A Novel Dual Iterative Q-Learning Method for Optimal Battery Management in Smart Residential Environments

Qinglai Wei, *Member, IEEE*, Derong Liu, *Fellow, IEEE*, and Guang Shi

***Abstract*—In this paper, a novel iterative Q-learning method called "dual iterative Q-learning algorithm" is developed to solve the optimal battery management and control problem in smart residential environments. In the developed algorithm, two iterations are introduced, which are internal and external iterations, where internal iteration minimizes the total cost of power loads in each period, and the external iteration makes the iterative Q-function converge to the optimum. Based on the dual iterative Q-learning algorithm, the convergence property of the iterative Q-learning method for the optimal battery management and control problem is proven for the first time, which guarantees that both the iterative Q-function and the iterative control law reach the optimum. Implementing the algorithm by neural networks, numerical results and comparisons are given to illustrate the performance of the developed algorithm.**

***Index Terms*—Adaptive critic designs, adaptive dynamic programming (ADP), approximate dynamic programming, neural networks, optimal control, Q-learning, smart grid.**

I. Introduction

With the rising cost, environmental concerns, and reliability issues, the need to develop optimal control and management systems in residential environments is continuously increasing. Smart residential energy systems, composed of power grids, battery systems, and residential loads, which are interconnected over a power management unit, provide end users with the optimal management of energy usage to improve the operation efficiency of power systems [1]–[3]. On the other hand, with the rapidly evolving technology of electric storage devices, energy-storage-based optimal management has attracted much attention [4]–[6]. Along with the development of smart grids, increasing intelligence is required in the optimal design of residential energy systems [7]–[9]. Hence, the intelligent optimization of battery management becomes a key tool for saving the power expense in smart residential environments.

Characterized by strong abilities of self-learning and adaptivity, adaptive dynamic programming (ADP), proposed by Werbos [10], [11], has demonstrated strong capability for finding the optimal control policy and solving the Hamilton–Jacobi–Bellman (HJB) equation forward in time [12]–[18]. Q-learning, proposed by Watkins [19], [20], is a typical ADP method and has been effectively applied to smart energy systems [21], [22]. In [23], Q-learning was denoted as action-dependent heuristic dynamic programming. In [24], a time-based Q-learning (TBQL) algorithm, inspired by [25], was proposed to obtain optimal control for residential energy systems. In [26] and [27], optimal control for residential energy systems was obtained by the TBQL algorithm, where renewable resources, including wind and solar energies, were taken into consideration. In previous TBQL algorithms, however, it is required that the time index t reaches infinity to obtain the optimal Q-function, which means that the optimal Q-function and optimal control law are time-invariant functions as $t \to \infty$. Since the residential load is a time-varying function, the optimal control law must be time varying, which means that the optimal Q-function is also time varying. In addition, in previous TBQL algorithms, properties of the algorithms, such as the convergence property, are not analyzed. In this case, the optimal control scheme cannot be guaranteed by the TBQL algorithms as $t \to \infty$, which limits the applications of the TBQL algorithms to a great extent. Hence, it is necessary to develop a new iterative Q-learning algorithm and establish corresponding property analysis, which therefore motivates our research.

In this paper, a new dual iterative Q-learning algorithm is developed to obtain the optimal management scheme for residential energy systems. First, the detailed iteration procedure of the dual iterative Q-learning algorithm is presented. Two iterations are introduced, which are external (i-iteration in brief) and internal (j-iteration in brief) iterations. The objective of i-iteration is to achieve the optimal Q-function, and the objective of j-iteration is to obtain the iterative control law sequence that minimizes the total cost in each period. Second, the convergence property of the dual iterative Q-learning algorithm is proven for the first time to guarantee that the iterative Q-function can reach the optimum under the iterative control laws. Furthermore, in order to facilitate the implementation of the dual iterative Q-learning algorithm, neural networks are employed to implement the developed algorithm to obtain the iterative Q-function and the iterative control laws. Finally,

Manuscript received January 30, 2014; revised May 28, 2014 and July 26, 2014; accepted August 31, 2014. Date of publication October 3, 2014; date of current version March 6, 2015. This work was supported in part by the National Natural Science Foundation of China under Grant 61034002, Grant 61374105, Grant 61233001, and Grant 61273140 and in part by the Beijing Natural Science Foundation under Grant 4132078.

The authors are with The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: qinglai.wei@ia.ac.cn; derong.liu@ia.ac.cn; shiguang2012@ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2014.2361485

numerical results and comparisons are given to show the effectiveness of the developed algorithm.

II. Problem Formulation

Here, smart residential energy systems with batteries will be described. The optimization objective of our research will be defined, and the corresponding principle of optimality will be introduced.

A. Notation

The list of notations used is reported as follows.

t, k	Time indexes.
i, j	Iteration indexes.
E_{bt}	Battery energy (kWh).
$E_b^{\min}$	Minimum storage energy of the battery (kWh).
$E_b^{\max}$	Maximum storage energy of the battery (kWh).
$\eta(\cdot)$	Charging/discharging efficiency of battery.
P_{bt}	Battery power output (kW).
$P_b^{\min}$	Minimum charging/discharging power of battery (kW).
$P_b^{\max}$	Maximum charging/discharging power of battery (kW).
P_{rate}	Rated power output of battery (kW).
P_{Lt}	Power of the residential load (kW).
P_{gt}	Power from the power grid (kW).
C_t	Electricity rate (cents/kWh).
E_b^o	Middle of storage limit (kWh).
m_1, m_2, r	Given positive constants in performance index function.
γ	Discount factor.
x_t	System state.
u_t	Control input.
$F(\cdot)$	System function.
$U(\cdot)$	Utility function.
$J(\cdot)$	Performance index function.
$Q(\cdot)$	Q-function.
$\arg\min$	Argument of the minimum.
$\min$	Minimum of the function.
λ	Period of residential load and electricity rate.
W_a	Hidden-output weight matrix of action network.
Y_a	Input-hidden weight matrix of action network.
W_c	Hidden-output weight matrix of critic network.
Y_c	Input-hidden weight matrix of critic network.

B. Smart Residential Energy Systems

The smart residential energy system described in [24] is composed of the power grid; the residential load; the battery system, which is located at the side of residential load (including a battery and a sinewave inverter); and the power management unit (controller). The schematic of the smart residential energy system can be described in Fig. 1.

The battery model used in this work is based on [24], [28], and [29], where the battery model is expressed as

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

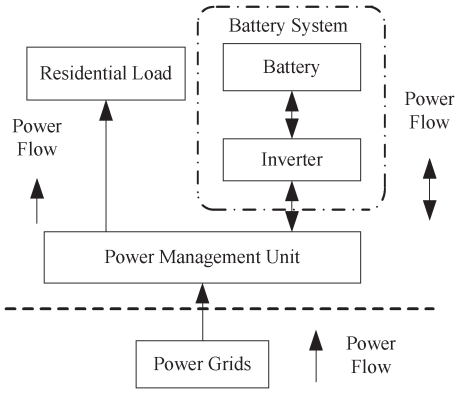

Fig. 1. Smart residential energy system.

Let $P_{bt} > 0$ denote battery discharging. Let $P_{bt} < 0$ denote battery charging, and let $P_{bt} = 0$ denote that the battery is idle. Let the efficiency of battery charging/discharging be derived as $\eta(P_{bt}) = 0.898 - 0.173|P_{bt}|/P_{\text{rate}}$, where we define $P_{\text{rate}} > 0$.

Remark 1: In this paper, the optimal battery control problem is treated as a discrete-time problem with the time step of 1 h, and it is assumed that the residential load varies hourly. Then, the battery power output P_{bt} satisfies $P_{bt}(\text{kW}) \times 1(\text{h}) = P_{bt}(\text{kWh})$, which is equal to the value of energy. Hence, the expression of the battery model in (1) is feasible.

