Value Iteration Adaptive Dynamic Programming for Optimal Control of Discrete-Time Nonlinear Systems

Qinglai Wei, *Member, IEEE*, Derong Liu, *Fellow, IEEE*, and Hanquan Lin

***Abstract*—In this paper, a value iteration adaptive dynamic programming (ADP) algorithm is developed to solve infinite horizon undiscounted optimal control problems for discrete-time nonlinear systems. The present value iteration ADP algorithm permits an arbitrary positive semi-definite function to initialize the algorithm. A novel convergence analysis is developed to guarantee that the iterative value function converges to the optimal performance index function. Initialized by different initial functions, it is proven that the iterative value function will be monotonically nonincreasing, monotonically nondecreasing, or nonmonotonic and will converge to the optimum. In this paper, for the first time, the admissibility properties of the iterative control laws are developed for value iteration algorithms. It is emphasized that new termination criteria are established to guarantee the effectiveness of the iterative control laws. Neural networks are used to approximate the iterative value function and compute the iterative control law, respectively, for facilitating the implementation of the iterative ADP algorithm. Finally, two simulation examples are given to illustrate the performance of the present method.**

***Index Terms*—Adaptive critic designs, adaptive dynamic programming (ADP), approximate dynamic programming, neural networks, neuro-dynamic programming, optimal control, reinforcement learning, value iteration.**

I. Introduction

DUE TO the increasing demands on system performance, production quality as well as economic operation and modern industrial processes are becoming more and more complicated and the degrees of automation of such processes are therefore significantly increasing [1]–[5]. As a result, the control of such complex processes is posing a great challenge due to the possible unavailability of sufficient quantitative knowledge about the process. Data-based control methods make use of the information obtained from the available process measurements to describe various complex behaviors, and thus have formed an efficient alternative for control and monitoring issues with complex industrial applications [6]–[9].

Although dynamic programming is a very useful tool in solving optimization and optimal control problems [10]–[12], it is often computationally untenable to run true dynamic programming. The difficulty lies in solving the time-varying Hamilton–Jacobi–Bellman (HJB) equation for which analytical solution is nearly impossible to obtain, i.e., as a result of the well-known "curse of dimensionality" [13]. Hence, many approaches were proposed to obtain the approximate solutions of HJB equation [14]–[17]. Among these approximate methods, adaptive dynamic programming (ADP), proposed by Werbos [18], [19], as an effective data-based approach to solve optimal control problems forward-in-time, has gained much attention from researchers [20]–[29]. In [30], a complex-valued ADP algorithm was discussed, where for the first time the optimal control problem of complex-valued nonlinear systems was successfully solved by ADP. In [31], based on neurocognitive psychology, a novel controller based on multiple actor-critic structures was developed for unknown systems and the proposed controller traded off fast actions based on stored behavior patterns with real-time exploration using current input–output data. In [32], an effective off-policy learning based integral reinforcement learning algorithm was presented, which successfully solved the optimal control problem for completely unknown continuous-time systems with unknown disturbances. Iterative methods are primary tools in ADP to obtain the solution of HJB equation indirectly and have received more and more attention [33]–[42].

Value iteration algorithms are a class of the most important iterative ADP algorithms [43]–[49]. Value iteration algorithms of ADP were given in [50] and [51]. In 2008, Al-Tamimi *et al*. [52] studied deterministic discrete-time affine nonlinear systems and a value iteration algorithm, which was referred to as heuristic dynamic programming (HDP), for finding the optimal control law. Starting from a zero initial value function, it was proven in [52] that the iterative value function was nondecreasing and bounded. When the iteration index increases to infinity, the iterative value function converges to the optimal performance index function which satisfies the HJB equation. In [53], value iteration algorithm is applied to solve optimal tracking control problems for nonlinear systems. In [54], value iteration of ADP, which was referred to as dual

Manuscript received May 4, 2015; revised August 22, 2015; accepted October 9, 2015. Date of publication November 2, 2015; date of current version February 12, 2016. This work was supported by the National Natural Science Foundation of China under Grant 61533017, Grant 61273140, Grant 61374105, and Grant 61233001. This paper was recommended by Associate Editor M. J. Er.

Q. Wei and H. Lin 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: qinglai.wei@ia.ac.cn; hanquan.lin@ia.ac.cn).

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/TCYB.2015.2492242

HDP (DHP), was implemented using RBF neural networks. In [55], value iteration of ADP was realized by globalized DHP. In [56], a Q-learning based value iteration algorithm was developed to obtain the optimal battery control law for smart residential grids. In recent years, value iteration algorithms have attracted more and more researchers [57]–[64].

But it is known that the existing value iteration algorithms, i.e., the traditional value iteration algorithms, possess inherent shortcomings and are not feasible for applications. First, nearly all the traditional value iteration algorithms are required to start from a zero initial condition. Other initial conditions are seldom discussed. In real-world applications, the initial performance index of control systems may not be zero. Second, it requires that the traditional value iteration algorithms implement infinite times to obtain the optimal control law. For real-world applications, however, the algorithm must be terminated within finite iterations. Third, iterative control laws obtained by traditional value iteration algorithms cannot guarantee to be stable control laws. It is only proven that as the iteration index increases to infinity, the converged optimal control is admissible. But for real-world applications, the algorithm must be terminated within finite iterations to find an effective iterative control law to control the system. Unfortunately, to the best of our knowledge, there are no discussions on the admissibility properties of the iterative control laws for value iteration algorithms. This means that the iterative control law achieved by the traditional value iteration algorithms may be invalid for real-world control systems. To overcome these disadvantages, a new iterative ADP algorithm will be developed with new termination criteria and analysis methods.

In this paper, a new value iteration ADP algorithm is developed to solve undiscounted optimal control problems of discrete-time nonlinear systems. In the developed algorithm, the zero initial condition is avoided. Starting with an arbitrary positive semi-definite function, it will be shown that the iterative value function will converge to the optimum. The convergence properties of the iterative value functions under different initial functions are analyzed. Furthermore, for the first time the admissibility properties of the iterative control laws for value iteration algorithms are developed. We emphasize that new termination criteria are established which guarantee the effectiveness of the achieved iterative control law by the developed value iteration algorithm.

This paper is organized as follows. In Section II, the problem formulations are presented. In Section III, the new value iteration algorithm is derived. The convergence properties of the iterative value functions and the admissibility properties of the iterative control laws are also presented in this section. In Section IV, two simulation examples are given to demonstrate the effectiveness of the developed control scheme. Finally, in Section V, this paper is concluded with a few pertinent remarks.

II. Problem Formulation

In this paper, we will study the following discrete-time nonlinear systems:

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

where $x_k \in \mathbb{R}^n$ is the state vector and $u_k \in \mathbb{R}^m$ is the control vector. Let x_0 be the initial state and $F(x_k, u_k)$ be the system function.

Let $\underline{u}_k = (u_k, u_{k+1}, \ldots)$ be an arbitrary sequence of controls from k to ∞. The performance index function for state x_0 under the control sequence $\underline{u}_0 = (u_0, u_1, \ldots)$ is defined as

$$J(x_0, \underline{u}_0) = \sum_{k=0}^{\infty} U(x_k, u_k) \tag{2}$$

where $U(x_k, u_k) > 0$, $\forall\, x_k, u_k \neq 0$, is the utility function.

We will study the optimal control problem for (1). The goal of this paper is to find an optimal control scheme which stabilizes (1) and simultaneously minimizes the performance index function (2). For convenience of analysis, results of this paper are based on the following assumptions.

Assumption 1: Equation (1) is controllable and the function $F(x_k, u_k)$ is Lipschitz continuous for x_k and u_k.

Assumption 2: The system state $x_k = 0$ is an equilibrium state of (1) under the control $u_k = 0$, i.e., $F(0, 0) = 0$.

Assumption 3: The feedback control $u_k = u(x_k)$ satisfies $u_k = u(x_k) = 0$ for $x_k = 0$.

Assumption 4: The utility function $U(x_k, u_k)$ is a continuous positive definite function of x_k and u_k.

Define the control sequence set as $\underline{\mathfrak{U}}_k = \big\{\underline{u}_k \colon \underline{u}_k = (u_k, u_{k+1}, \ldots),\ \forall\, u_{k+i} \in \mathbb{R}^m,\ i = 0, 1, \ldots\big\}$. Then, for an arbitrary control sequence $\underline{u}_k \in \underline{\mathfrak{U}}_k$, the optimal performance index function can be defined as

$$J^*(x_k) = \inf_{\underline{u}_k} \big\{J(x_k, \underline{u}_k) \colon \underline{u}_k \in \underline{\mathfrak{U}}_k\big\}. \tag{3}$$

According to Bellman's principle of optimality, $J^*(x_k)$ satisfies the following discrete-time HJB equation:

$$J^*(x_k) = \inf_{u_k} \big\{U(x_k, u_k) + J^*(F(x_k, u_k))\big\}. \tag{4}$$

Define the law of optimal control as

$$u^*(x_k) = \arg\inf_{u_k} \big\{U(x_k, u_k) + J^*(F(x_k, u_k))\big\}. \tag{5}$$

Hence, the HJB equation (4) can be written as

$$J^*(x_k) = U\big(x_k, u^*(x_k)\big) + J^*\big(F\big(x_k, u^*(x_k)\big)\big). \tag{6}$$

We can see that if we want to obtain the optimal control law $u^*(x_k)$, we must obtain the optimal performance index function $J^*(x_k)$. Generally, $J^*(x_k)$ is unknown before all the controls $u_k \in \mathbb{R}^n$ are considered. If we adopt the traditional dynamic programming method to obtain the optimal performance index function one step at a time, then we have to face the "the curse of dimensionality." In [52], a value iteration algorithm for affine nonlinear systems was proposed to obtain $J^*(x_k)$ iteratively, whereas the initial value function must set to zero to guarantee the convergence of the iterative value function which limits its applications. On the other hand, the admissibility properties of the iterative control law by value iteration cannot be guaranteed which makes the algorithm only implementable offline. To overcome these difficulties, and inspired by [52], a new iterative ADP algorithm is developed in this paper with convergence and admissibility properties.

III. Properties of the Value Iteration Algorithm of ADP

In this section, a new value iteration algorithm is developed to obtain the optimal control law for nonlinear system (1). New convergence analysis methods will be established in this section. Admissibility properties of the developed algorithm will be analyzed and new termination criteria of value iteration algorithms will be established.

