Adaptive Dynamic Programming for Discrete-Time Zero-Sum Games

Qinglai Wei, *Member, IEEE*, Derong Liu, *Fellow, IEEE*, Qiao Lin, and Ruizhuo Song, *Member, IEEE*

***Abstract*— In this paper, a novel adaptive dynamic programming (ADP) algorithm, called "iterative zero-sum ADP algorithm," is developed to solve infinite-horizon discrete-time two-player zero-sum games of nonlinear systems. The present iterative zero-sum ADP algorithm permits arbitrary positive semidefinite functions to initialize the upper and lower iterations. A novel convergence analysis is developed to guarantee the upper and lower iterative value functions to converge to the upper and lower optimums, respectively. When the saddle-point equilibrium exists, it is emphasized that both the upper and lower iterative value functions are proved to converge to the optimal solution of the zero-sum game, where the existence criteria of the saddle-point equilibrium are not required. If the saddle-point equilibrium does not exist, the upper and lower optimal performance index functions are obtained, respectively, where the upper and lower performance index functions are proved to be not equivalent. Finally, simulation results and comparisons are shown to illustrate the performance of the present method.**

***Index Terms*— Adaptive critic designs, adaptive dynamic programming (ADP), approximate dynamic programming, neurodynamic programming, optimal control, zero-sum game.**

I. Introduction

A LARGE class of real systems are controlled by more than one controller or decision maker with each using an individual strategy. These controllers often operate in a group with a general performance index function as a game [1]–[6]. Two-player zero-sum games, capturing two players' behaviors in which the success of one player in selecting strategies depends strictly on the choices of the other player, have been widely applied to decision making problems [7]–[9]. In these situations, many control schemes are presented in order to reach some form of optimality [10], [11]. Traditional approaches to deal with zero-sum games are to find the optimal solution or the saddle-point equilibrium of the games. There are many works discussing the existence criteria of the saddle-point equilibrium of zero-sum games [12]–[14]. In real-world applications, however, the existence criteria of the saddle-point equilibrium for zero-sum games are so difficult to satisfy that many applications of the zero-sum games are limited to linear systems [15]–[18]. For many zero-sum games of nonlinear systems, it is generally assumed that the saddle-point equilibrium exists, which is guaranteed by the assumption of L_2 gain [12], [19], and then, the optimal control of the zero-sum games for the nonlinear systems is obtained. Unfortunately, for real-world zero-sum games, especially for nonlinear systems, the L_2 gain cannot generally be guaranteed, which means the existence assumption of the saddle-point equilibrium for the zero-sum games cannot be realized. Thus, traditional optimal control for zero-sum games of nonlinear systems is actually difficult to apply. Therefore, a new method is necessary to obtain the saddle-point equilibrium of the zero-sum games without the complex existence criteria.

Dynamic programming is a systematic method for addressing dynamic optimization and optimal control problems [20]–[22]. However, due to the "curse of dimensionality" [23], it is often computationally untenable to run dynamic programming to obtain the optimal solution. Adaptive dynamic programming (ADP), proposed by Werbos [24], [25], overcomes the curse of dimensionality problem in dynamic programming by approximating the performance index function forward-in-time and becomes an important brainlike intelligent method of approximate optimal control for nonlinear systems [26]–[40]. Iterative methods, which include value and policy iterations [41]–[55], respectively, are primary tools in ADP to solve optimal zero-sum games [7], [56]. In [57], ADP was derived to solve the discrete-time zero-sum game for linear systems with applications to H_∞ control, in which the state and action spaces were continuous. In [58], ADP was used along with two-player policy iterations to solve the feedback strategies of a continuous-time zero-sum game that appeared in L_2-gain optimal control of nonlinear systems affine in inputs with the control policy having saturation constraints. In [59], a near optimal solution for discrete-time affine nonlinear control systems in the presence of partially unknown internal system dynamics and disturbances was solved by a zero-sum two-player ADP method, where the disturbance was considered as a control input of the system. In [60], an online adaptive policy learning algorithm based on ADP was proposed for learning the real-time solution to zero-sum games, which appeared in the H_∞ control problem. In [18], an online robust ADP algorithm was proposed for two-player zero-sum games of continuous-time

Manuscript received April 17, 2016; revised August 23, 2016; accepted September 6, 2016. Date of publication January 27, 2017; date of current version March 15, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61233001, Grant 61273140, Grant 61374105, Grant 61503379, Grant 61304079, Grant 61673054, Grant 61533017, and Grant U1501251, in part by the Fundamental Research Funds for the Central Universities under Grant FRF-TP-15-056A3, and in part by the Open Research Project from SKLMCCS under Grant 20150104.

Q. Wei and Q. Lin are with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the University of the Chinese Academy of Sciences, Beijing 100049, China (e-mail: qinglai.wei@ia.ac.cn; linqiao2014@ia.ac.cn).

D. Liu and R. Song are with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: derong@ustb.edu.cn; ruizhuosong@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/TNNLS.2016.2638863

unknown linear systems with matched uncertainties. Although ADP has achieved more and more attentions on solving zero-sum games, it is worth mentioning that the existing ADP algorithms, i.e., the traditional zero-sum ADP algorithms, possess inherent shortcomings, which are difficult for real applications.

First, linear and affine nonlinear systems with quadratic utility functions were generally considered in traditional zero-sum ADP algorithms [18], [57]–[59]. The zero-sum ADP algorithms for other types of systems were seldom considered. Second, we point out that nearly all the traditional zero-sum ADP algorithms [7], [18], [56]–[59] required satisfying the L_2-gain, which guarantees the existence of the saddle-point equilibrium of zero-sum games. However, the L_2-gain is difficult to satisfy in real applications, which causes the existence justification of the saddle-point equilibrium invalid. To the best of our knowledge, only in [61], the zero-sum ADP for continuous-time affine nonlinear systems with quadratic utility functions was proposed where the L_2-gain was not considered. The research on the zero-sum ADP algorithm for discrete-time nonlinear systems with a general performance index function has not been considered. This motivates our research.

In this paper, a new iterative zero-sum ADP algorithm is developed to solve infinite-horizon optimal control problems for discrete-time two-player zero-sum games of general nonlinear systems with general form utility functions. Initialized by arbitrary positive semidefinite functions for the upper and lower iterations, the upper and lower iterative value functions using the iterative zero-sum ADP algorithm can reach the upper and lower optimums of the zero-sum games, which satisfy the upper and lower Isaacs equations, respectively. A novel convergence analysis method is developed to show that the upper and lower iterative value functions converge to the upper and lower optimums, respectively. Optimality of the iterative zero-sum ADP algorithm will be presented. If the saddle-point equilibrium of the zero-sum game exists, we emphasize that both the upper and lower iterative value functions will converge to the optimal solution of the zero-sum game, where the existence criteria of the saddle-point equilibrium in the traditional zero-sum ADP algorithms are not required. Monotonicity of the upper and lower iterative value functions by the iterative zero-sum ADP algorithm is also presented. It is shown that under some mild constraints of the initial functions, the upper and lower iterative value functions can be monotonically nonincreasing, monotonically nondecreasing, or nonmonotonic and converge to their optimums. If the saddle-point equilibrium does not exist, the upper and lower optimal performance index functions are obtained, respectively, where it is proved that the converged upper and lower performance index functions are not equivalent. Finally, simulation results and comparisons are shown to illustrate the performance of the present method.

II. Problem Formulations

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

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

where $x_k \in \mathbb{R}^n$ is the state vector, and $u_k \in \mathbb{R}^m$ and $w_k \in \mathbb{R}^l$ are the control vectors of Players I and II, respectively. $F(x_k, u_k, w_k)$ is the system function. Let x_0 be the initial state. For convenience of analysis, results of system (1) are based on the following assumption.

Assumption 1: System (1) is controllable on a compact set $\Omega_x \subset \mathbb{R}^n$ containing the origin; the function $F(x_k, u_k, w_k)$ is Lipschitz continuous for x_k, u_k, and w_k; $x_k = 0$ is an equilibrium state of system (1) under the controls $u_k = 0$ and $w_k = 0$, i.e., $F(0, 0, 0) = 0$; the feedback control laws $u(x_k)$ and $w(x_k)$ are both continuous on Ω_x, such that $u_k = u(x_k) = 0$ and $w_k = w(x_k) = 0$, respectively, for $x_k = 0$.

Let $\mathcal{U}$ and $\mathcal{W}$ denote policy spaces of Players I and II, respectively. Let $u \in \mathcal{U}$ and $w \in \mathcal{W}$ be the control laws of Players I and II, respectively. Then, the infinite-horizon performance index function $J : \mathcal{U} \times \mathcal{W} \to \mathbb{R}$ for state x_0 can be defined as

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

where $u_k = u(x_k)$, $w_k = w(x_k)$, and we let the utility function $U(x_k, u_k, w_k)$ be a continuous function for x_k, u_k, and w_k, which is positive definite for x_k and u_k, and negative definite for w_k. The triplet $\{J; \mathcal{U}, \mathcal{W}\}$ constitutes the *normal form* of the zero-sum dynamic game (zero-sum game in brief) [12], in the context of which we can introduce the notion of a saddle-point equilibrium.

Definition 1: Given a zero-sum dynamic game $\{J; \mathcal{U}, \mathcal{W}\}$, a pair of control laws $(u^*, w^*) \in \mathcal{U} \times \mathcal{W}$ constitute a saddle-point solution [12] if, for all $(u, w) \in \mathcal{U} \times \mathcal{W}$

$$J(x_k, u^*, w) \le J^*(x_k) := J(x_k, u^*, w^*) \le J(x_k, u, w^*). \tag{3}$$

$J^*(x_k)$ is the optimal performance index function of the game.

