

Using Deep Reinforcement Learning in Optimal Energy Management for Residential House Aggregators with Uncertain User Behaviors

Yujun Lin

State Key Laboratory of Advanced
Electromagnetic Technology
School of Electrical and Electronic
Engineering
Huazhong University of Science and
Technology
Wuhan, China
0000-0002-5079-8642

Linfang Yan

State Grid (Suzhou) City & Energy
Research Institute
Suzhou, China
linfyan@foxmail.com

Hongxun Hui

State Key Laboratory of Internet of
Things for Smart City and Department
of Electrical and Computer
Engineering
University of Macau
Macau, China
hongxunhui@um.edu.mo

Yin Chen*

Department of Electronic and
Electrical Engineering
University of Strathclyde
Glasgow, U.K.
yin.chen.101@strath.ac.uk

Xia Chen

State Key Laboratory of Advanced
Electromagnetic Technology
School of Electrical and Electronic
Engineering
Huazhong University of Science and
Technology
Wuhan, China
cxhust@foxmail.com

Jinyu Wen

State Key Laboratory of Advanced
Electromagnetic Technology
School of Electrical and Electronic
Engineering
Huazhong University of Science and
Technology
Wuhan, China
jinyu.wen@hust.edu.cn

Abstract—In this study, the home energy management problem, which can be regarded as a high-dimensional optimization problem, for numerous residential houses, is addressed. The concept of the aggregator is utilized to reduce the state and action space and to handle the high dimensionality. A two-stage deep reinforcement learning (DRL)-based approach is proposed for the aggregators to track the schedule from a superior grid and guarantee the operation constraints. In the first stage, a DRL control agent is set to learn the optimal scheduling strategy interacting with the environment based on the soft-actor-critic framework and generate the aggregate control actions. In the second stage, the aggregate control actions are disaggregated to individual appliances considering the users' behaviors. The uncertainty of an electric vehicle's charging demand is quantitatively expressed based on the driver's experience. An aggregate anxiety concept is introduced to characterize the driver's anxiety on the electric vehicle's range and uncertain events. Finally, simulations are conducted to verify the effectiveness of the proposed approach under dynamic user behaviors, and comparisons show the superiority of the proposed approach over other benchmark methods.

Keywords—Home energy management, electric vehicles (EVs), deep reinforcement learning, soft actor-critic, dynamic user behaviors.

I. INTRODUCTION

In recent years, the deployment of distributed energy resources (DERs), such as rooftop solar panels, electric vehicles (EVs), and battery storage in smart grids, has increased to mitigate climate change and carbon emissions [1]. Although these resources enhance energy efficiency and grid reliability, their integration poses challenges in energy management, particularly in the management of numerous home appliances and DERs in residential areas [2]. However, the advent of smart energy devices and advanced

metering has enabled home energy management (HEM), which is employed to optimize appliance operations to reduce costs and maintain comfort [3].

In HEM research, model-based optimization approaches, which require detailed system modeling, and model-free methods (such as deep reinforcement learning (DRL)), which learn optimal schedules through interaction with the environment, are predominantly utilized. DRL has been adopted in numerous studies on the HEM problem, and its excellent control performance in a dynamic environment has been demonstrated. In [4], a multi-agent DRL with an attention mechanism was utilized in heating, ventilation, and air conditioning (HVAC) control to minimize the energy costs in a multizone commercial building; the mixed air temperature was used to describe the building temperature regulated by a group of HVAC systems. In [5], a DRL model that combined local HEM systems with a global server was proposed to optimize the scheduling of multiple smart homes and their appliances. In [6], personalized comfort was improved while reducing electricity costs and flattening the demand curve by incorporating human feedback and activity into the decision-making process. The aforementioned DRL-based HEM methods have exhibited promising performances in dynamic environments. However, individual EV-charging models have mostly been described solely based on the arrival time, departure time, and desired battery energy, and the distinct characteristics of drivers' individual behaviors, have been neglected. The effectiveness of DRL in dynamic HEM environments has been shown in previous studies; however, challenges persist, particularly in capturing diverse user needs and preferences and in addressing the high dimensionality of scheduling many appliances.