A. Derivation of the Value Iteration Algorithm

In the developed value iteration algorithm, the value function and control law are updated at every iteration, with the iteration index i increasing from 0 to infinity. For $x_k \in \mathbb{R}^n$, let the initial function $\Psi(x_k) \geq 0$ be an arbitrary positive semi-definite function. Then, let the initial value function be expressed as

$$V_0(x_k) = \Psi(x_k). \tag{7}$$

The iterative control law $v_0(x_k)$ can be computed as follows:

$$\begin{aligned} v_0(x_k) &= \arg\min_{u_k}\{U(x_k, u_k) + V_0(x_{k+1})\} \\ &= \arg\min_{u_k}\{U(x_k, u_k) + V_0(F(x_k, u_k))\} \end{aligned} \tag{8}$$

where $V_0(x_{k+1}) = \Psi(x_{k+1})$. The iterative value function can be updated as

$$V_1(x_k) = U(x_k, v_0(x_k)) + V_0(F(x_k, v_0(x_k))). \tag{9}$$

For $i = 1, 2, \ldots,$ the value iteration algorithm will iterate between

$$\begin{aligned} v_i(x_k) &= \arg\min_{u_k}\{U(x_k, u_k) + V_i(x_{k+1})\} \\ &= \arg\min_{u_k}\{U(x_k, u_k) + V_i(F(x_k, u_k))\} \end{aligned} \tag{10}$$

and

$$\begin{aligned} V_{i+1}(x_k) &= \min_{u_k}\{U(x_k, u_k) + V_i(x_{k+1})\} \\ &= U(x_k, v_i(x_k)) + V_i(F(x_k, v_i(x_k))). \end{aligned} \tag{11}$$

From the value iteration algorithm (7)–(11), we can see that the iterative value function $V_i(x_k)$ is used to approximate $J^*(x_k)$ and the iterative control law $v_i(x_k)$ is used to approximate $u^*(x_k)$. Therefore, when $i \to \infty$, the algorithm should be convergent which makes $V_i(x_k)$ and $v_i(x_k)$ converge to the optimal ones. In the next section, we will show such properties of the developed value iteration algorithm.

B. Convergence Properties of the Value Iteration Algorithm

In [52], for zero initial value function, it was proven that the iterative value function is monotonically nondecreasing and converges to the optimum. For arbitrary positive semi-definite initial functions, however, the analysis method in [52] is invalid. In [60], a "functional bound" method was proposed for the value iteration with zero initial value function. Inspired by [60], new convergence analysis methods for the value iteration algorithm are developed in this section.

Theorem 1: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). Let $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\delta}$, and $\overline{\delta}$ be constants that satisfy

$$0 < \underline{\gamma} \leq \overline{\gamma} < \infty \tag{12}$$

and

$$0 \leq \underline{\delta} \leq \overline{\delta} < 1 \tag{13}$$

respectively. If for any x_k, the constants $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\delta}$, and $\overline{\delta}$ make

$$\underline{\gamma} U(x_k, u_k) \leq J^*(F(x_k, u_k)) \leq \overline{\gamma} U(x_k, u_k) \tag{14}$$

and

$$\underline{\delta} J^*(x_k) \leq V_0(x_k) \leq \overline{\delta} J^*(x_k) \tag{15}$$

hold uniformly, then for $i = 0, 1, \ldots,$ the iterative value function $V_i(x_k)$ satisfies

$$\begin{aligned} &\left(1 + \frac{\underline{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)^i}\right) J^*(x_k) \leq V_i(x_k) \\ &\qquad \leq \left(1 + \frac{\overline{\delta} - 1}{\left(1 + \underline{\gamma}^{-1}\right)^i}\right) J^*(x_k). \end{aligned} \tag{16}$$

Proof: The theorem can be proven in two steps.

1) *Prove the Left-Hand Side of the Inequality (16):* Mathematical induction is employed to prove the conclusion. According to (15) and (16) obviously holds for $i = 0$. Now let $i = 1$. We have

$$\begin{aligned} V_1(x_k) &= \min_{u_k}\{U(x_k, u_k) + V_0(x_{k+1})\} \\ &\geq \min_{u_k}\left\{U(x_k, u_k) + \underline{\delta} J^*(x_{k+1})\right\} \\ &\geq \min_{u_k}\left\{\left(1 + \overline{\gamma}\frac{\underline{\delta} - 1}{1 + \overline{\gamma}}\right) U(x_k, u_k)\right. \\ &\qquad\left. + \left(\underline{\delta} - \frac{\underline{\delta} - 1}{1 + \overline{\gamma}}\right) J^*(x_{k+1})\right\} \\ &= \left(1 + \frac{\underline{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)}\right) \min_{u_k}\left\{U(x_k, u_k) + J^*(x_{k+1})\right\} \\ &= \left(1 + \frac{\underline{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)}\right) J^*(x_k). \end{aligned} \tag{17}$$

Assume that the conclusion holds for $i = l - 1$, $l = 1, 2, \ldots$. Then for $i = l$, we have

$$\begin{aligned} &V_{l+1}(x_k) \\ &= \min_{u_k}\{U(x_k, u_k) + V_l(F(x_k, u_k))\} \\ &\geq \min_{u_k}\left\{U(x_k, u_k) + \left(1 + \frac{\underline{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)^{l-1}}\right) J^*(F(x_k, u_k))\right. \\ &\qquad\left. + \frac{\underline{\delta} - 1}{(1 + \overline{\gamma})\left(1 + \overline{\gamma}^{-1}\right)^{l-1}}\left(\overline{\gamma} U(x_k, u_k) - J^*(F(x_k, u_k))\right)\right\} \\ &= \left(1 + \frac{\underline{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)^l}\right) \min_{u_k}\left\{U(x_k, u_k) + J^*(F(x_k, u_k))\right\} \\ &= \left(1 + \frac{\underline{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)^l}\right) J^*(x_k). \end{aligned} \tag{18}$$

2) *Prove the Right-Hand Side of the Inequality (16):* We also use mathematical induction to prove the conclusion. According to (15), (16) obviously holds for $i = 0$. Let $i = 1$. We have

$$\begin{aligned}
V_1(x_k) &= \min_{u_k}\{U(x_k, u_k) + V_0(x_{k+1})\} \\
&\le \min_{u_k}\big\{U(x_k, u_k) + \bar{\delta} J^*(x_{k+1})\big\} \\
&\le \min_{u_k}\Bigg\{\left(1 + \underline{\gamma}\frac{\bar{\delta} - 1}{1 + \underline{\gamma}}\right)U(x_k, u_k) \\
&\qquad + \left(\bar{\delta} - \frac{\bar{\delta} - 1}{1 + \underline{\gamma}}\right)J^*(x_{k+1})\Bigg\} \\
&= \left(1 + \frac{\bar{\delta} - 1}{\left(1 + \underline{\gamma}^{-1}\right)}\right)\min_{u_k}\big\{U(x_k, u_k) + J^*(x_{k+1})\big\} \\
&= \left(1 + \frac{\bar{\delta} - 1}{\left(1 + \underline{\gamma}^{-1}\right)}\right)J^*(x_k). \qquad (19)
\end{aligned}$$

Assume that the conclusion holds for $i = l - 1$, $l = 1, 2, \ldots$. Then for $i = l$, we have

$$\begin{aligned}
&V_{l+1}(x_k) \\
&= \min_{u_k}\big\{U(x_k, u_k) + V_l\big(x_{k+1}\big)\big\} \\
&\le \min_{u_k}\Bigg\{U(x_k, u_k) + \left(1 + \frac{\bar{\delta} - 1}{\left(1 + \underline{\gamma}^{-1}\right)^{l-1}}\right)J^*\big(x_{k+1}\big) \\
&\qquad + \frac{1 - \bar{\delta}}{\left(1 + \underline{\gamma}\right)\left(1 + \underline{\gamma}^{-1}\right)^{l-1}}\Big(J^*\big(x_{k+1}\big) - \underline{\gamma}U(x_k, u_k)\Big)\Bigg\} \\
&= \left(1 + \frac{\bar{\delta} - 1}{\left(1 + \underline{\gamma}^{-1}\right)^{l}}\right)\min_{u_k}\big\{U(x_k, u_k) + J^*\big(x_{k+1}\big)\big\} \\
&= \left(1 + \frac{\bar{\delta} - 1}{\left(1 + \underline{\gamma}^{-1}\right)^{l}}\right)J^*(x_k). \qquad (20)
\end{aligned}$$

The proof is completed. ■

Theorem 2: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). Let $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\delta}$, and $\overline{\delta}$ be constants that satisfy (12) and

$$0 \le \underline{\delta} \le 1 \le \overline{\delta} < \infty \qquad (21)$$

respectively. If for any x_k, the constants $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\delta}$, and $\overline{\delta}$ make (14) and (15) hold uniformly. Then, the iterative value function $V_i(x_k)$ satisfies

$$\left(1 + \frac{\underline{\delta} - 1}{(1 + \overline{\gamma}^{-1})^i}\right)J^*(x_k) \le V_i(x_k) \le \left(1 + \frac{\overline{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)^i}\right)J^*(x_k). \qquad (22)$$

Proof: The left-hand side of inequality (22) can be proven according to (17) and (18). Now, we prove the right-hand side of inequality (22) by mathematical induction. Inequality (22) obviously holds for $i = 0$. Let $i = 1$. We have

$$\begin{aligned}
V_1(x_k) &= \min_{u_k}\{U(x_k, u_k) + V_0(F(x_k, u_k))\} \\
&\le \min_{u_k}\Bigg\{U(x_k, u_k) + \bar{\delta}J^*(F(x_k, u_k)) \\
&\qquad + \frac{\bar{\delta} - 1}{(1 + \overline{\gamma})}\big(\overline{\gamma}U(x_k, u_k) - J^*(F(x_k, u_k))\big)\Bigg\} \\
&= \left(1 + \frac{\bar{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)}\right)J^*(x_k). \qquad (23)
\end{aligned}$$

Assume that the conclusion holds for $i = l - 1$, $l = 1, 2, \ldots$. Then for $i = l$, we have

$$\begin{aligned}
&V_{l+1}(x_k) \\
&= \min_{u_k}\{U(x_k, u_k) + V_l(F(x_k, u_k))\} \\
&\le \min_{u_k}\Bigg\{U(x_k, u_k) + \left(1 + \frac{\bar{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)^{l-1}}\right)J^*(x_{k+1}) \\
&\qquad + \frac{\bar{\delta} - 1}{(1 + \bar{\gamma})\left(1 + \bar{\gamma}^{-1}\right)^{l-1}}\big(\bar{\gamma}U(x_k, u_k) - J^*(x_{k+1})\big)\Bigg\} \\
&= \left(1 + \frac{\bar{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)^{l}}\right)J^*(x_k). \qquad (24)
\end{aligned}$$

Hence, (22) holds for $i = 0, 1, \ldots$. The mathematical induction is completed. ■

Note the difference between the right-hand sides of (16) and (22) are due to the difference between (13) and (21).

Theorem 3: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). Let $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\delta}$, and $\overline{\delta}$ be constants that satisfy (12) and

$$1 \le \underline{\delta} \le \overline{\delta} < \infty. \qquad (25)$$

If for any x_k, the constants $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\delta}$, and $\overline{\delta}$ make (14) and (15) hold uniformly, then the iterative value function $V_i(x_k)$ satisfies (16).