Given a zero-sum game $\{J; \mathcal{U}, \mathcal{W}\}$ in a normal form, we define the upper optimal performance index function as

$$\overline{J}^*(x_k) = \min_{u \in \mathcal{U}} \max_{w \in \mathcal{W}} J(x_k, u, w) \tag{4}$$

and the lower optimal performance index function can be defined as

$$\underline{J}^*(x_k) = \max_{w \in \mathcal{W}} \min_{u \in \mathcal{U}} J(x_k, u, w) \tag{5}$$

with the obvious inequality [10]–[13] $\underline{J}^*(x_k) \le \overline{J}^*(x_k)$. If the optimal performance index function exists, then we have

$$\overline{J}^*(x_k) = \underline{J}^*(x_k) = J^*(x_k). \tag{6}$$

According to the principle of optimality [12], the upper optimal performance index function $\overline{J}^*(x_k)$ satisfies the following discrete-time Isaacs equation:

$$\overline{J}^*(x_k) = \min_{u_k} \max_{w_k} \{U(x_k, u_k, w_k) + \overline{J}^*(F(x_k, u_k, w_k))\}. \tag{7}$$

The lower optimal performance index function $\underline{J}^*(x_k)$ satisfies the following discrete-time Isaacs equation:

$$\underline{J}^*(x_k) = \max_{w_k} \min_{u_k} \{U(x_k, u_k, w_k) + \underline{J}^*(F(x_k, u_k, w_k))\}. \tag{8}$$

Generally speaking, the existence of the saddle-point equilibrium of the zero-sum game $\{J; \mathcal{U}, \mathcal{W}\}$ is difficult to justify, especially for nonlinear systems. Thus, it is nearly impossible to obtain the saddle-point equilibrium by directly solving inequality (3). In this situation, obtaining the upper and lower optimal performance index functions is a necessary method to achieve the saddle-point equilibrium. Unfortunately, the upper and lower optimal performance index functions $\overline{J}^*(x_k)$ and $\underline{J}^*(x_k)$ are also difficult to obtain, due to the difficulty for solving the discrete-time Isaacs equations (7) and (8). To overcome this difficulty, a new iterative algorithm based on ADP will be developed.

III. Iterative ADP Algorithm for Discrete-Time Zero-Sum Games

In this section, the discrete-time zero-sum game for system (1) will be solved by ADP. New convergence and monotonicity analysis methods will be established in this section. Optimality analysis will be presented to show that the upper and lower iterative value functions will converge to the optimums. If the saddle-point equilibrium of the zero-sum game exists, the upper and lower iterative value functions are shown to converge to the optimal solution of the zero-sum game, where the existence criteria of the saddle-point equilibrium can effectively be avoided.

A. Derivations of the Iterative Zero-Sum ADP Algorithm

In the developed iterative zero-sum ADP 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 $\overline{\Psi}(x_k) \geq 0$ be an arbitrary positive semidefinite function. Then, let the upper initial value function be expressed as

$$\overline{V}_0(x_k) = \overline{\Psi}(x_k). \tag{9}$$

Then, the iterative control law $\overline{\omega}_0(x_k, u_k)$ can be computed as

$$\begin{aligned}\overline{\omega}_0(x_k, u_k) &= \arg\max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_0(x_{k+1})\} \\ &= \arg\max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_0(F(x_k, u_k, w_k))\}\end{aligned} \tag{10}$$

where $\overline{V}_0(x_{k+1}) = \overline{\Psi}(x_{k+1})$. The iterative control law $\overline{v}_0(x_k)$ can be obtained by

$$\begin{aligned}\overline{v}_0(x_k) = \arg\min_{u_k} \{&U(x_k, u_k, \overline{\omega}_0(x_k, u_k)) \\ &+ V_0(F(x_k, u_k, \overline{\omega}_0(x_k, u_k)))\}.\end{aligned} \tag{11}$$

Letting $\overline{\omega}_0(x_k) = \overline{\omega}_0(x_k, \overline{v}_0(x_k))$, we can obtain the upper iterative control pair $[\overline{v}_0(x_k), \overline{\omega}_0(x_k)]$.

For $i = 1, 2, \ldots$, the upper iterative value function can be updated as

$$\begin{aligned}\overline{V}_i(x_k) &= \min_{u_k}\max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_{i-1}(F(x_k, u_k, w_k))\} \\ &= U(x_k, \overline{v}_{i-1}(x_k), \overline{\omega}_{i-1}(x_k)) \\ &\quad + \overline{V}_{i-1}(F(x_k, \overline{v}_{i-1}(x_k), \overline{\omega}_{i-1}(x_k))).\end{aligned} \tag{12}$$

For $i = 1, 2, \ldots$, the iterative control law $\overline{\omega}_i(x_k, u_k)$ for the upper iterative value function can be computed as

$$\begin{aligned}\overline{\omega}_i(x_k, u_k) &= \arg\max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_i(x_{k+1})\} \\ &= \arg\max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_i(F(x_k, u_k, w_k))\}.\end{aligned} \tag{13}$$

The iterative control law $\overline{v}_i(x_k)$ can be obtained by

$$\begin{aligned}\overline{v}_i(x_k) = \arg\min_{u_k} \{&U(x_k, u_k, \overline{\omega}_i(x_k, u_k)) \\ &+ V_i(F(x_k, u_k, \overline{\omega}_i(x_k, u_k)))\}.\end{aligned} \tag{14}$$

Letting $\overline{\omega}_i(x_k) = \overline{\omega}_i(x_k, \overline{v}_i(x_k))$, the upper iterative control pair $[\overline{v}_i(x_k), \overline{\omega}_i(x_k)]$ can be constructed.

For $x_k \in \mathbb{R}^n$, let the initial function $\underline{\Psi}(x_k) \geq 0$ be an arbitrary positive semidefinite function. Let the lower initial value function be expressed as

$$\underline{V}_0(x_k) = \underline{\Psi}(x_k). \tag{15}$$

Then, the iterative control law $\underline{v}_0(x_k, w_k)$ can be computed as

$$\begin{aligned}\underline{v}_0(x_k, w_k) &= \arg\min_{u_k} \{U(x_k, u_k, w_k) + \underline{V}_0(x_{k+1})\} \\ &= \arg\min_{u_k} \{U(x_k, u_k, w_k) + \underline{V}_0(F(x_k, u_k, w_k))\}\end{aligned} \tag{16}$$

where $\underline{V}_0(x_{k+1}) = \underline{\Psi}(x_{k+1})$. The iterative control law $\underline{\omega}_0(x_k)$ can be obtained by

$$\begin{aligned}\underline{\omega}_0(x_k) = \arg\max_{u_k} \{&U(x_k, \underline{v}_0(x_k, w_k), w_k) \\ &+ V_0(F(x_k, \underline{v}_0(x_k, w_k), w_k))\}.\end{aligned} \tag{17}$$

Letting $\underline{v}_0(x_k) = \underline{v}_0(x_k, \underline{\omega}_0(x_k))$, we can obtain the lower iterative control pair $[\underline{v}_0(x_k), \underline{\omega}_0(x_k)]$.

For $i = 1, 2, \ldots$, the lower iterative value function can be updated as

$$\begin{aligned}\underline{V}_i(x_k) &= \max_{w_k}\min_{u_k} \{U(x_k, u_k, w_k) + \underline{V}_{i-1}(F(x_k, u_k, w_k))\} \\ &= U(x_k, \underline{v}_{i-1}(x_k), \underline{\omega}_{i-1}(x_k)) \\ &\quad + \underline{V}_{i-1}(F(x_k, \underline{v}_{i-1}(x_k), \underline{\omega}_{i-1}(x_k))).\end{aligned} \tag{18}$$

Then, the iterative control law $\underline{v}_i(x_k, w_k)$ can be computed as

$$\begin{aligned}\underline{v}_i(x_k, w_k) &= \arg\min_{u_k} \{U(x_k, u_k, w_k) + \underline{V}_i(x_{k+1})\} \\ &= \arg\min_{u_k} \{U(x_k, u_k, w_k) + \underline{V}_i(F(x_k, u_k, w_k))\}.\end{aligned} \tag{19}$$

The iterative control law $\underline{\omega}_i(x_k)$ can be obtained by

$$\begin{aligned}\underline{\omega}_i(x_k) = \arg\max_{u_k} \{&U(x_k, \underline{v}_i(x_k, w_k), w_k) \\ &+ V_i(F(x_k, \underline{v}_i(x_k, w_k), w_k))\}.\end{aligned} \tag{20}$$

Letting $\underline{v}_i(x_k) = \underline{v}_i(x_k, \underline{\omega}_i(x_k))$, we can obtain the lower iterative control pair $[\underline{v}_i(x_k), \underline{\omega}_i(x_k)]$.

From the iterative zero-sum ADP algorithm (9)–(20), the upper iterative value function $\overline{V}_i(x_k)$ is used to approximate the upper optimal performance index function $\overline{J}^*(x_k)$ and the lower iterative value function $\underline{V}_i(x_k)$ is used to approximate the lower optimal performance index function $\underline{J}^*(x_k)$.

Thus, the iteration (9)–(14) can be called "upper iterative zero-sum ADP algorithm." The iteration (15)–(20) can be called "lower iterative zero-sum ADP algorithm." Therefore, when $i \to \infty$, the developed iterative zero-sum ADP algorithm should be convergent, such that $\overline{V}_i(x_k)$ and $\underline{V}_i(x_k)$ converge to their optimal ones. If the saddle-point equilibrium exists of the zero-sum game, both the upper and lower iterative value functions are expected to converge to the saddle-point equilibrium of the zero-sum game. In Section III-B, we will show such properties of the developed iterative zero-sum ADP algorithm.

B. Property Analysis

In [62], considering nonlinear systems with single control input, it was proved that the iterative value function is monotonically nondecreasing and converges to the optimum by the iterative ADP algorithm if the algorithm is initialized by a zero initial value function. For arbitrary positive semidefinite initial functions and multicontroller systems, however, the analysis method in [62] is not applicable here. In [63] and [64], a "functional bound" method was proposed for the iterative ADP algorithm with single controller. In this paper, inspired by [63] and [64], convergence analysis methods for the zero-sum ADP algorithm will be developed in this section.

Theorem 1: For the zero-sum game $\{J; \mathcal{U}, \mathcal{W}\}$ of system (1), the upper and lower optimal performance index functions $\overline{J}^*(x_k)$ and $\underline{J}^*(x_k)$ in (4) and (5) are positive definite for x_k.

Proof: According to Assumption 1, we can easily derive

$$\underline{J}^*(x_k) = \overline{J}^*(x_k) = 0 \tag{21}$$

for $x_k = 0$. Letting $w_k \equiv 0$, $k = 0, 1, \ldots$, we can get

$$J(x_k, u, 0) = \sum_{i=k}^{\infty} U(x_i, u_i, 0) > 0 \tag{22}$$

for all $x_k \neq 0$ and all $u \in \mathcal{U}$, as the utility function is positive definite for x_k and u_k.

