Prioritized sum-tree experience replay TD3 DRL-based online energy management of a residential microgrid

Can Wang[a,*], Jiaheng Zhang[a], Aoqi Wang[b], Zhen Wang[a], Nan Yang[a], Zhuoli Zhao[d], Chun Sing Lai[c] and Loi Lei Lai[d]

[a]*the College of Electrical Engineering and New Energy, China Three Gorges University, Yichang 443002, China*
[b]*the State Grid Jinhua Power Supply Company, Jinhua 321017, China*
[c]*the Department of Electronic and Electrical Engineering, Brunel University London, London, UB8 3PH, UK*
[d]*the Department of Electrical Engineering, School of Automation, Guangdong University of Technology, Guangzhou, 510006, China*

ARTICLE INFO

Keywords:
Deep reinforcement learning (DRL)
Residential microgrid (RM)
Energy management
Uncertainty

ABSTRACT

Online energy management utilizing the real-time information of a residential microgrid (RM) can make full use of renewable energy and demand-side resources at the residential level. However, existing online energy management methods for RMs have poor robustness against environmental changes, which limits their applicability in highly uncertain scenarios. To address this, a novel online energy management method based on the prioritized sum-tree experience replay strategy with a double delayed deep deterministic policy gradient (PSTER-TD3) is proposed in this paper. First, we formulate the sequential scheduling decision problem as a Markov decision process (MDP) problem with the objective of minimizing residential energy costs while simultaneously ensuring household thermal comfort and minimizing range anxiety for electric vehicle usage. Then, using the proposed method, we determine the optimal online scheduling strategy under this objective. By integrating the prioritized experience replay strategy of the summation tree structure into TD3, the agent is able to learn the optimal scheduling strategy in complex environments, and its optimization performance and policy learning efficiency are significantly improved. In addition, its ability to handle multidimensional continuous action spaces helps achieve finer-grained optimization for RMs. The case study results demonstrate that the proposed method can effectively reduce the energy costs of residential microgrids while satisfying household thermal comfort requirements and reducing range anxiety for electric vehicle usage. Moreover, the optimization performance of the proposed method is robust when the uncertainty factors fluctuate violently in the environment.

1. Introduction

To address energy crises and environmental pollution issues, the penetration rate of distributed energy generation (DEG) at the residential level is being continuously increased [1]. As a small-scale power generation and distribution network that integrates household loads, renewable energy sources (RESs), and energy storage systems (ESSs), the residential microgrid (RM) is considered an effective solution for utilizing DEG [2], [3]. Given the flexible scheduling and rapid response characteristics of controllable RM units, an RM integrated with advanced communication and information technology can schedule residential demand-side resources in real time through a coordinated home energy management system (HEMS) [4], [5]. However, the intermittency and randomness of RES output, as well as the uncertainty of electricity market prices and household load energy demand, make the formulation of RM energy management plans much more difficult. Online energy management is a key technology that allows real-time online adjustment of controllable units based on the collected real-time status data of RMs. This approach balances the supply and demand of the system in real time to meet the energy comfort needs of households and optimize the operational costs of RMs [6], [7]. Therefore, providing high-quality online energy management strategies for RMs is crucial for the economical, efficient and stable operation of RMs.

Online energy management methods for RMs have been studied extensively, and a typical online optimization method is feedback-based model predictive control (MPC) [8]. This method improves robustness through rolling optimization. However, its performance is still directly affected by short-term prediction errors. To address this, some researchers have introduced the distributed ADMM algorithm [9] and Lyapunov optimization into the online optimization of networked microgrids (MGs) and smart homes; these methods do not rely on prediction information [10]. However, these model-based methods share the following shortcomings when they are applied to solve online optimization problems in RMs: (1) The performance of model-based online optimization methods depends on the specific case environment when the methods are applied to RMs considering the thermal comfort of households. Thus, their generalizability is poor [11]. (2) The reliability of these methods is limited by the model accuracy because the online optimization performance depends on the established dynamic physical model, and building

* Corresponding author.
E-mail address: xfcancan@163.com (C. Wang).

accurate models of RM components and operating environments is very difficult [12].

To overcome these drawbacks, researchers have become increasingly interested in utilizing deep reinforcement learning (DRL) techniques to design optimal energy scheduling strategies. DRL combines the universal function approximation capability of deep neural networks (DNNs) and the decision-making ability of reinforcement learning (RL) to form a model-free, data-driven approach [13]. Unlike in model-based methods, DRL agents can adaptively learn optimal action policies using feedback information from continuous interactions with the environment in scenarios without prior knowledge and explicit models. The trained DRL agent can make optimal action decisions in unknown environments in milliseconds based on the learned policy. Although DRL methods have been extensively studied and have achieved significant success in areas such as energy management, we must recognize some of their inherent drawbacks. To ensure appropriate performance, DRL methods often require large-scale annotated datasets to train the system. In situations where data collection costs are high, this can significantly increase computational expenses. Additionally, large-scale training demands substantial computational resources, especially when dealing with deep neural networks, leading to longer training times. Although DRL methods entail significant amounts of data and computational effort compared with those of traditional model-driven methods, they avoid the need to make specific models for complex situations such as highly nonlinear, partially observable or stochastically perturbed cases, so they are highly practical. Particularly in addressing complex problems with continuous action spaces, the autonomous exploration and learning capabilities of DRL methods, as well as their adaptation and generalization capabilities, are difficult to match with traditional methods.

Recently, value-function-based DRL methods have been successfully applied for solving optimal energy management problems. In [14], an online energy optimization method based on a deep Q network (DQN) was proposed. The method achieves optimal scheduling of electrical equipment in residential buildings. However, when noise and errors are present, the overestimation of the action value (Q value) often negatively impacts the convergence performance of the DQN [15]. To address this issue, in [16], an interruptible load demand response method based on dueling DQNs was developed and used to reduce the peak load and operating costs of the system. However, these methods use DNNs that often generate discrete Q-value estimates instead of continuous actions. As a result, they are applicable only to problems with discrete and low-dimensional action spaces [13]. The control decisions in RM online energy management are often multidimensional and continuous (such as in the charging/discharging of energy storage systems (ESSs)), and discretely processing the action space will distort the environmental feedback information received by the DRL agent and simultaneously limit the feasible domain of the action space [17].

Given these limitations, researchers have started exploring the application of policy-based DRL methods to address energy management problems with continuous action spaces. These methods utilize DNNs to directly output deterministic action values or probabilities for executing actions, enabling effective handling of continuous action problems and achieving finer-grained energy management. In [18], a home appliance scheduling method that applies trust region policy optimization (TRPO) was designed to participate in demand response programs with real-time electricity prices. However, computing conjugate gradients makes the computational process of this method highly complex. To improve the computational efficiency of the model, a real-time energy management method for microgrids based on proximal policy optimization (PPO) was proposed in [19]. However, the target policy and action policy of the on-policy approach are the same, which restricts the exploration capability of the agent, causing it to learn suboptimal action policies. The off-policy DRL method separates the target policy and the action policy and can obtain the global optimal value while maintaining exploration capability. In [20], a smart home scheduling method based on the deep deterministic policy gradient (DDPG) was developed. This method was used to minimize electricity costs and ensure household thermal comfort. However, similar to the DQN, the DDPG also overestimates Q-values.

Given the issues with the aforementioned energy management methods, an online RM energy management method based on the prioritized sum-tree experience replay strategy with a double delayed deep deterministic policy gradient (PSTER-TD3) is proposed in this paper. Unlike MPC methods, PSTER-TD3 directly approximates the optimal control policy through continuous interaction with the environment based on retrospective feedback. It does not rely on predictive inputs or modeling of environmental transition probabilities but rather trains deep neural networks by memorizing historical decision effects. This allows PSTER-TD3 to handle the uncertainty in state transitions and mitigate the impact of prediction errors, ensuring good control performance. The proposed method effectively improves the efficiency and quality of policy learning through three key technologies: tailoring double Q learning, policy delay updating and smooth target policy regularization under the actor-critic framework. Moreover, by integrating the priority experience replay strategy based on the sum tree structure into TD3, the agent can learn optimal energy management strategies in complex environments. The main contributions of this paper can be summarized as follows:

- A Markov decision process (MDP) with unknown state transition probabilities is established to describe the optimal energy management problem in an RM consisting of electric vehicles (EVs), photovoltaics (PVs), ESSs, heating, ventilation, and air conditioning (HVAC), and fixed loads. Unlike in [14] and [18]-[20], under the constructed MDP, the uncertainties of market electricity prices, outdoor temperature, PV generation, EV departure/arrival times, EV state of charge (SOC) upon arrival, and fixed load demands in the RM are comprehensively captured. Furthermore, the designed action space and reward function effectively consider the thermal comfort of households and the range anxiety in EV usage when the HEMS jointly schedules EVs, ESSs and HVAC systems.
- An RM online energy management method based on a novel DRL method called PSTER-TD3 is proposed. Unlike existing model-based methods [4], [5], and [7]-[10], the performance of the proposed method is not limited by the accuracy of physical model building or by specific environments. In addition, the proposed method offers higher optimization quality (including optimization accuracy and stability) and faster learning speed for energy management strategies than the DRL-based energy management methods used in [14], [19], and [20]. It also demonstrates greater robustness to uncertainties in the RM environment.
- The case studies based on real-world scenarios demonstrate that, compared to existing alternative state-of-the-art methods, even if there is some uncertainty in the environment, the proposed method can still effectively reduce the energy cost of the RM while meeting the household's thermal comfort requirements and reducing range anxiety in EV usage.

