-1}} \leq \overline{\delta}\mathcal{T}_{\hat{\nu}_i} \hat{V}_{\hat{\nu}_{i-1}}. \tag{66}$$

Similarly, we assume that there exist finite positive constants $\underline{\sigma} \leq 1$ and $\overline{\sigma} \geq 1$ that make

$$\underline{\sigma}\mathcal{T}\hat{V}_{\hat{\nu}_{i-1}} \leq \mathcal{T}_{\hat{\nu}_i}\hat{V}_{\hat{\nu}_{i-1}} \leq \overline{\sigma}\mathcal{T}\hat{V}_{\hat{\nu}_{i-1}} \tag{67}$$

hold uniformly, $\forall i = 1, 2, \ldots$ Combining (66) and (67), we obtain

$$\hat{V}_{\hat{\nu}_i} \leq \overline{\sigma}\ \overline{\delta}\mathcal{T}\hat{V}_{\hat{\nu}_{i-1}}. \tag{68}$$

On the other hand, considering (65), (67), and Assumption 1, we obtain

$$\begin{aligned}\hat{V}_{\hat{\nu}_i} &\geq \underline{\delta}\mathcal{T}_{\hat{\nu}_i}^K \hat{V}_{\hat{\nu}_{i-1}} \geq \underline{\sigma}\ \underline{\delta}\mathcal{T}\left(\mathcal{T}_{\hat{\nu}_i}^{K-1}\hat{V}_{\hat{\nu}_{i-1}}\right) \\ &\geq \cdots \geq \underline{\sigma}\ \underline{\delta}\mathcal{T}^K \hat{V}_{\hat{\nu}_{i-1}}.\end{aligned} \tag{69}$$

Therefore, the whole approximation errors in the value function and control policy update equations can be expressed by

$$\underline{\sigma}\ \underline{\delta}\mathcal{T}^K \hat{V}_{\hat{\nu}_{i-1}} \leq \hat{V}_{\hat{\nu}_i} \leq \overline{\sigma}\ \overline{\delta}\mathcal{T}\hat{V}_{\hat{\nu}_{i-1}}. \tag{70}$$

Using the notation in (20), (70) can be written as

$$\underline{\epsilon}\mathcal{T}^K \hat{V}_{\hat{\nu}_{i-1}} \leq \hat{V}_{\hat{\nu}_i} \leq \overline{\epsilon}\mathcal{T}\hat{V}_{\hat{\nu}_{i-1}}. \tag{71}$$

Then, we can establish the error bounds for the approximate optimistic policy iteration by the following theorem.

Theorem 6: Let Assumptions 1 and 2 hold. Suppose that $V^* \leq V_0 \leq \beta V^*$, $1 \leq \beta < \infty$, and that $V_0 \geq \mathcal{T}V_0$. The approximate value function $\hat{V}_{\hat{\nu}_i}$ satisfies the iterative error condition (71). Then, the value function sequence $\{\hat{V}_{\hat{\nu}_i}\}$ approaches V^* according to the following inequalities:

$$\begin{aligned}&\underline{\epsilon}\left[1 - \sum_{j=1}^{i} \frac{\lambda^j \underline{\epsilon}^{j-1}(1-\underline{\epsilon})}{(\lambda+1)^j}\right] V^* \leq \hat{V}_{\hat{\nu}_{i+1}} \\ &\leq \overline{\epsilon}\left[1 + \sum_{j=1}^{i} \frac{\lambda^j \overline{\epsilon}^{j-1}(\overline{\epsilon}-1)}{(\lambda+1)^j} + \frac{\lambda^i \overline{\epsilon}^i(\beta-1)}{(\lambda+1)^i}\right] V^*, \quad i \geq 0.\end{aligned} \tag{72}$$

Moreover, the approximate value function sequence $\{\hat{V}_{\hat{\nu}_i}\}$ converges to a finite neighborhood of V^* uniformly on Ω as $i \to \infty$

$$\frac{\underline{\epsilon}}{1+\lambda-\underline{\epsilon}\lambda}V^* \leq \lim_{i\to\infty} \hat{V}_{\hat{\nu}_i} \leq \frac{\overline{\epsilon}}{1+\lambda-\overline{\epsilon}\lambda}V^* \tag{73}$$

under the condition $\overline{\epsilon} < 1/\lambda + 1$.

