International Journal of Machine Learning and Cybernetics (2021) 12:1687–1703
https://doi.org/10.1007/s13042-020-01266-5

ORIGINAL ARTICLE

Deep reinforcement learning based home energy management system with devices operational dependencies

Caomingzhe Si[1] · Yuechuan Tao[2] · Jing Qiu[2] · Shuying Lai[2] · Junhua Zhao[1]

Received: 7 August 2020 / Accepted: 17 December 2020 / Published online: 24 January 2021

Abstract

Advanced metering infrastructure and bilateral communication technologies facilitate the development of the home energy management system in the smart home. In this paper, we propose an energy management strategy for controllable loads based on reinforcement learning (RL). First, based on the mathematical model, the Markov decision process of different types of home energy resources (HERs) is formulated. Then, two RL algorithms, i.e. deep Q-learning and deep deterministic policy gradient are utilized. Based on the living habits of the residents, the dependency modes for HERs are proposed and are integrated into the reinforcement learning algorithms. Through the case studies, it is verified that the proposed method can schedule HERs properly to satisfy the established dependency modes. The difference between the achieved result and the optimal solution is relatively small.

Keywords Home energy management system · Reinforcement learning · Dependency modes · Smart home

Abbreviations

HEMS	Home energy management system
HER	Home energy resource
EV	Electric vehicle
TCL	Thermostatically controlled load
RTP	Real time price
DCT	Demand charge tariff

Indices and sets

Θ^{NC}	Set of the non-interruptible appliance with constant power
Θ^{IC}	Set of the interruptible appliance with constant power
Θ^{ESS}	Set of the energy storage system
Θ^{TCL}	Set of the thermostatically controlled load
i	Appliance index
t	Time index

HEMS parameters

$\overline{E_i^{ESS}}\underline{E_i^{ESS}}$	Upward and downward energy storage limits
$E_i^{ESS,EXP}$	Expected energy storage state at the end of the day
$\overline{E_i^{EV}}$	Maximum capacity of the EVs
E_{i,t_arrive}^{EV}	Energy storage state of EVs when arriving
P_i^{rate}	Rated power of the constant power HERs
$\overline{P_i^{ESS,C}P_i^{ESS,D}}$	Charging and discharging power limits
P_t^{NS}	Energy consumption of the non-schedulable load
P_t^{PV}	Output of the PVs
R, C	Thermal resistance and heat ratio of air
t_i^s, t_i^e	Starting and ending time of the a
t_arrive	Arrival time of each EVs
T^{set}	Set indoor temperature
T^{min},T^{max}	Minimum and maximum indoor temperature
T_t^{out}	Outdoor temperature
WT_i	Working time of the HERs
W_i,D_i	Electricity consumption per unit distance and daily driving distance
η^C, η^D	Charging and discharging efficiency
λ_t^{RTP}	Real time electricity price

✉ Yuechuan Tao
tyc19950713@163.com

[1] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Longgang, Shenzhen 518172, China

[2] Australia School of Electrical and Information Engineering, The University of Sydney, Sydney, NSW 2006, Australia

λ^{DCT}	Demand charge tariff
φ	Weight co-efficiency
P'	Historical peak power recorded in the current billing cycle
δ	Temperature band

Thermal comfort model parameters

f^{cl}	Mean temperature of the outer surface of the clothed body
h^{c}	Heat transfer coefficient
I^{cl}	Thermal resistance of clothing
M	Metabolic rate
PMV_t	Predicted mean vote
PPD	Predicted percentage of dissatisfied
P^{v}	Vapor pressure in ambient air
rh	Relative air humidity
T^{a}	Indoor ambient air temperature
T^{mrt}	Mean radiant temperature
T^{cl}	Mean temperature of the outer surface of the clothed body
v^{ar}	Relative air velocity

HEMS variables

$E^{ESS}_{i,t}$	Energy storage state
$P^{ESS,C}_{i,t}$	Charging power of energy storage system
$P^{ESS,D}_{i,t}$	Discharging power of energy storage system
$P^{EV,CHA}_{i,m,t}$	Charging power of EVs
$P^{EV,DIS}_{i,m,t}$	Discharging power of EVs
P^{TCL}_{t}	Working power of TCLs
$\hat{P}_t$	Total energy consumption of the smart home
$P^{HER}_{i,t}$	Energy consumption of HERs
T^{in}_{t}	Indoor temperature
$t*$	Time when the HERs are turned on
$\delta_{i,t}$	Operation state of the HERs

1 Introduction

With the popularization of advanced metering infrastructure (AMI), home energy resources (HERs) become more capable to response the market signals, such as real-time price (RTP). In a smart home with an advanced home energy management system (HEMS), the HERs can be scheduled properly and the electricity bill can be reduced through active demand response (DR). On the other hand, the reshaping of the residents' energy consumption profile is beneficial to utility grids in terms of smoothing renewable energy output, clipping the peak load, reducing generation backup, and so on.

In recent years, with the development of automation devices and wireless two-way communication technology, HEMS becomes an essential part of DR. In the literature, the HEMS is usually based on an optimization algorithm. In the objective function, either time of use (TOU) pricing [1] or RTP [2, 3] can be considered. Reference [4] put forward a HEMS which optimally schedules the HERS through jointly considering RTP and the monthly basis peak power consumption penalty. The optimization is solved through Natural Aggregation Algorithm. Reference [5] proposed a multi-stage HEMS in a high rooftop PV penetrated environment. The proposed method contains three stages: forecasting, day-ahead scheduling, and actual operation. Reference [6] proposed a stochastic dynamic programming framework for the smart home with plug-in electric vehicles (PEV). The electricity cost is minimized while the charging demand requirement is satisfied. Apart from the energy consumption requirement, the comfort of the users was considered in the literature [7]. In Ref. [8] both energy cost and thermal discomfort cost are fully considered in a long-term horizon with heating, ventilation, and air conditioning load. In Ref. [9] a holistic model is proposed to center the preference of the users based on mixed-integer linear programming. In Refs. [10–12], a demand-side recommender system is proposed which provides a personalized recommendation to the residents. Besides the optimization method, the rule-based scheduling has been used for behavioral application by specifying conditions [13–16]. However, the rule-based algorithm shows difficulties in dealing with large data, especially in real-time control [17].

Artificial intelligence technologies have been utilized for HEMS. In Ref. [18] a distributed algorithm-based ANN was proposed to obtain accurate energy management decisions. Reference [19] proposed the demand response achieved by HEMS, aiming to obtain the optimal temperature scheduling where machine learning is integrated into the day-ahead electricity price and outdoor temperature forecasts. In Ref. [20] a supervised-learning-based strategy is put forward to substitute the conventional mathematical model to solve the office building energy management problem. In 2015, the AlphaGo developed by Google DeepMind proves that deep reinforcement learning (DRL) can solve the complex decision problem with massive space state [21]. Motivated by this remarkable milestone, the researchers began to focus on DRL in various researches areas. Reference [22] proposed a fully cooperative multi-agent reinforcement learning to solve the kinematic problem of redundant robots. Reference [23] proposed integration of data fusion and reinforcement learning techniques for the rank-aggregation problem. In Ref. [24], DRL is utilized in EVs charging navigation. In the energy management area, DRL becomes a promising tool in recent years. Reference [25] propose a multi-agent reinforcement learning framework for home energy management to achieve an efficient home-based demand response. To solve the hour-ahead scheduling problem, a finite Markov decision process with discrete time steps is formulated and the uncertainties such as PV output and electricity price are predicted through the extreme learning machine in a rolling time window. Reference [26] proposed an edge-cloud integrated solution for building demand response using DRL, where the learning is conducted at cloud infrastructure and the control is realized in the edge. With the edge-cloud integrated solution, a cost-effective automation system can be widely adopted for building demand response. Reference [27] aimed at minimizing the energy cost of heating, ventilation, and air-conditioner system in the multi-zone commercial building using DRL. The paper comprehensively considered the random zone occupancy, thermal comfort, and indoor air quality. Reference [28] proposed a real-time autonomous energy management strategy for a residential multi-energy system using a DRL based approach. The proposed approach is tailored to align with the nature of the problem by posing it in a multi-dimensional continuous state and action space. Reference [29] proposed a DRL based controller to manage the state of charge of a multi-energy storage system. The proposed controller is compared with benchmark DRL methods and other control technologies such as fuzzy logic and PID control. Hence, the application of DRL provides new opportunities for HEMS. Compared with conventional methods, DRL shows advantages in terms of short decision-making time and less dependence on the mathematical model.

In this paper, we proposed a HEMS based on DRL. The main contributions can be listed as follows:

First, based on the conventional mathematical model, the Markov decision process (MDP) of different types of HERs is formulated. The states, actions, transition functions, and reward functions are defined properly.

Second, DRL is applied in the HEMS problem. The HERs can be controlled in real-time rather than day-ahead scheduling. Two algorithms, deep Q-learning (DQN) and deep deterministic policy gradient (DDPG) are utilized respectively for comparison.

Third, based on the living habits of the residents, the dependency modes are put forward to satisfy the users' personalized settings. The dependency models are integrated into the DRL.

2 Mathematical model for HEMS

2.1 Classification of HERs

For a smart home containing N HERs which belongs to the set Θ, i.e. $|\Theta| = N$. Figure 1 shows the classification of HERs. According to the Fig. 1, the HERs are first classified into constant power appliances and inconstant power appliances. For the constant power appliances, they can only work at the rated power and they are further classified into interruptible and non-interruptible appliances. As for the interruptible appliances, they can be switched on and off such as the pool pump. But for the non-interruptible appliances, once they are switched on, they cannot be switched off before the task is finished, such as the rice cooker. For the inconstant power appliances, the power is flexible. Two typical inconstant power appliances are energy storage and thermostatically controlled load.

