A New Discrete-Time Iterative Adaptive Dynamic Programming Algorithm Based on Q-Learning*

Qinglai Wei[1] and Derong Liu[2]

[1] The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
qinglai.wei@ia.ac.cn

[2] School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China
derong@ustb.edu.cn

Abstract. In this paper, a novel Q-learning based policy iteration adaptive dynamic programming (ADP) algorithm is developed to solve the optimal control problems for discrete-time nonlinear systems. The idea is to use a policy iteration ADP technique to construct the iterative control law which stabilizes the system and simultaneously minimizes the iterative Q function. Convergence property is analyzed to show that the iterative Q function is monotonically non-increasing and converges to the solution of the optimality equation. Finally, simulation results are presented to show the performance of the developed algorithm.

Keywords: Adaptive critic designs, adaptive dynamic programming, approximate dynamic programming, Q-learning, policy iteration, neural networks, nonlinear systems, optimal control.

1 Introduction

Characterized by strong abilities of self-learning and adaptivity, adaptive dynamic programming (ADP), proposed by Werbos [25, 26], has demonstrated powerful capability to find the optimal control policy by solving the Hamilton-Jacobi-Bellman (HJB) equation forward-in-time and becomes an important brain-like intelligent optimal control method for nonlinear systems [4, 6–9, 12, 17, 23]. Policy and value iterations are basic iterative algorithms in ADP. Value iteration algorithm was proposed in [3]. In [2], the convergence of value iteration was proven. Policy iteration algorithms for optimal control of continuous-time (CT) systems were given in [1]. In [5], policy iteration algorithm for discrete-time nonlinear systems was developed. For many traditional iterative ADP algorithms, they require to build the model of nonlinear systems and then perform the ADP algorithms to derive an improved control policy [11, 16, 18–22, 24, 27, 28]. In contrast, Q-learning, proposed by Watkins [14, 15], is a typical data-based ADP algorithm. In [10], Q-learning was named action-dependent heuristic dynamic programming (ADHDP). For Q-learning algorithms, Q functions are used instead of value functions in

* This work was supported in part by the National Natural Science Foundation of China under Grants 61273140, 61304086, 61374105, and 61233001, and in part by Beijing Natural Science Foundation under Grant 4132078.

X. Hu et al. (Eds.): ISNN 2015, LNCS 9377, pp. 43–52, 2015.
DOI: 10.1007/978-3-319-25393-0_6

the traditional iterative ADP algorithms. Q functions depend on both system state and control, which means that they already include the information about the system and the utility function. Hence, it is easier to compute control policies from Q functions than the traditional performance index functions. Because of this merit, Q-learning algorithms are preferred to unknown and model-free systems to obtain the optimal control.

In this paper, inspired by [5], a novel Q-learning based policy iteration ADP algorithm is developed for discrete-time nonlinear systems. First, the procedure of the Q-learning based policy iteration ADP algorithm is described. Next, property analysis of the Q-learning based policy iteration ADP algorithm is established. It is proven that the iterative Q functions will monotonically non-increasing and converges to the optimal solution of the HJB equation. Finally, simulation results will illustrate the effectiveness of the developed algorithm.

The rest of this paper is organized as follows. In Section 2, the problem formulation is presented. In Section 3, the properties of the developed Q-learning based policy iteration ADP algorithm will be proven in this section. In Section 4, numerical results are presented to demonstrate the effectiveness of the developed algorithm. Finally, in Section 5, the conclusion is drawn.

2 Problem Formulation

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

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

where $x_k \in \mathbb{R}^n$ is the state vector and $u_k \in \mathbb{R}^m$ is the control vector. Let x_0 be the initial state and $F(x_k, u_k)$ be the system function. Let $\underline{u}_k = \{u_k, u_{k+1}, \ldots\}$ be an arbitrary sequence of controls from k to ∞. The performance index function for state x_0 under the control sequence $\underline{u}_0 = \{u_0, u_1, \ldots\}$ is defined as

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

where $U(x_k, u_k) > 0$, for $x_k, u_k \neq 0$, is the utility function. The goal of this paper is to find an optimal control scheme which stabilizes the system (1) and simultaneously minimizes the performance index function (2). For convenience of analysis, results of this paper are based on the following assumptions.