Proof: First, we prove the lower bound in (72). According to (71) and Assumption 1, we have

$$\hat{V}_{\hat{\nu}_1} \geq \underline{\epsilon}\mathcal{T}^K V_0 \geq \underline{\epsilon}\mathcal{T}^K V^* = \underline{\epsilon}V^* \tag{74}$$

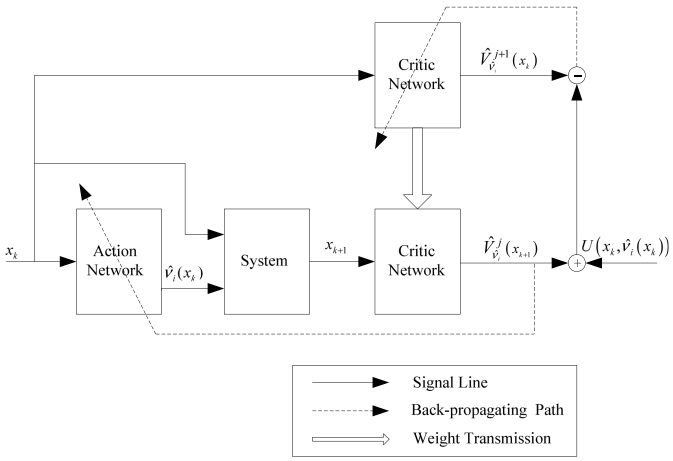

Fig. 1. Structure diagram of approximate optimistic policy iteration.

i.e., it holds for $i = 0$. Similarly, we obtain

$$\begin{aligned}\hat{V}_{\hat{v}_2} \geq \underline{\epsilon}\mathcal{T}^K \hat{V}_{\hat{v}_1} &= \underline{\epsilon} \min_u \left\{U(x,u) + \mathcal{T}^{K-1}\hat{V}_{\hat{v}_1}(f(x,u))\right\} \\ &\geq \underline{\epsilon} \min_u \left\{U(x,u) + \underline{\epsilon}\mathcal{T}^{K-1}V^*(f(x,u))\right\} \\ &= \underline{\epsilon} \min_u \{U(x,u) + \underline{\epsilon}V^*(f(x,u))\} \\ &\geq \underline{\epsilon} \min_u \left\{\left(1 - \lambda\frac{1-\underline{\epsilon}}{\lambda+1}\right)U(x,u) \right. \\ &\qquad \left. + \left(\underline{\epsilon} + \frac{1-\underline{\epsilon}}{\lambda+1}\right)V^*(f(x,u))\right\} \\ &= \underline{\epsilon}\left(1 - \frac{\lambda(1-\underline{\epsilon})}{\lambda+1}\right)V^*\end{aligned}$$

i.e., it holds for $i = 1$. Therefore, we can obtain the lower bound by induction. Similar to Theorem 2, we can obtain the upper bound in (72) and the conclusion in (73). ■

VI. Neural Network Implementation for Optimal Control

We have just proved that the approximate value iteration, approximate policy iteration, and approximate optimistic policy iteration algorithms can converge to a finite neighborhood of the optimal value function associated with the HJB equation. It should be mentioned that we consider approximation errors in both value function and control policy update equations at each iteration. This makes it feasible to use neural network approximation for solving undiscounted optimal control problems of nonlinear systems. It should be mentioned that kernel methods [58] and linear parametric architectures with learned basis functions [59]–[61] can also be applied. Since the optimistic policy iteration contains the value iteration and policy iteration, we only present a detailed implementation of the approximate optimistic policy iteration using neural networks in this section. The neural network implementation of approximate value iteration can be found in [54] and [62].

The whole structure diagram of the approximate optimistic policy iteration is shown in Fig. 1, where two multilayer feedforward neural networks are used. The critic neural network is used to approximate the value function, and the action neural network is used to approximate the control policy.