C. Optimization Objectives and Optimality Principles

In this paper, the power flow from the battery to the grid is not permitted, i.e., we define $P_{gt} \geq 0$, to guarantee the power quality of the grid. For convenience of analysis, we introduce delays in P_{bt} and P_{Lt}, and then, we can define the load balance as $P_{L(t-1)} = P_{b(t-1)} + P_{gt}$. The total performance index function expected to be minimized is defined as

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

where $0 < \gamma < 1$, and $E_b^o = 1/2(E_b^{\min} + E_b^{\max})$. The physical meaning of the first term of the performance index function is to minimize the total cost from the grid. The second term aims to make the stored energy of the battery close to the middle of storage limit, which avoids full charging/discharging of the battery. The third term is to prevent large charging/discharging power of the battery. Hence, the second and third terms aim to extend the lifetime of the battery. Let $x_{1t} = P_{gt}$ and $x_{2t} = E_{bt} - E_b^o$. Letting $u_t = P_{bt}$ and $x_t = [x_{1t}, x_{2t}]^T$, the equation of the residential energy system can be written as

$$x_{t+1} = F(x_t, u_t, t) = \begin{pmatrix} P_{Lt} - u_t \\ x_{2t} - u_t \eta(u_t) \end{pmatrix}. \tag{3}$$

Let $\underline{u}_t = (u_t, u_{t+1}, \ldots)$ denote the control sequence from t to ∞. Let $M_t = \begin{bmatrix} m_1 C_t^2 & 0 \\ 0 & m_2 \end{bmatrix}$. Let x_0 be the initial states. Then, the performance index function (2) can be written as $J(x_0, \underline{u}_0, 0) = \sum_{t=0}^{\infty} \gamma^t U(x_t, u_t, t)$, where $U(x_t, u_t, t) = x_t^T M_t x_t + r u_t^2$. Generally, functions of the residential load and the real-time electricity rate are periodic. For convenience of analysis, our discussion is based on the following assumption.

Assumption 1: The residential load P_{Lt} and the electricity rate C_t are periodic functions with the period $\lambda = 24$ h.

Define the control sequence set as $\underline{\mathfrak{U}}_t = \{\underline{u}_t : \underline{u}_t = (u_t, u_{t+1}, \ldots), \forall\, u_{t+i} \in \mathbb{R}^m, i = 0, 1, \ldots\}$. Then, the optimal performance index function can be defined as $J^*(x_t, t) = \min_{\underline{u}_t}\{J(x_t, \underline{u}_t, t) : \underline{u}_t \in \underline{\mathfrak{U}}_t\}$. Define the optimal Q-function as $Q^*(x_t, u_t, t)$, which satisfies $\min_{u_t} Q^*(x_t, u_t, t) = J^*(x_t, t)$. Hence, the Q-function is also called the action-dependent performance index function. According to [19] and [20], the optimal Q-function satisfies the following equation:

$$Q^*(x_t, u_t, t) = U(x_t, u_t, t) + \gamma \min_{u_{t+1}} Q^*(x_{t+1}, u_{t+1}, t+1). \quad (4)$$

Remark 2: According to Bellman's principle [30], $J^*(x_t, t)$ satisfies the following HJB equation:

$$J^*(x_t, t) = \min_{u_t}\{U(x_t, u_t, t) + J^*(F(x_t, u_t, t), t+1)\}. \quad (5)$$

From (5), we can see that the J-function, i.e., $J^*(x_t, t)$, only describes the quality of the states. If we desire to find an optimal control by the J-function, then the mathematical expressions of the residential energy system model and the utility function are both explicitly required. From (4), we can see that the Q-function also depends on the control input, and the control can be directly obtained by minimizing the Q-function [19], [20]. Because of these merits, Q-functions are preferred throughout this paper, and a novel iterative Q-learning algorithm will be developed.

Remark 3: It should be pointed out that if the residential environment, such as the battery model, is changed, the optimal control will change correspondingly. The control results under different elements of the battery will be discussed in Section V.

III. Dual Iterative Q-Learning Algorithm of ADP

Here, a new dual iterative Q-learning algorithm is developed to obtain the optimal control law for residential energy systems. A novel convergence analysis method will also be developed in this section.

A. Derivation of the Dual Iterative Q-Learning Algorithm

From (4), we can see that the optimal Q-function $Q^*(x_t, u_t, t)$ is a time-varying function, which means that for different time t, the optimal Q-function is different. This makes it difficult to obtain $Q^*(x_t, u_t, t)$. According to Assumption 1, for $\forall t = 0, 1, \ldots$, there exist $\varrho = 0, 1, \ldots$ and $\theta = 0, 1, \ldots, 23$ that satisfies $t = \varrho\lambda + \theta$. Let $k = \varrho\lambda$. Then, we have $P_{Lt} = P_{L(k+\theta)} = P_{L\theta}$ and $C_t = C_{k+\theta} = C_\theta$, respectively. Define $\mathcal{U}_k$ as the control sequence from k to $k + \lambda - 1$, i.e., $\mathcal{U}_k = (u_k, u_{k+1}, \ldots, u_{k+\lambda-1})$. Then, for $\forall k \in \{0, \lambda, 2\lambda, \ldots\}$, we can define a new utility function as

$$\Pi(x_k, \mathcal{U}_k) = \sum_{\theta=0}^{\lambda-1} \gamma^\theta U(x_{k+\theta}, u_{k+\theta}, \theta). \quad (6)$$

Then, (4) can be expressed as

$$Q^*(x_k, \mathcal{U}_k) = \Pi(x_k, \mathcal{U}_k) + \tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} Q^*(x_{k+\lambda}, \mathcal{U}_{k+\lambda}) \quad (7)$$

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

$$\mathcal{U}^*(x_k) = \arg\min_{\mathcal{U}_k}\{Q^*(x_k, \mathcal{U}_k)\}. \quad (8)$$

Based on the preparations above, the new dual iterative Q-learning algorithm of ADP can be developed. In the developed algorithm, two iterations are introduced, which are external iteration (i-iteration in brief) and internal iteration (j-iteration in brief), respectively. Let $i = 0, 1, \ldots$ be the external iteration index. Let $\Psi(x_k, u_k)$ be an arbitrary positive semidefinite function. Define the initial Q-function $Q_0(x_k, \mathcal{U}_k)$ as

$$Q_0(x_k, \mathcal{U}_k) = \Pi(x_k, \mathcal{U}_k) + \tilde{\gamma} \min_{u_{k+\lambda}} \Psi(x_{k+\lambda}, u_{k+\lambda}). \quad (9)$$

The iterative control law sequence $\mathcal{U}_0$ can be computed as

$$\mathcal{U}_0(x_k) = \arg\min_{\mathcal{U}_k} Q_0(x_k, \mathcal{U}_k). \quad (10)$$

The iterative Q-function can be updated as

$$Q_1(x_k, \mathcal{U}_k) = \Pi(x_k, \mathcal{U}_k) + \tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} Q_0(x_{k+\lambda}, \mathcal{U}_{k+\lambda}). \quad (11)$$

For $i = 1, 2, \ldots$, i-iteration will proceed between

$$\mathcal{U}_i(x_k) = \arg\min_{\mathcal{U}_k} Q_i(x_k, \mathcal{U}_k) \quad (12)$$

and

$$\begin{aligned} Q_{i+1}(x_k, \mathcal{U}_k) &= \Pi(x_k, \mathcal{U}_k) + \tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} Q_i(x_{k+\lambda}, \mathcal{U}_{k+\lambda}) \\ &= \Pi(x_k, \mathcal{U}_k) + \tilde{\gamma} Q_i(x_{k+\lambda}, \mathcal{U}_i(x_{k+\lambda})). \end{aligned} \quad (13)$$

Remark 4: The objective of i-iteration is to update the iterative performance index function to achieve the optimum. The idea of i-iteration (9)–(13) is inspired by the value iteration in [31], while there exist essential differences. First, for the value iteration in [31], the initial performance index function is required to be zero. In the developed Q iterative algorithm, $\Psi(x_k, u_k)$ can be an arbitrary positive semidefinite function. Second, for the value iteration in [31], for $\forall\, i = 0, 1, \ldots$, a single iterative control law is required to update the iterative performance index function, whereas for the developed Q iterative algorithm, an iterative control law sequence $\mathcal{U}_i(x_k)$ is needed, which means the iterative control law sequence cannot be directly obtained by solving (10) and (12). Hence, j-iteration is necessary to proceed.