Proof: The conclusion can be derived by (17)–(20) and the proof is omitted here. ■

Theorem 4: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). Let the constants $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\delta}$, and $\overline{\delta}$ satisfy (12) and

$$0 \le \underline{\delta} \le \overline{\delta} < \infty \qquad (26)$$

respectively. If for any x_k, the constants $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\delta}$, and $\overline{\delta}$ make (14) and (15) hold uniformly, then the iterative value function $V_i(x_k)$ converges to the optimal performance index function $J^*(x_k)$, that is

$$\lim_{i\to\infty} V_i(x_k) = J^*(x_k). \qquad (27)$$

Proof: According to the left-hand side of inequalities (16) and (22), letting $i \to \infty$, we can get

$$\lim_{i\to\infty}\left\{\left(1 + \frac{\underline{\delta} - 1}{\left(1 + \overline{\gamma}^{-1}\right)^i}\right)J^*(x_k)\right\} = J^*(x_k). \qquad (28)$$

On the other hand, according to the right-hand side of inequalities (16) and (22), letting $i \to \infty$, we can obtain

$$\begin{aligned}\lim_{i\to\infty}&\left\{\left(1+\frac{\overline{\delta}-1}{\left(1+\underline{\gamma}^{-1}\right)^{i}}\right)J^{*}(x_k)\right\}\\ &=\lim_{i\to\infty}\left\{\left(1+\frac{\overline{\delta}-1}{\left(1+\overline{\gamma}^{-1}\right)^{i}}\right)J^{*}(x_k)\right\}\\ &=J^{*}(x_k).\end{aligned}\tag{29}$$

According to (16), (22), (28), and (29), we have (27) immediately. The proof is completed. ■

Remark 1: From Theorem 4, we can see that the iterative value function will converge to the optimum as $i \to \infty$, which is independent of the initial value function $\Psi(x_k)$. Thus it is unnecessary to obtain the detailed values of $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\delta}$, and $\overline{\delta}$. This is an advantage of the developed algorithm. On the other hand, we should say that the initial value function affects the convergence process of the iterative value functions directly. It means that for different initial value functions, we will obtain different convergence processes. In the following, we will show these properties.

Corollary 1: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If for any x_k, the initial value function $\Psi(x_k) \le J^{*}(x_k)$, then $\forall\, i \ge 0$, we have $V_i(x_k) \le J^{*}(x_k)$ holds.

Corollary 2: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If for any x_k, the initial value function $\Psi(x_k) \ge J^{*}(x_k)$, then for $i \ge 0$, we have $V_i(x_k) \ge J^{*}(x_k)$ holds.

Theorem 5: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If for any x_k, the inequality

$$V_1(x_k) \le V_0(x_k)\tag{30}$$

holds, then the iterative value function $V_i(x_k)$ is a monotonically nonincreasing sequence for any $i \ge 0$, that is

$$V_{i+1}(x_k) \le V_i(x_k).\tag{31}$$

Proof: We prove this by mathematical induction. First, we let $i = 1$. According to (11) and (30), we have

$$\begin{aligned}V_2(x_k) &= \min_{u_k}\{U(x_k,u_k)+V_1(x_{k+1})\}\\ &\le \min_{u_k}\{U(x_k,u_k)+V_0(x_{k+1})\}\\ &= V_1(x_k).\end{aligned}\tag{32}$$

Assume that the conclusion holds for $i = l-1$, $l = 2, 3, \ldots$. Then for $i = l$ we have

$$\begin{aligned}V_{l+1}(x_k) &= \min_{u_k}\{U(x_k,u_k)+V_l(x_{k+1})\}\\ &\le \min_{u_k}\{U(x_k,u_k)+V_{l-1}(x_{k+1})\}\\ &= V_l(x_k).\end{aligned}\tag{33}$$

The proof is completed. ■

Theorem 6: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If for any x_k, the inequality

$$V_1(x_k) \ge V_0(x_k)\tag{34}$$

holds, then the iterative value function $V_i(x_k)$ is a monotonically nondecreasing sequence for $i \ge 0$, that is

$$V_{i+1}(x_k) \ge V_i(x_k).\tag{35}$$

Remark 2: If for any x_k, we let the initial value function $V_0(x_k) \equiv 0$, then the present value iteration algorithm is reduced to the traditional value iteration algorithm in [47], [52], [53], [55], [59], and [60]. In [60], using the functional bound method, the convergence of the iterative value function was proven with zero initial value function. In this paper, inspired by [60], we have proven that the iterative value function converges to the optimal performance index function with an arbitrary positive semidefinite initial value function. Furthermore, the monotonicity property for the traditional value iteration algorithm in [47], [52], [53], [55], and [59] can easily be justified by our developed value iteration algorithm. So, we can say that the traditional value iteration is a special case of the present value iteration algorithm.

Corollary 3: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If for any x_k, we have (30) holds, where $V_0(x_k)$ is expressed by (7), then for $i = 0, 1, \ldots,$ the iterative value function satisfies

$$V_i(x_k) \ge J^{*}(x_k).\tag{36}$$

Proof: According to Theorem 5, for $i = 0, 1, \ldots,$ we have

$$V_i(x_k) \ge V_{i+1}(x_k) \ge V_{i+2}(x_k) \ge \cdots\tag{37}$$

Then, for $l \ge i$, we can get

$$V_i(x_k) \ge V_l(x_k).\tag{38}$$

Let $l \to \infty$. According to (27), we can obtain

$$V_i(x_k) \ge \lim_{l\to\infty} V_l(x_k) = J^{*}(x_k).\tag{39}$$

The proof is completed. ■

Corollary 4: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If for any x_k, we have (34) holds, where $V_0(x_k)$ is expressed by (7), then for $i = 0, 1, \ldots,$ the iterative value function satisfies

$$V_i(x_k) \le J^{*}(x_k).\tag{40}$$

Remark 3: It should be pointed out that the converse propositions of Corollary 3 may not be true. For example, if we choose an initial value function $\Psi(x_k) \ge J^{*}(x_k)$, we cannot conclude that $V_{i+1}(x_k) \le V_i(x_k)$ holds for $i = 0, 1, \ldots$. If we choose an initial value function $\Psi(x_k) \le J^{*}(x_k)$, we cannot conclude that $V_{i+1}(x_k) \ge V_i(x_k)$ holds for $i = 0, 1, \ldots$. Hence, if we want the iterative value function to be monotonically nonincreasing (or nondecreasing) convergent to the optimum, it is not enough to choose an arbitrary initial value function $\Psi(x_k) \ge$ (or $\le$) $J^{*}(x_k)$ to guarantee the monotonicity of the iterative value functions. Additional initial conditions should be provided. In the following, two special initial conditions are provided to guarantee the monotonicity of the iterative value function.

Lemma 1: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). Let the initial value function $V_0(x_k) \equiv 0$.

Then for $i = 0, 1, \ldots,$ $V_i(x_k)$ is monotonically nondecreasing convergent to $J^*(x_k)$.

Before we proceed to the next theorem, the following definition is necessary.

Definition 1: A control law $u(x_k)$ is said to be admissible [52] with respect to (2) on Ω if $u(x_k)$ is continuous on Ω, $u(0) = 0$, $u(x_k)$ stabilizes (1) on Ω, and $\forall x_0 \in \Omega$, $J(x_0)$ is finite.

Theorem 7: For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). Let $\Psi(x_k)$ be an initial positive semi-definite function which satisfies

$$\Psi(x_k) = U(x_k, \bar{v}(x_k)) + \Psi(x_{k+1}) \tag{41}$$

where $\bar{v}(x_k)$ is an arbitrary admissible control law. Then for $i = 0, 1, \ldots,$ $V_i(x_k)$ is monotonically nonincreasing and converges to the optimum.

Proof: According to (41), we have

$$\begin{aligned} V_1(x_k) &= U(x_k, v_0(x_k)) + V_0(x_{k+1}) \\ &= \min_{u_k}\{U(x_k, u_k) + \Psi(x_{k+1})\} \\ &\le U(x_k, \bar{v}(x_k)) + \Psi(x_{k+1}) \\ &= \Psi(x_k). \end{aligned} \tag{42}$$

Hence, we obtain $V_1(x_k) \le V_0(x_k)$. Using the mathematical induction, we can prove that (31) holds for $i = 0, 1, \ldots.$ ■

Remark 4: In the above, it is shown that the iterative value function will converge to the optimum as $i \to \infty$. In real-world applications, however, the algorithm cannot be implemented for infinite number of iterations to obtain the optimal performance index function. The algorithm must be terminated within finite number of iterations and an iterative control law will be used to control the systems. For traditional value iteration algorithms [47], [52]–[55], [59], [60], if the iterative control law $v_i(x_k)$ makes the inequality $|V_{i+1}(x_k) - V_i(x_k)| \le \varepsilon$ hold, where ε is the computation precision, then the algorithm is terminated. We call $|V_{i+1}(x_k) - V_i(x_k)| \le \varepsilon$ the "convergence termination criterion." We usually regard the iterative control law $v_i(x_k)$ as the optimal one. Unfortunately, $v_i(x_k)$ may not be an admissible control law but only a uniformly ultimately bounded (UUB) one. The following theorem will show this property.

Definition 2: We say a solution is UUB [65], if there exists a compact set $\mathcal{X} \subset \mathbb{R}^n$, such that for all $x_{k_0} = x_0 \in \mathcal{X}$, there exists an ε and a number $T(\varepsilon, x_0)$ such that $\|x_k\| \le \varepsilon$ for all $k \ge k_0 + T$.

Theorem 8: Suppose that Assumptions 1–4 hold. For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If there exists a constant $\varepsilon > 0$ such that

$$|V_{i+1}(x_k) - V_i(x_k)| \le \varepsilon \tag{43}$$

then the state of the nonlinear system (1) is UUB under the iterative control law $v_i(x_k)$.

Proof: The theorem can be proven in two steps.

