Optimization Strategy Based on Deep Reinforcement Learning for Home Energy Management

Yuankun Liu, Dongxia Zhang, and Hoay Beng Gooi

***Abstract*—With the development of a smart grid and smart home, massive amounts of data can be made available, providing the basis for algorithm training in artificial intelligence applications. These continuous improving conditions are expected to enable the home energy management system (HEMS) to cope with the increasing complexities and uncertainties in the end-user side of the power grid system. In this paper, a home energy management optimization strategy is proposed based on deep Q-learning (DQN) and double deep Q-learning (DDQN) to perform scheduling of home energy appliances. The applied algorithms are model-free and can help the customers reduce electricity consumption by taking a series of actions in response to a dynamic environment. In the test, the DDQN is more appropriate for minimizing the cost in a HEMS compared to DQN. In the process of method implementation, the generalization and reward setting of the algorithms are discussed and analyzed in detail. The results of this method are compared with those of Particle Swarm Optimization (PSO) to validate the performance of the proposed algorithm. The effectiveness of applied data-driven methods is validated by using a real-world database combined with the household energy storage model.**

***Index Terms*—Deep reinforcement learning, demand response, home energy management system, smart grid.**

I. Introduction

In recent years, with the development of power electronics and distributed energy technologies, the physical structure of the end-user in the power grid has been changing. Since distributed solar photovoltaics, household energy storage, and electrical vehicles have emerged in large numbers, the uncertain and complicated operating environments need to be managed by the home energy management system (HEMS). The structure of HEMS can be seen in Fig. 1. By means of a control system, communication technologies, and optimization strategies, the equipment operations can be controlled and scheduled by the HEMS to improve the overall energy efficiency. Furthermore, the HEMS is also an extension of the smart grid, where the prosumers can cooperate with the grid company to rationally arrange power consumption schemes and demand response strategies through two-way communications [1]–[3].

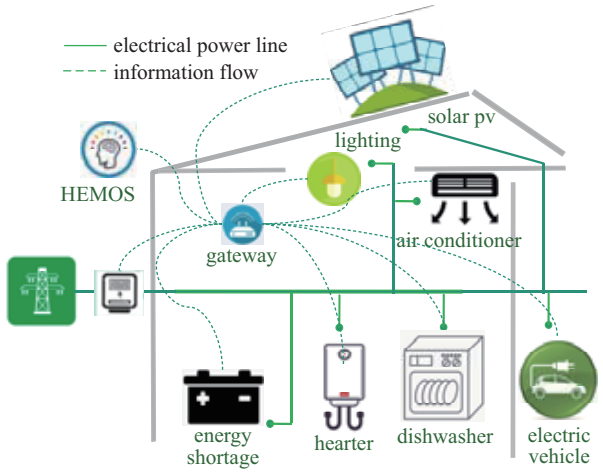

Fig. 1. The structure of HEMS.

The conventional HEMS is based on remote monitoring and remote control. The control strategies adopt simple "if-then" rules that are strict and passive, lack the predictive function, and cannot continuously learn the end users' behavior patterns. There is also little coordination among the internal subsystems. In addition, the current control strategy is difficult in meeting the demand for development due to the complexities and uncertainties. Therefore, it can be seen from the discussion above that the optimization strategy plays an essential role in the HEMS to satisfy the expectations for intellectualization and personalization of end-users in the power grid [4]–[6].

Numerous methods such as linear and dynamic programing, particle swarm optimization (PSO), fuzzy methods, and game theory have been proposed in the home energy management optimization strategy (HEMOS) fields [7]–[10]. The approaches mentioned above have the following characteristics:

1) The detailed appliances and environment models are essential in most HEMOSs. In this case, the accuracy of the fixed model cannot be guaranteed because the efficiency of the appliance and the environmental variables keep changing over time.

2) In the implementation of the above methods, various conditions with complex and uncertainty variables should be

Manuscript received November 5, 2019; revised February 28, 2020; accepted March 17, 2020. Date of online publication April 6, 2020; date of current version June 3, 2020.

Y. K. Liu (corresponding author, e-mail: yjs-lyk@epri.sgcc.com.cn) and D. X. Zhang are with China Electric Power Research Institute, Beijing 100192, China.

H. B. Gooi is with Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore.

DOI: 10.17775/CSEEJPES.2019.02890

2096-0042 © 2019 CSEE

considered, such as prosumer behavior that is not suitable for modeling.

3) The objective function and the constraint condition in some of these methods ignore a few unquantifiable factors or adopt inaccurate model formulas such as the users' satisfaction or comfort; therefore, there are disparities in different groups of people due to distinctive household appliances.

Machine learning technology can be used as an effective means for a smart home, especially in the field of HEMS [11]. There were some machine learning based approaches developed in this area. In reference [12], a time-of-use pricing (TOU) based method was proposed in the smart home energy management system. Reference [13] developed a novel methodology for home area energy management as a key vehicle for demand response, using electricity storage devices. In reference [14], a dynamic soft constraint method was proposed to enable the thermostatically controlled appliances to schedule work in both normal and abnormal situations. Reference [15] proposed an intelligent appliance control (IAC) algorithm to monitor and control the daily operation of these power-intensive appliances using their simulated load models. As a significant branch of machine learning, reinforcement learning (RL) has been proved to be applicable to power system decision-supporting for controlling and scheduling problems under a dynamic environment [16], [17]. Deep reinforcement learning (DRL) that combines deep learning (DL) with reinforcement learning has made huge progress in some applications [18]. These DL techniques allow for the learning of rich features from the massive amount of data and overcome the curse of dimensionality shortcomings, making RL suitable for handling large-scale problems. In reference [19], a fully automatic energy management algorithm was proposed based on reinforcement learning to learn how to make the best decision for consumers. A batch reinforcement learning method was introduced in reference [20] to arrange a group of household electric water heaters and was further applied to smart home energy management [21]. Unlike the traditional rule-based and model-based strategies, the DRL was used in reference [22] to learn the effective strategy for operating the HVAC systems in the building. In reference [23], the DRL framework was used to deal with the high-dimension and computational speed problems in order to overcome the difficulty in the building energy model in the classical model-based optimal control. Regarding the dissatisfaction cost, deep reinforcement learning was applied to control electric devices [24]. In reference [25], DRL was applied in HEMS by integrating the simulation environment with a machine learning platform and battery energy storage. The effectiveness of the DRL algorithm was verified by simulation in reference [26]. In addition, using Deep Q-learning for storage scheduling in microgrids was proposed in reference [27]. In reference [28], DRL was used in the area of smart cities and the Internet of Things (IoT). Since the storage device is important in the future microgrids, the application of batch RL was proposed to optimally schedule the operation of a storage device in energy management [29]. The capabilities of different deep learning techniques were investigated in reference [30] to extract relevant features, where DRL was proposed to schedule the thermostatically controlled loads. Moreover, multi-agent deep reinforcement learning was proposed in reference [31] to improve energy sharing in a community. Furthermore, DRL was applied in reference [32] to deal with the optimal control problem of building energy conservation; two deep reinforcement learning algorithms were applied and evaluated.

In this paper, a home energy management optimization strategy (HEMOS) is regarded as a smart agent; data-driven methods instead of model-based methods are applied to formulate the optimization problem. The contribution of this paper can be presented as follows:

1) A novel home energy management optimization strategy is proposed. DDQN and DQN are adopted to maximize the energy efficiency and DR potential of the equipment.

2) A demonstration is carried out in a real-world data set combined with an energy storage model to verify the validity and advancement of the proposed methods.

3) An intensive study is carried out to investigate the generalization and the reward settings of the algorithms in HEMS.