Let $j = 0, 1, \ldots, 23$ be the internal iteration index. For $i = 0$ and $j = 0$, let the initial iterative performance index be

$$Q_0^0(x_k, u_k) = \Psi(x_k, u_k). \quad (14)$$

For $i = 0$ and $j = 0, 1, \ldots, 23$, j-iteration will proceed between

$$u_0^j(x_k) = \arg\min_{u_k} Q_0^j(x_k, u_k) \quad (15)$$

and

$$\begin{aligned} Q_0^{j+1}(x_k, u_k) &= U(x_k, u_k, j) + \gamma \min_{u_{k+1}} Q_0^j(x_{k+1}, u_{k+1}) \\ &= U(x_k, u_k, j) + \gamma Q_0^j\left(x_{k+1}, u_0^j(x_{k+1})\right) \end{aligned} \quad (16)$$

where we let $x_{k+1}=\begin{pmatrix}P_{L(\lambda-1-j)}-u_k\\ x_{2k}-u_k\eta(u_k)\end{pmatrix}$. Let $U(x_k,u_k,j)=x_k^TM_{\lambda-1-j}x_k+ru_k^2$, where $M_{\lambda-1-j}=\begin{bmatrix}m_1C_{\lambda-1-j}^2 & 0\\ 0 & m_2\end{bmatrix}$. For $\forall\, i=1,2,\ldots$, we let $Q_i^0(x_k,u_k)=Q_{i-1}^{24}(x_k,u_k)$. For $j=0,1,\ldots,23$, j-iteration will proceed between

$$u_i^j(x_k)=\arg\min_{u_k}Q_i^j(x_k,u_k) \tag{17}$$

and

$$\begin{aligned}Q_i^{j+1}(x_k,u_k)&=U(x_k,u_k,j)+\gamma\min_{u_{k+1}}Q_i^j(x_{k+1},u_{k+1})\\&=U(x_k,u_k,j)+\gamma Q_i^j\left(x_{k+1},u_i^j(x_{k+1})\right).\end{aligned} \tag{18}$$

Then, for $\forall\, i=0,1,\ldots$, we can obtain the iterative control law sequence by

$$\mathcal{U}_i(x_k)=\left\{u_i^0(x_k),u_i^1(x_k),\ldots,u_i^{23}(x_k)\right\}. \tag{19}$$

Remark 5: The objective of j-iteration is to obtain the iterative control law sequence that minimizes the total cost in each period. From (17), we can see that for different j values, the iterative control law $u_i^j(x_k)$ is different. For $\forall\, i$, j-iteration (14)–(19) proceeds for a finite number of iterations, whereas i-iteration (10)–(13) proceeds for an infinite number of iterations. The optimal Q-function and iterative control law are desired to achieve according to i-iteration and j-iteration, and the developed algorithm is therefore called "dual iterative Q-learning algorithm."

B. Properties of the Dual Iterative Q-Learning Algorithm

Here, the convergence property of the dual iterative Q-learning algorithm will be investigated. First, we will show that iterative control law sequence $\mathcal{U}_i(x_k)$ obtained by j-iteration can minimize the total cost in each period.

Theorem 1: For $i=0,1,\ldots$ and $j=0,1,\ldots,23$, let the iterative Q-functions $Q_i(x_k,\mathcal{U}_k)$ and $Q_i^j(x_k,u_k)$ be obtained by (9)–(19). Then, we have

$$\min_{\mathcal{U}_k}Q_i(x_k,\mathcal{U}_k)=\min_{u_k}Q_i^{24}(x_k,u_k). \tag{20}$$

Proof: The statement can be proven by mathematical induction. First, for $i=0$, we have

$$\begin{aligned}&\min_{u_k}Q_0^{j+1}(x_k,u_k)\\&=\min_{u_k}\left(U(x_k,u_k,j)+\gamma\min_{u_{k+1}}Q_0^j(x_{k+1},u_{k+1})\right)\\&=\min_{u_k}\Big(U(x_k,u_k,j)+\gamma\min_{u_{k+1}}(U(x_{k+1},u_{k+1},j-1)+\cdots\\&\qquad+\gamma\min_{u_{k+j}}(U(x_{k+j},u_{k+j},0)\\&\qquad+\gamma\min_{u_{k+j+1}}\Psi(x_{k+j+1},u_{k+j+1}))\Big)\\&=\min_{(u_k,u_{k+1},\ldots,u_{k+j})}\left(\sum_{l=0}^{j}\gamma^lU(x_{k+l},u_{k+l},j-l)\right.\\&\qquad\left.+\gamma^{j+1}\min_{u_{k+j+1}}\Psi(x_{k+j+1},u_{k+j+1})\right).\end{aligned} \tag{21}$$

Let $j=23$. According to (6) and (9), we have

$$\begin{aligned}\min_{u_k}Q_0^{24}(x_k,u_k)&=\min_{\mathcal{U}_k}\left(\Pi(x_k,\mathcal{U}_k)+\tilde{\gamma}\min_{u_{k+\lambda}}\Psi(x_{k+\lambda},u_{k+\lambda})\right)\\&=\min_{\mathcal{U}_k}Q_0(x_k,\mathcal{U}_k).\end{aligned} \tag{22}$$

The conclusion holds for $i=0$. Assume that the conclusion holds for $i=\tau-1$, i.e., $\min_{\mathcal{U}_k}Q_{\tau-1}(x_k,\mathcal{U}_k)=\min_{u_k}Q_{\tau-1}^{24}(x_k,u_k)$. Then, for $i=\tau$, we have

$$\begin{aligned}&\min_{u_k}Q_\tau^{24}(x_k,u_k)\\&=\min_{u_k}\left(U(x_k,u_k,23)+\gamma\min_{u_{k+1}}Q_\tau^{23}(x_{k+1},u_{k+1})\right)\\&=\min_{u_k}\Big(U(x_k,u_k,23)+\gamma\min_{u_{k+1}}\Big(U(x_{k+1},u_{k+1},22)+\cdots\\&\quad+\gamma\min_{u_{k+23}}\Big(U(x_{k+23},u_{k+23},0)+\gamma\min_{u_{k+\lambda}}Q_\tau^0(x_{k+\lambda},u_{k+\lambda})\Big)\Big)\Big)\\&=\min_{\mathcal{U}_k}\left(\Pi(x_k,\mathcal{U}_k)+\tilde{\gamma}\min_{u_{k+\lambda}}Q_{\tau-1}^{24}(x_{k+\lambda},u_{k+\lambda})\right)\\&=\min_{\mathcal{U}_k}\left(\Pi(x_k,\mathcal{U}_k)+\tilde{\gamma}\min_{\mathcal{U}_{k+\lambda}}Q_{\tau-1}(x_{k+\lambda},\mathcal{U}_{k+\lambda})\right)\\&=\min_{\mathcal{U}_k}Q_\tau(x_k,\mathcal{U}_k).\end{aligned} \tag{23}$$

The mathematical induction is completed. ■

From Theorem 1, we can obtain the following corollary.

Corollary 1: Let $\mu(x_k)$ be an arbitrary control law. For $i=0,1,\ldots$ and $j=0,1,\ldots,23$, define a new performance index function as $\mathcal{Q}_i^{j+1}(x_k,u_k)=U(x_k,u_k,j)+\gamma\mathcal{Q}_i^j(x_{k+1},\mu(x_{k+1}))$ and define $Q_i^{j+1}(x_k,u_k)$ as in (18). For $\forall\, i=0,1,\ldots$, let $\mathcal{Q}_i^0(x_k,u_k)=Q_i^0(x_k,u_k)$. Then, for $\forall\, j=0,1,\ldots,23$, we have $Q_i^j(x_k,u_k)\le\mathcal{Q}_i^j(x_k,u_k)$.

From Theorem 1 and Corollary 1, for $\forall\, i=0,1,\ldots$, we can say that the total cost in each period can be minimized by the iterative control law sequence $\mathcal{U}_i(x_k)$ according to j-iteration (14)–(19). Next, the convergence property of i-iteration will be developed.