1) *Show That $V_i(x_k)$ is a Positive Definite Function for $t = 1, 2, \ldots$:* First, according to the definition of $\Psi(x_k)$, the value function $V_0(x_k) = \Psi(x_k)$ is a positive semi-definite function. Then, according to the definition of $V_1(x_k)$ in (9), let $x_k = 0$, and we obtain

$$V_1(0) = U(0, v_0(0)) + V_0(F(0, v_0(0))). \tag{44}$$

According to Assumptions 2–4, we have $V_0(0) = 0$, $F(0, 0) = 0$, and $U(0, 0) = 0$. Then, we can get $V_1(0) = 0$. On the other hand, according to Assumption 4, we have $V_1(x_k) > 0$, for any $x_k, u_k \neq 0$. Therefore, $V_1(x_k)$ is a positive definite function. Using the mathematical induction, we can easily obtain that $V_i(x_k)$ is a positive definite function for $i = 1, 2, \ldots.$ If $\Psi(x_k)$ is positive definite, $V_i(x_k)$ is a positive definite function for $i = 0, 1, \ldots.$

2) *Show That the State of the Nonlinear System (1) is UUB Under the Iterative Control Law $v_i(x_k)$:* According to (43), we have

$$\begin{aligned} -U(x_k, v_i(x_k)) - \varepsilon \le \Delta V_i(x_k) &= V_i(F(x_k, v_i(x_k))) - V_i(x_k) \\ &\le -U(x_k, v_i(x_k)) + \varepsilon. \end{aligned} \tag{45}$$

If

$$-U(x_k, v_i(x_k)) - \varepsilon \le \Delta V_i(x_k) \le 0 \tag{46}$$

we can easily prove that $V_i(x_k)$ is a Lyapunov function [66] and the system is asymptotically stable, since $V_i(x_k)$ is positive definite. Now we analyze the situation for $\Delta V_i(x_k) \le -U(x_k, v_i(x_k)) + \varepsilon$. As $V_i(x_k)$ is a positive definite function, there must exit two functions $\alpha(\|x_k\|)$ and $\beta(\|x_k\|)$, which belong to class $\mathcal{K}$ [66] and satisfy

$$0 < \alpha(\|x_k\|) \le V_i(x_k) \le \beta(\|x_k\|). \tag{47}$$

Define a new state set

$$\Omega_{x_k} = \left\{x_k \mid x_k \in \mathbb{R}^n \text{ and } U(x_k, v_i(x_k)) \le \varepsilon\right\}. \tag{48}$$

As $U(x_k, v_i(x_k))$ is a positive definite function, for any $x_k \in \Omega_{x_k}$, $\|x_k\|$ is finite, where $\|x_k\|$ is Euclidean norm. We can define

$$\varrho = \sup_{x_k \in \Omega_{x_k}} \{\|x_k\|\}. \tag{49}$$

As ε is finite, ϱ is finite. Then for any ϱ that satisfies (49), there exits a finite Γ, such that $\|\Gamma\| \ge \|\varrho\|$, which satisfies

$$\alpha(\|\Gamma\|) \ge \beta(\|\varrho\|). \tag{50}$$

Then, for all ϵ that satisfies $\epsilon \ge \|\Gamma\|$, there exists a $\delta(\epsilon)$, such that $\delta(\epsilon) \ge \|\varrho\|$, which satisfies $\beta(\delta) \le \alpha(\epsilon)$. Thus, there exists a state x_k, such that $\|\varrho\| \le \|x_k\| \le \delta(\epsilon)$, which satisfies

$$\alpha(\epsilon) \ge \beta(\delta) \ge V_i(x_k). \tag{51}$$

As $\|x_k\| \ge \|\varrho\|$, we have

$$V_i(x_{k+1}) - V_i(x_k) \le 0. \tag{52}$$

Hence, for any x_k that satisfies $\|\varrho\| \le \|x_k\| \le \delta(\epsilon)$, there exists a $T > 0$ that satisfies

$$\alpha(\epsilon) \ge \beta(\delta) \ge V_i(x_k) \ge V_i(x_{k+T}) \ge \alpha(\|x_{k+T}\|) \tag{53}$$

which obtains $\epsilon > \|x_{k+T}\|$. Therefore, for any x_k that satisfies $\|x_k\| \geq \|\varrho\|$, there exist a $T = 1, 2, \ldots$ that makes $\|x_{k+T}\| \leq \|\varrho\|$ hold. As $\|\Gamma\| \geq \|\varrho\|$, we can obtain $\|x_{k+T}\| \leq \|\Gamma\|$. According to Definition 2, we can draw the conclusion. The proof is completed. ■

It was proven in [52] that the optimal control law $u^*(x_k)$ is an admissible control law. Theorem 8 shows that the iterative control law $v_i(x_k)$ is only UUB. Hence, strictly speaking, $u^*(x_k)$ cannot be replaced by the iterative control law $v_i(x_k)$ and the algorithm cannot be terminated only by the convergence termination criterion (43). To overcome this difficulty, the properties of the iterative control law $v_i(x_k)$ will be analyzed, and new termination criteria of value iteration algorithms will be established.

Theorem 9: Suppose Assumptions 1–4 hold. For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If for any $x_k \neq 0$, the iterative control law $v_i(x_k)$ makes the following inequality:

$$V_{i+1}(x_k) - V_i(x_k) < U(x_k, v_i(x_k)) \tag{54}$$

hold, then the iterative control law $v_i(x_k)$ is an admissible control law.

Proof: According to (54), there must exist a constant $-\infty < \theta < 1$ that satisfies

$$V_{i+1}(x_k) - V_i(x_k) < \theta U(x_k, v_i(x_k)). \tag{55}$$

According to (11), the inequality (55) can be written as

$$V_i(x_{k+1}) - V_i(x_k) < (\theta - 1)U(x_k, v_i(x_k)). \tag{56}$$

Since $-\infty < \theta < 1$, we can get $V_i(x_{k+1}) - V_i(x_k) < 0$ which means $v_i(x_k)$ is a stable control law. On the other hand, according to (56), we can get

$$\begin{cases} V_i(x_{k+1}) - V_i(x_k) < (\theta - 1)U(x_k, v_i(x_k)) \\ V_i(x_{k+2}) - V_i(x_{k+1}) < (\theta - 1)U(x_{k+1}, v_i(x_{k+1})) \\ \quad\vdots \\ V_i(x_{k+N}) - V_i(x_{k+N-1}) < (\theta - 1)U(x_{k+N-1}, v_i(x_{k+N-1})). \end{cases} \tag{57}$$

As $v_i(x_k)$ is a stable control law, we have $\lim_{N\to\infty} V_i(x_{k+N}) = 0$. Let $N \to \infty$. We can get

$$V_i(x_k) > (1 - \theta)\sum_{j=0}^{\infty} U\big(x_{k+j}, v_i\big(x_{k+j}\big)\big). \tag{58}$$

As $V_i(x_k)$ is finite for any finite x_k and $-\infty < \theta < 1$, we can obtain that $\sum_{j=0}^{\infty} U(x_{k+j}, v_i(x_{k+j}))$ is finite, which proves the conclusion. ■

Remark 5: According to Theorem 9, new termination criteria of the value iteration algorithm can be established. The inequality (54) is called "admissibility termination criterion." We emphasize that admissibility termination criterion is an important termination criterion for the real-world applications of the developed value iteration algorithm. First, stability is a basic property of control systems, while the stability of the system cannot be guaranteed using the convergence criterion in [52]. Second, in [52], it required traditional value iteration algorithm to implement infinite times to reach the optimum, which makes it impossible to realize. Analyzing the property of the iterative control law is necessary to make the developed value iteration algorithm terminated within finite iterations. According to (43) and (54), we say that the value iteration algorithm can be terminated if and only if "convergence and admissibility termination criteria" are both satisfied and we declare that the admissibility termination criterion is a key criterion for the applications of the value iteration algorithm. On the other hand, if inequality (54) is never satisfied, it implies that the algorithm may never stop. Fortunately, this situation will not happen. The following theorem will show this property.

Theorem 10: Suppose Assumptions 1–4 hold. For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). Then, for all $x_k \neq 0$, there exists a finite $N > 0$ that satisfies

$$V_{N+1}(x_k) - V_N(x_k) < U(x_k, v_N(x_k)). \tag{59}$$

Proof: The conclusion can be proven by contradiction. Assume that (59) is false, and for all $N = 0, 1, \ldots,$ there exists an $\bar{x}_k \in \mathbb{R}^n$ that satisfies

$$V_{N+1}(\bar{x}_k) - V_N(\bar{x}_k) \geq U(\bar{x}_k, v_N(\bar{x}_k)). \tag{60}$$

Let $N \to \infty$. According to Theorem 4, we can get $\lim_{N\to\infty}(V_{N+1}(\bar{x}_k) - V_N(\bar{x}_k)) = 0$. According to (60), we can get

$$\lim_{N\to\infty} U(\bar{x}_k, v_N(\bar{x}_k)) = U(\bar{x}_k, v_\infty(\bar{x}_k)) = 0 \tag{61}$$

holds for $\bar{x}_k \in \mathbb{R}^n$. It contradicts the positive definiteness of $U(x_k, u_k)$. So the assumption is false and the conclusion holds. ■

Remark 6: An important property should be pointed out. For the value iteration algorithm (including traditional value iteration algorithms), if the iterative control law $v_i(x_k)$ is admissible, it cannot guarantee $v_j(x_k)$, $j > i$ to be admissible, although the iterative value function $V_j(x_k)$ is closer to the optimum than $V_i(x_k)$. In the simulation studies, we will show this property. If the iterative control law $v_i(x_k)$ is admissible, to guarantee the admissibility property of $v_j(x_k)$, $j > i$, new analysis method should be established.

Theorem 11: Suppose Assumptions 1–4 hold. For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). Let the iterative control law $v_i(x_k)$ be admissible that satisfies (54). If one of the following three conditions is satisfied.

1) For $j = 0, 1, \ldots,$ the iterative value function $V_{i+j}(x_k)$ satisfies

$$V_{i+1}(x_k) + V_{i+j}(x_k) \geq V_i(x_k) + V_{i+j+1}(x_k). \tag{62}$$

2) For all $j \geq i$, the iterative value function is convex for the iteration index j, i.e., $V_j(x_k)$ satisfies

$$V_j(x_k) \geq \frac{1}{2}\big(V_{j+1}(x_k) + V_{j-1}(x_k)\big). \tag{63}$$

3) Define $\Delta V_j(x_k)$ by

$$\Delta V_j(x_k) = V_j(x_k) - V_{j-1}(x_k) \tag{64}$$

and for all $j \geq i$, $\Delta V_j(x_k) \geq \Delta V_{j+1}(x_k)$.

Then, $v_j(x_k)$ is an admissible control law.

