A Multi-agent Reinforcement Learning based Data-driven Method for Home Energy Management

Xu Xu, *Student Member, IEEE,* Youwei Jia, *Member, IEEE,* Yan Xu, *Senior Member, IEEE,* Zhao Xu, *Senior Member, IEEE,* Songjian Chai, *Member, IEEE,* Chun Sing Lai, *Member, IEEE*

***Abstract*—This paper proposes a novel framework for home energy management (HEM) based on reinforcement learning in achieving efficient home-based demand response (DR). The concerned hour-ahead energy consumption scheduling problem is duly formulated as a finite Markov decision process (FMDP) with discrete time steps. To tackle this problem, a data-driven method based on neural network (NN) and *Q*-learning algorithm is developed, which achieves superior performance on cost-effective schedules for HEM system. Specifically, real data of electricity price and solar photovoltaic (PV) generation are timely processed for uncertainty prediction by extreme learning machine (ELM) in the rolling time windows. The scheduling decisions of the household appliances and electric vehicles (EVs) can be subsequently obtained through the newly developed framework, of which the objective is dual, i.e. to minimize the electricity bill as well as the DR induced dissatisfaction. Simulations are performed on a residential house level with multiple home appliances, an EV and several PV panels. The test results demonstrate the effectiveness of the proposed data-driven based HEM framework.**

***Index Terms*—Reinforcement learning, data-driven method, home energy management, finite Markov decision process, neural network, *Q*-learning algorithm, demand response**

NOMENCLATURE

i/Ω^{NS}	Index/set of non-shiftable appliances
j/Ω^{PS}	Index/set of power-shiftable appliances
m/Ω^{TS}	Index/set of time-shiftable appliances
n/Ω^{EV}	Index/set of EVs
t/T	Index of time slot
λ_t^G	Electricity price in time slot t
$P_{it}^{d,NS}/P_{jt}^{d,PS}/P_{mt}^{d,TS}$	Energy consumption of non-shiftable appliance i /power-shiftable appliance j/time-shiftable appliance m in time slot t
$P_{nt}^{d,EV}$	Energy consumption of EV n in time slot t
E_t^{PV}/E_t^{PVs}	Solar panel output/surplus solar energy in time slot t
$P_{j,max}^{d,PS}/P_{n,max}^{d,EV}$	Upper bound of Energy consumption for power-shiftable appliance j/EV n

I. INTRODUCTION

WITH recent advances in communication technologies and smart metering infrastructures, users can schedule their real-time energy consumption via the home energy management (HEM) system. Such actions of energy consumption scheduling are also referred to as demand response (DR), which balances supply and demand by adjusting elastic loads [1], [2].

Many research efforts have been paid on studying HEM system from the demand side perspective. In Ref. [3], a hierarchical energy management system is proposed for home microgrids with consideration of photovoltaic (PV) energy integration in day-ahead and real-time stages. Ref. [4] studies a novel HEM system in finding optimal operation schedules of home energy resources, aiming to minimize daily electricity cost and monthly peak energy consumption penalty. Authors in Ref. [5] propose a stochastic programming based dynamic energy management framework for the smart home with plug-in electric vehicle storage. The work presented in Ref. [6] proposes a new smart HEM system in terms of the quality of experience, which depends on the information of consumer's discontent for changing operations of home appliances. In Ref. [7], for the smart home equipped with heating, ventilation and air condition, Yu *et al.* investigate the issue of minimizing electricity bill and thermal discomfort cost simultaneously from the perspective of a long-term time horizon. Ref. [8] proposes a multi-time and multi-energy building energy management system, which is modeled as a non-linear quadratic programming problem. Ref. [9] utilizes an approximate dynamic programming method to develop a computationally efficient HEM system where temporal difference learning is adopted for scheduling distributed energy resources. The study reported in Ref. [10] introduces a new approach for HEM system to solve a DR problem which is formulated using the chance-constrained programming optimization, combining the particle swarm optimization method and the two-point estimation method. Till now, most studies related to HEM system adopt centralized optimization approaches. Generally, due to the assumption of the accurate uncertainty prediction, the optimization method is able to show the perfect performance. However, this assumption is not very reasonable in reality since the optimization model knows all environment information meanwhile removing all prediction errors. Besides, due to a large number of binary or integer variables involved, some of these methods may suffer from expensive computational cost.

As an emerging type of machine learning, reinforcement learning (RL) [11] shows excellent decision-making capability

X. Xu, and Y. Jia are with the Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China.

Y. Xu is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Z. Xu and S. J. Chai are with the Department of Electrical Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong.

C. S. Lai is with the School of Civil Engineering, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds LS2 9JT, UK and also with Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou 510006, China

in the absence of initial environment information. The deployment of RL in decision-makings has considerable merits. Firstly, RL seeks the optimal actions by interacting with the environment so it has no requirement for initial knowledge, which may be difficult to acquire in practice. Secondly, RL can be flexibly employed to different application objects by off-line training and on-line implementation, considering relative uncertainties autonomously. Thirdly, RL is easier to implement in real-life scenarios as compared with conventional optimization methods. The reason is that RL can obtain the optimal results in a look-up table, so its computational efficiency is fairly high. In recent literature, the RL has received growing interests for solving energy management problems. Vázquez-Canteli, et al. comprehensively summarizes the algorithms and modeling techniques for reinforcement learning for demand response. Interested readers can further refer to Ref. [12]. Ref. [13] proposes a batch RL based approach for residential DR of thermostatically controlled loads with predicted exogenous data. In Ref. [14], Wan *et al.* use deep RL algorithms to determine the optimal solutions of the EV charging/discharging scheduling problem. In Ref. [15], RL is adopted to develop a dynamic pricing DR method based on hierarchical decision-making framework in the electricity market, which considers both profits of the service provider and costs of customers. Ref. [16] proposes an hour-ahead DR algorithm to make optimal decisions for different home appliances. Ref. [17] proposes a residential energy management method considering peer-to-peer trading mechanism, where the model-free decision-making process is enhanced by the fuzzy Q-learning algorithm. In Ref. [18], a model-free DR approach for industrial facilities is presented based on the actor-critic-based deep reinforcement learning algorithm. Based on RL, Ref. [19] proposes a multi-agent based distributed energy management method for distributed energy resources in a microgrid energy market. Ref. [20] focuses on the deep RL based on-line optimization of schedules for the building energy management system. In Ref. [21], the deep neural network and the model-free reinforcement learning is utilized to manage the energy in a multi-microgrid system. Ref. [22] presents a novel RL based model for residential load scheduling and load commitment with uncertain renewable sources. In consideration of the state-of-the-art HEM methods in this field, there are still two significant limitations. Firstly, most HEM studies focus only on one category of loads, such as home appliance loads or electric vehicle (EV) loads, ignoring the coordinated decision-makings for diverse loads. This weakly reflects the operational reality. Secondly, the integration of renewables, especially solar PV generation, is rarely considered during the decision-making process. With the rapid growth of rooftop installation of residential PV panels [23], allocation of self-generated solar energy should be considered when scheduling residential energy consumption.