2. System description and problem formulation

2.1. System description

Fig. 1 illustrates the structure of the considered RM system, which includes PVs, ESSs, household electrical loads, and HEMSs. The RM is connected to the main power grid through the point of common coupling (PCC) and maintains a grid-connected operation mode. We categorize household electricity loads into controllable loads (including HVAC systems and EVs) and fixed loads [21]. The fixed load is the general term for all basic loads in an RM (such as refrigerators and TV sets) that have nonschedulable operation characteristics [2], [4]. In contrast, the operating time, power consumption, and interruption support of controllable loads can be flexibly adjusted.

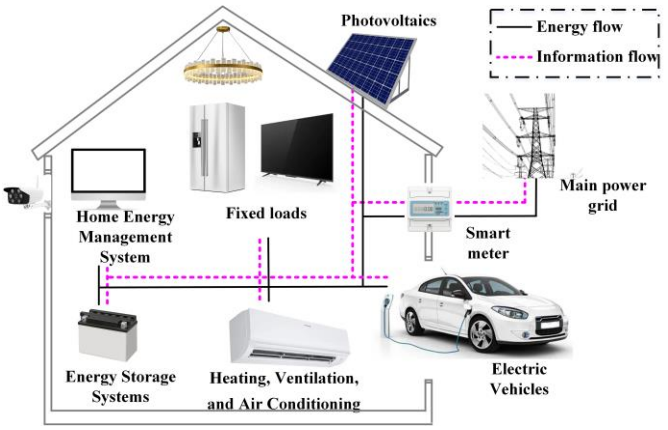

Fig. 1. The considered RM system structure.

Additionally, the power consumption of such loads can be continuously adjusted.

During the operation of an RM, the HEMS primarily serves two functions. First, it monitors, collects, and processes information about the various components and environmental conditions in the RM system. This information includes the PV generation power, market electricity prices, ambient temperature, and household electricity demand. Second, the operation of the adjustable equipment online is optimized through centralized control based on the collected local information. The optimization and scheduling objective is to achieve the optimal economic operation of the system while ensuring household comfort.

2.2. Mathematical model of RM components

2.2.1. EVs and ESSs

Due to the similar operational characteristics of EVs and ESSs, we discuss their mathematical models in the same section. Here, we assume that the arrival and departure of EVs occur at the beginning of the scheduling period. We define the moments when the EV arrives and departs from the household as t_a and t_d, respectively. The energy state of an EV can be expressed as follows [22]:

$$E_t^{\mathrm{EV}} = \begin{cases} E_{t-1}^{\mathrm{EV}} + \eta_{\mathrm{ch}}^{\mathrm{EV}} \cdot P_t^{\mathrm{EV}} \cdot \Delta t, \ P_t^{\mathrm{EV}} \geq 0, \\ E_{t-1}^{\mathrm{EV}} + 1/\eta_{\mathrm{dis}}^{\mathrm{EV}} \cdot P_t^{\mathrm{EV}} \cdot \Delta t, \ P_t^{\mathrm{EV}} < 0, \end{cases} \tag{1}$$

$$SoC_t^{\mathrm{EV}} = E_t^{\mathrm{EV}} / E_{\mathrm{CAP}}^{\mathrm{EV}}, SoC_{\min}^{\mathrm{EV}} \leq SoC_t^{\mathrm{EV}} \leq SoC_{\max}^{\mathrm{EV}}, \tag{2}$$

$$-P_{\max}^{\mathrm{EV}} \leq P_t^{\mathrm{EV}} \leq P_{\max}^{\mathrm{EV}}, \tag{3}$$

Equation (1) describes the energy conversion relationship during the charging and discharging process of an EV. Here, E_t^{EV} represents the energy stored in the EV battery during period t, and P_t^{EV} represents the charging or discharging power of the EV. If $P_t^{\mathrm{EV}} \geq 0$, the EV is in the charging state; otherwise, it is in the discharging state. Equation (2) imposes constraints on the state of charge of the EV to prevent overcharging or overdischarging of the EV. Here, $\eta_{\mathrm{ch}}^{\mathrm{EV}}$ and $\eta_{\mathrm{dis}}^{\mathrm{EV}}$ represent the charging and discharging efficiency coefficients of the EV, $SoC_{\max}^{\mathrm{EV}}$ and $SoC_{\min}^{\mathrm{EV}}$ represent the upper and lower limits of the EV state of charge, and $E_{\mathrm{CAP}}^{\mathrm{EV}}$ represents the battery capacity of the EV. Equation (3) restricts the charging/discharging power capacity of the EV, where $P_{\max}^{\mathrm{EV}}$ represents the maximum value of the charging/discharging power of the EV.

In addition, the state ξ_t^{EV} of the EV in period t can be defined as follows:

$$(\xi_{1,t}^{\mathrm{EV}}, \xi_{2,t}^{\mathrm{EV}}, \xi_{3,t}^{\mathrm{EV}}) = \begin{cases} (u_t^{\mathrm{EV}}, SoC_t^{\mathrm{EV}}, t), if \ t \in [t_a, t_d], \\ (u_t^{\mathrm{EV}}, 0, 0) \quad , \text{otherwise}, \end{cases} \tag{4}$$

where $\xi_{1,t}^{\mathrm{EV}}$ is the operational state of the EV, represented as a binary variable u_t^{EV}. If u_t^{EV}=1, the EV is connected to the charging facility in the RM; otherwise, the EV is unavailable. $\xi_{2,t}^{\mathrm{EV}}$ represents the charging/discharging progress of the EV, which refers to the state of charge SoC_t^{EV} of the EV battery in period t. $\xi_{3,t}^{\mathrm{EV}}$ is a specific attribute of the EV used to determine whether the EV is connected to the RM in period t.

The mathematical model of ESSs in RMs is similar to that of EVs, but unlike EVs, ESSs are always connected to the RM. Therefore, the main difference between the ESS mathematical model and the EV mathematical model is the charging/discharging period. The charging/discharging period of the ESS is the entire day.

2.2.2. HVAC

The load power of HVAC in period t, denoted as P_t^{AC}, can be continuously adjusted within the following range [23]:

$$0 \leq P_t^{\mathrm{AC}} \leq P_{\max}^{\mathrm{AC}}, \tag{5}$$

where $P_{\max}^{\mathrm{AC}}$ represents the maximum rated power of HVAC. Then, the state ξ_t^{AC} of HVAC in period t is defined as

$$(\xi_{1,t}^{\mathrm{AC}}, \xi_{2,t}^{\mathrm{AC}}, \xi_{3,t}^{\mathrm{AC}}) = (u_t^{\mathrm{AC}}, T_t^{\mathrm{AC}} - T_{\mathrm{set}}^{\mathrm{AC}}, T_{\mathrm{set}}^{\mathrm{AC}}), \tag{6}$$

where $\xi_{1,t}^{\mathrm{AC}}$ represents the operating state of HVAC, expressed as a binary variable u_t^{AC}; $\xi_{2,t}^{\mathrm{EV}}$ represents the difference $T_t^{\mathrm{AC}} - T_{\mathrm{set}}^{\mathrm{AC}}$ between the current indoor temperature T_t^{AC} during period t and the HVAC set temperature $T_{\mathrm{set}}^{\mathrm{AC}}$; and $\xi_{3,t}^{\mathrm{AC}}$ represents the temperature set value $T_{\mathrm{set}}^{\mathrm{AC}}$ of HVAC.

Furthermore, according to the equivalent thermal parameter model of HVAC, the relationship between the indoor temperature variation in residential buildings and P_t^{AC} can be represented as [18], [23]

$$T_{t+1}^{\mathrm{AC}} = \varsigma \cdot T_t^{\mathrm{AC}} + (1-\varsigma)(T_t^{\mathrm{out}} - \eta^{\mathrm{AC}} \cdot P_t^{\mathrm{AC}} \cdot \Delta t / \upsilon), \tag{7}$$

where T_t^{out} represents the outdoor temperature during period t, ς represents the heat dissipation coefficient of HVAC, η^{AC} represents the heat conversion efficiency, and υ represents the thermal conductivity of HVAC.

