

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

Yujun Lin, Linfang Yan, Hongxun Hui, *Member, IEEE*, Qiufan Yang, Jianyu Zhou, Yin Chen, Xia Chen, *Senior Member, IEEE*, Jinyu Wen, *Member, IEEE*

**Abstract**—This paper addresses the home energy management (HEM) problem for a large number of residential houses, which can be regarded as a high-dimensional optimization problem. To cope with the high-dimensional issue, the concept of the aggregator is utilized to reduce the state and action space. And a two-stage deep reinforcement learning (DRL) based approach is proposed for the aggregators to track the schedule from the 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 (SAC) 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 the EV charging demand is quantitatively described by the driver's experience. An aggregate anxiety concept is introduced to characterize both the driver's anxiety on the EV's range and uncertain events. Finally, simulation studies verify the effectiveness of the proposed approach under dynamic user behaviors, and the comparisons also show the superiority of the proposed approach over the method mentioned in benchmarks.

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

## I. INTRODUCTION

IN recent years, to deal with the climate change and reduce carbon emissions, there has been a substantial increase in the deployment of distributed energy resources (DERs) in smart grids, containing roof-top solar photovoltaic (PV), electric vehicles (EVs), battery energy storage (ES), heating ventilation air conditioning (HVAC) [1]–[4], et al. These DERs have the potential to improve the energy efficiency and reliability of the electrical grid, as well as to provide additional income for households

through selling excess energy to the grid. However, the integration of these DERs into the grid also introduces new challenges in energy management. Especially in a wide region that comprises a considerable number of residential users, managing numerous home appliances and DERs is a critical factor that influences the economic and secure operation of the power systems. Fortunately, with the growing availability of smart energy devices and advanced metering equipment [5], residential users are now capable of home energy management (HEM). With the aid of HEM, the operations of home appliances and DERs can be reasonably arranged to reduce electricity costs and ensure the comfort of consumers [6].

Research on HEM for residential houses has largely fallen into two categories: model-based and model-free methods. Model-based approaches, commonly used for HEM optimization, involve creating a mathematical model and using optimization algorithms to find the best solution, as seen in studies addressing electricity costs, user satisfaction, and temperature comfort [7], unpredictability in user behavior [8], distribution system scheduling [9], and robust multi-objective problems [10]. However, these methods necessitate model construction and parameter identification. Developing a scheduling strategy based on modeling requires the construction of models and the identification of parameters. This process requires detailed domain knowledge, and the performance may deteriorate due to model inaccuracy. Model-free methods for HEM utilize a single-agent deep reinforcement learning (DRL) approach to learn an optimal schedule plan. DRL achieves this by gaining experience through repeated interactions with the environment, rather than relying on accurate knowledge of the environment. DRL can be classified into two main types based on the action space: discrete methods and continuous methods. The Deep-Q-Network (DQN) algorithm [11]–[13] is a classical discrete control method that combines Q-learning and DL. This approach improves the adaptability of RL algorithms in large-scale continuous state space problems by introducing function approximation. However, DQN is limited by its discrete action space and inability to handle random policy problems. Continuous DRL, e.g. twin delayed deep deterministic (TD3) algorithm [14] and soft-actor-critic (SAC) [15] algorithm, can provide fine-grained control in continuous action space. By parameterizing the policy function, the algorithm can update the parameters of the policy network based on gradient ascent.

Numerous existing studies have adopted DRL in the HEM problem and demonstrated its excellent control performance in

This work is supported by the National Key Research and Development Program (No. 2023YFB2406600), and the National Natural Science Foundation of China (No. U22A6007 and No. 52222703). (Corresponding author: Yin Chen.)

Y. Lin, X. Chen and J. Wen are with the State Key Laboratory of Advanced Electromagnetic Engineering and Technology, and School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China. (e-mail: yjlin20@foxmail.com, cxhust@foxmail.com, jinyu.wen@hust.edu.cn).

L. Yan is with the State Grid (Suzhou) City & Energy Research Institute, Suzhou Jiangsu, China. (e-mail: linfyan@foxmail.com).

H. Hui is with State Key Laboratory of Internet of Things for Smart City and Department of Electrical and Computer Engineering, University of Macau, Macao, 999078, China (e-mail: hongxunhui@um.edu.mo).

Q. Yang is with Central China Branch of State Grid Corporation of China, Wuhan, China (e-mail: qiufang@foxmail.com).

J. Zhou is with the College of Electrical Engineering, Sichuan University, Chengdu, China (email: jianyu\_z@foxmail.com).

Y. Chen is with the Department of Electronic and Electrical Engineering, University of Strathclyde, G1 1XW Glasgow, U.K. (e-mail: cystrath@163.com).

a dynamic environment. In [16], multi-agent DRL with an attention mechanism is utilized in HVAC control to minimize energy costs in a multi-zone commercial building. The mixed air temperature was used to describe the building temperature regulated by a group of HVACs. In [17], a DRL model is proposed that combines local HEM systems with a global server to optimize the scheduling of multiple smart homes and their appliances. The authors of [18] aim to improve the personalized comfort while reducing the electricity cost 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 performance in dynamic environments. However, the individual electric vehicle (EV) charging models have mostly been described solely by the arrival time, departure time, and desired battery energy, neglecting the distinct characteristics of drivers' individual behaviors.

Furthermore, when considering the HEM problem for a significant number of residential houses, the optimization problem becomes high-dimensional, which can result in a significant computational burden for system operators [19]. Especially in DRL methods, the scheduling problem of a large number of household appliances can result in a high-dimensional action space, which can lead to computational challenges. As the dimensionality of the action space increases, a correspondingly greater number of samples is required to obtain an optimal scheduling strategy [20]. Moreover, the HEM problem for a large number of residential houses is particularly challenging, due to the difficulty in accurately modeling the dynamic characteristics and operational patterns of individual household appliances. The high variability in user behavior, preferences, and lifestyles across multiple households further exacerbates this challenge, leading to a highly complex optimization problem. As a result, it is difficult to directly design a scheduling approach that can fully capture the diverse needs and preferences of individual users. To address this problem, the concept of virtual power plants (VPPs) [21] has been considered one of the most promising and effective methods to coordinate the controller with individual residential homes that contain various household appliances. All of the appliances are managed and scheduled by a single aggregator, or a scheduling coordinator. The resources within the VPP can have different ownership, but the aggregator serves as the common commercial interface between the market and grid operators [22]. The authors of [23] propose a coordinated operation strategy for a VPP that consists of multiple DER aggregators in order to inspire these DER aggregators to provide energy and regulation services. In [24], a robust active dynamic aggregation model for the multi-energy systems is proposed to describe the maximum feasible region. In order to address the challenge of high-dimensional action space, the authors in [25] set the energy demands of EVs with the same deadline and at the same bus to be the same. In [26], the concept of the aggregator is utilized to mitigate the curse of dimensionality. In the first decision phase, the aggregator purchases energy from the electricity market. In the second decision phase, a heuristic dispatch algorithm is proposed to generate the charging plan for each single EV. By dividing the HEM

problem into two decision-making stages, the dimensionality of the action space is significantly reduced. The EV aggregator model in [27] is employed to predict the regulation capacity, and individual EVs response to the charging schedule based on operating states and laxities. Most of the above work utilizes aggregators to deal with the operation problems of DERs like EVs or HVACs, but rarely focuses on the coordination of HEM for a large number of residential houses containing various household appliances.