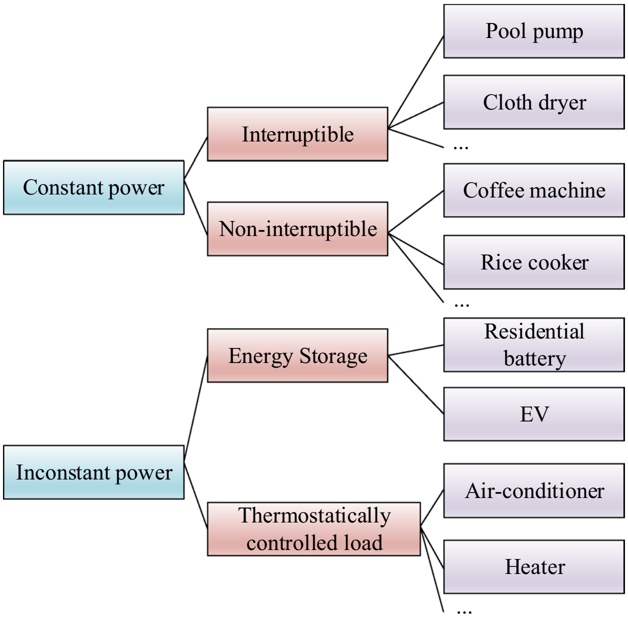


Fig. 1 Classification of the HERs

2.1.1 Non-interruptible appliance with constant power Θ^{NC}

In sub-set Θ^{NC}, the HERs working at the nominal power P_i^{rate}. The appliances are schedulable but cannot be interrupted once the machines start to work. The appliances should finish the assigned task within the pre-specified time range $\left(t_i^s, t_i^e\right)$. Typical HERs in this class includes rice cooker, printer, toaster, etc. The following operation constraints should be met:

$$\sum_{t=t*}^{t*+WT_i} \delta_{i,t}\Delta t \geq WT_i\left(\delta_{i,t+1} - \delta_{i,t}\right) \tag{1}$$

$$\delta_{i,t} \in \{0, 1\} \tag{2}$$

$$\delta_{i,t} = 0, t \notin \left(t_i^s, t_i^e\right) \tag{3}$$

where $t*$ represent the time when the HERs are turned on; WT_i represents the working time of the HERs; $\delta_{i,t}$ represents the operation state of the HERs (‘0’ indicates off and ‘1’ indicates on).

2.1.2 Interruptible appliance with constant power Θ^{IC}

In sub-set Θ^{IC}, the HERs working at the nominal power P_i^{rate}. However, the operation of the HERs can be interrupted and be resumed later. The appliances should finish the assigned task within the pre-specified time range $\left(t_i^s, t_i^e\right)$. Typical HERs in this class include dishwasher, pool pump, etc. The following operation constraints should be met:

$$\sum_{t=t_i^s}^{t_i^e} \delta_{i,t}\Delta t = WT_i \tag{4}$$

where $\delta_{i,t}$ represents the operation state of the HERs; WT_i represents the working time of the HERs.

2.1.3 *Energy storage system* Θ^{ESS}

In set Θ^{ESS}, HERs mainly includes two types of devices: residential battery and plugged-in electric vehicles (PEV). These two types of devices take responsibility for energy storage. The operating power is a continuous value in the range of $\left[-\overline{P_i^{ESS,D}}, \overline{P_i^{ESS,C}}\right]$, where $\overline{P_i^{ESS,D}}, \overline{P_i^{ESS,C}}$ donates the discharging and charging power of the appliances. The physical constraints for the devices in Θ^{ESS} can be described as:

$$E_{i,t+1}^{ESS} = E_{i,t}^{ESS} + P_{i,t}^{ESS,C}\eta^C\Delta t - P_{i,t}^{ESS,D}\Delta t/\eta^D \tag{5}$$

$$0 \le P_{i,t}^{ESS,C} \le \overline{P_i^{ESS,C}}, 0 \le P_{i,t}^{ESS,D} \le \overline{P_i^{ESS,D}} \tag{6}$$

$$\underline{E_i^{ESS}} \le E_{i,t+1}^{ESS} \le \overline{E_i^{ESS}} \tag{7}$$

$$E_{i,T}^{ESS} \ge E_i^{ESS,EXP} \tag{8}$$

where $E_{i,t}^{ESS}$ is the energy storage state; $P_{i,t}^{ESS,C}$ and $P_{i,t}^{ESS,D}$ are the charging and discharging power; η^C and η^D are the charging and discharging efficiency; $\overline{P_i^{ESS,C}}$ and $\overline{P_i^{ESS,D}}$ are the maximum charring and discharging power of the energy storage system;$\underline{E_i^{ESS}}$ and $\overline{E_i^{ESS}}$ is the upward and downward energy storage limits; $E_i^{ESS,EXP}$ is the expected energy storage state at the end of the day.

Equation (5) is the energy balance equations for ESS. Equation (6) is the charging and discharging power limits. Equation (7) is the energy storage state limits for security. Equation (8) ensures that the energy storage state of the ESS is not less than the pre-specified threshold at the end of the operation horizon. For the PEVs, the energy storage state at the end of the operation horizon should be able to support the next trip.

For PEV, the extra constraints regarding the driving behavior should be considered.

$$P_{i,t}^{EV,CHA} = 0, P_{i,t}^{EV,DIS} = 0 \,\text{When}\, t < t\_arrive \tag{9}$$

$$E_{i,t\_arrive}^{EV} = \overline{E_i^{EV}} - D_i \times W_i \tag{10}$$

where t_arrive the arrival time of each EVs; $P_{i,t}^{EV,CHA}$ and $P_{i,t}^{EV,DIS}$ are the charging and discharging power of EVs; E_{i,t_arrive}^{EV} is the energy storage state of EVs when arriving; $\overline{E_i^{EV}}$ is the maximum capacity of the EVs;D_i is the daily driving distance; W_i is the electricity consumption per unit distance.

2.1.4 Thermostatically controlled load Θ^{TCL}

The second sub-set contains the HERs which are interruptible appliances with power in the range of $\left[P_i^{\min}, P_i^{\max}\right]$. Typical HERs in this class are thermostatically controlled loads, such as air-conditioners, refrigerators, and heater. This type of HERs does not have a total energy consumption requirement, but the dissatisfaction of the users should be considered through evaluating the disutility function. In the mathematical model, we take a heating system as an example. The thermodynamic behavior for a heating system can be described as:

$$T_{t+1}^{in} = T_t^{in}e^{-1/RC} + \left(RP_t^{TCL} + T_t^{out}\right)\left(1 - e^{-1/RC}\right) \tag{11}$$

where T_t^{in} and T_t^{out} are the indoor and outdoor temperature; R refers to the thermal resistance; C refers to the heat ratio of air; P_t^{TCL} is the working power of TCLs.

Usually, an indoor temperature threshold will be given. When the acceptable temperature band is 2δ, the maximum and minimum indoor temperature can be described as:

$$T^{max} = T^{set} + \delta, T^{min} = T^{set} - \delta \tag{12}$$

However, the restriction of indoor temperature within a pre-specific range is not enough for a smart home. According to building environment science, residents’ thermal comfort can be modeled by considering indoor temperature, humidity, clothing condition, etc. In this paper, ISO 7730 Thermal Comfort Model is employed to evaluate the human’s thermal comfort degree through two indices: (1) predicted mean vote (PMV); and (2) predicted percentage of dissatisfied (PPD) [30]. PMV predicts the mean value of the votes of a large group of people on the ISO thermal sensation scale. According to PMV, the PPD is further calculated to reveal the percentage of people likely to feel uncomfortable.

The PMV can be calculated according to (13). The T^{cl} in (14) should be calculated according to (15) and (16). The mean temperature of the outer surface of the clothed body, f^{cl}, and vapor pressure in ambient air, P^v, are calculated according to (17).

$$PMV_t = \left[0.352 \cdot \exp(-0.042M) + 0.032\right] \cdot \begin{bmatrix} M - 0.35 \cdot (43 - 0.061M - P^v) - 0.42 \cdot (M - 50) - 0.0023 \cdot M \\ \cdot (44 - p^v) - 0.0014M \cdot (34 - T^a) - 3.4 \cdot 10^{-8} \cdot f^{cl} \cdot \left(\left(T^{cl} + 273\right)^4 - \left(T^{mrt} + 273\right)^4 \right) \\ -f^{cl} \cdot h^c \cdot \left(T^{cl} - T^a\right) \end{bmatrix} \tag{13}$$

where M represents metabolic rate (kcal/hr); P^v represents vapor pressure in ambient air (mmHg); T^a represents indoor ambient air temperature (°C); T^{cl} represents mean temperature of the outer surface of the clothed body (°C); T^{mrt} represents mean radiant temperature (°C); f^{cl} represents clothing surface area factor; h^c represents heat transfer coefficient (kcal/m^2 h °C).

$$T^{cl} = 35.7 - 0.032M - 0.18I^{cl} \cdot \left\{ 3.4 \cdot 10^{-8} \cdot f^{cl} \cdot \left[\left(T^{cl} + 273\right)^4 - \left(T^{mrt} + 273\right)^4 \right] \right\} + f^{cl} \cdot h^c \cdot \left(T^{cl} - T^{in}\right) \tag{14}$$

where I^{cl} thermal resistance of clothing (clo);

$$h^c = \begin{cases} 2.05 \cdot \left|T^{cl} - T^{in}\right|^{0.25} & , if\ 2.38\left|T^{cl} - T^{in}\right|^{0.25} > 10.4 \cdot \sqrt{v^{ar}} \\ 10.4 \cdot \sqrt{v^{ar}} & , if\ 2.38\left|T^{cl} - T^{in}\right|^{0.25} < 10.4 \cdot \sqrt{v^{ar}} \end{cases} \tag{15}$$

where v^{ar} represents relative air velocity (m/s).

$$f^{cl} = \begin{cases} 1 + 1290I^{cl} & , if\ I^{cl} \le 0.078 \\ 1.05 + 0.645I^{cl} & , if\ I^{cl} > 0.078 \end{cases} \tag{16}$$

$$P^v = rh \cdot 10 \cdot e^{(16.6536 - 4030.183)/(T^a + 273)} \tag{17}$$

where rh relative air humidity (%).