This research proves that the energy efficiency of end-users in a dynamic environment can be improved by DDQN and DQN. Specifically, it is evaluated that DDQN is more appropriate for cost minimization problems compared to DQN. The applied method has the generalization in HEMS fields such as without the PV scenario and without the EV scenario. It is analyzed that the reward setting should be adjusted according to actual situations. The DDQN is more suitable and effective than PSO in a home energy management system.

The rest of the paper is organized as follows.

The overview of the background and applied algorithms are explained in Section II. First, the characteristics of DL, RL, and their combination DRL are introduced, as well as the relationships between them. Then, mathematical principles of DQN and DDQN algorithms (both of them belong to DRL) are explained in detail.

In Section III, the research problem is described in mathematical formulas. First, data-set characteristics introduction and schedulable appliance analysis are conducted. Next, the scene is summarized as an optimization problem in mathematics; its objective functions and constraints are proposed. Then, the household energy storage model is proposed. Finally, the relationship between the training part and the execution part in DRL is introduced; the implementation of the algorithm is explained in detail from four aspects: agent, environment, reward, and action.

Results and discussions are included in Section IV. First, the learning process and overall trends of DQN and DDQN are displayed. Then, numerical results and comparisons of DQN and DDQN are presented. Next, the generalization and reward setting during the implementation are analyzed. Finally, the comparison of results between applied algorithms and the traditional one (PSO) is performed.

In Section V, conclusions and future study are provided.

II. Background and Applied Method

Many algorithms of the new generation of AI technologies are emerging and are in the process of development. DL, RL

and their combination, DRL, are representative methods [33], [34]. In this section, we provide a brief overview of DL, RL, and DRL. The applied DQN and DDQN algorithms are explained in detail.

A. RL, DL, and DRL

As shown in Fig. 2, there are four basic parts in RL: agent, environment, reward, and action. RL is a mapping between states and actions to maximize rewards and support agents in response to a dynamic and uncertain environment. Unlike other algorithm learning based on prior knowledge of external supervisors, RL is learning in interaction with the environment.

Fig. 2. Agent and Environment in RL.

The concept of DL stems from artificial neural networks. The function of DL is to unify feature extraction and learning in a multi-layer neural network. As a type of data-driven method, it overcomes the problem of over-rigidity of artificial extraction features, considering the complex environmental factors and contributing to solving nonlinear problems [35].

DRL combining DL with RL introduces neural networks to directly express and optimize value functions, strategies, or environmental models in an end-to-end manner. DRL can make full use of high-dimensional original input data to extract patterns and build models; in addition, it can be used as a basis for policy control. Compared with traditional reinforcement learning, deep reinforcement learning overcomes the inability to handle high-dimensional large-scales [36] and the success of the AlphaGo attributes to this improved algorithm.

B. Deep Q-learning (DQN)

In Q-learning, the Q-value of each state-action pair (the value of each action selected in each state) is stored in the Q-table and updated by a stochastic gradient descent method.

$$Q(s_{\mathrm{t}}, a_{\mathrm{t}}) \leftarrow Q(s_{\mathrm{t}}, a_{\mathrm{t}}) + \alpha(R_{\mathrm{t+1}} + \lambda \max_{a'} Q(s_{\mathrm{t+1}}, a') - Q(s_{\mathrm{t}}, a_{\mathrm{t}})) \tag{1}$$

In (1), α is the step-size control and $R_{\mathrm{t+1}} + \lambda \max_{a} Q(S_{\mathrm{t+1}}, a)$ is the expected rewards that can be obtained by performing an action a_{t} in state s_{t}. In the high dimension, the agent is too slow to learn the value of each state individually. The Q-learning becomes unrealistic when the state and action space are in high dimension.

The value function approximation was proposed in reference [36] in order to overcome this problem. By adjusting the parameter ω, the function conforms to the value function based on a certain strategy as shown in (2).

$$Q(s, a, w) \approx Q_{\pi}(s, a) \tag{2}$$

Through this method, the task is transformed into solving the parameter ω in the objective function:

$$L(w) = E[(R_{\mathrm{t+1}} + \lambda \max_{a'} Q(s_{\mathrm{t+1}}, a'; w') - Q(s_{\mathrm{t}}, a_{\mathrm{t}}; w))^2] \tag{3}$$

The stochastic gradient descent method is adopted to gradually approximate the parameters, helping the objective function converge to the minimum value.

C. Double Deep Q-learning (DDQN)

When selecting action a of the current state s, Q-learning depends on the ε-greedy method that directly selects the action corresponding to the maximum value function in the current state.

Q-learning is a model-free algorithm using environmental sampling to estimate the value function. In this process, there is an overestimation affecting the decision of the strategy; the selected action is not optimal. Therefore, DDQN was proposed in reference [37] to solve this problem. For DDQN, the action selection and action evaluation are different in the Q-value functions, and the corresponding parameter ω is different from DQN:

$$Y_{\mathrm{t}}^{\mathrm{DoubleQ}} = R_{\mathrm{t+1}} + \gamma Q(s_{\mathrm{t+1}}, \arg\max_{a'} Q(s_{\mathrm{t+1}}, a', w_{\mathrm{t}}); w'_{\mathrm{t}}) \tag{4}$$

The objective function of the DDQN is:

$$L(\omega)^{\mathrm{DoubleQ}} = E[(R_{\mathrm{t+1}} + \gamma Q(s_{\mathrm{t+1}}, \arg\max_{a'} Q(s_{\mathrm{t+1}}, a', w_{\mathrm{t}}); w'_{\mathrm{t}}) - Q(s_{\mathrm{t}}, a_{\mathrm{t}}, w_{\mathrm{t}}))^2] \tag{5}$$

In DDQN, the target network is transformed into a separately updated network. The independent updated network is employed to provide the target value. The primary network is used to select an action; meanwhile, instead of directly selecting the action corresponding to the maximum Q-value, a target Q-value is generated for evaluating the selected action by the target network. In addition, the target Q-value is calculated during the training. The detailed implementation of the algorithm is shown below.

Algorithm 1 Double deep Q-learning of HEMOS

Input: D-buffer; w-initial network parameters, w^{-}-a copy of w

Input: N_{r}-buffer maximum size; N_{b}-training batch size; N-target network replacement freq.

for episode $e \in \{1, 2, \cdots, M\}$ **do**
  Initialize frame sequence $x \leftarrow ()$
  for $t \in \{0, 1, \cdots\}$ **do**
    Set state $s \leftarrow x$, sample action $a \sim \pi_{\beta}$
    Sample next frame x^{t} from environment ε given (s, a) and receive the reward r, and append x^{t} to x
    if $|x| > N_{\mathrm{f}}$ **then**
      delete the oldest frame $x_{t\,\min}$ from x
    end if
    Set $s' \leftarrow x$, and add transition tuple (s, a, r, s') to D, replacing the oldest tuple if $|D| \geq N_{\mathrm{r}}$
    Sample a minibatch of N_{b} tuples $(s, a, r, s') \sim$ Unif (D)

Construct target values, one of each of the N_b tuples:
Define $a^{\max}(s';w) = \arg\max_{a'} Q(s',a';w)$
$y_j = \begin{cases} r & \text{if } s' \text{ is terminal} \\ r + \lambda Q(s', a^{\max}(s',w); w^-) & \text{otherwise} \end{cases}$
Do a gradient descent step with loss $\|y_j - Q(s,a;w)\|^2$
Replace target parameters $w^- \leftarrow w$ every N^- step
end for
end for

III. Problem Formulation

HEMS can be summarized as a mathematical optimization problem that is accompanied by complex environmental changes, including a variety of devices with their distinct characteristics. Optimal control plays a critical role in maximizing the energy efficiency and DR potential of the equipment.