In this article, we propose a solution that addresses the challenge of the high dimensionality of the HEM problem by utilizing an aggregation approach, and optimizing control actions at the aggregator level rather than considering individual appliance actions. Specifically, we propose a two-stage approach where, in the first decision stage, a DRL control agent is trained to learn an optimal scheduling strategy using the SAC framework, while in the second decision stage, the aggregate control actions are disaggregated to individual appliances by taking into account the users' individual behaviors. Portions of this work were presented in our previous paper at 2024 IEEE 7th Student Conference on Electric Machines and Systems (SCEMS) [28]. Compared to [28], the revised manuscript goes a step further by considering the driver's experience, charging preference, range anxiety, and time anxiety to describe the driver's individual behaviors for the charging problem of individual EVs in the households.

The main contributions of this paper are listed as follows.

- (1) Rather than managing each individual appliance's actions, we focus on controlling the aggregator's actions. We group individual household appliances in a region and control them as a collective through an aggregator, thereby reducing the dimensionality of the problem significantly.
- (2) The disaggregation method of aggregators with dynamic user behaviors is proposed to satisfy the electricity demand of household appliances. Unlike the existing studies [7]-[10], the household appliances compensate for each other to meet the power demand of the operator. Additionally, the individual charging scheduling plan of EVs is formulated considering the characteristics of drivers' individual behaviors.
- (3) A novel continuous SAC control framework is adopted to design the DRL-based approach for the scheduling problem of aggregators to obtain a fine-grained control. By repeatedly interacting with the environment, DRL can acquire experience and learn to optimize the scheduling without the need for constructing models or identifying parameters, making it a more flexible and adaptable approach.

The remainder of this paper is organized as follows. Section II introduces the system model. Section III introduces the proposed two-stage HEM method. Section IV provides the simulation results. Finally, section V concludes the article.

## II. SYSTEM OPERATION MODEL

Based on operating characteristics, the residential appliances can be classified into three categories: on-site power generation,

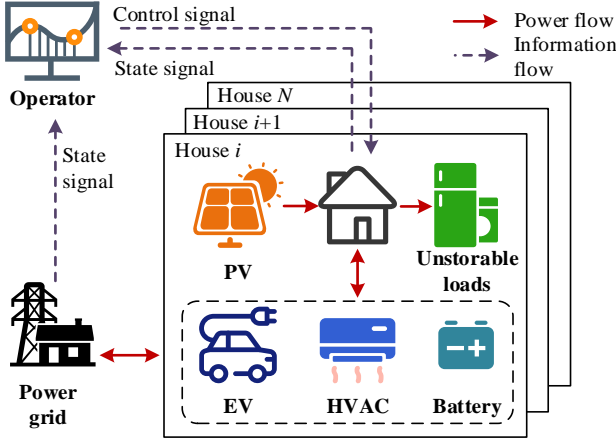


Fig. 1 Schematic of residential houses integrated with PV.

storable loads, and non-storable loads [29]. The on-site power generation provides power supply for the house, and the surplus energy can be sold to the electricity market or stored in the storable loads after meeting the household electricity demand. In this paper, we only consider a roof-top PV unit for a house as the sole source of on-site power generation for a household. The storable loads have the capacity to both store and release energy, and they can effectively shift and reduce demand, which contributes to load balancing in the grid. The storable loads mainly include HVAC systems, energy storage systems and smart lighting [30]. The non-storable loads, which typically consist of household appliances such as refrigerators, washing machines, and televisions, are unable to store energy and are unable to adjust their power consumption in real time. These loads are regarded as uncertain energy sources. Therefore, the main control objectives are storable loads, including batteries, HVACs, and EVs in this paper, as depicted in Fig. 1. In this section, the dynamic models of the storable loads in residential houses are presented.

#### A. Battery Model

In this paper, we take the small-capacity household batteries into consideration. The household batteries can release energy into the grid for profit, or support the HVACs and EVs at a high state of charge (SoC) level. The residual energy of 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  $\theta_i^{\text{BAT}}$  denotes the dissipation rate of battery  $i$ ;  $\eta_i^{\text{BAT}}$  denotes the conversion coefficient of battery  $i$ ;  $p_i^{\text{BAT}}$  denotes the power consumption of battery  $i$ ;  $e_i^{\text{BAT}}$  denotes the SoC of battery  $i$ .

And we also consider the operation constraints:

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

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

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

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 HVACs is to improve the comfort of residents by maintaining the indoor temperature within a reasonable range as

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

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

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

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

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

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

where  $p_i^{\text{HVAC}}$  denotes the power consumption of HVAC  $i$ ;  $e_i^{\text{HVAC}}$  denotes the indoor temperature of HVAC  $i$ ;  $\mathcal{G}_i^{\text{HVAC}}$  denotes the dissipation rate of HVAC  $i$ ;  $\eta_i^{\text{HVAC}}$  denotes the conversion coefficient of HVAC  $i$ ;  $\sigma_i^{\text{HVAC}}$  denotes the impact factor of the ambient temperature of HVAC  $i$ . The aforementioned factors are defined as

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

In addition to the above equality constraints, the state variables should be limited within a certain range as

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

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

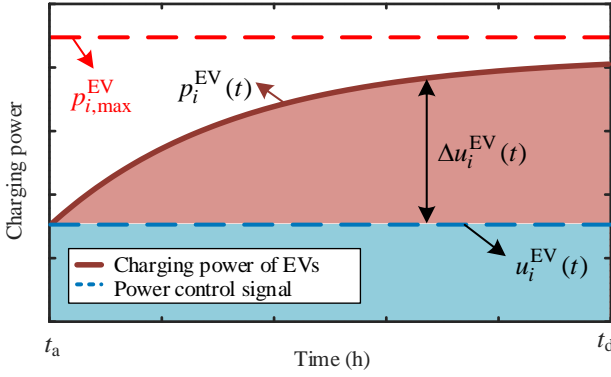


Fig. 2 The relationship between the charging power and control signal of EVs.

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

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

### C. EV Model

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

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

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

where  $p_i^{EV}$  denotes the power consumption of EV  $i$ ;  $e_i^{EV}$  denotes the SoC of EV  $i$ ;  $\theta_i^{EV}$  denotes the dissipation rate of EV  $i$ ;  $\eta_i^{EV}$  denotes the conversion coefficient of EV  $i$ .

The power consumption is limited by

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

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

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

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

For a residential house, the EV is only connected to the charging pile between the arrival time and the departure time every day. Within the charging time, the SoC of the EV is affected by the driving experience and driver's range anxiety (RA). RA refers to the anxiety degree of drivers that the EV range cannot cover the driving distance before the next charging. Thus RA is directly related to the SoC of EV at the departure

time. And the time anxiety (TA) is introduced to describe the anxiety degree of drivers about the uncertain events during charging. According to the above analysis, the driver's anxiety can be modeled by the expected SoC  $\tilde{e}_i^{EV}(t)$  during charging.