This paper introduces a novel two-stage aggregation

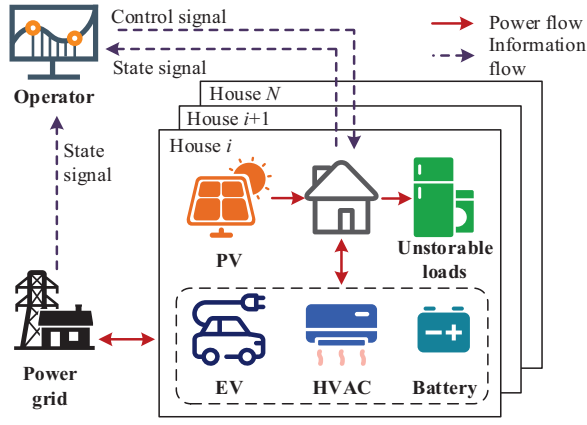


Fig. 1 Schematic of residential houses integrated with photovoltaic (PV).

approach for HEM within a virtual power plant framework optimized at the aggregator level to mitigate dimensionality issues. In the first stage, the soft-actor-critic (SAC) DRL framework is used to learn optimal scheduling strategies, whereas in the second stage, control actions are disaggregated to individual appliances by considering user behaviors. Previous aggregator models are extended by integrating driver experience, charging preferences, and anxieties into EV-charging behaviors.

II. SYSTEM OPERATION MODEL

The main control objectives in this study are storable loads, including batteries, HVAC systems, and EVs, as depicted in Fig. 1. In this section, the dynamic models of the storable loads in residential houses are presented, as referenced in [7].

A. Battery Model

In this study, we consider small-capacity household batteries. Household batteries can release energy into the grid for profit or support HVAC systems and EVs at high state of charge (SoC) levels. The residual energy of the batteries is calculated as

$$e_i^{\text{BAT}}(t) = \begin{cases} \theta_i^{\text{BAT}} e_i^{\text{BAT}}(t-1) + \eta_i^{\text{BAT}} p_i^{\text{BAT}}(t) \Delta t, & p_i^{\text{BAT}}(t) \geq 0 \\ \theta_i^{\text{BAT}} e_i^{\text{BAT}}(t-1) + \frac{1}{\eta_i^{\text{BAT}}} p_i^{\text{BAT}}(t) \Delta t, & p_i^{\text{BAT}}(t) < 0 \end{cases}, \quad (1)$$

where θ_i^{BAT} , η_i^{BAT} , p_i^{BAT} , and e_i^{BAT} denote the dissipation rate, conversion coefficient, power consumption, and SoC of battery i , respectively.

We also consider the following operation constraints:

$$p_i^{\text{BAT}}(t) \in [p_{i,\min}^{\text{BAT}}, p_{i,\max}^{\text{BAT}}] \quad ; \quad (2)$$

$$|p_i^{\text{BAT}}(t) - p_i^{\text{BAT}}(t-1)| \in [0, \delta_{i,\max}^{\text{BAT}}] \quad ; \quad (3)$$

$$e_i^{\text{BAT}}(t) \in [e_{i,\min}^{\text{BAT}}, e_{i,\max}^{\text{BAT}}] \quad , \quad (4)$$

where $p_{i,\min}^{\text{BAT}}$ and $p_{i,\max}^{\text{BAT}}$ are the lower and upper bounds of the power consumption of battery i , respectively; $\delta_{i,\max}^{\text{BAT}}$ is the ramping limitation of battery i ; $e_{i,\min}^{\text{BAT}}$ and $e_{i,\max}^{\text{BAT}}$ are the lower and upper bounds of the indoor temperature of battery i , respectively.

B. HVAC Model

The function of HVAC systems is to improve the comfort of residents by maintaining the indoor temperature within a reasonable range as

$$\theta(t) \in [\theta_{\min}, \theta_{\max}] \quad , \quad (5)$$

where θ denotes the indoor temperature, and θ_{\min} and θ_{\max} represent the lower and upper bounds of the temperature comfort zone, respectively.