Theorem 2: For $i=0,1,\ldots$, let $Q_{i+1}(x_k,\mathcal{U}_k)$ and $\mathcal{U}_i(x_k)$ be obtained by i-iteration (10)–(13). Then, the iterative Q-function $Q_i(x_k,\mathcal{U}_k)$ converges to its optimum, i.e.,

$$\lim_{i\to\infty}Q_i(x_k,\mathcal{U}_k)=Q^*(x_k,\mathcal{U}_k). \tag{24}$$

Proof: For functions $Q^*(x_k,\mathcal{U}_k)$, $\Pi(x_k,\mathcal{U}_k)$, and $Q_0(x_k,\mathcal{U}_k)$, inspired by [32], let $\underline{\varsigma}$, $\overline{\varsigma}$, $\underline{\delta}$, and $\overline{\delta}$ be constants that satisfy

$$\underline{\varsigma}\Pi(x_k,\mathcal{U}_k)\le\tilde{\gamma}\min_{\mathcal{U}_{k+\lambda}}Q^*(x_{k+\lambda},\mathcal{U}_{k+\lambda})\le\overline{\varsigma}\Pi(x_k,\mathcal{U}_k) \tag{25}$$

and

$$\underline{\delta}Q^*(x_k,\mathcal{U}_k) \le Q_0(x_k,\mathcal{U}_k) \le \overline{\delta}Q^*(x_k,\mathcal{U}_k) \tag{26}$$

respectively, where $0 < \underline{\varsigma} \le \overline{\varsigma} < \infty$ and $0 \le \underline{\delta} \le \overline{\delta} < \infty$. Since $Q^*(x_k,\mathcal{U}_k)$ is unknown, the values of $\underline{\varsigma}$, $\overline{\varsigma}$, $\underline{\delta}$, and $\overline{\delta}$ cannot be directly obtained. In the following, we will prove that for arbitrary constants $\underline{\varsigma}$, $\overline{\varsigma}$, $\underline{\delta}$, and $\overline{\delta}$, the iterative Q-function $Q_i(x_k,\mathcal{U}_k)$ will converge to the optimum, and the estimations for the values of these constants can be omitted. The proof proceeds in four steps. First, we show that if $0 \le \underline{\delta} \le \overline{\delta} < 1$, then for $\forall\, i = 0, 1, \ldots$, the iterative performance index function $Q_i(x_k,\mathcal{U}_k)$ satisfies

$$\left(1 + \frac{\underline{\delta} - 1}{(1+\overline{\varsigma}^{-1})^i}\right) Q^*(x_k,\mathcal{U}_k) \le Q_i(x_k,\mathcal{U}_k) \le \left(1 + \frac{\overline{\delta} - 1}{(1+\underline{\varsigma}^{-1})^i}\right) Q^*(x_k,\mathcal{U}_k). \tag{27}$$

Inequality (27) can be proven by mathematical induction. Let $i = 0$, and we have

$$\begin{aligned}
&Q_1(x_k,\mathcal{U}_k)\\
&= \Pi(x_k,\mathcal{U}_k) + \tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} \{Q_0(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\\
&\ge \Pi(x_k,\mathcal{U}_k) + \underline{\delta}\tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} \{Q^*(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\\
&\ge \left(1 + \overline{\varsigma}\frac{\underline{\delta} - 1}{1+\overline{\varsigma}}\right) \Pi(x_k,\mathcal{U}_k)\\
&\quad + \tilde{\gamma}\left(\underline{\delta} - \frac{\underline{\delta} - 1}{1+\overline{\varsigma}}\right) \min_{\mathcal{U}_{k+\lambda}} \{Q^*(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\\
&= \left(1 + \frac{\overline{\varsigma}(\underline{\delta}-1)}{(1+\overline{\varsigma})}\right)\left\{\Pi(x_k,\mathcal{U}_k) + \tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} \{Q^*(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\right\}\\
&= \left(1 + \frac{\underline{\delta} - 1}{(1+\overline{\varsigma}^{-1})}\right) Q^*(x_k,\mathcal{U}_k).
\end{aligned} \tag{28}$$

From the idea of (28), we can also get $Q_1(x_k,\mathcal{U}_k) \le (1 + (\overline{\delta} - 1/(1+\underline{\varsigma}^{-1}))Q^*(x_k,\mathcal{U}_k)$. Thus, (27) holds for $i = 0$. Assume that (27) holds for $i = l-1$, $l = 1, 2, \ldots$. Then, for $i = l$, we have

$$\begin{aligned}
&Q_{l+1}(x_k,\mathcal{U}_k)\\
&= \Pi(x_k,\mathcal{U}_k) + \tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} \{Q_l(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\\
&\ge \Pi(x_k,\mathcal{U}_k) + \tilde{\gamma}\left(1 + \frac{\overline{\varsigma}^{l-1}(\underline{\delta}-1)}{(1+\overline{\varsigma})^{l-1}}\right) \min_{\mathcal{U}_{k+\lambda}} \{Q^*(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\\
&\ge \left(1 + \frac{\overline{\varsigma}^{l}(\underline{\delta}-1)}{(1+\overline{\varsigma})^{l}}\right)\left\{\Pi(x_k,\mathcal{U}_k) + \tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} \{Q^*(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\right\}\\
&= \left(1 + \frac{\underline{\delta} - 1}{(1+\overline{\varsigma}^{-1})^{l}}\right) Q^*(x_k,\mathcal{U}_k).
\end{aligned} \tag{29}$$

From the idea of (29), we can also get $Q_{l+1}(x_k,\mathcal{U}_k) \le (1 + ((\overline{\delta} - 1)/(1+\underline{\varsigma}^{-1})^l)Q^*(x_k,\mathcal{U}_k)$. Hence, we obtain that (27) holds for $\forall i = 0, 1, \ldots$ The mathematical induction is completed. Second, we show that if $0 \le \underline{\delta} \le 1 \le \overline{\delta} < \infty$, then the iterative Q-function $Q_i(x_k,\mathcal{U}_k)$ satisfies

$$\left(1 + \frac{\underline{\delta} - 1}{(1+\overline{\varsigma}^{-1})^i}\right) Q^*(x_k,\mathcal{U}_k) \le Q_i(x_k,\mathcal{U}_k) \le \left(1 + \frac{\overline{\delta} - 1}{(1+\overline{\varsigma}^{-1})^i}\right) Q^*(x_k,\mathcal{U}_k). \tag{30}$$

The left-hand side of (30) can be proven according to (28) and (29). For the right-hand side of (30), letting $i = 0$, we have

$$\begin{aligned}
&Q_1(x_k,\mathcal{U}_k)\\
&= \Pi(x_k,\mathcal{U}_k) + \tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} \{Q_0(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\\
&\le \Pi(x_k,\mathcal{U}_k) + \overline{\delta}\tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} \{Q^*(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\\
&\quad + \frac{\overline{\delta} - 1}{(1+\overline{\varsigma})}\left(\overline{\varsigma}\Pi(x_k,\mathcal{U}_k) - \tilde{\gamma} \min_{\mathcal{U}_{k+\lambda}} \{Q^*(x_{k+\lambda},\mathcal{U}_{k+\lambda})\}\right)\\
&\le \left(1 + \frac{\overline{\delta} - 1}{(1+\overline{\varsigma}^{-1})}\right) Q^*(x_k,\mathcal{U}_k).
\end{aligned} \tag{31}$$

According to mathematical induction, we can obtain the right-hand side of (30). Third, for the situation $1 \le \underline{\delta} \le \overline{\delta} < \infty$, according to (28) and (29), we can prove that for $\forall\, i = 0, 1, \ldots$, the iterative performance index function $Q_i(x_k,\mathcal{U}_k)$ satisfies (27). Finally, considering the three situations above, for arbitrary constants $\underline{\varsigma}$, $\overline{\varsigma}$, $\underline{\delta}$, and $\overline{\delta}$, according to (27) and (30), we can easily obtain (24), as $i \to \infty$. ■

Corollary 2: For $i = 0, 1, \ldots$, let $Q_{i+1}(x_k,\mathcal{U}_k)$ and $\mathcal{U}_i(x_k)$ be obtained by i-iteration (10)–(13). 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 6: One important property should be mentioned. We say that the optimal control of the battery does not provide global optimal management of the whole smart home grids. The global optimal control can be obtained only if the system as a whole is discussed. In this paper, given the residential load and the electricity rate, the optimal battery control law for the residential energy system (3) is achieved to minimize the given performance index function in (2) under the assumptions of periodic residential load and electricity rate, i.e., Assumption 1. Hence, the approach presented in this paper in fact deoptimizes distributed system operation.

IV. Neural Network Implementation for the Dual Iterative Q-Learning Algorithm

Here, neural networks are introduced to implement the dual iterative Q-learning algorithm. There are two neural networks, which are critic and action networks, respectively, in the dual iterative Q-learning algorithm. Both neural networks are chosen as three-layer backpropagation networks. The whole structure diagram is shown in Fig. 2.

