Neurocomputing 149 (2015) 1461–1468

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Stability analysis of heuristic dynamic programming algorithm for nonlinear systems

Tao Feng [a], Huaguang Zhang [a,*], Yanhong Luo [a], Jilie Zhang [b]

[a] Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110004, PR China
[b] College of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, PR China

ARTICLE INFO

Article history:
Received 17 March 2014
Received in revised form
15 August 2014
Accepted 20 August 2014
Communicated by Hongyi Li
Available online 6 September 2014

Keywords:
Convergence
Stability
Heuristic dynamic programming (HDP)
Optimal control
Value-iteration

ABSTRACT

In this paper, a value-iteration based heuristic dynamic programming (HDP) algorithm is developed to solve the optimal control for the continuous time affine nonlinear systems. First, a rigorous convergence proof of the HDP algorithm is given. Second, stability issues of the HDP algorithm for nonlinear systems are investigated. It is commonly believed that the main drawback of the HDP algorithm is that only the limit function of the iterative control sequence is proved to be stabilized, thus infinite iterations are executed. To confront this problem, we present a novel stability result for the HDP algorithm, which indicates that the resulting iterative control laws after finite iterations can guarantee the closed-loop stability. A similar stability result is also obtained for the discrete time nonlinear systems. Therefore, the practicality of the HDP algorithm is greatly improved. Single neural network (NN) structure is employed to implement the algorithm. It should be pointed that the algorithm can be implemented without knowing the internal dynamics of the systems. Finally, two numerical examples are given to demonstrate the effectiveness of the developed methods.

1. Introduction

One of the most basic design principles in the feedback control design is to guarantee the closed-loop stability of the nonlinear systems. Optimal control aims to design a feedback control law, which not only guarantees the system closed-loop stability, but also follows the optimal manner according to an overall performance index. In the past decades, a mountain of work has been done for the optimal control of nonlinear systems. Dynamic programming [1], which is proved to be a powerful method, has been extensively applied to generate the optimal control for nonlinear systems. However, one notable drawback of this method is the computing cost with the increasing dimension of the nonlinear systems, which is referred to as the "curse of dimensionality". Approximate dynamic programming (ADP) [2] methods have been proposed to circumvent this difficulty. Different from the DP methods, ADP solves the optimal control problems forward-in-time [2,3].

The optimal control for linear systems with respect to a quadric performance index can be achieved by solving the algebra Riccati equation (ARE). However, for nonlinear systems, the optimal feedback control depends on obtaining the solution to the Hamilton–Jacobi–Bellman (HJB) equation, which is challenging to solve directly due to its inherently nonlinear nature. To confront this difficulty, iterative methods have been proposed to obtain the solution of the HJB equation indirectly which can be roughly sorted into two classes [4]: policy-iteration and value-iteration. For the policy-iteration algorithm [5–11], all the iterative control laws stabilize the system, however, an initial stabilized control law is required, which is often difficult to obtain in practical applications.

For the value-iteration algorithm, an initial stabilized control law is not required. Zhang et al. [12] studied the near-optimal control for a class of discrete-time affine nonlinear systems with control constraints by the iterative DHP method. Al-Tamimi et al. [13] derived a value-iteration based HDP algorithm to solve the optimal control problems and provided a full rigorous convergence proof. In [14], the HDP algorithm has been used to solve the non-affine nonlinear systems with respect to a discounted performance index. An iterative value-iteration based ADP method has been proposed in [15] to solve a class of nonlinear zero-sum differential games. An iterative DHP algorithm has been proposed in [16] for optimally controlling a large class of nonlinear discrete-time systems affected by an unknown time variant delay and system uncertainties. In [17], Huang et al. proposed an optimal tracking control scheme based on HDP algorithm by transforming the original tracking problem into a regulation problem with respect to the state tracking error. A SN-DHP based technique has been developed in [18] to find the near optimal controller for

* Corresponding author.
E-mail addresses: sunnyfengtao@163.com (T. Feng), hgzhang@ieee.org (H. Zhang), neuluo@neu.edu.cn (Y. Luo), jilie0226@home.swjtu.edu.cn (J. Zhang).

http://dx.doi.org/10.1016/j.neucom.2014.08.046
0925-2312/

unknown affine nonlinear discrete-time systems. An on-line learning and control approach based on ADP for wind farm control and integration with the grid has been investigated in [19]. In [20], a greedy HDP algorithm has been developed to solve the zero-sum game problems for affine discrete-time systems, which can be used to solve the Hamilton–Jacobi–Isaacs equation associated with H_∞ optimal regulation control problems. The data-driven ADP methods have also received considerable attention recently. A model-free optimal control scheme for a class of linear discrete-time systems with multiple delays in state, control and output vectors has been developed in [21], where the optimal control can be obtained using only measured input/output data from systems by ADP technology. For more details, see [22–24] and references therein. However, most of the research on value-iteration focuses on the discrete-time nonlinear systems, the value-iteration based HDP algorithm for the continuous-time nonlinear systems remains unstudied. This motivates our work.

The main drawback of the value-iteration algorithm is that only the limit function of the iterative control sequence has been proved to be stabilized while the iterative control laws may be not [25]. This greatly limits the applications of the value-iteration algorithm. Li et al. [26] proposed general value-iteration (GVI) algorithm and Wei et al. [28] proposed a stable θ-ADP scheme, but the initial values of both are difficult to be obtained. Convergence of the ADP algorithm does not mean that the iterative control laws provide the closed-loop stability of the considered nonlinear systems. The closed-loop stability of the nonlinear systems must be guaranteed when the optimality is achieved. However, it is worthy noting that in the existing references, say [12–14,16–18], the optimal iterative control laws obtained by the value-iteration algorithm are indeed stabilized, rather than just the limit function of the iterative control sequence. A theoretical explanation for this phenomenon has not yet been given, to our best knowledge. In this paper, novel stability results for iterative control laws are proposed. It is proved that for the infinite horizon problem, the resulting iterative control laws after finite iterations can guarantee the closed-loop stability of the nonlinear systems, which greatly increases the practicability of the value-iteration based HDP algorithm.

The rest of the paper is organized as follows. In Section 2, the value-iteration based HDP algorithm for the continuous-time affine nonlinear systems is developed and a rigorous convergence proof is given. Novel stability results of the HDP algorithm for the continuous-time nonlinear systems are proposed. In Section 3, stability issues of the HDP algorithm for discrete-time nonlinear systems are investigated. NN implementations of the HDP algorithm are given in Section 4. Two simulation examples are employed in Section 5 to demonstrate the effectiveness of the developed methods.