To address the above issues, this paper proposes a novel multi-agent reinforcement learning based data-driven HEM method. The hour-ahead home energy consumption scheduling problem is formulated as a finite Markov decision process (FMDP) with discrete time steps. The bi-objective of the formulated problem is to minimize the electricity bill as well as DR induced dissatisfaction cost. The main contributions of this paper are threefold.

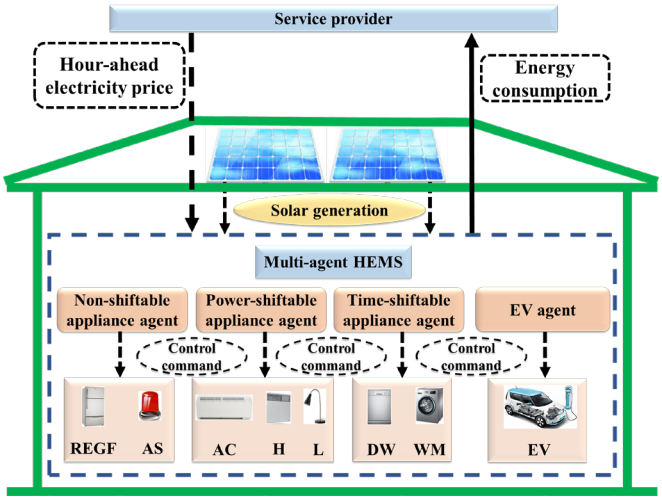


Fig. 1. Structure of our proposed HEM system (REFG: refrigerator; AS: alarm system; AC: air conditioner; H: heating; L: lighting; WM: washing machine; DW: dishwashing; EV: electric vehicle).

1) Under the data-driven framework, we propose a novel model-free and adaptable HEM method based on extreme learning machine (ELM) and Q-learning algorithm. To our best knowledge, such method is rarely investigated before. The test results show that the proposed HEM method can not only achieve promising performance in terms of reducing electricity cost for householders but also improve the computational efficiency.
2) The conventional HEM methods are based on optimization algorithms with the assumption of perfect uncertainty prediction. However, this assumption is infeasible and unreasonable since the prediction errors are unavoidable. By contrast, our proposed model-free data-driven based HEM method can overcome the future uncertainties by the ELM based NN and discover the optimal DR decisions by the learning capability of the Q-learning algorithm.
3) In confronting with different types of loads in a residential house (e.g. non-shiftable loads, power-shiftable loads, time-shiftable loads and EV charging loads), a multi-agent Q-learning algorithm based RL method is developed to tackle the HEM problem involved with multiple loads. In this way, optimal energy consumption scheduling decisions for various home appliances and EV charging can be obtained in a fully decentralized manner.

The remainder of this paper is organized as follows. Section II models the home energy consumption scheduling problem as a FMDP. Then our proposed solution approach is presented in Section III. In Section IV, test results are given to demonstrate the effectiveness of our proposed methodology. Finally, Section V concludes the paper.

II. Problem Modelling

As illustrated in Fig. 1, this paper considers four agents in a HEM system, which correspond to non-shiftable appliance load, power-shiftable appliance load, time-shiftable appliance

TABLE I
STATE SET, ACTION SET AND REWARD FUNCTION OF EACH AGENT

Item ID	State set	Action set	Reward function
REFG	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	1 - "on"	Eq. (1)
AS	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	1 - "on"	Eq. (1)
AC1	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	{0.7, 0.8, …, 1.4} - "power ratings"	Eq. (2)
AC2	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	{0.7, 0.8, …, 1.4} - "power ratings"	Eq. (2)
H	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	{0.5, 0.6, …, 1.5} - "power ratings"	Eq. (2)
L1	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	{0.2, 0.3, …, 0.6} - "power ratings"	Eq. (2)
L2	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	{0.2, 0.3, …, 0.6} - "power ratings"	Eq. (2)
WM	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	0 - "off" 1 - "on"	Eq. (3)
DW	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	0 - "off" 1 - "on"	Eq. (3)
EV	$\{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$	{0, 3, 6} - "charging rates"	Eq. (4)

REFG: refrigerator; AS: alarm system; AC: air conditioner; H: heating; L: light; WM: wash machine; DW: dishwashing; EV: electric vehicle

load, and EV load, respectively. In this paper, we envision the proposed HEM system includes multiple agents, which are virtual to control different kinds of smart home appliances in a decentralized manner. It should note that smart meters are assumed to be installed on smart home appliances to monitor the devices and receive the control command given by the agents. In each time slot, we determine the hour-ahead energy consumption actions for home appliances and EVs. Specifically, in time slot t, the agent observes the state s_t and chooses the action a_t. After taking this action, the agent observes the new state s_{t+1} and chose a new action a_{t+1} for the next time slot $t+1$. This hour-ahead energy consumption scheduling problem can be formulated as a FMDP, where the outcomes are partly controlled by the decision-maker and partly random. The FMDP of our problem contains five tuples, i.e., $(\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{R}(\cdot,\cdot), \gamma, \theta)$, where $\boldsymbol{S}$ denotes the state set, $\boldsymbol{A}$ denotes the finite action set, $\boldsymbol{R}(\cdot,\cdot)$ denotes the reward set, γ denotes the discount rate and θ denotes the learning rate. The details about the FMDP formulation are described as follows.

A. State

The state s_t can describe the current situation in the FMDP. In this paper, the state s_t in time slot t can be defined as a vector, defined as,

$$s_t = \{(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G), (E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})\}$$

where s_t consists of two types of information,

1) $(\lambda_t^G, \lambda_{t+1}^G, \dots, \lambda_T^G)$ indicate the current electricity price λ_t in and predicted future electricity prices from the next time slot $t+1$ to the end time slot T. In each time slot, the hour-ahead electricity price can be informed by a service provider.

2) $(E_t^{PV}, E_{t+1}^{PV}, \dots, E_T^{PV})$ indicate the current solar panel output E_t^{PV} and predicted future solar panel outputs from the next time slot $t+1$ to the end time slot T. In this paper, we assume that the householder owns a residential solar system, including the solar panels and the inverter systems [24], operating at the maximum power point (MPP) [25]. In each intraday time slot, the householder needs to allocate the generated solar energy to the home appliances sequentially. Consider that non-shiftable appliances always consume fixed energy and play an important role in ensuring the convenience and safety of the living environment, so these appliances are served first. Also, the comfort level of the living environment should be taken into account, so the surplus self-generated solar energy E_t^{PVS} is delivered according to the descending order of the dissatisfaction coefficients of the remaining home appliances. Finally, the surplus solar energy will be curtailed or sold to the utility grid with wholesale market clearing price.