A. Action Network

For $\forall\, i = 0, 1, \ldots$, the role of the action network is to approximate the iterative control law sequence $\mathcal{U}_i(x_k)$ defined in (12).

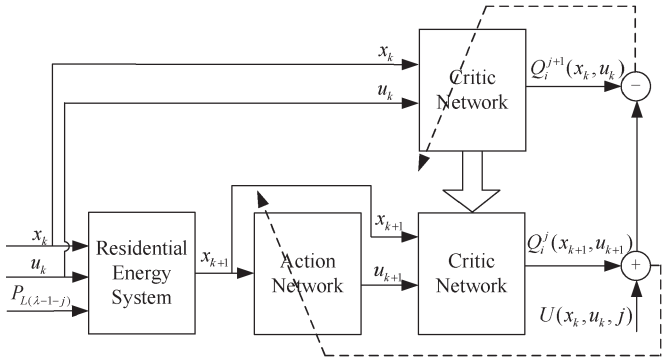

Fig. 2. Structure diagram of the dual iterative Q-learning algorithm.

The target of the action network can be defined as (15) and (17). The action network can be constructed by two input neurons, ten sigmoidal hidden neurons, and one linear output neuron. Let $l = 0, 1, \ldots$ be the training step. The output of the action network can be expressed as $\hat{u}_i^{j,l}(x_k) = W_{ai}^{jT}(l)\sigma(\mathcal{Z}_a(x_k))$, where $\mathcal{Z}_a(x_k) = Y_a^T x_k$, and $\sigma(\cdot)$ is a sigmoid function [25]. To enhance the training speed, only the hidden-output weight $W_{ai}^j(l)$ is updated during the neural network training, whereas the input-hidden weight is fixed [33]. According to [25], the action network weight update is expressed as follows:

$$W_{ai}^j(l+1) = W_{ai}^j(l) - \beta_a \left[\frac{\partial E_{ai}^j(l)}{\partial W_{ai}^j(l)}\right] \tag{32}$$

where $E_{ai}^j(l) = (1/2)(e_{ai}^j(l))^2$, $e_{ai}^j(l) = \hat{u}_i^{j,l}(x_k) - u_i^j(x_k)$, and $\beta_a > 0$ is the learning rate of the action network.

B. Critic Network

For $\forall\, i = 0, 1, \ldots$ and $j = 0, 1, \ldots, 23$, the goal of the critic network is to obtain $Q_i(x_k, \mathcal{U}_k)$ by updating $Q_i^{j+1}(x_k, u_k)$ in (18), iteratively. The critic network can be constructed by three input neurons, 15 sigmoidal hidden neurons, and one linear output neuron. Let $Z_{ck} = [x_k^T, u_k]^T$ be the input vector of the critic network. Then, the output of the critic network can be expressed as $\hat{Q}_i^{j+1,l}(x_k, u_k) = W_{ci}^{jT}(l)\sigma(\mathcal{Z}_{ck})$, where $\mathcal{Z}_{ck} = Y_c^T Z_{ck}$, and $\sigma(\cdot)$ is a sigmoid function [25]. During the neural network training, the hidden-output weight $W_{ci}^j(l)$ is updated, whereas the input-hidden weight Y_c is fixed. According to [25], the critic network weight update is expressed as follows:

$$W_{ci}^j(l+1) = W_{ci}^j(l) - \alpha_c \left[\frac{\partial E_{ci}^j(l)}{\partial W_{ci}^j(l)}\right] \tag{33}$$

where $E_{ci}^j(l) = (1/2)(e_{ci}^j(l))^2$, $e_{ci}^j(l) = \hat{Q}_i^{j+1,l}(x_k, u_k) - Q_i^{j+1}(x_k, u_k)$, and $\alpha_c > 0$ is the learning rate of the critic network.

C. Training Phase

Here, the dual iterative Q-learning algorithm implemented by action and critic networks is explained step by step and shown in Algorithm 1.

Algorithm 1 Dual iterative Q-learning algorithm.

Initialization:

1: Collect an array of system data for the residential energy system (3).
2: Give a positive semidefinite function $\Psi(x_k, u_k)$.
3: Give the computation precision $\varepsilon > 0$.

Iteration:

4: Let $i = 0$. For $j = 0$, let $Q_0^0(x_k, u_k) = \Psi(x_k, u_k)$.
5: For $j = 0, 1, \ldots, 23$, train the action and critic networks to obtain $u_0^j(x_k)$ and $Q_0^{j+1}(x_k, u_k)$ that satisfy (15) and (16), respectively.
6: Let $Q_0(x_k, \mathcal{U}_k) = Q_0^{24}(x_k, u_k)$. Obtain $\mathcal{U}_0(x_k)$ and $Q_1(x_k, \mathcal{U}_k)$ by (10) and (11), respectively.
7: Let $i = i + 1$.
8: For $j = 0, 1, \ldots, 23$, train the action and critic networks to obtain $u_i^j(x_k)$ and $Q_i^{j+1}(x_k, u_k)$ that satisfy (17) and (18), respectively.
9: Let $Q_i(x_k, \mathcal{U}_k) = Q_i^{24}(x_k, u_k)$. Obtain $\mathcal{U}_i(x_k)$ and $Q_{i+1}(x_k, \mathcal{U}_k)$ by (12) and (13), respectively.
10: If $|Q_i(x_k, \mathcal{U}_k) - Q_{i-1}(x_k, \mathcal{U}_k)| \le \varepsilon$, then goto next step. Otherwise, goto Step 7.
11: For $j = 0, 1, \ldots, 23$, solve $u_i^j(x_k)$ by (17) and obtain $\mathcal{U}_i(x_k) = (u_i^0(x_k), \ldots, u_i^{23}(x_k))$.
12: **return** $Q_i(x_k, \mathcal{U}_k)$ and $\mathcal{U}_i(x_k)$.

V. Numerical Analysis

Here, the performance of the dual iterative Q-learning algorithm will be examined by numerical experiments. Comparisons will also be given to show the superiority of the developed algorithm. The profiles of the residential load demand and the real-time electricity rate are taken from [24], [26], and [27], where the residential load demand and the real-time electricity rate for one week (168 h) are shown in Fig. 3(a) and (c), respectively. We can see that the residential load demand and the real-time electricity rate are both periodic-like functions with the period $\lambda = 24$. The average trajectories of the residential load demand and the electricity rate are shown in Fig. 3(b) and (d). In this paper, we use average residential load demand and average electricity rate as the periodic residential load demand and electricity rate.

We assume that the supply from the power grid guarantees the residential load demand at any time. Define the capacity of the battery as 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. The rated power output of the battery and the maximum charging/discharging rate is 16 kW. The initial level of the battery is 60 kWh. Let the performance index function be expressed as in (2), where we set $m_1 = 1$, $m_2 = 0.2$, $r = 0.1$, and $\gamma = 0.995$. Let the initial function $\Psi(x_k, u_k) = [x_k^T, u_k] I [x_k^T, u_k]^T$, where I is the identity matrix with a suitable dimension. Let the initial state be $x_0 = [8, 60]^T$. After normalizing the data of the residential load demand and the electricity rate [25], [34], we implement the developed dual iterative Q-learning algorithm by neural networks for $i = 20$

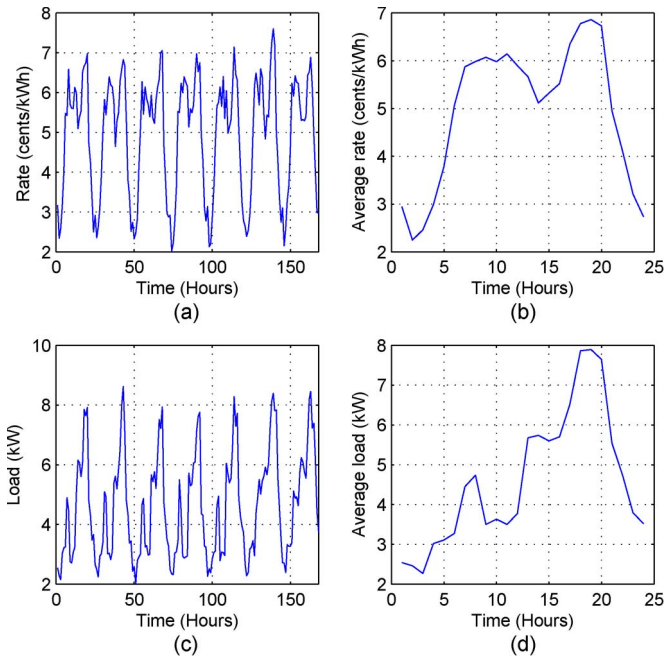

Fig. 3. Residential load demand and electricity rate. (a) Residential load demand for 168 h. (b) Average residential load demand. (c) Real-time electricity rate for 168 h. (d) Average electricity rate.