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

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

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

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

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

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

According to [14] and [15], the optimal Q function satisfies the Q-Bellman equation

$$Q^*(x_k, u_k) = U(x_k, u_k) + \min_{u_{k+1}} Q^*(x_{k+1}, u_{k+1}). \tag{4}$$

The optimal performance index function satisfies

$$J^*(x_k) = \min_{u_k} Q^*(x_k, u_k). \tag{5}$$

The optimal control law $u^*(x_k)$ can be expressed as

$$u^*(x_k) = \arg\min_{u_k} Q^*(x_k, u_k). \tag{6}$$

From (5), we know that if we obtain the optimal Q function $Q^*(x_k, u_k)$, then the optimal control law $u^*(x_k)$ and the optimal performance index function $J^*(x_k)$ can be obtained. However, the optimal Q function $Q^*(x_k, u_k)$ is generally an unknown and non-analytic function, which cannot be obtained directly by (4). Hence, a discrete-time Q learning algorithm is developed in [15] to solve for the Q function iteratively.

3 Discrete-Time Policy Iteration ADP Algorithm Based on Q-Learning

In this section, the Q-learning based policy iteration ADP algorithm will be developed to obtain the optimal controller for discrete-time nonlinear systems. Convergence and optimality proofs will also be given to show that the iterative Q function will converge to the optimum.

3.1 Derivation of the Discrete-Time Policy Iteration ADP Algorithm Based on Q-Learning

In the developed policy iteration algorithm, the Q function and control law are updated by iterations, with the iteration index i increasing from 0 to infinity. Let $v_0(x_k)$ be an arbitrary admissible control law [5]. For $i = 0$, let $Q_0(x_k, u_k)$ be the initial iterative Q function constructed by $v_0(x_k)$, i.e.,

$$Q_0(x_k, v_0(x_k)) = \sum_{j=0}^{\infty} U(x_{k+j}, v_0(x_{k+j})). \tag{7}$$

Thus, initial iterative Q function satisfies the following generalized Q-Bellman equation

$$Q_0(x_k, u_k) = U(x_k, u_k) + Q_0(x_{k+1}, v_0(x_{k+1})). \tag{8}$$

Then, the iterative control law is computed by

$$v_1(x_k) = \arg\min_{u_k} Q_0(x_k, u_k). \tag{9}$$

For $i = 1, 2, \ldots$, let $Q_i(x_k, u_k)$ be the iterative Q function constructed by $v_i(x_k)$, which satisfies the following generalized Q-Bellman equation

$$Q_i(x_k, u_k) = U(x_k, u_k) + Q_i(x_{k+1}, v_i(x_{k+1})), \tag{10}$$

and the iterative control law is updated by

$$v_{i+1}(x_k) = \arg\min_{u_k} Q_i(x_k, u_k). \tag{11}$$

3.2 Properties of the Policy Iteration Based Deterministic *Q*-Learning Algorithm

For the policy iteration algorithm of discrete-time nonlinear systems [5], it shows that the iterative value function is monotonically non-increasing and converges to the optimum. In this subsection, inspired by [5], we will show that the iterative Q function will also be monotonically non-increasing and converges to its optimum.

Theorem 1. *For $i = 0, 1, \ldots$, let $Q_i(x_k, u_k)$ and $v_i(x_k)$ be obtained by (8)–(11). If Assumptions 1–4 hold, then the iterative Q function $Q_i(x_k, u_k)$ is monotonically non-increasing and converges to the optimal Q function $Q^*(x_k, u_k)$, as $i \to \infty$, i.e.,*

$$\lim_{i\to\infty} Q_i(x_k, u_k) = Q^*(x_k, u_k), \tag{12}$$

which satisfies the optimal Q-Bellman equation (4).