B. Action

In this study, the action denotes the energy consumption scheduling of each home appliance as well as EV battery charging, described as follows.

1) Action set for non-shiftable appliance agent: Non-shiftable appliances, e.g., refrigerator and alarm system, require high reliability to ensure daily-life convenience and safety, so their demands must be satisfied and cannot be scheduled. Therefore, only one action, i.e., "on", can be taken by the non-shiftable appliance agent.

2) Action set for power-shiftable appliance agent: Power-shiftable appliances, such as air conditioner, heating and light, can operate flexibly by consuming energy in a predefined range. Hence, power-shiftable agents can choose discrete actions, i.e., 1,2,3, …, which indicate the power ratings at different levels.

3) Action set for time-shiftable appliance agent: The time-shiftable loads can be scheduled from peak periods to off-peak periods to reduce the electricity cost and avoid peak energy

usage. Time-shiftable appliances, including wash machine and dishwasher, have two operating points, "on" and "off".

4) Action set for EV agent: As the EV user, the householder would like to reduce electricity cost by scheduling EV battery charging. It should be noted that in this paper, EV battery discharging is not considered since it can significantly shorten the useful lifetime of EV battery [26]. As suggested by Ref. [27], the EV charger can provide discrete charging rates.

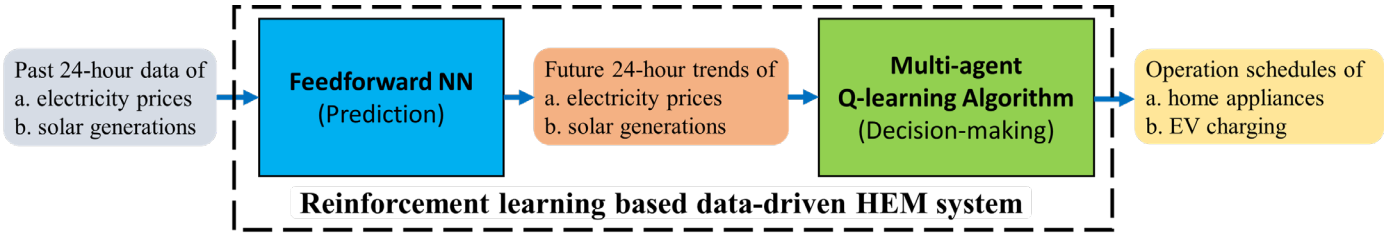


Fig. 2. Schematic of the reinforcement learning based data-driven HEM system.

C. Reward

The reward represents the inverse utility cost of each agent, described as follows.

1) The reward of non-shiftable appliance agent

$$r_{it}^{NS} = -\lambda_t^G [P_{it}^{d,NS} - E_{it}^{PVs}]^+ \quad i \in \Omega^{NS}, t = \{1,2,...,T\} \tag{1}$$

The reward of non-shiftable appliance agent only concerns on electricity cost since the non-shiftable loads are immutable. Note that $[\cdot]^+$ represents the projection operator onto the non-negative orthant, i.e., $[x]^+ = max(x,0)$.

2) The reward of power-shiftable appliance agent

$$r_{jt}^{PS} = -\lambda_t^G [P_{jt}^{d,PS} - E_{jt}^{PVs}]^+ - \alpha_j^{PS} (P_{j,\max}^{d,PS} - P_{jt}^{d,PS})^2$$
$$j \in \Omega^{PS}, t = \{1,2,...,T\} \tag{2}$$

where the first term denotes the electricity cost and the second term is the dissatisfaction cost caused by reducing power ratings of power-shiftable appliances. This dissatisfaction cost is defined by a quadratic function [28] with an appliance-dependent coefficient α_j^{PS}, which can be adjusted to achieve a trade-off between the electricity cost and the satisfaction level.

3) The reward of time-shiftable appliance agent

$$r_{mt}^{TS} = -\lambda_t^G [u_{mt} P_{mt}^{d,TS} - E_{mt}^{PVs}]^+ - \alpha_m^{TS} (t_m^s - t_m^{ini})^2$$
$$m \in \Omega^{TS}, t = [t_m^{ini}, t_m^{end}] \tag{3}$$

where u_{mt} is the binary variable representing the operating point of the time-shiftable appliance m in time slot t, i.e., $u_{mt} = 1$ (on) or $u_{mt} = 0$ (off). When the time-shiftable loads are scheduled, dissatisfaction cost of householder would be raised due to the waiting time for them to start. Therefore, electricity bill (first term) and dissatisfaction cost (second term) should be taken into account simultaneously for operating time-shiftable appliances. α_m^{TS} is the dissatisfaction coefficient describing the tolerance of waiting time for the appliance m and it is determined by personal dependence on devices. Thus, a higher α_m^{TS} means that the waiting for appliance m start is more likely to cause dissatisfaction. Note that time-shiftable appliance m should start to operate during its normal working period $[t_m^{ini}, t_m^{end}]$.

4) The reward of EV agent

$$r_{nt}^{EV} = -\lambda_t^G P_{nt}^{d,EV} - \alpha_n^{EV} (P_{n,\max}^{d,EV} - P_{nt}^{d,EV})^2$$
$$n \in \Omega^{EV}, t \in [t_n^{arr}, t_n^{dep}] \tag{4}$$

where the first two terms of (4) describe that the EV owner needs to pay the electricity cost ($\lambda_t^G P_{nt}^{d,EV}$) during the period $[t_n^{arr}, t_n^{dep}]$. Besides, the second term of (4) represents the cost of "charging anxiety" with the dissatisfaction coefficient α_n^{EV}, which describes the fear that the EV has insufficient energy to get its destination without underfilled EV battery.

D. Total Reward of HEM System

After giving the rewards (see Eqs. (1)-(4)) of all agents in the proposed HEM system, the total reward R can be acquired, described as follows,

$$R = -\sum_{t \in T} \left\{ \begin{array}{l} \lambda_t^G \left(\begin{array}{l} [P_{it}^{d,NS} - E_{it}^{PVs}]^+ - [P_{jt}^{d,PS} - E_{jt}^{PVs}]^+ \\ -[u_{mt} P_{mt}^{d,TS} - E_{mt}^{PVs}]^+ - P_{nt}^{d,EV} \end{array} \right) \\ - \left(\begin{array}{l} \alpha_j^{PS} (P_{j,\max}^{d,PS} - P_{jt}^{d,PS})^2 - \alpha_m^{TS} (t_m^s - t_m^{ini})^2 \\ -\alpha_n^{EV} (P_{n,\max}^{d,EV} - P_{nt}^{d,EV})^2 \end{array} \right) \end{array} \right\} \tag{5}$$

E. Action-value Function

The quality of the action a_t under the state s_t, i.e., energy consumption scheduling in time slot t, can be evaluated by the the expected sum of future rewards for the horizon of K time steps, given as follows,

$$Q_\pi(s,a) = \mathrm{E}_\pi \left[\sum_{k=0}^{K} \gamma^k \cdot r_{t+1} \mid s_t = s, a_t = a \right] \tag{6}$$

where $Q_\pi(s,a)$ represents the action-value function and π is the policy mapping from a system state to an energy consumption schedule. $\gamma \in [0,1]$ is the discount rate denoting the relative importance of future rewards for the current reward. When $\gamma = 0$, the agent seems to be shortsighted since it only cares about the current reward, while $\gamma = 1$ indicates that the agent is foresighted and it considers future rewards. To balance the trade-off between current reward and future reward, setting a fraction in the range $[0,1]$ for γ is suggested.