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

where  $\alpha$  and  $\beta$  are shape parameters.

A larger  $\alpha$  leads to a higher SoC at the departure time, and a larger  $\beta$  indicates a higher SoC during charging, which exactly characterize the RA and TA, respectively.

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

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

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

where  $u_i^{EV}(t)$  is the power control signal to EV  $i$ , which is introduced in detail in Sec III. Fig. 2 depicts the relationship between the charging power  $p_i^{EV}(t)$  and control signal  $u_i^{EV}(t)$  of EVs. And the variables  $a_1$  and  $a_2$  in (18) are defined as

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

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

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

## III. PROPOSED METHOD

In this section, the proposed two-stage HEM algorithm is introduced. The first step involves the design of an aggregation model for household appliances to reduce the dimension of the problem. Then the decision making process of the HEM problem is formalized as an MDP process. 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 produce the aggregate control actions. In the second decision stage, these aggregate control actions are then decomposed into actions for individual appliances while taking into account user behavior.

### A. Aggregation Model

The appliance model of residential house, introduced in Section II, containing batteries, HVACs, and EVs can be uniformly modeled as

$$e_i(t) = \theta_i e_i(t-1) + \eta_i p_i(t) \Delta t + \sigma_i T_{amb}(t) \quad (21)$$

$$p_i(t) \in [p_{i,min}, p_{i,max}] \quad (22)$$

$$|p_i(t) - p_i(t-1)| \in [0, \delta_{i,\max}] \quad (23)$$

$$e_i(t) \in [e_{i,\min}, e_{i,\max}] \quad (24)$$

where  $\theta_i$  denotes the dissipation rate of appliance  $i$ ;  $\eta_i$  denotes the conversion coefficient of appliance  $i$ ;  $p_i$  denotes the power consumption of appliance  $i$ ;  $e_i$  denotes the SoC of appliance  $i$ ;  $p_{i,\min}$  and  $p_{i,\max}$  are the lower and upper bounds of the power consumption of appliance  $i$ , respectively;  $\delta_{i,\min}$  is the ramping limitation of appliance  $i$ ;  $e_{i,\min}$  and  $e_{i,\max}$  are the lower and upper bounds of the indoor temperature of appliance  $i$ , respectively.

The operation costs of the individual appliances are given as a quadratic function [31] and can be written as

$$c_i^{\text{ope}}(p_i(t)) = a_{2,i} p_i(t)^2 + a_{1,i} p_i(t) + a_{0,i} \quad (25)$$

where  $a_{0,i}$ ,  $a_{1,i}$  and  $a_{2,i}$  are the operational cost coefficients of appliance  $i$ . Furthermore, the users' acoustic discomfort degree is considered [32]. The sound pressure level of each household appliance is determined by the per-unit value of active power as

$$c_i^{\text{ADD}}(p_i(t)) = (\zeta_i \frac{p_i(t)}{P_{i,\max}})^{0.67} \quad (26)$$

where  $\zeta_i$  is the acoustic weight coefficient of appliance  $i$ .

Thus the aggregator model can be obtained by calculating the weighted average value of the parameters as

$$E_M(t) = \theta_M E_M(t-1) + \eta_M P_M(t) \Delta t + \sigma_M T_{\text{amb}}(t) \quad (27)$$

$$P_M(t) \in [P_{M,\min}, P_{M,\max}] \quad (28)$$

$$|P_M(t) - P_M(t-1)| \in [0, \delta_{M,\max}] \quad (29)$$

$$E_M(t) \in [E_{M,\min}, E_{M,\max}] \quad (30)$$

where  $P_M$  and  $E_M$  are the power consumption and residual energy of aggregator  $M$ , respectively;  $P_{M,\min}$  and  $P_{M,\max}$  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  $M$ ;  $\theta_M$  denotes the dissipation rate of aggregator  $M$ ;  $\eta_M$  denotes the conversion coefficient of aggregator  $M$ ;  $\sigma_M$  denotes the impact factor of the ambient temperature of aggregator  $M$ .

The aggregate cost for tracking error is calculated as

$$C^{\text{pnl}}(P_M(t)) = - \left| P_D(t) - \sum_{M=1}^N P_M(t) - P_L(t) \right| \quad (31)$$

where  $P_D(t)$  is the power demand;  $P_L(t)$  is the power of local loads, including the roof-top PV units and other non-storable loads.

The aggregate operation cost of aggregator  $M$  is given as

$$C_M^{\text{ope}}(P_M(t)) = A_{2,M} P_M(t)^2 + A_{1,M} P_M(t) + A_{0,M} \quad (32)$$

where  $A_{0,M}$ ,  $A_{1,M}$  and  $A_{2,M}$  are the approximation of operational cost coefficients of aggregator  $M$ . The aggregate acoustic discomfort cost of aggregator  $M$  is given as

$$C_M^{\text{ADD}}(P_M(t)) = (\zeta_M \frac{P_M(t)}{P_{M,\max}})^{0.67} \quad (33)$$

where  $\zeta_M$  is the approximation of the acoustic weight coefficient of aggregator  $M$ .

The approximation parameters of aggregators can be divided

into two types:  $\{P_{M,\min}, P_{M,\max}, E_{M,\min}, E_{M,\max}, \delta_{M,\max}\}$ , which is associated with the operation bounds, the approximation is calculated by directly summing the corresponding parameters;  $\{\theta_M, \eta_M, \sigma_M, A_{0,M}, A_{1,M}, A_{2,M}, B_M, \zeta_M\}$ , which is associated with the dynamic models, the approximation is calculated by the weighted average of the corresponding individual parameters. The weights can be determined by the rated active power capacity of appliances [33].

### B. DRL 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 produce the aggregate control actions. The decision-making process for the HEM problem is formalized as a Markov decision process (MDP) in which the operator optimizes the cumulative reward while operating in an uncertain environment. The MDP is defined by a set of five tuples,  $\{S, A, P, R, \gamma\}$ .  $S$  denotes the set of environment states observed by the DRL agent.  $A$  denotes the set of actions.  $P$  denotes the transition probability from any state  $s \in S$  to any  $s' \in S$  for any action  $a \in A$ .  $R$  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;  $P_{M,\min}(t)$  and  $P_{M,\max}(t)$  are the lower and upper bounds of the power consumption 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  $\varepsilon(t) = P_D(t) - \sum_{M=1}^N P_M(t) - P_L(t)$ .

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

$$P_M(t) = P_{M,\min} + a_{M,t}(P_{M,\max} - P_{M,\min}) \quad (34)$$

In this way, the power consumption  $P_M(t)$  is naturally limited within the range  $[P_{M,\min}, P_{M,\max}]$ . 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 be transited from  $s_t$  to  $s_{t+1}$  with the probability  $P(s_t, s_{t+1}) = \Pr(s_{t+1} | s_t, a_t)$ .