Proof. The statement can be proven in two steps.
1) Show that the iterative Q function $Q_i(x_k, u_k)$ is monotonically non-increasing as i increases, i.e.,

$$Q_{i+1}(x_k, u_k) \le Q_i(x_k, u_k). \tag{13}$$

According to (11), we have

$$Q_i(x_k, v_{i+1}(x_k)) = \min_{u_k} Q_i(x_k, u_k) \le Q_i(x_k, v_i(x_k)). \tag{14}$$

For $i = 0, 1, \ldots$, define a new iterative Q function $\mathcal{Q}_{i+1}(x_k, u_k)$ as

$$\mathcal{Q}_{i+1}(x_k, u_k) = U(x_k, u_k) + Q_i(x_{k+1}, v_{i+1}(x_{k+1})), \tag{15}$$

where $v_{i+1}(x_{k+1})$ is obtained by (11). According to (14), we can obtain

$$\begin{aligned}\mathcal{Q}_{i+1}(x_k, u_k) &= U(x_k, u_k) + Q_i(x_{k+1}, v_{i+1}(x_{k+1}))\\ &= U(x_k, u_k) + \min_{u_{k+1}} Q_i(x_{k+1}, u_{k+1})\\ &\le U(x_k, u_k) + Q_i(x_{k+1}, v_i(x_{k+1}))\\ &= Q_i(x_k, u_k).\end{aligned} \tag{16}$$

Now we prove inequality (13) by mathematical induction. For $i = 0, 1, \ldots$, as

$$\begin{aligned} Q_i(x_{k+1}, v_i(x_{k+1})) - Q_i(x_k, v_i(x_k)) \\ &= -U(x_k, v_i(x_k)) \\ &< 0, \end{aligned} \tag{17}$$

we have $v_i(x_k)$ is a stable control. Thus, we have $x_{\mathcal{N}} = 0$ for $\mathcal{N} \to \infty$. According to Assumptions 1–4, we have $v_{i+1}(x_{\mathcal{N}}) = v_i(x_{\mathcal{N}}) = 0$, which obtains

$$Q_{i+1}(x_{\mathcal{N}}, v_{i+1}(x_{\mathcal{N}})) = \mathcal{Q}_{i+1}(x_{\mathcal{N}}, v_{i+1}(x_{\mathcal{N}})) = Q_i(x_{\mathcal{N}}, v_i(x_{\mathcal{N}})) = 0, \tag{18}$$

and

$$Q_{i+1}(x_{\mathcal{N}-1}, u_{\mathcal{N}-1}) = \mathcal{Q}_{i+1}(x_{\mathcal{N}-1}, u_{\mathcal{N}-1}) = Q_i(x_{\mathcal{N}-1}, u_{\mathcal{N}-1}) = U(x_{\mathcal{N}-1}, u_{\mathcal{N}-1}). \tag{19}$$

Let $k = \mathcal{N} - 2$. According to (11),

$$\begin{aligned} Q_{i+1}(x_{\mathcal{N}-2}, u_{\mathcal{N}-2}) &= U(x_{\mathcal{N}-2}, u_{\mathcal{N}-2}) + Q_{i+1}(x_{\mathcal{N}-1}, v_{i+1}(x_{\mathcal{N}-1})) \\ &= U(x_{\mathcal{N}-2}, u_{\mathcal{N}-2}) + Q_i(x_{\mathcal{N}-1}, v_{i+1}(x_{\mathcal{N}-1})) \\ &= \mathcal{Q}_{i+1}(x_{\mathcal{N}-2}, u_{\mathcal{N}-2}) \\ &\le Q_i(x_{\mathcal{N}-2}, u_{\mathcal{N}-2}). \end{aligned} \tag{20}$$

So, the conclusion holds for $k = \mathcal{N} - 2$. Assume that the conclusion holds for $k = \ell + 1$, $\ell = 0, 1, \ldots$. For $k = \ell$ we can get

$$\begin{aligned} Q_{i+1}(x_\ell, u_\ell) &= U(x_\ell, u_\ell) + Q_{i+1}(x_{\ell+1}, v_{i+1}(x_{\ell+1})) \\ &\le U(x_\ell, u_\ell) + Q_i(x_{\ell+1}, v_{i+1}(x_{\ell+1})) \\ &= \mathcal{Q}_{i+1}(x_\ell, u_\ell) \\ &\le Q_i(x_\ell, u_\ell). \end{aligned} \tag{21}$$

Hence, we can obtain that for $i = 0, 1, \ldots$, the inequality (13) holds, for x_k, u_k. The proof of mathematical induction is completed.