Then PPD is calculated as the function of PMV as (18).

$$PPD = 100 - 95 \cdot \exp\left(-0.03353 \cdot PMV^4 - 0.2179 \cdot PMV^2\right) \tag{18}$$

In the application, the calculation of PMV requires the sensing of v^{ar},rh,P^v,T^{mrt},T^a. In real-time control, the ISO 7730 model shows high requirements in sensors and information communications. Hence, the calculation of ISO 7730 model can be simplified as (19), where the parameters can be obtained through [31].

$$PMV_t = \alpha T_t^a + \beta P_t^v - \sigma \tag{19}$$

The parameters in the thermal comfort model include M (metabolic rate); T^a (indoor ambient air temperature); T^{mrt} (mean radiant temperature); I^{cl} (thermal resistance of clothing); v^{ar} (relative air velocity); rh (relative air humidity). Among these parameters, T^a, T^{mrt}, v^{ar} and rh can be measured by the advanced sensing equipment. The parameter M and I^{cl} can be estimated based on the historical data of the occupants. In the simulation, the temperature and humidity data are provided by Guangzhou Central Meteorological Observatory, China [30]. Based on the real data, Gaussian noise with different deviation are added to enrich the original data. As for the air velocity, it is assumed that it is less than 0.2 m/s. Thus, an air velocity profile is randomly generated in the range of [0, 0.2]. Similarly, Gaussian noise is added. The parameter M differs from different people and is set in the range of [80, 200] cal. As for I^{cl}, since the heating system in winter is considered, the residents wear sweaters with a thermal resistance of 0.25–0.36 Clo.

The other parameters such as: P^v (vapor pressure in ambient air); T^{cl} (mean temperature of the outer surface of the clothed body); f^{cl} (the mean temperature of the outer surface of the clothed body); h^c (heat transfer coefficient) can be calculated according to the other parameters in formula (14)–(17).

2.2 Home energy management model

The energy consumption of the smart home during the whole operation horizon can be expressed as vector: $\hat{\mathbf{P}} = \left[\hat{P}_1, \hat{P}_2, ..., \hat{P}_T\right]$. The energy consumption is calculated according (20), which contains three parts: the energy consumption of HERs, $\mathbf{P}_i^{HER} = \left[P_{i,1}^{HER}, P_{i,2}^{HER}, ..., P_{i,T}^{HER}\right]$, the energy consumption of non-schedulable devices, $\mathbf{P}_i^{NS} = \left[P_{i,1}^{NS}, P_{i,2}^{NS}, ..., P_{i,T}^{NS}\right]$ and the output of PVs, $\mathbf{P}^{PV} = \left[P_1^{PV}, P_2^{PV}, ..., P_T^{PV}\right]$.

$$\hat{P}_t = P_t^{NS} + \sum_{i \in \Theta} P_{i,t}^{HER} - P_t^{PV} \tag{20}$$

In the references, HEMS mainly focuses on minimizing the electricity bill based on TOU or RTP tariffs. Under this framework, the electricity bill can be reduced, and the peak load can be shifted to the off-peak pricing time slots. In recent years, a new type of tariff called demand charge tariff (DCT) is introduced to the HEMS. The DCT is charged according to the maximum power recorded during the billing cycle. Under the context of RTP and DCT, the objective function for HEMS can be derived according to (21).

$$\min F = \sum_t \left(\lambda_t^{RTP} \widehat{P}_t \Delta t \right) + \varphi \cdot \lambda^{DCT} \max \left[\left(\max \left(\widehat{\mathbf{P}} \right) - P' \right), 0 \right] \tag{21}$$

s.t. Equation (1)–(11)

where λ_t^{RTP} is the real-time electricity price; max $(\bullet)$ is the maximum function returning the maximum value in the vectors; P' is the historical peak power recorded in the current billing cycle; λ^{DCT} is the DCT charges; φ is the weight co-efficiency.

In the objective function, the energy consumption cost based on RTP and the DCT charges will both facilitate the peak shifting. However, these two cost terms will be conflicted in some cases. For example, being stimulated by the RTP, all the HERs are scheduled to work with low electricity prices to save electricity bills to the largest extent. But, under this circumstance, the residents will be charged by DCT. Hence, DCT is responsible for preventing the over-shifting of the demand which creates a new peak.

To be noted, in the objective function (21), the satisfaction of the residents has not been considered. Besides, the objective function is non-linear because of the DCT cost term, which brings difficulties in optimization.

3 Deep reinforcement learning

3.1 Deep Q learning

In the Q-learning algorithm, the agents choose the actions based on the Q-value in the Q-table. The Q-table is updated iteratively according to the following equation:

$$Q(s,a) = Q(s,a) + \beta \left(R + \gamma \max_{a'} Q\left(s', a'\right) - Q(s,a) \right) \tag{22}$$

where s, a represent the state and action; r represent the reward; β is the learning rate; γ is the discount factor.

However, in some problems, the state space is very large which results in insufficient memory for Q-table. Therefore, DQN is proposed to solve the high dimensional observation problems. In DQN, the mapping relationship between $Q_\pi(s,a)$ and $\langle s,a \rangle$ is expressed by a deep neural network. The target value of the deep neural network and the loss function can be expressed as:

$$y_j = \begin{cases} R_j & , \quad if\_end \ \ is \ \ true \\ R_j + \gamma \max_{a'} Q'\left(\phi\left(s'_j\right), a'_j, w' \right) & , \quad if\_end \ \ is \ \ false \end{cases} \tag{23}$$

$$L(w) = \frac{1}{M} \sum_{j=1}^{M} \left(y_j - Q\left(s_j, a_j | w\right) \right)^2 \tag{24}$$

where w represents the parameters in the deep neural network.

DQN is capable of solving discrete problems but shows difficulties in solving the continuous problems. In this paper, the actions for the non-interruptible appliance with constant power and the actions for the interruptible appliance with constant power are discrete. The actions for these appliances contain switching off and switching on. However, for the appliances with inconstant power, namely energy storage systems and TCLs, the actions are the working power which is continuous values. There are two solutions for the DQN algorithm to deal with the continuous action problem. In the first solution, the agent can sample a set of actions from the action space and then determine the action which can obtain the largest Q-value. This method converts the continuous action problem to a discrete action problem, which takes the advantage of efficiency. However, this method cannot guarantee optimal action. The second solution is to solve an optimization problem as: $a = \arg\max_a Q(s,a)$. The optimization problem can be solved through gradient ascend. However, this method causes a larger computation burden, because the agent needs to update the action a iteratively. In this paper, we have applied the first solution to solve continuous action problem when using DQN, which is the simplest and effective method.

3.2 Deep deterministic policy gradient

Compared with DQN, DDPG is more capable of solving the high dimension continuous action space problem. In the policy-based reinforcement learning algorithms, we focus on the direct learning of the policy rather than the action value and the state value. The policy $\pi_\theta(s,a)$ represent the probability to choose action under state s as (25), where θ is the parameter of the policy.

$$\pi_\theta(s,a) = P(a|s,\theta) \tag{25}$$

First, we aim to maximize the reward considering one step from the initial state as (26).