The indoor temperature is affected by multiple factors such as the previous indoor temperature, ambient temperature, air humidity, and active power of the HVAC system. Based on the energy storage characteristics of the HVAC system, the dynamic model can be presented as follows:

$$\begin{aligned} \theta(t+1) &= \theta(t) - \frac{1}{R_{\text{hv}} C_{\text{hv}}} (\theta(t) - \theta_{\text{amb}}(t) + \eta_{\text{hv}} R_{\text{hv}} p(t)) \Delta t \\ &= (1 - \frac{\Delta t}{R_{\text{hv}} C_{\text{hv}}}) \theta(t) - \frac{\eta_{\text{hv}}}{C_{\text{hv}}} p \Delta t + \frac{\Delta t}{R_{\text{hv}} C_{\text{hv}}} \theta_{\text{amb}}(t) \quad , \quad (6) \\ &= \mathcal{G} \theta(t) + \eta p(t) \Delta t + \sigma \theta_{\text{amb}}(t) \end{aligned}$$

where $\theta(t)$ and $\theta_{\text{amb}}(t)$ are the indoor temperature and ambient temperature at timeslot t , respectively; R_{hv} is the equivalent thermal resistance; C_{hv} is the equivalent heat capacity; η_{hv} is the efficiency coefficient; p is the power consumption; Δt is the time interval. The dynamic model of the HVAC system can be expressed in the unified form of an ES as

$$\begin{aligned} e_i^{\text{HVAC}}(t) &= \mathcal{G}_i^{\text{HVAC}} e_i^{\text{HVAC}}(t-1) \\ &\quad + \eta_i^{\text{HVAC}} p_i^{\text{HVAC}}(t) \Delta t + \sigma_i^{\text{HVAC}} T_{\text{amb}}(t) \end{aligned} \quad , \quad (7)$$

where p_i^{HVAC} , e_i^{HVAC} , $\mathcal{G}_i^{\text{HVAC}}$, η_i^{HVAC} , and σ_i^{HVAC} denote the power consumption, indoor temperature, dissipation rate, conversion coefficient, and impact factor of the ambient temperature of HVAC system i , respectively. These factors are defined as

$$\begin{cases} e_i^{\text{HVAC}}(t) = T(t), \theta_i^{\text{HVAC}} = 1 - \frac{1}{R_{\text{hv}} C_{\text{hv}}} \\ \eta_i^{\text{HVAC}} = \frac{\eta_{\text{hv}}}{C_{\text{hv}}}, \sigma_i^{\text{HVAC}} = \frac{\Delta t}{R_{\text{hv}} C_{\text{hv}}} \end{cases} \quad . \quad (8)$$

In addition to these aforementioned equality constraints, the state variables should be limited within a certain range as follows:

$$p_i^{\text{HVAC}}(t) \in [p_{i,\min}^{\text{HVAC}}, p_{i,\max}^{\text{HVAC}}] \quad ; \quad (9)$$

$$|p_i^{\text{HVAC}}(t) - p_i^{\text{HVAC}}(t-1)| \in [0, \delta_{i,\max}^{\text{HVAC}}] \quad ; \quad (10)$$

$$e_i^{\text{HVAC}}(t) \in [e_{i,\min}^{\text{HVAC}}, e_{i,\max}^{\text{HVAC}}] \quad , \quad (11)$$

where $p_{i,\min}^{\text{HVAC}}$ and $p_{i,\max}^{\text{HVAC}}$ are the lower and upper bounds of the power consumption of HVAC system i , respectively; $\delta_{i,\max}^{\text{HVAC}}$ is the ramping limitation of HVAC system i ; $e_{i,\min}^{\text{HVAC}}$ and $e_{i,\max}^{\text{HVAC}}$ are the lower and upper bounds of the indoor temperature of HVAC system i , respectively.

C. EV Model

The dynamic EV user behaviors of multiple residential users are uncertain and time-varying. In this study, a driver's experience, charging preferences, and charging habits are jointly considered to describe the EV-charging model.

The SoC of the EV can be calculated in a unified form of the ES as

$$e_i^{\text{EV}}(t) = \theta_i^{\text{EV}} e_i^{\text{EV}}(t-1) + \eta_i^{\text{EV}} p_i^{\text{EV}}(t) \Delta t \quad , (12)$$

where p_i^{EV} , e_i^{EV} , θ_i^{EV} , and η_i^{EV} denote the power consumption, SoC, dissipation rate, and conversion coefficient of EV i , respectively.