As $Q_i(x_k, u_k)$ is a non-increasing and lower bounded sequence, i.e., $Q_i(x_k, u_k) \ge 0$, the limit of the iterative Q function $Q_i(x_k, u_k)$ exists as $i \to \infty$, i.e.,

$$Q_\infty(x_k, u_k) = \lim_{i \to \infty} Q_i(x_k, u_k). \tag{22}$$

2) Show that the limit of the iterative Q function $Q_i(x_k, u_k)$ satisfies the optimal Q-Bellman equation, as $i \to \infty$.

According to (21), we can obtain

$$\begin{aligned} Q_\infty(x_k, u_k) &= \lim_{i \to \infty} Q_{i+1}(x_k, u_k) \le Q_{i+1}(x_k, u_k) \le \mathcal{Q}_{i+1}(x_k) \\ &= U(x_k, u_k) + Q_i(x_{k+1}, v_{i+1}(x_{k+1})) \\ &= U(x_k, u_k) + \min_{u_k} Q_i(x_{k+1}, u_{k+1}). \end{aligned} \tag{23}$$

Letting $i \to \infty$, we obtain

$$Q_\infty(x_k, u_k) \le U(x_k, u_k) + \min_{u_{k+1}} Q_\infty(x_{k+1}, u_{k+1}). \tag{24}$$

Let $\zeta > 0$ be an arbitrary positive number. There exists a positive integer p such that

$$Q_p(x_k, u_k) - \zeta \le Q_\infty(x_k, u_k) \le Q_p(x_k, u_k). \tag{25}$$

Hence, we can get

$$\begin{aligned} Q_\infty(x_k, u_k) &\ge Q_p(x_k, u_k) - \zeta \\ &= U(x_k, u_k) + Q_p(x_{k+1}, v_p(x_{k+1})) - \zeta \\ &\ge U(x_k, u_k) + Q_\infty(x_{k+1}, v_p(x_{k+1})) - \zeta \\ &\ge U(x_k, u_k) + \min_{u_{k+1}} Q_\infty(x_{k+1}, u_{k+1}) - \zeta. \end{aligned} \tag{26}$$

Since ζ is arbitrary, we have

$$Q_\infty(x_k, u_k) \ge U(x_k, u_k) + \min_{u_{k+1}} Q_\infty(x_{k+1}, u_{k+1}). \tag{27}$$

Combining (24) and (27), we obtain

$$Q_\infty(x_k, u_k) = U(x_k, u_k) + \min_{u_{k+1}} Q_\infty(x_{k+1}, u_{k+1}). \tag{28}$$

According to the definition of the optimal Q function in (4), we have $Q_\infty(x_k, u_k) = Q^*(x_k, u_k)$. The proof is completed.

4 Simulation Study

We now examine the performance of the developed policy iteration algorithm in a nonlinear torsional pendulum system [13]. The dynamics of the pendulum is as follows

$$\begin{bmatrix} x_{1(k+1)} \\ x_{2(k+1)} \end{bmatrix} = \begin{bmatrix} 0.1x_{2k} + x_{1k} \\ -0.49\sin(x_{1k}) - 0.1f_d x_{2k} + x_{2k} \end{bmatrix} + \begin{bmatrix} 0 \\ 0.1 \end{bmatrix} u_k, \tag{29}$$