Algorithm 1 Value Iteration ADP Algorithm

```
Initialization:
    Choose randomly an array of initial states $x_0$;
    Choose a computation precision $\varepsilon$;
    Give a positive semi-definite function $\Psi(x_k)$;
Iteration:
    Let the iteration index $i = 0$ and $V_0(x_k) = \Psi(x_k)$;
    Compute the initial iterative control law $v_0(x_k)$ by (8). Obtain the value function $V_1(x_k)$ by (9).
    If $\forall x_k$, $V_1(x_k) \le V_0(x_k)$, then goto Step 4; Otherwise, goto Step 6.

    Block 1.
    Compute the initial iterative control law $v_i(x_k)$ by (10) and obtain the value function $V_{i+1}(x_k)$ by (11).
    If $\forall x_k$, $|V_{i+1}(x_k) - V_i(x_k)| \le \varepsilon$, then goto Step 9. Otherwise, let $i = i + 1$ and goto Step 4.

    Block 2.
    Compute the iterative control law $v_i(x_k)$ by (10) and obtain the value function $V_{i+1}(x_k)$ by (11).
    If $\forall x_k$, $|V_{i+1}(x_k) - V_i(x_k)| \le \varepsilon$, then goto next step. Otherwise, let $i = i + 1$ and goto Step 6.
    If $\forall x_k$, $V_{i+1}(x_k) - V_i(x_k) < U(x_k, v_i(x_k)))$, then goto next step. Otherwise, let $i = i + 1$ and goto Step 6.
    return $v_i(x_k)$ and $V_i(x_k)$.
```

Proof: First, if (62) holds, we have

$$\big(V_{i+j+1}(x_k) - V_{i+j}(x_k)\big) - (V_{i+1}(x_k) - V_i(x_k)) \le 0. \quad (65)$$

Then, we can get

$$\begin{aligned} &V_{i+j+1}(x_k) - U\big(x_k, v_{i+j}(x_k)\big) - V_{i+j}(x_k) \\ &\quad \le V_{i+1}(x_k) - U(x_k, v_i(x_k)) - V_i(x_k) + U(x_k, v_i(x_k)) \end{aligned} \quad (66)$$

which implies

$$\begin{aligned} V_{i+j}&(x_{k+1}) - V_{i+j}(x_k) \\ &\le V_i(x_{k+1}) - V_i(x_k) + U(x_k, v_i(x_k)) \\ &< 0. \end{aligned} \quad (67)$$

According to Theorem 9, we can draw the conclusion.

If (63) or (64) holds, it is very easy to draw the conclusion using the idea of (65)–(67) and the details are omitted here. ■

Lemma 2: Suppose Assumptions 1–4 hold. For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If the inequality (30) holds, then for $i = 0, 1, \ldots,$ the iterative control law $v_i(x_k)$ is an admissible control law.

Corollary 5: Suppose Assumptions 1–4 hold. For $i = 0, 1, \ldots,$ let $V_i(x_k)$ and $v_i(x_k)$ be obtained by (7)–(11). If the initial value function $\Psi(x_k)$ satisfies (41), then the iterative control law $v_i(x_k)$ is an admissible control law.

C. Summary of the Value Iteration ADP Algorithm

Now, we summarize the value iteration ADP algorithm in Algorithm 1.

Remark 7: We can see that the developed value iteration algorithm is classified into two blocks in which the termination criteria are different. When for any x_k, $V_1(x_k) \le V_0(x_k)$ holds, the value iteration algorithm is then implemented in block 1 in which only convergence termination criterion is considered. Otherwise, the value iteration algorithm is implemented in block 2, in which two termination criteria must be considered. Thus, if $V_1(x_k) \le V_0(x_k)$ holds, the algorithm becomes simpler. However, the initial value function is also more difficult to determine than the one in block 2.

Remark 8: For traditional value iteration algorithms, zero initial condition is used to implement the algorithm. According to Theorem 6, we know that the algorithm is implemented in block 2, while in [47], [52]–[55], [59], and [60] only the convergence termination criterion is considered in those algorithms. In this situation, we say that the admissibility property of the iterative control law cannot be guaranteed only by the convergence termination criterion. In simulation studies, we will show this property. In this paper, admissibility termination criterion is established based on the value iteration algorithm which guarantees the validity of the achieved iterative control. This makes the value iteration algorithm possess more potential for applications.

IV. Simulation Studies

To evaluate the performance of our value iteration algorithm, we choose two examples with quadratic utility functions for numerical experiments.

Example 1: The first example is a discretized inverted pendulum system [67]. The dynamics of the pendulum is expressed as

$$\begin{bmatrix} x_{1(k+1)} \\ x_{2(k+1)} \end{bmatrix} = \begin{bmatrix} x_{1k} + 0.1x_{2k} \\ 0.1\dfrac{g}{\ell}\sin(x_{1k}) + (1 - 0.1\kappa\ell)x_{2k} \end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{0.1}{m\ell^2} \end{bmatrix} u_k \quad (68)$$

where $m = 1/2$ kg and $\ell = 1/3$ m are the mass and length of the pendulum bar, respectively. Let $\kappa = 0.2$ and $g = 9.8$ m/s^2 be the frictional factor and the gravitational acceleration, respectively.

Let the performance index function be expressed by (2). The utility function is the quadratic form $U(x_k, u_k) = x_k^{\mathsf{T}} Q x_k + u_k^{\mathsf{T}} R u_k$, where $Q = I_1$, $R = I_2$, and I_1, I_2 denote the identity matrices with suitable dimensions. Let $x_0 = [1, -1]^{\mathsf{T}}$. Let the state space be $\Theta = \{x_k \mid -1 \le x_{1k} \le 1, -1 \le x_{2k} \le 1\}$. We choose $p = 10\,000$ states in Θ to implement the developed value iteration algorithm to obtain the optimal control law. Neural networks are used to implement the present value iteration algorithm. The critic network and the action network are chosen as three-layer back-propagation neural networks with the structures of 2–8–1 and 2–8–1, respectively. For each iteration step, the critic network and the action network are trained for 2000 steps under the learning rate 0.01 so that the neural network training errors become less than 10^{-6}. The weights updating rules of the neural networks can be seen in [69] and omitted here. To illustrate the effectiveness of the algorithm,

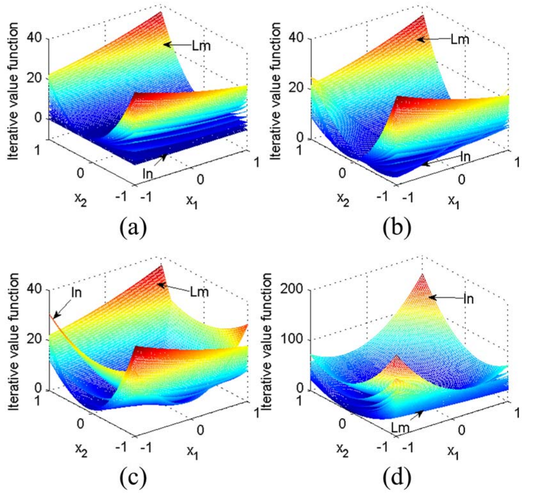

Fig. 1. Convergent curves of iterative value functions with $\Psi^j(x_k)$, $j = 1, \ldots, 4$. (a) $\Psi^1(x_k)$. (b) $\Psi^2(x_k)$. (c) $\Psi^3(x_k)$. (d) $\Psi^4(x_k)$.

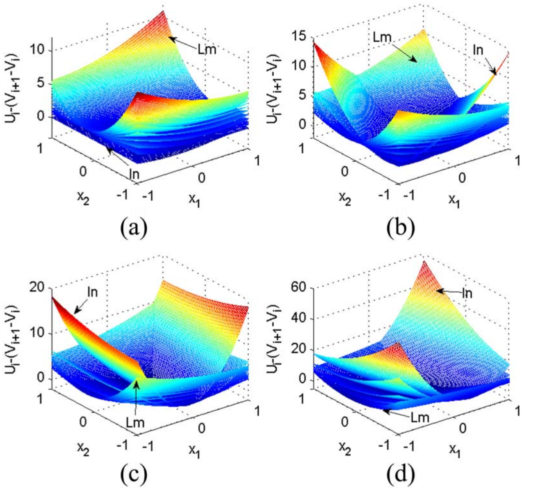

Fig. 2. Plots of $U(x_k, v_i(x_k)) - (V_{i+1}(x_k) - V_i(x_k))$ with $\Psi^j(x_k), j = 1, \ldots, 4$. (a) $\Psi^1(x_k)$. (b) $\Psi^2(x_k)$. (c) $\Psi^3(x_k)$. (d) $\Psi^4(x_k)$.

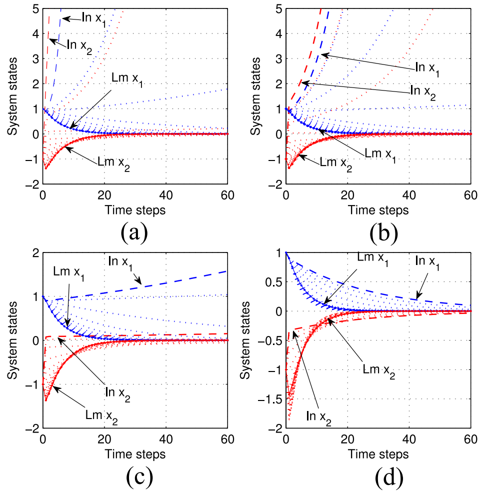

Fig. 3. Iterative state trajectories with $\Psi^j(x_k)$, $j = 1, \ldots, 4$. (a) $\Psi^1(x_k)$. (b) $\Psi^2(x_k)$. (c) $\Psi^3(x_k)$. (d) $\Psi^4(x_k)$.

four different initial value functions are considered. Let the initial value function be the quadratic form which are expressed by $\Psi^j(x_k) = x_k^{\mathsf{T}} P_j x_k$, $j = 1, 2, 3, 4$. Let $P_1 = 0$. Let P_2–P_4 be initialized by arbitrary positive definite matrices with the forms $P_2 = \begin{bmatrix} 9.56 & -5.39 \\ -5.39 & 4.31 \end{bmatrix}$, $P_3 = \begin{bmatrix} 7.09 & -1.14 \\ -1.14 & 21.26 \end{bmatrix}$, and $P_4 = \begin{bmatrix} 28.22 & 12.67 \\ 12.67 & 38.79 \end{bmatrix}$, respectively.

Implement the value iteration algorithm for 25 iterations to reach the computation precision $\varepsilon = 0.01$. The convergence plots of the iterative value functions initialized by $\Psi^j(x_k)$, $j = 1, 2, 3, 4$, are shown in Fig. 1(a)–(d), respectively, where "In" denotes initial iterations and "Lm" denotes limiting iterations.

As $\Psi^1(x_k) \equiv 0$, we know that the value iteration (7)–(11) is reduced to the traditional value iteration algorithm [52]. In [52], it has been shown that the iterative value function is monotonically nondecreasing and converges to the optimum. For $\Psi^1(x_k)$, as $V_1(x_k) \geq V_0(x_k)$, according to Theorem 6 and Corollary 3, we know that $V_{i+1}(x_k) \geq V_i(x_k)$ $i = 0, 1, \ldots,$ and $V_i(x_k) \leq J^*(x_k)$, which can be justified from Fig. 1(a). Thus, the property of the traditional value iteration [52] can be justified by the developed value iteration (7)–(11). However, the convergence property with nonzero initial value iteration was not presented in [52]. From Fig. 1(a)–(d), initialized by an arbitrary positive semi-definite function, we can see that the iterative value function will converge to the optimum. Furthermore, for $\Psi^4(x_k)$, we have $V_1(x_k) \leq V_0(x_k)$ and then the iterative value function $V_i(x_k)$ is monotonically nonincreasing. For $i = 0, 1, \ldots,$ we have $V_i(x_k) \geq J^*(x_k)$. Based on $\Psi^j(x_k)$, $j = 1, \ldots, 4$, the plots of $U(x_k, v_i(x_k)) - (V_{i+1}(x_k) - V_i(x_k))$ are shown in Fig. 2(a)–(d), respectively. In Fig. 2, U_i denotes $U(x_k, v_i(x_k))$ and V_{i+1} and V_i denote $V_{i+1}(x_k)$ and $V_i(x_k)$, respectively.