2.2.3. Power balance

To maintain the power balance of the RM, the total generating power of the RM must be equal to the total power consumed at any time during the whole scheduling cycle, that is,

$$P_t^{\mathrm{grid}} + P_t^{\mathrm{PV}} = P_t^{\mathrm{FL}} + P_t^{\mathrm{AC}} + P_t^{\mathrm{EV}} + P_t^{\mathrm{ESS}}, \tag{8}$$

where P_t^{grid} represents the power interaction between the RM and the main power grid, P_t^{PV} denotes the PV output power, and P_t^{FL} represents the power demand of the fixed load.

2.3. Optimization objective function

The optimization problem considered in this paper requires minimizing the expected total energy cost of the RM over T scheduling periods. This problem can be expressed as

$$\textbf{(P1)} \ \min \ \mathbb{E}\left[\sum_{t=1}^{T}\left(\mathrm{Cost}_t^{\mathrm{ESS}} + \mathrm{Cost}_t^{\mathrm{grid}} + \mathrm{Cost}_t^{\mathrm{EV}}\right)\right], \tag{9a}$$

$$\text{s.t.} \ \mathrm{Cost}_t^{\mathrm{ESS}} = c^{\mathrm{ESS}} \cdot \left|P_t^{\mathrm{ESS}}\right| \cdot \Delta t, \tag{9b}$$

$$\mathrm{Cost}_t^{\mathrm{grid}} = \left(\left((\lambda_t^+ - \lambda_t^-) \cdot \left|P_t^{\mathrm{grid}}\right|\right)/2 + \left((\lambda_t^+ + \lambda_t^-) \cdot P_t^{\mathrm{grid}}\right)/2\right) \cdot \Delta t, \tag{9c}$$

$$\mathrm{Cost}_t^{\mathrm{EV}} = c^{\mathrm{EV}} \cdot \left|\frac{m_k}{100}\right| \cdot \frac{P_t^{\mathrm{EV}}}{E_{\mathrm{CAP}}^{\mathrm{EV}}} \cdot \Delta t, \tag{9d}$$

$$(1)-(8). \tag{9e}$$

where $\mathrm{Cost}_t^{\mathrm{ESS}}$ represents the operational cost of the ESSs, c^{ESS} represents the cost coefficient of the ESS operation, and $\mathrm{Cost}_t^{\mathrm{grid}}$ represents the power interaction cost between the RM and the main power grid [20]. λ_t^+ and λ_t^- represent the prices of electricity purchased and sold by the RM to/from the main power grid, respectively. $\mathrm{Cost}_t^{\mathrm{EV}}$ represents the battery degradation cost of EV charging/discharging [22], c^{EV} represents the total cost of the battery, and m_k represents the linearly approximated slope of the battery life cycle function.

The models constructed above can provide a basis for the construction of representations of state and action variables and the design of reward functions in subsequent MDP. They also provide physical interpretations and mathematical relationship descriptions for the parameter variables involved in the MDP, thereby establishing a mapping relationship between the states,

action variables, and reward functions considered in this paper and specific physical states. This makes it easier to understand the state, action variables, and reward functions. Additionally, introducing models of relevant state and action variables helps in analyzing and understanding the behavior of the method proposed in this paper in specific scenarios, thereby enhancing the interpretability of the results.

3. Reformulating the problem as an MDP

To facilitate the use of the DRL-based method, we reformulate problem (P1) defined in Equation (9) as an MDP problem. From a mathematical perspective, an MDP is generally defined as a quintuple (S, A, P, R, γ). Here, S represents the set of all states that can be perceived by the agent in the environment, A represents the set of actions that the agent can perform, $P: S \times A \times S \to [0,1]$ is defined as the state transition probability distribution function, R represents the immediate reward obtained by the agent when performing the action in a specific state, and $\gamma \in [0,1)$ is defined as the discount factor. The MDP formula based on the mapping of the optimization problem under consideration is established as described below.

3.1. State space

The environmental information elements observed by the HEMS agent in period t include the output power P_t^{PV} of PVs, the purchase/sale prices λ_t^+ and λ_t^-, the fixed load demand power P_t^{FL}, the outdoor temperature T_t^{out}, the charge state SoC_t^{ESS} of the ESSs, the EV state tuple ξ_t^{EV} shown in formula (4) and the HVAC state tuple ξ_t^{AC} shown in formula (6). Therefore, the state space s_t can be defined as follows:

$$s_t = \left\{P_t^{\mathrm{PV}}, \lambda_t^+, \lambda_t^-, P_t^{\mathrm{FL}}, SoC_t^{\mathrm{ESS}}, T_t^{\mathrm{out}}, \xi_t^{\mathrm{EV}}, \xi_t^{\mathrm{AC}}\right\}. \tag{10}$$

Notably, certain elements in state s_t have inherent stochastic characteristics. These elements, which are not influenced by the actions of the HEMS agent, can be defined as exogenous state feature variables. In this study, the exogenous state feature variables are $\{P_t^{\mathrm{PV}}, \lambda_t^+, \lambda_t^-, P_t^{\mathrm{FL}}, T_t^{\mathrm{out}}, t_a, t_d\}$, where P_t^{PV} is related to the equipment status, location, and weather conditions and λ_t^+ and λ_t^- are influenced by supply-demand conditions. Additionally, T_t^{out} is affected by different weather conditions. The stochastic nature of these factors introduces significant uncertainty. Moreover, the moments t_a and t_d when the EV arrives and departs from the household and the SoC of the battery when the EV arrives are influenced by user behavior patterns and traffic conditions. These factors are also difficult to accurately predict and model, and their randomness further increases the dynamic uncertainty of the environment.

3.2. Action space

The scheduling decision variables of the HEMS include the charge/discharge power P_t^{EV} of the EVs, the output power P_t^{AC} of the HVAC, the charge/discharge power P_t^{ESS} of the ESSs, and the interaction power P_t^{grid} between the RM and the main power grid. Note that when P_t^{EV}, P_t^{AC} and P_t^{ESS} have been determined, P_t^{grid} can be directly obtained based on the power balance equation (as shown in Equation (8)). To simplify the action space, we do not consider P_t^{grid} in a_t. In summary, the a_t of the HEMS agent can be expressed as

$$a_t = \{P_t^{\mathrm{AC}}, P_t^{\mathrm{EV}}, P_t^{\mathrm{ESS}}\}. \tag{11}$$

3.3. State transition dynamics

The state transition function represents the probability distribution of the environment transitioning from state s_t to the next state s_{t+1} when the HEMS agent performs a given action a_t in period t. It can be expressed as follows:

$$P(s_t, s_{t+1}) = \Pr(s_{t+1} | s_t, a_t). \tag{12}$$

In this study, the state transition is influenced not only by the actions of the HEMS agent but also by the inherent stochasticity of the exogenous state features, as discussed in Section 3.1. Therefore, describing the state transition probability P using an accurate probability distribution model is difficult; that is, P in the established MDP formula is unknown, which reflects the uncertainty of the system. To address this issue, we adopt a DRL method. DRL can implicitly learn probability distribution characteristics based on the historical data of the random parameters of the system.

3.4. Reward

According to the energy cost considered in Equation (9), the base electricity reward r_t^{elec} for the HEMS agent during period t is set as follows:

$$r_t^{\mathrm{elec}} = \mathrm{Cost}_t^{\mathrm{ESS}} + \mathrm{Cost}_t^{\mathrm{grid}} + \mathrm{Cost}_t^{\mathrm{EV}}. \tag{13}$$

In addition, to ensure that the HVAC output can maintain an indoor temperature within a comfortable range, we define thermal comfort using the concept of an acceptable temperature range. Specifically, this range is defined as the maximum positive and negative deviation around the preferred temperature set by the user. If the indoor temperature deviates beyond this range from the set temperature, thermal discomfort is considered to occur. We represent this by calculating the absolute deviation between the actual temperature and the set temperature (beyond a deadband threshold $\Delta T_{\mathrm{thes}}^{\mathrm{AC}}$), and we add a penalty term r_t^{AC} for household thermal discomfort to the base reward, forming part of the negative reward. This is expressed as follows:

$$r_t^{\mathrm{AC}} = \omega_1 \max\left(0, \left|T_{\mathrm{set}}^{\mathrm{AC}} - T_t^{\mathrm{AC}}\right| - \Delta T_{\mathrm{thes}}^{\mathrm{AC}}\right), \tag{14}$$

where $\Delta T_{\mathrm{thes}}^{\mathrm{AC}}$ represents the threshold $\left|T_{\mathrm{set}}^{\mathrm{AC}} - T_t^{\mathrm{AC}}\right|$ of the difference between room temperature T_t^{AC} and the HVAC set temperature $T_{\mathrm{set}}^{\mathrm{AC}}$. ω_1 represents the thermal comfort weight factor, and it is measured in units of \$/°C, which allows the thermal discomfort term to be measured in the same units, \$, as the

