Iterative Q-Learning-Based Nonlinear Optimal Tracking Control

Qinglai Wei, Ruizhuo Song, Yancai Xu and Derong Liu

***Abstract*—A new Q-learning algorithm is developed for a class of discrete-time nonlinear systems in this paper to solve the infinite horizon optimal tracking problems. Using system transformations, the optimal tracking problem is transformed to be an optimal regulation problem. Thereafter, for the regulation system, the new Q-learning algorithm is developed in order to obtain the optimal control law. Convergence of the iterative Q functions and the admissibility of the iterative control law are analyzed. In the end, two corresponding simulation examples are presented to illustrate the performance of the newly developed algorithm.**

I. Introduction

In the past several decades, the optimal control problems especially for nonlinear systems have always been the focus in the control field [6]. As is known to all, dynamic programming is a very useful tool while solving the optimal control problems. Nevertheless, considering the "curse of dimensionality" , when trying to obtain the optimal solution, it is very likely to be computationally untenable to perform dynamic programming. Correspondingly, The adaptive dynamic programming (ADP) algorithm was proposed by [1], [2] as a solution to optimal control problems in a forward-in-time way. Policy and value iterations are two primary iterative ADP algorithms [3]. In [4], policy iteration algorithms are firstly used for optimal control of continuous-time (CT) systems, which have continuous states and action spaces. In [5], the optimal control law for multiple actor-critic structures was effectively obtained using shunting inhibitory artificial neural network (SIANN). Policy iteration for zero-sum and non-zero-sum games was discussed in [7]–[9]. In [10], the multi-agent optimal control was obtained using fuzzy approximation structures. In [11], while solving problems of discrete-time (DT) nonlinear systems, policy iteration algorithm was developed. Thereafter, value iteration algorithm was presented for the optimal control problems of discrete-time nonlinear systems in [12]. For deterministic discrete-time affine nonlinear systems, [13] studied the value iteration algorithm. It was proven that the iterative value function is non-decreasing and bounded, and hence converges to the optimum as the iteration index increases to infinity. In [6], [14], [15], value iteration algorithms with approximation errors were analyzed. Based on the framework of policy and value iteration algorithms, more investigations on iterative ADP algorithms have been developed [16]–[32].

Q-learning, proposed by Watkins [33], [34], is a representative data-based ADP algorithms. In Q-learning algorithms, the Q function depends on both system state and control [35], which means that it already includes the information about the system and the utility function. A new policy iteration Q-learning algorithm is established in this paper for a large class of discrete-time-nonlinear systems. According to system transformation processes, the corresponding optimal tracking problem is effectively transformed into an optimal regulation one. The corresponding tracking error system is presented. According to the tracking error and the reference tracking control, the performance index function is displayed. Next, the policy iteration Q-learning algorithm for the transformed system is derived. The convergence and stability properties are analyzed. It is shown that the nonlinear system can be stabilized by any of the iterative control laws. The iterative Q function is nonincreasing in a monotonic way and converges to the optimal Q function, which is proven. Neural networks are employed to implement the policy iteration Q-learning algorithm by approximating the iterative Q function and iterative control law, respectively. At the end, simulation results will illustrate the good effectiveness of the developed algorithm.

II. Problem Statement

The following discrete-time nonlinear systems are consider in this paper.

$$x(k+1)=f(x(k))+gu(k), \tag{1}$$

where $x(k)\in\mathbb{R}^n$ is the state vector and $u(k)\in\mathbb{R}^m$ is the control vector. Let f and g denote system function. For infinite-time optimal tracking problem, the control objective is to design optimal feedback control $u(x(k))$ for system (1) such that the state $x(k)$ track the specified desired trajectory $x_r(k)\in\mathbb{R}^n, k=0,1,\ldots$. In this paper, we assume that the control gain matrix g satisfies $\operatorname{rank}\{g\}\geq n$ for the convenience of our analysis. Let $u_r(k)$ be the reference control. Then, the reference control $u_r(k)$ should satisfy

$$x_r(k+1)=f(x_r(k))+gu_r(k), \tag{2}$$

and the reference control talked above can be computed by the following equation

$$u_r(k)=g^{+}(x_r(k+1)-f(x_r(k))), \tag{3}$$

where g^{+} is the Moore-Penrose pseudo-inverse matrix of g. Define the tracking error as

$$y(k)=x(k)-x_r(k). \tag{4}$$

This work was supported in part by the National Natural Science Foundation of China under Grants 61374105, 61273140, 61503379, 61304079, 61533017, and U1501251.