4) Reward: Since the control objective of the aggregators is to cover the power demand and minimize the operation cost, when the system state is transited from  $s_t$  to  $s_{t+1}$ , the DRL agent will receive a reward  $r_t$ :

$$r_t = \omega^{\text{pnl}} C^{\text{pnl}}(t) - \sum_{M=1}^N \left[ \omega^{\text{cost}} C_M^{\text{ope}}(t) + \omega^{\text{ADD}} C_M^{\text{ADD}}(t) \right] \quad (35)$$

where the reward function is divided into three parts: the cost for tracking error, the operation cost, and the acoustic discomfort cost.  $\omega^{\text{pnl}}$ ,  $\omega^{\text{cost}}$ , and  $\omega^{\text{ADD}}$  are the weight coefficient 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

$$\max J = \mathbb{E}_{(s_t, a_t) \sim \pi} \left( \sum_{t=0}^T \gamma^t r(s_t, a_t) | s_t = s, a_t = a \right) \quad (36)$$

---

**Algorithm 1:** Training process of the proposed algorithm

---

**Input:**  $\phi, \theta$

1. Initialize the actor network  $\pi_\phi$  randomly.
  2. Initialize the critic networks  $q_\theta$  randomly.
  3. Initialize an empty replay buffer  $\mathbf{B}$ .
  4. **for** each episode **do**
  5.   **for** each state transition step **do**:
  6.     Obtain decision  $a_t$  according current  $s_t$  using  $\pi_\phi$ .
  7.     Execute  $a_t$ .
  8.     Obtain reward  $r_t$  and next state  $s_{t+1}$ .
  9.     Store the transition  $\{s_t, a_t, r_t, s_{t+1}\}$  into buffer  $\mathbf{B}$ .
  10.   **end for**
  11.   **for** each gradient step **do**:
  12.     Update weights of critic network using gradient in (41).
  13.     Update weights of actor network using gradient in (42).
  14.     Update temperature parameter using gradient in (43).
  15.     Update weights of target critic network.
  16.   **end for**
  17. **end for**
  18. **Output:**  $\phi, \theta$
- 

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

This paper adopts SAC [34], the state-of-the-art continuous control model-free RL algorithm, to cope with the high sample complexity and improve the stability of model-free DRL methods. In comparison to standard RL methods, the SAC incorporates the entropy value of the policy into the rewards as

$$\max J = \mathbb{E}_{(s_t, a_t) \sim \rho_\pi} \left( \sum_{t=0}^T \gamma^t [r(s_t, a_t) + \alpha \cdot H(\pi(\cdot | s_t))] \right) \quad (37)$$

where  $H(\pi(\cdot | s))$  is the entropy of policy  $\pi$ .  $\alpha$  is the temperature parameter, which determines the relative significance of entropy with respect to reward. The maximum entropy RL framework improves the exploration efficiency. The SAC algorithm endeavors to find a policy satisfying

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{(s_t, a_t) \sim \pi} \left( \sum_{t=0}^T \gamma^t [r(s_t, a_t) + \alpha \cdot H(\pi(\cdot | s_t))] \right) \quad (38)$$

The Q-value function in policy critic is calculated as follows

$$q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim \rho} V(s_{t+1}) \quad (39)$$

where  $V(s_t)$  is the soft state value function and is denoted by

$$V(s_t) = \mathbb{E}_{a_t \sim \pi} \left( q(s_t, a_t) - \alpha \log(\pi(a_t | s_t)) \right) \quad (40)$$

To accommodate the challenges posed by continuous state and action spaces, the soft q-function has been parameterized using a neural network, with the parameter  $\theta$  as  $q_\theta(s_t, a_t)$ . The parameters of the critic network are trained by minimizing the squared residual error through

$$J_q(\theta) = \mathbb{E}_{(s_t, a_t) \sim D} \left[ \frac{1}{2} \left( q_\theta(s_t, a_t) - (r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim \rho} V_\theta(s_{t+1})) \right)^2 \right] \quad (41)$$

where  $D$  is the experience replay buffer and  $V_\theta(s_{t+1})$  is the estimated state value using a target network.

The parameter of the actor network is trained by minimizing the expected Kullback-Leibler (KL) divergence as

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim D} \left[ \mathbb{E}_{a_t \sim \pi_\phi} \left[ \alpha \log \pi_\phi(a_t | s_t) - q_\theta(s_t, a_t) \right] \right] \quad (42)$$

The learning objective for the parameter  $\alpha$  is updated as

$$J(\alpha) = \mathbb{E}_{a_t \sim \pi_\phi} \left[ -\alpha \log \pi_\phi(a_t | s_t) - \alpha \bar{H} \right] \quad (43)$$

where  $\bar{H}$  is the minimum expected value of target entropy.

Based on the above theory, we propose a power dispatch strategy that leverages a learning architecture with the SAC framework. The actor network is responsible for determining the power dispatch action  $a_{M,t}$  based on the states  $s_{M,t}$ . The joint action  $a_t = (a_{1,t}, a_{2,t}, \dots, a_{N,t})$  and the current state determine the next state of the environment and the reward function of the DRL agent. The information regarding the state, action, and reward, denoted by  $s_{M,t}$ ,  $a_{M,t}$ ,  $r_{M,t}$ , and the next states  $s_{M,t+1}$ , are recorded and stored in the relay buffer  $D_i$  to be used for the training the DRL agent. The SAC algorithm consists of two neural networks, the actor network  $\pi_\phi$  and the soft critic network  $q_\theta$ , which are updated through interactions with the environment, as specified in (41)-(43). The pseudocode of the proposed DRL-based algorithm is depicted in pseudocode form in Algorithm 1. The algorithm consists of two identical actor networks, denoted as  $\pi_\phi$  and  $\pi_\phi$ , which interact with the environment and determine the power dispatch  $a_{M,t}$ . Additionally, the algorithm employs soft critic networks,  $q_\theta$ , which evaluate the quality of the actions  $a_{M,t}$  under  $s_{M,t}$  by approximating the action-value function using the soft Bellman equation. The critic networks  $q_\theta$  take the state-action pair  $(s_t, a_t)$  as input and output the estimated Q-values. The SAC algorithm uses a target network and a replay buffer to train the actor and critic networks. The target network is a copy of the actor or critic network that is used to compute the target value in the training process, and the replay buffer stores the state-action-reward-next state information for the purpose of training.

### C. Disaggregation Control of the household appliances (Decision Stage II)

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

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, power range and so on.