The objective of the energy consumption scheduling problem is to find the optimal policy π^*, i.e., a sequence of optimal operating actions for each home appliance and EV battery, to maximize the action-value function.

III. Proposed Data-driven based Solution Method

In this paper, the proposed reinforcement learning based data-driven method is comprised of two parts (see Fig. 2), (i) an ELM based feedforward NN is trained for predicting the future trends of electricity price and PV generation, (ii) a multi-agent Q-learning algorithm based RL method is developed for making hour-ahead energy consumption decisions. Details of this data-driven based solution method are given in the following subsections.

A. ELM based Feedforward NN for Uncertainty Prediction

As a well-studied training algorithm, ELM algorithm has become a popular topic in the fields of load forecasting [29], electricity price forecasting [30] and renewable generation forecasting [31]. Since the input weights and biases of the hidden layer are randomly assigned and free to be tuned further when using ELM algorithm, some exceptional features can be obtained, e.g., fast learning speed and good generalization. To deal with the uncertainties of electricity prices and solar generations, we propose an ELM based feedforward NN to dynamically predict future trends of these two uncertainties. Specifically, at each hour, the inputs of the trained feedforward NN are past 24-hour electricity price data and solar generation data, and its outputs are the forecasted future 24-hour trends of electricity prices and solar generations. This predicted information will be fed into the decision-making process of energy consumption scheduling, as described in the following subsection.

B. Multi-agent Q-learning Algorithm for Decision-making

After acquiring the predicted future electricity prices and solar panel outputs, we employ the Q-learning algorithm to use this information to find the optimal policy π^*. As an emerging machine learning algorithm, Q-learning algorithm is widely used for the decision-making process to gain the maximum cumulative rewards [32]. The basic mechanism of this algorithm is to construct a Q-table where Q-value $Q(s_t, a_t)$ of each state-action pair is updated in each iteration until the convergence condition is satisfied. In this way, the optimal action with optimal Q-value in each state can be selected. The optimal Q-value $Q_\pi^*(s_t, a_t)$ can be obtained by using Bellman equation [33], given as below,

$$Q_\pi^*(s_t, a_t) = r(s_t, a_t) + \gamma * \max Q(s_{t+1}, a_{t+1}) \tag{7}$$

The Q-value can be updated in terms of reward, learning rate and discount factor, described as follows,

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \theta \begin{bmatrix} r(s_t, a_t) \\ +\gamma * \max(Q(s_{t+1}, a_{t+1})) \\ -Q(s_t, a_t) \end{bmatrix} \tag{8}$$

where $\theta \in [0,1]$ denotes the learning rate indicating to what extent the new Q-value can overturn the old one. When $\theta = 0$, the agent exploits the prior information exclusively, whereas $\theta = 1$ indicates that the agent considers only the current estimate and overlooks the prior information. A value of a decimal between 0 and 1 should be applied to θ, trading off the new Q-value and old Q-value.

C. Implementation Process of Proposed Solution Method

Algorithm 1 demonstrates the implementation process of our proposed solution approach for solving the FMDP problem as described in Section II. Specifically, in the initial time slot, i.e., $t = 1$, the HEM system initializes power rating, dissatisfaction coefficient, discount rate, and learning rate. In each time slot, the trained DFM is used to forecast future 24-hour electricity prices as well as solar panel outputs, as shown in Algorithm 2. Upon obtaining the predicted information, the multi-agent

Algorithm 1 Proposed Data-driven based Solution Method

1. Initialize power rating, using time, dissatisfaction coefficient α, discount factor γ and learning rate θ
2. **For** time slot $t = 1{:}T$ **do**
3. | **For** HEM system **do**
4. | | Execute **Algorithm 2**
5. | **End for**
6. | **Receive** extracted information about future electricity prices and solar generations
7. | **For** each agent **do** ▷ Sort descending by α
8. | | Execute **Algorithm 3**
9. | **End for**
10. **End for**

Algorithm 2 Feedforward NN (Features Extraction)

1. Update the input electricity price data $\{\lambda_{t-23}^G, \dots, \lambda_t^G\}$ and solar generation data $\{E_{t-23}^{PV}, \dots, E_t^{PV}\}$
2. Extract the future trends of electricity prices and solar generations
   $\{\lambda_{t+1}^G, \lambda_{t+2}^G, \dots, \lambda_T^G\} \leftarrow \text{NN}(\{\lambda_{t-23}^G, \dots, \lambda_t^G\})$
   $\{E_{t+1}^{PV}, E_{t+2}^{PV} \dots, E_T^{PV}\} \leftarrow \text{NN}(\{E_{t-23}^{PV}, \dots, E_t^{PV}\})$
3. Output the extracted information

Algorithm 3 Q-learning Algorithm (Decision-making)

1. Initialize Q-value Q arbitrarily
2. **Repeat** for each episode σ
3. | Initialize the state s_t
4. | **Repeat**
5. | | Choose the action a_t for the current state s_t by using ε-greedy policy
6. | | Observe the current reward $r_t(s_t, a_t)$ and the next state s_{t+1}
7. | | Update the Q-value $Q(s_t, a_t)$ via Eq. (8)
8. | **Until** s_{t+1} is terminal
9. **Until** termination criterion, i.e., $|Q^{(\sigma)} - Q^{(\sigma-1)}| \leq \tau$, is satisfied
10. Output the optimal policy π^*, i.e., $\{a_t^*, a_{t+1}^*, \dots, a_T^*\}$
11. Execute the optimal action a_t^* for current time slot t

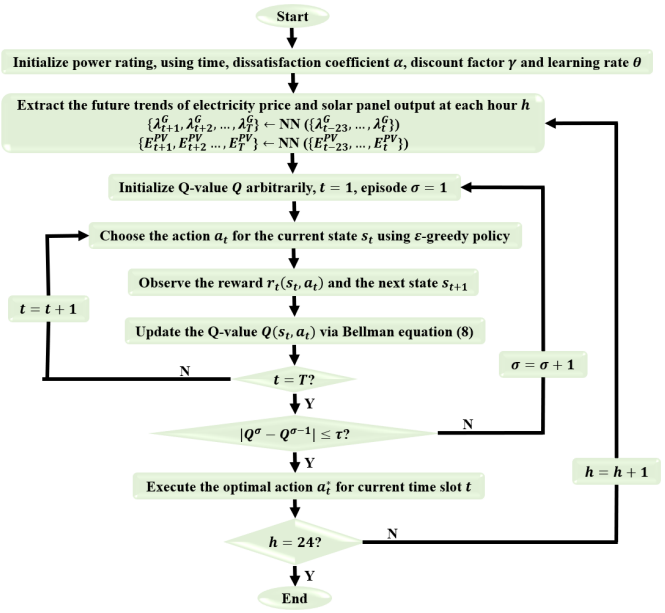


Fig. 3. Flowchart of implementing our proposed solution method for each agent.