Q. Wei and Y. Xu are with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (phone: +86-10-82544761; fax: +86-10-82544799; email: qinglai.wei@ia.ac.cn). R. Song and D. Liu are with the School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China (phone: +86-10- 62332905; e-mail: ruizhuosong@ustb.edu.cn; derong@ustb.edu.cn).

Then, the tracking error system can be expressed as

$$\begin{aligned} y(k+1) &= x(k+1) - x_r(k+1) \\ &= a(y(k)) + gc(k) \end{aligned} \tag{5}$$

where $a(y(k)) = f(x_r(k) + y(k)) - f(x_r(k))$ and the tracking error control input $c(k) = u(k) - u_r(k)$.

To find an optimal-tracking-error control $c(k)$ is our objective, which makes the tracking error system (5) stable, and makes the following performance index function minimum

$$J(y(0), \underline{c}(0)) = \sum_{k=0}^{\infty} U(y(k), c(k)) \tag{6}$$

where $\underline{c}(0) = \{c(0), c(1), \dots\}$ is the tracking error control input sequence, and $U(y(k), c(k)) > 0$, for $\forall\, y(k), c(k) \neq 0$, is the utility function.

Then we can define its optimal performance index function to be

$$J^*(y(k)) = \min_{\underline{c}(k)} \big\{ J(y(k), \underline{c}(k)) \colon \underline{c}(k) \in \underline{\mathfrak{U}}(k) \big\}, \tag{7}$$

where $\underline{\mathfrak{U}}(k) = \big\{ \underline{c}(k) \colon \underline{c}(k) = (c(k), c(k+1), \ldots),\ \forall c(k+i) \in \mathbb{R}^m, i = 0, 1, \dots \big\}$.

On the other hand, define Q-Bellman equation as

$$\begin{aligned} Q^*(y(k), c(k)) =& U(y(k), c(k)) \\ &+ \min_{c(k+1)} Q^*(y(k+1), c(k+1)). \end{aligned} \tag{8}$$

Therefore, the optimal performance index function satisfies

$$J^*(y(k)) = \min_{c(k)} Q^*(y(k), c(k)) \tag{9}$$

and the optimal tracking control input is expressed as

$$c^*(y(k)) = \arg\min_{c(k)} Q^*(y(k), c(k)). \tag{10}$$

We know that the optimal Q function $Q^*(y(k), c(k))$ is generally an unknown and non-analytic function, which cannot be obtained directly by (8). Hence, a discrete-time policy iteration Q learning algorithm will be presented to obtain the approximate optimal Q function iteratively.

III. Policy Iteration Q-Learning Algorithm for Optimal Tracking Control

In this section, the policy iteration Q-learning algorithm will be developed while obtaining the optimal tracking controller for discrete-time nonlinear systems. Stability and convergence proofs will be given to show the iterative Q-learning algorithm properties.

In the developed policy iteration Q-learning algorithm, the iterative tracking error control law and iterative Q function are updated by iterations, as the iteration index $i \to \infty$. For $i = 0$, by an arbitrary admissible tracking control $c^{[0]}(k)$ [13], the initial iterative Q function $Q^{[0]}(y(k), c(k))$ is constructed by the following generalized Q-Bellman equation

$$\begin{aligned} Q^{[0]}(y(k), c(k)) =& U(y(k), c(k)) \\ &+ Q^{[0]}(y(k+1), c^{[0]}(y(k+1))). \end{aligned} \tag{11}$$