A neural network can be used to approximate some smooth function on a prescribed compact set. The value function $V_{v_i}^{j+1}(x_k)$ in (56) is approximated by the critic neural network

$$\hat{V}_{\hat{v}_i}^{j+1}(x_k) = \left(W_{c(i)}^{j+1}\right)^T \phi\left(\left(Y_{c(i)}^{j+1}\right)^T x_k\right) \tag{75}$$

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

$$V_{\hat{v}_i}^{j+1}(x_k) = U(x_k, \hat{v}_i(x_k)) + \hat{V}_{\hat{v}_i}^{j}(x_{k+1}) \tag{76}$$

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

$$e_{c(i)}^{j+1}(x_k) = \hat{V}_{\hat{v}_i}^{j+1}(x_k) - V_{\hat{v}_i}^{j+1}(x_k) \tag{77}$$

and the performance function to be minimized is defined by

$$E_{c(i)}^{j+1}(x_k) = \frac{1}{2}\left(e_{c(i)}^{j+1}(x_k)\right)^2. \tag{78}$$

The control policy $v_{i+1}(x_k)$ in (57) is approximated by the action neural network

$$\hat{v}_{i+1}(x_k) = W_{a(i+1)}^T \phi\left(Y_{a(i+1)}^T x_k\right). \tag{79}$$

The target function of the action neural network is defined by

$$d_{i+1}(x_k) = \arg\min_{u_k} \left\{U(x_k, u_k) + \hat{V}_{\hat{v}_i}(x_{k+1})\right\}. \tag{80}$$

Then, the error function for training the action neural network is given by

$$e_{a(i+1)}(x_k) = \hat{v}_{i+1}(x_k) - d_{i+1}(x_k). \tag{81}$$

The weights of the action neural network are updated to minimize the following performance function:

$$E_{a(i+1)}(x_k) = \frac{1}{2}(e_{a(i+1)}(x_k))^T e_{a(i+1)}(x_k). \tag{82}$$

We use the gradient descent method to tune the weights of neural networks on a training set constructed from the compact set Ω. The details of this tuning method can be found in [54]. Some other tuning methods can also be used, such as Newton's method and the Levenberg–Marquardt method [62], in order to increase the convergence rate of neural network training.

A detailed process of the approximate optimistic policy iteration is given in Algorithm 1, where the approximate value iteration can be regarded as a special case. If we have an initial stabilizing control policy, the algorithm can iterate between Step 4 and Step 5 directly. It should be mentioned that Algorithm 1 runs in an offline manner. Note that it can also be implemented online but a persistence of excitation condition is usually required.

VII. Simulation Study

In this section, we provide a simulation example to demonstrate the effectiveness of the algorithms developed in

Algorithm 1 Approximate Optimistic Policy Iteration

Step 1. Initialization:
Initialize critic and action neural networks;
Select an initial value function V_0 satisfying $V_0 \geq \mathcal{T}V_0$;
Set policy evaluation steps K and maximum number of iteration steps $i_{\max}$.

Step 2. Set $i = 0$. Update the control policy $\hat{v}_1(x_k)$ by minimizing (82) on a training set $\{x_k\}$ randomly selected from the compact set $\mathbf{\Omega}$.

Step 3. Set $i = 1$.

Step 4. For $j = 0, 1, \ldots, K-1$, update the value function $\hat{V}_{\hat{v}_i}^{j+1}(x_k)$ by minimizing (78) on a training set $\{x_k\}$ randomly selected from the compact set $\mathbf{\Omega}$. After convergence, set $\hat{V}_{\hat{v}_i}(x_k) = \hat{V}_{\hat{v}_i}^{K}(x_k)$.

Step 5. Update the control policy $\hat{v}_{i+1}(x_k)$ by minimizing (82) on a training set $\{x_k\}$ randomly selected from the compact set $\mathbf{\Omega}$.

Step 6. Set $i \leftarrow i + 1$.

Step 7. Repeat Steps 4–6 until the convergence conditions are met.