A. Data-set Characteristics and Schedulable Appliance Analysis

The large real-world open-source data set recorded by Pecan Street. Inc. is used in this paper [38]. The data set contains electrical submeters that have accumulated over many years. Samples of 4 days of power consumption are randomly selected in Fig. 3. As four main controllable components, the power consumption time series of an air conditioner, heater, dishwasher, and an electric car are shown in different color curves. The blue curves are the total household energy consumption that is the sum of non-controllable loads and controllable loads. The solar PV generation is shown in the green color curve. As shown in Fig. 3, the total household energy consumption, solar photovoltaic, air conditioners, and electric vehicles that have a greater impact on the energy efficiency of users are selected in this case. The data used in the algorithm training is collected in one house every 15 minutes for a period of one year.

The electricity price data is selected by the local power operator, including Time-of-use plan and PV on-grid price. The necessary simplified modifications were made according to the needs of the algorithm. The above electricity price model can be regarded as a fixed electricity price. The random price fluctuation from −0.01$/kWh to 0.03$/kWh is provided based on the fixed price (the changing price) in order to test the generalization ability of the algorithm.

Since the appliances have different operational patterns, the appliances should be categorized. Based on the analysis of the physical structure and the usage habits, the electrical appliances of residents can be divided into three categories as follows:

1) Base load, which does not have the ability to reduce and shift, can be regarded as a fixed demand for electricity usage.
2) Time-shift load, which includes devices such as washing machines and dishwashers, has two statuses (open and closed); the running time can be changed.
3) Power-shift load, which can be flexible in a given usage time interval, such as an air conditioner.

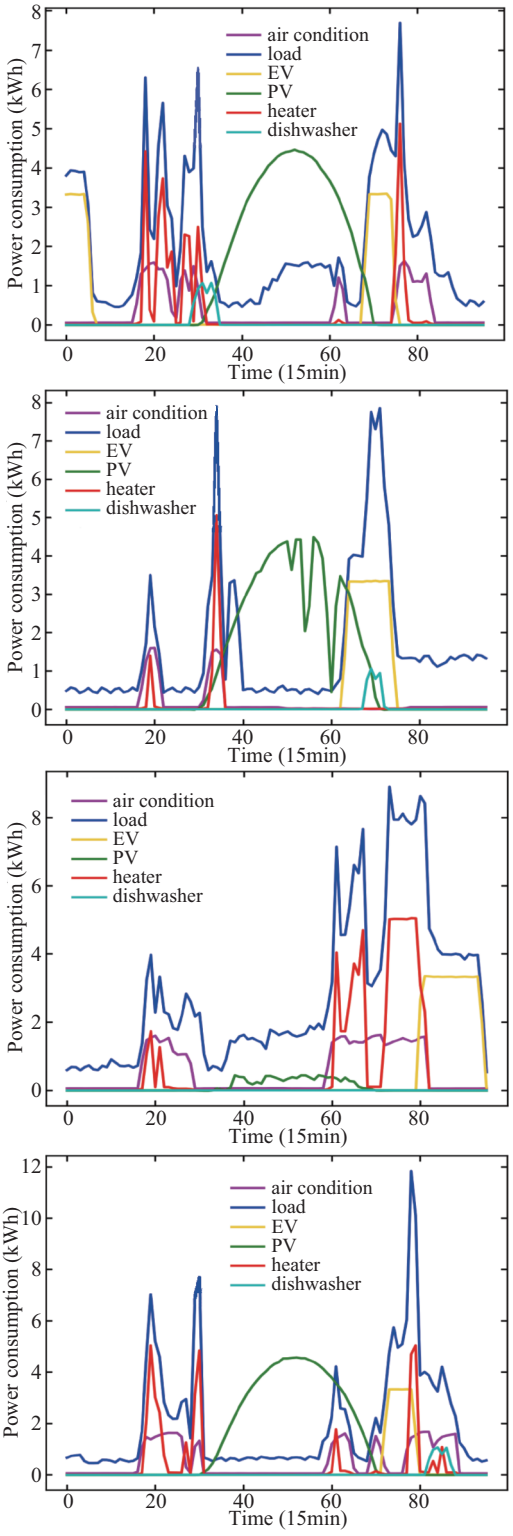

Fig. 3. Power usage behavior randomly selected.

Since the physical constraints and control conditions of the latter two types of electrical appliances are quite different, they need different considerations in the algorithm design.

B. Objective Function and Constraints

The main objective is to schedule the energy consumption of home appliances to help prosumers reduce the electricity cost

based on the operational requirements of home appliances. If the demand is higher than the local clean energy supply, the objective function is:

$$\min C = \sum_{t=1}^{\mathrm{T}} \lambda_{\mathrm{t}}^{-} \Bigg(P_{\mathrm{t}}^{\text{fixed}} + a_{\mathrm{t,d}} \sum_{d=1}^{\mathrm{n}} P_{\mathrm{t,d}} + a_{\mathrm{t,\ cha}} P_{\mathrm{t}}^{\mathrm{cha}} - a_{\mathrm{t,\ disc}} P_{\mathrm{t}}^{\mathrm{disc}} - P_{\mathrm{t}}^{\mathrm{PV}} \Bigg) \tag{6}$$

Otherwise, the objective function is:

$$\max C = \sum_{t=1}^{\mathrm{T}} \lambda_{\mathrm{t}}^{+} \Bigg(P_{\mathrm{t}}^{\mathrm{PV}} - P_{\mathrm{t}}^{\text{fixed}} - a_{\mathrm{t,d}} \sum_{d=1}^{\mathrm{n}} P_{\mathrm{t,d}} - a_{\mathrm{t,\ cha}} P_{\mathrm{t}}^{\mathrm{cha}} + a_{\mathrm{t,\ disc}} P_{\mathrm{t}}^{\mathrm{disc}} \Bigg) \tag{7}$$

$$\text{s.t.} \quad \sum_{t=1}^{\mathrm{T}} P_{\mathrm{d}} \Delta t = E_{\mathrm{d}}, \forall d \in N, \forall t \in N \tag{8}$$

$$a_{\mathrm{t,d}} = \{1, 0\}, \forall a \in A, \forall d \in N, \forall t \in N \tag{9}$$

$$a_{\mathrm{t,\ cha}}, a_{\mathrm{t,\ disc}} = \{1, 0\}, a_{\mathrm{t,\ cha}} \wedge a_{\mathrm{t,\ disc}} = 0, \quad \forall a \in A, \forall t \in N \tag{10}$$

$$\lambda_{\mathrm{t}}^{+}, \lambda_{\mathrm{t}}^{-} \geq 0, \forall t = [1 : T] \in N \tag{11}$$

In the above two equations, λ_{t}^{-} and λ_{t}^{+} are the real-time electricity price of the purchased electricity and the electricity price of selling back to the power grid. Eqs. (6) and (7) are associated with application d at time t, where $a_{\mathrm{t,d}} = 1$ represents the on-state of application at the interval of time Δt; $a_{\mathrm{t,d}} = 0$ represents that it is closed. The same meaning in the charging and discharging of energy shortage is applied to $a_{\mathrm{t,\ cha}}$ and $a_{\mathrm{t,\ disc}}$.

$P_{\mathrm{t}}^{\text{fixed}}$ is the base load consumption. $P_{\mathrm{t}}^{\mathrm{cha}}$ and $P_{\mathrm{t}}^{\mathrm{disc}}$ are the charging power and discharging power. $P_{\mathrm{t}}^{\mathrm{PV}}$ is the generated power from the solar PV. All of them are at an interval of time, Δt.