$$J_1(\theta) = V_{\pi\theta}\left(s_1\right) = E_{\pi\theta} \tag{26}$$

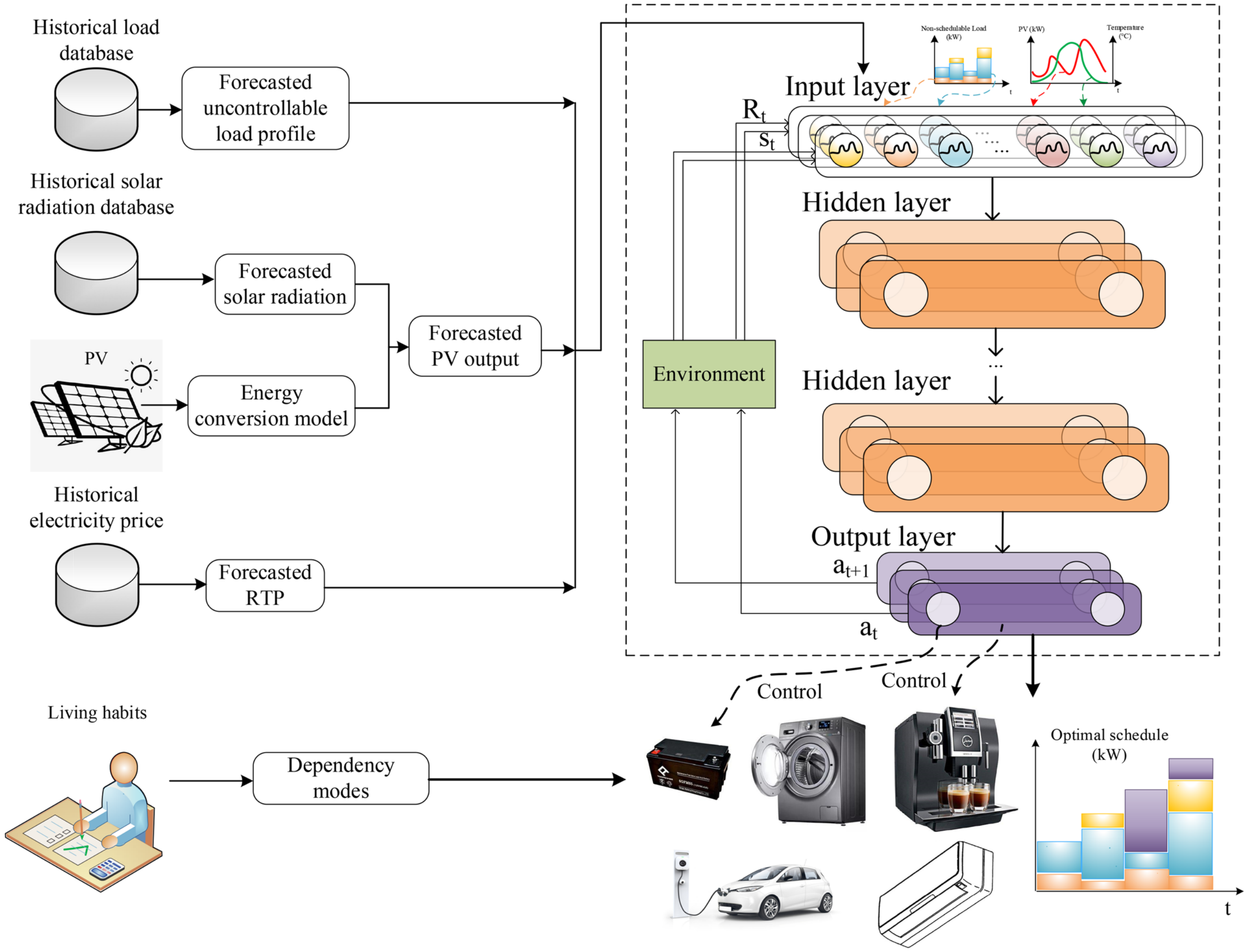


Fig. 2 Proposed DRL framework for HEMS

For the problems that do not have a specific initial state, but the initial state follows the distributions:$s \sim d_{\pi\theta}(s)$,the average value of the policy is defined as:

$$J_{avR}(\theta) = \sum_{s} d_{\pi\theta}(s) \sum_{a} \pi_\theta(s,a) R(a|s) \tag{27}$$

The gradient can be derived as:

$$\nabla_\theta J(\theta) = E_{\pi\theta}\left[\nabla_\theta \log \pi_\theta(s,a) R(s,a)\right] \tag{28}$$

The reward $R(s, a)$ in (28) can be obtained through Monte Carlo by calculating an average value. However, the space state may be very large in some problems. Thus, the Q value is introduced to estimate the value of the policy. The Q value is a mapping from the current action and state, which is expressed by a deep neural network known as a critic network. The structure of the critic network is similar to the DQN. Finally, the action is determined by Q value and the policy with the largest probability. The gradient can be further transferred to (29).

$$\nabla_\theta J(\theta) = E_{s\sim\rho^\pi}\left[\nabla_\theta \pi_\theta(s) \nabla_a Q_\pi(s,a)|_{a=\pi_\theta(s)}\right] \tag{29}$$

3.3 Implement details

The HEMS under the DRL framework is shown in Fig. 2. Based on the historical data, the environment variables, such as solar radiation and RTP, will be forecasted through the regression algorithm. These variables will be regarded as states for DRL. The other states in DRL will be further defined in the following section. The output of DRL is the actions that will be used to control the HERs. Furthermore, based on the user's living habits, the dependency modes of the devices will be established. The method to integrate the dependency modes to the DRL framework will be introduced in section IV. In this paper, we consider on-line learning for the HEMS. However, the reward is very low during the early training stage. Since the training process of reinforcement learning is through "trial and error". An inappropriate try may lead to an extremely bad result. To prevent the loss at the early stage of the training, an off-line pre-training can be conducted in a simulated environment before the on-line application. Therefore, before the on-line training, the parameters in neural networks have already been given proper initial value. The on-line training will finally converge through fine-tuning. In this way, the loss can be reduced to a large extent.

To implement the DRL in HEMS, the Markov decision process (MDP) should be defined for the specific problem through the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}r(.,.), \mathcal{R}(.,.) \rangle$, where $\mathcal{S}$ represents the state set; $\mathcal{A}$ represents the actions set; $\mathcal{P}r(.,.)$ represents the transition function; $\mathcal{R}(.,.)$ represents reward function.

3.3.1 Definition of state set

For the HERs, the operations will be affected by the observation of the real-time electricity price, λ_t^{RTP}, the output of PVs P_t^{PV}, and the non-schedulable loads, P_t^{NS}. These observations are usually considered as stochastic variables in the optimizations and should be regarded as states which are influence by the random exogenous data ζ_t in the environment.

According to the experience in traditional HEMS, the action of HERs not only depends on the current operation environment but also affected by the future operation environment, i.e. $\left[\lambda_{t+}^{RTP}, P_{t+}^{PV}, P_{t+}^{NS}\right]$. However, the future operation environment cannot be directly observed. Therefore, we use the long short-term memory (LSTM) algorithm to forecast the future home operation environment. The forecasted future home operation environment will be regarded as the states in the DRL context. Thus, the forecasting methodology of the home operation environment is embedded in the DRL. To be noted, the forecasting is not day-ahead but occurs every time interval of the decision in the term of rolling time window.

For the devices in Θ^{NC} and Θ^{IC}, the current working time,$\chi_{i,t}$, indicating whether the devices have finished the assigned task, should be considered as an important state. Thus, the state set can be expressed as:

$$\mathcal{S}^{\Theta^{NC}} or \mathcal{S}^{\Theta^{IC}} = \left[\lambda_{t+}^{RTP}, P_{t+}^{PV}, P_{t+}^{NS}, \chi_{i,t}\right] \tag{30}$$

For the devices in Θ^{TCL}, the total working time is not specified. However, the temperature should be considered as an essential state to make sure the satisfaction of the residents. Thus, the state set can be expressed as:

$$\mathcal{S}^{\Theta^{TCL}} = \left[\lambda_{t+}^{RTP}, P_{t+}^{PV}, P_{t+}^{NS}, T_t\right] \tag{31}$$

For the devices in Θ^{ESS}, the energy storage state should be considered as an essential state. Thus, the state set can be expressed as:

$$\mathcal{S}^{\Theta^{ESS}} = \left[\lambda_{t+}^{RTP}, P_{t+}^{PV}, P_{t+}^{NS}, E_{i,t}^{ESS}\right] \tag{32}$$

For the PEVs that also belongs to Θ^{ESS}, the connected state, indicating whether the PEVs have arrived home, should be considered as an additional state. Thus, the state set can be expressed as:

$$\mathcal{S}^{\Theta^{ESS},PEV} = \left[\lambda_{t+}^{RTP}, P_{t+}^{PV}, P_{t+}^{NS}, E_t^{EV}, \mu_t^{EV}\right] \tag{33}$$

3.3.2 Definition of action set

Based on the current state, the agents choose corresponding actions. The action alternatives for the devices are different. For the devices in Θ^{NC} and Θ^{IC}, the power is constant, and only “on” and “off” can be selected. In Θ^{NC}, the devices are non-interruptible, which means the agents cannot select “off” action once the devices are open. Hence, the action set is compact and discrete, which can be expressed as:

$$\mathcal{A}^{\Theta^{NC}} = \left[o^{on}\right] \tag{34}$$

$$\mathcal{A}^{\Theta^{IC}} = \left[o^{on}, o^{off}\right] \tag{35}$$

For the devices in Θ^{TCL} and Θ^{ESS}, the power is inconstant. Therefore, the action set is compact and continuous. which can be expressed as:

$$\mathcal{A}^{\Theta^{TCL}} = \left[P_{i,t}^{TCL}\right] \tag{36}$$

For the devices in Θ^{ESS}, the energy flow can be bidirectional, i.e. charging and discharging. The action set can be expressed as:

$$\mathcal{A}^{\Theta^{ESS}} = \left[P_{i,t}^{ESS,C}, P_{i,t}^{ESS,D}\right] \tag{37}$$

3.3.3 Definition of the transition function

The state transition from t to t + 1 is controlled partly by the action a_t and partly by the random exogenous data ζ_t, which can be expressed as:

$$\begin{cases} = f^T\left(\lambda_{t-}^{RTP}, P_{t-}^{PV}, P_{t-}^{NS}, \mu_t^{EV}, \zeta_t\right) \\ E_{i,t+1}^{ESS} = E_{i,t}^{ESS} + P_{i,t}^{ESS,C}\eta^C\Delta t - P_{i,t}^{ESS,D}\Delta t/\eta^D \\ T_{t+1}^{in} = T_t^{in}e^{-1/RC} + \left(RP_t^{TCL} + T_t^{out}\right)\left(1 - e^{-1/RC}\right) \end{cases} \tag{38}$$

3.3.4 Definition of the reward function

In this paper, the reward function is defined based on three components

Component I The first part is the electricity bill. Since the minimum cost is expected, the electricity bill based on RTP and DCT is considered as a negative reward in monetary.

$$r_1 = -\sum_t \left(\lambda_t^{RTP} \widehat{P}_t \Delta t\right) - \varphi \cdot \lambda^{DCT} \max\left[\left(\max\left(\widehat{\mathbf{P}}\right) - P'\right), 0\right] \tag{39}$$