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

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

The pseudocode of the proposed disaggregation algorithm is shown in Algorithm 2. Initially, the power control signal of each storable load is allocated in proportion to the power range as

$$u_i(t) = \begin{cases} \frac{P_{i,\max}}{P_{M,\max}} P_M(t), & P_M(t) \geq 0 \\ \frac{P_{i,\min}}{P_{M,\min}} P_M(t), & P_M(t) < 0 \end{cases} \quad (44)$$

The disaggregation algorithm (44) guarantees the feasibility

---

**Algorithm 2:** Disaggregation algorithm

---

**Input:** the power control signal to aggregator  $M$   $P_M(t)$ , the driver's anxiety

1. **for** each DER  $i$  in aggregator  $M$  **do**
  2.     Eq. (44)
  3. **end for**
  4. **for** each EV  $i$  in aggregator  $M$  **do**
  5.     Eq. (45)(46)
  6. **end for**
  7. **for** each DER  $i$  in aggregator  $M$  **do**
  8.     **Iteration** (48)-(53)
  9. **end for**
  10. **Output:**  $p_i(t)$
- 

because the individual power value  $u_i(t)$  is always within the power range  $[p_{i,\min}, p_{i,\max}]$ . Furthermore, there are EVs' charging demand as

$$p_i^{\text{EV}}(t) = \begin{cases} u_i^{\text{EV}}(t) + \Delta u_i^{\text{EV}}(t), & p_i^{\text{EV}}(t) < p_{i,\max}^{\text{EV}} \\ p_{i,\max}^{\text{EV}}, & p_i^{\text{EV}}(t) \geq p_{i,\max}^{\text{EV}} \end{cases} \quad (45)$$

where the charging requirement of EVs is calculated as (18)-(20).

It can be observed that to satisfy the EVs' charging demand, there is extra power produced as

$$P_{\text{extra},M}(t) = \sum_{i \in \text{agg}_M} [p_i^{\text{EV}}(t) - u_i^{\text{EV}}(t)] \quad (46)$$

The extra power should be complemented by other energy resources in the same aggregator in proportion to the power range as

$$\Delta u_i(t) = \frac{p_{i,\max}}{P_{M,\max}} P_{\text{extra},M}(t) \quad (47)$$

However, after adding  $\Delta u_i(t)$  to each load, the feasibility guaranteed by (44) may be lost. Furthermore, we have not considered the SoC/temperature boundaries in the analysis above. Thus we design an iteration process to allocate the extra power while ensuring the feasibility. The iteration rules of the allocation approach are listed as follows:

1) The initial state  $u_i^{(0)}(t)$  is determined by (44), and the initial extra power  $P_{\text{extra},M}^{(0)}(t)$  is determined by (46). The initial state of the power range  $p_{i,\max}^{(0)} = p_{i,\max}$ ,  $P_{M,\max}^{(0)} = P_{M,\max}$ . The initial state of the power allocated to each individual load is calculated as

$$\Delta u_i^{(0)}(t) = \frac{p_{i,\max}^{(0)}}{P_{M,\max}^{(0)}} P_{\text{extra},M}^{(0)}(t) \quad (48)$$

2) During each iteration step (the number of iterations is named iter), for each storable load with  $p_{i,\max} > 0$ , we have

$$u_i^{(\text{iter}+1)}(t) = \begin{cases} u_i^{(\text{iter})}(t) + \Delta u_i^{(\text{iter})}(t), & u_i^{(\text{iter}+1)}(t) < p_{i,\max} \\ p_{i,\max}, & u_i^{(\text{iter}+1)}(t) \geq p_{i,\max} \\ (e_{i,\max} - e_i(t)) / \Delta t, & e_i(t+1) \geq e_{i,\max} \end{cases} \quad (49)$$

The loads with  $u_i^{(\text{iter}+1)}(t) \geq p_{i,\max}$  or  $e_i(t+1) \geq e_{i,\max}$  are

predicted to reach the power and SoC range, which are called the saturated loads. And we set the power range of the saturated loads to 0, so as to avoid allocating power to them.

$$p_{i,\max}^{(\text{iter}+1)} = 0, u_i^{(\text{iter}+1)}(t) \geq p_{i,\max} \cup e_i(t+1) \geq e_{i,\max} \quad (50)$$

$$P_{M,\max}^{(\text{iter}+1)} = \sum_{i \in \text{agg}_M} p_{i,\max}^{(\text{iter}+1)} \quad (51)$$

3) Calculate the extra power by minus  $u_i^{(\text{iter})}(t)$  of aggregator  $M$  as

$$P_{\text{extra},M}^{(\text{iter}+1)}(t) = P_{\text{extra},M}^{(\text{iter})}(t) - \sum_{i \in \text{agg}_M} u_i^{(\text{iter}+1)}(t) \quad (52)$$

$$\Delta u_i^{(\text{iter}+1)}(t) = \frac{p_{i,\max}^{(\text{iter}+1)}}{P_{M,\max}^{(\text{iter}+1)}} P_{\text{extra},M}^{(\text{iter}+1)}(t) \quad (53)$$

4) If  $P_{\text{extra},M}^{(\text{iter})}(t) = 0$  is satisfied in all aggregators, the extra power has been allocated completely. The output of the algorithm is sent to each units in step 3. Otherwise,  $\text{iter} = \text{iter} + 1$ , and return to 2).

From the iteration process, we can observe that the charging demand of EVs and the power schedule defined in the first decision stage are both satisfied as long as the batteries and HVACs have residual energy to release. If all the residual energy is exhausted, the actual power consumption will deviate from the aggregate schedule in the first decision stage. Then the schedule plan may need to be reformulated to meet the charging demand.

#### IV. CASE STUDIES

In this section, simulation case studies are conducted to validate the effectiveness of the proposed operation strategy and disaggregation strategy.

##### A. Environment Setup

In the simulation, 8 aggregators are considered, each of which contains approximately 1000 households. The simulation environment is based on real-world data and includes various parameters and operating cost coefficients, as shown in Table II. Each household is equipped with an EV, HVAC and an energy storage battery. The photovoltaic station's output power is sourced from an eastern Chinese plant, and the first month of each quarter (30 days) is used as the training set, totaling 120 days. The remaining days are used for evaluation purposes.

The critic network and the actor network both consist of four serial fully connected layers. Each layer comprises 128 hidden units. The proposed aggregator operator is implanted by Python with Pytorch. The hyper-parameters of the learning algorithm and proposed approach are listed in Table I.  $\alpha$  is the temperature parameter, which determines the relative significance of entropy with respect to reward.  $\alpha$  is set to  $10^{-2}$  to provide a good balance between exploration and exploitation, ensuring that the agent could effectively learn optimal strategies without becoming overly conservative.

The proposed method is compared with two benchmarks as:



Table I  
HYPER-PARAMETERS OF LEARNING ALGORITHM

Parameters	Value
Learning rate for actor	$10^{-3}$
Learning rate for critic	$10^{-2}$
Learning rate for $\alpha$	$10^{-2}$
Training episodes	12000
Discounted factor	0.99
Mini-batch size	256
Replay buffer size	10000
Step in each episode M	50

Table II  
PARAMETERS OF THE STORABLE LOADS

HVAC	
Parameters	Value
Min./Max. indoor temperature	20/24 °C
Initial temperature	(22, 0.5 <sup>2</sup> ) °C
Max. power consumption	(30, 5 <sup>2</sup> ) kW
Temperature dissipation rate	(0.98, 0.005 <sup>2</sup> ), no more than 1.
Conversion efficiency	(0.1, 0.001 <sup>2</sup> ), no less than 0.08.
EV	
SoC range	[0, 1]
Battery capacity	(50, 0.5 <sup>2</sup> ) kWh
Max. charge/discharge power	(12, 0.5 <sup>2</sup> ) kW
Initial SoC at the arrival time	(30, 5 <sup>2</sup> ) %
Energy dissipation rate	(0.99, 0.001 <sup>2</sup> ), no more than 1.
Conversion efficiency	(0.95, 0.01 <sup>2</sup> ), no more than 1.
Battery	
SoC range	[0, 1]
Battery capacity	(50, 0.5 <sup>2</sup> ) kWh
Max. charge/discharge power	(15, 5 <sup>2</sup> ) kW
Initial SoC	(50, 5 <sup>2</sup> ) %
Energy dissipation rate	(0.99, 0.001 <sup>2</sup> ), no more than 1.
Conversion efficiency	(0.95, 0.01 <sup>2</sup> ), no more than 1.