In different consumption profiles, $P_{\mathrm{t,d}}$ is set with different associated constraints as follows: For the power-shift load, in the specified time interval, the constraints are shown in (12) and (13), where the $p(P_{\mathrm{d}}|t)$ is the probability of the application d to be active in time t and $\exists \varepsilon_{\mathrm{d}} \in R$ is the constant consumption amount in the optimization horizon $T_{\mathrm{d}}^{\mathrm{h}} = [\theta - \sigma, \theta + \sigma]$. θ is the typical start time, σ is the tolerable adjustment range for the users.

$$\sum_{\mathrm{t}} P_{\mathrm{d}} \leq \varepsilon_{\mathrm{d}} \quad \text{if } p(P_{\mathrm{d}}|t) \in (0, 1] \tag{12}$$

$$\sum_{\mathrm{t}} P_{\mathrm{d}} = \varepsilon_{\mathrm{d}} \quad \text{if } p(P_{\mathrm{d}}|t) = 0 \tag{13}$$

For the time-shift load, a minimum amount of power $\exists \varepsilon_{\mathrm{d}} \in R$ must be consumed in the optimization horizon and the usage time $T_{\mathrm{d}}^{\mathrm{u}} \in N$ cannot be disturbed. The continuous usage time is included in the optimization space, $T_{\mathrm{d}}^{\mathrm{u}} \subseteq T_{\mathrm{d}}^{\mathrm{h}}$. The constrains is $\sum_{t}^{T_{\mathrm{u}}} P_{\mathrm{d}} = \varepsilon_{\mathrm{d}}$.

Equations (6) and (7) should be integrated into the Deep RL algorithm. $a_{\mathrm{t}} = \{a_{\mathrm{t,d}}, a_{\mathrm{t,\ cha}}, a_{\mathrm{t,\ disc}}\}$ is the binary action vector. $a_{\mathrm{t,d}}$, $a_{\mathrm{t,\ cha}}$, and $a_{\mathrm{t,\ disc}}$ are the actions of electrical applications that are determined by the output of the neural network shown in Fig. 5. During the training, these actions promote the decision-making to gradually obtain the $\max_{a'} Q(s_{\mathrm{t+1}}, a'; w')$ and $Q(s_{\mathrm{t+1}}, \arg\max_{a'} Q(s_{\mathrm{t+1}}, a', w_{\mathrm{t}}); w_{\mathrm{t}}')$ in (3) and (5). Therefore $a_{\mathrm{t,d}} \sum_{d=1}^{\mathrm{n}} P_{\mathrm{t,d}}$ in (6) and (7) are optimally scheduled.

The conventional method specifies more detailed constraints and discomfort functions of users. The method in this paper transfers these constraints and conditions into reward functions which is further detailed in Section III-D.

C. Household Energy Storage Model

Household energy storage is a development trend and a controllable load that can transfer the load. The energy storage model is proposed in this study. The relationship between the state of charge (SOC) and the energy storage with the charging power and the discharging power are described as follows:

$$\mathrm{SOC}(t+1) = \mathrm{SOC}(t) + a_{\mathrm{t,\ cha}} \frac{\eta^{\mathrm{ch}} \Delta t}{Q_{\text{storage}}} \boldsymbol{P}_{\text{storage}}^{\mathrm{ch}}(t) + a_{\mathrm{t,\ disc}} \frac{\Delta t}{\eta^{\mathrm{disc}} Q_{\text{storage}}} \boldsymbol{P}_{\text{storage}}^{\mathrm{disc}}(t) \tag{14}$$

where $\boldsymbol{P}_{\text{storage}}^{\mathrm{ch}}(t)$ and $\boldsymbol{P}_{\text{storage}}^{\mathrm{dis}}(t)$ are the charging power and the discharging power, $\boldsymbol{P}_{\text{storage}}^{\mathrm{ch}}(t) \geq 0$ and $\boldsymbol{P}_{\text{storage}}^{\mathrm{disc}}(t) \leq 0$; η^{dis} and η^{ch} are the discharging and charging efficiency of the energy storage; Δt is the length of the time step and is the capacity of the energy storage. The charging or discharging situation depends on (10). Charging or discharging is not performed at the same time, such as $a_{\mathrm{t,\ cha}} \wedge a_{\mathrm{t,\ disc}} = 0$. The energy storage model is subject to the following constraints:

$$\mathrm{SOC}^{\min} \leq \mathrm{SOC}(t+1) \leq \mathrm{SOC}^{\max} \tag{15}$$

The energy storage model parameters are provided in Table I.

TABLE I
ENERGY STORAGE MODEL PARAMETERS

Parameter	Value
Battery type	NCA
Rated voltage (V)	46.7
Battery capacity (Ah)	150
Charging and discharging efficiencies	90%

D. Implementation Details

The main advantage of RL or DRL is that the model can learn from the default environment and can adapt to the dynamic environment. After the model is completely trained using the off-line database, it can be exploited on-line in a real environment. A framework for implementation of the algorithm applied in decision-making and control is illustrated in Fig. 4.

There are two parts in the framework: training and execution. The training part is the main content of this research. The training part oversees learning the knowledge and the execution part and puts the learned knowledge into practice to make optimized decisions in a real physical environment. If there are emergencies, the agents will interact with the new environment. Through the adjusting of actions, the agent

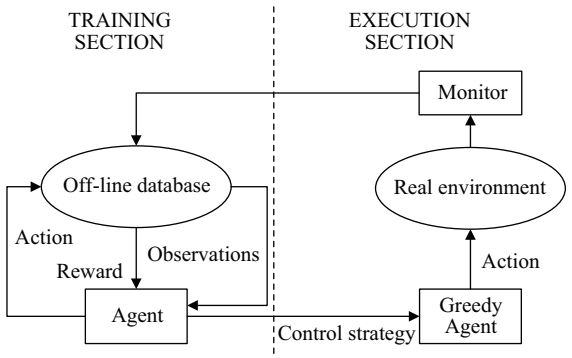

Fig. 4. Framework for RL in the home energy management optimization decision and control.

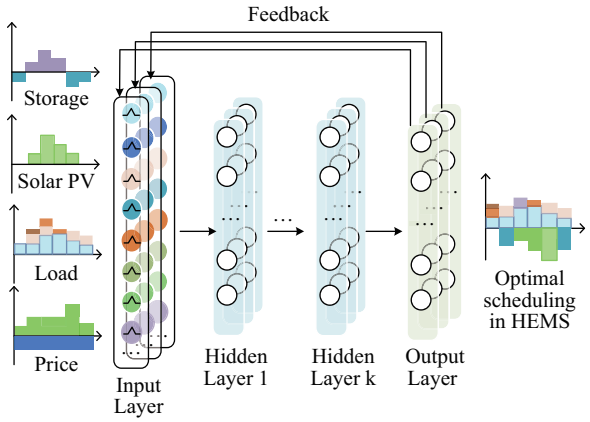

Fig. 5. The general architecture of a home energy management optimization strategy based on deep reinforcement learning.

gradually increases the obtained reward, and restores the optimization effect.

As mentioned above, agent, environment, reward, and action are the four basic parts in RL. In addition, the details of the algorithm implementation would be explained according to these four parts. The general architecture is shown in Fig. 5.

1) Agent

HEMOS of HEMS is regarded as an agent and learns from the environment to choose actions in order to maximize future rewards. Depending on the existing home automation systems, the agent is used to realize the switching or adjustment of the device (the charging and discharging for energy storage). In the algorithm design described above, the optimization decisions highly depend on the environment and rewards.

