Optimal Constrained Self-learning Battery Sequential Management in Microgrid Via Adaptive Dynamic Programming

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

Abstract—This paper concerns a novel optimal self-learning battery sequential control scheme for smart home energy systems. The main idea is to use the adaptive dynamic programming (ADP) technique to obtain the optimal battery sequential control iteratively. First, the battery energy management system model is established, where the power efficiency of the battery is considered. Next, considering the power constraints of the battery, a new non-quadratic form performance index function is established, which guarantees that the value of the iterative control law cannot exceed the maximum charging/discharging power of the battery to extend the service life of the battery. Then, the convergence properties of the iterative ADP algorithm are analyzed, which guarantees that the iterative value function and the iterative control law both reach the optimums. Finally, simulation and comparison results are given to illustrate the performance of the presented method.

Index Terms—Adaptive critic designs, adaptive dynamic programming (ADP), approximate dynamic programming, battery management, energy management system, neuro-dynamic programming, optimal control, smart home.

I. INTRODUCTION

NOWADAYS, the need of the smart grid is continuously increasing [1]−[3]. Smart home energy management system is an important component of the smart grid. In smart home energy management systems, the intelligent optimal control of battery is a key technology for saving the power consumptions. In [4], a battery management method was proposed by battery dynamics modeling. In [5], the development of battery management systems was summarized which was applied in the smart grid and electric vehicles. In [6], the operating

Q. L. Wei is 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 Chinese Academy of Sciences, Beijing 100049, China (e-mail: qinglai.wei@ia.ac.cn).

D. R. Liu and R. Z. 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).

Y. Liu is with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yu.liu@ia.ac.cn).

schedule of battery energy storage system was solved by the particle swarm optimization approach. However, in previous researches on the battery management, the properties of the battery management, such as convergence and optimality, were not provided, which limited the applications of the battery control. Adaptive dynamic programming (ADP), proposed by Werbos [7], [8], has been widely used in optimal energy management [6], [9]−[15]. There are several synonyms of ADP, including "adaptive critic designs" [16], "approximate dynamic programming" [11], "neuro-dynamic programming" [17], and "relaxing dynamic programming" [18].

Iterative methods are widely used in ADP to obtain the solution of Hamilton-Jacobi-Bellman (HJB) equation indirectly [14], [19]−[34]. Policy and value iterations are two primary iterative ADP algorithms [35]. Policy iteration algorithms for optimal control of continuous-time (CT) systems with continuous states and action spaces were first given in [36], [37]. In [38], a complex-valued ADP algorithm was discussed, where for the first time the optimal control problem of complex-valued nonlinear systems was successfully solved by ADP. In [39], based on neurocognitive psychology, a novel controller based on multiple actor-critic structures was developed for unknown systems and the proposed controller traded off fast actions based on stored behavior patterns with real-time exploration using current input-output data. In [40], an effective off-policy learning based integral reinforcement learning (IRL) algorithm was presented, which successfully solved the optimal control problem for completely unknown continuous-time systems with unknown disturbances. In [41], a policy iteration algorithm for discrete-time nonlinear systems was developed. Value iteration algorithms for optimal control of discrete-time nonlinear systems were given in [17]. In [18] and [42], the convergence properties of the value iteration were proposed. Value iteration algorithms are generally initialized by a "zero" performance index function [18], [42], [43], which guarantees the convergence properties of the iterative value functions. In [44], a Q-learning based ADP algorithm was proposed to obtain the optimal control law for the battery, which solved the optimal energy management for the microgrid of smart homes. In [9], considering renewable electricity, such as electricity generated from wind and solar energies, the optimal control for the battery was solved by ADP method. In [10], a particle swarm optimization (PSO) method was proposed to pre-train the weights of the action and critic neural networks, which facilitated the implementation of

the ADP method for the optimal control of the battery. In [45], an effective dual Q-learning based iterative ADP algorithm was developed to obtain the optimal battery management for the microgrid of smart homes, where the convergence and optimality of the dual Q-learning based ADP algorithm were proven to guarantee the optimal battery control. However, in [45], the charging/discharging constraints of the battery were not considered in the performance index function. Actually, for all the energy management systems of the microgrid of smart homes, the charging and discharging power of the battery cannot reach infinity. Hence, the optimal control of the battery with power constraints of the battery is a key technique for real-world smart home energy management systems.

In this paper, inspired by [45], a new iterative ADP algorithm is developed to solve the optimal battery control for the smart home energy management system, where the charging/discharging constraints of the battery are considered. First, the models of the smart home energy systems and the battery are established, where the efficiency of the battery is considered. Second, inspired by [37], a new non-quadratic performance index function is constructed, where the charging/discharging power of the battery is defined in the performance index function. Then, the iterative ADP algorithm is derived for the optimal control law of the battery. Via the system transformation and the definition of the performance index function, the expression of the iterative sequential control law for the battery can be obtained. The convergence and optimality of the algorithm are presented, which guarantees that the iterative value function will converge to the optimal performance index function, as the iteration index increases to infinity.

The rest of this paper is organized as follows. In Section II, the problem formulation is presented. The model of the smart home energy system is constructed. The operation principle of the battery is introduced. The optimization objectives of the control problem are also declared. In Section III, the iterative ADP algorithm for battery management system is established. According to the system transformation and the optimality principle, the iterative ADP algorithm is derived. The convergence properties will also be proven in this section. In Section IV, numerical results are presented to demonstrate the effectiveness of the developed algorithm. Finally, in Section V, the conclusion is drawn.

II. PROBLEM FORMULATION

In this section, smart home energy systems with a battery will be described. The optimization objectives of our research will be defined and the corresponding principle of optimality will be introduced.

A. Smart Home Energy Systems

In this paper, the optimal battery control problem is treated as a discrete time problem with the time step of 1 hour and it is assumed that the home load varies hourly. The home load P_{Lt} and the electricity rate C_t are periodic functions with the period $\lambda = 24$ hours. The battery will make decisions to meet the demand of the home load, according to the real-time electricity rate. There are three operational modes for the battery of the home energy system, which are charging mode, idle mode, and discharging mode, respectively.