According to (4), for all $x_k \neq 0$, the optimal upper performance index function satisfies

$$\begin{aligned} \overline{J}^*(x_k) &= \min_{u\in\mathcal{U}} \max_{w\in\mathcal{W}} J(x_k, u, w) \\ &\geq \min_{u\in\mathcal{U}} J(x_k, u, 0) > 0. \end{aligned} \tag{23}$$

According to (5), for all $x_k \neq 0$, the optimal lower performance index function satisfies

$$\begin{aligned} \underline{J}^*(x_k) &= \max_{w\in\mathcal{W}} \min_{u\in\mathcal{U}} J(x_k, u, w) \\ &= \max_{w\in\mathcal{W}} J(x_k, \underline{\mu}^*, w) \\ &\geq J(x_k, \underline{\mu}^*, 0) > 0. \end{aligned} \tag{24}$$

The proof is complete. ■

Let $\epsilon \geq 0$ be a nonnegative real number. Define a state set Ω_ϵ as

$$\Omega_\epsilon = \{x_k : x_k \in \Omega_x, \|x_k\| \leq \epsilon\}. \tag{25}$$

Then, we can derive the following corollary.

Corollary 1: Let Ω_ϵ be defined as in (25). For any $\epsilon > 0$, we can derive the following.

1) $\inf\{\overline{J}^*(x_k) : x_k \in \Omega_x \backslash \Omega_\epsilon, \epsilon > 0\} > 0$.
2) $\inf\{\underline{J}^*(x_k) : x_k \in \Omega_x \backslash \Omega_\epsilon, \epsilon > 0\} > 0$.

Lemma 1: For $i = 0, 1, \ldots$, let the upper and lower iterative value functions $\overline{V}_i(x_k)$ and $\underline{V}_i(x_k)$ be updated by (12) and (18), respectively, where $\overline{V}_0(x_k)$ and $\underline{V}_0(x_k)$ satisfy (9) and (15). Then, for any $i = 0, 1, \ldots$, the upper and lower iterative value functions $\overline{V}_i(x_k)$ and $\underline{V}_i(x_k)$ are positive definite functions of x_k.

Proof: The statement can be proved by mathematical induction. We first consider the upper iterative value function. For $i = 0$, according to (13), we have

$$\overline{V}_1(x_k) = \min_{u_k} \max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_0(x_{k+1})\}. \tag{26}$$

According to Assumption 1, it is easy to know $\overline{V_1}(x_k) = 0$ for $x_k = 0$. For any $x_k \neq 0$, we have

$$\begin{aligned} \overline{V}_1(x_k) &= \min_{u_k} \max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_0(x_{k+1})\} \\ &\geq \min_{u_k} \{U(x_k, u_k, 0) + \overline{V}_0(F(x_k, u_k, 0))\} \\ &> 0. \end{aligned} \tag{27}$$

Assume that the statement holds for $i = l-1$, $l = 1, 2, \ldots$, i.e., $\overline{V}_{l-1}(x_k) > 0$, $\forall x_k \neq 0$ and $\overline{V}_{l-1}(x_k) = 0$, for $x_k = 0$.

Then, for $i = l$, according to Assumption 1, we have $\overline{V}_{l+1}(x_k) = 0$ and $x_k = 0$. For all $x_k \neq 0$, we can get

$$\begin{aligned} \overline{V}_{l+1}(x_k) &= \min_{u_k} \max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_l(x_{k+1})\} \\ &\geq \min_{u_k} \{U(x_k, u_k, 0) + \overline{V}_l(F(x_k, u_k, 0))\} \\ &> 0. \end{aligned} \tag{28}$$

Thus, we have $\overline{V}_{i+1}(x_k)$ is positive definite for any $i = 0, 1, \ldots$ Next, we consider the lower iterative value function. For $i = 0$, according to (19), we have

$$\underline{V}_1(x_k) = \max_{w_k} \min_{u_k} \{U(x_k, u_k, w_k) + \underline{V}_0(x_{k+1})\}. \tag{29}$$

According to Assumption 1, it is easy to know $\underline{V}_1(x_k) = 0$ for $x_k = 0$. For any $x_k \neq 0$, we have

$$\begin{aligned} \underline{V}_1(x_k) &= \max_{w_k} \{U(x_k, \underline{v}_1(x_k, w_k), w_k) \\ &\qquad + \underline{V}_0(F(x_k, \underline{v}_1(x_k, w_k), w_k))\} \\ &\geq U(x_k, \underline{v}_1(x_k, 0), 0) \\ &\quad + \underline{V}_0(F(x_k, \underline{v}_1(x_k, 0), 0)) \\ &> 0. \end{aligned} \tag{30}$$

Assume that the statement holds for $i = l-1$, $l = 1, 2, \ldots$, i.e., $\underline{V}_{l-1}(x_k) > 0$, $\forall x_k \neq 0$. Then, for $i = l$, according to Assumption 1, we have $\underline{V}_{l+1}(x_k) = 0$ and $x_k = 0$. For all $x_k \neq 0$, we can get

$$\begin{aligned} \underline{V}_{l+1}(x_k) &= \max_{w_k} \{U(x_k, \underline{v}_l(x_k, w_k), w_k) \\ &\qquad + \underline{V}_l(F(x_k, \underline{v}_l(x_k, w_k), w_k))\} \\ &\geq U(x_k, \underline{v}_l(x_k, 0), 0) \\ &\quad + \underline{V}_l(F(x_k, \underline{v}_l(x_k, 0), 0)) \\ &> 0. \end{aligned} \tag{31}$$

Thus, $\underline{V}_{i+1}(x_k)$ is positive definite for any $i = 0, 1, \ldots$. The mathematical induction is complete. ■

Lemma 2: For $i = 0, 1, \ldots$, let the upper and lower iterative value functions $\overline{V}_i(x_k)$ and $\underline{V}_i(x_k)$ be updated by (12) and (18), respectively, where $\overline{V}_0(x_k)$ and $\underline{V}_0(x_k)$ satisfy (9) and (15), respectively. If for all $x_k \in \Omega_x$, $\overline{\Psi}(x_k) \geq \underline{\Psi}(x_k)$, then for any $i = 0, 1, \ldots$ we have

$$\underline{V}_i(x_k) \leq \overline{V}_i(x_k). \tag{32}$$

Proof: First, the conclusion obviously holds for $i = 0$. For $i = 1$, we have

$$\begin{aligned}\overline{V}_1(x_k) &= \min_{u_k}\max_{w_k}\ \{U(x_k, u_k, w_k) + \overline{V}_0(x_{k+1})\} \\ &\geq \max_{w_k}\min_{u_k}\ \{U(x_k, u_k, w_k) + \overline{V}_0(x_{k+1})\}. \end{aligned}\tag{33}$$

According to $\overline{\Psi}(x_k) \geq \underline{\Psi}(x_k)$ and (33), we can derive

$$\begin{aligned}\overline{V}_1(x_k) &\geq \max_{w_k}\min_{u_k}\ \{U(x_k, u_k, w_k) + \overline{V}_0(x_{k+1})\} \\ &\geq \max_{w_k}\min_{u_k}\ \{U(x_k, u_k, w_k) + \underline{V}_0(x_{k+1})\} \\ &= \underline{V}_1(x_k). \end{aligned}\tag{34}$$

Inequality (32) holds for $i = 1$. Using the mathematical induction, we can derive inequality (32) for any $i = 0, 1, \ldots$ The proof is complete. ■

From Theorem 1, the upper and lower optimal performance index functions, i.e., $\overline{J}^*(x_k)$ and $\underline{J}^*(x_k)$, are positive definite functions of x_k, which satisfy (7) and (8), respectively. On the other hand, according to Lemma 1, we can derive that the upper and lower iterative value functions, i.e., $\overline{V}_i(x_k)$ and $\underline{V}_i(x_k)$, $i = 0, 1, \ldots$, in the iterative ADP algorithm (9)–(20) are also positive definite functions of x_k. Now, we derive an important theorem.

Theorem 2: For $i = 0, 1, \ldots$, let $\overline{V}_i(x_k)$ and $[\overline{v}_i(x_k), \overline{w}_i(x_k)]$ be obtained by the upper iteration (9)–(14). If the optimal upper performance index function $\overline{J}^*(x_k)$ can be defined for all $x_k \in \Omega_x$, then the upper iterative value function $\overline{V}_i(x_k)$ converges to the upper optimal performance index function $\overline{J}^*(x_k)$ in (7) for all $x_k \in \Omega_x$, as $i \to \infty$.

Proof: The statement will be proved by the following three steps. First, let $x_k \in \Omega_\epsilon$, where Ω_ϵ is defined in (25). According to Theorem 1 and Lemma 1, as $J^*(x_k)$ and $\overline{V}_i(x_k)$, $i = 1, 2, \ldots$, are both positive definite functions of x_k, we can derive that

$$\overline{J}^*(x_k) = \overline{V}_i(x_k) = 0 \tag{35}$$

for $x_k \in \Omega_\epsilon$ and $\epsilon = 0$.

Second, for any $i = 0, 1, \ldots$, let $x_k \in \Omega_x \backslash \Omega_\epsilon$ for an arbitrary $\epsilon > 0$. For a compact set Ω_x, the upper optimal performance index function $\overline{J}^*(x_k)$ is upper bounded. For any $\epsilon > 0$ and $x_k \in \Omega_x \backslash \Omega_\epsilon$, according to Corollary 1, $\overline{J}^*(x_k)$ is bounded, i.e., $0 < \overline{J}^*(x_k) \leq \overline{\mathcal{M}}_{\sup}$, for positive constants $0 < \overline{\mathcal{M}}_{\sup} < \infty$. As $0 \leq \overline{\Psi}(x_k) < \infty$, $\forall x_k \in \Omega_x$, for all $x_k \in \Omega_x \backslash \Omega_\epsilon$ there exist a positive constant β, such that

$$\overline{\Psi}(x_k) \leq \beta \overline{J}^*(x_k) \tag{36}$$

where $1 \leq \beta < \infty$. For the control laws, $u \in \mathcal{U}$ and $w \in \mathcal{W}$, respectively. As Ω_x is a compact set, for any $x_k \in \Omega_x$, we can derive that u_k and w_k are both finite controls. Thus, there exists a constant $\eta > 0$, such that $U(x_k, u_k, w_k) > -\eta$. Then, we have

$$U(x_k, u_k, w_k) + \eta > 0. \tag{37}$$

For all $x_k \in \Omega_x \backslash \Omega_\epsilon$, there exists a constant $0 < \gamma < \infty$, such that

$$\gamma\,(U(x_k, u_k, w_k) + \eta) \geq \overline{J}^*(F(x_k, u_k, w_k)). \tag{38}$$

For $i = 0, 1, \ldots$, we will prove the following inequality:

$$\begin{aligned}\overline{V}_i(x_k) \leq &\left(1 + \frac{\beta - 1}{(1 + \gamma^{-1})^i}\right)\overline{J}^*(x_k) \\ &+ \left(\sum_{j=1}^{i} \frac{1}{(1 + \gamma^{-1})^j}\right)(\beta - 1)\eta \end{aligned}\tag{39}$$

for all $x_k \in \Omega_x \backslash \Omega_\epsilon$, where $\sum_j^i (\cdot) = 0$ for $j > i$.