2. HDP algorithm for continuous-time nonlinear systems

Consider the affine continuous-time nonlinear system of form

$$\dot{x}(t)=f(x(t))+g(x(t))u(x(t)),\quad x(0)=x_0, \tag{1}$$

where $x(t)\in\mathbb{R}^n$ is the state vector and $u(t)\in\mathbb{R}^m$ is the input vector, $f(x(t))\in\mathbb{R}^n$ and $g(x(t))\in\mathbb{R}^{n\times m}$. It is assumed that $f(x(t))+g(x(t))u(x(t))$ is Lipschitz continuous on a set $\Omega\subseteq\mathbb{R}^n$ which contains the origin, and that the dynamical system is stabilizable on Ω, which means that there exists a continuous control function $u(x(t))\in\mathbb{R}^m$ such that the system is asymptotically stable on Ω.

We consider the following quadric performance index:

$$J(x(t))=\int_t^\infty x(\tau)^TQx(\tau)+u(x(\tau))^TRu(x(\tau))\,d\tau, \tag{2}$$

where the state weighting matrix $Q\in\mathbb{R}^{n\times n}$ is nonnegative definite and the inputs weighting matrix $R\in\mathbb{R}^{m\times m}$ is positive definite. The objective is to find the control law $u(x(t))$ which minimizes the infinite-horizon cost function (2). Note that the control law $u(x(t))$ needs to be stabilized and guarantees that (2) is finite, i.e., the control law must be admissible [5].

2.1. Value-iteration based HDP algorithm for continuous-time nonlinear systems

In this subsection, we propose the value-iteration based HDP algorithm for continuous-time nonlinear systems and give the convergence proof. Note that the key difference between the HDP algorithm and the general policy-iteration algorithm with $k=1$ (which is in fact a variant of the value-iteration algorithm) in [10] is that the initial control law is not necessary stabilized.

Defining the Hamiltonian of the problem as

$$H(x(t),u(t),\partial V/\partial x)=x(t)^TQx(t)+u(x(t))^TRu(x(t))+\left(\frac{\partial V}{\partial x}\right)^T(f(x(t))+g(x(t))u(x(t))), \tag{3}$$

then we can start with an initial value $V_0(x(t))\geq 0$, and then solves for u_0 as

$$u_0(x(t))=\arg\min_{v(x(t))} H(x(t),v(x(t)),\partial V_0/\partial x), \tag{4}$$

then we update the cost function as

$$V_1(x(t))=\int_t^{t+h} x(\tau)^TQx(\tau)+u_0(x(\tau))^TRu_0(x(\tau))\,d\tau+V_0(x(t+h)), \tag{5}$$

where $h>0$ is the sampling period.

The value-iteration based HDP algorithm iterates between the following two steps:

- *Value update step*: update the value using

$$V_{i+1}(x(t))=\int_t^{t+h} x(\tau)^TQx(\tau)+u_i(x(\tau))^TRu_i(x(\tau))\,d\tau+V_i(x(t+h)). \tag{6}$$

- *Policy improvement step*: determine the improved policy using

$$u_{i+1}(x(t))=\arg\min_{v(x(t))} H(x(t),v(x(t)),\partial V_{i+1}(x(t))/\partial(x(t))). \tag{7}$$

In the above recurrent iteration, i is the iteration index. The cost function and control law are updated until they converge to the optimal values. The following convergence theorem is inspired by the innovative work of [26,27].

Theorem 1. *Suppose the condition*

$$0\leq J^*(x(t+h))\leq\theta\int_t^{t+h} x(\tau)^TQx(\tau)+u(x(\tau))^TRu(x(\tau))\,d\tau \tag{8}$$

holds uniformly for some $0<\theta<\infty$ *and that* $0\leq\delta J^*\leq V_0\leq\omega J^*$, $0\leq\delta\leq 1$, $1\leq\omega\leq\infty$. *The control law sequence* $\{u_i\}$ *and value function sequence* $\{V_i\}$ *are iteratively updated by (6) and (7). Then the value function* V_i *approaches the optimal value function* $J^*(x(t))$ *according to the inequalities*

$$\left[1+\frac{\delta-1}{(1+\theta^{-1})^i}\right]J^*(x(t))\leq V_i(x(t))\leq\left[1+\frac{\omega-1}{(1+\theta^{-1})^i}\right]J^*(x(t)). \tag{9}$$

Define $V_\infty(x(t))=\lim_{i\to\infty}V_i(x(t))$, *then* $V_\infty(x(t))=J^*(x(t))$.

Proof. The proof is given in the Appendix.

Remark 1. Let the iteration index i go to infinite, then we see

$$V_\infty(x(t)) = \min_{u(x(t))}\left\{\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau + V_\infty(x(t+h))\right\}$$
$$= \int_t^{t+h} x(\tau)^T Qx(\tau)+u_\infty(x(\tau))^T Ru_\infty(x(\tau))\,d\tau + V_\infty(x(t+h)), \tag{10}$$

where

$$u_\infty(x(t)) = \arg\min_{u(x(t))}\left\{\int_t^{t+h} x(\tau)^T Qx(\tau) + u(x(\tau))^T Ru(x(\tau))\,d\tau + V_\infty(x(t+h))\right\}. \tag{11}$$

Let $h\to 0^+$, then we obtain

$$\dot V_\infty(x(t)) = \dot V_\infty^+(x(t)) = \lim_{h\to 0^+}\frac{1}{h}[V_\infty(x(t+h))-V_\infty(x(t))]$$
$$= -[x(t)^T Qx(t)+u_\infty(x(t))^T Ru_\infty(x(t))] < 0, \tag{12}$$

where $\dot V_\infty^+(x(t))$ is the right derivative of $\dot V_\infty(x(t))$. Then we readily see that $V_\infty(x(t))$ is a Lyapunov function. By the Lyapunov stability criteria, the system (1) using the iterative control law (11) is stable. This result can be viewed as the continuous time case of that in [13].

Remark 2. Generally speaking, infinite iterations are required to obtain the optimal value $V_\infty(x(t))$, then obtain the optimal control by Eq. (4). In practical applications, since we cannot implement the iteration until $i\to\infty$, we run the algorithm with a prespecified accuracy ε to test the convergence of the cost function sequence. If $|V_{i+1}(x(t))-V_i(x(t))|<\varepsilon$, then we consider that the cost function sequence has converged sufficiently and $V_i(x(t))$ is the optimal value, $u_i(x(t))$ is the optimal control.