2) Environment

The environment refers to dynamic energy consumption, production, and electricity price. The data was accumulated for a period of one year and collected every 15 minutes from one house. The electric car, heater, and dishwasher are considered as time-shift loads; the former one can be interrupted while the latter two cannot be interrupted until the end of usage. The air-conditioner is considered as a power-shift load. Energy storage built into the model is involved in the home energy management system as a controllable load, broadening the possibilities of this scenario. PV generation is preferred for local usage and regarded as a non-curtailable resource.

The observations, including total load and each controllable appliance load, PV generation, SOC and electricity price, are regarded as the inputs to the neural networks at every time interval. The fixed electricity price and random changing price are set as different inputs and the results are compared in order to verify the generalization performance of the algorithm. All the observations of the agent are shown on the left side of Fig. 5. The Q-value for each discretized action is represented by the outputs of the neural networks, causing a changing amount of states to maximize rewards with continuous feedback. All the actions of the agent are shown on the right side of Fig. 5.

3) Reward

The rewards are the key to RL. The agent can be guided by a reasonable value function to advance in the "right direction." The setting of the value function in this study is primarily based on (6) and (7). In addition, the PV's priority in local consumption and equipment usage satisfaction are incorporated into the value function. The objective of the strategy is that the more rewards the agent obtains, the higher the benefit from real-world energy optimization.

This scenario can be summarized as a multi-objective optimization problem. Multiple-task reward is adopted to solve this problem with five components, which are shown in (16), (17), (18), (19) and (20). The five joint reward components are summed up to perform this multiple-task optimized scheduling.

In (16), C is the total cost derived from (6) and ξ_1 is the weighting factor for total rewards.

$$r_1 = \xi_1 C \tag{16}$$

For energy storage, the discharging action will be rewarded when the electricity price of the public power grid is high. Conversely, the charging of energy storage will receive a reward when the electricity price of the public power grid is low, which is shown in (17). In (17), p'_{disc} is the electric price at the discharging time, p'_{nor} is the electric price in normal time (here we set it at 0.07\$/kWh), ξ_2 and ξ_3 are the weighting factors which contribute to the total rewards. Since the life cycle is related to the number of charging or discharging, a small amount of negative reward would be obtained by performing charging and discharging actions in this case in order to reduce the times of energy storage usage, which is shown in (18). In (18), n is the number of actions in one day, ξ_4 and ξ_5 are the weighting factors.

$$r_2 = \begin{cases} -\xi_2 & \text{if } p'_{\text{disc}} < p'_{\text{nor}} \\ \xi_3 & \text{if } p'_{\text{disc}} > p'_{\text{nor}} \end{cases} \tag{17}$$

$$r_3 = \begin{cases} -\xi_4 n & \text{if } n < 10 \\ -\xi_5 n & \text{if } n \geq 10 \end{cases} \tag{18}$$

For solar PV, the priority is local consumption. Rewards will be awarded if the energy is generated by solar photovoltaics and locally consumed (including the storing of renewable energy into local storage), which is shown in (19). In (19), P_{use} is the local consumption from solar PV, P_{sell} is the selling back to the power company, ξ_6 and ξ_7 are the weighting factors.

$$r_4 = \xi_6 P_{\text{use}} + (-\xi_7 P_{\text{sell}}) \quad P_{\text{pv}} = P_{\text{use}} + P_{\text{sell}} \tag{19}$$

For controllable appliances satisfaction, each appliance is divided into 4 to 6 usage intervals per day according to the historical statistics. The interval not only has a requirement for usage amount but also is a finite delay time window for shifting load. Each usage interval needs to meet the satisfaction amount as a check. The shifting of the load is also performed within the finite delay time interval. The historical statistical usage of each load interval is taken as the observation of the agent. Moreover, a negative reward will be given if the relevant load does not pass the satisfaction check within the interval. In (20), P_{d} is the actual consumption of appliance $d, P_{\mathrm{T}i}$ is the satisfied consumption amount from historical statistics

$$r_5 = \begin{cases} -\xi_8 & \text{if } \sum_t P_{\mathrm{d}} < P_{\mathrm{T}i} \\ \xi_9 & \text{if } \sum_t P_{\mathrm{d}} \geq P_{\mathrm{T}i} \end{cases} \tag{20}$$

Unlike the traditional method of setting the dissatisfied function, the proposed method encapsulates consumers' satisfaction into the rewards. Therefore, the obtained rewards can reflect the satisfaction situation. Through continuous interaction with the environment, the agent gradually learns the user's electricity consumption habits and therefore tends to meet user needs when scheduling appliances. Unlike the fixed dissatisfied function, the proposed method is more adaptable to the dynamic environment.

The setting of reward in this study is a framework and a basic principle. The reward settings are given in the following example, verifying the rationality of the structure. For the value setting of the rewards, different emphases would lead to quantitative differences, showing the diverse effects.

4) Action

The output of the neural network is the Q-value combining action, which is the on or off status of the time-shift load, the consumption adjustment of the power-shift load, and the charging or discharging of energy storage. The actions are subject to power balance, the physical constraint of devices, demand satisfaction, and other constraints.

IV. Results and Discussion

In this section, the performance of the applied methods in the experimental environment is analyzed and validated by unifying the real-world data and energy storage models. Some comparisons are performed and their results are discussed.

Our simulation environment is Python 3.6 on a laptop with 8.0 GB RAM, GTX 1060, and an Intel i7. For the neural network, three hidden layers (100 neurons, 512 neurons, and 50 neurons) are built. Two rectifier linear units (ReLU) and one Sigmoid are applied as activation functions in the neural networks.

The whole off-line training process of DDQN for 1000 episodes is almost 110 minutes to 130 minutes. The computational efficiency of the DQN is better than DDQN in our simulation, which is almost 100 minutes to 120 minutes. The system parameters are provided in Table II

Most the components in the environment are provided by the dataset. We made a summary of the adopted data in the dataset to describe the technical characteristics of the components in Table III. The electric price adopted is listed in Table IV.

TABLE II
System Parameters

Parameter	Value	
Episode	1000	
Learning rate	0.001	
Discount rate	0.95	
Exploration rate	Exploration Max	1
	Exploration Min	0.1
	Decay Rate	0.995

TABLE III
Technical Characteristics of the Controllable Components

Load type	Electric appliance	Maximum power (kWh)	Average daily usage time (hour)
Power-shift	Air conditioner	1.64	2.34
Time-shift	Heater	1.08	1.89
	Electric car	3.35	1.64
	Dishwasher	5.12	0.59

TABLE IV
The Electric Price Plan Situation

Time	Purchase price from power company ($/kWh)	Sell price from Solar PV generation ($/kWh)
6:00–14:00	0.07	0.04
14:00–22:00	0.12	0.04
22:00–6:00 (next day)	0.02	0.04

A. The Learning Process and Overall Trends

Due to the characteristics of the algorithm, the agent learns to gradually adapt to the environment and obtains more rewards. There are a lot of random choices at the beginning; after many iterations, the agent learns to choose the converging trend and possibilities that are close to the optimization objective. Both the DQN and DDQN have achieved favorable training results, as shown in Fig. 6.