B. Battery Model

The battery model used in this work is based on [6], [44], [46] where battery efficiency is considered to extend the battery's lifetime as far as possible. Let E_{bt} be the battery energy at time t and let $\eta(\cdot)$ be the charging/discharging efficiency of the battery. Then, the battery model can be expressed as

$$E_{b(t+1)} = E_{bt} - P_{bt} \times \eta(P_{bt}) \tag{1}$$

where P_{bt} is the battery power output at time t. Let $P_{bt} > 0$ denote battery discharging. Let $P_{bt} < 0$ denote battery charging and let $P_{bt} = 0$ denote battery idle. The efficiency of battery charging/discharging is derived by [6], [44], [46], which is expressed as

$$\eta(P_{bt}) = 0.898 - 0.173|P_{bt}|/P_{\text{rate}} \tag{2}$$

where $P_{\text{rate}} > 0$ is the rated power output of the battery. To extend the battery's lifetime, two constraints need to be considered:

1) The storage limit is considered:

$$E_b^{\min} \le E_{bt} \le E_b^{\max} \tag{3}$$

where $E_b^{\min}$ and $E_b^{\max}$ are the minimum and maximum storage energy of the battery, respectively.

2) The charging and discharging power limits are considered:

$$P_b^{\min} \le P_{bt} \le P_b^{\max} \tag{4}$$

where $P_b^{\min}$ and $P_b^{\max}$ are the minimum and maximum charging/discharging powers of the battery, respectively.

C. Optimization Objectives

Given the home load and real-time electricity rate, the objective of the optimal control is to find the optimal battery charging/discharging/idle control law at each time step which minimizes the total expense of the power from the grid while considering the battery limitations. To find the optimal control law, the load balance should be considered. Let P_{Tt} be the power of the home load at time t and let P_{gt} be the power from the power grid.

To establish the equation of the home energy system, we introduce a delay in P_{bt} and then we have $P_{Tt} = P_{b(t-1)}\eta(P_{b(t-1)}) + P_{gt}$. We denote $P_{L(t-1)} = P_{Tt}$ and then we can define the load balance as

$$P_{L(t-1)} = P_{b(t-1)}\eta(P_{b(t-1)}) + P_{gt}. \tag{5}$$

In this paper, the power flow from the battery to the grid is not permitted, i.e., we define $P_{gt} \ge 0$, to guarantee the power quality of the grid. To extend the lifetime of the battery, according to (3), we desire that the stored energy of the battery is close to the middle of storage limit E_b^o, where

$$E_b^o = \frac{1}{2}(E_b^{\min} + E_b^{\max}). \tag{6}$$

In [45], a quadratic form performance index function was proposed, which was expected to be minimized

$$\sum_{t=0}^{\infty} \left(m_1(C_t P_{gt})^2 + m_2(E_{bt} - E_b^o)^2 + m_3(P_{bt})^2 \right) \quad (7)$$

where m_1, m_2, and m_3 are given positive constants. In the performance index function (7), the first term aims to minimize the total cost from the grid. The second term avoids fully charging/discharging of the battery and the third term aims to minimize the charging/discharging power of the battery.

III. ITERATIVE ADP ALGORITHM FOR BATTERY MANAGEMENT SYSTEM

In [45], a Q-learning based iterative ADP algorithm was developed to obtain the optimal control law of the battery, which minimized the performance index function (7). In this section, inspired by [45], considering the power efficiency and the power constraint of the battery, a new system function will be constructed and a new performance index function will be established.

A. System Transformations

First, we define the system states. Let $x_{1t} = P_{gt}$ and $P_{bt} = u_t$. According to (5), we have

$$x_{1,t+1} = P_{Lt} - u_t \eta(u_t). \quad (8)$$

Let $x_{2t} = E_{bt} - E_b^o$. According to (1), we can get

$$x_{2,t+1} = x_{2,t} - u_t \eta(u_t). \quad (9)$$

Letting $x_t = [x_{1t}, x_{2t}]^T$, the equation of the home energy system can be written as

$$x_{t+1} = F(x_t, u_t, t) = \begin{pmatrix} P_{Lt} - u_t \eta(u_t) \\ x_{2t} - u_t \eta(u_t) \end{pmatrix}. \quad (10)$$

In [45], the performance index function was defined as (7). However, in (7), the power constraint of the battery was not considered. In this paper, inspired by [37], a non-quadratic performance index function will be defined for the battery management system, which is expressed as

$$\sum_{t=0}^{\infty} \left(m_1(C_t P_{gt})^2 + m_2(E_{bt} - E_b^o)^2 + m_3 \int_0^{P_{bt}} \left(\Phi^{-1}(s) \right)^T ds \right) \quad (11)$$

where $\Phi(\cdot)$ is a monotonic odd function with its first derivative bounded by a constant $\mathcal{M}$. An example is the hyperbolic tangent $\Phi(\cdot) = \tanh(\cdot)$. R is a positive definite matrix. In (11), the upper bound of the parameter s is not larger than P_{bt}. However, in (11), the power efficiency of the battery is not described. Thus, we further define a new performance index function as

$$\sum_{t=0}^{\infty} \left(m_1(C_t P_{gt})^2 + m_2(E_{bt} - E_b^o)^2 + m_3 \int_0^{P_{bt}\eta(P_{bt})} \left(\Phi^{-1}(s) \right)^T ds \right). \quad (12)$$

Remark 1: The smart home energy system (10) is different from the one in [45]. First, in equation (3) of [45], the battery efficiency of the battery was not considered in the power balance of the load, while in (10) of this paper, the battery efficiency is considered both in the power balance of the load and the battery power balance. Second, the performance index function defined in [45] did not consider the efficiency of the battery. This makes it possible for the optimal control law to exceed the max/min power of the battery, which may make the optimal control invalid. In this paper, both the max/min power of the battery and the efficiency of the battery are considered in the performance index function (12). Thus, the models of the system (10) and performance index function (12) are more reasonable.

B. Optimality Principle

Let $\underline{u}_t = (u_t, u_{t+1}, \ldots)$ denote the control sequence from t to ∞. Let $M_t = \begin{bmatrix} m_1 C_t^2 & 0 \\ 0 & m_2 \end{bmatrix}$. The performance index function (12) can be written as

$$J(x_t, \underline{u}_t, t) = \sum_{i=t}^{\infty} \gamma^i U(x_i, u_i, i) \quad (13)$$

where the utility function is expressed as

$$U(x_t, u_t, t) = x_t^T M_t x_t + m_3 \int_0^{u_t \eta(u_t)} \left(\Phi^{-1}(s) \right)^T ds. \quad (14)$$

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

$$J^*(x_t, t) = \inf_{\underline{u}_t} \left\{ J(x_t, \underline{u}_t, t) \colon \underline{u}_t \in \underline{\mathfrak{U}}_t \right\}. \quad (15)$$

According to Bellman's principle of optimality [47], we can obtain the following discrete-time HJB equation

$$J^*(x_t, t) = \inf_{u_t} \left\{ U(x_t, u_t, t) + J^*(x_{t+1}, t+1) \right\}. \quad (16)$$

The optimal sequential control law can be expressed as

$$u^*(x_t, t) = \inf_{u_t} \left\{ U(x_t, u_t, t) + J^*(x_{t+1}, t+1) \right\}. \quad (17)$$

Remark 2: We can see that the home energy system (10) is a nonlinear dynamic system. The optimal battery control is actually an infinite horizon optimal control problem for nonlinear system with a non-quadratic performance index function. In this situation, many static mathematical programming methods, such as linear programming, are not effective. Dynamic programming is a powerful method to solve these problems. However, if we adopt the traditional dynamic programming method to obtain the optimal performance index function one step at a time, then we have to face the "curse of dimensionality". Thus, a new iterative ADP algorithm will be developed in this paper.