Q-learning algorithm is adopted to make ideal energy scheduling decisions for different residential appliances and EV battery charging iteratively, as shown in Algorithm 3. Specifically, in each episode σ, the agent observes the state s_t and then chooses an action a_t using the exploration and exploitation mechanism. To realize the exploration and exploitation, the ε-greedy policy ($\varepsilon \in [0,1]$) [34] is adopted so the agent can either execute a random action form the set of available actions with probability ε or select an action whose current Q-value is maximum, with probability $1-\varepsilon$. After taking an action, the agent acquires an immediate reward $r(s_t, a_t)$, observes the next state s_{t+1} and updates the Q-value $Q(s_t, a_t)$ via Eq. (8). This process is repeated until the state s_{t+1} is terminal. After one episode, the agent checks the episode termination criterion, i.e., $|Q^{\sigma} - Q^{\sigma-1}| \leq \tau$, where τ is a system-dependent parameter to control the accuracy of the convergence. If this termination criterion is not satisfied, the agent will move to the next episode and repeat the above process. Finally, each agent will gain optimal actions for each coming hour, i.e., $h = 1, 2, \ldots, 24$. Note that only the optimal action for the current hour is taken. The above procedure will be repeated until the end hour, namely, $h = 24$. Besides, the flowchart in Fig. 3 clearly depicts the process.

IV. Test Results

A. Case Study Setup

In this study, real-world data is utilized for training our proposed feedforward NN. The hourly data electricity prices and solar generations from January 1, 2017 to December 31, 2018 lasting 730 days are collected from PJM [35]. After a number of accuracy tests, the trained feedforward NN for electricity price data consists of three layers, i.e., one input layer with 24 neurons, one hidden layer with 40 neurons and one output layer with 24 neurons, and the trained feedforward NN for solar generation data also includes three layers, i.e., one input layer with 24 neurons, one hidden layer with 20 neurons and one output layer with 24 neurons. The number of training episode is 50,000. As for the parameters related to the Q-learning algorithm. The discount rate γ is set to 0.9, so the obtained strategy is foresighted. To ensure that the agent can call all state-action pairs and learn new knowledge from the system, the learning rate θ as well as turning parameter ε are both set to 0.1.

TABLE II
Parameters of Each House Appliance and EV Battery

Item ID	Dissatisfaction coefficient	Power rating (kWh)	Using time
REFG	-	0.5	24h
AS	-	0.1	24h
AC1	0.05	0- 1.4	24h
AC2	0.08	0 - 1.4	24h
H	0.12	0 - 1.5	24h
L1	0.02	0 - 0.6	6pm - 11pm
L2	0.03	0 - 0.6	6pm - 11pm
WM	0.1	0.7	7pm - 10pm
DW	0.06	0.3	8pm - 10pm
EV	0.04	0 - 6	11pm - 7am

REFG: refrigerator; AS: alarm system; AC: air conditioner; H: heating; L: light; WM: wash machine; DW: dishwashing; EV: electric vehicle.

In this paper, simulations are conducted on a detached residential house with two same solar panels, two non-shiftable appliances (REFG and AS), five power-shiftable appliances (AC1, AC2, H, L1 and L2), two time-shiftable appliances (WM and DW) and one EV. Detailed parameters of these home appliances and the EV battery are listed in Table II. Besides, our proposed HEM method can be applied to residential houses with more home appliances and renewable resources. All simulations are implemented by using MATLAB with an Intel Core i7 of 2.4 GHz and 12GB memory.

B. Performance of the Proposed Feedforward NN

Fig. 4 and Fig. 5 show performance of the proposed feedforward NN for extracting features of electricity prices as well as solar generations, respectively. For the hour-ahead horizon, the mean absolute percentage error (MAPE) of the forecasted PV output is 8.82%, and the MAPE of the forecasted electricity price is 9.34%. In these two figures, the blue line represents the extracted future values, and the red line indicates the actual values. It can be observed that both extracted trends of electricity prices and solar generations are generally similar to actual ones, though small errors can be observed from some time slots. Therefore, the proposed feedforward NN can generate accurate and reasonable forecasting values, which can benefit the following decision-making process for energy consumption scheduling.

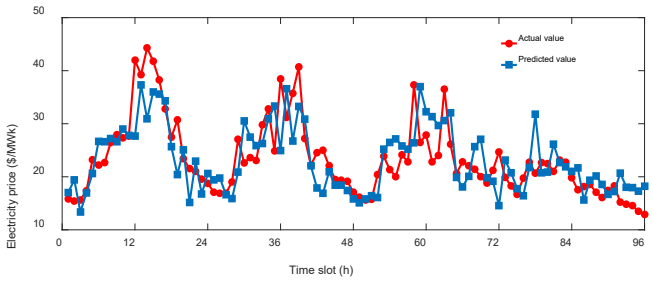


Fig. 4. Comparison of the actual and predicted electricity prices on January 1-4, 2019.

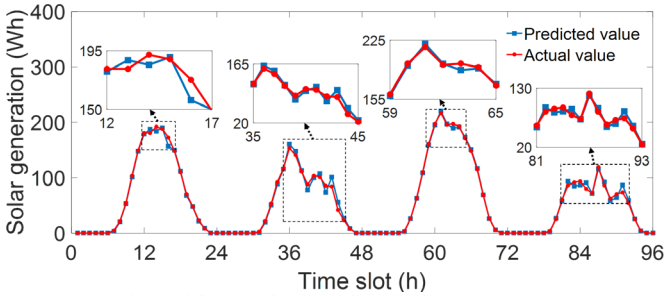


Fig. 5. Comparison of the actual and predicted solar generations on January 1-4, 2019.