2.2. Stability analysis of the value-iteration based HDP algorithm

The main advantage of the value-iteration algorithm over the policy-iteration algorithm is that the initial control law needs not to be stabilized. However, the notable drawback of existing results of the value-iteration algorithm is that only the limit function of the iterative control sequences is proved to be stabilized. This can be achieved only when infinite iterations are executed, which greatly restricts its practical application. Although we can obtain the optimal value $V_i(x(t))$ and the optimal control $u_i(x(t))$, where the iteration index i is determined according to Remark 2, but whether $u_i(x(t))$ can stabilizes the nonlinear system (1) still remains unsolved. After all, $u_i(x(t))$ does not necessarily equate to $u_\infty(x(t))$. However, in the existing work [12–14,16–18], its worth noting that the resulting iterative control laws after finite iterations do stabilize the nonlinear systems when acceptable approximation precision to the optimality is achieved. The following theorem verifies this fact for the first time.

Theorem 2. *For value-iteration based HDP algorithm (6) and (7), there exists a finite iterative index i^* such that for any $i\ge i^*$, the iterative value functions $\{V_i(x(t))\}_i^\infty$ are a series of Lyapunov functions and the nonlinear system (1) using the iterative control law $u_i(x(t))$ is stable.*

Proof. It is easy to prove that $V_i(x(t))$, for any i, is a positive define differentiable function. From Theorem 1, $\lim_{i\to\infty} V_i(x(t)) = J^*(x(t))$, which implies

$$\lim_{i\to\infty}[V_{i+1}(x(t))-V_i(x(t))] = 0. \tag{13}$$

Since that

$$\lim_{i\to\infty}\int_t^{t+h} x(\tau)^T Qx(\tau)+u_i(x(\tau))^T Ru_i(x(\tau))\,d\tau = \int_t^{t+h} x(\tau)^T Qx(\tau)+u^*(x(\tau))^T Ru^*(x(\tau))\,d\tau$$

holds, and for $\forall x(t)\neq 0$ and any given sampling period h, the utility function of ith iterative control law has a lower bound

$$\int_t^{t+h} x(\tau)^T Qx(\tau)+u_i(x(\tau))^T Ru_i(x(\tau))\,d\tau > \int_t^{t+h} x(\tau)^T Qx(\tau)\,d\tau. \tag{14}$$

Then we get the conclusion that there exists a finite iterative index i^*, for any $i\ge i^*$, it holds that

$$|V_{i+1}(x(t))-V_i(x(t))| < \int_t^{t+h} x(\tau)^T Qx(\tau)+u_i(x(\tau))^T Ru_i(x(\tau))\,d\tau, \tag{15}$$

according to (6), we get

$$V_i(x(t+h))-V_i(x(t)) = V_{i+1}(x(t))-V_i(x(t)) - \int_t^{t+h} x(\tau)^T Qx(\tau)+u_i(x(\tau))^T Ru_i(x(\tau))\,d\tau$$
$$\le |V_{i+1}(x(t))-V_i(x(t))| - \int_t^{t+h} x(\tau)^T Qx(\tau)+u_i(x(\tau))^T Ru_i(x(\tau))\,d\tau$$
$$< 0. \tag{16}$$

Similarly, let $h\to 0^+$, we obtain

$$\dot V_i(x(t)) = \dot V_i^+(x(t)) = \lim_{h\to 0^+}\frac{1}{h}[V_i(x(t+h))-V_i(x(t))] < 0, \tag{17}$$

where $\dot V_i^+(x(t))$ denotes the right derivative of $V_i(x(t))$. Then we readily see that for any $i\ge i^*$, $\{V_i(x(t))\}_i^\infty$ are proved to be a series of Lyapunov functions. By the Lyapunov stability criteria, the system using the iterative control laws $\{u_i(x(t))\}_{i>i^*}^\infty$ is stable. □

Remark 3. Theorem 2 indicates that the value-iteration based HDP algorithm (6) and (7) produce a series of stabilized iterative control laws as the iterative value functions converge to the optimal value, rather than only the limit function of the iterative control sequence.

3. HDP algorithm for discrete-time nonlinear systems

Consider the affine discrete-time nonlinear system of form

$$x_{k+1} = f(x_k)+g(x_k)u(x_k),\quad k=1,2,\ldots \tag{18}$$

where $x_k\in\mathbb{R}^n$ is the state vector and $u(x_k)\in\mathbb{R}^m$ is the input vector, let x_0 be the initial state and $f+gu$ be the system function. Without loss of generality, that $x=0$ is an equilibrium state of the system under the control $u_k=0$.

It is desired for us to find an admissible control law $u(x_k)$ which minimizes the infinite-horizon cost function as follows:

$$J(x_k,u) = \sum_{j=k}^{\infty} x_j^T Qx_j + u(x_j)^T Ru(x_j), \tag{19}$$

where the state weighting matrix $Q\in\mathbb{R}^{n\times n}$ is nonnegative definite, and the inputs weighting matrix $R\in\mathbb{R}^{m\times m}$ is positive definite.

The value-iteration based HDP algorithm for discrete-time nonlinear systems is given as follows.

For any initial value $V_0(x_k)\ge 0$, solve for u_0 as follows:

$$u_0(x_k) = \arg\min_{u(x_k)}\{x_k^T Qx_k + u(x_k)^T Ru(x_k) + V_0(f(x_k)+g(x_k)u(x_k))\}.$$

Then update the cost function as

$$V_1(x_k) = x_k^T Qx_k + u_0(x_k)^T Ru_0(x_k) + V_0(f(x_k)+g(x_k)u_0(x_k)).$$

So the value-iteration based HDP algorithm iterates between the following two steps:

- *Value update step*: update the value using

$$V_{i+1}(x_k) = x_k^T Q x_k + u_i(x_k)^T R u_i(x_k) + V_i(f(x_k)+g(x_k)u_i(x_k)). \quad (20)$$

- *Policy improvement step*: determine the improved policy using

$$u_{i+1}(x_k) = \arg\min_{u(x_k)}\{x_k^T Q x_k + u(x_k)^T R u(x_k) + V_{i+1}(f(x_k)+g(x_k)u(x_k))\}. \quad (21)$$

In the above recurrent iteration, i is the iteration index, while k is the time index. The cost function and control law are updated until they converge to the optimal optimum values. The convergence proof of the algorithm has been given in [26].