After 25 iterations, the iterative value functions $V_i(x_k)$ satisfy the convergence criterion. From Fig. 2, we can also see that the function $U(x_k, v_i(x_k)) - (V_{i+1}(x_k) - V_i(x_k))$ are larger than zero after 25 iterations, which satisfies the admissibility criterion (54). From Theorem 9, we know that the iterative control law is admissible. Let the termination time $T_f = 60$. The trajectories of system states and control are shown in Figs. 3 and 4, respectively. We can see that the iterative states and controls are convergent to their optimums and the converged control laws are admissible.

In [52], it has been shown that the iterative value function is convergent to the optimum as $i \to \infty$, while the admissibility of the iterative control law was not guaranteed. This makes the iteration not to stop until $i \to \infty$, which is impossible to realize. From Figs. 2–4, we can see that if the convergence and admissibility criteria are satisfied, then the algorithm can be terminated and an admissible optimal control law can be obtained. Hence, we say that the developed value iteration algorithm with convergence and termination criteria possesses more potential for applications than traditional value iteration algorithm.

Policy iteration algorithm [68] is a basic iterative ADP algorithm. To show the effectiveness of the developed value iterative algorithm, in the following, comparisons with the discrete-time policy iteration algorithm will be presented. The detailed iteration process of the discrete-time policy iteration algorithm was described in [68]. To implement the

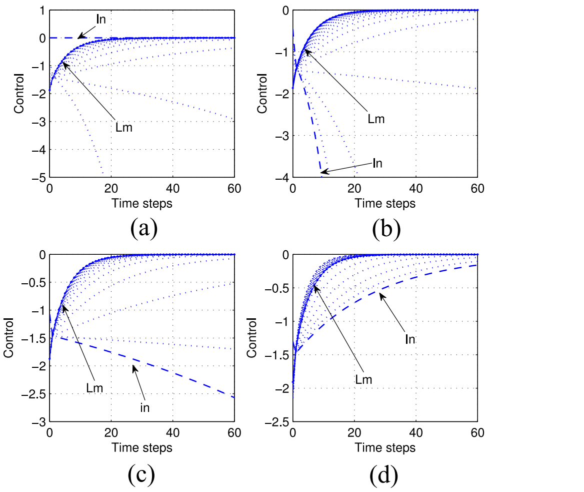

Fig. 4. Iterative control trajectories with $\Psi^j(x_k)$, $j = 1, \ldots, 4$. (a)$\Psi^1(x_k)$. (b) $\Psi^2(x_k)$. (c) $\Psi^3(x_k)$. (d) $\Psi^4(x_k)$.

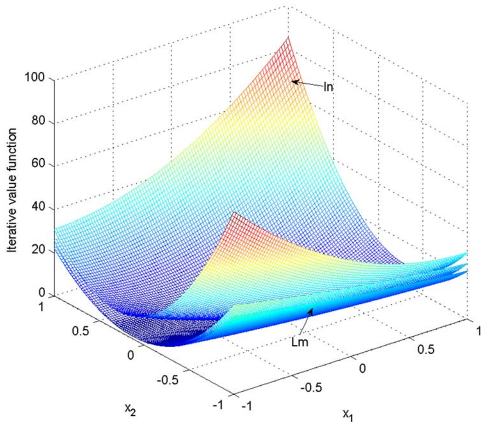

Fig. 5. Iterative value function for discrete-time policy iteration.

policy iteration algorithm, an initial admissible control law is required. We can use action network to obtain the initial admissible control law. Let the output of the action network be expressed as $\nu(x_k) = W_{a,\text{initial}}\sigma(Y_{a,\text{initial}}x_k + b_{a,\text{initial}})$, where $W_{a,\text{initial}}$ is the weight matrix between the hidden layer and output layer, $Y_{a,\text{initial}}$ is the weight matrix between the input layer and hidden layer, and $b_{a,\text{initial}}$ is the threshold value. According to [68, Algorithm 1], the weights for the initial admissible control can be obtained as

$$Y_{a,\text{initial}} = \begin{bmatrix} -4.19 & 0.10 & -5.98 & 2.26 & 0.66 & 1.96 & 0.84 & -1.65 \\ -0.74 & 4.14 & 2.68 & -3.65 & 0.18 & -0.09 & 0.26 & 0.03 \end{bmatrix}^{\mathsf{T}}$$

$$W_{a,\text{initial}} = [0.07,\ 0.01,\ 0,\ 0,\ -2.77,\ -0.04,\ -1.59,\ 0.11]$$

and

$$b_{a,\text{initial}} = [4.36,\ 2.9,\ 3.01,\ -0.64,\ -0.48,-0.33,1,\ -1.32]^{\mathsf{T}}.$$

Initialized by $\nu(x_k)$, implementing the policy iteration algorithm for six iterations to reach the computation precision $\varepsilon = 0.01$. The convergence plots of the iterative value function are shown in Fig. 5. The corresponding trajectories of iterative states and controls are shown in Fig. 6(a) and (b), respectively.

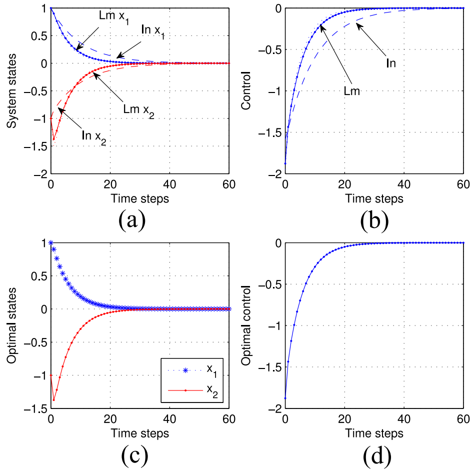

Fig. 6. States and control trajectories. (a) Iterative states for discrete-time policy iteration. (b) Iterative controls for discrete-time policy iteration. (c) Optimal states. (d) Optimal control.

From the simulation results we can see that the iterative value functions are convergent to the optimum after six iterations, which is the same as the value iteration algorithm in this paper. The optimal state and control trajectories are shown in Fig. 6(c) and (d), respectively. Thus, the effectiveness of the developed value iteration algorithm can be justified. However, we say that the iteration process between the developed value iteration algorithm and the policy iteration algorithm [68] are inherently different. First, the policy iteration algorithm is initialized by an admissible control law, i.e., $\nu(x_k)$. The developed value iteration algorithm is initialized by a positive semi-definite function and the admissible control law is not necessary. Second, in each iteration of policy iteration algorithms, it requires solving a generalized HJB (GHJB) equation, such as

$$V_i(x_k) = U(x_k, v_i(x_k)) + V_i(x_{k+1}) \tag{69}$$

to update the iterative value function. In the developed value iteration algorithm, the GHJB equation (69) is not required. Third, we say that any of the iterative control laws in the policy iteration algorithm is admissible. For the developed value iteration algorithm, the admissibility for each iterative control laws cannot be guaranteed. However, if the admissibility criterion (54) is satisfied, then the admissibility of the iterative control law for the developed value iteration algorithm can be guaranteed. Generally, the admissible control law for nonlinear system is difficult to obtain, while the initial positive semi-definite function can easily be chosen. Thus, we can say that the developed value iteration algorithm possesses more potential for applications than policy iteration algorithm [68].

Example 2: We now examine the performance of the developed algorithm in a discretized torsional pendulum system [69]. The dynamics of the pendulum is given as follows:

$$\begin{bmatrix} x_{1(k+1)} \\ x_{2(k+1)} \end{bmatrix} = \begin{bmatrix} 0.1x_{2k} + x_{1k} \\ -\frac{0.1\text{Mgl}}{\mathcal{J}}\sin(x_{1k}) + \left(\frac{0.1 - 0.1f_d}{\mathcal{J}}\right)x_{2k} \end{bmatrix} + \begin{bmatrix} 0 \\ \left(\frac{0.1}{\mathcal{J}}\right) \end{bmatrix} u_k \tag{70}$$

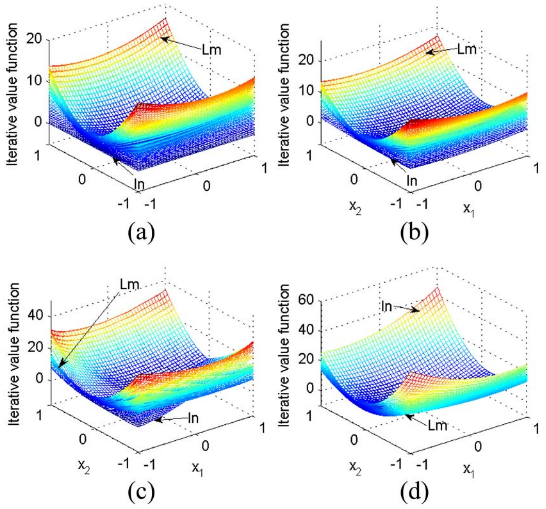

Fig. 7. Convergent curves of iterative value functions with $\tilde{\Psi}^j(x_k)$, $j = 1, \ldots, 4$. (a) $\tilde{\Psi}^1(x_k)$. (b) $\tilde{\Psi}^2(x_k)$. (c) $\tilde{\Psi}^3(x_k)$. (d) $\tilde{\Psi}^4(x_k)$.

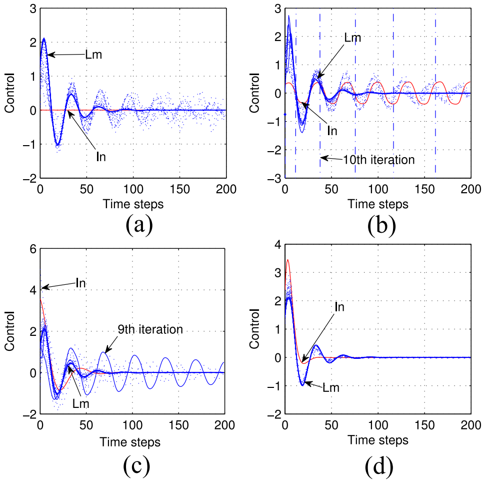

Fig. 8. Trajectories of iterative controls with $\tilde{\Psi}^j(x_k)$, $j = 1, \ldots, 4$. (a) $\tilde{\Psi}^1(x_k)$. (b) $\tilde{\Psi}^2(x_k)$. (c) $\tilde{\Psi}^3(x_k)$. (d) $\tilde{\Psi}^4(x_k)$.