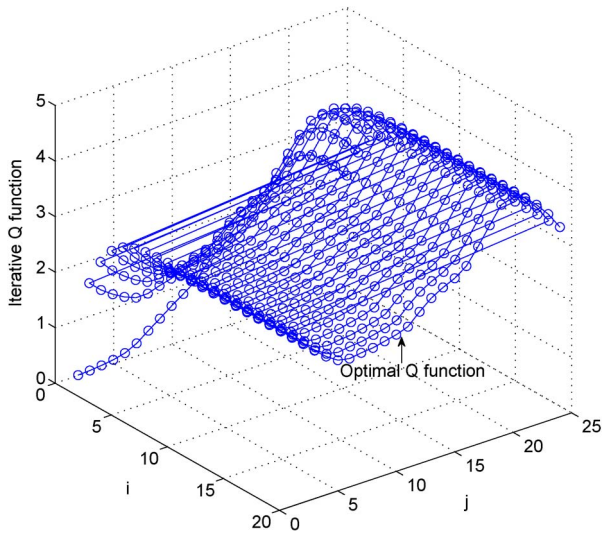

Fig. 4. Trajectory of the iterative Q-function.

iterations to guarantee the computation precision $\varepsilon = 10^{-4}$. The learning rates of the action and critic networks are 0.01, and the training precision of the neural networks is 10^{-6}. Let $Q_i^j(x_0, \bar{u}) = \min_u Q_i^j(x_0, u)$. The trajectory of $Q_i^j(x_0, \bar{u})$ is shown in Fig. 4. After $i = 20$ iterations, for $\forall\, j = 0, 1, \ldots, 23$, we can get $Q_{i+1}^j(x_0, \bar{u}) = Q_i^j(x_0, \bar{u})$, which means that the iterative Q-function is convergent to the optimum. According to one week's residential load demand and electricity rate, the optimal control of the battery is shown in Fig. 5.

In the following, the TBQL algorithm [24] and the particle swarm optimization (PSO) algorithm [27] will be compared to illustrate the superiority of the developed dual iterative Q-learning algorithm. For $\forall\, t = 0, 1, \ldots$, the goal of the TBQL algorithm [24], [25] is to design an iterative control that

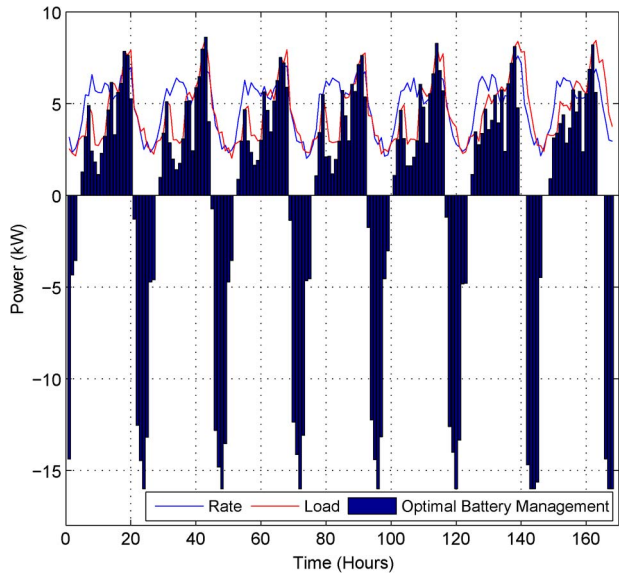

Fig. 5. Optimal control of the battery in one week.

satisfies the following optimality equation: $Q(x_{t-1}, u_{t-1}, t-1) = U(x_t, u_t, t) + \gamma Q(x_t, u_t, t)$. Let the initial function and the structures of the action and critic networks, which implement the TBQL algorithm, be the same as those in our example. For the PSO algorithm [27], let $\mathcal{G} = 30$ be the swarm size. The position of each particle at time t is represented by $x_{\ell t}$, $\ell = 1, 2, \ldots, \mathcal{G}$ and its movement by the velocity vector $v_{\ell t}$. Then, the update rule of PSO can be expressed as

$$x_{\ell t} = x_{\ell(t-1)} + \nu_{\ell t}$$
$$\nu_{\ell t} = \omega \nu_{\ell(t-1)} + \varphi_1 \rho_1^T \left(p_\ell - x_{\ell(t-1)}\right) + \varphi_2 \rho_2^T \left(p_g - x_{\ell(t-1)}\right).$$

Let the inertia factor be $\omega = 0.7$. Let the correction factors $\rho_1 = \rho_2 = [1, 1]^T$. Let φ_1 and φ_2 be random numbers in [0,1]. Let p_ℓ be the best position of particles, and let p_g be the global best position. Implement the TBQL algorithm for 100 time steps and the PSO algorithm for 100 iterations. Let the real-time cost function be $R_{ct} = C_t P_{gt}$, and the corresponding real-time cost functions are shown in Fig. 6(a), where the term "original" denotes "no battery system." The comparison of the total cost for 168 h is displayed in Table I. From Table I, the superiority of our dual iterative Q-learning algorithm can be verified. The trajectories of the battery energy by dual iterative Q-learning and TBQL algorithms are shown in Fig. 6(b). We can see that using the TBQL algorithm, the battery is fully charged each day, whereas the battery level is more reasonable by the dual iterative Q-learning algorithm.

In the above optimizations, we give more importance to the electricity rate than the cost of the battery system, i.e., m_1 in the performance index function is large. On the other hand, the discharging rate and depth are also important for the battery system to be kept "alive" as long as possible. Hence, we enlarge parameters m_2 and r in the performance index function. Let $m_2 = 1$, $r = 1$, and let m_1 be unchanged. The iterative Q-function is shown in Fig. 7. The optimal battery control can be seen in Fig. 8, and the battery energy under the new performance index function can be seen in Fig. 9(a). Enlarging m_2 and r, we can see that the value of the iterative Q-function is

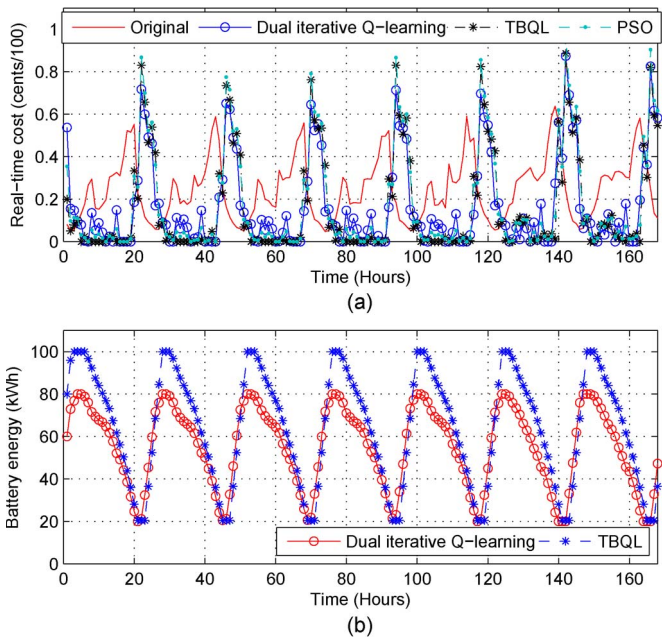

Fig. 6. Numerical comparisons. (a) Real-time cost comparison among dual iterative Q-learning, TBQL, and PSO algorithms. (b) Battery energy comparison between dual iterative Q-learning and TBQL algorithms.