C. Derivations of the Iterative ADP Algorithm

From (10) we know that the battery management system is time-varying. It means that the control law is also time-varying. This makes the controller design difficult. To overcome this difficulty, in [45], [48], by defining a new sequence of control for a period, the time-varying optimal control was transformed into a time-invariant one, which significantly relaxed the computation burden. In this paper, inspired by [45], [48], we will define the sequence of control for a period, where the constraints of the battery are considered. For any $k = 0, 1, \ldots$, we define $\mathcal{U}_k$ as the control sequence from k to $k + \lambda - 1$, i.e., $\mathcal{U}_k = (u_k, u_{k+1}, \ldots, u_{k+\lambda-1})$. We can define a new utility function as

$$\Lambda(x_k, \mathcal{U}_k) = \sum_{\theta=0}^{\lambda-1} U(x_{k+\theta}, u_{k+\theta}, \theta). \qquad (18)$$

Then, for any $k = 0, 1, \ldots$, the optimal performance index function is obtained as

$$J^*(x_k) = \min_{\mathcal{U}_k} \left\{ \Lambda(x_k, \mathcal{U}_k) + \bar{\gamma} J^*(x_{k+\lambda}) \right\} \qquad (19)$$

where $\bar{\gamma} = \gamma^\lambda$. The optimal sequential control law sequence can be expressed by

$$\mathcal{U}^*(x_k) = \arg \min_{\mathcal{U}_k} \{ \Lambda(x_k, \mathcal{U}_k) + \bar{\gamma} J^*(x_{k+\lambda}) \}. \qquad (20)$$

Define an iteration index $i = 0, 1, \ldots$. The iterative value function is defined as

$$V_{i+1}(x_k) = \min_{\mathcal{U}_k} \{ \Lambda(x_k, \mathcal{U}_k) + \bar{\gamma} V_i(x_{k+\lambda}) \} \qquad (21)$$

where $V_0(x_k) = \Psi(x_k)$ and $\Psi(x_k)$ is a positive semi-definite function. The iterative sequential control law sequence $\mathcal{U}_i$ can be computed as follows

$$\mathcal{U}_i(x_k) = \arg \min_{\mathcal{U}_k} \{ \Lambda(x_k, \mathcal{U}_k) + \bar{\gamma} V_i(x_{k+\lambda}) \}. \qquad (22)$$

For any $i = 0, 1, \ldots$, we define a new iteration index $j = 0, 1, \ldots, 23$. We can get

$$V_i^{j+1}(x_k) = \min_{u_k} \{ U(x_k, u_k, j) + \gamma V_i^j(x_{k+1}) \}$$
$$= U(x_k, u_i^j(x_k), j) + \gamma V_i^j(x_{k+1}) \qquad (23)$$

and

$$u_i^j(x_k) = \arg \min_{u_k} \{ U(x_k, u_k, j) + \gamma V_i^j(x_{k+1}) \} \qquad (24)$$

where we let $V_i^0(x_k) = V_i(x_k)$ and $V_{i+1}(x_k) = V_i^{23}(x_k)$.

The system function is defined as

$$x_{k+1} = f(x_k, j) + g v_k \qquad (25)$$

where $f(x_k, j) = [P_{L(\lambda-1-j)}, x_{2k}]^T$, $g = [-1, -1]^T$, and $v_k = u_k \eta(u_k)$. According to the principle of optimality, for any $i = 0, 1, \ldots$ and $j = 0, 1, \ldots, 23$, the iterative control law $v_i^j(x_k)$ satisfies

$$\frac{\partial V_i^{j+1}(x_k)}{\partial v_i^j(x_k)} = 0. \qquad (26)$$

Then, we can obtain that

$$v_i^j(x_k) = -\Phi\left(\frac{1}{2} m_3^{-1} g^T \frac{dV_i^j(x_{k+1})}{dx_{k+1}} \right). \qquad (27)$$

We can see that for any $i = 0, 1, \ldots$, we use $j = 0, 1, \ldots, \lambda - 1$ iterations to obtain the optimal sequential control law sequence for a day.

Remark 3: From (27), we can see that the expression of the iterative sequential control law in this paper is different from the one in [45]. In [45], the iterative sequential control law was obtained by minimizing a Q function, such as

$$u_i^j(x_k) = \arg \min_{u_k} Q_i^j(x_k, u_k). \qquad (28)$$

Generally, the Q function is not an analytical function. In this situation, the iterative control law in [45] does not possess an analytical expression. In this paper, according to the iterative value function $V_i^j(x_k)$, the expressions of the iterative control law can be obtained according to (27), which is simpler than the one in [45]. This is a merit of the algorithm developed in this paper. On the other hand, in [45], the iterative control law could be obtained directly by minimizing the iterative Q function, where the system function was not required. However, in this paper, to obtain the expression of the iterative control law, the system function (25) is required. This is a disadvantage of the method in this paper.

Now, we let $v_k = \mathcal{G}(u_k) = u_k \eta(u_k)$. Then, we can obtain

$$u_k = \mathcal{G}^{-1}(v_k). \qquad (29)$$

Hence, for any $i = 0, 1, \ldots$, and $j = 0, 1, \ldots, 23$, we can get $u_i^j(x_k) = \mathcal{G}^{-1}(v_i^j(x_k))$. For a given $v_i^j(x_k)$, if there exist several iterative control laws $u_i^j(x_k)$ that satisfy (29), we use the iterative control law that minimizes the norm $\|u_i^j(x_k)\|$.

D. Properties of the Iterative ADP Algorithm

In this subsection, the convergence and optimality properties of the proposed iterative ADP algorithm will be developed. It will be shown that the iterative value function and iterative control law will converge to their optimums as the iteration index i increases to infinity.