Lemma 1. *Suppose the condition*

$$0 \le J^*(f(x_k)+g(x_k)u(x_k)) \le \theta U(x_k, u(x_k)) \quad (22)$$

holds uniformly for some $0<\theta<\infty$ *and that* $0 \le \delta J^* \le V_0 \le \omega J^*$, $0 \le \delta \le 1$, $1 \le \omega \le \infty$. *The control law sequence* $\{u_i\}$ *and value function sequence* $\{V_i\}$ *are iteratively updated by (20) and (21). Then the value function* V_i *approaches* J^* *according to the inequalities*

$$\left[1+\frac{\delta-1}{(1+\theta^{-1})^i}\right]J^*(x_k) \le V_i(x_k) \le \left[1+\frac{\omega-1}{(1+\theta^{-1})^i}\right]J^*(x_k). \quad (23)$$

Define $V_\infty(x_k) = \lim_{i\to\infty} V_i(x_k)$, *then* $V_\infty(x_k) = J^*(x_k)$.

According to the lemma given above, we can also prove that the value-iteration based HDP algorithm (21) and (20) produce a series of stabilized iterative control laws as the iterative value functions converge to the optimal value, which is shown in the following theorem.

Theorem 3. *For the value-iteration algorithm (20) and (21), there exists a finite* i^* *such that for* $i \ge i^*$, *the iterative value functions* $\{V_i(x_k)\}_i^\infty$ *are a series of Lyapunov functions and the system using the iterative control laws* $\{u_i(x_k)\}_i^\infty$ *is stable for any given initial value function.*

Proof. It is easy to prove that $V_i(x_k)$, $\forall i$ is a positive define function. From Lemma 1, $\lim_{i\to\infty} V_i(x_k) = J^*(x_k)$, which implies

$$\lim_{i\to\infty}[V_{i+1}(x_k) - V_i(x_k)] = 0. \quad (24)$$

Since $\lim_{i\to\infty} U(x_k, u_i(x_k)) = U(x_k, u^*(x_k))$, then for $\forall x_k \ne 0$, it holds that

$$U(x_k, u(x_k)) > x_k^T Q x_k. \quad (25)$$

Using (24), we see that there exists a finite iterative index i^*, for any $i \ge i^*$, it holds that

$$|V_{i+1}(x_k) - V_i(x_k)| < x_k^T Q x_k < U(x_k, u(x_k)). \quad (26)$$

According to (20), we have

$$V_i(x_{k+1}) - V_i(x_k) = V_{i+1}(x_k) - V_i(x_k) - U(x_k, u_i(x_k)) \le |V_{i+1}(x_k) - V_i(x_k)| - U(x_k, u_i(x_k)) < 0. \quad (27)$$

Therefore, for $i \ge i^*$, $V_i(x_k)$ is proved to be a Lyapunov function. By the Lyapunov stability criteria, the system using the iterative control laws $\{u_i(x(t))\}_{i>i^*}^\infty$ is stable. □

Remark 4. Apparently, Theorem 3 can be viewed as the generalization of the conclusion of that in [13], where only the limit function of the iterative control sequence is proved to be stabilized.

Remark 5. Consider the linear discrete-time systems, the value function is of form $V_i(x_k) = x_k^T P_i x_k$ [13], where P_i is symmetric positive definite, then we have

$$V_i(x_{k+1}) - V_i(x_k) = x_k^T P_{i+1} x_k - x_k^T P_i x_k - x_k^T Q x_k - u_i^T(x_k) R u_i(x_k) = x_k^T (P_{i+1} - P_i - Q) x_k - u_i^T(x_k) R u_i(x_k). \quad (28)$$

Obviously, for $x_k \ne 0$ and any given symmetric positive definite matrix Q, if $P_{i+1} - P_i < Q$, then $V_i(x_k) = x_k^T P_i x_k$ is a Lyapunov function. This condition is very easy to satisfy since $P_i \to P^*$ as $i \to +\infty$, i.e., $\|P_{i+1} - P_i\| \to 0$.

4. Implementation of HDP algorithm via single neural network

It is well known that neural networks can be used to approximate smooth functions on prescribe compact sets [29,30]. The cost function $V_i(x(t))$ is approximated at each step by a critic neural network

$$\hat{V}_i(x(t)) = (W_i^L)^T \phi(x(t)) = \sum_{j=1}^{L} \omega_i^j \phi_j(x(t)), \quad (29)$$

where the activation function $\phi_j(x(t)) : \Omega \to \mathbb{R}$ is continuous, $\phi_j(x(t))|_{x=0} = 0$, the neural network weights of the ith step are ω_i^j, and L is the number of hidden layer neurons. The vector $\phi(x(t)) \equiv [\phi_1(x(t))\ \phi_2(x(t))\ \cdots\ \phi_L(x(t))]^T$ is the vector activation function and $W_i^L \equiv [\omega_i^1(x(t))\ \omega_i^2(x(t))\ \cdots\ \omega_i^L(x(t))]^T$ is the weight vector at the ith step.

The critic weights are tuned at each step to minimize the residual error between $\hat{V}_i(x(t))$ and the target function defined in (30) in a least squares sense over a set of points within a compact set Ω

$$E(x(t), x(t+h), W_i^L, \hat{u}_i(x(t))) = \int_t^{t+h} x(\tau)^T Q x(\tau) + \hat{u}_i(x(\tau))^T R \hat{u}_i(x(\tau))\, d\tau + \hat{V}_i(x(t+h)) = \int_t^{t+h} x(\tau)^T Q x(\tau) + \hat{u}_i(x(\tau))^T R \hat{u}_i(x(\tau))\, d\tau + (W_i^L)^T \phi(x(t+h)). \quad (30)$$

The residual error becomes

$$e_L = (W_{i+1}^L)^T \phi(x(t)) - E(x(t), x(t+h), W_i^L, \hat{u}_i(x(t))). \quad (31)$$

To find the least squares solution, the method of weighted residuals is used. Then the weights W_i^L are determined by projecting the residual error onto $(\partial e_L(x(t))/\partial W_i^L)$ and setting the result to zero $\forall x \in \Omega$, i.e.,

$$\left\langle \frac{\partial e_L(x(t))}{\partial W_i^L}, e_L(x(t)) \right\rangle = 0, \quad (32)$$

where $\langle f, g\rangle = \int_\Omega f g^T dx$ is a Lebesgue integral. When expanded, (32) becomes

$$0 = \int_\Omega \phi(x(t))(\phi^T(x(t)) W_{i+1}^L - E^T(x(t), x(t+h), W_i^L, \hat{u}_i(x(t)))\, dx(t). \quad (33)$$

Selecting activation functions $\{\phi_j(x(t))\}^L$ are linearly independent on the compact set $\Omega \subseteq \mathbb{R}^n$, then $\int_\Omega \phi(x(t))\phi^T(x(t))\, dx(t)$ is of full rank and invertible, so the unique solution for W_{i+1}^L exists and is given as

$$W_{i+1}^L = \left(\int_\Omega \phi(x(t))\phi^T(x(t))\, dx(t)\right)^{-1} \times \int_\Omega \phi(x(t)) E^T(x(t), x(t+h), W_i^L, \hat{u}_i(x(t)))\, dx(t). \quad (34)$$

Updating the value function neural network until the neural network weights converges, then the policy is updated as

$$\hat{u}_{i+1}(x) = -\frac{1}{2} R^{-1} g(x(t))^T \left(\frac{\partial \phi(x(t))}{\partial x(t)}\right)^T W_{i+1}^L. \quad (35)$$

Note that (6) and (7) do not need the internal dynamics, and $g(x(t))$ is only needed to update the control using (35), so the method works for a system with partially unknown dynamics.