Component II The thermal comfort of the residents is considered as the second component in the reward function. It is expected that the predicted percentage of dissatisfied (PPD) should be as lower as possible.

$$\begin{aligned} & r_2 = -\Xi_1 \\ & PPD = -\Xi_1\left(100 - 95 \cdot \exp\left(-0.03353 \cdot PMV^4 - 0.2179 \cdot PMV^2\right)\right) \end{aligned} \tag{40}$$

Component II Physical constraints for the appliances need to be considered in the reward function. Once the constraints will be violated according to control signals, a large penalty factor should be given.

$$r_3 = \begin{cases} 0 & , if \ E_{i,T}^{ESS} \geq E_i^{ESS,EXP} \\ -\Xi_2 & , else \end{cases} \tag{41}$$

$$r_3 = \begin{cases} 0 & , if \ E_{i,t}^{ESS} \in \left[\underline{E_i^{ESS}}, \overline{E_i^{ESS}}\right] \\ -\Xi_2 & , else \end{cases} \tag{42}$$

$$r_3 = \begin{cases} 0 & , if \ T_t \in \left[T^{\min}, T^{\max}\right] \\ -\Xi_2 & , else \end{cases} \tag{43}$$

$$r_3 = \begin{cases} 0 & , if \ \sum_{t=t_i^s}^{t_i^e} \delta_{i,t} \Delta t = WT_i \\ -\Xi_2 & , else \end{cases} \tag{44}$$

Appliance operational dependency based on living habits

In a smart home, different operation dependency requirements based on the living habits of the residents will be imposed. For example, the clothes dryer is expected to start working shortly after the completion of the washing machine. To realize the individualized settings for HEMS, following five dependency mode are designed to be integrated into the deep reinforcement learning in this paper:

3.3.5 Dependency mode I

The appliance $\mathcal{A}$ must be started after/before the completion of appliance $\mathcal{B}$ plus a time shift. Hence, dependency mode I can be expressed as:

$$t_B^{cmpt} + t_{A,B}^{shft} \leq t_A^{start} \tag{45}$$

$$\text{or } t_A^{start} \leq t_B^{cmpt} + t_{A,B}^{shft} \tag{46}$$

where t^{cmpt} represents the completion time of the appliance; t^{start} represents the starting time of the appliance; t^{shft} represents a time shift.

3.3.6 Dependency mode II

The appliance $\mathcal{A}$ must be started after/before the start of appliance $\mathcal{B}$ plus a time shift. Hence, dependency mode II can be expressed as:

$$t_B^{start} + t_{A,B}^{shft} \leq t_A^{start} \text{ or } t_A^{start} \leq t_B^{start} + t_{A,B}^{shft} \tag{47}$$

3.3.7 Dependency mode III

The appliance $\mathcal{A}$ must be completed after/before the completion of appliance $\mathcal{B}$ plus a time shift. Hence, Dependency Mode III can be expressed as:

$$t_B^{cmpt} + t_{A,B}^{shft} \leq t_A^{cmpt} \text{ or } t_A^{cmpt} \leq t_B^{cmpt} + t_{A,B}^{shft} \tag{48}$$

3.3.8 Dependency mode IV

The appliance $\mathcal{A}$ must be completed after/before the start of appliance $\mathcal{B}$ plus a time shift. Hence, dependency mode IV can be expressed as:

$$t_B^{start} + t_{A,B}^{shft} \leq t_A^{cmpt} \text{ or } t_B^{start} + t_{A,B}^{shft} \leq t_A^{cmpt} \ or \ t_A^{start} \leq t_B^{cmpt} + t_{A,B}^{shft} \tag{49}$$

3.3.9 Dependency mode V

The overlapped running time of appliances $\mathcal{A}$ and $\mathcal{B}$ must be smaller or larger than a threshold τ. Hence, dependency mode V can be expressed as:

$$\left|\left\{t | \delta_{A,t} = 1, \forall t \in (1, T)\right\}\right| \cap \left|\left\{t | \delta_{B,t} = 1, \forall t \in (1, T)\right\}\right| \leq \tau \tag{50}$$

$$\text{or} \left|\left\{t | \delta_{A,t} = 1, \forall t \in (1, T)\right\}\right| \cap \left|\left\{t | \delta_{B,t} = 1, \forall t \in (1, T)\right\}\right| \geq \tau \tag{51}$$

There are two methods to integrate the dependency into the DRL framework. In the first method, when the dependency is violated, give a large negative reward. Hence, the dependency modes are treated as constraints. However, when adding more dependencies, more episodes will be used to reach the convergence. In the second method, the devices

Table 1 Settings and parameters for HERs

Controllable constant power appliance				
Device	Operation duration	Operation time range	Power	Interruptible
Pool pump	3 h	[9:00,18:00]	1.5	Yes
Washing machine	2 h	[10:00,19:00]	0.9	Yes
Cloth dryer	1.5 h	[10:00,19:00]	2.5	Yes
Dish washer	1 h	[19:00,7:00+]	1	Yes
Coffee machine	15 min	[6:00,8:30]	0.8	No
Bread maker	15 min	[6:00,8:30]	0.5	No
Rice cooker	50 min	[16:00,19:00]	0.6	No
Dehumidifier	30 min	[9:00,21:00]	0.3	No
BESS				
Operation duration	Power capacity		Energy capacity	
[0:00, 23:00]	2 kW		8 kW	
SOC Lower limit	SOC Upper limit		SOC_deisre	
10%	90%		40%	
PEV				
Operation duration	Power capacity		Energy capacity	
[17:00*,9:00+*]	1.7 kW		8 kW	
SOC lower limit	SOC Upper limit	SOC_arrive	SOC_deisre	
10%	90%	60%	80%*	
Air-conditioner				
Operation duration	power range	T_{set}	T_{min}	T_{max}
[11:00,4:00+]	0–2 kW	23	21	25

in a dependency can be regarded as a pack. For example, the coffee machine and the bread maker are expected to finish the tasks at the same time. Then, in the second method, these two devices can be regarded as one device during the scheduling. However, when the dependencies are complex, it is difficult to pack different devices. Hence, in the application, these two methods can be utilized jointly.

When the user enters a new dependency, a logic check should be conducted to ensure that the new input dependency does not have a conflict with the previous set dependency mode. The logic check algorithm is not the scope of this paper. Therefore, we assume that all the dependencies have no conflict. In the revised paper, we have stated this assumption.

4 Case studies

4.1 Experiment setting

The proposed method is tested in a smart home consisting of a 2 kW-capacity rooftop PV, a residential BESS, PEV, and multiple controllable loads including interruptible constant power loads, non-interruptible constant power loads and an air-conditioner. The parameters and scheduling information of all the controllable devices are summarized in Table 1. The mark “+” in Table 1 indicates the “next day”. For example, the operation time range of dishwasher is set from 19:00 to the next day 7:00. The mark "*" in Table 1 indicates the uncertain inputs. For example, the operation time of PEV starts from 17:00, but it is uncertain. The scheduling interval is set to 5 min. The DCT is set to be $8.1 kW/month according to.[4]. The simulations were completed by a PC with an Intel Core (TM) i7-9750 CPU@2.60 GHz with 16.00 GB RAM.

4.2 Comparisons between DQN, DDPG and optimization result

In this paper, two reinforcement learning algorithms, i.e. DQN and DDPG, are applied in the HEMS problem. The evolution processes of DQN and DDPG are shown in Figs. 3 and 4, respectively. It can be concluded that DDPG shows better convergence capability than DQN. DQN reaches convergence after around 600 episodes while DDPG reaches convergence after around 100 episodes.

Figure 5 shows the comparison between DRL and conventional optimization methods in daily cost in winter. Furthermore, the daily cost and peak are summarized in Tables 2 and 3. It can be concluded that there is a slight difference between the DRL method and the optimization method. The optimization result is the optimal control solution for HERs towards a specific objective function. Hence,