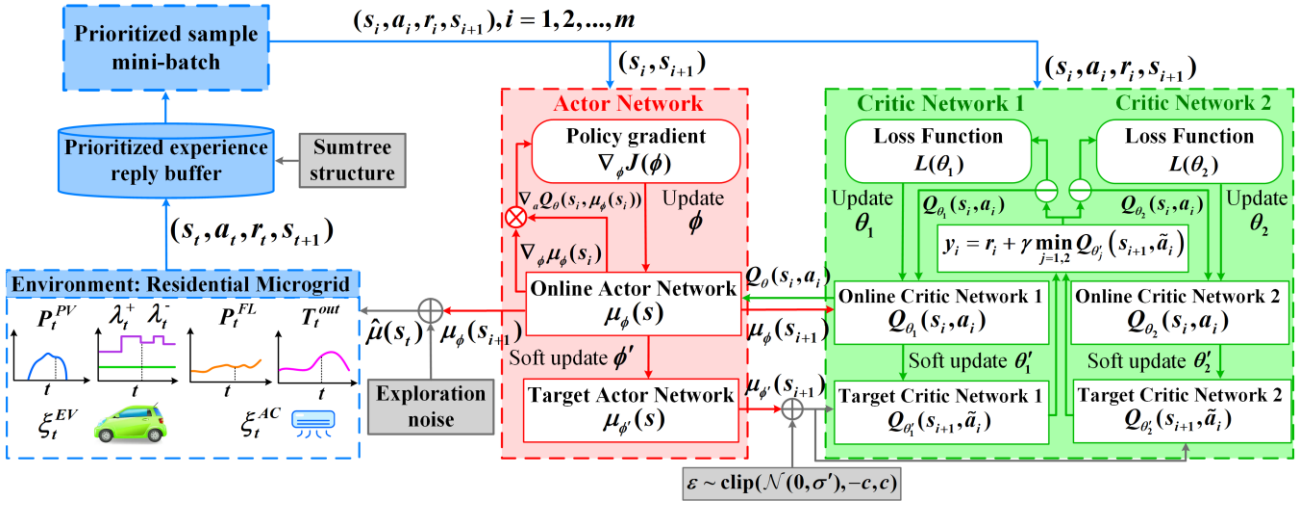

Fig. 2. PSTER-TD3-based RM energy management framework.

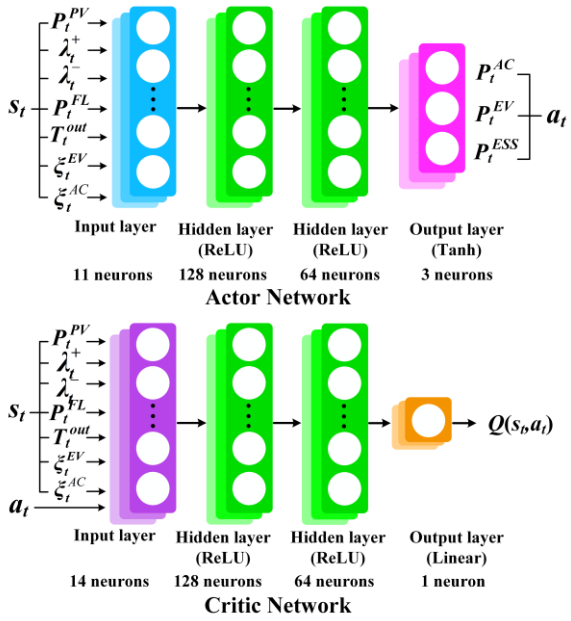

Fig. 3. Actor and critic network architecture of the proposed PSTER-TD3.

energy cost. The household's thermal discomfort is related to the maximum deviation threshold $\Delta T_{\text{thes}}^{\text{AC}}$ between the current indoor temperature T_t^{AC} and the set HVAC temperature $T_{\text{set}}^{\text{AC}}$. If $\left|T_{\text{set}}^{\text{AC}} - T_t^{\text{AC}}\right|$ is greater than $\Delta T_{\text{thes}}^{\text{AC}}$, the thermal discomfort gradually increases; otherwise, the thermal discomfort is 0. We choose the above definition because it provides a quantifiable, standardized explanation of thermal comfort and can determine a comfortable temperature range considering the household environment. Additionally, through equation (7), it is directly associated with the control of the HVAC system to reflect the direct temperature control functionality of the HVAC system.

To reflect the user's anxiety about having insufficient energy in the EV battery to reach the destination, we square the difference between the EV battery capacity (maximum range of the EV) and the remaining charge at the time of departure (actual remaining range at departure) to calculate the penalty for range anxiety. The less charge the vehicle has at the time of departure, indicating a shorter actual remaining range, the more anxious the owner is about not having enough energy to reach the destination. We then incorporate this penalty term, denoted as r_t^{EV}, into the reward function, forming part of the negative reward. It is expressed as [22], [24]

$$r_t^{\text{EV}} = \omega_2 (E_{\text{CAP}}^{\text{EV}} - E_t^{\text{EV}})^2,\ t = t_d, \tag{15}$$

where ω_2 represents the weight factor of the range anxiety term, and its unit is defined as \$/kWh^2, which allows the range anxiety term to be measured in the same units, \$, as the energy cost. We choose this definition because it is an effective and realistic way of increasing the penalty for severe electricity shortages by squaring the difference, preventing the case in which user travel demands cannot be met due to a severe electricity shortage. Additionally, through equations (1) and (2), it is directly associated with the control of EV charging and discharging, enabling controllable range anxiety.

Therefore, the reward function derived from the optimization objective is $r_t = -(r_t^{\text{elec}} + r_t^{\text{AC}} + r_t^{\text{EV}})$. By introducing weights for penalty terms, we ensure that both penalty terms and cost terms have the same units of measurement and can balance and adjust the relative influences of the two penalty terms.

Through the above clear definitions, user perceptions can be quantified and incorporated into the optimization framework of PSTER-TD3. Specifically, we first incorporate the operation of the HVAC system into the optimization framework and design a penalty term for household thermal discomfort to penalize behaviors deviating from the human comfort zone, thereby meeting thermal comfort requirements. Next, range anxiety is used to describe the concerns of EV owners regarding whether there is sufficient power to meet travel needs. We design a penalty term for range anxiety to penalize the anxiety regarding insufficient energy in the EV battery to reach the destination, managing the charging/discharging of the EV battery to minimize the risk of not having enough power to complete the intended journey. Finally, the thermal discomfort penalty term and range anxiety penalty term are explicitly incorporated into the MDP formulation and reward design, superimposed with the household energy cost term, and mapped into the reward function of the DRL method. During the interaction between the agent and the environment, training is conducted to learn a policy that maximizes the cumulative rewards. This policy ensures not only a reduction in energy costs but also a decrease in thermal discomfort and range anxiety, effectively meeting the requirements for household comfort.

3.5. Action-value function

The utility of the HEMS agent in executing scheduling action a following policy π under a given state s is evaluated by the cumulative discount reward $Q_\pi(s,a) = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \gamma^t r_t | s_t = s, a_t = a]$ within T time steps, where $Q_\pi(s,a)$ is the action value function. The HEMS agent aims to find the optimal scheduling policy π^* among all feasible policies to maximize the reward or minimize the energy cost. This can be represented as $\pi^* = \text{argmax}_{a \in A} Q_\pi(s,a)$.

4. Proposed online energy management method

4.1. Preliminaries

As a representative algorithm based on policy gradient DRL, DDPG is known for its ability to effectively handle control tasks with high-dimensional continuous action and state spaces. DDPG is based on the deterministic policy gradient (DPG) algorithm combined with the actor-critic framework and DQN extension [17]. Although DDPG has many advantages, one drawback is that the critic network in DDPG is always updated in the gradient direction of increasing Q-values, which can lead to overestimation of the Q-values. Furthermore, high variance in the calculation of the target value makes it difficult to stably obtain the optimal action strategy. Although the above shortcomings seriously affect the policy learning quality of DDPG, they also provide motivation for research on DRL methods that perform better than current DRL methods. These better-performing methods can contribute to the learning of better action strategies for HEMS agents in complex RM scheduling environments.