Step 8. Obtain the approximate optimal control policy $\hat{v}_i(x_k)$.

this paper. Consider the following discrete-time nonlinear system $x_{k+1} = h(x_k) + g(x_k)u_k$:

$$h(x_k) = \begin{bmatrix} 0.9x_{1k} + 0.1x_{2k} \\ -0.05(x_{1k} + x_{2k}(1 - (\cos(2x_{1k}) + 2)^2)) + x_{2k} \end{bmatrix}$$
$$g(x_k) = \begin{bmatrix} 0 \\ 0.1\cos(2x_{1k}) + 0.2 \end{bmatrix} \tag{83}$$

$x_k = [x_{1k} \quad x_{2k}]^T \in \mathbb{R}^2$, and $u_k \in \mathbb{R}$, $k = 0, 1, \ldots$. Define the cost function as

$$J(x_0, u) = \sum_{k=0}^{\infty} \left(x_k^T Q x_k + u_k^T R u_k \right) \tag{84}$$

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

To implement the developed algorithms, we choose three-layer feedforward neural networks as function approximation structures. The structures of the critic and action neural networks are both chosen as 2–8–1. The initial weights of the critic and action neural networks are chosen randomly in $[-0.1, 0.1]$. The maximum number of iteration steps is selected as $i_{\max} = 20$. The compact set $\mathbf{\Omega}$ or the operation region of the system is selected as $-1 \leq x_1 \leq 1$ and $-1 \leq x_2 \leq 1$. The training set $\{x_k\}$ is constructed by randomly choosing 1000 samples from the compact set $\mathbf{\Omega}$ at each iteration.

The initial value function is selected as $V_0 = 2x_{1k}^2 + 2x_{2k}^2$. According to Fig. 2, it can be observed that $V_0 \geq V_1$ holds for all states in the compact set $\mathbf{\Omega}$. We implement Algorithm 1 by letting $K = 1$, $K = 3$, and $K = 10$, respectively. For the initial state $x_0 = [1, -1]^T$, the convergence curve of the value function sequence $\{\hat{V}_{\hat{v}_1}^j\}$ is shown in Fig. 3. We can see that $\{\hat{V}_{\hat{v}_1}^j\}$ is a monotonically nonincreasing sequence, and it is basically convergent at $K = 10$. Thus,

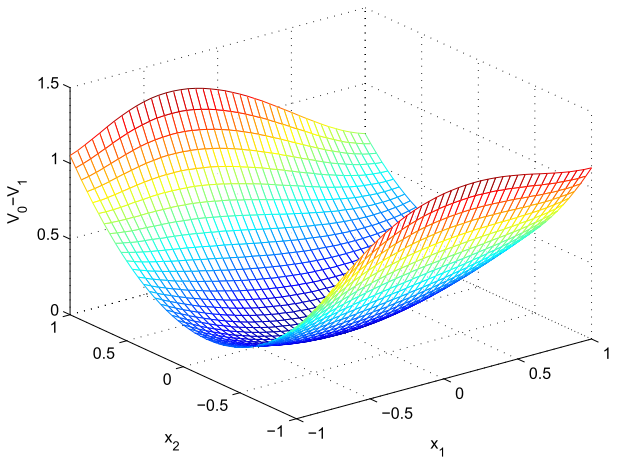

Fig. 2. 3-D plot of $V_0 - V_1$ in the compact set $\mathbf{\Omega}$.

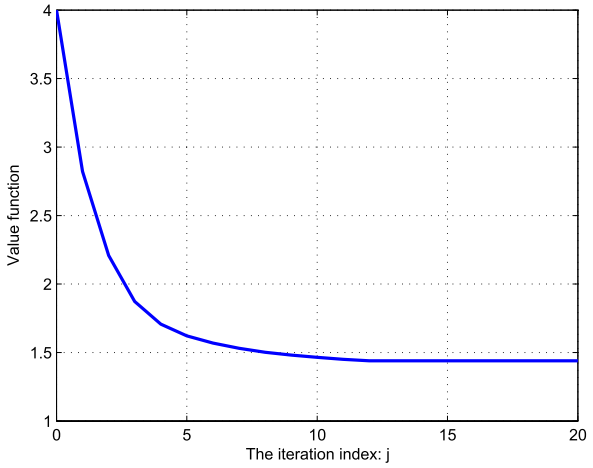

Fig. 3. Convergence curve of the value function $\hat{V}_{\hat{v}_1}^j$ at x_0.