Theorem 1: For $i = 0, 1, \ldots$, and $j = 0, 1, \ldots, 23$, let the iterative value function $V_i^j(x_k)$ and the iterative control law $v_i^j(x_k)$ be obtained by (23) and (24). Then, the iterative control law sequence can be expressed by

$$\mathcal{U}_i(x_k) = \{ u_i^{23}(x_k), u_i^{22}(x_{k+1}), \ldots, u_i^0(x_{k+23}) \}. \qquad (30)$$

Proof: The statement can be proven by mathematical induction. For any $i = 0, 1, \ldots$, we have

$$V_i^{j+1}(x_k)$$
$$= \min_{u_k} \left(U(x_k, u_k, j) + \gamma V_i^j(x_{k+1}) \right)$$
$$= \min_{u_k} \left(U(x_k, u_k, j) + \gamma \min_{u_{k+1}} \left(U(x_{k+1}, u_{k+1}, j-1) \right. \right.$$
$$\left. \left. + \cdots + \gamma \min_{u_{k+j}} \left(U(x_{k+j}, u_{k+j}, 0) + \gamma V_i^0(x_{k+j+1}) \right) \right) \right)$$

$$= \min_{(u_k, u_{k+1}, \ldots, u_{k+j})} \left(\sum_{l=0}^{j} \gamma^l U(x_{k+l}, u_{k+l}, j-l) \right.$$
$$\left. + \gamma^{j+1} V_i(x_{k+j+1}) \right). \tag{31}$$

First, let $j = 0$. According to (24), we can derive that

$$u_i^0(x_k) = \arg \min_{u_k} \{ U(x_k, u_k, 0) + \gamma V_i^0(x_{k+1}) \} \tag{32}$$

where $x_{k+1} = f(x_k, 0) + g v_k$. The function $f(x_k, 0)$ can be expressed as $f(x_k, 0) = [P_{L(23)}, x_{2k}]^T$. For $j = 0, 1, \ldots, 23$, (24) holds. Hence, the sequential sequence of iterative control law in (31) can be expressed as

$$(u_k, u_{k+1}, \ldots, u_{k+j})$$
$$= (u_i^{23}(x_k), u_i^{22}(x_{k+1}), \ldots, u_i^{\lambda-1-j}(x_{k+j})). \tag{33}$$

Let $j = 23$. We can obtain that (30) holds. ∎

From Theorem 1, for $i = 0, 1, \ldots$, we can say that the total cost in each period can be minimized by the iterative sequential control law sequence $\mathcal{U}_i(x_k)$ according to the local iteration (23)−(27). In [18], [20], an effective "functional bound" method was proposed by Rantzer for the iterative ADP algorithm. Next, inspired by [18], [20], the convergence property of the iterative value function $V_i(x_k)$ will be developed.

Theorem 2: For $i = 0, 1, \ldots$, let $V_{i+1}(x_k)$ and $\mathcal{U}_i(x_k)$ be obtained by (21) and (22). Then, the iterative value function $V_i(x_k)$ converges to the optimal performance index function, i.e.,

$$\lim_{i \to \infty} V_i(x_k) = J^*(x_k). \tag{34}$$

Proof: According to the control law sequence $\mathcal{U}_k = (u_k, u_{k+1}, \ldots, u_{k+\lambda-1})$, we can obtain

$$x_{k+\lambda} = \mathcal{F}(x_k, \mathcal{U}_k). \tag{35}$$

Inspired by [18], [20], assume that there are constants ψ_1, ψ_2, and λ, such that $0 \leq \psi_1 \leq 1 \leq \psi_2 < \infty$ and $0 < \chi < \infty$, which satisfy $J^*(\mathcal{F}(x_k, \mathcal{U}_k)) \leq \chi U(x_k, \mathcal{U}_k)$, and $\psi_1 J^*(x_k) \leq V_0(x_k) \leq \psi_2 J^*(x_k)$.

For $i = 0, 1, \ldots$, the iterative value function $V_i(x_k)$ satisfies

$$\left(1 + \frac{\psi_1 - 1}{(1 + \chi^{-1})^i} \right) J^*(x_k) \leq V_i(x_k)$$
$$\leq \left(1 + \frac{\psi_2 - 1}{(1 + \chi^{-1})^i} \right) J^*(x_k). \tag{36}$$

Inequality (36) can be proven by mathematical induction. Let $i = 0$. From (21), we have

$$V_1(x_k) = \min_{\mathcal{U}_k} \{ U(x_k, \mathcal{U}_k) + \bar{\gamma} V_0(x_{k+\lambda}) \}$$
$$\geq \min_{u_k} \left\{ \left(1 + \chi \frac{\psi_1 - 1}{1 + \chi} \right) U(x_k, \mathcal{U}_k) \right.$$
$$\left. + \bar{\gamma} \left(\psi_1 - \frac{\psi_1 - 1}{1 + \chi} \right) J^*(x_{k+\lambda}) \right\}$$
$$\geq \left(1 + \frac{\psi_1 - 1}{(1 + \chi^{-1})} \right) \min_{u_k} \{ U(x_k, \mathcal{U}_k) + \bar{\gamma} J^*(x_{k+\lambda}) \}$$
$$= \left(1 + \frac{\delta - 1}{(1 + \bar{\gamma}^{-1})} \right) J^*(x_k). \tag{37}$$

Following a similar procedure, we can prove inequality (36) using mathematical induction for $i = 0, 1, \ldots$. Letting $i \to \infty$, we can obtain

$$\lim_{i \to \infty} \left\{ \left(1 + \frac{\psi_1 - 1}{(1 + \chi^{-1})^i} \right) J^*(x_k) \right\}$$
$$= \lim_{i \to \infty} \left\{ \left(1 + \frac{\psi_2 - 1}{(1 + \chi^{-1})^i} \right) J^*(x_k) \right\}$$
$$= J^*(x_k). \tag{38}$$

∎

Corollary 1: For $i = 0, 1, \ldots$, let $V_{i+1}(x_k)$ and $\mathcal{U}_i(x_k)$ be obtained by (21) and (22). Then the iterative control law sequence $\mathcal{U}_i(x_k)$ converges to the optimal control law sequence, i.e., $\lim_{i \to \infty} \mathcal{U}_i(x_k) = \mathcal{U}^*(x_k)$.