To investigate the impact of the prediction accuracy on the solution results, we compare the performance on operation cost between case 1-3 and optimal solution. Note that the optimal solution can be obtained by conventional optimization method based on the perfect prediction. Each case includes two kinds of predicted information, e.g. PV generation and electricity price. Table III lists the MAPE of the prediction result in case 1-3. Fig. 6 is plotted to demonstrate the comparison result. As shown in this figure, with the increase of the prediction accuracy, the operation cost obtained by the *Q*-learning algorithm based RL method is lower, which becomes closer to the optimal solution. Therefore, the prediction accuracy has a direct effect on the optimal result. In this paper, we introduce the ELM based NN to dynamically produce the future uncertainties to provide state statues for the agent in *Q*-learning algorithm. However, developing more accurate prediction model to reduce the prediction error is out of the scope of this paper, since more explanatory variable should be included, e.g., load demand, historical data, power trading direction, energy policy, etc.

TABLE III
MAPE OF PREDICTION RESULT IN CASE 1-3

	Case 1	Case 2	Case 3
MAPE of PV generation prediction	4.41%	8.82%	17.64%
MAPE of electricity price prediction	4.67%	9.34%	18.68%

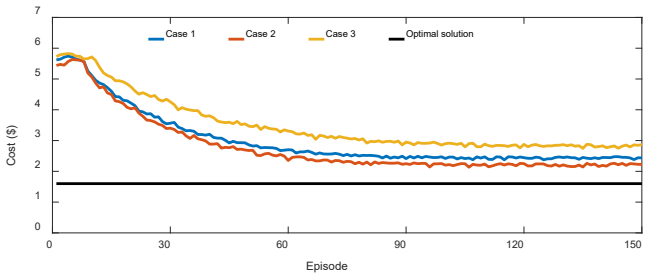


Fig. 6. Comparison of operation cost with different prediction accuracy.

C. Performance of Multi-agent Q-learning Algorithm

TABLE IV
COMPUTATIONAL EFFICIENCY PERFORMANCE WITH DIFFERENT NUMBER OF STATE-ACTION PAIR

No. of state-action pair (No. of state status * No. of action status)	Computation time (s)
$3.6*10^2$ (24*15)	1.315
$3.6*10^3$ (24*150)	2.067
$3.6*10^4$ (24*1,500)	7.385
$3.6*10^5$ (24*15,000)	317.992

Table IV lists the performance on computation time, considering a different number of state-action pair. As shown in this table, with the increase of the state-action pair, the *Q*-learning algorithm takes more time to fill up the *Q*-table and find the optimal *Q*-value. It should be noted that the state space is fixed (24 state statuses), so the state-action pair increases with larger considered action space (power ratings of home appliance). For example, 15 action statues correspond to 15 power ratings, i.e., 0 kWh, 0.1 kWh, 0.2 kWh, …, 1.4 kWh, and 150 action statues correspond to 150 power ratings, i.e., 0 kWh, 0.01 kWh, 0.02 kWh, …, 1.4 kWh. Therefore, more accurate energy consumption scheduling poses a minor effect on optimal results. Besides, it is difficult to achieve precise control of the power rating for most home appliances. In this regard, it is reasonable to consider a small number of state-action pairs for each home appliance in each time slot, resulting in short search time.

Fig. 7 depicts the convergence of the Q-value for each power-shiftable agent on January 1, 2019. It can be seen from this figure that each power-shiftable appliance agent converges to the maximum Q-value. In the beginning, the Q-value is low since the agent takes poor actions, then it becomes high as the agent discovers the actions by learning them through trials and errors, finally reaching the maximum Q-value.

To illustrate the effectiveness of our proposed HEM algorithm, Fig. 8 is plotted to show the energy consumption of all power-shiftable appliances in each time slot. As shown in this figure, the energy consumption is high during the first five time slots. Then these five appliances reduce their energy consumption since the electricity price increases from the time slot 6. As the electricity price reaches its maximum in around time slot 13, the energy consumption of each appliance decreases to its minimum value. Finally, with the price goes down, the energy consumption starts to increase.

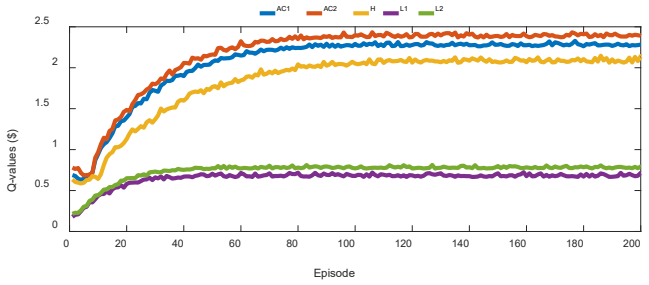


Fig. 7. The convergence of Q-value for power-shiftable agents on January 1, 2019.

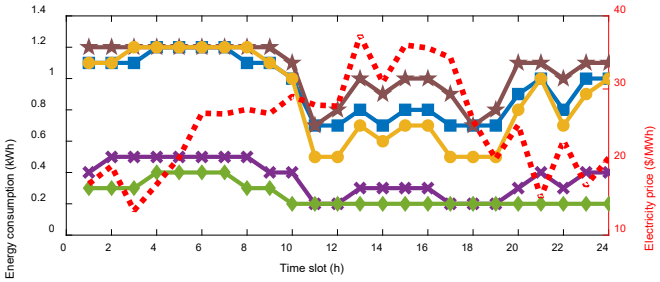


Fig. 8. Energy consumption of five power-shiftable appliances in each time slot.

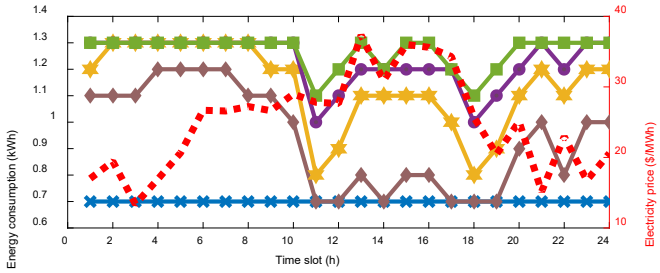
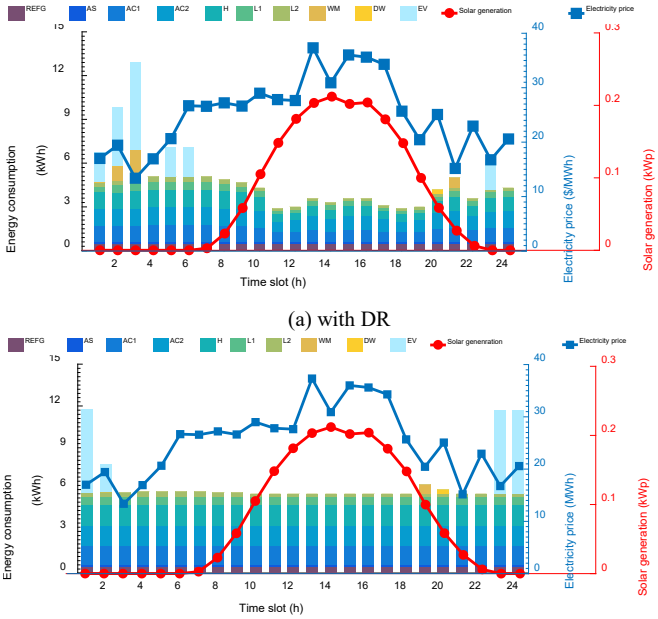


Fig. 9. Energy consumption of AC1 with changing dissatisfaction coefficient α during 24-time slots without consideration of solar generation.