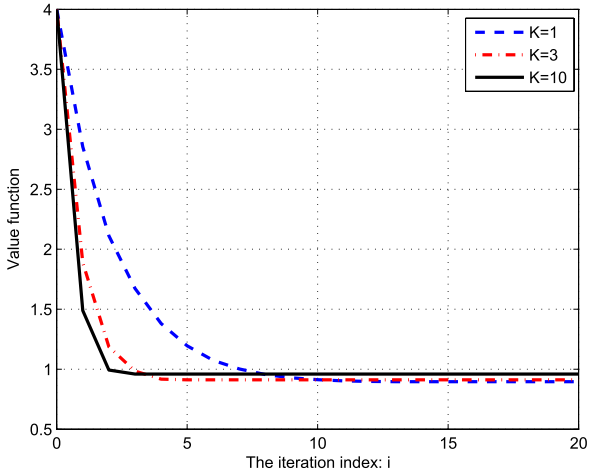

Fig. 4. Convergence curves of the value function $\hat{V}_{\hat{v}_i}$ at x_0 when $K = 1$, $K = 3$, and $K = 10$.

the algorithm for $K = 10$ can be regarded as the approximate policy iteration. The algorithms for $K = 1$ and $K = 3$ are the approximate value iteration and approximate optimistic policy iteration, respectively. After implementing the algorithms for $i_{\max} = 20$, the convergence curves of the value functions $\hat{V}_{\hat{v}_i}$ at x_0 are shown in Fig. 4. It can be observed that all

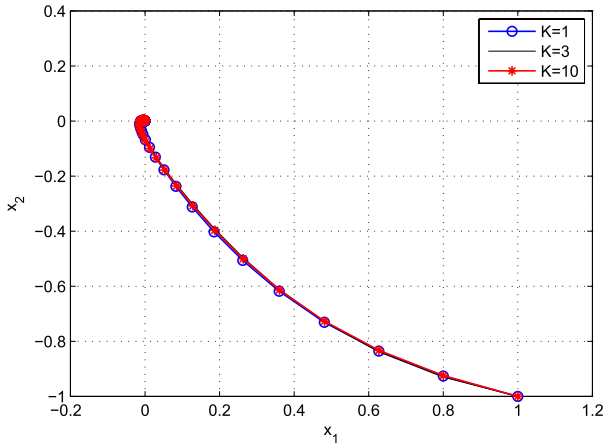

Fig. 5. State trajectories.

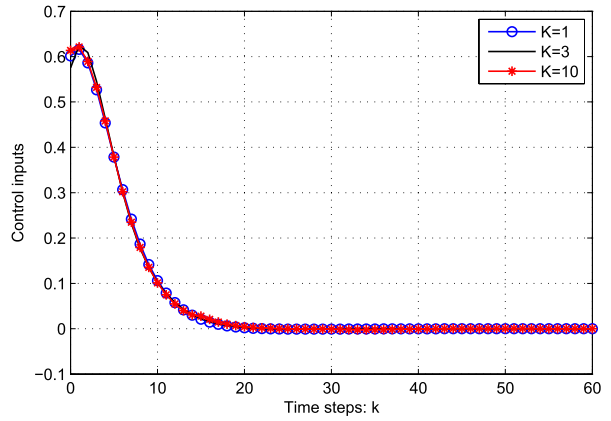

Fig. 6. Control inputs.

the value functions are basically convergent with the iteration index $i > 10$, and the obtained approximate optimal value functions at $i = 20$ are quite close. Although there exist approximation errors in both value function and control policy update steps, the approximate value function can converge to a finite neighborhood of the optimal value function.

Finally, we apply the obtained approximate optimal control policies $\hat{v}_{20}$ by the algorithms to (83) for 60 time steps. The corresponding state trajectories are displayed in Fig. 5, and the control inputs are displayed in Fig. 6. Examining the results, it is observed that all the control policies obtain very good performance, and the differences between the three trajectories are quite small.