where $f_d = 0.2$ is the rotary inertia and frictional factor. Let the initial state be $x_0 = [1, -1]^T$. The utility function is expressed as $U(x_k, u_k) = x_k^T Q x_k + u_k^T R u_k$, where $Q = I$, $R = I$ and I denotes the identity matrix with suitable dimensions. Choose the critic and action networks as back propagation (BP) networks with the structures of 3–12–1 and 2–12–1, respectively. We randomly choose $p = 20000$ training data to implement the developed algorithm to obtain the optimal control law. For each iteration step, the critic network and the action network are trained for 1000 steps using the learning rate of $\alpha_c = \beta_a = 0.01$ so that the neural network training error becomes less than 10^{-5}. Implementing the developed Q-learning based policy iteration adaptive dynamic programming algorithm for $i = 25$ iterations to reach the computation precision $\varepsilon = 0.01$. The plots of the iterative function $Q_i(x_k, v_i(x_k))$ are shown in Fig. 1.

For nonlinear system (29), the iterative Q function is monotonically non-increasing and converges to its optimum by the Q-learning based policy iteration ADP algorithm. The corresponding iterative trajectories of system states and controls are shown in Figs. 2 and 3, respectively.

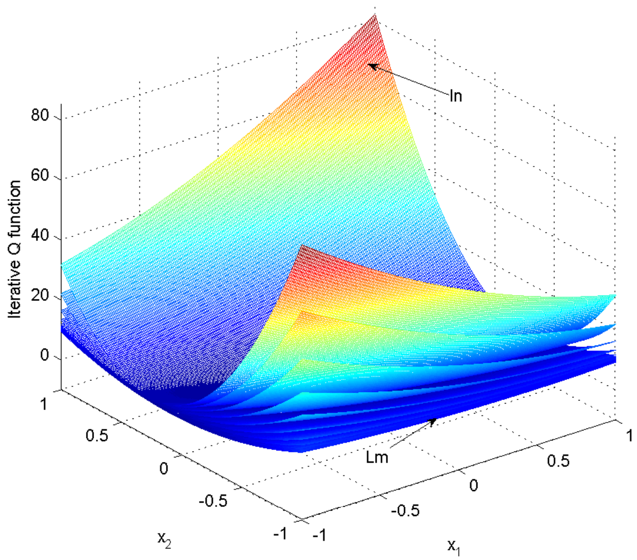

Fig. 1. The plots of the iterative Q function

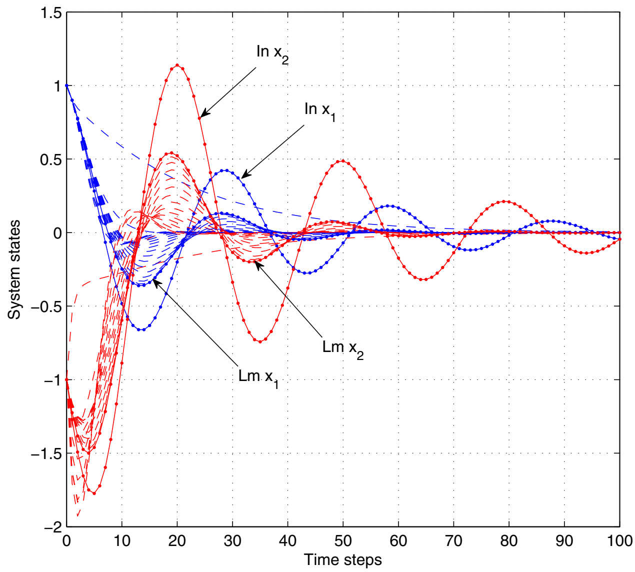

Fig. 2. The iterative state trajectories

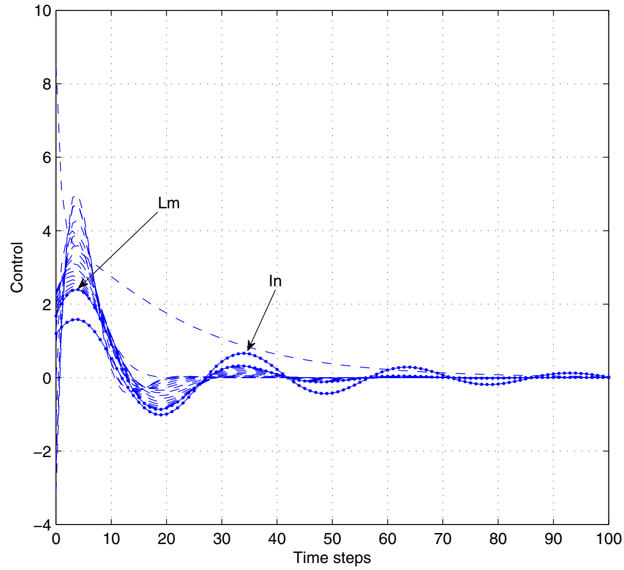

Fig. 3. The iterative control trajectories

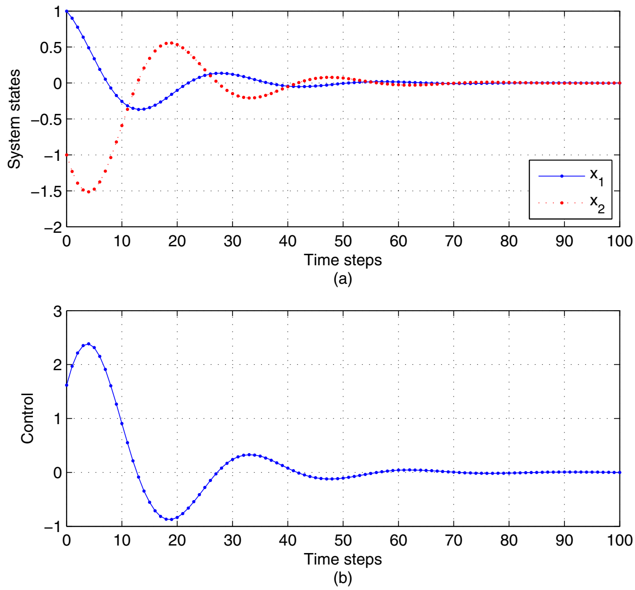

Fig. 4. The optimal state and control trajectories

From Figs. 2 and 3, we can see that the iterative system states and controls are both convergent to their optimal ones. The nonlinear system (29) can be stabilized under an arbitrary iterative control law $v_i(x_k)$, where the stability properties of the developed

Q-learning based policy iteration ADP algorithm can be verified. The optimal states and control trajectories are shown in Fig. 4.

5 Conclusions