Then, the iterative tracking control is computed by

$$c^{[1]}(y(k)) = \arg\min_{c(k)} Q^{[0]}(y(k), c(k)). \tag{12}$$

For $i = 1, 2, \cdots$, the iterative Q function $Q^{[i]}(y(k), c(k))$ satisfies the following generalized Q-Bellman equation

$$\begin{aligned} Q^{[i]}(y(k), c(k)) =& U(y(k), c(k)) \\ &+ Q^{[i]}(y(k+1), c^{[i]}(y(k+1))) \end{aligned} \tag{13}$$

and the iterative tracking control is updated by

$$c^{[i+1]}(y(k)) = \arg\min_{c(k)} Q^{[i]}(y(k), c(k)). \tag{14}$$

IV. Properties of the Policy Iteration Q-Learning Algorithm

In this section, the detailed property analysis of the developed policy iteration Q-learning algorithm will be given .

Theorem 1: Let $Q^{[i]}(y(k), c(k))$ and $c^{[i]}(y(k))$ be updated by the policy iteration Q-learning algorithm, for $i = 0, 1, 2, \dots$ (11)–(14). Then for $\forall i = 0, 1, \dots$, the iterative tracking error control $c^{[i]}(y(k))$ makes the tracking error system (5) stable.

Proof: Define Lyapunov candidate as follows

$$V^{[i]}(y(k)) = Q^{[i]}(y(k), c^{[i]}(y(k))). \tag{15}$$

Then we have

$$\begin{aligned} &V^{[i]}(y(k+1)) - V^{[i]}(y(k)) \\ &\quad= Q^{[i]}(y(k+1), c^{[i]}(y(k+1))) - Q^{[i]}(y(k), c^{[i]}(y(k))) \\ &\quad= -U(y(k), c^{[i]}(y(k))) \\ &\quad< 0. \end{aligned} \tag{16}$$

Then $c^{[i]}(y(k)))$ can make the tracking error system (5) stable. ■

In the following theorems, the convergence property of the policy iteration Q-learning algorithm will be proven.

Theorem 2: For $i = 0, 1, \dots$, let $Q^{[i]}(y(k), c(k))$ and $c^{[i]}(y(k))$ be updated by the policy iteration Q-learning algorithm (11)–(14). Then the iterative Q function $Q^{[i]}(y(k), c(k))$ is monotonically non-increasing, i.e.,

$$Q^{[i+1]}(y(k), c(k)) \leq Q^{[i]}(y(k), c(k)). \tag{17}$$

Proof: According to (14), we have

$$\begin{aligned} Q^{[i]}(y(k), c^{[i+1]}(y(k))) &= \min_{c(k)} Q^{[i]}(y(k), c(k)) \\ &\leq Q^{[i]}(y(k), c^{[i]}(y(k))). \end{aligned} \tag{18}$$

For $i = 0, 1, \dots$, define a new iterative Q function $\mathcal{Q}^{[i+1]}(y(k), c(k))$ as

$$\begin{aligned} \mathcal{Q}^{[i+1]}(y(k), c(k)) =& U(y(k), c(k)) \\ &+ Q^{[i]}(y(k+1), c^{[i+1]}(y(k+1))), \end{aligned} \tag{19}$$

where $c^{[i+1]}(y(k))$ is obtained by (14). According to (18) for $\forall y(k), c(k)$, we can obtain

$$\begin{aligned}
&\mathcal{Q}^{[i+1]}(y(k), c(k)) \\
&\quad = U(y(k), c(k)) + Q^{[i]}(y(k+1), c^{[i+1]}(y(k+1))) \\
&\quad = U(y(k), c(k)) + \min_{c(k+1)} Q^{[i]}(y(k+1), c(k+1)) \\
&\quad \leq U(y(k), c(k)) + Q^{[i]}(y(k+1), c^{[i]}(y(k+1))) \\
&\quad = Q^{[i]}(y(k), c(k)). \qquad (20)
\end{aligned}$$