The optimal control of nonlinear continuous-time systems can be obtained by going through the following steps:

1. Initialize the weights W_i^L and set the computation precision ε.
2. Compute the iterative performance control law $u_0(x(t))$ by (4), then obtain the iterative performance value $V_1(x(t))$ by (5).
3. Let $i=i+1$. Update the action $\hat{u}_i(x(t))$ by (35), then compute the weights W_{i+1}^L using (34), and obtain the value function $\hat{V}_{i+1}(x(t))$ by (29).
4. If $|\hat{V}_{i+1}(x(t))-\hat{V}_i(x(t))|<\varepsilon$ holds, go to step 5. Else, go to step 3.
5. Return the optimal value and control law.

The NN implementation of the value-iteration based HDP algorithm for discrete-time nonlinear systems can be found in [13].

5. Numerical simulation

In this section, two examples are given to demonstrate the effectiveness of the proposed methods. The basis functions are generated from a fourth-order polynomial as

$$\phi(x)=\{x_1^2, x_1x_2, x_2^2, x_1^4, x_1^3x_2, x_1^2x_2^2, x_1x_2^3, x_2^4\}, \tag{36}$$

which can be constructed from the expansion of the polynomial [5]

$$\sum_{j=1}^{N/2}\left(\sum_{i=1}^{n} x_i(t)\right)^{2j}, \tag{37}$$

where N is the order of approximation and n is the dimension of the system. The polynomial approximation is also used in the standard Weierstrass high-order approximation theorem.

Example 1 (*Continuous-time case*). Consider the following nonlinear continuous-time system:

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix}=\begin{bmatrix}-0.5\ x_1(t)+x_2(t)+2\ x_2^3(t)\\ -2(x_1(t)+x_2(t))\end{bmatrix}+\begin{bmatrix}0\\ \sin\ x_1(t)\end{bmatrix}u(x(t)). \tag{38}$$

The initial state $x_0=[0.5\ -0.5]^T$ and the performance index is given as

$$J(x(0))=\int_0^{\infty} x(\tau)^TQx(\tau)+u(x(\tau))^TRu(x(\tau))\,d\tau, \tag{39}$$

where $Q=I_2$ and $R=1$.

The sampling period $h=0.1$ s and the NN training error is set to 10^{-6}, then the weights of the value function converge to

$$W^*=[0.7333, 0.1333, 0.2833, -0.0335, 0.0448, -0.0660, 0.0744, 0.2934].$$

The evolution process of the weights is shown in Fig. 1.

The states trajectories are shown in Fig. 2. Apparently, the states converge to the origin, which indicates that the system is stable under the resulting control shown in Fig. 3.

In practical applications, a stabilized initial control for a nonlinear system is usually difficult to obtain. Compared with the general policy-iteration algorithm with $k=1$ (which is in fact a variant of the value-iteration algorithm) in [10], the stabilized initial control is not required using our method.

Example 2 (*Discrete-time case*). Consider the discrete-time nonlinear system $x_{k+1}=f(x_k)+g(x_k)u(x_k)$, where

$$f(x_k)=\begin{bmatrix}\sin\ x_{1k}\\ x_{1k}x_{2k}\end{bmatrix},\quad g(x_k)=\begin{bmatrix}0.4\\ -0.2\end{bmatrix}.$$

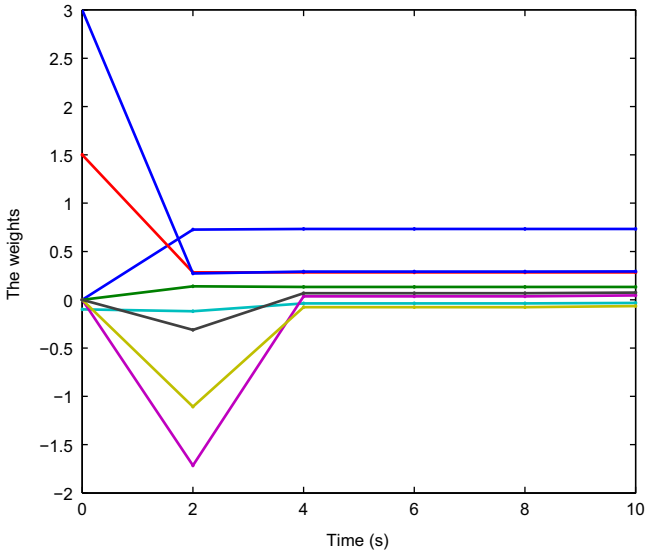


Fig. 1. The evolution process of the weights.

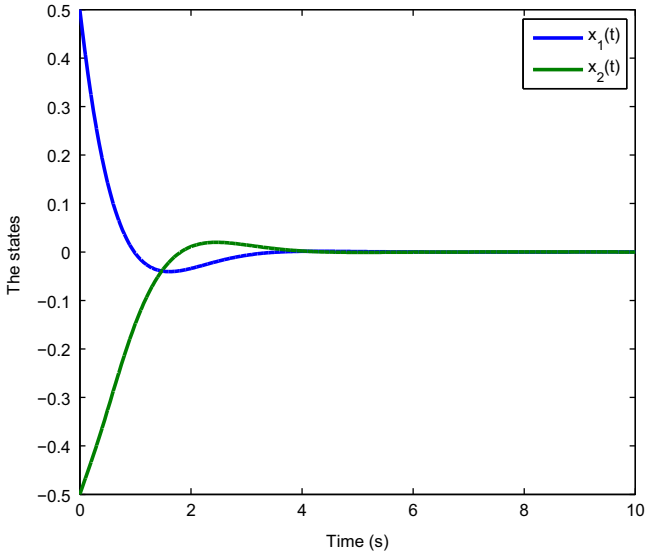


Fig. 2. The evolution process of the states.