Mathematical induction is employed to prove the conclusion. According to (9) and (36), inequality (39) obviously holds for $i = 0$. Now, let $i = 1$. We have

$$\begin{aligned}\overline{V}_1(x_k) &= \min_{u_k}\max_{w_k}\ \{U(x_k, u_k, w_k) + \overline{V}_0(F(x_k, u_k, w_k))\} \\ &\leq \min_{u_k}\max_{w_k}\ \{U(x_k, u_k, w_k) + \beta\overline{J}^*(F(x_k, u_k, w_k))\} \\ &\leq \min_{u_k}\max_{w_k}\Big\{U(x_k, u_k, w_k) + \beta\overline{J}^*(F(x_k, u_k, w_k)) \\ &\qquad + \frac{\beta - 1}{1 + \gamma^{-1}}(U(x_k, u_k, w_k) + \eta) \\ &\qquad + \frac{\beta - 1}{1 + \gamma}\overline{J}^*(F(x_k, u_k, w_k))\Big\} \\ &= \left(1 + \frac{\beta - 1}{1 + \gamma^{-1}}\right) \\ &\quad \times \min_{u_k}\max_{w_k}\ \{U(x_k, u_k, w_k) + \overline{J}^*(F(x_k, u_k, w_k))\} \\ &\quad + \frac{\beta - 1}{1 + \gamma^{-1}}\eta \\ &= \left(1 + \frac{\beta - 1}{1 + \gamma^{-1}}\right)\overline{J}^*(x_k) + \frac{\beta - 1}{1 + \gamma^{-1}}\eta. \end{aligned}\tag{40}$$

Assume that the conclusion holds for $i = \ell - 1$, $\ell = 1, 2, \ldots$ Then, for $i = \ell$, we can obtain inequality (41), as shown at the top of next page.

Thus, inequality (39) holds for $i = \ell$. The mathematical induction is complete. According to (39), letting $i \to \infty$, for $x_k \in \Omega_x \backslash \Omega_\epsilon$, we can obtain

$$0 < \lim_{i\to\infty} \overline{V}_i(x_k) \leq \overline{J}^*(x_k) + \gamma\,(\beta - 1)\eta. \tag{42}$$

It means that the upper iterative value function is upper bounded as $i \to \infty$.

Third, we will prove that the upper iterative value function is a sum of series of positive terms. According to Lemma 1, for $i = 1, 2, \ldots$, as $\overline{V}_i(x_k)$ is positive definite of x_k, there exists a function $0 < \lambda_i(x_k, u_k, w_k) < \infty$, $\forall x_k, u_k, w_k$,

$$
\begin{aligned}
\overline{V}_{\ell}(x_k) &= \min_{u_k}\max_{w_k}\ \{U(x_k,u_k,w_k)+\overline{V}_{\ell-1}(x_{k+1})\}\\
&\le \min_{u_k}\max_{w_k}\left\{U(x_k,u_k,w_k)+\left(1+\frac{\beta-1}{(1+\gamma^{-1})^{\ell-1}}\right)\overline{J}^*(F(x_k,u_k,w_k))+\left(\sum_{j=1}^{\ell-1}\frac{1}{(1+\gamma^{-1})^j}\right)(\beta-1)\eta\right\}\\
&\le \min_{u_k}\max_{w_k}\left\{U(x_k,u_k,w_k)+\left(1+\frac{\beta-1}{(1+\gamma^{-1})^{\ell-1}}\right)\overline{J}^*(F(x_k,u_k,w_k))+\left(\sum_{j=1}^{\ell-1}\frac{1}{(1+\gamma^{-1})^j}\right)(\beta-1)\eta\right.\\
&\qquad\left.+\frac{\beta-1}{(1+\gamma^{-1})^{\ell-1}(1+\gamma)}(\gamma\,(U(x_k,u_k,w_k)+\eta)-\overline{J}^*(F(x_k,u_k,w_k)))\right\}\\
&= \min_{u_k}\max_{w_k}\left\{\left(1+\frac{\beta-1}{(1+\gamma^{-1})^{\ell}}\right)(U(x_k,u_k,w_k)+\overline{J}^*(F(x_k,u_k,w_k)))\right.\\
&\qquad\left.+\left(\sum_{j=1}^{\ell-1}\frac{1}{(1+\gamma^{-1})^j}\right)(\beta-1)\eta+\frac{1}{(1+\gamma^{-1})^{\ell}}(\beta-1)\eta)\right\}\\
&= \left(1+\frac{\beta-1}{(1+\gamma^{-1})^{\ell}}\right)\ \min_{u_k}\max_{w_k}\ \{(U(x_k,u_k,w_k)+\overline{J}^*(F(x_k,u_k,w_k)))\}+\left(\sum_{j=1}^{\ell}\frac{1}{(1+\gamma^{-1})^j}\right)(\beta-1)\eta\\
&= \left(1+\frac{\beta-1}{(1+\gamma^{-1})^{\ell}}\right)\overline{J}^*(x_k)+\left(\sum_{j=1}^{\ell}\frac{1}{(1+\gamma^{-1})^j}\right)(\beta-1)\eta
\end{aligned}
\tag{41}
$$

such that

$$
\begin{aligned}
\overline{V}_{i+1}(x_k) &= U(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))+\overline{V}_i(F(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))\\
&= U(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))+\lambda_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))\\
&\quad+(\overline{V}_i(F(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k)))\\
&\quad-\lambda_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k)))
\end{aligned}
\tag{43}
$$

where $U(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))+\lambda_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))>0$ and $\overline{V}_i\big(F(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))\big)-\lambda_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))>0$.

Let $0<r_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))<1$, $\forall x_k,u_k,w_k$, be a positive function, such that

$$
\begin{aligned}
&\overline{V}_i(F(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k)))-\lambda_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))\\
&\quad=r_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))V_i(F(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))).
\end{aligned}
\tag{44}
$$

Letting

$$
\begin{aligned}
&U_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))\\
&\quad=U(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))+\lambda_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))
\end{aligned}
\tag{45}
$$

we have

$$
\begin{aligned}
\overline{V}_{i+1}(x_k) &= U(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))+\overline{V}_i(F(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))\\
&= U(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))+\lambda_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))\\
&\quad+(\overline{V}_i(F(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))\\
&\quad-\lambda_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k)))\\
&= U_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))+r_i(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k))\\
&\quad\times\overline{V}_i(F(x_k,\overline{v}_i(x_k),\overline{\omega}_i(x_k)).
\end{aligned}
\tag{46}
$$

According to (46), we can derive

$$
\begin{aligned}
\overline{V}_{i+1}(x_k) &= \sum_{j=0}^{i}\left(\prod_{l=0}^{j}r_{i+1-l}(x_{i+1-l},\overline{v}_{i+1-l}(x_{i+1-l}),\right.\\
&\qquad\left.\times\,\overline{\omega}_{i+1-l}(x_{i+1-l}))\right)\\
&\quad\times U_{i-j}(x_{k+j},\overline{v}_{i-j}(k+j),\overline{\omega}_{i-j}(k+j))
\end{aligned}
\tag{47}
$$

where we define $U_0(\cdot)=\Psi(\cdot)$. According to (47), the iterative value function $\overline{V}_{i+1}(x_k)$ is a sum of series of positive terms and forms a monotonically, which is a nondecreasing sequence. As $\overline{V}_i(x_k)$ in (42) is upper bounded for all $x_k\in\Omega_x\backslash\Omega_\epsilon$, we can derive that limit of the upper iterative value function $\overline{V}_i(x_k)$ exists, as $i\to\infty$. Letting

$$
\lim_{i\to\infty}\overline{V}_i(x_k)=\overline{V}_\infty(x_k)
\tag{48}
$$

according to (12), we can derive that

$$
\overline{V}_\infty(x_k)=\min_{u_k}\max_{w_k}\ \{U(x_k,u_k,w_k)+\overline{V}_\infty(F(x_k,u_k,w_k))\}.
\tag{49}
$$

According to (7), we can derive that $\overline{V}_\infty(x_k)=J^*(x_k)$ for all $x_k\in\Omega_x\backslash\Omega_\epsilon$. Thus, for all $x_k\in\Omega_x$, the upper iterative value function $\overline{V}_i(x_k)$ converges to $\overline{J}^*(x_k)$, $x_k\in\Omega_x$. The proof is complete. ■

Corollary 2: For $i=0,1,\ldots,$ let $\underline{V}_i(x_k)$ and $[\underline{v}_i(x_k),\underline{\omega}_i(x_k)]$ be obtained by (15)–(20). Then, if the lower optimal performance index function $\underline{J}^*(x_k)$ can be defined for all $x_k\in\Omega_x$, then the lower iterative value

function $\underline{V}_i(x_k)$ converges to the lower optimal performance index function $\underline{J}^*(x_k)$ in (8), as $i \to \infty$.

The conclusion can be proved according to the idea of (35)–(49) and the detail is omitted here.