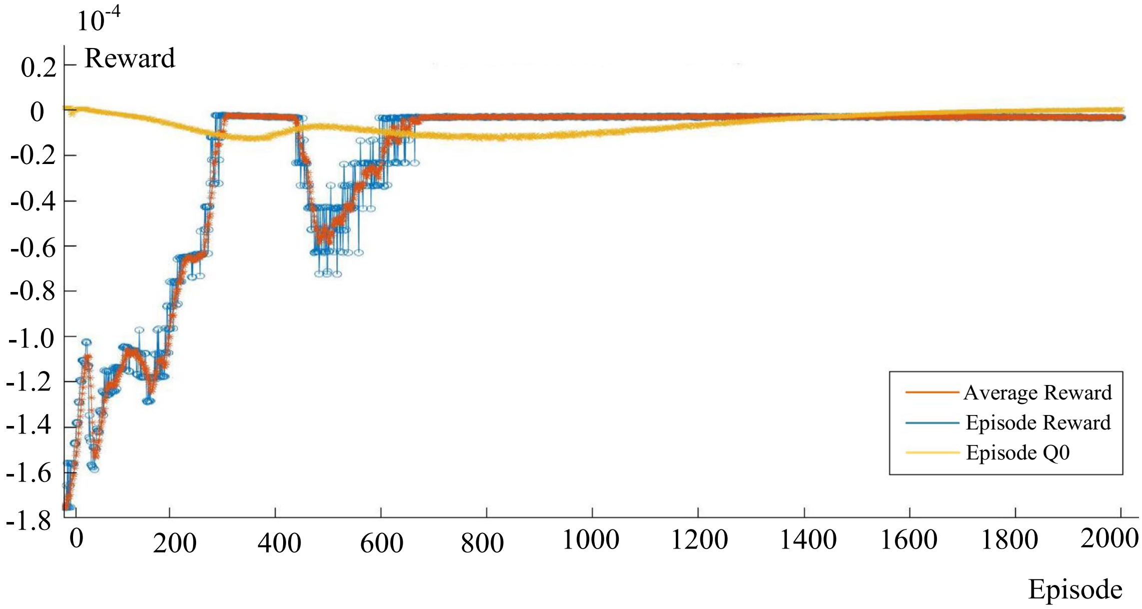


Fig. 3 Reward evolution of DQN

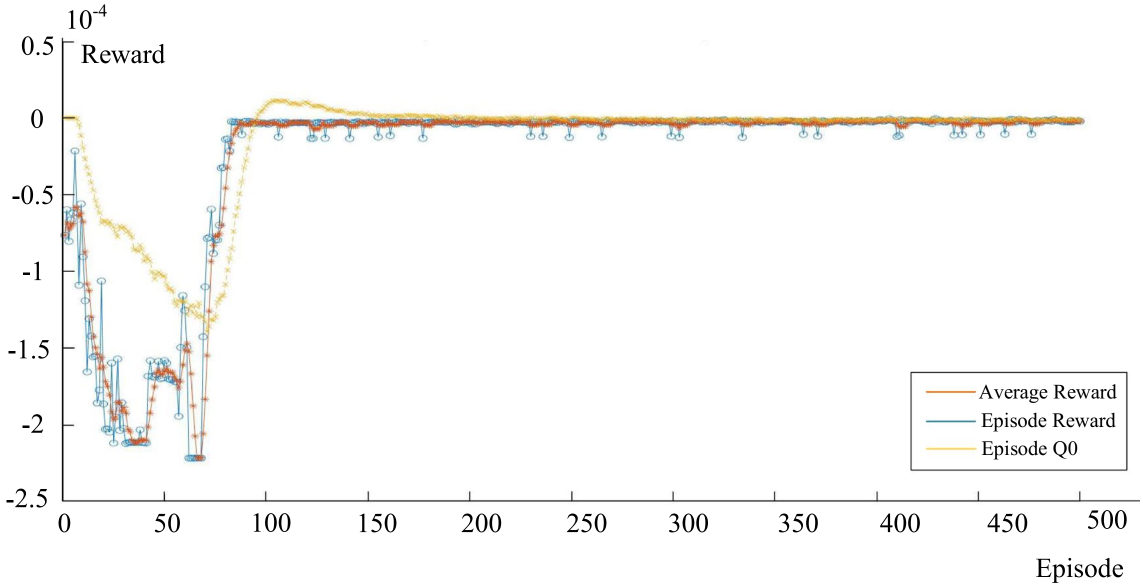


Fig. 4 Reward evolution of DDPG

if the learning result of a model-free method is closed to the optimization result, it means that the machine can automatically choose a relatively satisfying solution. Therefore, according to Fig. 5, both DDPG and DQN algorithm can reach a satisfying result compared with the optimization method. For the mean daily cost, the DQN algorithm is 7.8% higher than the optimization method while the DDPG algorithm 2.3% higher than the optimization method. For the mean peak load, the DQN algorithm is 3.4% higher than the optimization method while the DDPG algorithm 2.6% higher than the optimization method.

In Table 3, the training episode and decision-making time of the optimization method, DQN algorithm and DDPG algorithm with different numbers of HERs are given. From the home level, the number of controllable devices will not be extremely large. To reveal the scalability of our work, the number of HERs is increased to 30. With the increasing of the number of HERs, the average training episode of DQN

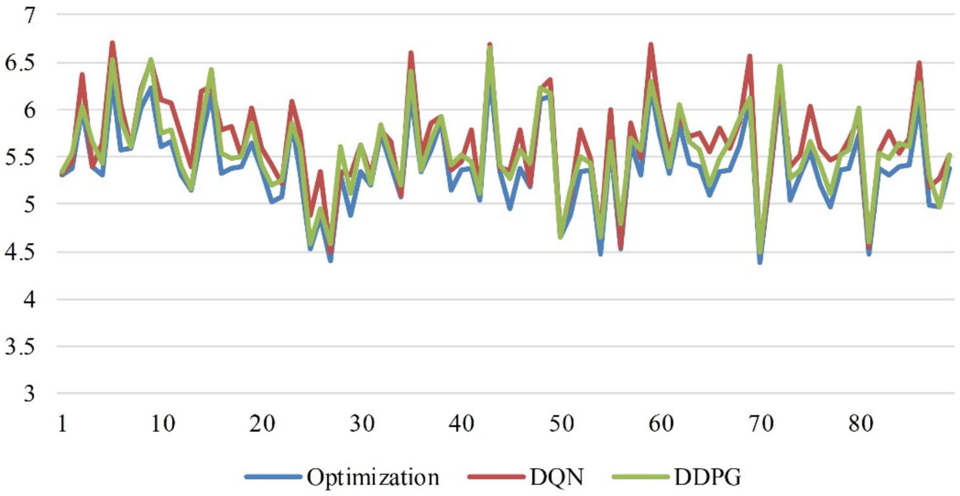


Fig. 5 Comparison between optimization result and reinforcement learning result

Table 2 Daily cost and peak load result from optimization, DQN and DDPG

	Optimization	DQN	DDPG
Daily cost ($)			
Mean	2.57	2.77	2.63
St. dev.	1.76	1.57	1.48
Peak			
Mean	5.31	5.49	5.45
St. dev.	1.01	0.54	0.67

and DDPG increased. When the number of HERs reaches 30, the DQN takes 1251 episodes to converge on average. Compared with DQN, DDPG shows superior convergence capability (289 episodes on average) when the number of HERs reaches 30. Compared with the optimization method, the DRL shows advantages in decision making time. For the average decision time, all the methods under the RL framework take very little time. It takes only several milliseconds for each time of forwarding propagation. However, for the optimization method, the optimization process needs to be re-run for every decision and the computational time will increase with the problem scale. When the problem scale

Table 3 Training Episode and decision-making time of optimization, DQN and DDPG

	Method	Number of HERs		
		11	20	30
Average training episode	Optimization	N/A	N/A	N/A
	DQN	615	924	1251
	DDPG	83	167	289
Average decision time	Optimization	2 min 39 s	5 min 29 s	8 min 12 s
	DQN	< 1 s	< 1 s	< 1 s
	DDPG	< 1 s	< 1 s	< 1 s

Table 4 Established dependency modes

No	Dependency	Reason
1	$t_{BM}^{cmpt}+0\le t_{CM}^{cmpt}\le t_{BM}^{cmpt}+0$	The user expects that the coffee machine and bread maker can complete simultaneously
2	$\left\lvert\left\{t\middle\vert\delta_{A,t}=1,\forall t\in(1,T)\right\}\right\rvert\cap\left\lvert\left\{t\middle\vert\delta_{B,t}=1,\forall t\in(1,T)\right\}\right\rvert\le 30$	The overlap working time of the pool pump and cloth dryer should within 30 min due to the noise
3	$t_{CD}^{start}\le t_{WM}^{cmpt}+30$	The user expects that the clothes dryer can start working shortly after the completion of the washing machine
4	$t_{CD}^{cmpt}\le t_{DH}^{start}+0$	The dehumidifier cannot start working before the completion of the cloth drier