4.2. Proposed PSTER-TD3 method

The PSTER-TD3 method framework we developed to solve RM optimization operation problems is shown in Fig. 2. First, the proposed method calculates the target value by truncating the double Q-learning process, addressing the overestimation problem for critic networks in DDPG. In the proposed method, two identical and independent critic networks are used to estimate the Q value. Their target value calculations can be formulated as follows, where j represents the jth critical network:

$$y_t = r_t + \gamma \min_{j=1,2} Q_{\theta_j'}\left(s_{t+1}, \mu_{\phi'}(s_{t+1})\right). \tag{16}$$

Second, the proposed method introduces a regularization term in the output of the target policy to reduce the variance of the target value calculation in DDPG. Specifically, random noise ε is added to the target action used to calculate the target value, smoothing out the variation in the Q function along the action:

$$\tilde{a}_t \leftarrow \mu_{\phi'}(s_{t+1}) + \varepsilon,\ \varepsilon \sim \text{clip}(\mathcal{N}(0, \sigma'), -c, c), \tag{17}$$

where clip(·) represents the truncation function and truncating random noise limits the amplitude of changes in the action after noise processing. Therefore, Equation (18) can be rewritten as

Algorithm 1 Training Process of PSTER-TD3

Input: RM's environmental status information on training day s_t ; action space information a_t

Output: The parameters $Q_{\theta_j'}(s_t, a_t)$ and corresponding weight parameters θ_1' and θ_2' of the two trained critic target networks; The parameters $\mu_{\phi'}(s_t)$ and corresponding weight parameters ϕ' of the trained actor target network

Random initialization: Two critic network parameters $Q_{\theta_1}(s,a)$ and $Q_{\theta_2}(s,a)$, as well as their parameters θ_1 and θ_2; Actor network parameter μ_ϕ and its parameter ϕ

Initialization: $\theta_j' \leftarrow \theta_j$, $\phi' \leftarrow \phi$, experience buffer pool $\mathcal{D}$ with capacity N_p; The structure of the experience pool is a sum tree with priority parameters β_1 and β_2

for episode = 1: M^{train} **do**

  Initialize exploration noise $\mathcal{N}_t$

  Obtain the initial observation status s_1 from RM

  for t = 1: $T_{\max}$ **do**

    Select actions based on the current policy and exploration noise based on $\hat{\mu}(s_t) = \mu_\phi(s_t) + \mathcal{N}_t$

    Execute action a_t in the RM environment to receive reward r_t and the next new state s_{t+1}

    Store experience samples (s_t, a_t, r_t, s_{t+1}) in the sum tree and set priority $p_t = \max_{i<t} p_i$

    for i = 1: m **do**

      Use the probability $p(i)$ obtained from equation (20) to sample experience i from the sum tree

      Calculate the priority importance sampling weight ω_i through equation (21)

      Calculate the TD error δ_i of experience i through equation (19)

      Update the priority value corresponding to leaf nodes in the sum tree using $|\delta_i|$ based on $p_i = 1/\text{rank}(i)$

    end for

    Use the target network to obtain actions $\tilde{a}_t \leftarrow \mu_{\phi'}(s_{t+1}) + \varepsilon$, $\varepsilon \sim \text{clip}(\mathcal{N}(0, \sigma'), -c, c)$

    Calculate the target Q value of the online critic network according to equation (18)

    Update the online critic network parameters according to equations (22) and (24)

    if t **mod** d **then**

      Update actor network parameters according to equations (23) and (25)

      Soft update the target network parameters according to equations (26) and (27)

    end if

  end for

end for

$$y_t = r_t + \gamma \min_{j=1,2} Q_{\theta_j'}(s_{t+1}, \tilde{a}_t). \tag{18}$$

Third, the proposed method weakens the negative impact of the target value calculation variance on actor network learning in DDPG by reducing the update frequency of the actor network parameter ϕ and the target network parameters ϕ' and θ'. Specifically, in the proposed method, the critic network completes $d (d \geq 2)$ updates before updating the actor network once. This update method can better stabilize the training of the actor network. The

Algorithm 2 PSTER-TD3-based RM Online Operation Strategy

Input: RM environment status s_t on test day, test period E^{test}

Output: The scheduling action decision a_t of the HEMS agent in each period

Read the online actor network parameters ϕ^* trained according to Algorithm 1

for episode = 1: E^{test} **do**

  Obtain the initial status s_1 of the test day

  for t = 1: $T_{\max}$ **do**

    Set the action of the HEMS agent to $a_t = \mu_{\phi^*}(s_t)$

    Perform action a_t and interact with the RM environment to obtain reward value r_t, and observe the RM environment transfer to a new state s_{t+1}

  end for

end for

actor and critic network architecture of the proposed PSTER-TD3 method is shown in Fig. 3. The actor network of the proposed method takes the environmental state observation s_t of the RM as input and outputs continuous scheduling action a_t based on the Q value estimated by the critic network. The critic network takes state observation s_t and action a_t as inputs and outputs an estimated Q value.

Furthermore, to enable the model to fully explore the important scheduling experience obtained from the interactions between the agent and the environment during training, we introduce priority experience replay (PER) in the proposed method. PER assigns a priority weight value to each empirical sample based on its importance to model training; the larger the priority weight value is, the higher the sampling probability. In the proposed PER mechanism, the temporal difference error (TD error) δ_i of the empirical sample i is used to measure the importance of the sample for training the proposed method, where δ_t is defined as the estimated Q value error between the online critic network and the target critic network:

$$\delta_i = r_i + \gamma Q_{\theta_j'}(s_{i+1}, \tilde{a}_i) - Q_{\theta_j}(s_i, a_i). \tag{19}$$

The larger the absolute value of the TD error δ_i of empirical sample i is, the greater its contribution to the gradient update of the neural network. This also means that the more assistance it provides for learning the proposed method policy, the greater its importance. We define the probability of sampling experience sample i from experience buffer pool $\mathcal{D}$ as

$$p(i) = p_i^{\beta_1} / \sum_k p_k^{\beta_1}, \tag{20}$$

where $\beta_1 \in [0,1]$ represents the sampling weight coefficient of priority, which is used to control the degree of priority usage. If $\beta_1 = 0$, then the sampling rule at this time completely follows a uniform distribution. In Equation (22), we calculate the priority $p_i = 1/\text{rank}(i)$ of empirical sample i based on the sorting priority strategy. Here, $\text{rank}(i)$ represents the sequence number of the replay unit sorted from highest to lowest based on the absolute value $|\delta_i|$ of the TD error in the empirical sample. In addition, we introduce importance sampling (IS) to correct the model updating deviation caused by the change in the sample experience distribution. The IS weight ω_i of sample i is

$$\omega_i = (N_p \cdot p(i))^{-\beta_2} / \max_k \omega_k, \tag{21}$$

where $\beta_2 \in [0,1]$ represents the IS weight adjustment coefficient. To ensure the stability of the model during training, the IS weights are normalized using $\max_k \omega_k$. Then, the IS weight ω_i is included in the calculation of the critic network loss function. Therefore, the calculation formula of the critic network loss function corrected by the proposed method is rewritten as

$$L(\theta_j) = \frac{1}{m} \sum_{i=1}^{m} \omega_i (y_i - Q_{\theta_j}(s_i, a_i))^2. \tag{22}$$

Similarly, the actor network update equation of the proposed method is rewritten as

$$\nabla_{\phi} J(\phi) = \frac{1}{m} \sum_{i=1}^{m} \nabla_a Q_\theta(s_i, a)\Big|_{a=\mu_\phi(s_i)} \nabla_\phi \mu_\phi(s_i). \tag{23}$$

The weight parameter update formulas for the proposed online critic and actor networks can be expressed as

$$\theta_j \leftarrow \theta_j + \alpha^{\theta_j} \nabla_{\theta_j} L(\theta_j), \tag{24}$$

$$\phi \leftarrow \phi + \alpha^{\phi} \nabla_{\phi} J(\phi), \tag{25}$$

where α^{θ_j} and α^{ϕ} represent the corresponding learning rates of the online critic and actor networks, respectively. Notably, the "soft update" strategy is used to update the weight parameters of the proposed method's target network:

$$\theta'_j \leftarrow \tau\theta_j + (1-\tau)\theta'_j, \tag{26}$$

$$\phi' \leftarrow \tau\phi + (1-\tau)\phi', \tag{27}$$

where $\tau(\tau << 1)$ represents the soft update coefficient and θ'_j and ϕ' represent the corresponding weight parameters of the critic and actor target networks, respectively.

To further improve the sampling efficiency and model convergence speed, the proposed method introduces PER based on a sum-tree structure, known as PSTER. The empirical sample data structure based on the sum tree is shown in Fig. 4. Here, the prioritized actual data of the experience sample are stored only in the lowest leaf node. Based on the binary tree structure of the sum tree, the sum of the priority values of the two leaf nodes is stored in the corresponding parent nodes of each pair of child nodes. Then, this summation method continues until it converges to the root node. Therefore, the root node priority value, represented as s_p, is the sum of the priority values stored by all leaf nodes.

When it is necessary to use empirical samples, the sampling probability interval [0, s_p] with a total interval length of s_p is first divided evenly into subintervals with the same number of small batch sampling times m, which are then sorted by priority interval from smallest to largest. Then, one number s_{rand} is randomly selected from each of the m subintervals. Starting from the root node, the rules are followed from left to right and from top to bottom to traverse the child nodes of s_{rand} point by point. The above process is repeated until the leaf node is retrieved, and the corresponding priority values retrieved by each random number are used to determine the experience sample.