TABLE I
COST COMPARISON

	Original	PSO	TBQL	Dual iterative Q-learning
Total cost (cents)	4124.13	3029.96	2866.64	2797.86
Saving		26.53%	30.49%	32.16%

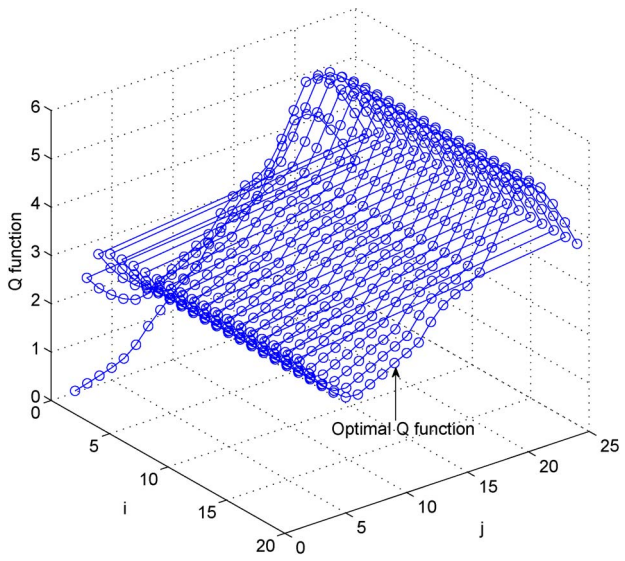

Fig. 7. Trajectory of the iterative Q-function.

enhanced. The battery output power is reduced, and the battery energy is closer to E^o, which extends the lifetime of the battery. However, the total cost of one week is 2955.35 cents, which means the cost saving is reduced.

On the other hand, the battery model is important to the optimal control law of the battery. To illustrate the effectiveness of the developed algorithm, different elements of the battery will be considered. For convenience of analysis, we let $m_1 = 1$, $m_2 = 0.2$, $r = 0.1$. First, let the efficiency of

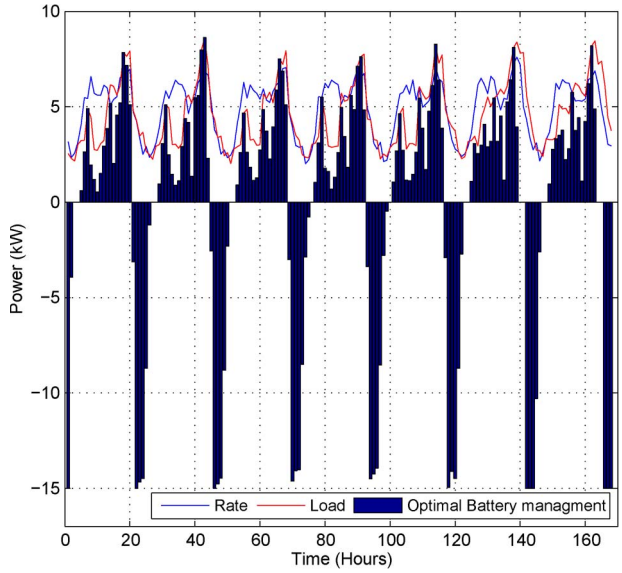

Fig. 8. Optimal control of the battery under $m_1 = 1$, $m_2 = 1$, and $r = 1$.

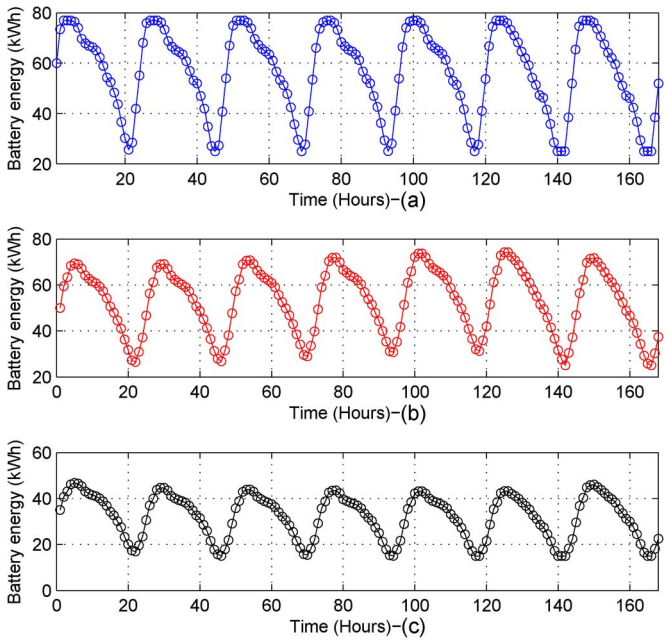

Fig. 9. Batteries' energy. (a) New performance index function with $m_1 = 1$, $m_2 = 1$, and $r = 1$. (b) Battery I. (c) Battery II.

battery charging/discharging be reduced as $\eta(P_{bt}) = 0.698 - 0.173|P_{bt}|/P_{\text{rate}}$, and let the capacity of the battery be 80 kWh. Define the battery as Battery I. Implementing the developed dual iterative Q-learning algorithm with Battery I, the trajectory of $Q_i^j(x_0, \bar{u})$ is shown in Fig. 10. We can see that the iterative Q-function is also convergent to the optimum after $i = 20$ iterations, and the values of the Q-functions are larger than those in Fig. 4, which indicates that the optimization ability decreases. The optimal control trajectory for Battery I is shown in Fig. 11. The battery energy of Battery I is shown in Fig. 9(b), and the total cost of one week is 2914.70 cents.

Next, we keep on reducing the performance of the battery. Let the capacity of the battery decrease to 60 kWh. Let the rated power output of the battery and the maximum charging/

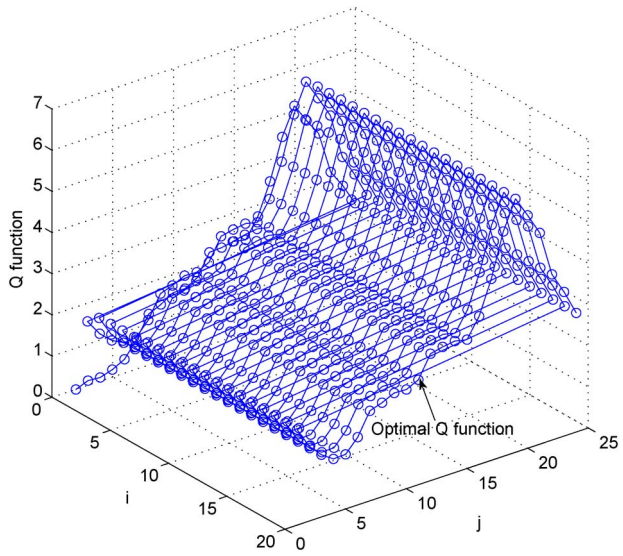

Fig. 10. Trajectory of the iterative Q-function.

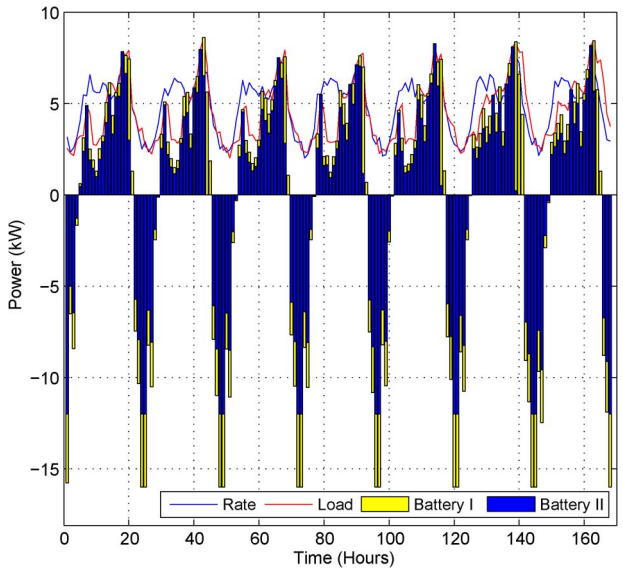

Fig. 11. Optimal control of the battery in one week.

discharging rate be 12 kW. Define the battery as Battery II. The optimal control trajectory for Battery II is shown in Fig. 11. The battery energy of Battery II is shown in Fig. 9(c), and the total cost of one week is 3027.17 cents.

From the numerical results, we can see that for different battery models, the developed dual iterative Q-learning algorithm will make the iterative performance index function converge to the optimum and obtain the optimal battery control law. We can also see that as the performance of the battery decreases, the optimization ability of the battery also decreases.