The power consumption is limited as

$$p_i^{\text{EV}}(t) \begin{cases} \in [p_{i,\min}^{\text{EV}}, p_{i,\max}^{\text{EV}}], t \in [T_a, T_d] \\ = 0, \text{ otherwise} \end{cases} \quad ; (13)$$

$$|p_i^{\text{EV}}(t) - p_i^{\text{EV}}(t-1)| \in [0, \delta_{i,\max}^{\text{EV}}] \quad ; (14)$$

$$e_i^{\text{EV}}(t) \in [e_{i,\min}^{\text{EV}}, e_{i,\max}^{\text{EV}}] \quad , (15)$$

where $p_{i,\min}^{\text{EV}}$ and $p_{i,\max}^{\text{EV}}$ are the lower and upper bounds of the power consumption of EV i , respectively; $\delta_{i,\max}^{\text{EV}}$ is the ramping limitation of EV i ; $e_{i,\min}^{\text{EV}}$ and $e_{i,\max}^{\text{EV}}$ are the lower and upper bounds of the indoor temperature of EV i , respectively; T_a and T_d are the arrival and departure times of EV i , respectively.

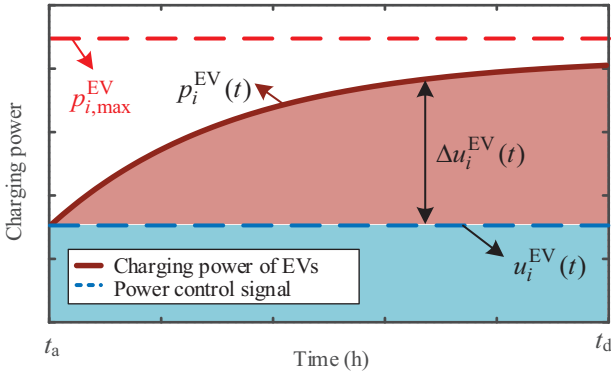


Fig. 2 Relationship between the charging power and control signal of EVs.

For a residential house, the EV is only connected to the charging pile between the arrival and departure times every day. During the charging time, the SoC of the EV is affected by the driver's experience and range anxiety (RA), which refers to the driver's anxiety that the EV range cannot cover the driving distance before the next charging. Thus, the RA is directly related to the SoC of the EV at the departure time. Time anxiety (TA) is introduced to describe the degree of driver anxiety regarding uncertain events during charging. According to this analysis, the driver's anxiety can be modeled by applying the expected SoC, $\tilde{e}_i^{\text{EV}}(t)$, during charging.

$$\tilde{e}_i^{\text{EV}}(t) = \frac{\alpha(e^{-\beta(t-t_a)/(t_d-t_a)} - 1)}{e^{-\beta} - 1}, t \in [T_a, T_d] \quad , (16)$$

where α and β are shape parameters.

A larger α leads to a higher SoC at the departure time, and a larger β indicates a higher SoC during charging; these exactly characterize the RA and TA, respectively.

The SoC of batteries and the temperature of houses can change flexibly within a certain range, and the residual energy can be released to support the EVs. Therefore, for the disaggregation operation strategy, we prioritize the charging requirements of the EVs. The power dispatched to the EV can be expressed as

$$p_i^{\text{EV}}(t) = \begin{cases} u_i^{\text{EV}}(t) + \Delta u_i^{\text{EV}}(t), p_i^{\text{EV}}(t) < p_{i,\max}^{\text{EV}} \\ p_{i,\max}^{\text{EV}}, p_i^{\text{EV}}(t) \geq p_{i,\max}^{\text{EV}} \end{cases} \quad ; (17)$$

$$\Delta u_i^{\text{EV}}(t) = \frac{\alpha(e^{-\beta a_1(t)} - 1)}{e^{-\beta} - 1} \frac{(e^{-\beta a_2(t)} - 1)}{e^{-\beta} - 1} \quad , (18)$$

where $u_i^{\text{EV}}(t)$ is the power control signal for EV i and discussed in detail in Section III. Fig. 2 depicts the relationship between the charging power, $p_i^{\text{EV}}(t)$, and control signal, $u_i^{\text{EV}}(t)$, of EVs. Variables a_1 and a_2 in (18) are respectively defined as