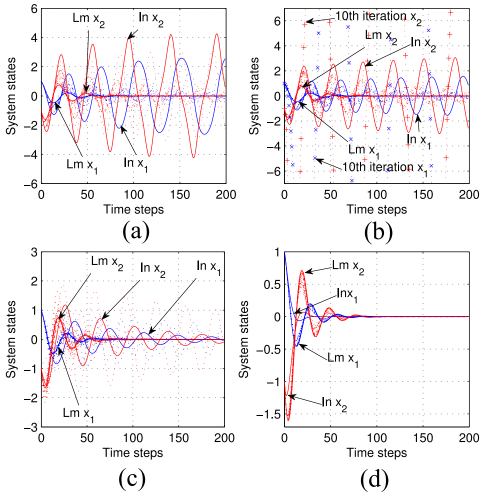

Fig. 9. Trajectories of iterative system states with $\tilde{\Psi}^j(x_k)$, $j = 1, \ldots, 4$. (a) $\tilde{\Psi}^1(x_k)$. (b) $\tilde{\Psi}^2(x_k)$. (c) $\tilde{\Psi}^3(x_k)$. (d) $\tilde{\Psi}^4(x_k)$.

where $M = 1/3$ kg and $l = 3/2$ m are the mass and length of the pendulum bar, respectively. Let $\mathcal{J} = 4/3\ Ml^2$ and $f_d = 0.2$ be the rotary inertia and frictional factor, respectively. Let $g = 9.8$ m/s^2 be the gravitational acceleration. Let the initial state be $x_0 = [1, -1]^\mathsf{T}$. The utility function is chosen the same as Example 1 with $Q = 0.2I_1$ and $R = 0.2I_2$.

Neural networks are also used to implement the present value iteration algorithm, where the structures of the critic network and the action network are the same as the ones in Example 1. We choose $p = 20\,000$ states in Θ to implement the developed value iteration algorithm to obtain the optimal control law. For each iteration step, the critic network and the action network are trained for $10\,000$ steps using the learning rate of 0.005 so that the neural network training errors become less than 10^{-6}. To illustrate the effectiveness of the algorithm, we also choose four different initial value functions with the form $\tilde{\Psi}^j(x_k) = x_k^\mathsf{T}\tilde{P}_j x_k$, $j = 1, 2, 3, 4$. Let $\tilde{P}_1 = 0$. Let $\tilde{P}_2$–$\tilde{P}_4$ be positive matrices given by $\tilde{P}_2 = \begin{bmatrix} 0.52 & -0.26 \\ -0.26 & 0.80 \end{bmatrix}$, $\tilde{P}_3 = \begin{bmatrix} 10.30 & -7.61 \\ -7.61 & 5.74 \end{bmatrix}$, and $\tilde{P}_4 = \begin{bmatrix} 27.82 & 4.93 \\ 4.93 & 7.53 \end{bmatrix}$. Implement the value iteration algorithm for 40 iterations, and the convergence curves of the iterative value functions initialized by $\tilde{\Psi}^j(x_k)$, $j = 1, 2, 3, 4$, are shown in Fig. 7(a)–(d), respectively.

After 40 iterations, we can see that the iterative value functions converge to the optimum. For $\tilde{\Psi}^1(x_k)$, the developed value iterative algorithm is reduced to the traditional one [52]. As $V_1(x_k) \geq V_0(x_k)$, we have that the iterative value function $V_i(x_k)$ is monotonically nondecreasing and for $i = 0, 1, \ldots,$ $V_i(x_k) \leq J^*(x_k)$. Then the convergence property of the traditional value iteration algorithm [52] can be justified by our developed algorithm.

On the other hand, for $\tilde{\Psi}^4(x_k)$, we have $V_1(x_k) \leq V_0(x_k)$ and then the iterative value function $V_i(x_k)$ is monotonically nonincreasing and for $i = 0, 1, \ldots,$ $V_i(x_k) \geq J^*(x_k)$. The iterative control laws obtained by the value iteration algorithm, which are initialized by $\tilde{\Psi}^j(x_k)$, $j = 1, 2, 3, 4$, are displayed in Fig. 8(a)–(d), respectively. Implementing the iterative control laws to the control system (70) for $T_f = 200$ time steps, we can obtain the system state trajectories, shown in Fig. 9(a)–(d), respectively. For $\tilde{\Psi}^4(x_k)$, we have $V_1(x_k) \leq V_0(x_k)$. From Figs. 8(d) and 9(d), we can see for any $i = 0, 1, \ldots, v_i(x_k)$ is an admissible control law. For different initial value functions, all the iterative states and control converge to their optimums. The optimal system states are shown in Fig. 10(a) and the optimal control is shown in Fig. 10(b).

In this paper, it is shown that if an iterative control law $v_i(x_k)$ is admissible, it is not sure that $v_{i+j}(x_k)$ is also admissible. In order to clearly show the simulation results, we copy the state and control trajectories for ninth and tenth iterations in Figs. 8(b) and 9(b) to Fig. 11(a) and (b), respectively, where we can see that the ninth iterative control law $v_9(x_k)$ initialized by $\tilde{\Psi}^2(x_k)$ is admissible, while the next iterative control law $v_{10}(x_k)$ is not admissible. From Fig. 11(c), we can see that the iterative value function $V_{10}(x_k)$ does not satisfy (63). From Fig. 11(d), we can see that the admissibility criterion (54) is

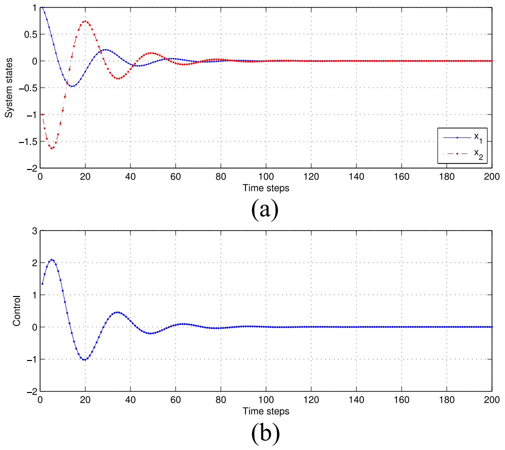

Fig. 10. Optimal trajectories. (a) Optimal states. (b) Optimal control.

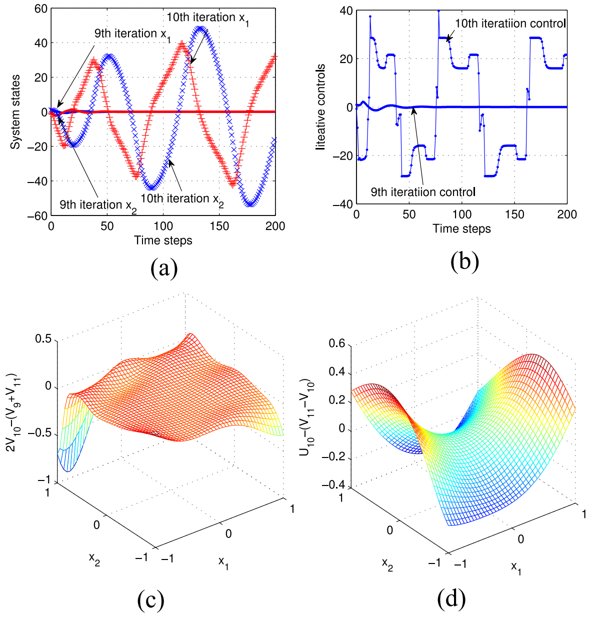

Fig. 11. Tenth simulation results initialized by $\tilde{\Psi}^2(x_k)$. (a) Iterative states. (b) Iterative controls. (c) Error of value function by (63). (d) Error obtained by (54).

not satisfied either. Hence the admissibility of $v_{10}(x_k)$ cannot be guaranteed.

We have pointed out that the traditional value iteration algorithm is terminated by convergence criterion. For $\tilde{\Psi}^3(x_k) = x_k^{\mathsf{T}} P_3 x_k$, from Fig. 12(a), we can see that $|V_{10}(x_k) - V_9(x_k)| < 0.25$. If we define $\varepsilon = 0.25$, then the iterative algorithm can stop. However, from Fig. 8(c) we can see that the iterative control $v_9(x_k)$ is not an admissible control law. We copy Fig. 9(c) to Fig. 13, where the system states under the iterative control law $v_9(x_k)$ is emphasized. We can see that the system is not stable under $v_9(x_k)$. Hence, we confirm that the convergence termination criterion cannot guarantee the admissibility property of the iterative control. From Fig. 12(b), we can see that the function $U(x_k, v_9(x_k)) - (V_{10}(x_k) - V_9(x_k))$ is not larger than zero for all x_k, which means that the admissibility criterion (54) is not satisfied. In this point of view, we say that initialized by an arbitrary positive semi-definite function, the developed value iteration algorithm can be terminated if the convergence and admissibility criteria are both satisfied. Therefore, we declare that the developed value iteration algorithm possesses more potential for applications comparing with traditional value and policy iteration algorithms.

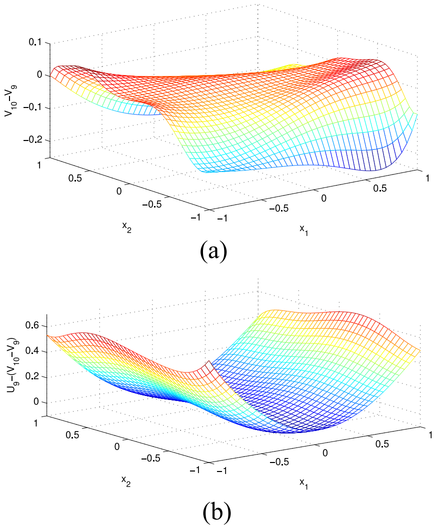

Fig. 12. Ninth simulation results initialized by $\tilde{\Psi}^3(x_k)$. (a) Iterative error. (b) Error obtained by (54).

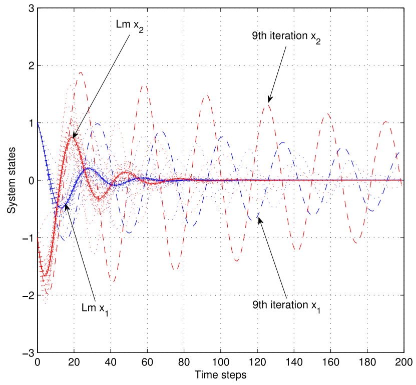

Fig. 13. Trajectories of iterative system states initialized by $\tilde{\Psi}^3(x_k)$.

V. Conclusion

In this paper, a new value iterative ADP algorithm is developed to find the infinite horizon optimal control for discrete-time nonlinear systems. It is proven that the iterative value function is convergent to the optimum under an arbitrary positive semi-definite function. For different initial value functions, the detailed convergence properties are also developed. For the first time the admissibility properties of the iterative control laws for value iteration algorithms are analyzed and new termination criteria of the value iteration algorithms are established which guarantee the effectiveness of the achieved iterative control law. Finally, two simulation examples are given to illustrate the performance of the present method.