VIII. Conclusion

In this paper, we established error bounds for value iteration, policy iteration, and optimistic policy iteration for undiscounted discrete-time nonlinear systems by defining a new error condition at each iteration. We considered approximation errors in both value function and control policy update equations. It was shown that the iterative approximate value function converges to a finite neighborhood of the optimal value function under some mild conditions. The results provided theoretical guarantees for using neural network approximation for solving undiscounted optimal control problems. To implement the developed algorithms, the critic and action neural networks were used to approximate the value function and the control policy, respectively. A simulation example was given to demonstrate the effectiveness of the developed algorithms. It should be mentioned that the system model is assumed to be known *a priori*. It will be desirable to extend the developed results to unknown nonlinear systems.

References

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

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

[3] 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.

[4] 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, 1992, ch. 13.

[5] J. Si, A. G. Barto, W. B. Powell, and D. Wunsch, Eds., *Handbook of Learning and Approximate Dynamic Programming*. New York, NY, USA: IEEE Press, 2004.

[6] F. L. Lewis and D. Liu, Eds., *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. Hoboken, NJ, USA: Wiley, 2013.

[7] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.

[8] D. Liu, Y. Zhang, and H. Zhang, "A self-learning call admission control scheme for CDMA cellular networks," *IEEE Trans. Neural Netw.*, vol. 16, no. 5, pp. 1219–1228, Sep. 2005.

[9] D. Liu, H. Javaherian, O. Kovalenko, and T. Huang, "Adaptive critic learning techniques for engine torque and air–fuel ratio control," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 988–993, Aug. 2008.

[10] 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.

[11] 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.

[12] 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, Jun. 2013.

[13] 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.

[14] 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.

[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] 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.

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

[18] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Circuits Syst. Mag.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

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

[20] M. L. Puterman and M. C. Shin, "Modifed policy iteration algorithms for discounted Markov decision problems," *Manage. Sci.*, vol. 24, no. 11, pp. 1127–1137, 1978.

[21] 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.

[22] T. Dierks, B. T. Thumati, and S. Jagannathan, "Optimal control of unknown affine nonlinear discrete-time systems using offline-trained neural networks with proof of convergence," *Neural Netw.*, vol. 22, nos. 5–6, pp. 851–860, Jul./Aug. 2009.

[23] 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.

[24] D. Liu, D. Wang, and X. Yang, "An iterative adaptive dynamic programming algorithm for optimal control of unknown discrete-time nonlinear systems with constrained inputs," *Inf. Sci.*, vol. 220, pp. 331–342, Jan. 2013.

[25] F.-Y. Wang, N. Jin, D. Liu, and Q. Wei, "Adaptive dynamic programming for finite-horizon optimal control of discrete-time nonlinear systems with ε-error bound," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 24–36, Jan. 2011.

[26] A. Heydari and S. N. Balakrishnan, "Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 145–157, Jan. 2013.

[27] H. Zhang, Q. Wei, and Y. Luo, "A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 937–942, Aug. 2008.

[28] 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.

[29] H. Zhang, R. Song, Q. Wei, and T. Zhang, "Optimal tracking control for a class of nonlinear discrete-time systems with time delays based on heuristic dynamic programming," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 1851–1862, Dec. 2011.

[30] 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.

[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. Zhang, D. Liu, Y. Luo, and D. Wang, *Adaptive Dynamic Programming for Control: Algorithms and Stability*. London, U.K.: Springer-Verlag, 2013.

[33] R. A. Howard, *Dynamic Programming and Markov Processes*. Cambridge, MA, USA: MIT Press, 1960.

[34] 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.

[35] 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.

[36] Z. Chen and S. Jagannathan, "Generalized Hamilton–Jacobi–Bellman formulation-based neural network control of affine nonlinear discrete-time systems," *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 90–106, Jan. 2008.

[37] 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.

[38] J. N. Tsitsiklis, "On the convergence of optimistic policy iteration," *J. Mach. Learn. Res.*, vol. 3, pp. 59–72, Jul. 2002.