$$a_1(t) = \frac{t - T_a}{T_d - T_a}, t \in [T_a, T_d] \quad ; (19)$$

$$a_2(t) = \frac{\tilde{e}_i^{\text{EV}}(T_d) - e_i^{\text{EV}}(t)}{\tilde{e}_i^{\text{EV}}(T_d)}, t \in [T_a, T_d] \quad , (20)$$

where t_a and t_d are the arrival and departure times of EV i , respectively, and $\tilde{e}_i^{\text{EV}}(t)$ is the expected SoC of EV i during charging, as defined in (16).

III. PROPOSED METHOD

In this section, the proposed two-stage HEM algorithm is described. The first step involves designing an aggregation model for household appliances to reduce the high dimension of the problem. Subsequently, the decision-making process for the HEM problem is formalized as a Markov decision process (MDP). In the first stage of decision-making, a DRL control agent utilizing the SAC framework is employed to learn the optimal scheduling strategy by interacting with the environment and to produce aggregate control actions. In the second decision stage, these aggregate control actions are decomposed into actions for individual appliances by considering the user behavior.

A. Aggregation Model

The approximation parameters of aggregators can be categorized into two types:

1) $\{P_{M,\min}, P_{M,\max}, E_{M,\min}, E_{M,\max}, \delta_{M,\max}\}$, where $P_{M,\min}(t)$ and $P_{M,\max}(t)$ are the lower and upper bounds of the power consumption of aggregator M , respectively; $E_{M,\min}$ and $E_{M,\max}$ are the lower and upper bounds of the residual energy of aggregator M , respectively; $\delta_{M,\max}$ is the ramping limitation of aggregator M .

The variables are associated with the operation bounds, and the approximation is performed by directly summing the corresponding parameters.

2) $\{\theta_M, \eta_M, \sigma_M, A_{0,M}, A_{1,M}, A_{2,M}, B_M, \zeta_M\}$, which is

associated with the dynamic models. The approximation is performed by using the weighted average of the corresponding individual parameters. The detailed calculation method for the approximation parameters of the aggregators can be found in [8].

B. Control of the Aggregators (Decision Stage I)

In the first decision stage, a DRL control agent utilizing the SAC framework is employed to learn the optimal scheduling strategy by interacting with the environment and to produce the aggregate control actions. The decision-making process for the HEM problem is formalized as a MDP in which the operator optimizes the cumulative reward while operating in an uncertain environment. The MDP is defined by using a set of five tuples, $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma\}$, where \mathcal{S} denotes the set of environmental states observed by the DRL agent, \mathcal{A} denotes the set of actions, \mathcal{P} denotes the transition probability from any state $s \in \mathcal{S}$ to any $s' \in \mathcal{S}$ for any action $a \in \mathcal{A}$, \mathcal{R} denotes the immediate reward set, and $\gamma \in [0,1]$ denotes the discount rate that penalizes future rewards.

1) State: The operation problem is solved by the DRL agent based on the local observation, $s_{M,t}$:

$$s_{M,t} = \{P_M(t), E_M(t), P_{M,\min}(t), P_{M,\max}(t), P_D(t), \varepsilon(t)\},$$

where $P_M(t)$ and $E_M(t)$ are the power consumption and residual energy of aggregator M , respectively; $P_D(t)$ is the power demand; $\varepsilon(t)$ is the power deviation between the power demand and the actual power consumption, and

$$\varepsilon(t) = P_D(t) - \sum_{M=1}^{N_{\text{agg}}} P_M(t) - P_L(t).$$

2) Action: Action $a_{M,t} \in [0,1]$ is defined as the power output rate:

$$P_M(t) = P_{M,\min} + a_{M,t}(P_{M,\max} - P_{M,\min}). \quad (21)$$

Thus, the power consumption, $P_M(t)$, is limited within the $[P_{M,\min}, P_{M,\max}]$ range. The joint action at time step t can be expressed as $a_t = (a_{1,t}, a_{2,t}, \dots, a_{N,t})$.