References

[1] P. Y. Chen, S. M. Cheng, and K. C. Chen, "Optimal control of epidemic information dissemination over networks," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2316–2328, Dec. 2014.

[2] Q. Wei, D. Liu, G. Shi, and Y. Liu, "Optimal multi-battery coordination control for home energy management systems via distributed iterative adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, vol. 42, no. 7, pp. 4203–4214, Jul. 2015.

[3] S. Yin, S. X. Ding, X. Xie, and H. Luo, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6418–6428, Nov. 2014.

[4] S. Yin, G. Wang, and X. Yang, "Robust PLS approach for KPI-related prediction and diagnosis against outliers and missing data," *Int. J. Syst. Sci.*, vol. 45, no. 7, pp. 1375–1382, Jul. 2014.

[5] S. Yin, X. Zhu, and O. Kaynak, "Improved PLS focused on key-performance-indicator-related fault diagnosis," *IEEE Trans. Ind. Electron.*, vol. 62, no. 3, pp. 1651–1658, Mar. 2015.

[6] D. Molina, G. K. Venayagamoorthy, J. Liang, and R. G. Harley, "Intelligent local area signals based damping of power system oscillations using virtual generators and approximate dynamic programming," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 498–508, Feb. 2013.

[7] Q. Wei, R. Song, and P. Yan, "Data-driven zero-sum neuro-optimal control for a class of continuous-time unknown nonlinear systems with disturbance using ADP," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[8] S. Yin, X. Li, H. Gao, and O. Kaynak, "Data-based techniques focused on modern industry: An overview," *IEEE Trans. Ind. Electron.*, vol. 62, no. 1, pp. 657–667, Jan. 2015.

[9] S. Yin and Z. Huang, "Performance monitoring for vehicle suspension system via fuzzy positivistic C-means clustering based on accelerometer measurements," *IEEE/ASME Trans. Mechatronics*, vol. 20, no. 5, pp. 2613–2620, Oct. 2015.

[10] J. Li *et al.*, "Efficient video stitching based on fast structure deformation," *IEEE Trans. Cybern.*, to be published.

[11] 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.

[12] S. Li, H. Lu, and X. Shao, "Human body segmentation via data-driven graph cut," *IEEE Trans. Cybern.*, vol. 44, no. 11, pp. 2099–2108, Nov. 2014.

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

[14] H. Zhang, C. Qin, B. Jiang, and Y. Luo, "Online adaptive policy learning algorithm for H_∞ state feedback control of unknown affine nonlinear discrete-time systems," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2706–2718, Dec. 2014.

[15] A. Heydari, "Revisiting approximate dynamic programming and its convergence," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2733–2743, Dec. 2014.

[16] D. V. Prokhorov and D. C. Wunsch, "Adaptive critic designs," *IEEE Trans. Neural Netw.*, vol. 8, no. 5, pp. 997–1007, Sep. 1997.

[17] Y. Tang, H. Gao, and J. Kurths, "Distributed robust synchronization of dynamical networks with stochastic coupling," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 61, no. 5, pp. 1508–1519, May 2014.

[18] P. J. Werbos, "Advanced forecasting methods for global crisis warning and models of intelligence," *Gen. Syst. Yearbook*, vol. 22, pp. 25–38, 1977.

[19] P. J. Werbos, "A menu of designs for reinforcement learning over time," in *Neural Networks for Control*, W. T. Miller, R. S. Sutton, and P. J. Werbos, Eds. Cambridge, MA, USA: MIT Press, 1991, pp. 67–95.

[20] S. Mehraeen, T. Dierks, S. Jagannathan, and M. L. Crow, "Zero-sum two-player game theoretic formulation of affine nonlinear discrete-time systems using neural networks," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1641–1655, Dec. 2013.

[21] H. Xu and S. Jagannathan, "Neural network-based finite horizon stochastic optimal control design for nonlinear networked control systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 472–485, Mar. 2015.

[22] X. Zhong, H. He, H. Zhang, and Z. Wang, "Optimal control for unknown discrete-time nonlinear Markov jump systems using adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2141–2155, Dec. 2014.

[23] Y. Tang, H. He, J. Wen, and J. Liu, "Power system stability control for a wind farm based on adaptive dynamic programming," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 166–177, Jan. 2015.

[24] M. Palanisamy, H. Modares, F. L. Lewis, and M. Aurangzeb, "Continuous-time Q-learning for infinite-horizon discounted cost linear quadratic regulator problems," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 165–176, Feb. 2015.

[25] M. Fairbank, D. Prokhorov, and E. Alonso, "Clipping in neurocontrol by adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 10, pp. 1909–1920, Oct. 2014.

[26] Q. Wei, R. Song, and Q. Sun, "Nonlinear neuro-optimal tracking control via stable iterative Q-learning algorithm," *Neurocomputing*, vol. 168, pp. 520–528, Nov. 2015.

[27] Q. Wei, D. Liu, and F. L. Lewis, "Optimal distributed synchronization control for continuous-time heterogeneous multi-agent differential graphical games," *Inf. Sci.*, vol. 317, pp. 96–113, Oct. 2015.

[28] D. Liu, D. Wang, F.-Y. Wang, H. Li, and X. Yang, "Neural-networkbased 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.

[29] 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.

[30] R. Song, W. Xiao, H. Zhang, and C. Sun, "Adaptive dynamic programming for a class of complex-valued nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1733–1739, Sep. 2014.

[31] R. Song *et al.*, "Multiple actor-critic structures for continuous-time optimal control using input–output data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 851–865, Apr. 2015.

[32] R. Song, F. L. Lewis, Q. Wei, and H. Zhang, "Off-policy actor-critic structure for optimal control of unknown systems with disturbances," *IEEE Trans. Cybern.*, to be published.

[33] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780–1792, Jul. 2014.

[34] H. Modares, F. L. Lewis, and M. B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, Jan. 2014.

[35] S. Bhasin *et al.*, "A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems," *Automatica*, vol. 49, no. 1, pp. 82–92, Jan. 2013.

[36] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, "Adaptive dynamic programming," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 32, no. 2, pp. 140–153, May 2002.

[37] Q. Wei and D. Liu, "Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 4, pp. 1020–1036, Oct. 2014.

[38] H. Zhang, Q. Wei, and D. Liu, "An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games," *Automatica*, vol. 47, no. 1, pp. 207–214, Jan. 2011.

[39] Y. Jiang and Z. P. Jiang, "Robust adaptive dynamic programming and feedback stabilization of nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 882–893, May 2014.

[40] Q. Wei, D. Liu, and X. Yang, "Infinite horizon self-learning optimal control of nonaffine discrete-time nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 4, pp. 866–879, Apr. 2015.

[41] 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.

[42] 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.

[43] 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.

[44] Q. Wei and D. Liu, "Numerical adaptive learning control scheme for discrete-time nonlinear systems," *IET Control Theory Appl.*, vol. 7, no. 11, pp. 1472–1486, Jul. 2013.

[45] Q. Wei and D. Liu, "An iterative ϵ-optimal control scheme for a class of discrete-time nonlinear systems with unfixed initial state," *Neural Netw.*, vol. 32, no. 6, pp. 236–244, Aug. 2012.

[46] Q. Wei, D. Wang, and D. Zhang, "Dual iterative adaptive dynamic programming for a class of discrete-time nonlinear systems with time-delays," *Neural Comput. Appl.*, vol. 23, nos. 7–8, pp. 1851–1863, Dec. 2013.

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

[48] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Belmont, MA, USA: Athena Scientific, 2007.

[49] W. B. Powell, *Approximate Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2007.

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

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

[52] 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.

[53] 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.

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

[55] 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.

[56] Q. Wei, D. Liu, and G. Shi, "A novel dual iterative *Q*-learning method for optimal battery management in smart residential environments," *IEEE Trans. Ind. Electron.*, vol. 62, no. 4, pp. 2509–2518, Apr. 2015.

[57] 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, Jun. 2009.

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

[59] Q. Wei, H. Zhang, and J. Dai, "Model-free multiobjective approximate dynamic programming for discrete-time nonlinear systems with general performance index functions," *Neurocomputing*, vol. 72, nos. 7–9, pp. 1839–1848, Mar. 2009.

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

[61] J. M. Lee and J. H. Lee, "Approximate dynamic programming-based approaches for input-output data-driven control of nonlinear processes," *Automatica*, vol. 41, no. 7, pp. 1281–1288, Jul. 2005.

[62] L. Busoniu, R. Babuska, B. D. Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Boca Raton, FL, USA: CRC Press, 2010.

[63] 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.

[64] 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.

[65] H. K. Khalil, *Nonlinear System*. New York, NY, USA: Prentice-Hall, 2002.

[66] X. Liao, L. Wang, and P. Yu, *Stability of Dynamical Systems*. Amsterdam, The Netherlands: Elsevier Press, 2007.

[67] R. Beard, "Improving the closed-loop performance of nonlinear systems," Ph.D. dissertation, Rensselaer Polytechnic Inst., Troy, NY, USA, 1995.

[68] 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.

[69] 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.

Qinglai Wei (M'11) received the B.S. degree in automation, the M.S. degree in control theory and control engineering, and the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2002, 2005, and 2008, respectively.

From 2009 to 2011, he was a Post-Doctoral Fellow with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, where he is currently an Associate Professor. He has authored one book, and published over 50 international journal papers. His current research interests include adaptive dynamic programming, neural-networks-based control, optimal control, and nonlinear systems and their industrial applications.

Dr. Wei was a recipient of the Outstanding Paper Award of *Acta Automatica Sinica* in 2011 and the Zhang Siying Outstanding Paper Award of Chinese Control and Decision Conference in 2015. He has been an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS since 2014 and *Acta Automatica Sinica* since 2015. He has been a Secretary of the IEEE Computational Intelligence Society Beijing Chapter since 2015. He was the Registration Chair of the 2014 IEEE World Congress on Computational Intelligence, the 2013 International Conference on Brain Inspired Cognitive Systems (BICSs), and the 8th International Symposium on Neural Networks (ISNNs). He was the Publication Chair of the 9th ISNN in 2012 and 5th International Conference on Information Science and Technology in 2015. He was the Finance Chair of the 4th International Conference on Intelligent Control and Information Processing in 2013, and the Publicity Chair of the BICS in 2012. He was a Guest Editor of several international journals.

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 Computer Science in 2006. He is currently a Full Professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China. He has published 15 books (six research monographs and nine edited volumes).

Prof. Liu was a recipient of 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 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 was selected for the "100 Talents Program" by the Chinese Academy of Sciences in 2008. He is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. He is a fellow of the International Neural Network Society.

Hanquan Lin received the B.S. degree in automation from Nankai University, Tianjin, China, in 2013. He is currently pursuing the M.S. degree in control theory and control engineering with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

He is with the University of Chinese Academy of Sciences, Beijing. His current research interests include neural networks, reinforcement learning, multiagent systems, and machine learning.