Now we prove inequality (17) by mathematical induction. For $i = 0, 1, \ldots$, we have that $c^{[i+1]}(y(k))$ is a stable control input. Then, we have $y(k) \to 0$, for $\forall k \to \infty$. Without loss of generality, let $y(N) = 0$, where $N \to \infty$. We have $c^{[i+1]}(y(N)) = c^{[i]}(y(N)) = 0$, which obtains

$$\begin{aligned}
Q^{[i+1]}(y(N), c^{[i+1]}(y(N))) &= \mathcal{Q}^{[i+1]}(y(N), c^{[i+1]}(y(N))) \\
&= Q^{[i]}(y(N), c^{[i]}(y(N))) \\
&= 0 \qquad (21)
\end{aligned}$$

and

$$\begin{aligned}
Q^{[i+1]}(y(N-1), c(N-1)) &= \mathcal{Q}^{[i+1]}(y(N-1), c(N-1)) \\
&= Q^{[i]}(y(N-1), c(N-1)) \\
&= U(y(N-1), c(N-1)). \qquad (22)
\end{aligned}$$

Let $k = N-2$, we have

$$\begin{aligned}
Q^{[i+1]}(y(N-2), c(N-2)) &= U(y(N-2), c(N-2)) \\
&\quad + Q^{[i+1]}(y(N-1), c^{[i+1]}(y(N-1))) \\
&= U(y(N-2), c(N-2)) \\
&\quad + Q^{[i]}(y(N-1), c^{[i+1]}(y(N-1))) \\
&= \mathcal{Q}^{[i+1]}(y(N-2), c(N-2)) \\
&\leq Q^{[i]}(y(N-2), c(N-2)). \qquad (23)
\end{aligned}$$

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

$$\begin{aligned}
&Q^{[i+1]}(y(L), c(L)) \\
&\quad = U(y(L), c(L)) + Q^{[i+1]}(y(L+1), c^{[i+1]}(y(L+1))) \\
&\quad \leq U(y(L), c(L)) + Q^{[i]}(y(L+1), c^{[i+1]}(y(L+1))) \\
&\quad = \mathcal{Q}^{[i+1]}(y(L), c(L)) \\
&\quad \leq Q^{[i]}(y(L), c(L)). \qquad (24)
\end{aligned}$$

Hence, we can obtain that for $i = 0, 1, \ldots$, the inequality (17) holds for $\forall y(k), c(k)$. The proof is completed. ■

Theorem 3: For $i = 0, 1, \ldots$, let $Q^{[i]}(y(k), c(k))$ and $c^{[i]}(y(k))$ be updated by the policy iteration Q-learning algorithm (11)–(14). Let

$$Q^{\infty}(y(k), c(k)) = \lim_{i \to \infty} Q^{[i]}(y(k), c(k)). \qquad (25)$$

Then $Q^{\infty}(y(k), c(k))$ satisfies the optimal Q-Bellman equation, as $i \to \infty$, i.e.,

$$Q^{\infty}(y(k), c(k)) = U(y(k), c(k)) + \min_{c(k+1)} Q^{\infty}(y(k+1), c(k+1)) \qquad (26)$$

Proof: According to (24), we can obtain

$$\begin{aligned}
&Q^{\infty}(y(k), c(k)) \\
&\quad = \lim_{i \to \infty} Q^{[i+1]}(y(k), c(k)) \leq Q^{[i+1]}(y(k), c(k)) \\
&\quad \leq \mathcal{Q}^{[i+1]}(y(k)) \\
&\quad = U(y(k), c(k)) + Q^{[i]}(y(k+1), c^{[i+1]}(y(k+1))) \\
&\quad = U(y(k), c(k)) + \min_{c(k+1)} Q^{[i]}(y(k+1), c(k+1)) \qquad (27)
\end{aligned}$$

Letting $i \to \infty$, we can obtain

$$\begin{aligned}
Q^{\infty}(y(k), c(k)) \leq & U(y(k), c(k)) \\
& + \min_{c(k+1)} Q^{\infty}(y(k+1), c(k+1)) \qquad (28)
\end{aligned}$$