Remark 4: In [45], for a dual Q-learning based iterative ADP algorithm, inspired by [18], [20], it was proven that the iterative Q function converges to the optimal Q function as the iteration index $i \to \infty$. In this paper, inspired by [18], [20], we have shown that the iterative value function $V_i(x_k)$ with the constraints of the battery can also converge to the optimal performance index function $J^*(x_k)$. We should point out that the implementation of the above two iterative ADP algorithms are different. First, for the dual iterative Q-learning algorithm in [45], the system model is not necessary. This is a remarkable advantage of the dual iterative Q-learning algorithm. However, in each iteration of the dual iterative Q-learning algorithm, it is required to search both of the state and control spaces to update the iterative Q function. Hence in this case, the computation load is high. In the proposed iterative ADP algorithm (21)−(27), we can see that the iterative value function can be updated only in state space. Thus, the computation burden of the present algorithm is relaxed. This is an advantage of the iterative ADP algorithm in this paper. On the other hand, from (27) we can see that the system model is necessary to obtain the iterative control law. This is a disadvantage of the algorithm.

IV. SIMULATION ANALYSIS

In this section, the performance of the iterative ADP algorithm with constraints will be examined by numerical experiments. The simulation results for the developed iterative ADP algorithm with constraints will be compared with the dual Q-learning algorithm in [45]. The profiles of the home load demand (kW) and the real-time electricity rate (in cents) are taken from [44], where the home load demand and the real-time electricity rate for four weeks (672 hours) are shown in Figs. 1 (a) and (c), respectively. We can see that the home load demand and the real-time electricity rate are quasi-periodic functions and the periods are $\lambda = 24$ hours. According to the functions of the home load demand and the real-time electricity rate for 672 hours, we can obtain the average trajectories of the home load demand and the electricity rate which are shown in Figs. 1 (b) and (d).
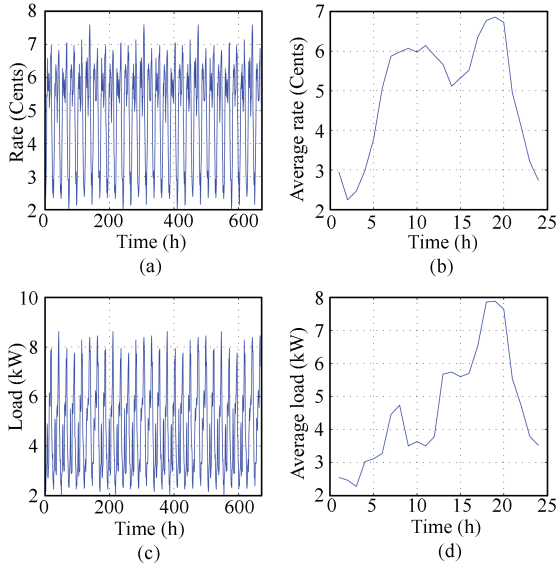
Fig. 1. Home load demand and electricity rate. (a) Home load demand for 672 hours. (b) Average home load demand in a day. (c) Real-time electricity rate for 672 hours. (d) Average electricity rate in a day.

Let the capacity of the battery be chosen as in [45], which is 100 kWh. Let the upper and lower storage limits of the battery be $E_b^{\min} = 20$ kWh and $E_b^{\max} = 80$ kWh, respectively, which are the same as those in [45]. The rated power output of the battery namely the maximum charging/discharging rate is 10 kW. Let the initial battery energy be 65 kWh. Let the performance index function be expressed as in (7), where we set $m_1 = 0.9$, $m_2 = 0.5$ and $m_3 = 0.5$. Let the initial function $\Psi(x_k) = x_k^T \Pi x_k$, where Ψ is arbitrarily chosen as a positive semi-definite matrix with

$$\Pi = \begin{bmatrix} 0.2894 & 0.6852 \\ 0.6852 & 2.3505 \end{bmatrix}.$$

Neural networks are used to implement the iterative ADP algorithm. There are two neural networks, which are critic and action networks, respectively. Both neural networks are chosen as three-layer back-propagation (BP) network. The structures of the critic and action networks are chosen as $2-8-1$ and $2-8-1$, respectively. The training methods of the neural networks are shown in [45], [49] and omitted here. Now we let the real maximum charging/discharging power be 4.5 kW. According to these data, we implement the iterative ADP algorithm $(21)-(27)$ for 15 iterations to guarantee the computation precision $\varepsilon = 10^{-2}$. The plots of the iterative value function are shown in Fig. 2.

From Fig. 2, we can see that under the power constraints of the battery, the iterative value function converges to the optimum in 15 iterations. The optimal battery charging/discharging management is shown in Fig. 3. From Fig. 3, we can see that the charging and discharging power of the battery cannot exceed the bound of 4.5 kW, which shows the effectiveness of the developed algorithm. The plot of battery energy is shown in Fig. 4. From Fig. 4, we can see that the energy of the battery does not exceed the maximum energy of the battery.
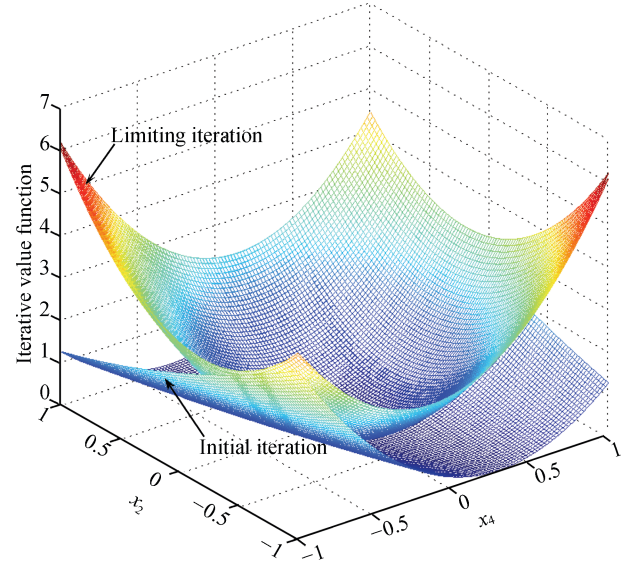

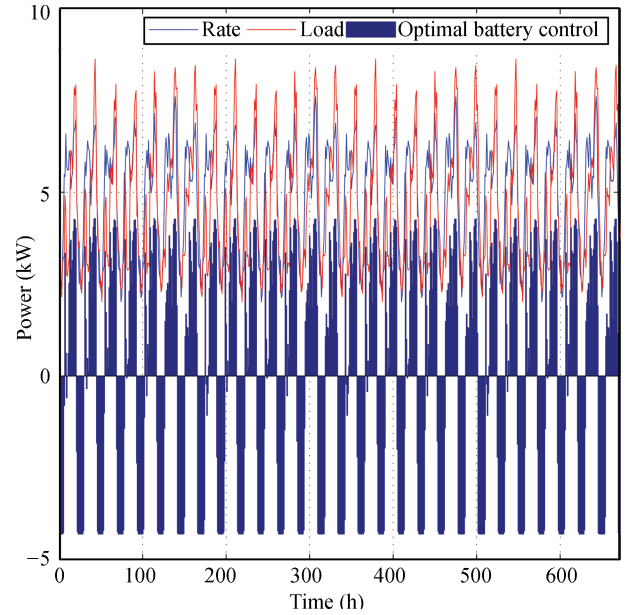Fig. 2. The trajectories of the iterative value function.