1) Perfect Information Optimum (PIO): PIO refers to an offline optimization method, where all the uncertain information, including future outdoor temperature, driver behavior, and aggregator parameters, can be accurately predicted, enabling the optimal operation problem to be resolved using an optimization solver such as Cplex.

2) Model Predictive Control (MPC): MPC relies on the prediction of environment information for a short-term future and assumes that the controller has knowledge of the distribution of each equipment's parameters, which are predicted based on the average value. As a result, the operation problem is transformed into a deterministic mathematical model and only the first step's schedule is executed.

### B. Training Performance

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

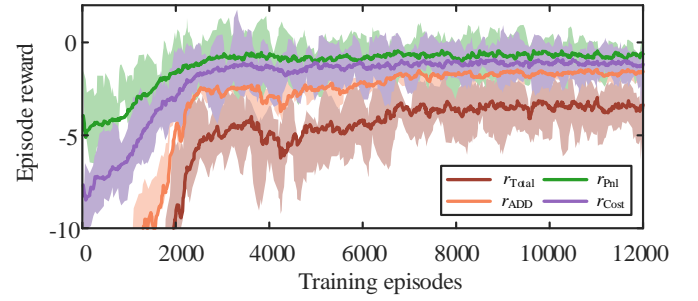


Fig. 3 Training performance of the proposed SAM algorithm.

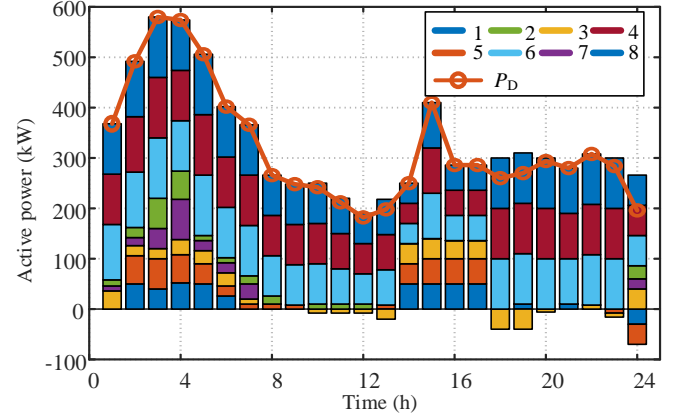


Fig. 4 The energy consumption schedules for the aggregators.

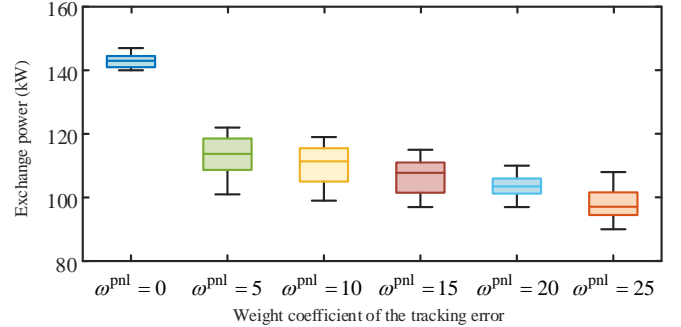


Fig. 5 Control performance with different weight coefficients.

$$\begin{cases} r_{\text{pnl}} = \omega^{\text{pnl}} C^{\text{pnl}}(t) \\ r_{\text{cost}} = -\sum_{M=1}^{N_{\text{agg}}} \omega^{\text{cost}} C_M^{\text{ope}}(t) \\ r_{\text{ADD}} = -\sum_{M=1}^{N_{\text{agg}}} \omega^{\text{ADD}} C_M^{\text{ADD}}(t) \end{cases} \quad (54)$$

Fig. 3 depicts that the proposed algorithm is capable of learning a stable operation strategy through interaction with the environment within the first 6000 episodes. The above results demonstrate that the SAC approach is effective in finding optimal policies for the ES aggregators.

### C. Operation Performance

The proposed algorithm has improved the ability of the aggregators to track the power demand curve. The actual power of aggregators and the power demand curve are shown in Fig. 4. The orange curve denotes the power demand curve, and the



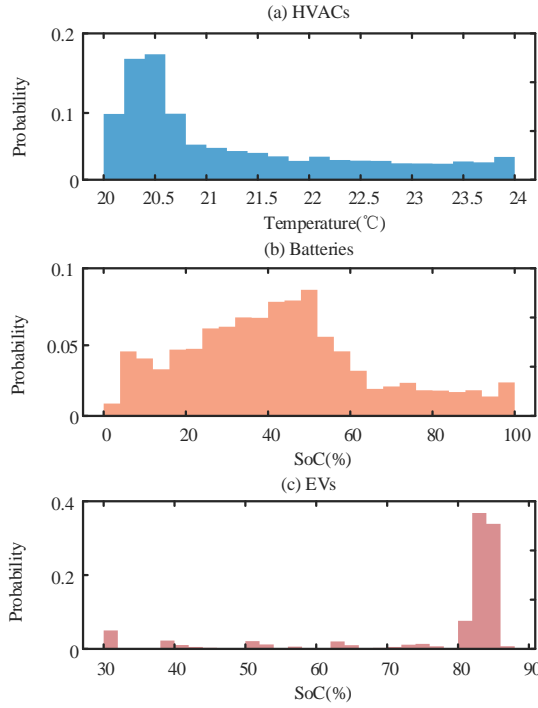


Fig. 6 Disaggregation results of individual appliances.

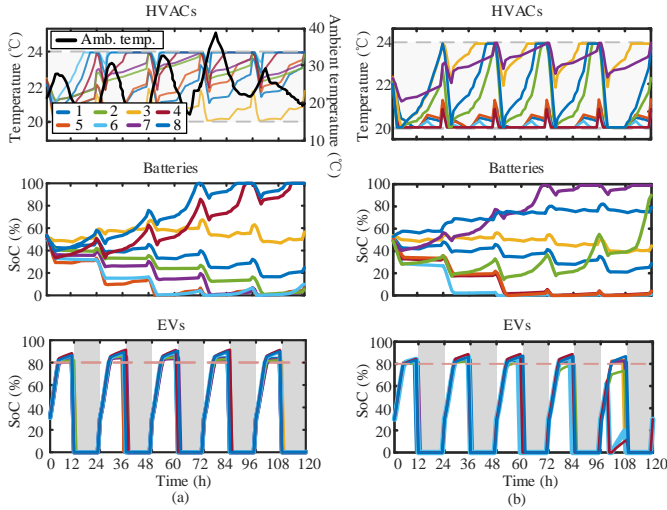


Fig. 7 Disaggregation results of individual ESs: (a) the proposed method, (b) the MPC method.

actual power of aggregators are illustrated by different colored bar chart. The results show that the DRL-based algorithm has good performance in terms of tracking accuracy and stability, demonstrating its effectiveness in solving the HEM problem. As shown in Fig. 5, the tracking error decreases as the weight coefficient  $\omega^{\text{pnl}}$  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 taking into account the impact of residential users' behavior on the energy consumption.