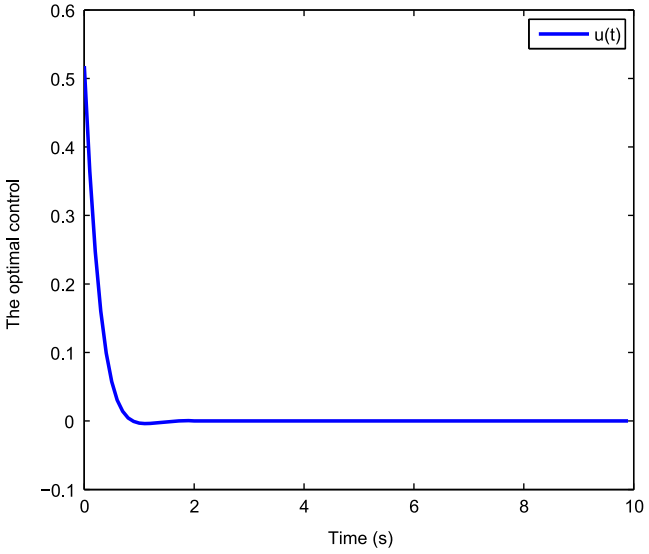


Fig. 3. The evolution process of the optimal control.

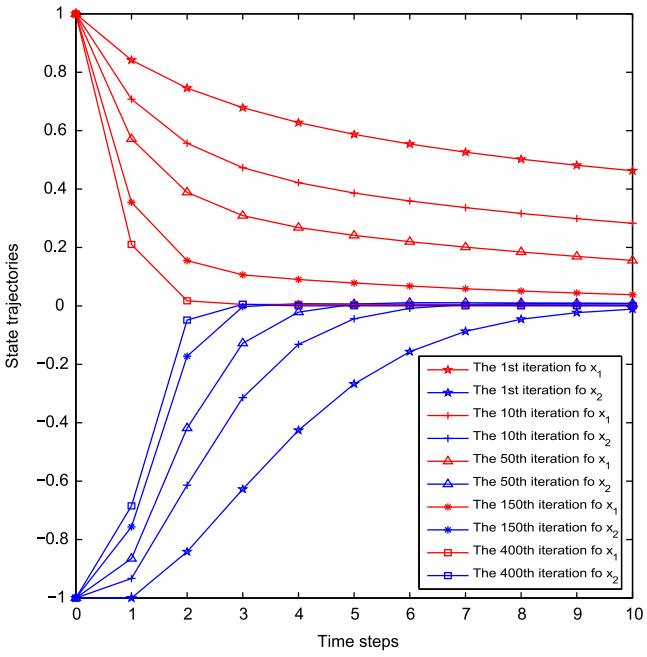


Fig. 4. The state trajectories.

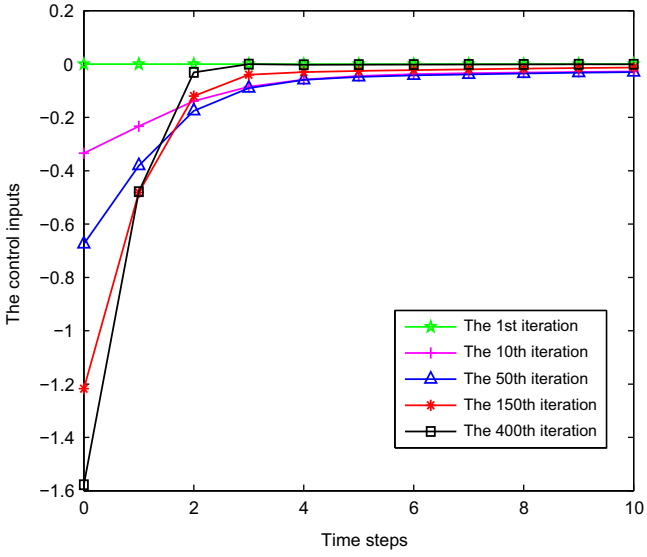


Fig. 5. The control inputs.

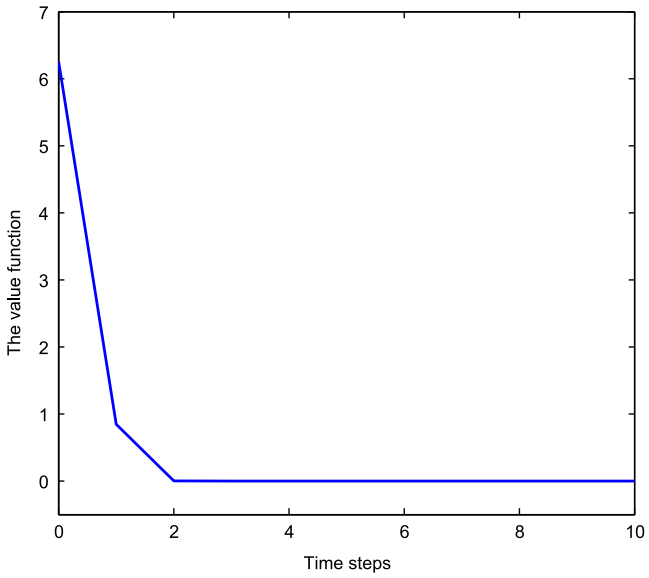


Fig. 6. The value function.

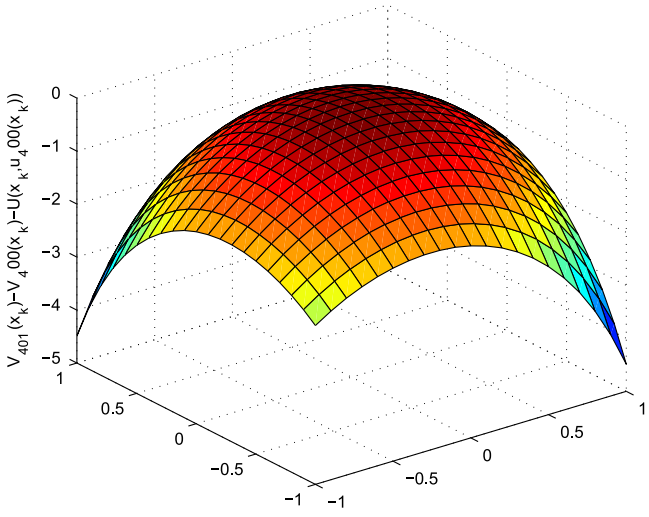


Fig. 7. The error between $V_{401}(x_k)-V_{400}(x_k)$ and $U(x_k,u_{400}(x_k))$.

The initial state is $x_0=[1\ -1]^T$ and the performance index is chosen as (19) and the weight matrices are chosen as $Q=I_2, R=1$. The NN training error is selected as 10^{-6}.