[39] D. P. Bertsekas, "Weighted sup-norm contractions in dynamic programming: A review and some new applications," Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. LIDS-P-2884, May 2012.

[40] D. P. Bertsekas, "Approximate policy iteration: A survey and some new methods," *J. Control Theory Appl.*, vol. 9, no. 3, pp. 310–335, 2011.

[41] L. Busoniu, D. Ernst, B. De Schutter, and R. Babuska, "Approximate reinforcement learning: An overview," in *Proc. IEEE Symp. Adapt. Dyn. Program. Reinforcement Learn.*, Paris, France, Apr. 2011, pp. 1–8.

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

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

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

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

[46] T. J. Perkins and D. Precup, "A convergent form of approximate policy iteration," in *Advances in Neural Information Processing Systems*. Vancouver, BC, Canada: MIT Press, 2002.

[47] B. Li and J. Si, "Approximate robust policy iteration using multilayer perceptron neural networks for discounted infinite-horizon Markov decision processes with uncertain correlated transition matrices," *IEEE Trans. Neural Netw.*, vol. 21, no. 8, pp. 1270–1280, Aug. 2010.

[48] C. Thiery and B. Scherrer, "Performance bound for approximate optimistic policy iteration," INRIA, Rocquencourt, France, Tech. Rep. INRIA-00480952, 2010.

[49] B. Scherrer, V. Gabillon, M. Ghavamzadeh, and M. Geist, "Approximate modified policy iteration," in *Proc. 29th Int. Conf. Mach. Learn.*, Edinburgh, Scotland, Jun./Jul. 2012, pp. 1207–1214.

[50] R. J. Leake and R.-W. Liu, "Construction of suboptimal control sequences," *SIAM J. Control*, vol. 5, no. 1, pp. 54–63, 1967.

[51] A. Rantzer, "Relaxed dynamic programming in switching systems," *IEE Proc. Control Theory Appl.*, vol. 153, no. 5, pp. 567–574, Sep. 2006.

[52] B. Lincoln and A. Rantzer, "Relaxing dynamic programming," *IEEE Trans. Autom. Control*, vol. 51, no. 8, pp. 1249–1260, Aug. 2006.

[53] 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.

[54] 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.

[55] 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.

[56] D. P. Bertsekas, *Abstract Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 2013.

[57] A. Almudevar and E. F. de Arruda, "Optimal approximation schedules for a class of iterative algorithms, with an application to multigrid value iteration," *IEEE Trans. Autom. Control*, vol. 57, no. 12, pp. 3132–3146, Dec. 2012.

[58] X. Xu, D. Hu, and X. Lu, "Kernel-based least squares policy iteration for reinforcement learning," *IEEE Trans. Neural Netw.*, vol. 18, no. 4, pp. 973–992, Jul. 2007.

[59] S. Mahadevan and M. Maggioni, "Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes," *J. Mach. Learn. Res.*, vol. 8, no. 10, pp. 2169–2231, 2007.

[60] X. Xu, Z. Huang, D. Graves, and W. Pedrycz, "A clustering-based graph Laplacian framework for value function approximation in reinforcement learning," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2613–2625, Dec. 2014.

[61] Z. Huang, X. Xu, and L. Zuo, "Reinforcement learning with automatic basis construction based on isometric feature mapping," *Inf. Sci.*, vol. 286, pp. 209–227, Dec. 2014.

[62] 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.

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 the General Motors Research and Development Center, Warren, MI, USA, from 1993 to 1995, and 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, and now he serves as the Associate Director of The State Key Laboratory of Management and Control for Complex Systems at the Institute of Automation. He has authored 15 books, including six research monographs and nine edited volumes.

Dr. Liu is a fellow of the International Neural Networks Society. 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 the 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 currently the Editor-in-Chief of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

Hongliang Li received the B.S. degree in mechanical engineering and automation from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing.

He is also with the University of Chinese Academy of Sciences, Beijing. His current research interests include machine learning, neural networks, reinforcement learning, adaptive dynamic programming, and game theory.

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

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. His current research interests include adaptive dynamic programming, neural networks and learning systems, and complex systems and intelligent control.