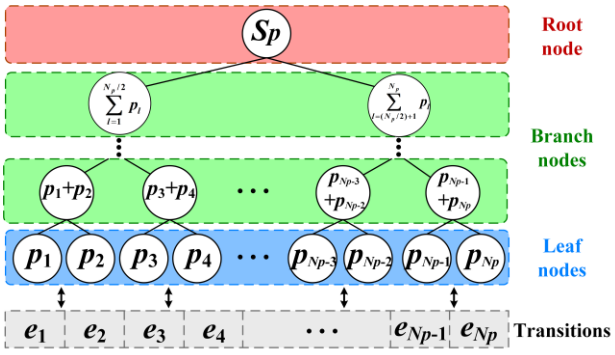

Fig. 4. Experience sample priority storage structure based on sum-tree.

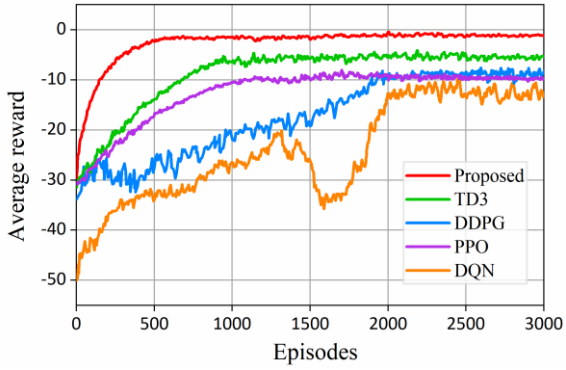

Fig. 5. The average reward for each DRL method training process.

Table 1
Operating parameters of RM key components.

Type	Symbol and Value
EV	$P_{\max}^{\text{EV}}$ =3kW, $E_{\text{CAP}}^{\text{EV}}$ =16kWh, $\eta_{\text{ch}}^{\text{EV}} = \eta_{\text{dis}}^{\text{EV}}$ =0.98
	$SoC_{\max}^{\text{EV}}$ =1.0, $SoC_{\min}^{\text{EV}}$ =0.1
	t_a~clip($\mathcal{N}(18,1^2)$,15, 21) , t_d~cilp($\mathcal{N}(8,1^2)$,6,11)
HVAC	$T_{\text{set}}^{\text{AC}}$ =24℃, η^{AC} =2.0, ς =0.7, υ =7.27×10^{-3}kW/℃
	$P_{\max}^{\text{AC}}$ =2kW, $\Delta T_{\text{thes}}^{\text{AC}}$ =2℃
ESS	$\eta_{\text{dis}}^{\text{ESS}} = \eta_{\text{ch}}^{\text{ESS}}$ =0.95, $P_{\max}^{\text{ESS}}$ =2kWh
	$SoC_{\max}^{\text{ESS}}$ =1.0, $SoC_{\min}^{\text{ESS}}$ =0.1, $E_{\text{CAP}}^{\text{ESS}}$ =6kWh

Table 2
PSTER-TD3 hyperparameter setting information.

Parameters	Value
critic Network Learning rate α^{θ_1} and α^{θ_2}	1×10^{-3}
actor Network Learning rate α^{ϕ}	1×10^{-4}
Discount factor γ	0.99
Soft update coefficient τ	5×10^{-3}
Policy update interval d	2
Minimum batch sampling size m	128
Experience buffer pool capacity N_p	1×10^5
Noise clipping coefficient c	0.5
Priority sampling weight β_1	0.6
Weight adjustment coefficient of Importance sampling β_2	0.4

To address uncertainty, the PSTER-TD3 method proposed in this paper leverages real-world historical data and adaptive learning strategies through environmental interactions to gain experience and approximate the optimal decision-making strategy. It implicitly learns the probability distribution characteristics of the interactions based on the historical data of the system's random parameters and then handles the issue of unknown state transition probabilities. Additionally, the proposed method introduces an experience replay strategy based on a prioritized sum tree, which automatically distinguishes the supervisory values of different experience samples when facing complex dynamic environments. It automatically selects the most effective training data from a large amount of encountered data and dynamically adjusts the distribution of sample weights. This enables the model training to automatically focus on high-value data in regions of the state space with significant dynamic changes, enhancing the data utilization efficiency, accelerating policy convergence, and addressing environmental uncertainty in a targeted manner. Moreover, this structure reinforces the agent's ability to discover and adapt to the inherent dynamic patterns of the environment by reusing critical experiences. This helps mitigate the impact of state variable dynamics and randomness, thereby strengthening the robustness of the policy.

4.3. Application process

The process of applying the proposed method includes two stages, offline training and online deployment, the input and output variables of which are shown in Table 3. First, the proposed PSTER-TD3 method is used for offline training of the DNN, as shown in Algorithm 1. After training, the optimal parameters of the PSTER-TD3 model can be obtained. Then, the trained PSTER-TD3 model is deployed on the HEMS agent to provide online optimization operation strategies for the RM.

The online optimization operation strategy of the RM based on PSTER-TD3 is shown in Algorithm 2. For each period of the selected test day, the

Table 3
The input and output variables for offline training and online deployment.

Stage	Category	Variable	Description
Offline training	Input	$s_t=\left\{\begin{matrix}P_t^{\mathrm{PV}},\lambda_t^{+},\lambda_t^{-},P_t^{\mathrm{FL}},\\ SoC_t^{\mathrm{ESS}},T_t^{\mathrm{out}},\xi_t^{\mathrm{EV}},\xi_t^{\mathrm{AC}}\end{matrix}\right\}.$	The RM's environmental status information on the training day
		$a_t=\{P_t^{\mathrm{AC}},P_t^{\mathrm{EV}},P_t^{\mathrm{ESS}}\}.$	The RM's action space information on the training day
	Output	$Q_{\theta_j'}(s_t,a_t)$	The parameters of the trained critic target network
		θ_1' 、 θ_2'	The weight parameters of the trained critic target network
		$\mu_{\phi'}(s_t)$	The parameters of the trained actor target network
		ϕ'	The weight parameters of the trained actor target network
Online deployment	Input	$s_t=\left\{\begin{matrix}P_t^{\mathrm{PV}},\lambda_t^{+},\lambda_t^{-},P_t^{\mathrm{FL}},\\ SoC_t^{\mathrm{ESS}},T_t^{\mathrm{out}},\xi_t^{\mathrm{EV}},\xi_t^{\mathrm{AC}}\end{matrix}\right\}.$	The RM environment status on the test day
		E^{test}	The test period
	Output	$a_t=\{P_t^{\mathrm{AC}},P_t^{\mathrm{EV}},P_t^{\mathrm{ESS}}\}.$	The scheduling action decision of the HEMS agent in each period

Table 4
Comparison of computational performance between different methods

Methods	MILP	MPC	DQN	PPO	DDPG	TD3	Proposed
Training (h)	-	-	2.85	3.47	3.72	3.28	2.94
Operation (s)	1.035	0.497	0.406	0.428	0.459	0.393	0.342

Table 5
Test performance indicators for different methods

Methods	*MACE*(%)	*MTD*(℃)	*RANGE*	*STD*
MILP	33.46	0.00	0.21	0.076
MPC	37.50	4.27	0.26	0.095
DQN	29.17	10.82	0.39	0.137
PPO	44.79	9.71	0.20	0.051
DDPG	40.63	8.65	0.31	0.113
TD3	26.04	5.36	0.27	0.086
Proposed	20.83	0.00	0.15	0.048

HEMS agent trained with Algorithm 1 for the optimal online actor network parameters is read. Based on the observed initial state s_1 of the RM environment at this moment, the control action a_t is determined based on the optimal policy $\mu_{\phi'}(s_t)$ learned from PSTER-TD3. Then, reward r_t is observed, and the environmental state transitions to s_{t+1}. Finally, the above process is repeated until all the scheduling tasks for the testing period are completed online.

To address variations in the size of the elements or parameters of the problem, the method proposed in this paper can adapt to different scales of state and action spaces through sampling buffers and neural network function approximation. It can handle variations in the size of elements or parameters of the problem to a certain extent without needing to be retrained from scratch, and it has strong robustness. These variations mainly involve variations in the number of states, action dimensions, and environmental parameters. Specifically, 1) when the number of states varies, the proposed PSTER-TD3 method uses an experience replay buffer based on importance sampling to store historical state-transition data and samples these data in batches according to priority from the buffer to train the neural network. The sampling scope is not dependent on the specific number of states. Therefore, the addition or removal of states has little impact on the method. 2) When the dimension of actions varies, in the PSTER-TD3 method, the output layer dimension of the policy network (actor) corresponds to the dimension of actions. It is sufficient to adjust the output layer while keeping the other parameters unchanged, and retraining is not necessary. 3) When the environmental parameters vary, the proposed PSTER-TD3 method utilizes a priority experience replay mechanism based on the sum tree, which focuses on learning transitions with large TD errors and pays attention to these difficult-to-learn but crucial samples to enhance generalizability. Combined with exploratory random action noise and slower policy updates, the proposed method demonstrates reliable performance even under changing environmental parameters.