Theorem 3: For $i = 0, 1, \ldots,$ let $\overline{V}_i(x_k)$ and $[\overline{v}_i(x_k), \overline{\omega}_i(x_k)]$ be obtained by (9)–(14). If for all $x_k \in \Omega_x$, inequality

$$\overline{V}_1(x_k) \le \overline{V}_0(x_k) \tag{50}$$

holds, then the upper iterative value function $\overline{V}_i(x_k)$ is a monotonically nonincreasing sequence for $i = 1, 2, \ldots,$ that is

$$\overline{V}_i(x_k) \le \overline{V}_{i-1}(x_k). \tag{51}$$

Proof: We prove this by mathematical induction. First, inequality (51) obviously holds for $i = 1$. We let $i = 2$. According to (12) and (50), for all $x_k \in \Omega_x$, we have

$$\begin{aligned}\overline{V}_2(x_k) &= \min_{u_k} \max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_1(x_{k+1})\} \\ &\le \min_{u_k} \max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_0(x_{k+1})\} \\ &= \overline{V}_1(x_k). \end{aligned} \tag{52}$$

Thus, inequality (51) holds for $i = 2$. Assume that the conclusion holds for $i = l$, $l = 1, 2, \ldots$ Then, for $i = l + 1$, we have

$$\begin{aligned}\overline{V}_{l+1}(x_k) &= \min_{u_k} \max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_l(x_{k+1})\} \\ &\le \min_{u_k} \max_{w_k} \{U(x_k, u_k, w_k) + \overline{V}_{l-1}(x_{k+1})\} \\ &= \overline{V}_l(x_k). \end{aligned} \tag{53}$$

Thus, inequality (51) holds for any $i = 1, 2, \ldots$ The proof is complete. ■

Corollary 3: For $i = 0, 1, \ldots,$ let $\overline{V}_i(x_k)$ and $[\overline{v}_i(x_k), \overline{\omega}_i(x_k)]$ be obtained by (9)–(14). If for all $x_k \in \Omega_x$, inequality

$$\overline{V}_1(x_k) \ge \overline{V}_0(x_k) \tag{54}$$

holds, then the upper iterative value function $\overline{V}_i(x_k)$ is a monotonically nondecreasing sequence for $i = 1, 2, \ldots,$ that is

$$\overline{V}_i(x_k) \ge \overline{V}_{i-1}(x_k). \tag{55}$$

Corollary 4: For $i = 0, 1, \ldots,$ let $\underline{V}_i(x_k)$ and $[\underline{v}_i(x_k), \underline{\omega}_i(x_k)]$ be obtained by (15)–(20). If for all $x_k \in \Omega_x$, inequality

$$\underline{V}_1(x_k) \le \underline{V}_0(x_k) \tag{56}$$

holds, then the lower iterative value function $\underline{V}_i(x_k)$ is a monotonically nonincreasing sequence for $i = 1, 2, \ldots,$ that is

$$\underline{V}_i(x_k) \le \underline{V}_{i-1}(x_k). \tag{57}$$

Corollary 5: For $i = 0, 1, \ldots,$ let $\underline{V}_i(x_k)$ and $[\underline{v}_i(x_k), \underline{\omega}_i(x_k)]$ be obtained by (15)–(20). If for all $x_k \in \Omega_x$, inequality

$$\underline{V}_1(x_k) \ge \underline{V}_0(x_k) \tag{58}$$

holds, then the lower iterative value function $\underline{V}_i(x_k)$ is a monotonically nondecreasing sequence for $i = 1, 2, \ldots,$ that is

$$\underline{V}_i(x_k) \ge \underline{V}_{i-1}(x_k). \tag{59}$$

Theorem 4: For $i = 0, 1, \ldots,$ let $\overline{V}_i(x_k)$ and $[\overline{v}_i(x_k), \overline{\omega}_i(x_k)]$ be obtained by (9)–(14). If for all $x_k \in \Omega_x$, inequality (50) holds, then the upper iterative value function $\overline{V}_i(x_k)$ satisfies

$$\overline{V}_i(x_k) \ge \overline{J}^*(x_k). \tag{60}$$

If for all $x_k \in \Omega_x$, inequality (54) holds, then the upper iterative value function $\overline{V}_i(x_k)$ satisfies

$$\overline{V}_i(x_k) \le \overline{J}^*(x_k). \tag{61}$$

Proof: According to Theorem 3, for any $i = 0, 1, \ldots,$ we have

$$\overline{V}_i(x_k) \ge \overline{V}_{i+1}(x_k) \ge \overline{V}_{i+2}(x_k) \ge \cdots \tag{62}$$

Thus, for any $i = 0, 1, \ldots,$ we can obtain

$$\overline{V}_i(x_k) \ge \lim_{l \to \infty} \overline{V}_l(x_k) = \overline{J}^*(x_k). \tag{63}$$

According to (62) and (63), we can easily derive (61). The proof is complete. ■

Corollary 6: For $i = 0, 1, \ldots,$ let $\underline{V}_i(x_k)$ and $[\underline{v}_i(x_k), \underline{\omega}_i(x_k)]$ be obtained by (15)–(20). If for all $x_k \in \Omega_x$, inequality (56) holds, then for $i = 0, 1, \ldots,$ the lower iterative value function satisfies

$$\underline{V}_i(x_k) \ge \underline{J}^*(x_k). \tag{64}$$

If for all $x_k \in \Omega_x$, inequality (58) holds, then for $i = 0, 1, \ldots,$ the lower iterative value function satisfies

$$\underline{V}_i(x_k) \le \underline{J}^*(x_k). \tag{65}$$

Remark 1: According to Theorems 2–4, the convergence and monotonicity properties of the iterative zero-sum ADP algorithm are analyzed, which guarantee that the upper and lower iterative value functions are convergent to their optimums. If the saddle-point equilibrium of the zero-sum game exists, then both the upper and lower iterative value functions are expected to converge to the optimal solution of the zero-sum game. Next, the optimality of the iterative zero-sum ADP algorithm will be analyzed.

Theorem 5: For $i = 0, 1, \ldots,$ let $\overline{V}_i(x_k)$ and $[\overline{v}_i(x_k), \overline{\omega}_i(x_k)]$ be obtained by (9)–(14) and let $\underline{V}_i(x_k)$ and $[\underline{v}_i(x_k), \underline{\omega}_i(x_k)]$ be obtained by (15)–(20). We can derive the following.

1) If the saddle-point equilibrium of the zero-sum game exists, then the upper and lower iterative value functions will both converge to the saddle-point equilibrium.
2) If the upper and lower iterative value functions converge to the same function, then the converged value function is the saddle-point equilibrium of the zero-sum game.

Proof: First, if the saddle-point equilibrium of the zero-sum game exists, then $\overline{J}^*(x_k) = \underline{J}^*(x_k) = J^*(x_k)$. According to Theorem 2 and Corollary 2, we can derive $\lim_{i\to\infty} \overline{V}_i(x_k) = \overline{J}^*(x_k)$ and $\lim_{i\to\infty} \underline{V}_i(x_k) = \underline{J}^*(x_k)$, $\forall x_k \in \Omega_x$, respectively. We can easily derive $\lim_{i\to\infty} \overline{V}_i(x_k) = \lim_{i\to\infty} \underline{V}_i(x_k) = J^*(x_k)$, $\forall x_k \in \Omega_x$.

On the other hand, if the upper and lower performance index functions converge to the same function, that is

$$\lim_{i\to\infty} \underline{V}_i(x_k) = \lim_{i\to\infty} \overline{V}_i(x_k) = J^o(x_k). \tag{66}$$

Then, we can get

$$\begin{aligned} J^o(x_k) &= \min_{u_k} \max_{w_k} \{U(x_k, u_k, w_k) + J^o(x_{k+1})\} \\ &= \max_{w_k} \min_{u_k} \{U(x_k, u_k, w_k) + J^o(x_{k+1})\} \end{aligned} \tag{67}$$

which means $J^o(x_k) = J^*(x_k)$, $\forall x_k \in \Omega_x$. The proof is complete. ∎

From Theorem 5, if the saddle point of the game exists, i.e., $\overline{V}_i(x_k) = \underline{V}_i(x_k) = J^*(x_k)$, then it is shown that the upper and lower iterative value functions will both converge to the optimum. According to Theorem 5, we can derive the following corollary.

Corollary 7: For $i = 0, 1, \ldots,$ let $\overline{V}_i(x_k)$ and $[\overline{v}_i(x_k), \overline{\omega}_i(x_k)]$ be obtained by (9)–(14) and let $\underline{V}_i(x_k)$ and $[\underline{v}_i(x_k), \underline{\omega}_i(x_k)]$ be obtained by (15)–(20). We can derive the following.

1) If the saddle-point equilibrium of the zero-sum game does not exist, then the upper and lower iterative value functions cannot converge to the same function.
2) If the upper and lower iterative value functions converge to different functions, then the saddle-point equilibrium of the zero-sum game does not exist.

Remark 2: In this paper, according to Theorem 5, if the saddle-point equilibrium of the zero-sum game exists, it is proved that both the upper and lower iterative value functions converge to the optimal solution of the zero-sum game, where the existence criteria of the saddle-point equilibrium in the traditional zero-sum ADP algorithms [7], [18], [56]–[59] are not required. We emphasize that this is an important contribution of the developed algorithm. On the other hand, if the saddle-point equilibrium does not exist, according to Corollary 7, we can derive that the upper and lower iterative value function must converge to different functions. Thus, the existence justification of the saddle-point equilibrium for the zero-sum games is not necessary for the developed iterative zero-sum ADP. This is another advantage of the developed algorithm.

IV. Simulation Example

We consider the performance of the developed algorithm in a Van der Pol's oscillator system [65] with modifications, where a new control w is added to the system. The dynamics of the modified Van der Pol's oscillator system is given as follows:

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ (1 - x_1^2)x_2 - x_1 \end{pmatrix} + Bu + Cw \tag{68}$$

where $B = \left[\begin{smallmatrix} 3.5 & 0 \\ 0 & 3.5 \end{smallmatrix}\right]$ and $C = \left[\begin{smallmatrix} 4 & 0 \\ 0 & 3 \end{smallmatrix}\right]$. Discretizing the system with the sampling interval $\Delta T = 0.1$ s leads to

$$\begin{aligned} \begin{pmatrix} x_{1(k+1)} \\ x_{2(k+1)} \end{pmatrix} &= \begin{pmatrix} x_{1k} + \Delta T x_{2k} \\ -\Delta T x_{1k} + (1 + \Delta T)x_{2k} - \Delta T x_{1k}^2 x_{2k} \end{pmatrix} \\ &\quad + \Delta T B u_k + \Delta T C w_k. \end{aligned} \tag{69}$$