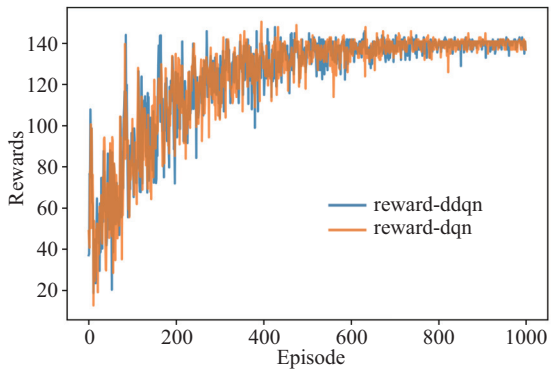

Fig. 6. The rewards training process of two algorithms.

Simultaneously, the cost of electricity consumption for users is reduced as shown in Fig. 7. As can be seen from Fig. 7, the costs did not converge as good as the rewards. The reason is that the proposed method adopts days as the training episodes ("time steps" in Fig. 7) and the different days have significant differentiation in adjustable space and base costs.

The user can obtain economic benefits through the control of the heater, air conditioner, dishwasher, and electric vehicle, and the charging and discharging of the household energy storage. Both the applied algorithms are effective, which can be seen from Figs. 6 and 7. The overall load curve tends to be smooth

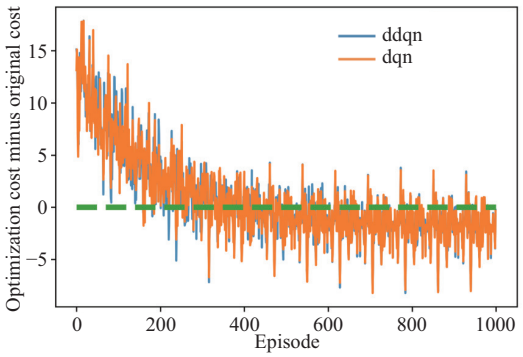

Fig. 7. With the training process, the cost is gradually reduced.

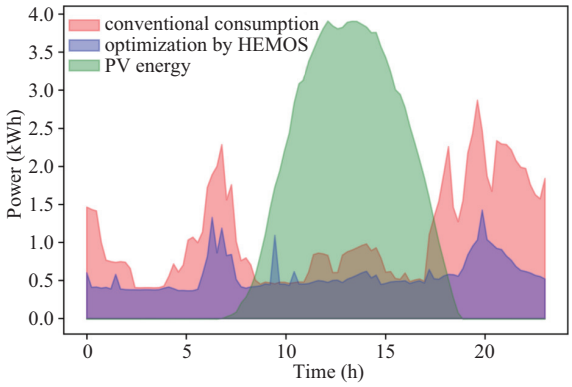

Fig. 8. Optimization results using DDQN (one typical day selected).

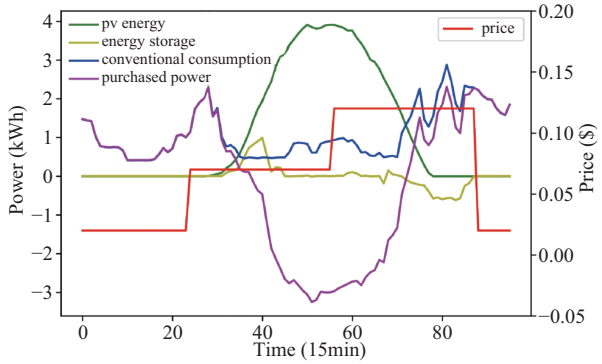

Fig. 9. Facing the dynamic fluctuation of electricity price, PV generation and electrical energy consumption, the operation of energy storage after learning by HEMOS (typical day selected).

and the load peak is reduced as shown in Fig. 8.

The household energy storage is taken as an example to describe the dynamic process of optimization shown in Fig. 9. Since the number of charging and discharging is related to the equipment life, a negative reward would be given when charging or discharging. The energy storage would capture the environmental changes and start charging to store the energy when the electricity price is low and the PV energy generation rises. In addition, the energy storage discharges when the electricity price is high and the electricity consumption almost meets the peak in one day.

B. Numerical Results and Comparison

After training by our applied algorithm, the agent can adapt to the changing environment and complete the optimization problem of high-dimensional large-scale data both in DQN and DDQN. The results for the cost minimization problem are summarized in Table V.

TABLE V
DAILY COST MINIMIZATION RESULTS FOR 50 DAYS WITH 15 MINUTES RESOLUTION USING DDQN AND DQN

Description	Algorithm	Fixed price		Changing price	
		Mean	St.dev.	Mean	St.dev.
Cost ($/day)		2.956	2.492	3.807	2.703
Minimized Cost ($/day)	DDQN	1.211	2.535	2.072	2.854
	DQN	1.217	2.615	2.081	2.872

As can be seen in Table V and Figs. 6 and 7, DDQN and DQN can handle optimization problems in HEMS; the optimization effects are very close. For random changing electrical price signals, DDQN shows more advantages [14] is the representative work in the HEMS area. By using the proposed load control scheme in [14], the user's average daily payment decreases by 25%. Our proposed method shows better optimization results than [14], which is shown in Table V.

C. Generalization and Reward Setting Analysis

During the implementation of the applied algorithm, there are two things (the generalization and the reward setting) that need to be discussed and analyzed in detail.

1) Generalization

Generalization has always been a difficulty in the current DRL algorithm. Although agents can carry out complex tasks after training, it is difficult for them to transfer their knowledge or experience to a new environment. This needs to be analyzed and discussed in the home energy management area. In this research, two new environments of the home energy system are designed to analyze the generalization by training and testing agents. As a trend of future transportation, electric vehicles are becoming more and more popular. However, not all the charging locations of the electric vehicle are within the control range of HEMS, which without electric vehicles is an essential scenario. Since not every house roof is suitable for installing solar photovoltaic power generation equipment, HEMS without solar PV should be a common scenario.

The real-world open source data set recorded by Pecan Street is adopted in this research. It is assumed that removing some parts of the data can have a negligible impact on the basic load pattern. The scenarios between normal (without PV and without EV) are tested with the same house data in order to highlight the comparisons. In addition, the scenarios are built by removing some data and keeping the same basic load pattern. Finally, our method is applied to the dataset corresponding to the different scenarios to estimate the generalization. Two layers of Dropout in the multiple layers of the neural network are set in order to improve the generalization. Dropout refers to the temporary discarding of neural network units according to a certain probability during the training of deep learning. The dropout can effectively be altered through the occurrence of over-fitting and achieve regularization to a certain extent; there is a detailed introduction in reference [39].

Table VI demonstrates that the algorithm applied is equally applicable to home energy management in the scenarios

TABLE VI
DAILY COST MINIMIZATION RESULTS FOR 50 DAYS WITH 15 MINUTES RESOLUTION BY DDQN IN 3 DIFFERENT SITUATIONS

Situation		Mean	St.dev.
Normal	Original cost ($/day)	2.956	2.492
	Minimized cost ($/day)	1.211	2.535
Without PV	Original cost ($/day)	8.392	1.951
	Minimized cost ($/day)	6.923	1.747
Without EV	Original cost ($/day)	0.550	2.078
	Minimized cost ($/day)	−0.106	2.179

without PV and EV. PV is the distributed energy, which can reduce dependence on external energy and create the possibility of profitability. Compared with other scenarios, the cost of the without PV is more than in other scenarios. EV is a high energy-consuming device, which accounts for a large proportion of energy in the whole environment. Compared with other scenarios, the cost of the without EV is less than in other scenarios.

2) Reward Setting

The reward is the core concept in RL and guides agents to learn in the environment. There may be a bad effect on the algorithm implementation if the main objectives cannot be accurately reflected by the design of the reward rules.