VI. Concluding Remarks and Future Work

Given the residential load and the real-time electricity rate, the objective of the optimal control in this paper 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. The main idea of the developed dual iterative Q-learning algorithm is to update the iterative performance index function and iterative control laws by the ADP technique according to i-iteration and j-iteration, respectively. For the first time, the convergence and optimality of the developed algorithm are developed. Neural networks are introduced to implement the developed dual iterative Q-learning algorithm. Finally, the effectiveness of the developed algorithm is justified by numerical results.

As is known, renewable sources, such as solar and wind energies, are important elements to reduce the total cost of residential energy systems and extend the life of the battery. As renewable sources possess more uncertainties, how to deal with the uncertainties is a key problem to implementing our Q-learning algorithm to renewable-source-based residential energy systems, and it is also our future research topic.

References

[1] F. D. Angelis *et al.*, "Optimal home energy management under dynamic electrical and thermal constraints," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1518–1527, Aug. 2013.

[2] L. Jian *et al.*, "Regulated charging of plug-in hybrid electric vehicles for minimizing load variance in household smart microgrid," *IEEE Trans. Ind. Electron.*, vol. 60, no. 8, pp. 3218–3226, Aug. 2013.

[3] X. Lu, K. Sun, J. M. Guerrero, J. C. Vasquez, and L. Huang, "State-of-charge balance using adaptive droop control for distributed energy storage systems in DC microgrid applications," *IEEE Trans. Ind. Electron.*, vol. 61, no. 6, pp. 2804–2815, Jun. 2014.

[4] J. M. Guerrero, P. C. Loh, T. L. Lee, and M. Chandorkar, "Advanced control architectures for intelligent microgrids—Part II: Power quality, energy storage, AC/DC microgrids," *IEEE Trans. Ind. Electron.*, vol. 60, no. 4, pp. 1263–1270, Apr. 2013.

[5] Z. Amjadi and S. S. Williamson, "Power-electronics-based solutions for plug-in hybrid electric vehicle energy storage and management systems," *IEEE Trans. Ind. Electron.*, vol. 57, no. 2, pp. 608–616, Feb. 2010.

[6] H. Rahimi-Eichi, F. Baronti, and M. Y. Chow, "Online adaptive parameter identification and state-of-charge coestimation for lithium-polymer battery cells," *IEEE Trans. Ind. Electron.*, vol. 61, no. 4, pp. 2053–2061, Apr. 2014.

[7] S. N. Motapon, L. A. Dessaint, and K. Al-Haddad, "A comparative study of energy management schemes for a fuel-cell hybrid emergency power system of more-electric aircraft," *IEEE Trans. Ind. Electron.*, vol. 61, no. 3, pp. 1320–1334, Mar. 2014.

[8] A. Chaouachi, R. M. Kamel, R. Andoulsi, and K. Nagasaka, "Multiobjective intelligent energy management for a microgrid," *IEEE Trans. Ind. Electron.*, vol. 60, no. 4, pp. 1688–1699, Apr. 2013.

[9] R. Smolenski, J. Bojarski, A. Kempski, and P. Lezynski, "Time-domain-based assessment of data transmission error probability in smart grids with electromagnetic interference," *IEEE Trans. Ind. Electron.*, vol. 61, no. 4, pp. 1882–1890, Apr. 2014.

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

[11] P. J. Werbos, "A menu of designs for reinforcement learning over time," in *Neural Networks for Control*, W. T. Miller, R. S. Sutton, and P. J. Werbos, Eds. Cambridge, MA, USA: MIT Press, 1991, pp. 67–95.

[12] A. Heydari and S. N. Balakrishnan, "Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 24, no. 1, pp. 145–157, Jan. 2013.

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

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

[15] Q. Wei and D. Liu, "A novel iterative θ-adaptive dynamic programming for discrete-time nonlinear systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 4, pp. 1176–1190, Sep. 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 Trans. Ind. Electron.*, vol. 61, no. 11, pp. 6399–6408, Nov. 2014.

[17] Q. Wei and D. Liu, "Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 14, pp. 1020–1036, Nov. 2013.
[18] H. Zhang, F. L. Lewis, and Z. Qu, "Lyapunov, adaptive, optimal design techniques for cooperative systems on directed communication graphs," *IEEE Trans. Ind. Electron.*, vol. 59, no. 7, pp. 3026–3041, Jul. 2012.
[19] C. Watkins, "Learning from Delayed Rewards," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 1989.
[20] C. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3/4, pp. 279–292, May 1992.
[21] 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.
[22] Y. Liang, L. He, X. Cao, and Z. J. Shen, "Stochastic control for smart grid users with flexible demand," *IEEE Trans. Smart Grid*, vol. 4, no. 4, pp. 2296–2308, Apr. 2013.
[23] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Syst. Technol.*, vol. 32, no. 6, pp. 76–105, Dec. 2012.
[24] T. Huang and D. Liu, "A self-learning scheme for residential energy system control and management," *Neural Comput. Appl.*, vol. 22, no. 2, pp. 259–269, Feb. 2013.
[25] J. Si and Y.-T. Wang, "On-line learning control by association and reinforcement," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 264–276, Mar. 2001.
[26] M. Boaro *et al.*, "Adaptive dynamic programming algorithm for renewable energy scheduling and battery management," *Cognitive Comput.*, vol. 5, no. 2, pp. 264–277, Jun. 2013.
[27] D. Fuselli *et al.*, "Action dependent heuristic dynamic programming for home energy resource scheduling," *Int. J. Elect. Power Energy Syst.*, vol. 48, pp. 148–160, Jun. 2013.
[28] 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, Mar. 2007.
[29] T. Yau, L. N. Walker, H. L. Graham, and R. Raithel, "Effects of battery storage devices on power system dispatch," *IEEE Trans. Power App. Syst.*, vol. PAS-100, no. 1, pp. 375–383, Jan. 1981.
[30] R. E. Bellman, *Dynamic Programming*. Princeton, NJ, USA: Princeton Univ. Press, 1957.
[31] 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. Part B, Cybern.*, vol. 38, no. 4, pp. 943–949, Aug. 2008.
[32] B. Lincoln and A. Rantzer, "Relaxing dynamic programming," *IEEE Trans. Autom. Control*, vol. 51, no. 8, pp. 1249–1260, Aug. 2006.
[33] T. Dierks and S. Jagannathan, "Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update," *IEEE Trans. Neural Netw. Learning Syst.*, vol. 23, no. 7, pp. 1118–1129, Apr. 2012.
[34] S. Aksoy and R. M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognit. Lett.*, vol. 22, no. 5, pp. 563–582, Apr. 2001.

Qinglai Wei (M'11) received the B.S. degree in automation, the M.S. degree in control theory and control engineering, and the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2002, 2005, and 2008, respectively.

From 2009 to 2011, he was a Postdoctoral Fellow with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, where he is currently an Associate Professor with The State Key Laboratory of Management and Control for Complex Systems. His research interests include neural-network-based control, adaptive dynamic programming, optimal control, and nonlinear systems and their industrial applications.

Derong Liu (S'91–M'94–SM'96–F'05) received the Ph.D. degree in electrical engineering from the University of Notre Dame, Notre Dame, IN, USA, in 1994.

From 1993 to 1995, he was a Staff Fellow with the General Motors Research and Development Center, Warren, MI, USA. From 1995 to 1999, he was an Assistant Professor with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA. In 1999, he joined the University of Illinois at Chicago, Chicago, IL, USA, where he became a Full Professor of electrical and computer engineering and of computer science in 2006. He is the author of 14 books (six research monographs and eight edited volumes).

Dr. Liu currently serves as the Editor-in-Chief of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. He received the Faculty Early Career Development Award from the National Science Foundation in 1999, the University Scholar Award from the University of Illinois in 2006, and the Overseas Outstanding Young Scholar Award from the National Natural Science Foundation of China in 2008. He was selected for the "100 Talents Program" by the Chinese Academy of Sciences in 2008.

Guang Shi received the B.S. degree in automation from Zhejiang University, Hangzhou, China, in 2012. He is currently working toward the Ph.D. degree in The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

His research interests include neural networks, adaptive dynamic programming, optimal control, and energy management in smart grids.