Letting $\zeta > 0$ be an absolute arbitrary positive number, a positive integer p must exist such that

$$Q^{[p]}(y(k), c(k)) - \varepsilon \leq Q^{\infty}(y(k), c(k)) \leq Q^{[p]}(y(k), c(k)). \qquad (29)$$

Hence, we can get

$$\begin{aligned}
&Q^{\infty}(y(k), c(k)) \\
&\quad \geq Q^{[p]}(y(k), c(k)) - \zeta \\
&\quad = U(y(k), c(k)) + Q^{[p]}(y(k+1), v^{[p]}(y(k+1))) - \zeta \\
&\quad \geq U(y(k), c(k)) + Q^{\infty}(y(k+1), v^{[p]}(y(k+1))) - \zeta \\
&\quad \geq U(y(k), c(k)) + \min_{c(k+1)} Q^{\infty}(y(k+1), c(k+1)) - \zeta \qquad (30)
\end{aligned}$$

Since ζ is arbitrary, we have

$$\begin{aligned}
Q^{\infty}(y(k), c(k)) \geq & U(y(k), c(k)) \\
& + \min_{c(k+1)} Q^{\infty}(y(k+1), c(k+1)) \qquad (31)
\end{aligned}$$

Combining (28) and (31), we can obtain

$$\begin{aligned}
Q^{\infty}(y(k), c(k)) = & U(y(k), c(k)) \\
& + \min_{c(k+1)} Q^{\infty}(y(k+1), c(k+1)) \qquad (32)
\end{aligned}$$

■

V. Simulation Study

In the simulation study section, it is the performance of the developed algorithm in a inverted pendulum system [36] with modifications which need to be examined. The dynamics of the pendulum is expressed as

$$\begin{aligned}
\begin{bmatrix} \dot{x}_1(k) \\ \dot{x}_2(k) \end{bmatrix} &= \begin{bmatrix} x_1(k) + 0.1x_2(k) \\ 0.1\dfrac{g}{\ell}\sin(x_1(k)) + (1 - 0.1\kappa\ell)x_2(k) \end{bmatrix} \\
&\quad + \frac{0.1}{m\ell^2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} u(k). \qquad (33)
\end{aligned}$$

where $m = 1/2\,\mathrm{kg}$ and $\ell = 1/3\,\mathrm{m}$ are the mass and length of the pendulum bar, respectively. Let $\kappa = 0.2$ and $g = 9.8\,\mathrm{m/s}^2$ be the frictional factor and the gravitational acceleration, respectively. Discretization of the system function with the sampling interval $\Delta t = 0.025$s, and let the desired state trajectory be expressed as

$$x_r(k) = \big[\sin(k \cdot \Delta t), \cos(k \cdot \Delta t)\big]. \tag{34}$$

Let the initial state be $x_0 = [1, -1]^T$. Neural networks are widely applied for the developed iterative Q-learning algorithm. The critic network and the action network are chosen as three-layer-BP NNs with the structures of 4–12–1 and 2–12–2, respectively. We choose 500 states and 500 controls in Ω_x and Ω_u, respectively, to train the action and critic networks. Implement the developed policy iteration Q-learning algorithm for 30 iterations. The critic network and the action network have to be trained for 3000 steps with the learning rate of $\alpha_c = \beta_a = 0.01$ inside each iteration step, therefore the training error of the neural network becomes less than 10^{-5}. The plots of the iterative Q function $Q^{[i]}(y(k), c^{[i]}(y(k)))$ are presented in Fig. 1.