Four aspects are being considered the most in this study: the economic interests, clean energy usage, satisfaction of household appliances usage, and the life of household electrical devices. Following the requirement in this research, the reward settings include 4 aspects. The first one is directly related to (6) and (7). The second one is the priority in solar PV local consumption. The third one is to meet the usage demand of users. The fourth one is to reduce usage times of controllable electrical appliances, especially for energy storage.

Optimization results are influenced by reward design. Therefore, the reward setting of energy storage is taken as an example to illustrate.

As can be seen from Fig. 10, rewards of −0.5, −0.2, and 0 are given, respectively, when the energy storage is being operated every time. The knowledge learned by the agent is not the same in the three different cases.

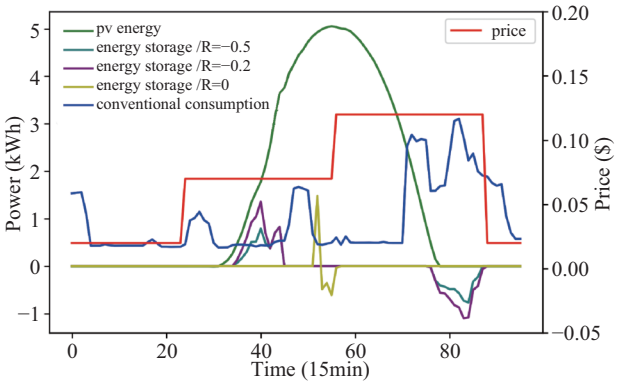

Fig. 10. The effect of different reward settings on energy storage (typical day selected).

It can be seen from the comparisons of reward = −0.5 and reward = −0.2 that the greater the punishment, the more conservative the action of the agent. The charging action follows the solar PV generation; the discharge time is concentrated at the peak time of daily power consumption. The agent concentrates on the action during the interval of the electricity price changing when no penalty is given (rewards = 0).

The effect of using energy storage on reducing electricity costs is illustrated in Table VII. Different reward settings can lead to different results each time the accumulator is operated. Reward = −0.2 is the most suitable case in our test. For different scenarios, the reward setting needs to be comprehensively discussed in terms of various factors. However, it is difficult to sum up a consensus and generally recognize the setting method.

TABLE VII
DAILY COST MINIMIZATION RESULTS FOR 50 DAYS WITH 15 MINUTES RESOLUTION BY DDQN IN 3 DIFFERENT REWARDS SETTINGS

Reward setting of energy storage	Minimized cost ($/day)	Reduced percentage
−0.5	1.720	41.8%
−0.2	1.211	59.0%
0	34.871	−1079.54%

D. Comparison with Traditional Algorithms

Particle Swarm Optimization (PSO) is widely used in solving the optimization problem as a traditional algorithm. PSO can find the optimal result in the solution space by simulating the group cooperative behavior [40]. The logic and principle of PSO are quite different from the DRL algorithm, which is not discussed in detail in this paper.

The results of the proposed DRL algorithm are compared with those of PSO to validate its performance. As can be seen from Fig. 11, PSO can obtain convergence at the beginning. In contrast, the DDQN converges after many iterations. The optimization results of the two algorithms are very close. Compared to PSO, DDQN is more effective in our case, as shown in Table VIII.

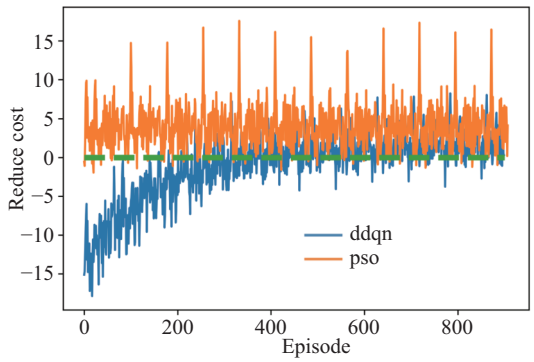

Fig. 11. The comparison of DDQN and PSO in training processes.

TABLE VIII
DAILY COST MINIMIZATION RESULTS FOR 50 DAYS WITH 15 MINUTES RESOLUTION USING DDQN AND PSO

Algorithm	Minimized cost ($/day)	St.dev.
DDQN	1.211	2.535
PSO	1.693	2.126

It should be highlighted that the PSO relies on the fixed environmental model while the DRL is a continuously data-

driven algorithm Theoretically, the DRL has more advantages in the actual dynamical environments.

V. Conclusion and Future Work

In this paper, the deep Q-learning and double deep Q-learning methods of reinforcement learning are applied in decision-making support for home energy management optimization strategies.

A large amount of multi-dimensional real-world data and a household energy storage model are used to validate the applied algorithm. In addition to the electricity price signal in the real world, random dynamic pricing is designed to test and verify the ability of agents responding to dynamic changes.

Experimental results illustrate that deep Q-learning and double deep Q-learning are effective in forming the strategies of home energy management. It shows that double deep Q-learning is more suitable for scheduling of energy resources compared to deep Q-learning both in stability and effectiveness. The applied method has the generalization in HEMS fields such as the scenarios without the PV or without the EV. It was analyzed that the reward setting should be adjusted according to actual situations. In our case, the results of DDQN are compared with those of PSO to validate the performance of the algorithm applied. DDQN shows more advantages both in practice and in theory.

However, further research is still needed. Our optimization is based on globally observable conditions and does not always exist in the real world. In many cases, since the algorithm is operational under partial observable conditions due to constraints, the algorithm and implementation strategy should be discussed and analyzed. Thus, special consideration should be given to the analysis of the comparison between the model-driven methods and data-driven optimization methods because they have advantages in different scenarios.

References

[1] B. Zhou, W. T. Li, K. W. Chan, Y. J. Cao, Y. H. Kuang, X. Liu, and X. Wang, "Smart home energy management systems: Concept, configurations, and scheduling strategies," *Renewable and Sustainable Energy Reviews*, vol. 61, pp. 30–40, Aug. 2016.

[2] B. L. R. Stojkoska and K. V. Trivodaliev, "A review of Internet of Things for smart home: challenges and solutions," *Journal of Cleaner Production*, vol. 140, pp. 1454-1464, Jan. 2017.

[3] M. Shakeri, M. Shayestegan, H. Abunima, S. M. Salim Reza, M. Akhtaruzzaman, A. R. M. Alamoud, K. Sopian, and N. Amin, "An intelligent system architecture in Home Energy Management Systems (HEMS) for efficient demand response in smart grid," *Energy and Buildings*, vol. 138, pp. 154–164, Mar. 2017.

[4] B. P. Esther and K. S Kumar, "A survey on residential demand side management architecture, approaches, optimization models and methods," *Renewable and Sustainable Energy Reviews*, vol. 59, pp. 342–351, Jan. 2016.

[5] X. Jin, K. Baker, D. Christensen, and S. Isley, "Foresee: a user-centric home energy management system for energy efficiency and demand response," *Applied Energy*, vol. 205, pp. 1583–1595, Nov. 2017.

[6] P. Chavali, P. Yang, and A. Nehorai, "A distributed algorithm of appliance scheduling for home energy management system," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 282–290, Jan. 2014.

[7] M. R. Alam, M. St-Hilaire, and T. Kunz, "Computational methods for residential energy cost optimization in smart grids: a survey," *ACM Computing Surveys*, vol. 49, no. 1, pp. 2, Apr. 2016.

[8] E. Loukarakis, C. J. Dent, and J. W. Bialek, "Decentralized multi-period economic dispatch for real-time flexible demand management," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 672–684, Jan. 2016.

[9] A. H. Mohsenian-Rad and A. Leon-Garcia, "Optimal residential load control with price prediction in real-time electricity pricing environments," *IEEE Transactions on Smart Grid*, vol. 1, no. 2, pp. 120–133, Sep. 2010.