Fig. 3. Optimal management of battery with constraints in four weeks.

In [45], a dual Q-learning based iterative ADP algorithm was developed to solve the optimal battery management where the constraints for the power of the battery were not considered. The performance index function is defined as in (7), where the parameters are kept unchanged. Choosing the initial Q function as $Q_0 = [x_k, u_k]^T \bar{\bar{\Pi}} [x_k, u_k]$, where $\bar{\bar{\Pi}}$ is chosen as

$$\bar{\bar{\Pi}} = \begin{bmatrix} 0.3479 & 0.8594 & 0.4914 \\ 0.8594 & 2.8694 & 1.4634 \\ 0.4914 & 1.4634 & 4.1274 \end{bmatrix}.$$

We implement the dual Q-learning based iterative ADP algorithm [45] for 15 iterations which makes the iterative Q learning algorithm converge to the optimum. The optimal battery management with no constraints is shown in Fig. 5. From Fig. 5, we can see that if the power constraints are
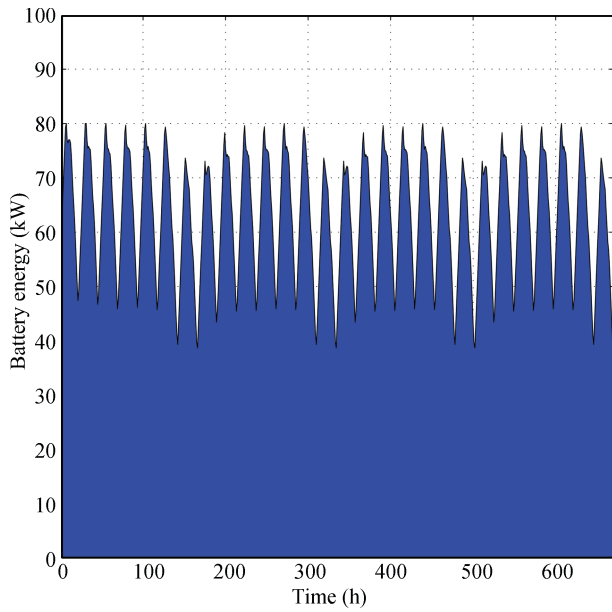
Fig. 4.   Optimal battery energy with constraints in four weeks.
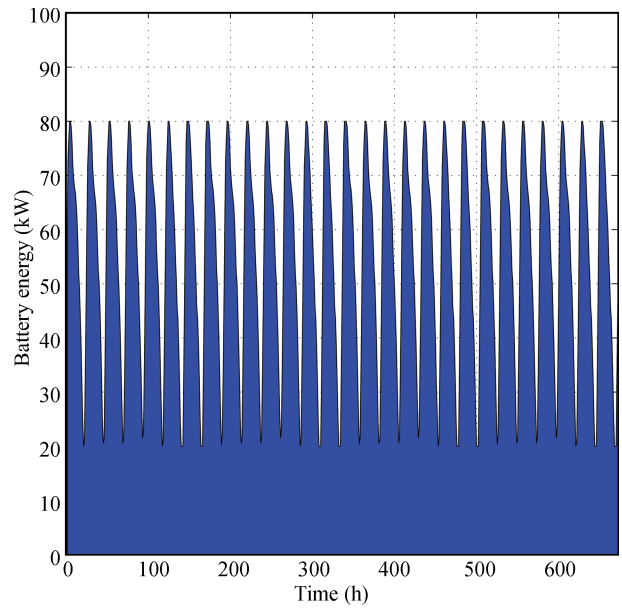

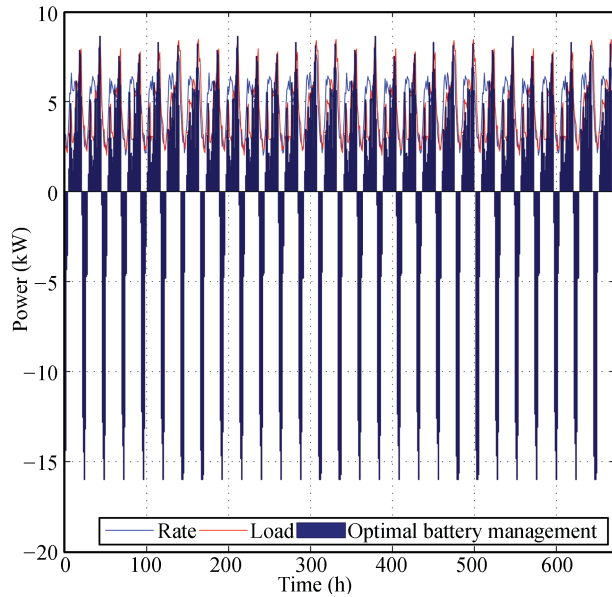Fig. 6.   Optimal battery energy with no constraints.

Fig. 5.   Optimal management of batteries with no constraints in four weeks.
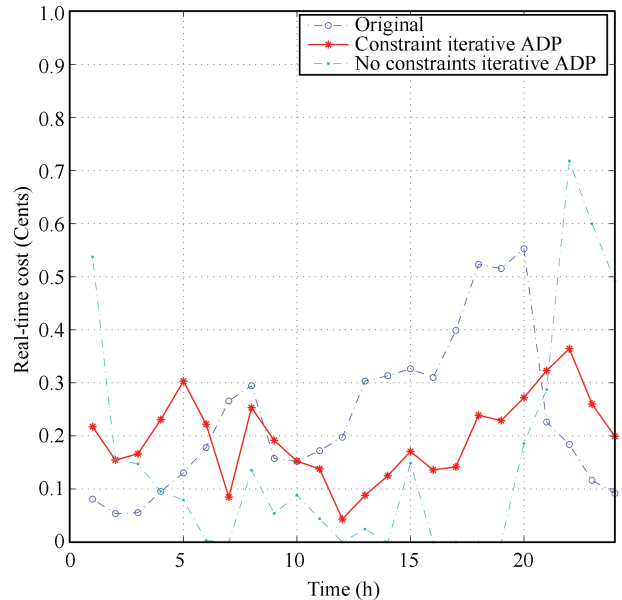

Fig. 7.   Numerical comparisons in a day.