The disaggregation performance of aggregator 1 is presented in Fig. 6. The simulation results demonstrate the effectiveness

of the proposed two-stage HEM algorithm in controlling the indoor temperature, as well as the SoCs of batteries and EVs. The proposed two-stage HEM algorithm allows the HVACs and batteries to release more energy to support EV charging 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 of HVACs, batteries, and EVs to meet power demand and maintain indoor temperature comfort (20~24 °C).

The simulation results in Fig. 7 show that when utilizing the MPC method, the indoor temperature of residential homes belonging to aggregators 1, 4, 5, and 6 remain around the lower bound due to prediction errors, and the SoCs of batteries for aggregators 4, 5, and 6 decrease to 0 after  $t = 48\text{h}$ . These findings suggest that the active power dispatched to these aggregators by the predictive controller is insufficient to maintain normal operations for residential users. Additionally, on the 5<sup>th</sup> day, i.e. after  $t = 96\text{h}$ , the EVs of the aggregators 4, 5 and 6 fail to reach the desired SoC level of 80% before the anxious time.

#### D. Comparison with Benchmarks

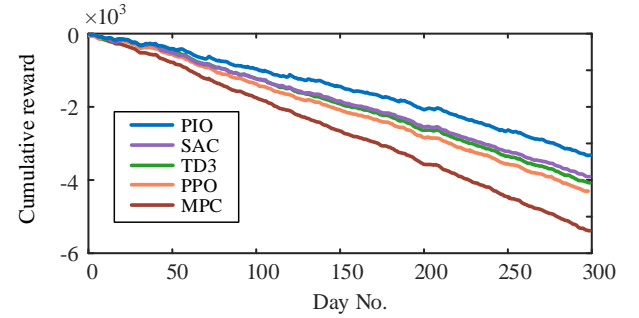


Fig. 8 Comparison of cumulative reward.

The comparison of the performance of the proposed algorithms and the benchmarks can also be shown in terms of cumulative reward in Fig. 8. In addition to PIO and MPC, we have also taken into consideration TD3 [14] and PPO [35] as benchmarks. The test set contains 300 days from the database. The PIO algorithm has the best performance, which is considered the ideal result for comparison. The MPC method is highly depended on the predictive accuracy, and we have exhibited in Fig. 7 (b) that in the highly uncertain environment, the MPC method cannot guarantee the long-term operation of ES aggregators. Consequently, the cumulative reward of the MPC method is the lowest. The control effectiveness of SAC exceeds that of other off-policy and on-policy DRL algorithms, including TD3 and PPO. SAC performs well and approaches the ideal outcome of PIO, with its advantage stemming from non-policy updates and the maximum entropy framework.

#### V. CONCLUSION

This paper focuses on aggregators' problem of defining a schedule plan for a large number of residential users, in absence of an exact model of each energy source and load. The proposed DRL-based approach is evaluated using numerical simulations

based on real-world data, demonstrating its ability to solve the HEM problem economically and effectively by covering the schedule plan from the superior grid. The results demonstrate the effectiveness of the proposed algorithm in controlling indoor temperature, battery SoCs, and EV charging, while considering individual user requirements and uncertainty. Furthermore, a comparison with existing 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, considering the behavior uncertainties of the residents and the grid constraints. In the future, this work will be further extended to the energy management of multi-energy buildings including heat pumps and electric boilers.