The algorithm runs 401 iteration steps to make sure that the given compute precision 10^{-6} has been achieved. The convergence process of state trajectories and the control inputs is shown in Figs. 4 and 5, respectively. The corresponding value function is given in Fig. 6. We can see that although the compute precision is not achieved, the state trajectories still converge to the origin, which indicates the corresponding optimal control are stabilized. In Fig. 7, (27) is satisfied by showing the 3-D plot of error between $V_{401}(x_k)-V_{400}(x_k)$ and $U(x_k,u_{400}(x_k))$, the error is less than zero globally in the square area $[-1;1]\times[-1;1]$, which implies the iterative control law $u_{400}(x_k)$ is stabilized. This further confirms our developed theory.

In the work [26,28], to obtain stabilized iterative control laws, the initial value function used for iterations can only be obtained by recurrent algorithms. By contrast, our method can obtain stabilized iterative control laws using any given initial value functions (for simplicity, let $V_0(x)=0$). Therefore, our method is more simple.

6. Conclusion

In this paper, the value-iteration based HDP algorithm has been employed to solve the optimal control for the continuous time nonlinear systems. Stability issues of the value-iteration based heuristic dynamic programming (HDP) algorithm for nonlinear systems have been investigated. Novel stability results for the HDP algorithm has been presented, which indicates that the resulting iterative control laws after finite iterations can guarantee the closed-loop stability of the nonlinear systems, rather than only the limit function of the iterative control sequence. Therefore, the practicality of the HDP algorithm has been greatly improved. Single neural network (NN) structure has been employed to implement the proposed HDP algorithm without knowing the internal dynamics of the systems. Two numerical examples have been given to demonstrate the effectiveness of the developed methods.

The value-iteration based HDP algorithm considered in this paper is a partial model-free method. It is known that we can transform the original nonlinear system into a new augment system by using a compensator, which allows us to implement the value-iteration based HDP algorithm without knowing any knowledge of the original nonlinear system. This is left to the future study.

Acknowledgments

The research was supported by the National Natural Science Foundation of China (61034005, 61273027, and 61203046) and National High Technology Research and Development Program of China (2012AA040104).

Appendix

Proof of Theorem 1. The assumption (8) implies that

$$\frac{\delta-1}{1+\theta}\left[\theta\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau - J^*(x(t+h))\right]\le 0. \quad (40)$$

Next, we will demonstrate the left hand side of the inequality (9) by mathematical induction. For $i=1$, we obtain

$$\begin{aligned}
V_1(x(t)) &= \min_u\left[V_0(x(t+h))+\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau\right]\\
&\ge \min_u\left[\delta J^*(x(t+h))+\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau\right]\\
&\ge \min_u\left[\left(\delta-\frac{\delta-1}{\theta+1}\right)J^*(x(t+h))\right.\\
&\quad\left.+\left(1+\theta\frac{\delta-1}{\theta+1}\right)\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau\right]\\
&= \frac{1+\delta\theta}{\theta+1}\min_u\left[J^*(x(t+h))+\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau\right]\\
&= \left[1+\frac{\delta-1}{1+\theta^{-1}}\right]J^*(x(t)).
\end{aligned}$$

Assume that for $i-1$, it holds that

$$\left[1+\frac{\delta-1}{(1+\theta^{-1})^{i-1}}\right]J^*\le V_{i-1},$$

then, we have

$$\begin{aligned}
V_i(x(t)) &= \min_u\left[V_{i-1}(x(t+h))+\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau\right]\\
&\ge \min_u\left\{\left[1+\frac{\delta-1}{(1+\theta^{-1})^{i-1}}\right]J^*(x(t+h))\right.\\
&\quad\left.+\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau\right\}\\
&\ge \min_u\left\{\left[1+\frac{(\delta-1)\theta^i}{(\theta+1)^i}\right]\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau\right.\\
&\quad\left.+\left[1+\frac{\delta-1}{(\theta^{-1}+1)^{i-1}}-\frac{(\delta-1)\theta^{i-1}}{(\theta+1)^i}\right]J^*(x(t+h))\right\}\\
&= \left[1+\frac{(\delta-1)\theta^i}{(\theta+1)^i}\right]\min_u\left[J^*(x(t+h))\right.\\
&\quad\left.+\int_t^{t+h} x(\tau)^T Qx(\tau)+u(x(\tau))^T Ru(x(\tau))\,d\tau\right]\\
&= \left[1+\frac{\delta-1}{(1+\theta^{-1})^i}\right]J^*(x(t)).
\end{aligned}$$

The left hand side of (9) is proved and the right hand side can be shown by the same way.

Let the iteration index i go to infinity, then we obtain

$$\lim_{i\to\infty}\left[1+\frac{\delta-1}{(1+\theta^{-1})^i}\right]J^*(x(t))=J^*(x(t))$$

and

$$\lim_{i\to\infty}\left[1+\frac{\omega-1}{(1+\theta^{-1})^i}\right]J^*(x(t))=J^*(x(t)).$$