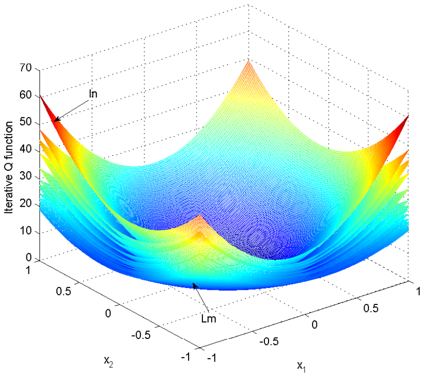

Fig. 1. The iterative Q function

From Fig. 1 we can see that given an arbitrary tracking error admissible control law $\tilde{c}^{[0]}(y(k))$, the iterative Q function is monotonically non-increasing and converges to the optimum. Thus the monotonicity and optimality of the iterative Q function can be justified for nonlinear systems. In Fig. 2, the trajectories of the iterative control laws are shown. In Fig. 3, the tracking errors of the system under the corresponding tracking control law can be seen. Besides, in Fig. 4, the trajectories of the system states are shown. From Figs. 2–4, we can see that for $\forall\, i = 0, 1, \ldots,$ the tracking error can be stabilized and the system states can track the desired trajectories. Hence, the admissibility property of the developed algorithm can be justified.

REFERENCES

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

[2] 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: MIT Press, 1991, pp. 67–95.

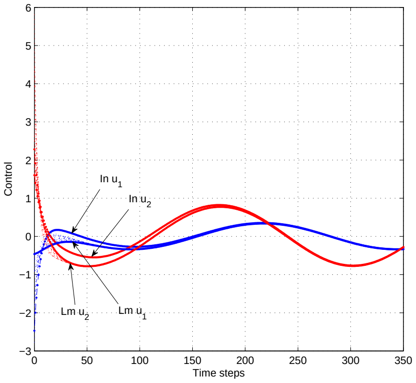

Fig. 2. The trajectories of the iterative controls

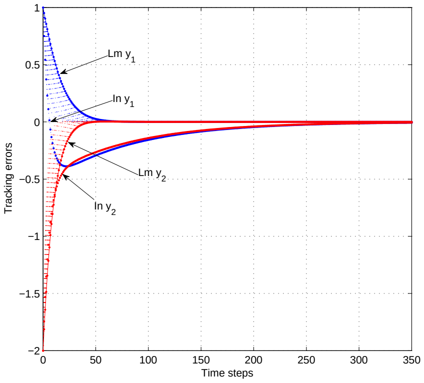

Fig. 3. The trajectories of tracking errors

[3] 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 Systems*, vol. 32, no. 6, pp. 76–105, Dec. 2012.

[4] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, “Adaptive dynamic programming,” *IEEE Transactions on Systems, Man, and Cybernetics–Part C: Applications and Reviews*, vol. 32, no. 2, pp. 140–153, May 2002.

[5] R. Song, F. L. Lewis, Q. Wei, and H. Zhang, “Off-policy actor-critic structure for optimal control of unknown systems with disturbances,” *IEEE Transactions on Cybernetics*, article in press, 2015. DOI:10.1109/TCYB.2015.2421338

[6] Q. Wei, D. Liu, and H. Lin, “Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems,” *IEEE Transactions on Cybernetics*, vol. 46, no. 3, pp. 840–853, Mar.

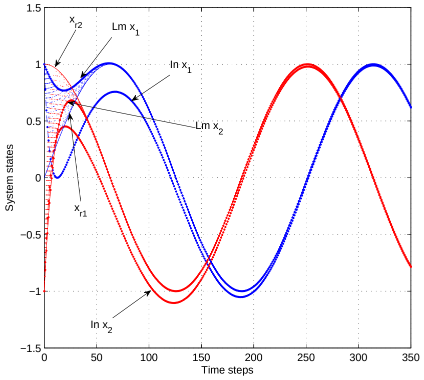

Fig. 4. The trajectories of the system states

2016.

[7] K. G. Vamvoudakis and F. L. Lewis, "Multi-player non-zero-sum games: online adaptive learning solution of coupled Hamilton-Jacobi equations," *Automatica*, vol. 47, no. 8, pp. 1556–1569, Aug. 2011.

[8] 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 Transactions on Cybernetics*, vol. 43, no. 1, pp. 206–216, Feb. 2013.

[9] 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.

[10] H. Zhang, J. Zhang, G. Yang, and Y. Luo, "Leader-based optimal coordination control for the consensus problem of multi-agent differential games via fuzzy adaptive dynamic programming," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 1, pp. 152–163, Jan. 2015.