3) State transition: The system state can transition from s_t to s_{t+1} with probability $P(s_t, s_{t+1}) = \Pr(s_{t+1}|s_t, a_t)$.

4) Reward: Because the control objective of the aggregators is to cover the power demand and minimize the operational cost, when the system state transitions from s_t to s_{t+1} , the DRL agent receives reward r_t :

$$r_t = \omega^{\text{pnl}} C^{\text{pnl}}(t) - \sum_{M=1}^{N_{\text{agg}}} \left[\omega^{\text{cost}} C_M^{\text{ope}}(t) + \omega^{\text{ADD}} C_M^{\text{ADD}}(t) \right], \quad (22)$$

where the reward function is divided into three parts: the cost of the tracking error, operation cost, and acoustic discomfort cost. ω^{pnl} , ω^{cost} , and ω^{ADD} are the weight coefficients for the three parts.

5) Objective function: The objective of the DRL agent is to maximize the expected value of rewards for the horizon of T time steps as

$$\max J = \mathbb{E}_{(s_t, a_t) \sim \pi} \left(\sum_{t=0}^T \gamma^t r(s_t, a_t) \mid s_t = s, a_t = a \right), \quad (23)$$

where π is the control policy that generates action a_t according to state s_t ; discounted rate γ determines the effects of the future reward on the current reward.

The SAC approach [9], which is a state-of-the-art continuous-control model-free reinforcement-learning

algorithm, is adopted to cope with high sample complexity and improve the stability of model-free DRL methods.

C. Disaggregation Control of Household Appliances (Decision Stage II)

In the second decision stage, the aggregator generates a schedule plan for the individual household appliances to follow the aggregate power schedule defined in the first decision stage. The process of decomposing an aggregate schedule into individual appliance schedules is defined below.

The aggregator meets this power schedule as closely as possible by using a disaggregation algorithm in three steps:

1) The disaggregation algorithm collects the appliances' state information, including SoCs, users' individual behaviors, and power ranges.

2) The individual power schedule plan is generated by the algorithm.

3) Each control power value is sent to the corresponding unit.

The iteration process guarantees that both the charging demand of the EVs and the power schedule defined in the first decision stage are satisfied as long as the batteries and HVAC systems have residual energy to release. If all residual energy is exhausted, the actual power consumption deviates from the aggregate schedule in the first decision stage. Subsequently, the schedule plan may need to be reformulated to meet the charging demand.

IV. CASE STUDIES

This section presents simulation case studies conducted to validate the effectiveness of the proposed operation and disaggregation strategies.

A. Environment Setup

Both the critic and actor networks consist of four serial fully connected layers. Each layer comprises 128 hidden units. The proposed aggregator operator is implemented by using Python and PyTorch. The proposed method is compared with two benchmarks: Perfect Information Optimum and Model Predictive Control.

B. Training Performance

In Fig. 3, the curves in the shaded region illustrate the average and real daily episode rewards. The total reward is divided into three parts: tracking-error cost, operational cost, and acoustic discomfort cost, which are listed as follows:

$$\begin{cases} r_{\text{pnl}} = \omega^{\text{pnl}} C^{\text{pnl}}(t) \\ r_{\text{cost}} = -\sum_{M=1}^{N_{\text{agg}}} \omega^{\text{cost}} C_M^{\text{ope}}(t) \\ r_{\text{ADD}} = -\sum_{M=1}^{N_{\text{agg}}} \omega^{\text{ADD}} C_M^{\text{ADD}}(t) \end{cases} \quad (24)$$

Fig. 3 displays that the proposed algorithm can learn a stable operation strategy by interacting with the environment within the first 6000 episodes. The results reveal that the SAC approach is effective for determining optimal policies for ES aggregators.

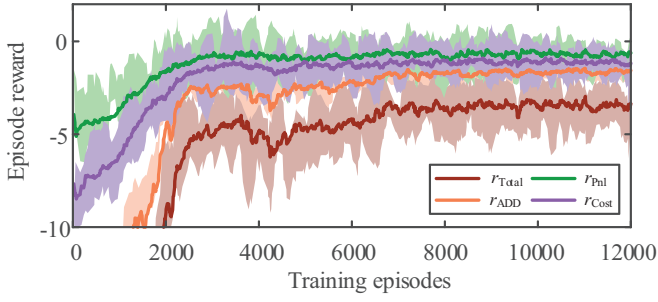


Fig. 3 Training performance of the proposed algorithm.