## REFERENCES

- [1] X. Chen, Y. Liu, Q. Wang, J. Lv, J. Wen, and X. Chen, "Pathway toward carbon-neutral electrical systems in China by mid-century with negative CO<sub>2</sub> 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 Journal of Emerging and Selected Topics in Power Electronics*, vol. 9, no. 6, pp. 7223-7236, Dec. 2021.
- [3] S. Wang, J. Zhai, H. Hui, "Optimal Energy Flow in Integrated Electricity and Gas Systems with Injection of Alternative Gas," *IEEE Trans. Sustain. Energy*, vol. 14, no. 3, pp. 1540-1557, Dec. 2023.
- [4] Y. Chen, L. Xu, A. Egea-Álvarez and B. Marshall, "Accurate and General Small-Signal Impedance Model of LCC-HVDC in Sequence Frame," *IEEE Trans. Power Del.*, vol. 38, no. 6, pp. 4226-4241, Aug. 2023.
- [5] S. Barja-Martinez, F. Rucker, M. Aragües-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.
- [6] M. Rastegar, "Impacts of residential energy management on reliability of distribution systems considering a customer satisfaction model," *IEEE Trans. Power Syst.*, vol. 33, no. 6, pp. 6062-6073, Nov. 2018.
- [7] A. Anvari-Moghaddam, H. Monsef and A. Rahimi-Kian, "Optimal Smart Home Energy Management Considering Energy Saving and a Comfortable Lifestyle," *IEEE Trans. Smart Grid*, vol. 6, pp. 324-332, 2015.
- [8] S. Dash, R. Sodhi and B. Sodhi, "A Bilayer Clustered-Priority-Driven Energy Management Model for Inclining Block Rate Tariff Environment," *IEEE Trans. Indus. Informat.*, vol. 18, pp. 3936-3946, 2022.
- [9] A. Rajaei, S. Fattaheian-Dehkordi, M. Fotuhi-Firuzabad, M. Moeini-Aghaie, and M. Lehtonen, "Developing a Distributed Robust Energy Management Framework for Active Distribution Systems," *IEEE Trans. Sustain. Energy*, vol. 12, pp. 1891-1902, 2021.
- [10] M. Tostado-Véliz, S. Kamel, F. Aymen, and F. Jurado, "A novel hybrid lexicographic-IGDT methodology for robust multi-objective solution of home energy management systems," *Energy*, vol. 253, p. 124146, 2022.
- [11] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246-5257, Sep. 2019.
- [12] H. Li, Z. Wan and H. He, "Constrained EV Charging Scheduling Based on Safe Deep Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 11, pp. 2427-2439, Jan. 2020.
- [13] E. Mocanu, D. C. Mocanu, P.H. Nguyen, A. Liotta, M. E. Webber, and M. Gibescu, "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698-3708, Jul. 2019.
- [14] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," *International Conference on Machine Learning*, PMLR, 2018: 1587-1596.
- [15] T. Li et al., "Mechanism Analysis and Real-time Control of Energy Storage Based Grid Power Oscillation Damping: A Soft Actor-Critic Approach," *IEEE Trans. Sustain. Energy*, vol. 12, no. 4, pp. 1915-1926, Oct. 2021.
- [16] 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.
- [17] S. Lee and D. Choi, "Federated Reinforcement Learning for Energy Management of Multiple Smart Homes With Distributed Energy Resources," *IEEE Trans. Indus. Informat.*, vol. 18, pp. 488-497, 2022.
- [18] 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, Dec. 2022.
- [19] 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.
- [20] L. Yang and M. Wang, "Sample-optimal parametric Q-learning using linearly additive features," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6995-7004.
- [21] E. Mashhour, S. M. Moghaddas-Tafreshi, "Bidding strategy of virtual power plant for participating in energy and spinning reserve markets—part I: problem formulation," *IEEE Trans. Power Syst.*, 26, (2), pp. 949-956, Dec. 2011.
- [22] M. Giuntoli and D. Poli, "Optimized Thermal and Electrical Scheduling of a Large Scale Virtual Power Plant in the Presence of Energy Storages," *IEEE Trans. Smart Grid*, vol. 4, pp. 942-955, Jan. 2013.
- [23] Z. Yi, Y. Xu, H. Wang, and L. Sang, "Coordinated Operation Strategy for a Virtual Power Plant With Multiple DER Aggregators," *IEEE Trans. Sustain. Energy*, vol. 12, pp. 2445-2458, Jan. 2021.
- [24] H. Zhao, B. Wang, X. Wang, Z. Pan, H. Sun, Z. Liu, and Q. Guo, "Active Dynamic Aggregation Model for Distributed Integrated Energy System as Virtual Power Plant," *Journal of Modern Power Systems and Clean Energy*, vol. 8, pp. 831-840, Jan. 2020.
- [25] J. Jin and Y. Xu, "Optimal Policy Characterization Enhanced Actor-Critic Approach for Electric Vehicle Charging Scheduling in a Power Distribution Network," *IEEE Trans. Smart Grid*, vol. 12, pp. 1416-1428, Jan. 2021.
- [26] S. Vandael, B. Claessens, D. Ernst, T. Holvoet, and G. Deconinck, "Reinforcement Learning of Heuristic EV Fleet Charging in a Day-Ahead Electricity Market," *IEEE Trans. Smart Grid*, vol. 6, pp. 1795-1805, Jan. 2015.
- [27] M. Wang, Y. Mu, Q. Shi, H. Jia, and F. Li, "Electric Vehicle Aggregator Modeling and Control for Frequency Regulation Considering Progressive State Recovery," *IEEE Trans. Smart Grid*, vol. 11, pp. 4176-4189, Jan. 2020.
- [28] Y. Lin, L. Yan, H. Hui, Y. Chen, X. Chen, and J. Wen, "Using Deep Reinforcement Learning in Optimal Energy Management for Residential House Aggregators with Uncertain User Behaviors," in *Proc. 2024 IEEE 7th Student Conference on Electric Machines and Systems (SCEMS)*, IEEE, Macau, China, 2024: 1-6.
- [29] W. Zou, Y. Sun, D. Gao, X. Zhang, J. Liu, "A review on integration of surging plug-in electric vehicles charging in energy-flexible buildings: Impacts analysis, collaborative management technologies, and future perspective," *Applied Energy*, vol. 331, 2023, p. 120393.
- [30] S. Wang, H. Hui, Y. Ding, C. Ye and M. Zheng, "Operational Reliability Evaluation of Urban Multi-Energy Systems With Equivalent Energy Storage," *IEEE Trans. Indus. Appl.*, vol. 59, no. 2, pp. 2186-2201, Mar. 2023.
- [31] Z. Wang, W. Wu and B. Zhang, "A Distributed Quasi-Newton Method for Droop-Free Primary Frequency Control in Autonomous Microgrids," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 2214-2223, Jan. 2018.
- [32] Y. Deng, Y. Zhang, F. Luo, and G. Ranzi, "Many-Objective HEMS Based on Multi-Scale Occupant Satisfaction Modelling and Second-Life BESS Utilization," *IEEE Trans. Sustain. Energy*, vol. 13, pp. 934-947, Dec. 2022.
- [33] L. Subramanian, V. Debusschere, H. B. Gooi, and N. Hadsaid, "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.
- [34] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "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.
- [35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," Jul. 2017. [Online]. Available: arXiv:1707.06347.



**Yujun Lin** received the B.S. degrees in electrical engineering from Chongqing University, Chongqing, China, in 2020. He is currently working toward his Ph.D. degree in electrical engineering at Huazhong University of Science and Technology (HUST), Wuhan, China. His research interests include the distributed control and optimization of energy storage clusters.



**Linfang Yan** received the B.E. and Ph.D. degrees in electrical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2017 and 2022, respectively. He is currently with the State Grid (Suzhou) City & Energy Research Institute, Suzhou Jiangsu, China. His research interests include deep reinforcement learning, electric vehicle charging, smart home, P2P energy trading, distributed control and hybrid energy storage.



**Hongxun Hui** (Member, IEEE) received the B.E. and Ph.D. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2015 and 2020, respectively. From 2018 to 2019, he was a Visiting Scholar with the Advanced Research Institute, Virginia Tech, Blacksburg, VA, USA, and the CURENT Center, University of Tennessee, Knoxville, TN, USA. He is currently a Research Assistant Professor with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao, China. His research interests include optimization and control of power system, demand response, and Internet of Things technologies for smart energy.



**Qiufan Yang** received the B.E. and Ph.D. degrees in electrical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2019 and 2024, respectively. He is currently Central China Branch of State Grid Corporation of China, Wuhan, China. His research interests include dc microgrids, distributed control, hybrid energy storage.



**Jianyu Zhou** received the B.E. and Ph.D. degrees in electrical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2017 and 2022, respectively. He is currently with the College of Electrical Engineering, Sichuan University, Chengdu, China. His research interests include dc microgrids, distributed control, hybrid energy storage.



**Yin Chen** received the B.S. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2009, the M.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 2014, and the Ph.D. degree in electrical engineering from University of Strathclyde, Glasgow, U.K., in 2020. He is currently an Associate Researcher with University of Strathclyde. His research interests include modeling of power electronic converters, grid integration of renewable power, and stability analysis of the HVDC transmission systems.



**Xia Chen** (Senior Member, IEEE) received the B.S. degree in power system and its automaton from the Wuhan University of Technology, Wuhan, China, in 2006, and the M.S. and Ph.D. degrees in electrical engineering from the Huazhong University of Science and Technology (HUST), Wuhan, in 2008 and 2012, respectively. She was a Postdoctoral Research Fellow with the University of Hong Kong, Hong Kong, from 2012 to 2015. She is currently a Professor with the School of Electrical

and Electronic Engineering, HUST. Her research interests include energy storage control and operation, renewable energy integration technologies, and new smart grid device.



**Jinyu Wen** (Member, IEEE) received the B.Eng. and Ph.D. degrees all in electrical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 1992 and 1998, respectively. He was a visiting student from 1996 to 1997 and research fellow from 2002 to 2003 all at the University of Liverpool, UK, and a senior visiting researcher at the University of Texas at Arlington, USA in 2010. From 1998 to 2002 he was a director engineer in XJ Electric Co. Ltd. in China. In 2003 he joined the HUST and now is a professor at HUST. His current research interests include renewable energy integration, energy storage application, dc grid, and power system operation and control.