Therefore, $V_\infty(x(t))=J^*(x(t))$, $u_\infty(x(t))=u^*(x(t))$. □

References

[1] R.E. Bellman, Dynamic Programming, Princeton University Press, Princeton, NJ, 1957.
[2] P.J. Werbos, Approximate dynamic programming for real-time control and neural modeling, in: D.A. White, D.A. Sofge (Eds.), Handbook of Intelligent Control, Van Nostrand Reinhold, New York, 1992.
[3] H. Zhang, D. Liu, Y. Luo, D. Wang, Adaptive Dynamic Programming for Control-Algorithms and Stability, Springer-Verlag, London, 2013.
[4] F. Wang, H. Zhang, D. Liu, Adaptive dynamic programming: an introduction, Comput. Intell. Mag. 4 (2) (2009) 39–47.
[5] R. Beard, G. Saridis, J. Wen, Galerkin approximations of the generalized Hamilton–Jacobi–Bellman equation, Automatica 33 (12) (1997) 2158–2177.
[6] K. Doya, Reinforcement learning in continuous time and space, Neural Comput. 12 (2000) 219–245.
[7] R. Howard, Dynamic Programming and Markov Processes, MIT Press, Cambridge, MA, 1960.
[8] D. Kleinman, On a iterative technique for Riccati equation computations, IEEE Trans. Autom. Control 3 (1) (1968) 114–115.
[9] M. Abu-Khalaf, F.L. Lewis, Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach, Automatica 41 (5) (2005) 779–791.
[10] D. Vrabie, F.L. Lewis, Generalized policy iteration for continuous time systems, in: Proceedings of International Joint Conference on Neural Networks 2009, pp. 3224–3231.
[11] D. Liu, Q. Wei, Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems, IEEE Trans. Neural Netw. Learn. Syst. 25 (3) (2014) 621–634.
[12] H. Zhang, Y. Luo, D. Liu, Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints, IEEE Trans. Neural Netw. 20 (9) (2009) 1490–1503.
[13] A. Al-Tamimi, F.L. Lewis, M. Abu-Khalaf, Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof, IEEE Trans. Syst. Man Cybern. Part B Cybern. 38 (4) (2008) 943–949.
[14] D. Wang, D. Liu, Q. Wei, D. Zhao, N. Jin, Optimal control of unknown non affine nonlinear discrete-time systems based on adaptive dynamic programming, Automatica 48 (8) (2012) 1825–1832.
[15] H. Zhang, Q. Wei, D. Liu, An iterative adaptive dynamic programming method for solving a class of nonlinear zero-sum differential games, Automatica 47 (1) (2011) 207–214.
[16] B. Wang, D. Zhao, C. Alippi, D. Liu, Dual heuristic dynamic programming for nonlinear discrete-time uncertain systems with state delay, Neurocomputing 134 (2014) 222–229.
[17] Y. Huang, D. Liu, Neural-network-based optimal tracking control scheme for a class of unknown discrete-time nonlinear systems using iterative ADP algorithm, Neurocomputing 125 (2014) 46–56.
[18] D. Wang, D. Liu, Neuro-optimal control for a class of unknown nonlinear dynamic systems using SN-DHP technique, Neurocomputing 121 (2013) 218–225.
[19] Y. Tang, H. He, Z. Ni, J. Wen, X. Sui, Reactive power control of grid-connected wind farm based on adaptive dynamic programming, Neurocomputing 125 (2014) 125–133.
[20] D. Liu, H. Li, D. Wang, Neural-network-based zero-sum game for discrete-time nonlinear systems via iterative adaptive dynamic programming algorithm, Neurocomputing 11 (2013) 92–100.
[21] J. Zhang, H. Zhang, Y. Luo, T. Feng, Model-free optimal control design for a class of linear discrete-time systems with multiple delays using adaptive dynamic programming, Neurocomputing 135 (2014) 163–170.
[22] F.L. Lewis, K.G. Vamvoudakis, Reinforcement learning for partially observable dynamic processes: adaptive dynamic programming using measured output data, IEEE Trans. Syst. Man Cybern. Part B 41 (1) (2011) 14–24.
[23] Q. Wei, D. Liu, Numerical adaptive learning control scheme for discrete-time nonlinear systems, IET Control Theory Appl. 7 (11) (2013) 1472–1486.

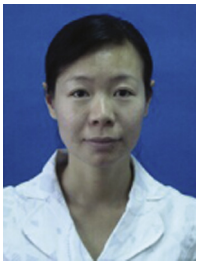
[24] D. Zhao, Z. Hu, Z. Xia, C. Alippi, Y. Zhu, D. Wang, Full-range adaptive cruise control based on supervised adaptive dynamic programming, Neurocomputing 125 (2014) 57–67.
[25] D. Liu, Q. Wei, Finite-approximation-error-based optimal control approach for discrete-time nonlinear systems, IEEE Trans. Cybern. 43 (2) (2013) 779–789.
[26] H. Li, D. Liu, Optimal control for discrete-time affine non-linear systems using general value iteration, Control Theory Appl. IET 6 (18) (2012) 2725–2736.
[27] B. Lincoln, A. Rantzer, Relaxing dynamic programming, IEEE Trans. Autom. control 51 (8) (2006) 1249–1260.
[28] Q. Wei, D. Liu, Adaptive dynamic programming with stable value iteration algorithm for discrete-time nonlinear systems, in: WCCI 2012 IEEE World Congress on Computational Intelligence 2012, pp. 10–15.
[29] D.V. Prokhorov, D.C. Wunsch, Adaptive critic designs, IEEE Trans. Neural Netw. 8 (5) (1997) 997–1007.
[30] T. Hanselmann, L. Noakes, A. Zaknich, Continuous-time adaptive critics, IEEE Trans. Neural Netw. 18 (3) (2007) 631–647.

Tao Feng received the B.S. degree in Mathematics and Applied Mathematics from China University of Petroleum (East China), Dongying, China, in 2008 and the M.S. degree in Fundamental Mathematics from Northeastern University, China, in 2011. Now, he is a Ph.D. candidate in the College of Information Science and Engineering, Northeastern University. His main research interests include approximate dynamic programming, inverse optimal control and multi-agent systems.

Huaguang Zhang (SM'04) received the B.S. and M.S. degrees in control engineering from Northeastern Electric Power University, Jilin, China, 1982 and 1985, respectively, and he received the Ph.D. degree in thermal power engineering and automation from Southeast University, Nanjing, China, in 1991.

He joined the Department of Automatic Control, Northeastern University, Shenyang, China, in 1992, as a Postdoctoral Fellow. Since 1994, he has been a Professor and the Head of the Electric Automation Institute, Northeastern University. He has authored three English monographs, and holds 30 patents. His main research interests are neural network-based control, fuzzy control, chaos control, nonlinear control, signal processing, adaptive dynamic programming (ADP) and their industrial applications.

Zhang was a recipient of the Nationwide Excellent Post-Doctor, the Outstanding Youth Science Foundation Award from the National Natural Science Foundation Committee of China in 2003, the Cheung Kong Scholar Award from the Education Ministry of China in 2005, and the IEEE Transactions on Neural Networks Outstanding Paper Award in 2012. He was an Associate Editor of Automatica, IEEE Transactions on Cybernetics and Neurocomputing. He is the Deputy Director of the Intelligent System Engineering Committee of Chinese Association of Artificial Intelligence.



Yanhong Luo received the B.S. degree in automation control and the M.S. and Ph.D. degrees in control theory and control engineering from Northeastern University, Shenyang, China, in 2003, 2006, and 2009, respectively. She is currently an Associate Professor with Northeastern University. Her current research interests include fuzzy controls, neural networks adaptive controls, approximate dynamic programming, and their industrial applications.

Jilie Zhang received the B.S. degree in automation from Liaoning University of Technology, Jinzhou, China, in 2007 and the M.S. degree in control theory and control Engineering from Kunming University of Science and Technology, Kunming, China, in 2010. He received the Ph.D. degree in the College of Information Science and Engineering, Northeastern University, PR China, in 2014. At present, he works in Southwest Jiaotong University. His main research interests include fault diagnosis, approximate dynamic programming, reinforcement learning, game theory and multi-agent system.