V. Conclusion

In this paper, a new iterative ADP algorithm is developed to solve the optimal battery control problem for the smart home energy systems. Considering the efficiency and the charging/discharging constraints of the battery, the model of the smart home energy system is constructed. A new non-quadratic form performance index function is established, which guarantees the iterative control amplitude not to exceed the upper bound of the battery. Iterative ADP algorithm is developed, where in each iteration, the expressions of the iterative control law can be obtained. The convergence properties of the iterative ADP algorithm are given, which guarantees that the iterative value function and the iterative control law both reach the optimal ones. Finally, simulation and comparison results are given to illustrate the performance of the presented method.

not considered, the maximum battery power reaches 16 kW. The energy of the battery is shown in Fig. 6. However, adding the constraints of battery, we can see that the power of the battery does not exceed 4.5 kW, which prevents the large charging/discharging of the battery. In Fig. 6, we can see that the battery reaches its maximum and minimum energies of the battery under the un-constrained optimal battery management. In Fig. 4, the battery nearly does not reach the maximum and minimum energies of the battery. Hence, the proposed algorithm is preferred for the long term operation of the battery. Fig. 7 shows the real-time cost comparisons in a day. We can see that the real-time cost using the iterative ADP with constraints is smaller than the iterative ADP with no constraint, which shows the effectiveness of the developed method.

In this paper, to extend the lifetime of the battery, we aim to draw the stored energy of the battery close to the middle of storage limit and minimize the large charging/discharging power of the battery. On the other hand, the frequency of the battery is not considered in this paper. Since frequent quick charging/discharging frequency may also damage the battery, how to avoid frequent charging/discharging of the battery will be our main future research topic.

REFERENCES

[1] H. P. Li, C. Z. Zang, P. Zeng, H. B. Yu, and Z. W. Li, "A stochastic programming strategy in microgrid cyber physical energy system for energy optimal operation," *IEEE/CAA J. Automat. Sin.*, vol. 2, no. 3, pp. 296−303, Jul. 2015.

[2] G. Chen and E. N. Feng, "Distributed secondary control and optimal power sharing in microgrids," *IEEE/CAA J. Automat. Sin.*, vol. 2, no. 3, pp. 304−312, Jul. 2015.

[3] W. Wei, F. Liu, and S. W. Mei, "Energy pricing and dispatch for smart grid retailers under demand response and market price uncertainty," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1364−1374, May 2015.

[4] A. Szumanowski and Y. H. Chang, "Battery management system based on battery nonlinear dynamics modeling," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 3, pp. 1425−1432, May 2008.

[5] H. Rahimi-Eichi, U. Ojha, F. Baronti, and M. Y. Chow, "Battery management system: an overview of its application in the smart grid and electric vehicles," *IEEE Ind. Electron. Maga.*, vol. 7, no. 2, pp. 4−16, Jun. 2013.

[6] T. Y. Lee, "Operating schedule of battery energy storage system in a time-of-use rate industrial user with wind turbine generators: a multipass iteration particle swarm optimization approach," *IEEE Trans. Energy Convers.*, vol. 22, no. 3, pp. 774−782, Sep. 2007.

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

[8] P. J. Werbos, W. T. Miller, and R. S. Sutton, *Neural Networks for Control*. Cambridge: MIT Press, 1991.

[9] M. Boaro, D. Fuselli, F. De Angelis, D. R. Liu, Q. L. Wei, and F. Piazza, "Adaptive dynamic programming algorithm for renewable energy scheduling and battery management," *Cognit. Comput.*, vol. 5, no. 2, pp. 264−277, Jun. 2013.

[10] D. Fuselli, F. De Angelis, M. Boaro, S. Squartini, Q. L. Wei, D. R. Liu, and F. Piazza, "Action dependent heuristic dynamic programming for home energy resource scheduling," *Int. J. Electr. Power Energy Syst.*, vol. 48, pp. 148−160, Jun. 2013.

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

[12] J. Q. Liang, D. D. Molina, G. K. Venayagamoorthy, and R. G. Harley, "Two-level dynamic stochastic optimal power flow control for power systems with intermittent renewable generation," *IEEE Trans. Power Syst.*, vol. 28, no. 3, pp. 2670−2678, Aug. 2013.

[13] S. Mohagheghi, G. K. Venayagamoorthy, and R. G. Harley, "Fully evolvable optimal neurofuzzy controller using adaptive critic designs," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 6, pp. 1450−1461, Dec. 2008.

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

[15] Z. Zhang and D. B. Zhao, "Clique-based cooperative multiagent reinforcement learning using factor graphs," *IEEE/CAA J. Automat. Sin.*, vol. 1, no. 3, pp. 248−256, Jul. 2014.

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

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

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

[19] Q. L. Wei, F. Y. Wang, D. R. Liu, and X. Yang, "Finite-approximation-error-based discrete-time iterative adaptive dynamic programming," *IEEE Trans. Cybernet.*, vol. 44, no. 12, pp. 2820−2833, Dec. 2014.

[20] A. Rantzer, "Relaxed dynamic programming in switching systems," *IET Proc.-Contr. Theor. Appl.*, vol. 153, no. 5, pp. 567−574, Oct. 2006.

[21] H. Modares and F. L. Lewis, "Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning," *IEEE Trans. Automat. Control*, vol. 59, no. 11, pp. 3051−3056, Nov. 2014.

[22] 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.

[23] H. G. Zhang, T. Feng, G. H. Yang, and H. J. Liang, "Distributed cooperative optimal control for multiagent systems on directed graphs: an inverse optimal approach," *IEEE Trans. Cybernet.*, vol. 45, no. 7, pp. 1315−1326, Jul. 2015.

[24] M. Kumar, K. Rajagopal, S. N. Balakrishnan, et al, "Reinforcement learning based controller synthesis for flexible aircraft wings," *IEEE/CAA J. Automat. Sin.*, vol. 1, no. 4, pp. 435−448, Oct. 2014.

[25] R. Kamalapurkar, J. R. Klotz, and W. E. Dixon, "Concurrent learning-based approximate feedback-Nash equilibrium solution of N-player nonzero-sum differential games," *IEEE/CAA J. Automat. Sin.*, vol. 1, no. 3, pp. 239−247, Jul. 2014.