Let the performance index function be expressed by (2). The utility function is the quadratic form $U_1(x_k, u_k, w_k) = x_k^{\mathrm{T}} Q_1 x_k + u_k^{\mathrm{T}} R_1 u_k + w_k^{\mathrm{T}} S_1 w_k$, where $Q_1 = I_1$, $R_1 = I_2$, $S_1 = -5I_3$, and I_1, I_2, and I_3 denote the identity matrices with suitable dimensions. For the upper and lower iterative zero-sum ADP algorithms, we use three BP neural networks, including a critic network and two action networks, to implement the upper and lower iterative ADP algorithms, respectively. The structure of the critic network is 2–8–1. The structures of the action networks are chosen as 2–8–2 and 2–8–2, respectively. The weight updating rules of the neural networks can be seen in [66] and [67] and omitted here. For each iteration, the critic network and the action networks are trained for 3000 steps under the learning rate 0.01, so that the neural network training errors become less than 10^{-6}. For the upper iterative zero-sum ADP algorithm, the critic network and two action networks are used to approximate upper iterative value function $\overline{V}_i(x_k)$ and upper iterative control law pair $[\overline{v}_i(x_k), \overline{\omega}_i(x_k)]$, respectively. For lower iterative zero-sum ADP algorithm, the critic network and two action networks are used to approximate lower iterative value function $\underline{V}_i(x_k)$ and lower iterative control law pair $[\underline{v}_i(x_k), \underline{\omega}_i(x_k)]$, respectively. To illustrate the effectiveness of the algorithm, four different initial value functions are considered. Let the upper initial value function be the quadratic form, which are expressed by $\overline{\Phi}^j(x_k) = x_k^{\mathrm{T}} \overline{\mathcal{P}}_j x_k$, $j = 0, 1$. Let $\overline{\mathcal{P}}_0 = \left[\begin{smallmatrix} 7.98 & -1 \\ -1 & 25.97 \end{smallmatrix}\right]$ and $\overline{\mathcal{P}}_1 = \left[\begin{smallmatrix} 8.98 & 2 \\ 2 & 30 \end{smallmatrix}\right]$. Let the lower initial value function be the quadratic form, which are expressed by $\underline{\Phi}^j(x_k) = x_k^{\mathrm{T}} \underline{\mathcal{P}}_j x_k$, $j = 0, 1$. Let $\underline{\mathcal{P}}_0 = \left[\begin{smallmatrix} 24.98 & -0.5 \\ -0.5 & 9 \end{smallmatrix}\right]$ and $\underline{\mathcal{P}}_1 = 0$. Implement the iterative zero-sum ADP algorithm for 25 iterations to reach the computation precision $\varepsilon = 0.01$. The convergence plots of the upper and lower iterative value functions, i.e., $\overline{V}_i(x_k)$ and $\underline{V}_i(x_k)$, which are initialized by $\overline{\Phi}^0(x_k)$ and $\underline{\Phi}^0(x_k)$, respectively, are shown in Fig. 1(a) and (b), respectively. The differences between the upper and lower iterative value functions are shown in Fig. 1(c). We can see that differences between the upper and lower iterative value functions converge to zero. Initialized by $\overline{\Phi}^1(x_k)$ and $\underline{\Phi}^1(x_k)$, the convergence plots of the upper and lower iterative value functions are shown in Figs. 1(d) and 2(a), respectively. The differences between the upper and lower iterative value functions are shown in Fig. 2(b), which shows that the differences between the upper and lower iterative value functions converge to zero. Hence, the saddle-point equilibrium of the zero-sum game, which is the optimal performance index function of the zero-sum game, exists. The plot of the optimal performance index function is shown in Fig. 2(c).

Initialized by $\overline{\Phi}^1(x_k)$, we obtain $\overline{V}_1(x_k) \leq \overline{V}_0(x_k)$. From Fig. 1(d), the upper iterative value function is monotonically nonincreasing and converges to the optimum, which verifies Theorems 2 and 3. Initialized by $\underline{\Phi}^1(x_k)$, we obtain

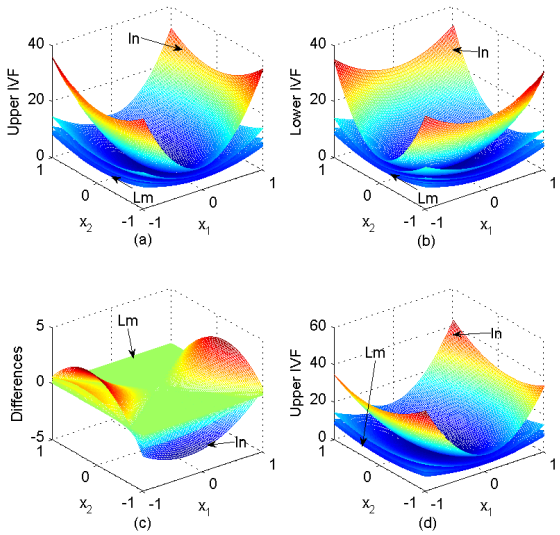

Fig. 1. Convergence plots of the iterative value functions. (a) $\overline{V}_i(x_k)$ with $\overline{\Phi}^0(x_k)$. (b) $\underline{V}_i(x_k)$ with $\underline{\Phi}^0(x_k)$. (c) Plots of $\overline{V}_i(x_k) - \underline{V}_i(x_k)$. (d) $\overline{V}_i(x_k)$ with $\overline{\Phi}^1(x_k)$.

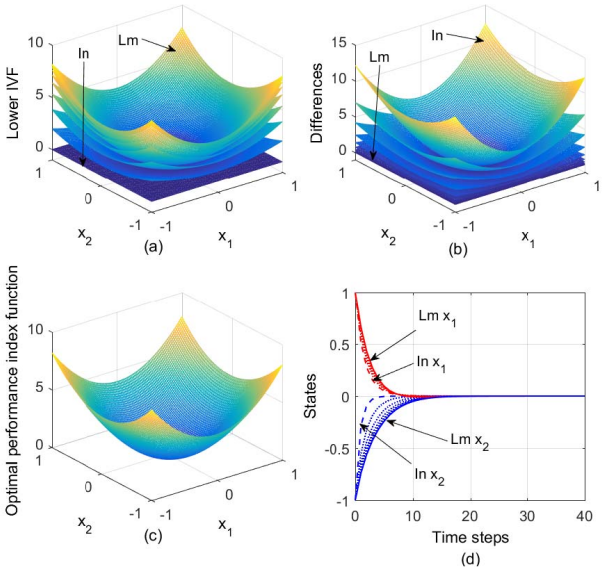

Fig. 2. Convergence plots of the iterative value functions and states. (a) $\underline{V}_i(x_k)$ with $\underline{\Phi}^1(x_k)$. (b) Plots of $\overline{V}_i(x_k) - \underline{V}_i(x_k)$. (c) Optimal performance index function. (d) States by the upper iteration with $\overline{\Phi}^0(x_k)$.

$\underline{V}_1(x_k) \geq \underline{V}_0(x_k)$. From Fig. 2(a), the lower iterative value function is monotonically nondecreasing and converges to the optimum, which verifies Theorems 2 and 3. As $\overline{\Psi}^1(x_k) \geq \underline{\Psi}^1(x_k)$, from Lemma 2, $\overline{V}_i(x_k) \geq \underline{V}_i(x_k)$, $\forall i = 0, 1, \ldots$, which can be verified from Fig. 2(b).

The state and control trajectories by the upper iteration with $\overline{\Phi}^0(x_k)$ are shown in Figs. 2(d) and 3(a) and (b), respectively. The state and control trajectories by the lower iteration with $\underline{\Phi}^1(x_k)$ are shown in Figs. 3(c) and (d) and 4(a), respectively, where the upper and lower iterative states and controls converge to their optimums. The optimal state trajectories are shown in Fig. 4(b). The optimal controls are shown in Fig. 4(c) and (d), respectively.

On the other hand, if the saddle-point equilibrium does not exist, then the traditional methods for solving the optimal

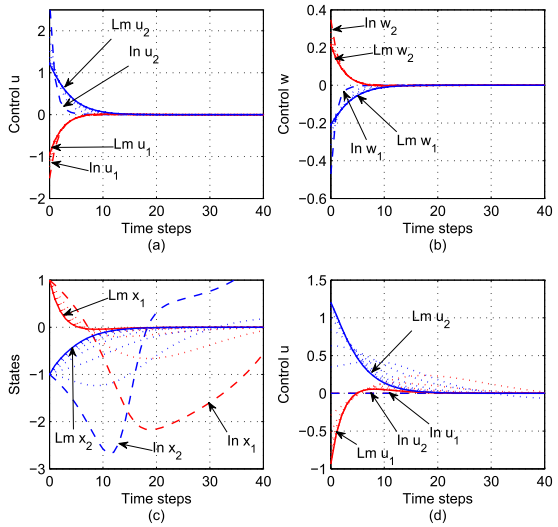

Fig. 3. Trajectories of states and controls. (a) Control u by the upper iteration with $\overline{\Phi}^0(x_k)$. (b) Control w by the upper iteration with $\overline{\Phi}^0(x_k)$. (c) States by the lower iteration with $\underline{\Phi}^1(x_k)$. (d) Control u by the lower iteration with $\underline{\Phi}^1(x_k)$.

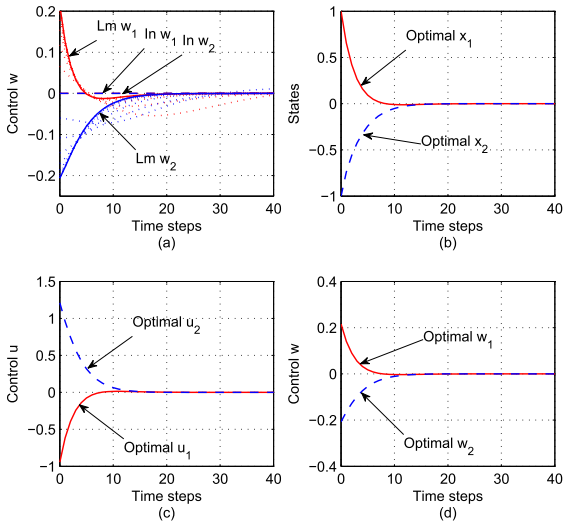

Fig. 4. Iterative and optimal trajectories. (a) Control w by the lower iteration with $\underline{\Phi}^1(x_k)$. (b) Optimal states. (c) Optimal control u. (d) Optimal control w.

control of the zero-sum game become invalid. In this situation, the developed iterative zero-sum ADP algorithm can also find the upper and lower optimal solution of the game. Now, we change the utility function to $U_4(x_k, u_k, w_k) = x_k^{\mathrm{T}} Q_4 x_k + u_k^{\mathrm{T}} R_4 u_k + w_k^{\mathrm{T}} S_4 w_k$, where $Q_4 = 0.9 I_4$, $R_4 = 0.8 I_5$, and $S_4 = -1.3 I_6$. To illustrate the effectiveness of the algorithm, we also choose four different initial value functions. Let the upper initial value function be expressed by $\overline{\Phi}^j(x_k) = x_k^{\mathrm{T}} \overline{\mathcal{P}}_j x_k$, $j = 2, 3$. Let $\overline{\mathcal{P}}_2 = \left[\begin{smallmatrix} 3 & -0.2 \\ -0.2 & 1 \end{smallmatrix}\right]$ and $\overline{\mathcal{P}}_3 = \left[\begin{smallmatrix} 10.98 & -1.75 \\ -1.75 & 8.97 \end{smallmatrix}\right]$. Let the lower initial value function be the quadratic form, which are expressed by $\underline{\Phi}^j(x_k) = x_k^{\mathrm{T}} \underline{\mathcal{P}}_j x_k$, $j = 2, 3$. Let $\underline{\mathcal{P}}_2 = \left[\begin{smallmatrix} 0.3 & 0 \\ 0 & 6 \end{smallmatrix}\right]$ and $\underline{\mathcal{P}}_3 = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0.5 \end{smallmatrix}\right]$.