4.4. Global Optimality Analysis of the PSTER-TD3 Algorithm

As a type of memoryless stochastic process, the convergence theory of Markov chains provides theoretical tools for analyzing the convergence of optimization algorithms. The proposed PSTER-TD3 algorithm generates a sequence of solutions by repeating actions such as state sampling, action selection, reward evaluation, and policy/value function updates. The behavior in each round depends only on the current state of the solution and is independent of historical states, thus satisfying the Markov property. Therefore, the sequence of solutions generated by the algorithm can be regarded as a Markov chain. Based on the theory of Markov chains, if the algorithm ensures that the sequence of solutions has a positive probability of visiting all states within a finite time and that each new solution is always better than or equal to the previous one, then the algorithm will converge to the global optimal solution with probability 1. Therefore, this paper uses Markov chain theory to demonstrate that the proposed algorithm can converge to the global optimum.

If the proposed algorithm converges to the global optimum with probability 1, it needs to satisfy the following two conditions [27]:

(1) Any two points $(\pi|q)_1,(\pi|q)_2$ in the feasible solution space can be reached through state transitions;

(2) The sequence of experienced solutions $M_1,M_2,\cdots,M_n$ is monotonic.

To ensure that (1) holds, we need to prove that (i) the Markov chain of experienced solution sequences is finite; (ii) the Markov chain of experienced solution sequences is homogeneous; and (iii) the Markov chain of experienced solution sequences is ergodic. The details are as follows:

For the *n*th round of the experience solution $M_n=\left\{(\prod|\mathrm{Q})_1,(\prod|\mathrm{Q})_2,\ldots(\prod|\mathrm{Q})_D\right\}$, D represents the dimensionality of the experience samples sampled from the experience replay pool, and $(\prod|\mathrm{Q})_d$ is the *d*th experience solution.

According to $D<\infty$, M_n is finite; thus, condition (i) holds. In the proposed algorithm, state sampling, action selection, reward evaluation, and policy/value function updates are performed independently in stochastic processes. Each update of the policy/value function is chosen optimally based on the expected return through gradient ascent. The update of the policy/value function in round $n+1$ depends only on the cumulative reward evaluation in round n, which is independent of the transition probabilities between other states and the number of iterations. Thus, condition (ii) holds.

To prove that condition (iii) holds, the Markov chain needs to satisfy irreducibility and aperiodicity and be ergodic. We refer to the following

definitions in the proof [28]:

a) The transition probability matrix between different solutions in the sequence is $P_{ij} = P\{M_{n+1} = j \mid M_n = i, n \geq 1\}$; if for any i, j, there exists an $n \geq 1$ such that $P_{ij}^k \geq 0$, then the Markov chain is irreducible;

b) If the greatest common divisor of the nonempty set $U = \{n \mid n \geq 1, P_{ij}^n > 0, \forall i, j\}$ is 1, then the Markov chain is aperiodic;

c) If the recurrent state i satisfies $U_i = \sum_n^{\infty} nP_{ij}^n < +\infty$, then i is called positively recurrent. If i is aperiodic, then the Markov chain is ergodic.

The specific proof is as follows: Since the transition probability matrix P_{ij} depends only on i, j and all elements in M_n are positive, for any i, j, there exists an $n \geq 1$ such that $P_{ij}^k \geq 0$. According to definition a), the Markov chain satisfies irreducibility. Based on irreducibility, for $U = \{n \mid n \geq 1, P_{ij}^n > 0, \forall i, j\}$, there exists $n = 1$ such that the greatest common divisor of set U is 1. According to definition b), the Markov chain satisfies aperiodicity. The behaviors of state sampling, action selection, reward evaluation, and policy/value function updates all lead to state transitions, which can be represented by transition matrices S'、A'、R' and O', and all of them are between 0 and 1, as the transition probabilities are defined as $P': S' \times A' \times R' \times O' \to [0,1]$. Therefore, the transition probabilities satisfy $0 < P_{ij} < 1$. Let $\varepsilon = \max\{P_{ij} : \forall i, j \in H\}$; then, by the Cauchy-Riemann equation, there exists $n \geq 1$ such that for any state F, it satisfies $U_i = \sum_n^{\infty} nP_{ij}^n \leq \sum_n^{\infty} n\left(\max\left(P_{ij}\right)\right)^n < \infty$. Given that i is aperiodic, according to definition c), the Markov chain is ergodic. In conclusion, condition (iii) holds.

The state sampling, action selection, reward evaluation, and policy/value function updates in the proposed algorithm all adhere to a policy of selecting and retaining better solutions. Furthermore, the sequence of policy/value function solutions generated by the algorithm can be regarded as a finite homogeneous Markov chain, and each round only transitions and updates the policy/value function when a better Q-value estimate is found. Therefore, in each iteration round of TD3, the newly generated policy/value function must be superior to the old estimate. Thus, it can be concluded that the sequence of policy/value function solutions generated by the TD3 algorithm converges monotonically, satisfying condition (2).

Overall, the proposed algorithm converges to the global optimal solution of the problem with probability 1.

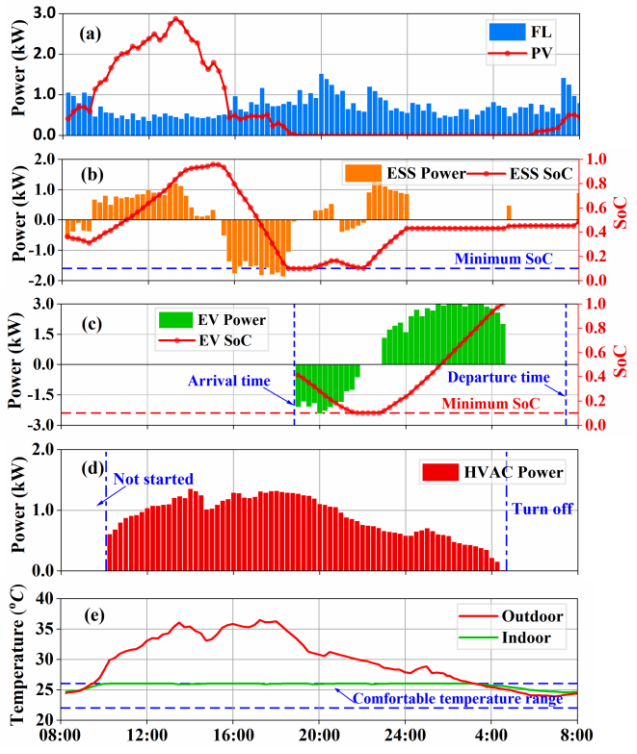

Fig. 6. The scheduling results of the proposed method on a certain testing day (a) Required power and photovoltaic power for fixed loads. (b) ESS charging and discharging power and SoC variation. (c) EV charging and discharging power and SoC variation. (d) HVAC output power. (e) Comparison of indoor and outdoor temperatures.

5. Case study

5.1. Experimental settings

- Datasets and Parameter Setup: We evaluate the performance of the proposed method on actual data. The actual photovoltaic output, load demand, and outdoor temperature data are provided by the famous Pecan Street database [25]. The data in this database from March 2, 2017 to June 29, 2017 are used as the training set, and the data from July 2 to August 30 are used as the test set. The detailed parameter settings for EVs, HVAC systems, and ESSs are shown in Table 1. In this study, we assume that EVs have relatively fixed arrival and departure times, and their arrival and departure times are modeled as truncated normal distributions [22]. Moreover, we assume that the SoC of a battery is sampled from the truncated normal distribution $\text{clip}(\mathcal{N}(0.5, 0.1^2), 0.2, 0.8)$ when an EV arrives [22], [24]. To ensure the compatibility of the input data in the simulation environment, parameters for the driving distances and charging patterns of EVs are sourced from traffic data in the same region as Pecan Street. This ensures that the data for EVs in the simulation environment are matched with those of other energy-consuming devices. The Adam optimizer [26] is used to train the actor and critic network parameters of PSTER-TD3. The settings of the other hyperparameters for the proposed method are displayed in Table 2.
- Benchmark Method: To verify the performance of the proposed method, we compare it with the following benchmark methods. The traditional online energy management method MPC, in which the long short-term memory neural network (LSTM) is used for temporal prediction of uncertain parameters in the future rolling period, is used as one of the benchmarks. We also compared the proposed method with benchmark methods based on the DQN, PPO, DDPG, and TD3 algorithms.