[26] Q. M. Zhao, H. Xu, and S. Jagannathan, "Near optimal output feedback control of nonlinear discrete-time systems based on reinforcement neural network learning," *IEEE/CAA J. Automat. Sin.*, vol. 1, no. 4, pp. 372−384, Oct. 2014.

[27] Q. L. Wei, R. Z. Song, and P. F. Yan, "Data-driven zero-sum neuro-optimal control for a class of continuous-time unknown nonlinear systems with disturbance using ADP," *IEEE Trans. Neur. Net. Lear. Syst.*, vol. 27, no. 2, pp. 444−458, Feb. 2016.

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

[29] Q. L. Wei and D. R. Liu, "A novel iterative θ-adaptive dynamic programming for discrete-time nonlinear systems," *IEEE Trans. Automat. Sci. Eng.*, vol. 11, no. 4, pp. 1176−1190, Oct. 2014.

[30] Q. L. Wei and D. R. 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.

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

[32] Q. L. Wei, D. R. Liu, and X. Yang, "Infinite horizon self-learning optimal control of nonaffine discrete-time nonlinear systems," *IEEE Trans. Neur. Net. Lear. Syst.*, vol. 26, no. 4, pp. 866−879, Apr. 2015.

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

[34] Q. L. Wei and D. R. Liu, "A novel policy iteration based deterministic Q-learning for discrete-time nonlinear systems," *Sci. China Inform. Sci.*, vol. 58, no. 12, pp. 1−15, Dec. 2015.

[35] F. L. Lewis, D. Vrabie, 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.

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

[37] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach," *Automatica*, vol. 41, no. 5, pp. 779−791, May 2005.

[38] R. Z. Song, W. D. Xiao, H. G. Zhang, and C. Y. Sun, "Adaptive dynamic programming for a class of complex-valued nonlinear systems," *IEEE Trans. Neur. Net. Lear. Syst.*, vol. 25, no. 9, pp. 1733−1739, Sep. 2014.

[39] R. Z. Song, F. Lewis, Q. L. 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. Neur. Net. Lear. Syst.*, vol. 26, no. 4, pp. 851−865, Apr. 2015.

[40] R. Z. Song, F. L. Lewis, Q. L. Wei, and H. G. Zhang, "Off-policy actor-critic structure for optimal control of unknown systems with disturbances," *IEEE Trans. Cybernet.*, vol. 46, no. 5, pp. 1041−1050, May 2016.

[41] D. R. Liu and Q. L. Wei, "Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems," *IEEE Trans. Neur. Net. Lear. Syst.*, vol. 25, no. 3, pp. 621−634, Mar. 2014.

[42] 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.

[43] H. G. Zhang, Q. L. Wei, and Y. H. 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] T. Huang and D. R. Liu, "A self-learning scheme for residential energy system control and management," *Neur. Comput. Appl.*, vol. 22, no. 2, pp. 259−269, Feb. 2013.

[45] Q. L. Wei, D. R. 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.

[46] T. Yau, L. N. Walker, H. L. Graham, A. Gupta, and R. Raithel, "Effects of battery storage devices on power system dispatch," *IEEE Trans. Power Apparatus Syst.*, vol. PAS-100, no. 1, pp. 375−383, Jan. 1981.

[47] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[48] Q. L. Wei, D. R. 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.

[49] J. Si and Y. T. Wang, "Online learning control by association and reinforcement," *IEEE Trans. Neur. Net.*, vol. 12, no. 2, pp. 264−276, Mar. 2001.

Derong Liu (S'91−M'94−SM'96−F'05) received the Ph.D. degree in electrical engineering from the University of Notre Dame in 1994. He was a staff fellow with General Motors Research and Development Center, Warren, MI, from 1993 to 1995. He was an assistant professor in the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, from 1995 to 1999. He joined the University of Illinois at Chicago in 1999, and became a full professor of electrical and computer engineering and of computer science in 2006. He was selected for the "100 Talents Program" by the Chinese Academy of Sciences in 2008. He has published 15 books (six research monographs and nine edited volumes). Currently, he is the Editor-in-Chief of the *IEEE Transactions on Neural Networks and Learning Systems*. He received the Michael J. Birck Fellowship from the University of Notre Dame (1990), the Harvey N. Davis Distinguished Teaching Award from Stevens Institute of Technology (1997), the Faculty Early Career Development (CAREER) award from the National Science Foundation (1999), the University Scholar Award from University of Illinois (2006-2009), and the Overseas Outstanding Young Scholar Award from the National Natural Science Foundation of China (2008). He is a Fellow of the IEEE and a Fellow of the INNS.

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 2009−2011, he was a postdoctoral 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 an associate professor of the institute. He has authored one book, and published over 50 international journal papers. His research interests include adaptive dynamic programming, neural-networks-based control, optimal control, nonlinear systems and their industrial applications.

Dr. Wei is an associate editor of *IEEE Transaction on Systems Man, and Cybernetics: Systems* since 2016, *Information Sciences* since 2016, *Acta Automatica Sinica* since 2015, and has been holding the same position for *IEEE Transactions on Neural Networks and Learning Systems* during 2014-2015. He is the secretary of IEEE Computational Intelligence Society (CIS) Beijing Chapter since 2015. He was registration chair of the 12th World Congress on Intelligent Control and Automation (WCICA2016), 2014 IEEE World Congress on Computational Intelligence (WCCI2014), the 2013 International Conference on Brain Inspired Cognitive Systems (BICS 2013), and the Eighth International Symposium on Neural Networks (ISNN 2011). He was the publication chair of 5th International Conference on Information Science and Technology (ICIST2015) and the Ninth International Symposium on Neural Networks (ISNN 2012). He was the finance chair of the 4th International Conference on Intelligent Control and Information Processing (ICICIP 2013) and the publicity chair of the 2012 International Conference on Brain Inspired Cognitive Systems (BICS 2012). He was guest editors for several international journals. He was a recipient of the Outstanding Paper Award of Acta Automatica Sinica in 2011 and Zhang Siying Outstanding Paper Award of Chinese Control and Decision Conference (CCDC) in 2015. Corresponding author of this paper.

Yu Liu received the B.S. degree in automation, the M.S. degree in pattern recognition and intelligent systems, from Beijing Institute of Technology, Beijing, China, in 2001 and 2004 respectively, and the Ph.D. degree in pattern recognition and intelligent systems, from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2010. His research interests include artificial intelligence, streaming data analysis, and energy management in smart grids.

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. Her research interests include optimal control, neural-networks-based control, nonlinear control, wireless sensor networks, adaptive dynamic programming and their industrial application.