Fig. 9 demonstrates the results of daily energy consumption for AC1 with different five dissatisfaction coefficients. We can see that as the dissatisfaction coefficient increases, the daily energy consumption goes up since the dissatisfaction coefficient can be regarded as a penalty factor. This creates a trade-off between saving electricity bill and decreasing dissatisfaction caused by reducing power rating of AC1. Besides, this figure also shows that the agent leans to increase energy consumption for low dissatisfaction during the off-peak time slots and decrease energy consumption for low electricity cost during the on-peak time slots. These observations verify that the proposed method can be applied to consumers for helping them manage their individual energy consumption.



(a) with DR

(b) without DR

Fig. 10. Energy consumption of all appliances on January 1, 2019, (a) with and (b) without DR.

TABLE V
COMPARISON OF ELECTRICITY COST WITH AND WITHOUT DR

Item ID	Electricity cost ($)	
	With DR	Without DR
REFG	0.492	0.492
AS	0.098	0.098
AC1	0.836	1.378
AC2	0.942	1.378
H	0.731	1.476
L1	0.301	0.591
L2	0.223	0.591
WM	0.023	0.051
DW	0.012	0.012
EV	0.399	1.262
Total	4.057	7.329

REFG: refrigerator; AS: alarm system; AC: air conditioner; H: heating; L: light; WM: wash machine; DW: dishwashing; EV: electric vehicle

Fig. 10. gives the daily energy consumption of each home appliance and EV in two different cases with and without DR, along with the electricity prices and solar panel outputs. With DR mechanism, more energy is consumed when the price is low, and the load demand is reduced when the price is high, as shown in Fig. 10(a). Thus, the power-shiftable or time-shiftable loads can be reduced or scheduled to off-peak periods, maintaining the overall energy consumption at a low level during the on-peak periods. By contrast, for the case without DR, as shown in Fig. 10(b), no reduction or shift on energy consumption can be observed. The comparison of electricity costs in these two cases is listed in Table V, which shows that the electricity cost can be significantly reduced with DR.

D. Numerical Comparison with Genetic Algorithm

To evaluate the performance of our proposed *Q*-learning algorithm-based solution method, the genetic algorithm (GA) [36] based solution method is compared as a benchmark. The benchmark problem is a mixed-integer nonlinear programming

problem, and its objective function is to minimize the electricity cost and dissatisfaction cost, as given by equations (1)-(4). We can see from Fig. 11 that our proposed solution method (red line) shows a poor performance at the initial training stage since it is undergoing trials and errors. However, after experiencing more iterations, our method adapts to the learning environment and adjusts its policy via exploration and exploitation mechanism. For a longer time, it outperforms the GA based solution method (blue line). The reason is that the RL agent not only considers the current reward but also the future rewards so it can learn from the environment while the GA algorithm has low learning capability. Note that the black dashed line in Fig. 11 is plotted as the benchmark to show the optimal result obtained by the conventional optimization method, which knows all environment information and removes prediction errors.

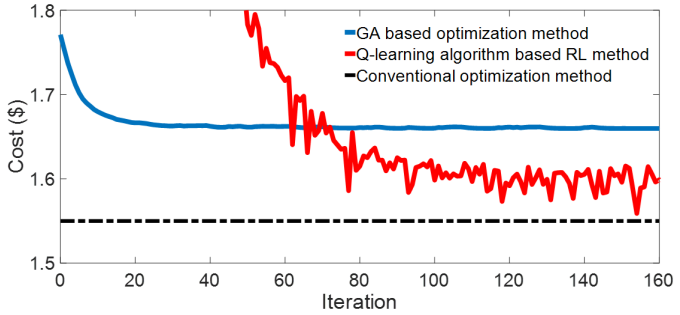


Fig. 11. Optimization performance comparison of the three methods for scheduling AC1 loads.

TABLE VI
COMPARISON ON COMPUTATION EFFICIENCY BY GA OPTIMIZATION METHOD AND *Q*-LEARNING ALGORITHM BASED RL METHOD

	Average computation time of running 1000 times
GA based optimization method	46.296 s
Q-learning algorithm based RL method	1.107 s

Besides, Table VI is added to compare the computation efficiency between the proposed solution method and benchmark. It can be observed that our proposed solution method is able to significantly reduce the computation time. The reasons can be summarized into two aspects: 1) GA algorithm is based on Darwin's theory of evolution, which is a slow gradual process that works by making changes to the making slight and slow changes. Moreover, GA usually makes slight changes to its solutions slowly until getting the best solution. 2) In the *Q*-learning algorithm, the agent chooses an action using the exploration and exploitation mechanism, so it is fast by employing the ε-greedy policy to explore and exploit the optimum from the look-up table. Note that only a small number of state-action pairs need to be searched by the *Q*-learning algorithm, resulting in a high computation efficiency. In this regard, considering the adaptivity of model-free RL to the external environment, it is suggested to accept our proposed well-performing solution method for HEM system.

V. CONCLUSION

Based on a feedforward NN and Q-learning algorithm, this paper proposes a new multi-agent RL based data-driven method for HEM system. Specifically, ELM is employed to train the feedforward NN to predict future trends of electricity price and solar generation according to real-world data. Then, the predicted information is fed into the multi-agent *Q*-learning algorithm based decision-making process for scheduling the energy consumption of different home appliances and EV charging. To implement the proposed HEM method, the FMDP is utilized to model the hour-ahead energy consumption scheduling problem with the objective of minimizing the electricity bill as well as DR included dissatisfaction. Simulations are performed on a residential house with multiple home appliances, an EV and several PV panels. The test results show that the proposed HEM method can not only achieve promising performance in terms of reducing electricity cost for householders but also improve the computational efficiency. In the future, energy storage for rooftop solar PV systems will also be considered in the HEM system. Besides, more effective uncertainty prediction model will be developed to facilitate the decision-making process of DR.

REFERENCES

[1] H. Shareef, M. S. Ahmed, A. Mohamed, and E. Al Hassan, "Review on home energy management system considering demand responses, smart technologies, and intelligent controllers," *IEEE Access,* vol. 6, pp. 24498-24509, 2018.