[10] N. G. Paterakis, O. Erdinç, I. N. Pappi, A. G. Bakirtzis, and J. P. S. Catalão, "Coordinated operation of a neighborhood of smart households comprising electric vehicles, energy storage and distributed generation," *IEEE Transactions on Smart Grid*, vol. 7, no. 6, pp. 2736–2747, Nov. 2016.

[11] W. X. Li, T. Logenthiran, V. T. Phan, and W. L. Woo, "Implemented IoT-based Self-Learning Home Management System (SHMS) for Singapore," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2212–2219, Jan. 2018.

[12] Z. G. Pan, H. B. Sun, and Q. L. Guo, "TOU-based optimal energy management for smart home," in *Proceedings of IEEE PES ISGT Europe 2013*, 2013, pp. 1–5.

[13] Z. M. Wang, C. H. Gu, F. R. Li, P. Bale, and H. B. Sun, "Active demand response using shared energy storage for household energy management," *IEEE Transactions on Smart Grid*, vol. 4, no. 4, pp. 1888–1897, Dec. 2013.

[14] Z. G. Pan, Q. L. Guo, and H. B. Sun, "A dynamic soft constraint method for thermostatically controlled appliances scheduling," in *Proceedings of 2016 IEEE Power and Energy Society General Meeting*, 2016, pp. 1–5.

[15] R. Mehta, D. Srinivasan, and P. Verma, "Intelligent appliance control algorithm for optimizing user energy demand in smart homes," in *Proceedings of 2017 IEEE Congress on Evolutionary Computation*, 2017, pp. 1255–1262.

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning*: *An Introduction*, 2nd ed., New York: MIT Press, 2018.

[17] M. Glavic, R. Fonteneau, and D. Ernst, "Reinforcement learning for electric power system decision and control: past considerations and perspectives," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 6918–6927, Jul. 2017.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," $Nature$, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[19] D. O'Neill, M. Levorato, A. Goldsmith, and U. Mitra, "Residential demand response using reinforcement learning," in *Proceedings of the 2010 First IEEE International Conference on Smart Grid Communications*, 2010, pp. 409–414.

[20] F. Ruelens, B. J. Claessens, S. Vandael, S. Iacovella, P. Vingerhoets, and R. Belmans, "Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning," in *Proceedings of 2014 Power Systems Computation Conference*, 2014, pp. 1–7.

[21] H. Berlink and A. H. R. Costa, "Batch reinforcement learning for smart home energy management," in *Proceedings of the 24th International Conference on Artificial Intelligence*, 2015, pp. 2561–2567.

[22] T. S. Wei, Y. Z. Wang, and Q. Zhu, "Deep reinforcement learning for building HVAC control," in *Proceedings of the 54th Annual Design Automation Conference 2017*, 2017, pp. 22.

[23] Z. A. Zhang, A. Chong, Y. Q. Pan, C. L. Zhang, S. L. Lu, and K. P. Lam, "A deep reinforcement learning approach to using whole building energy model for HVAC optimal control," in *Proceedings of 2018 Building Performance Modeling Conference and SimBuild*, 2018, pp. 675–682.

[24] M. N. H. Nguyen, T. Le Pham, N. H. Tran, and C. S. Hong, "Deep reinforcement learning based smart building energy management," in *Proceedings of South Korea software comprehensive academic conference, 2017*, pp. 871–873.

[25] J. R. Vázquez-Canteli, S. Ulyanin, J. Kämpf, and Z. Nagy, "Fusing TensorFlow with building energy simulation for intelligent energy management in smart cities," *Sustainable Cities and Society*, vol. 45, pp. 243–257, Feb. 2019.

[26] Z. Q. Wan, H. P. Li, and H. B. He, "Residential energy management with deep reinforcement learning," in *Proceedings of 2018 International Joint Conference on Neural Networks*, 2018, pp. 1–7.

[27] V. François-Lavet, D. Taralla, D. Ernst, and R. Fonteneau, "Deep reinforcement learning solutions for energy microgrids management," in *Proceedings of European Workshop on Reinforcement Learning*, 2016.

[28] M. Mohammadi, A. Al-Fuqaha, M. Guizani, and J. S. Oh, "Semisupervised deep reinforcement learning in support of IoT and smart city services,"*IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 624–635, Apr. 2018.

[29] B. V. Mbuwir, F. Ruelens, F. Spiessens, and G. Deconinck, "Battery energy management in a microgrid using batch reinforcement learning," *Energies*, vol. 10, no. 11, pp. 1846, Nov. 2017.
[30] F. Ruelens, B. J. Claessens, P. Vrancx, F. Spiessens, and G. Deconinck, "Direct load control of thermostatically controlled loads based on sparse observations using deep reinforcement learning," *CSEE Journal of Power and Energy Systems*, vol. 5, no. 4, pp. 423–432, Dec. 2019.
[31] A. Prasad and I. Dusparic, "Multi-agent deep reinforcement learning for zero energy communities," arXiv: 1810.03679v1, 2018.
[32] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Slootweg, "On-line building energy optimization using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.
[33] D. X. Zhang, X. Q. Han, and C. Y. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, pp. 362–370, Sep. 2018.
[34] Z. Zhang, D. Zhang and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview,"*CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213–225, Mar. 2020.
[35] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning*, Cambridge: MIT Press, 1998.
[36] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.
[37] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double *Q*-learning," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 2094–2100.
[38] W. J. Cole, J. D. Rhodes, W. Gorman, K. X. Perez, M. E. Webber, and T. F. Edgar, "Community-scale residential air conditioning control for effective grid management," *Applied Energy*, vol. 130: 428–436, Oct. 2014.
[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[40] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-International Conference on Neural Networks*, 1995, pp. 1942–1948.

Yuankun Liu received the M.S. degree in Electrical Engineering from North China Electric Power University in 2017. He is currently pursuing the Ph.D. degree in Power System and its Automation with China Electric Power Research Institute. His current research interests are machine learning and smart grid.

Dongxia Zhang received the M.S. degree in Electrical Engineering from the Taiyuan University of Technology, Taiyuan, Shanxi, China, in 1992 and her Ph.D. degree in Electrical Engineering from Tsinghua University, Beijing, China, in 1999. From 1992 to 1995, she was a Lecturer with Taiyuan University of Technology. Since 1999, she has been working at China Electric Power Research Institute. She is the co-author of four books, and more than 40 articles. Her research interests include power system analysis and planning, big data and AI applications in power systems. She is an Associate Editor of Proceedings of the CSEE.

Hoay Beng Gooi received the B.S. degree in Electrical Engineering from National Taiwan University, Taiwan, China, in 1978; the M.S. degree in Power Engineering from the University of New Brunswick, Fredericton, NB, Canada, in 1980; and the Ph.D. degree in Power Engineering from the Ohio State University, Columbus, OH, USA, in 1983. From 1983 to 1985, he was an Assistant Professor with the Department of Electrical Engineering, Lafayette College, Easton, PA, USA. From 1985 to 1991, he was a Senior Engineer with Empros (now Siemens), Minneapolis, MN, USA, where he was responsible for the design and testing coordination of domestic and international energy management system projects. In 1991, he joined the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, as a Senior Lecturer, where he has been an Associate Professor since 1999 and the Deputy Head of Power Engineering Division from 2008 to 2014. Starting 2020, he serves as a Co-Director of Singapore Power-NTU Joint Lab. He has served as Associate Editor, IEEE Transactions on Power Systems since 2016. His current research interests include microgrid energy management systems, electricity markets, spinning reserve, energy storage, and renewable energy sources.