C. Operation Performance

The proposed algorithm improves the ability of the aggregators to track the power demand curve. The actual power of the aggregators and the power demand curve are shown in Fig. 4. The orange curve represents the power demand curve, and the actual power of the aggregators is indicated by using a differently colored bar chart. The results show that the DRL-based algorithm performs well in terms of tracking accuracy and stability; thus, the algorithm's effectiveness in solving the HEM problem is demonstrated. As depicted in Fig. 5, the tracking error decreases as the weight coefficient, ω^{pnl} , increases. The proposed algorithm not only reduces the operational cost but also improves the acoustic comfort level of the users by balancing the power demand and supply while considering the impact of residential users' behavior on the energy consumption.

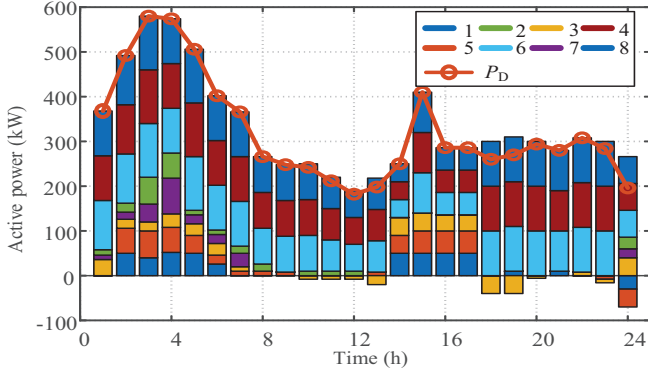


Fig. 4 Energy consumption schedules for the aggregators.

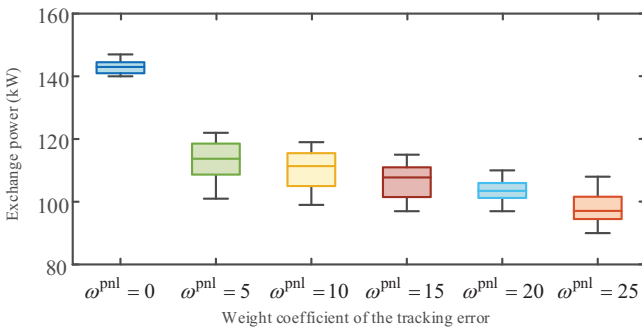


Fig. 5 Control performance with different weight coefficients.

The disaggregation performance of aggregator 1 is presented in Fig. 6. The simulation results display the effectiveness of the proposed two-stage HEM algorithm in controlling the indoor temperature and the SoCs of batteries and EVs. The proposed two-stage HEM algorithm allows the HVAC systems and batteries to release more energy to support the EV-charging demand when drivers are more concerned about uncertainty, thus reducing the anxiety penalty. As a result, the SoC levels of EVs are able to exceed the desired level of 80% before the anxious time, demonstrating the effective coordination between HVAC systems, batteries, and EVs to meet the power demand and maintain indoor temperature comfort (20–24 °C).

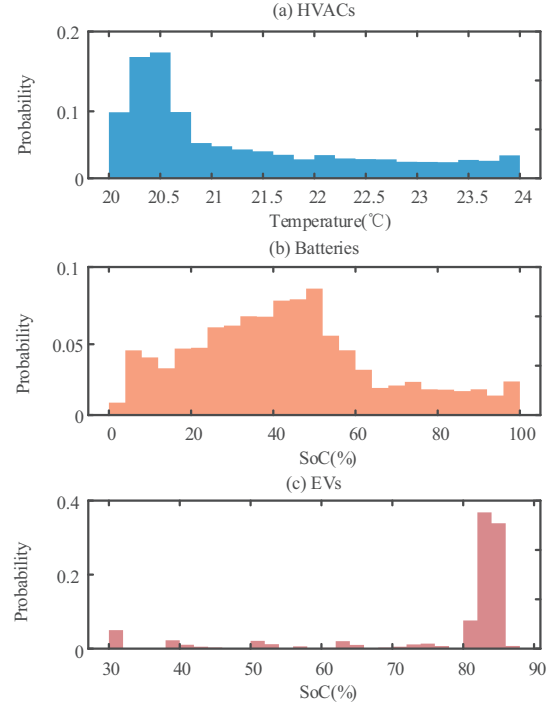


Fig. 6 Disaggregation results of individual appliances.