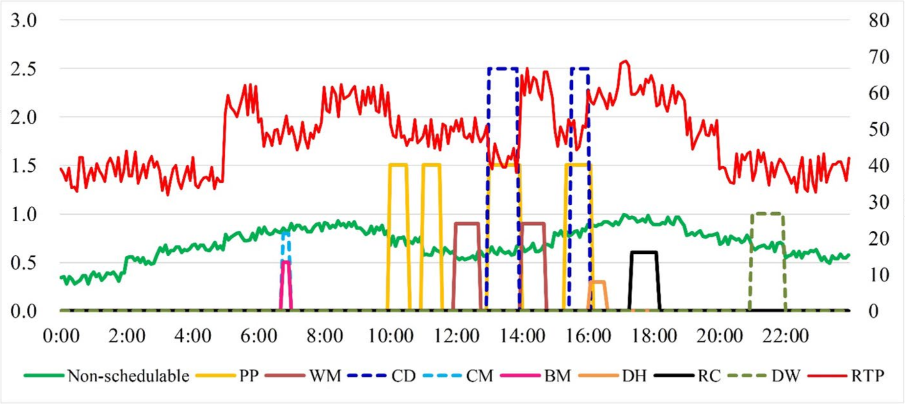


Fig. 6 Scheduling of the controllable load without dependencies

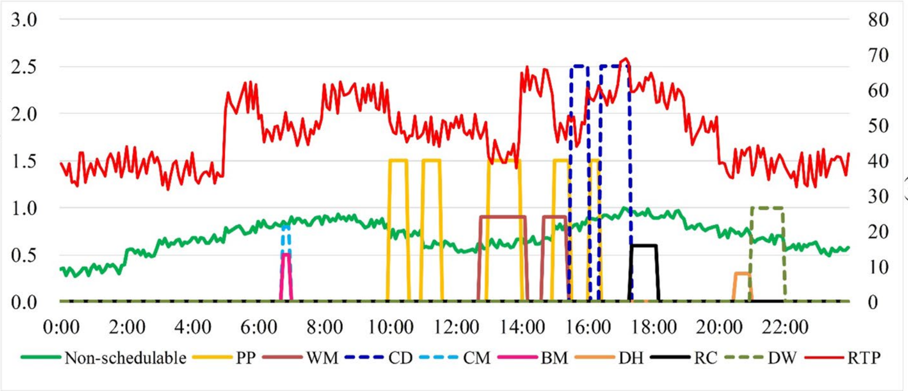


Fig. 7 Scheduling of the controllable load with dependencies

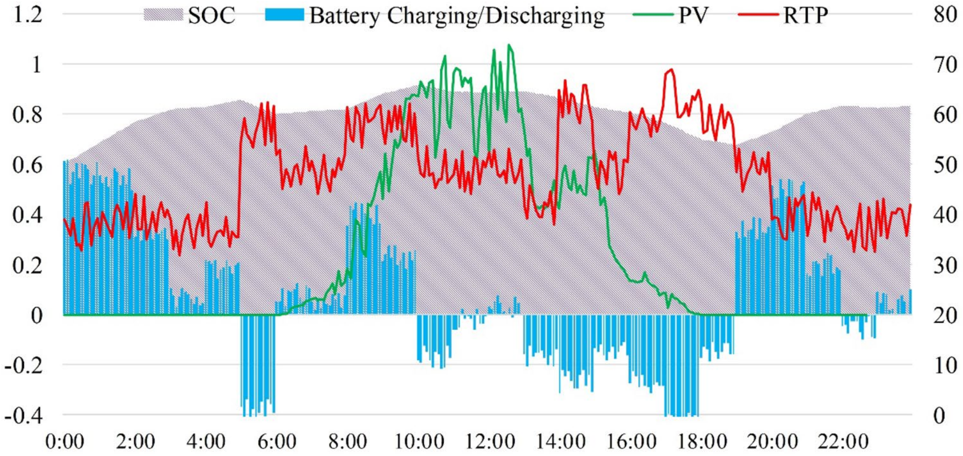


Fig. 8 Scheduling of the residential BESS

becomes large, it brings difficulties to the real-time control of HERs.

4.3 HERs scheduling results

In this paper, the operation dependencies based on the residents' living habit is integrated into the DRL framework. Four dependencies are considered in the simulations listed in Table 4. The residents can change the dependencies and the dependency modes can be regarded as atomic. The residents can create personalized dependencies by compositing the atomic.

The scheduling of controllable load without considering dependency is shown in Fig. 6, while the scheduling of controllable load with considering dependency is shown in Fig. 7. The operation time of cloth drier, the dehumidifier is

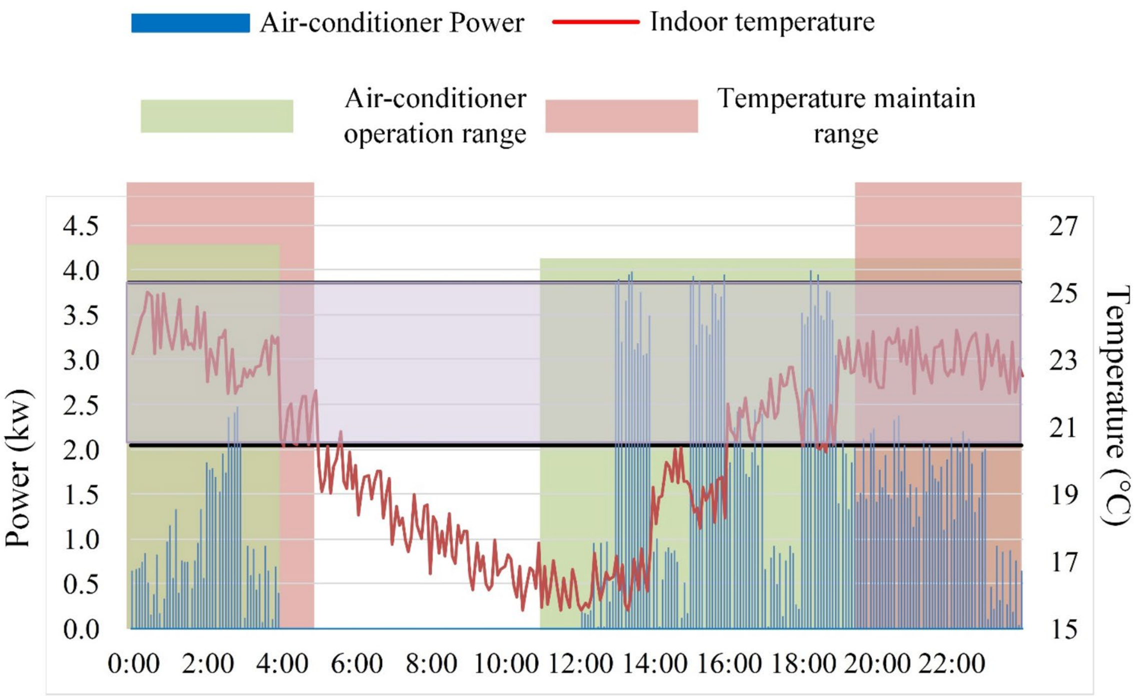


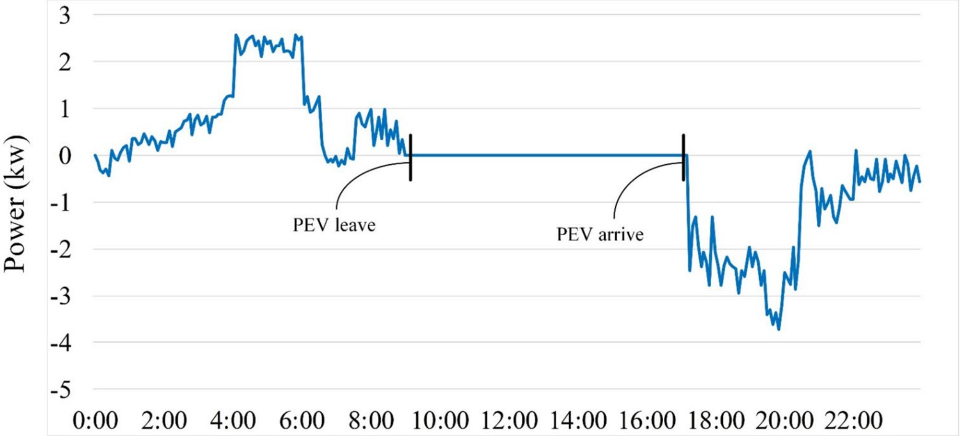
Fig. 9 Scheduling of the air-conditioner

shifted to meet the dependency constraints which means the resident's living habits are sufficiently respected. Besides, it can be found that for both cases, the controllable is scheduled at the time when RTP is relatively low.

The operation strategy for residential BESS is shown in Fig. 8. The positive value represents charging while the negative value represents discharging. Conventionally, the residential BESS is utilized to absorb excessive rooftop PV output during noon. An interesting finding is that the BESS is discharging during noon when the PV output is the highest. That is because many controllable loads are scheduled at noon, such as the water pump and washing machine. At midnight, around 0:00–2:00, the BESS is charging through purchasing relative cheap electricity from utility gird indicating that the BESS is obtaining profit interest arbitrage. Hence, in an automatic smart home, the role of the residential BESS may be redefined in the future. From the authors' view, in a future smart home, the BESS will take the responsibility of: (1) interest arbitrage (2) providing flexible and reliable energy supply.

The operation strategy for the air-conditioner is shown in Fig. 9. The setting of the air-conditioner mainly focuses on three parts. First, the allowed operation time of the air-conditioner is set from 11:00 to 4:00+. Second, the indoor temperature should be maintained within the allowed range from 19:30 to 3:30+. Third, the allowed range of the indoor

Fig. 10 Scheduling of the PEV



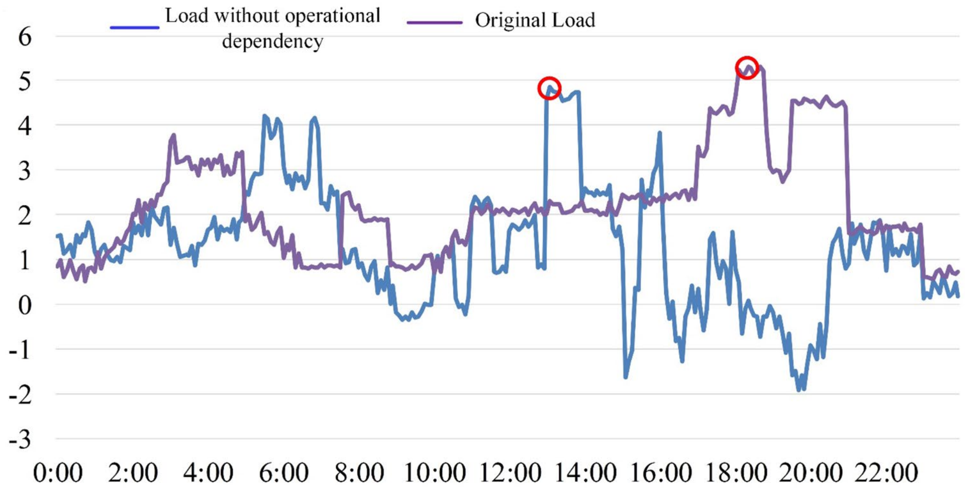


Fig. 11 Net load profile without dependencies

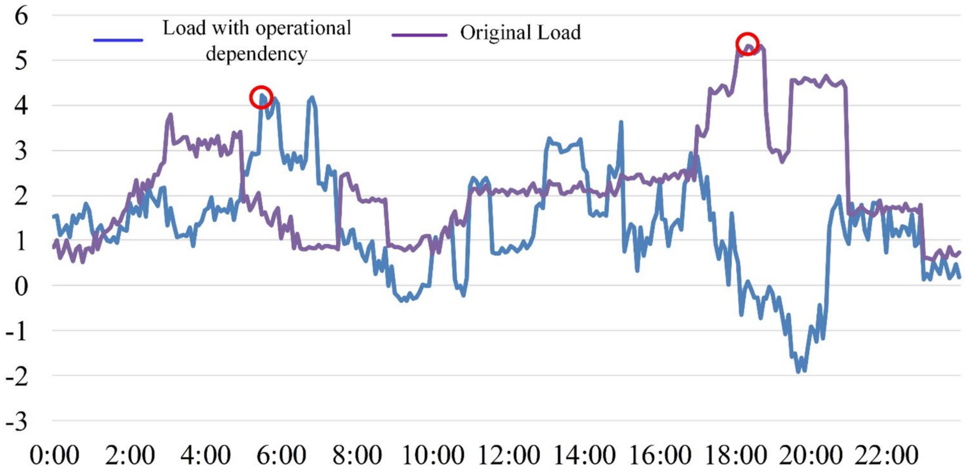


Fig. 12 Net load profile with dependencies

temperature is set from 21 to 25 °C. According to the scheduling result, the air-conditioner begins to warm-up from 12:00 and the indoor temperature begins to rise gradually. Till 20:00, the indoor temperature starts to fluctuate around 23 °C because the thermal comfort of the residents is fully considered in the DRL framework.