In this paper, an effective policy iteration adaptive dynamic programming algorithm based on Q-learning is developed to solve optimal control problems for infinite horizon discrete-time nonlinear systems. The iterative Q functions is proven to be monotonically non-increasing and converges to the optimum as the iteration index increases to infinity. Finally, simulation results are presented to illustrate the performance of the developed algorithm.

References

1. Abu-Khalaf, M., Lewis, F.L.: Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach. Automatica 41(5), 779–791 (2005)
2. Al-Tamimi, A., Lewis, F.L., Abu-Khalaf, M.: Discrete-time nonlinear HJB solution using approximate dynamic programming: convergence proof. IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics 38(4), 943–949 (2008)
3. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming. Athena Scientific, Belmont (1996)
4. Jiang, Y., Jiang, Z.P.: Robust adaptive dynamic programming with an application to power systems. IEEE Transactions on Neural Networks and Learning Systems 24(7), 1150–1156 (2013)
5. Liu, D., Wei, Q.: Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. IEEE Transactions on Neural Networks and Learning Systems 25(3), 621–634 (2014)
6. Modares, H., Lewis, F.L.: Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning. IEEE Transactions on Automatic Control 59(11), 3051–3056 (2014)
7. Modares, H., Lewis, F.L., Naghibi-Sistani, M.B.: Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks. IEEE Transactions on Neural Networks and Learning systems 24(10), 1513–1525 (2013)
8. Modares, H., Lewis, F.L.: Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning. Automatica 50(7), 1780–1792 (2014)
9. Kiumarsi, B., Lewis, F.L., Modares, H., Karimpur, A., Naghibi-Sistani, M.B.: Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. Automatica 50(4), 1167–1175 (2014)
10. Prokhorov, D.V., Wunsch, D.C.: Adaptive critic designs. IEEE Transactions on Neural Networks 8(5), 997–1007 (1997)
11. Song, R., Xiao, W., Zhang, H., Sun, C.: Adaptive dynamic programming for a class of complex-valued nonlinear systems. IEEE Transactions on Neural Networks and Learning Systems 25(9), 1733–1739 (2014)
12. Song, R., Lewis, F.L., Wei, Q., Zhang, H., Jiang, Z.-P., Levine, D.: Multiple actor-critic structures for continuous-time optimal control using input-output data. IEEE Transactions on Neural Networks and Learning Systems 26(4), 851–865 (2015)