V. CONCLUSION

The focus in this study is the aggregator problem of defining a schedule plan for numerous residential users in the absence of an exact model of each energy source and load. The proposed DRL-based approach is evaluated by performing numerical simulations based on real-world data, and its ability to solve the HEM problem economically and effectively by covering the schedule plan from a superior grid is demonstrated. The results reveal the effectiveness of the proposed algorithm in controlling the indoor temperature, battery SoCs, and EV charging while considering individual user requirements and uncertainty. Furthermore, a comparison with existent methods indicates that the proposed algorithm outperforms other approaches in terms of energy efficiency and user comfort.

In conclusion, the proposed DRL-based approach provides a novel solution to the HEM problem because the behavioral uncertainties of residents and the grid constraints are considered. In the future, this work will be extended to the energy management of multi-energy buildings, including heat pumps and electric boilers.

ACKNOWLEDGMENT

This work was supported by the National Key Research and Development Program (No. 2023YFB2406600) and the National Natural Science Foundation of China (No. U22A6007 and No. 52222703).

REFERENCES

- [1] X. Chen, Y. Liu, Q. Wang, J. Lv, J. Wen, X. Chen, et. al., "Pathway Toward Carbon-neutral Electrical Systems in China by Mid-century with Negative CO₂ Abatement Costs Informed by High-resolution Modeling," *Joule*, vol. 5, no. 10, pp. 2715-2741, 2021.
- [2] Y. Chen, L. Xu, A. Egea-Álvarez, B. Marshall, M. H. Rahman, and A. D. Oluwole, "MMC Impedance Modeling and Interaction of Converters in Close Proximity," *IEEE J. Emerg. Sel. Top. Power Electron.*, vol. 9, no. 6, pp. 7223-7236, Dec. 2021.
- [3] S. Barja-Martinez, F. Rucker, M. Aragues-Penalba, R. Villafafila-Robles, I. Munne-Collado, and P. Lloret-Gallego, "A Novel Hybrid Home Energy Management System Considering Electricity Cost and Greenhouse Gas Emissions Minimization," *IEEE Trans. Ind. Appl.*, vol. 57, pp. 2782-2790, 2021.
- [4] L. Yu, Y. Sun, Z. Xu, C. Shen, D. Yue, T. Jiang, and X. Guan, "Multi-Agent Deep Reinforcement Learning for HVAC Control in Commercial Buildings," *IEEE Trans. Smart Grid*, vol. 12, pp. 407-419, Jan 2021.
- [5] S. Lee and D. Choi, "Federated Reinforcement Learning for Energy Management of Multiple Smart Homes With Distributed Energy Resources," *IEEE Trans. Ind. Inform.*, vol. 18, pp. 488-497, 2022.
- [6] S. S. Shuvo and Y. Yilmaz, "Home Energy Recommendation System (HERS): A Deep Reinforcement Learning Method Based on Residents' Feedback and Activity," *IEEE Trans. Smart Grid*, vol. 13, pp. 2812-2821, 2022.
- [7] Z. Yi, Y. Xu, W. Gu, and Z. Fei, "Distributed Model Predictive Control Based Secondary Frequency Regulation for a Microgrid With Massive Distributed Resources," *IEEE Trans. Sustain. Energy*, vol. 12, pp. 1078-1089, 2021.
- [8] L. Subramanian, V. Debusschere, H. B. Gooi, et al., "A Distributed Model Predictive Control Framework for Grid-friendly Distributed Energy Resources," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 727-738, Jan. 2021.
- [9] T. Haarnoja, A. Zhou, P. Abbeel, et al. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," *International Conference on Machine Learning*, Stockholm, Sweden, 2018, pp. 1861-1870.
- [10] J. Schulman et al., "Proximal Policy Optimization Algorithms," Jul. 2017. [Online]. Available: arXiv:1707.06347.