[11] D. Liu and Q. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 621–634, Mar. 2014.

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

[13] A. Al-Tamimi, F. L. Lewis, and M. Abu-Khalaf, "Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof," *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, vol. 38, no. 4, pp. 943–949, Aug. 2008.

[14] Q. Wei, F. Wang, D. Liu, and X. Yang, "Finite-approximation-error based discrete-time iterative adaptive dynamic programming," *IEEE Transactions on Cybernetics*, vol. 44, no. 12, pp. 2820–2833, Dec. 2014.

[15] D. Liu and Q. Wei, "Finite-approximation-error-based optimal control approach for discrete-time nonlinear systems," *IEEE Transactions on Cybernetics*, vol. 43, no. 2, pp. 779–789, Apr. 2013.

[16] Q. Wei and D. Liu, "Data-driven neuro-optimal temperature control of water gas shift reaction using stable iterative adaptive dynamic programming," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 11, pp. 6399–6408, Nov. 2014.

[17] R. Song, F. L. Lewis, Q. Wei, H. Zhang, Z. P. Jiang, and D. Levine, "Multiple actor-critic structures for continuous-time optimal control using input-output data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 4, pp. 851–865, Apr. 2015.

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

[19] 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 Transactions on Neural Networks and Learning Systems*, vol. 27, no. 2, pp. 444–458, Feb. 2016.

[20] Q. Wei and D. Liu, "Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 4, pp. 1020–1036, Nov. 2014.

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

[22] Q. Wei, R. Song, Q. Sun, and W. Xiao, "Off-policy IRL optimal tracking control for continuous-time chaotic systems," *Chinese Physics B*, vol. 24, no. 9, pp. 090504:1–090504:6, Sep. 2015.

[23] Q. Wei, D. Liu, and Q. Lin, "Discrete-time local iterative adaptive dynamic programming: Terminations and admissibility analysis," *IEEE Transactions on Neural Networks and Learning Systems*, article in press, 2016. DOI: 10.1109/TNNLS.2016.2593743

[24] Q. Wei, D. Liu, Q. Lin, and R. Song, "Discrete-time optimal control via local policy iteration adaptive dynamic programming," *IEEE Transactions on Cybernetics*, article in press, 2016. DOI: 10.1109/T-CYB.2016.2586082

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

[26] Q. Wei and D. Liu, "A novel iterative θ-adaptive dynamic programming for discrete-time nonlinear systems," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 4, pp. 1176–1190, Oct. 2014.

[27] Q. Wei, F. L. Lewis, Q. Sun, P. Yan, and R. Song, "Discrete-time deterministic Q-Learning: A novel convergence analysis," *IEEE Transactions on Cybernetics*, article in press, 2016. DOI: 10.1109/T-CYB.2016.2542923

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

[29] Q. Wei and D. Liu, "A novel policy iteration based deterministic Q-learning for discrete-time nonlinear systems," *Science China Information Sciences*, vol. 58, no. 12, pp 1–15, Dec. 2015.

[30] Q. Wei, D. Liu, and X. Yang, "Infinite horizon self-learning optimal control of nonaffine discrete-time nonlinear systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 4, pp. 866–879, Apr. 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 Transactions on Neural Networks and Learning systems*, vol. 24, no. 10, pp. 1513–1525, Oct. 2013.

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

[33] C. Watkins, *Learning from Delayed Rewards*. Ph.D. Thesis, Cambridge University, Cambridge, England, 1989.

[34] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3–4, pp. 279–292, May 1992.

[35] Q. Wei, F. L. Lewis, Q. Sun, P. Yan, and R. Song, "Discrete-time deterministic Q-learning: A novel convergence analysis," *IEEE Transactions on Cybernetics*, article in press, 2016. DOI: 10.1109/T-CYB.2016.2542923

[36] R. Beard, *Improving the Closed-Loop Performance of Nonlinear Systems*, Ph.D. Thesis, Rensselaer Polytechnic Institute, Troy, NY, 1995.