13. Si, J., Wang, Y.-T.: On-line learning control by association and reinforcement. IEEE Transactions on Neural Networks 12(2), 264–276 (2001)
14. Watkins, C.: Learning from Delayed Rewards. Ph.D. Thesis, Cambridge University, Cambridge (1989)
15. Watkins, C., Dayan, P.: Q-learning. Machine Learning 8(3-4), 279–292 (1992)
16. Wei, Q., Liu, D.: An iterative ϵ-optimal control scheme for a class of discrete-time nonlinear systems with unfixed initial state. Neural Networks 32, 236–244 (2012)
17. Wei, Q., Zhang, H., Dai, J.: Model-free multiobjective approximate dynamic programming for discrete-time nonlinear systems with general performance index functions. Neurocomputing 72(7–9), 1839–1848 (2009)
18. Wei, Q., Liu, D.: A novel iterative θ-adaptive dynamic programming for discrete-time nonlinear systems. IEEE Transactions on Automation Science and Engineering 11(4), 1176–1190 (2014)
19. Wei, Q., Liu, D.: Data-driven neuro-optimal temperature control of water gas shift reaction using stable iterative adaptive dynamic programming. IEEE Transactions on Industrial Electronics 61(11), 6399–6408 (2014)
20. Wei, Q., Liu, D.: Neural-network-based adaptive optimal tracking control scheme for discrete-time nonlinear systems with approximation errors. Neurocomputing 149(3), 106–115 (2015)
21. Wei, Q., Liu, D., Shi, G., Liu, Y.: Optimal multi-battery coordination control for home energy management systems via distributed iterative adaptive dynamic programming. IEEE Transactions on Industrial Electronics (2015) (article in press)
22. Wei, Q., Liu, D., Shi, G.: A novel dual iterative Q-learning method for optimal battery management in smart residential environments. IEEE Transactions on Industrial Electronics 62(4), 2509–2518 (2015)
23. Wei, Q., Liu, D.: Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification. IEEE Transactions on Automation Science and Engineering 11(4), 1020–1036 (2014)
24. Wei, Q., Liu, D., Yang, X.: Infinite horizon self-learning optimal control of nonaffine discrete-time nonlinear systems. IEEE Transactions on Neural Networks and Learning Systems 26(4), 866–879 (2015)
25. Werbos, P.J.: Advanced forecasting methods for global crisis warning and models of intelligence. General Systems Yearbook 22, 25–38 (1977)
26. Werbos, P.J.: A menu of designs for reinforcement learning over time. In: Miller, W.T., Sutton, R.S., Werbos, P.J. (eds.) Neural Networks for Control, pp. 67–95. MIT Press, Cambridge (1991)
27. Xu, X., Lian, C., Zuo, L., He, H.: Kernel-based approximate dynamic programming for real-time online learning control: An experimental study. IEEE Transactions on Control Systems Technology 22(1), 146–156 (2014)
28. Zhang, H., Wei, Q., Luo, Y.: A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm. IEEE Transactions on System, Man, and cybernetics–Part B: Cybernetics 38(4), 937–942 (2008)