Implement the iterative zero-sum ADP algorithm for 25 iterations to reach the computation precision $\varepsilon = 0.01$. The convergence plots of the upper and lower iterative value functions, i.e., $\overline{V}_i(x_k)$ and $\underline{V}_i(x_k)$, which are initialized by $\overline{\Phi}^2(x_k)$ and $\underline{\Phi}^2(x_k)$, respectively, are shown in Fig. 5(a) and (b),

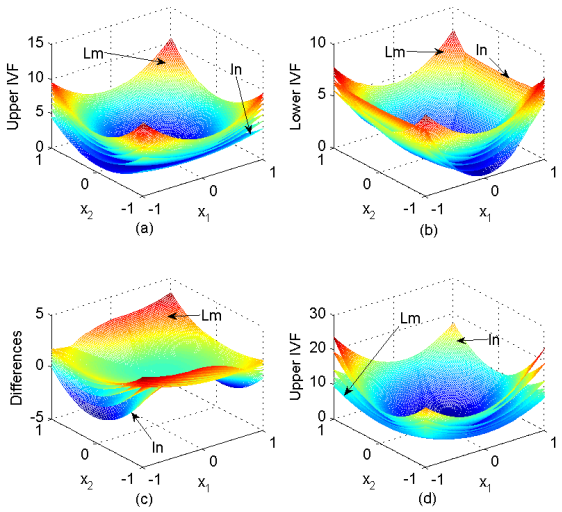

Fig. 5. Convergence plots of the iterative value functions. (a) $\overline{V}_i(x_k)$ with $\overline{\Phi}^2(x_k)$. (b) $\underline{V}_i(x_k)$ with $\underline{\Phi}^2(x_k)$. (c) Plots of $\overline{V}_i(x_k) - \underline{V}_i(x_k)$. (d) $\overline{V}_i(x_k)$ with $\overline{\Phi}^3(x_k)$.

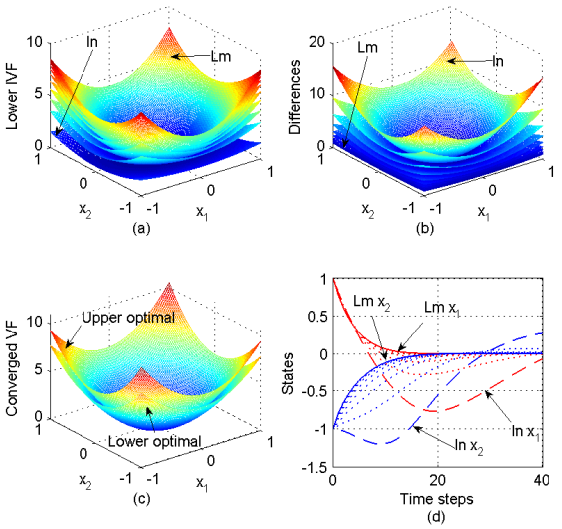

Fig. 6. Convergence plots of the iterative value functions and states. (a) $\underline{V}_i(x_k)$ with $\underline{\Phi}^3(x_k)$. (b) Plots of $\overline{V}_i(x_k) - \underline{V}_i(x_k)$. (c) Upper and lower optimal performance index functions. (d) States by upper iteration with $\overline{\Phi}^2(x_k)$.

respectively. The differences between the upper and lower iterative value functions are shown in Fig. 5(c). We can see that differences between the upper and lower iterative value functions do not converge to zero. Initialized by $\overline{\Phi}^3(x_k)$ and $\underline{\Phi}^3(x_k)$, the convergence plots of the upper and lower iterative value functions are shown in Figs. 5(d) and 6(a), respectively. The differences between the upper and lower iterative value functions are shown in Fig. 6(b), which shows that the differences between the upper and lower iterative value functions do not converge to zero. Hence, the saddle-point equilibrium of the zero-sum game does not exist. The converged upper and lower value function are shown in Fig. 6(c).

Initialized by $\overline{\Phi}^2(x_k)$, we can get $\overline{V}_1(x_k) \geq \overline{V}_0(x_k)$. According to Theorems 2 and 3, the upper iterative value function is monotonically nondecreasing and converges to the upper optimum, which can be verified by Fig. 5(a).

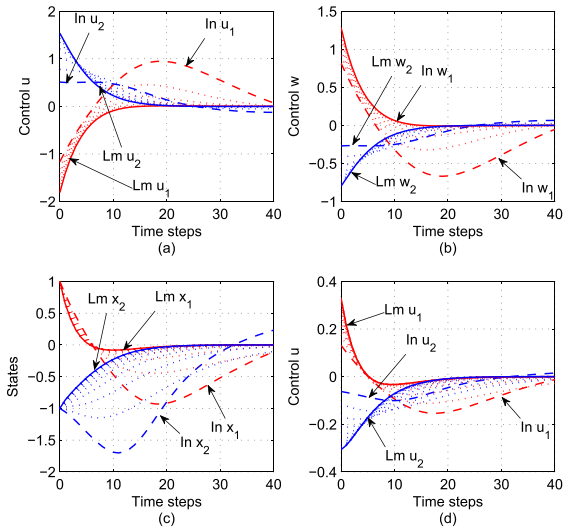

Fig. 7. Trajectories of states and controls. (a) Control u by the upper iteration with $\overline{\Phi}^2(x_k)$. (b) Control w by the upper iteration with $\overline{\Phi}^2(x_k)$. (c) States by the lower iteration with $\overline{\Phi}^3(x_k)$. (d) Control u by the lower iteration with $\overline{\Phi}^3(x_k)$.

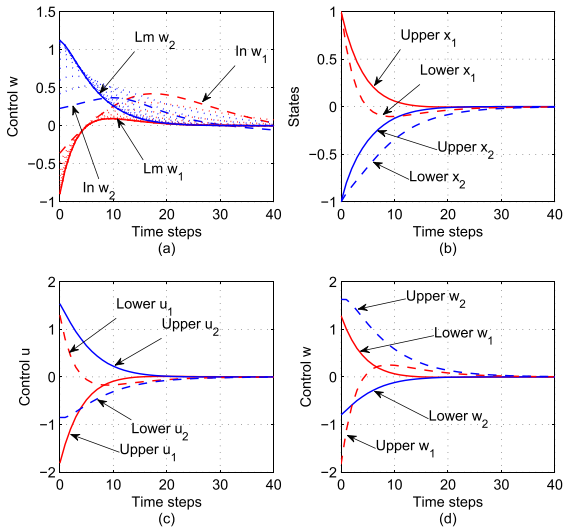

Fig. 8. Iterative and optimal trajectories. (a) Control w by the lower iteration with $\overline{\Phi}^3(x_k)$. (b) Upper and lower optimal states. (c) Upper and lower optimal control u. (d) Upper and lower optimal control w.

Initialized by $\overline{\Phi}^3(x_k)$ and $\underline{\Phi}^3(x_k)$, we can get $\overline{V}_1(x_k) \leq \overline{V}_0(x_k)$ and $\underline{V}_1(x_k) \geq \underline{V}_0(x_k)$, respectively. According to Theorems 2 and 3 and Corollary 5, the upper and lower iterative value functions are monotonically nonincreasing and monotonically nondecreasing, respectively, and converge to the upper and lower optimums, respectively. These properties can be verified by Figs. 5(d) and 6(a), respectively. The upper and lower optimal performance index functions are shown in Fig. 6(c), where the saddle-point equilibrium of the zero-sum game does not exist.

The state and control trajectories by the upper iteration with $\overline{\Phi}^2(x_k)$ are shown in Figs. 6(d) and 7(a) and (b), respectively. The state and control trajectories by the lower iteration with $\underline{\Phi}^3(x_k)$ are shown in Figs. 7(c) and (d) and 8(a), respectively, where the upper and lower iterative states and controls

converge to their optimums. However, the upper and lower optimal control pairs are not the same. The upper and lower optimal state trajectories are shown in Fig. 8(b). The upper and lower optimal controls are shown in Fig. 8(c) and (d), respectively. The differences of the state and control trajectories between the upper and lower iterations can be noticed, which also verify the nonexistence of saddle-point equilibrium of the zero-sum game.

V. Conclusion

In this paper, a new iterative ADP algorithm is developed for solving infinite-horizon optimal control problems for zero-sum games of discrete-time nonlinear systems. The iterative zero-sum ADP algorithm is separated into two iteration procedures, i.e., upper and lower iterations, for solving upper and lower optimal performance index functions, respectively. If the saddle-point equilibrium of the zero-sum game exists, it is proved that both the upper and lower iterative value functions converge to the optimal solution of the zero-sum game, where the existence criteria of the saddle-point equilibrium are not required. With some constraints on the initial upper and lower functions, the monotonicity of the upper and lower iterative value functions by the iterative zero-sum ADP algorithm can be guaranteed. If the upper and lower iterative value functions do not converge to the same function, it is proved that the saddle-point equilibrium does not exist. Finally, a simulation example is given to illustrate the performance of the present method.

References

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

[2] Q. Wei, D. Liu, G. Shi, and Y. Liu, "Multibattery optimal coordination control for home energy management systems via distributed iterative adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, vol. 62, no. 7, pp. 4203–4214, Jul. 2015.

[3] K. G. Vamvoudakis, F. L. Lewis, and G. R. Hudas, "Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality," *Automatica*, vol. 48, no. 8, pp. 1598–1611, 2012.