The operation strategy for PEV is shown in Fig. 10. In the conventional mathematical model, the departure time and the arrival time of the PEV should be fully considered as a stochastic model. However, in the DQL framework, the stochastic model does not need to be considered anymore, because whether PEV is at home is considered as a state in the DQL, indicating only when PEV is available, the process can give a charging or discharging order to the PEV. Hence, in the real application, the action is generated according to the current state of PEV and does not rely on the stochastic model. According to the simulation result, the PEV will discharge between 18:00 and 20:00 and charge between 4:00 and 8:00.

After considering all the HERs, the final net load without considering operational dependency and the final net load considering operational dependency are shown Figs. 11 and 12, respectively. It can be concluded that the load profile is changed totally. The peak load is shifted and reduced. The peak load is marked by red circles. In Fig. 11, the peak is shifted from 19:00 to 13:00, while in Fig. 12 the peak is shifted to 5:00. If the peak is the highest in the billing period, DCT will be charged. According to the comparison between Figs. 11 and 12, the introducing of dependency operation does not mean that more DCT.

4.4 Future works

In this paper, the proposed method shows difficulties to control the HERs properly under the emergent situation, such as an outage. However, in the smart home, the fast response to the emergency is an essential section. At the current stage, the control strategy under emergency can be solved through a simple rule-based method. For example, when the outage happens, the battery will be automatically switched to the backup mode, which provides energy to the inelastic load, such as light. The other loads, such as the air-conditioner is automatically switched off. In the future, the researches will focus on how to formulate a powerful control strategy for HERs which can perform well in both normal and emergent situations.

5 Conclusion

In this paper, we have proposed a HEMS based on DRL and two algorithms, i.e. DQL and DDPG, are utilized. First, we provided a detailed classification of HERs based on the mathematical model. Then, based on the proposed DRL framework for HEMS, the MDP is defined in detail. To better accommodate the user's living habits, five operational dependency modes are put forward and the method to integrate the dependency modes to the DRL framework is introduced. According to the case studies, it shows that DDPG performs better especially in convergence capacity. Compared with the optimization results, the results from DRL are very closed to the optimal solution. However, the decision-making time for DRL is much smaller than optimization, which shows advantages in real-time control. Furthermore, through DRL, the operational dependencies can be satisfied.

References

1. Zhang Y, Hajiesmaili MH, Cai S, Chen M, Zhu Q (2016) Peak-aware online economic dispatching for microgrids. IEEE Trans Smart Grid 9(1):323–335
2. Mohsenian-Rad A-H, Leon-Garcia A (2010) Optimal residential load control with price prediction in real-time electricity pricing environments. IEEE Trans Smart Grid 1(2):120–133
3. Tsui KM, Chan S-C (2012) Demand response optimization for smart home scheduling under real-time pricing. IEEE Trans Smart Grid 3(4):1812–1821
4. Luo F, Kong W, Ranzi G, Dong ZY (2019) Optimal home energy management system with demand charge tariff and appliance operational dependencies. IEEE Trans Smart Grid 11(1):4–14
5. Luo F, Ranzi G, Wan C, Xu Z, Dong ZY (2018) A multistage home energy management system with residential photovoltaic penetration. IEEE Trans Ind Inf 15(1):116–126
6. Wu X, Hu X, Yin X, Moura SJ (2016) Stochastic optimal energy management of smart home with PEV energy storage. IEEE Trans Smart Grid 9(3):2065–2075
7. Anvari-Moghaddam A, Monsef H, Rahimi-Kian A (2014) Optimal smart home energy management considering energy saving and a comfortable lifestyle. IEEE Trans Smart Grid 6(1):324–332
8. Yu L, Jiang T, Zou Y (2017) Online energy management for a sustainable smart home with an HVAC load and random occupancy. IEEE Trans Smart Grid 10(2):1646–1659
9. Hou X, Wang J, Huang T, Wang T, Wang P (2019) Smart home energy management optimization method considering energy storage and electric vehicle. IEEE Access 7:144010–144020
10. Luo F, Ranzi G, Wang X, Dong ZY (2016) Service recommendation in smart grid: vision, technologies, and applications. In: Proceedings of 9th International Conference on Service Science (ICSS), pp 31–38
11. Luo F, Ranzi G, Kong W, Dong ZY, Wang S, Zhao J (2017) Non-intrusive energy saving appliance recommender system for smart grid residential users. IET Gener Transm Distrib 11(7):1786–1793
12. Luo F, Ranzi G, Wang X, Dong ZY (2017) Social information filtering-based electricity retail plan recommender system for smart grid end users. IEEE Trans Smart Grid 10(1):95–104
13. Kawakami T, Yoshihisa T, Fujita N, Tsukamoto M (2013) A rule-based home energy management system using the Rete algorithm. In: Proceedings of IEEE 2nd Global Conference on Consumer Electronics (GCCE), pp 162–163
14. Yoshihisa T, Fujita N, Tsukamoto M (2012) A rule generation method for electrical appliances management systems with home EoD. In: Proceedings of The 1st IEEE Global Conference on Consumer Electronics, pp 248–250
15. Althaher SZ, Mutale J (2012) Management and control of residential energy through implementation of real time pricing and demand response. In: Proceedings of IEEE Power and Energy Society General Meeting, pp 1–7
16. Ahmed MS, Shareef H, Mohamad A, Abd Ali J, Mutlag AH Rule base home energy management system considering residential demand response application. In: Proceedings of Applied Mechanics and Materials, pp 526–531
17. Shareef H, Ahmed MS, Mohamed A, Al Hassan E (2018) Review on home energy management system considering demand responses, smart technologies, and intelligent controllers. IEEE Access 6:24498–24509
18. Liu Y, Yuen C, Yu R, Zhang Y, Xie S (2015) Queuing-based energy consumption management for heterogeneous residential demands in smart grid. IEEE Trans Smart Grid 7(3):1650–1659
19. Hong Y-Y, Lin J-K, Wu C-P, Chuang C-C (2012) Multi-objective air-conditioning control considering fuzzy parameters using immune clonal selection programming. IEEE Trans Smart Grid 3(4):1603–1610
20. Kim Y-J (2020) A supervised-learning-based strategy for optimal demand response of an HVAC system in a multi-zone office building. IEEE Trans Smart Grid 11(5):4212–4226
21. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529(7587):484–489
22. Perrusquía A, Yu W, Li X (2020) Multi-agent reinforcement learning for redundant robot control in task-space. Int J Mach Learn Cybern. https://doi.org/10.1007/s13042-020-01167-7
23. Keyhanipour AH, Moshiri B, Rahgozar M, Oroumchian F, Ansari AA (2016) Integration of data fusion and reinforcement learning techniques for the rank-aggregation problem. Int J Mach Learn Cybern 7(6):1131–1145
24. Mocanu E, Mocanu DC, Nguyen PH, Liotta A, Webber ME, Gibescu M, Slootweg JG (2018) On-line building energy optimization using deep reinforcement learning. IEEE Trans Smart Grid 10(4):3698–3708

25. Xu X, Jia Y, Xu Y, Xu Z, Chai S, Lai CS (2020) A multi-agent reinforcement learning based data-driven method for home energy management. IEEE Trans Smart Grid 11(4):3201–3211
26. Zhang X, Biagioni D, Cai M, Graf P, Rahman S (2020) An edge-cloud integrated solution for buildings demand response using reinforcement learning. IEEE Trans Smart Grid 12(1):420–431
27. Yu L, Sun Y, Xu Z, Shen C, Yue D, Jiang T, Guan X (2020) Multi-agent deep reinforcement learning for HVAC control in commercial buildings. IEEE Trans Smart Grid 12(1):407–419
28. Ye Y, Qiu D, Wu X, Strbac G, Ward J (2020) Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning. IEEE Trans Smart Grid 11(4):3068–3082
29. Gorostiza FS, Gonzalez-Longatt F (2020) Deep reinforcement learning-based controller for SOC management of multi-electrical energy storage system. IEEE Trans Smart Grid 11(6):5039–5050
30. Luo F, Dong ZY, Meng K, Wen J, Wang H, Zhao J (2016) An operational planning framework for large-scale thermostatically controlled load dispatch. IEEE Trans Ind Inf 13(1):217–227
31. Buratti C, Ricciardi P, Vergoni M (2013) HVAC systems testing and check: a simplified model to predict thermal comfort conditions in moderate environments. Appl Energy 104:117–127

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.