[2] Y. Chen, Y. Xu, Z. Li, and X. Feng, "Optimally coordinated dispatch of combined-heat-and-electrical network with demand response," *IET Generation, Transmission & Distribution,* 2019.

[3] F. Luo, G. Ranzi, S. Wang, and Z. Y. Dong, "Hierarchical energy management system for home microgrids," *IEEE Transactions on Smart Grid,* 2018.

[4] F. Luo, W. Kong, G. Ranzi, and Z. Y. Dong, "Optimal Home Energy Management System with Demand Charge Tariff and Appliance Operational Dependencies," *IEEE Transactions on Smart Grid,* 2019.

[5] X. Wu, X. Hu, X. Yin, and S. Moura, "Stochastic optimal energy management of smart home with PEV energy storage," *IEEE Transactions on Smart Grid,* vol. 9, no. 3, pp. 2065-2075, 2016.

[6] V. Pilloni, A. Floris, A. Meloni, and L. Atzori, "Smart home energy management including renewable sources: A qoe-driven approach," *IEEE Transactions on Smart Grid,* vol. 9, no. 3, pp. 2006-2018, 2016.

[7] L. Yu, T. Jiang, and Y. Zou, "Online energy management for a sustainable smart home with an HVAC load and random occupancy," *IEEE Transactions on Smart Grid,* 2017.

[8] S. Sharma, Y. Xu, A. Verma, and B. K. Panigrahi, "Time-Coordinated Multi-Energy Management of Smart Buildings under Uncertainties," *IEEE Transactions on Industrial Informatics,* 2019.

[9] C. Keerthisinghe, G. Verbič, and A. Chapman, "A fast technique for smart home management: ADP with temporal difference learning," *IEEE Transactions on smart grid,* vol. 9, no. 4, pp. 3291-3303, 2016.

[10] Y. Huang, L. Wang, W. Guo, Q. Kang, and Q. Wu, "Chance constrained optimization in a home energy management system," *IEEE Transactions on Smart Grid,* vol. 9, no. 1, pp. 252-260, 2016.

[11] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning* (no. 4). MIT press Cambridge, 1998.

[12] J. R. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Applied energy,* vol. 235, pp. 1072-1089, 2019.

[13] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Transactions on Smart Grid,* vol. 8, no. 5, pp. 2149-2159, 2016.

[14] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-Free Real-Time EV Charging Scheduling Based on Deep Reinforcement Learning," *IEEE Transactions on Smart Grid,* 2018.
[15] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach," *Applied Energy,* vol. 220, pp. 220-230, 2018.
[16] R. Lu, S. H. Hong, and M. Yu, "Demand Response for Home Energy Management using Reinforcement Learning and Artificial Neural Network," *IEEE Transactions on Smart Grid,* 2019.
[17] S. Zhou, Z. Hu, W. Gu, M. Jiang, and X.-P. Zhang, "Artificial intelligence based smart energy community management: A reinforcement learning approach," *CSEE Journal of Power and Energy Systems,* vol. 5, no. 1, pp. 1-10, 2019.
[18] X. Huang, S. H. Hong, M. Yu, Y. Ding, and J. Jiang, "Demand response management for industrial facilities: A deep reinforcement learning approach," *IEEE Access,* vol. 7, pp. 82194-82205, 2019.
[19] E. Foruzan, L.-K. Soh, and S. Asgarpoor, "Reinforcement learning approach for optimal distributed energy management in a microgrid," *IEEE Transactions on Power Systems,* vol. 33, no. 5, pp. 5749-5758, 2018.
[20] E. Mocanu *et al.*, "On-line building energy optimization using deep reinforcement learning," *IEEE Transactions on Smart Grid,* 2018.
[21] Y. Du and F. Li, "Intelligent Multi-microgrid Energy Management based on Deep Neural Network and Model-free Reinforcement Learning," *IEEE Transactions on Smart Grid,* 2019.
[22] T. Remani, E. Jasmin, and T. I. Ahamed, "Residential load scheduling with renewable generation in the smart grid: A reinforcement learning approach," *IEEE Systems Journal,* 2018.
[23] A. McCabe, D. Pojani, and A. B. van Groenou, "Social housing and renewable energy: Community energy in a supporting role," *Energy Research Social Science,* vol. 38, pp. 110-113, 2018.
[24] M. A. J. E. C. Green, NJ, Prentice-Hall, Inc., . 288 p., "Solar cells: operating principles, technology, and system applications," 1982.
[25] G. Walker, "Evaluating MPPT converter topologies using a MATLAB PV model," *Journal of Electrical Electronics Engineering, Australia,* vol. 21, no. 1, p. 49, 2001.
[26] S. M. Rezvanizaniani, Z. Liu, Y. Chen, and J. Lee, "Review and recent advances in battery health monitoring and prognostics technologies for electric vehicle (EV) safety and mobility," *Journal of Power Sources,* vol. 256, pp. 110-124, 2014.
[27] C. Le Floch, E. C. Kara, and S. Moura, "PDE modeling and control of electric vehicle fleets for ancillary services: A discrete charging case," *IEEE Transactions on Smart Grid,* vol. 9, no. 2, pp. 573-581, 2016.
[28] M. Yu and S. H. Hong, "Incentive-based demand response considering hierarchical electricity market: A Stackelberg game approach," *Applied Energy,* vol. 203, pp. 267-279, 2017.
[29] M. Rafiei, T. Niknam, J. Aghaei, M. Shafie-Khah, and J. P. Catalão, "Probabilistic load forecasting using an improved wavelet neural network trained by generalized extreme learning machine," *IEEE Transactions on Smart Grid,* vol. 9, no. 6, pp. 6961-6971, 2018.
[30] S. Chai, Z. Xu, and Y. Jia, "Conditional Density Forecast of Electricity Price based on Ensemble ELM and Logistic EMOS," *IEEE Transactions on Smart Grid,* 2018.
[31] W. Fu, K. Wang, C. Li, and J. Tan, "Multi-step short-term wind speed forecasting approach based on multi-scale dominant ingredient chaotic analysis, improved hybrid GWO-SCA optimization and ELM," *Energy Conversion Management,* vol. 187, pp. 356-377, 2019.
[32] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning,* vol. 8, no. 3-4, pp. 279-292, 1992.
[33] H. J. Kappen, "Optimal control theory and the linear bellman equation," 2011.
[34] M. Tokic and G. Palm, "Value-difference based exploration: adaptive control between epsilon-greedy and softmax," in *Annual Conference on Artificial Intelligence*, 2011, pp. 335-346: Springer.
[35] M. Attar, O. Homaee, H. Falaghi, and P. Siano, "A novel strategy for optimal placement of locally controlled voltage regulators in traditional distribution systems," *International Journal of Electrical Power & Energy Systems,* vol. 96, pp. 11-22, 2018.
[36] J. R. Koza, "Genetic programming," 1997.