[4] D. Zhao, Q. Zhang, D. Wang, and Y. Zhu, "Experience replay for optimal control of nonzero-sum game systems with unknown dynamics," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 854–865, Mar. 2016.

[5] R. Song, F. L. Lewis, and Q. Wei, "Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer nonzero-sum games," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published. doi: 10.1109/TNNLS.2016.2582849.

[6] 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.

[7] Y. Fu and T. Chai, "Online solution of two-player zero-sum games for continuous-time nonlinear systems with completely unknown dynamics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 12, pp. 2577–2587, Dec. 2016, doi: 10.1109/TNNLS.2015.2496299.

[8] 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.

[9] 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, no. 13, pp. 92–100, 2013.

[10] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. New York, NY, USA: Academic, 1982.

[11] G. Owen, *Game Theory*, New York, NY, USA: Academic, 1982.

[12] T. Basar and P. Bernhard, *H∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, 2nd ed. Boston, MA, USA: Birkhäuser, 1995.

[13] D. P. Bertsekas, *Convex Analysis and Optimization*. Belmont, MA, USA: Athena Scientific, 2003.

[14] R. Goebel, "Convexity in zero-sum differential games," *SIAM J. Control Optim.*, vol. 40, no. 5, pp. 1491–1504, May 2002.

[15] H. Xu and K. Mizukami, "Linear-quadratic zero-sum differential games for generalized state space systems," *IEEE Trans. Autom. Control*, vol. 39, no. 1, pp. 143–147, Jan. 1994.

[16] J. Engwerda, "Uniqueness conditions for the affine open-loop linear quadratic differential game," *Automatica*, vol. 44, no. 2, pp. 504–511, Feb. 2008.

[17] X. Yang and J. Gao, "Linear—Quadratic uncertain differential game with application to resource extraction problem," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 4, pp. 819–826, Aug. 2016, doi: 10.1109/TFUZZ.2015.2486809.

[18] Y. Fu, J. Fu, and T. Chai, "Robust adaptive dynamic programming of two-player zero-sum games for continuous-time linear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 12, pp. 3314–3319, Dec. 2015.

[19] A. J. van der Schaft, "L_2-gain analysis of nonlinear systems and nonlinear state-feedback H_∞ control," *IEEE Trans. Autom. Control*, vol. 37, no. 6, pp. 770–784, Jun. 1992.

[20] Z. Sunberg, S. Chakravorty, and R. S. Erwin, "Information space receding horizon control for multisensor tasking problems," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1325–1336, Jun. 2016, doi: 10.1109/TCYB.2015.2445744.

[21] K. V. Berkel, B. D. Jager, T. Hofman, and M. Steinbuch, "Implementation of dynamic programming for optimal control problems with continuous states," *IEEE Trans. Control Syst. Technol.*, vol. 23, no. 3, pp. 1172–1179, May 2015.

[22] K. Deng *et al.*, "Model predictive control of central chiller plant with thermal energy storage via dynamic programming and mixed-integer linear programming," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 2, pp. 565–579, Apr. 2015.

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

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

[25] 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.

[26] 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.

[27] Y. Jiang and Z.-P. Jiang, "Global adaptive dynamic programming for continuous-time nonlinear systems," *IEEE Trans. Autom. Control*, vol. 60, no. 11, pp. 2917–2929, Nov. 2015.

[28] Z. Ni, H. He, X. Zhong, and D. V. Prokhorov, "Model-free dual heuristic dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1834–1839, Aug. 2015.

[29] H. Zargarzadeh, T. Dierks, and S. Jagannathan, "Optimal control of nonlinear continuous-time systems in strict-feedback form," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2535–2549, Oct. 2015.

[30] Z. Ni, H. He, D. Zhao, X. Xu, and D. V. Prokhorov, "GrDHP: A general utility function representation for dual heuristic dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 614–627, Mar. 2015.

[31] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 10, pp. 1513–1525, Oct. 2013.

[32] B. Kiumarsi and F. L. Lewis, "Actor—Critic-based optimal tracking for partially unknown nonlinear discrete-time systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 1, pp. 140–151, Jan. 2015.

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

[34] A. Heydari, "Feedback solution to optimal switching problems with switching cost," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2009–2019, Oct. 2016, doi: 10.1109/TNNLS.2015.2388672.

[35] R. Song, F. Lewis, Q. Wei, H.-G. Zhang, Z.-P. Jiang, and D. Levine, "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.

[36] 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.*, vol. 46, no. 5, pp. 1041–1050, May 2016, doi: 10.1109/TCYB.2015.2421338.

[37] 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.

[38] Q. Wei, D. Liu, and H. Lin, "Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems," *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 840–853, Mar. 2016.

[39] 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.

[40] 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.

[41] Q. Wei, F. L. Lewis, D. Liu, R. Song, and H. Lin, "Discrete-time local value Iteration adaptive dynamic programming: Convergence analysis," *IEEE Trans. Syst., Man, Cybern., Syst.*, to be published, doi: 10.1109/TSMC.2016.2623766.

[42] Q. Wei, D. Liu, F. L. Lewis, Y. Liu, and J. Zhang "Mixed iterative adaptive dynamic programming for optimal battery energy control in smart residential microgrids," *IEEE Trans. Ind. Electron.*, 2017.

[43] 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.

[44] H. Zhang, L. Cui, X. Zhang, and Y. Luo, "Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 2226–2236, Dec. 2011.

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

[46] D. P. Bertsekas, "Value and policy iterations in optimal control and adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published. doi: 10.1109/TNNLS.2015.2503980.

[47] 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.

[48] Q. Wei, F. L. Lewis, Q. Sun, P. Yan, and R. Song, "Discrete-time deterministic Q-learning: A novel convergence analysis," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2542923.

[49] 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.

[50] 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.

[51] Q. Wei, D. Liu, Q. Lin, and R. Song, "Discrete-time optimal control via local policy iteration adaptive dynamic programming," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2586082.

[52] Q. Wei, D. Liu, and Q. Lin, "Discrete-time local value iteration adaptive dynamic programming: Admissibility and termination analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2016.2593743.

[53] Q. Wei and D. Liu, "Numerical adaptive learning control scheme for discrete-time non-linear systems," *IET Control Theory Appl.*, vol. 7, no. 18, pp. 1472–1486, Jul. 2013.

[54] D. Wang, D. Liu, and H. Li, "Policy iteration algorithm for online design of robust control for a class of continuous-time nonlinear systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 2, pp. 627–632, Apr. 2014.

[55] 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, 2012.

[56] 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.*, vol. 27, no. 2, pp. 444–458, Feb. 2016.

[57] A. Al-Tamimi, M. Abu-Khalaf, and F. L. Lewis, "Adaptive critic designs for discrete-time zero-sum games with application to H_∞ control," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 1, pp. 240–247, Feb. 2007.

[58] M. Abu-Khalaf, F. L. Lewis, and J. Huang, "Neurodynamic programming and zero-sum games for constrained control systems," *IEEE Trans. Neural Netw.*, vol. 19, no. 7, pp. 1243–1252, Jul. 2008.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

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

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

[65] A. Heydari and S. N. Balakrishnan, "Fixed-final-time optimal tracking control of input-affine nonlinear systems," *Neurocomputing*, vol. 129, pp. 528–539, Apr. 2014.

[66] 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.

[67] 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.

Qinglai Wei (M'11) received the B.S. degree in automation, and the Ph.D. degree in control theory and control engineering, from the Northeastern University, Shenyang, China, in 2002 and 2009, 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. He is currently a Professor with the Institute of Automation, Chinese Academy of Sciences. He has authored one book, and published over 60 international journal papers. His current research interests include adaptive dynamic programming, neural-networks-based control, optimal control, nonlinear systems and their industrial applications.

Dr. Wei was a recipient of Shuang-Chuang Talents Jiangsu Province, China, in 2014, the Outstanding Paper Award of Acta Automatica Sinica in 2011, the Zhang Siying Outstanding Paper Award of Chinese Control and Decision Conference in 2015, and the Young Researcher Award of Asia Pacific Neural Network Society in 2016. He was the Registration Chair of the 12th World Congress on Intelligent Control and Automation in 2016, the 2014 IEEE World Congress on Computational Intelligence in 2014, the 2013 International Conference on Brain Inspired Cognitive Systems in 2013, and the Eighth International Symposium on Neural Networks (ISNN) in 2011. He was the Publication Chair of fifth International Conference on Information Science and Technology in 2015 and the Ninth ISNN in 2012. He was the Finance Chair of the fourth International Conference on Intelligent Control and Information Processing in 2013 and the Publicity Chair of the 2012 International Conference on Brain Inspired Cognitive Systems in 2012. He has been an Associate Editor of the IEEE TRANSACTION ON SYSTEMS MAN, AND CYBERNETICS: SYSTEMS since 2016, the *Information Sciences* since 2016, the *Neurocomputing* since 2016, the *Optimal Control Applications and Methods* since 2016, the *Acta Automatica Sinica* since 2015, and has been holding the same position for the IEEE *Transactions on Neural Networks and Learning Systems* from 2014 to 2015. He has been the Secretary of the IEEE Computational Intelligence Society Beijing Chapter, since 2015. 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 the 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, where he was a Full Professor of Electrical and Computer Engineering, and Computer Science in 2006. He was selected for the 100 Talents Program by the Chinese Academy of Sciences in 2008. He served as an Associate Director of the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 2010 to 2015. He is currently a Full Professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing. He has authored 17 books.

Dr. Liu is a fellow of the International Neural Network Society. He is an Elected Administrative Committee Member of the IEEE Computational Intelligence Society. He was a recipient of 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, the Overseas Outstanding Young Scholar Award from the National Natural Science Foundation of China in 2008, and the Outstanding Achievement Award from Asia Pacific Neural Network Assembly in 2014. He was the General Chair of the 2014 IEEE World Congress on Computational Intelligence and the 2016 World Congress on Intelligent Control and Automation. He is the Editor-in-Chief of the *Artificial Intelligence Review*.

Qiao Lin received the bachelor's degree in automatic control from the Huazhong University of Science and Technology, Wuhan, China, in 2014. She is currently pursuing the master's degree with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Her current research interests include adaptive dynamic programming, data-driven control, adaptive control, and neural network-based control.

Ruizhuo Song (M'11) received the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2012.

She is currently an Associate Professor with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China. Her current research interests include optimal control, neural-networks-based control, nonlinear control, wireless sensor networks, adaptive dynamic programming and their industrial application.