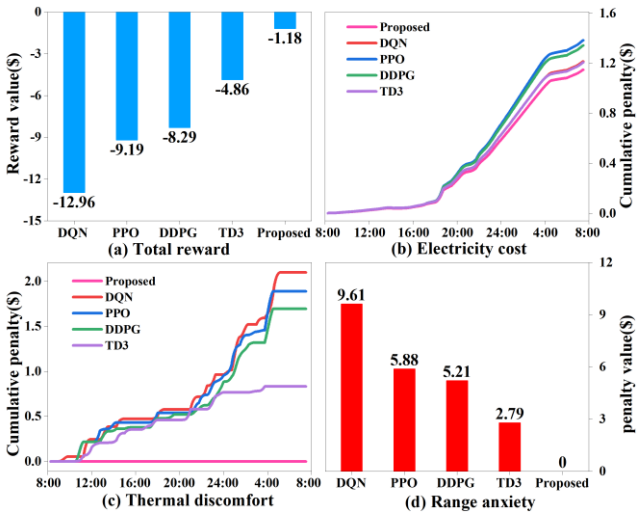

Fig. 7. The scheduling results of each DRL method on a certain testing day

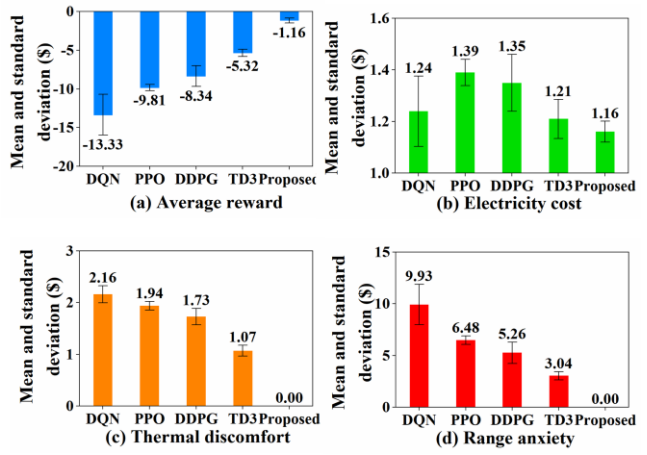

Fig. 8. Comparison of rewards for testing different DRL methods.

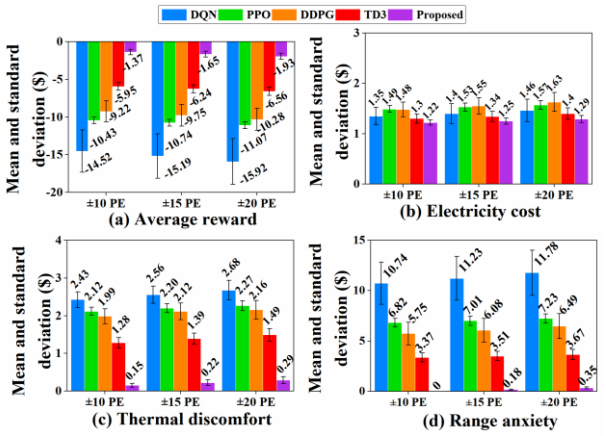

Fig. 9. Comparison of rewards for testing various DRL methods under different prediction errors.

- Performance Metrics: To comprehensively evaluate the optimization performance of various online energy management methods, we adopt the following performance evaluation indicators. The mean absolute optimization error (MACE) and mean temperature deviation (MTD) are used to evaluate the optimization accuracy of each method. The range and standard deviation (STD) of the optimization errors are used to evaluate the stability of the scheduling results. The definitions of each metric are as follows:

$$MACE=\frac{1}{E^{\text{test}}}\cdot\sum_{i=1}^{E^{\text{test}}}(\sigma_i^{\text{bias}}/c_i^{\text{PIO}})\times 100\% \tag{28}$$

$$MTD=\sum_{i=1}^{E^{\text{test}}}T_i^{\text{dev}}\Big/E^{\text{test}} \tag{29}$$

$$RANGE=\max\{\sigma_1^{\text{bias}},...,\sigma_{E^{\text{test}}}^{\text{bias}}\}-\min\{\sigma_1^{\text{bias}},...,\sigma_{E^{\text{test}}}^{\text{bias}}\} \tag{30}$$

$$STD=\sqrt{\sum_{i=1}^{E^{\text{test}}}\left(\sigma_i^{\text{bias}}-\sigma_m^{\text{bias}}\right)^2\Big/E^{\text{test}}} \tag{31}$$

MACE represents the mean absolute percentage error between the energy cost c_i^{daily} of test day i and the solution c_i^{PIO} of the Perfect Information Optimum (PIO) strategy [18], where σ_i^{bias} represents the optimization bias, that is, $\left|c_i^{\text{daily}}-c_i^{\text{PIO}}\right|$. MTD represents the average deviation between the indoor temperature on the test day and the comfortable temperature range (22°C to 26°C), where T_i^{dev} represents the sum of the deviations between the indoor temperature on test day i and the comfortable temperature range. RANGE represents the range of optimization errors generated during testing. STD represents the standard deviation of optimization errors during testing, where σ_m^{bias} represents the mean of all optimization errors.

5.2. Numerical results

5.2.1. Training performance

Fig. 5 compares the average reward obtained during training by the proposed method and other DRL methods. Fig. 5 illustrates that the performance of the proposed PSTER-TD3 is superior to the performances of the other DRL methods. Specifically, PSTER-TD3 converges at approximately the 600th round, while TD3/DDPG/PPO/DQN converge at approximately the 1000th/2000th/1100th/2200nd rounds. PSTER-TD3 explores the optimal action strategy approximately 1.7/3.3/1.8/3.7 times faster than TD3/DDPG/PPO/DQN. The average reward (-1.12) when PSTER-TD3 converges is greater than the average rewards (-5.29/-8.30/-9.76/-13.31) when TD3/DDPG/PPO/DQN converge. These results also indicate that the proposed method has lower learning costs than the compared methods. Compared to TD3, the superior performance of PSTER-TD3 is due to its ability to replay action experiences associated with high TD errors more frequently during training with less time complexity. This greatly aids in improving the agent's action policy and enhancing the algorithm's convergence speed. Compared to DDPG/PPO/DQN, the adoption of a series of key techniques in PSTER-TD3 (see Section 4.2) effectively alleviates the issues of overestimation and high variance of Q-values. Specifically, compared to DDPG/PPO/DQN, (i) the higher quality of the convergent solutions in PSTER-TD3 is attributed to the double Q-network architecture, which reduces the overestimation of Q-values and enhances the robustness of Q-value estimation, leading to more accurate and robust state value evaluation. (ii) The increased stability in the convergence process of PSTER-TD3 is due to the introduction of a regularization term in the output of the target policy, which smooths the target policy actions. This helps prevent drastic policy oscillations and divergence during the learning process, thereby improving the stability of policy updates. Additionally, by reducing the update frequency of the actor network and target network parameters, PSTER-TD3 mitigates the instability caused by constant changes in the targets, further enhancing the stability of learning and control. Compared with the DQN, the proposed method can more fully utilize the information in the entire multidimensional continuous action space, which significantly improves the learning quality of the policy. Overall, the proposed PSTER-TD3 can adaptively learn stable optimized scheduling policies through good training, and its policy learning speed and quality are higher than those of other DRL methods.

5.2.2. Scheduling results on a test day

To verify the effectiveness of the proposed method, we present the 24-hour scheduling results for a test day, as shown in Fig. 6. Figs. 6 (a) and (b) demonstrate that ESSs try to absorb the remaining electricity generated by PV during the peak period of PV output to prevent electricity from being sold. The ESSs discharge during peak electricity consumption periods (16:00-20:00) with high electricity purchase prices, which helps reduce electricity purchase costs and absorb PV output. ESSs do not immediately charge a large amount when the SoC is at a low level; instead, the ESSs shift the charging period to a low electricity price period (22:00-6:00). Similarly, the EVs discharge during the high electricity price period from 19:00 to 22:00 (Fig. 6 (c)), and their charging process shifts to the low electricity price period as much as possible. In addition, Fig. 6 (c) demonstrates that introducing battery degradation costs enables the HEMS to purposely control EV charging and discharging actions and SoC fluctuations, which helps to extend the service life of the battery. Moreover, when an EV is about to leave, the SoC of its battery approaches 1, indicating that the EV has effectively completed the charging task while participating in the demand response process. In addition, Figs. 6 (d) and (e) demonstrate that due to the low temperature at 08:00 and to reduce energy costs, the HVAC system did not start until approximately 10:00. However, the HVAC system also maintains the indoor temperature precisely within a comfortable temperature range.

Furthermore, to better illustrate the performance of the proposed method, we present the scheduling results of each DRL method for 24 hours on this test day in Fig. 7. As shown in Fig. 7(a), compared with the total reward obtained by TD3/DDPG/PPO/DQN (-4.86/-8.29/-9.19/-12.96), the total reward obtained by the proposed method (-1.18) is greater, indicating that the proposed method achieves the best overall performance when considering various metrics. Additionally, from Fig. 7(b), (c), and (d), it can be observed that while achieving the lowest electricity costs, the proposed method also minimizes thermal discomfort and range anxiety, thereby maximizing the satisfaction of user comfort requirements. This finding implies that the proposed method not only effectively regulates the indoor temperature but also ensures that the EV battery has sufficient charge before departure. In summary, compared to other DRL methods, the proposed PSTER-TD3 method demonstrates superior energy management, cost-effectiveness, and user comfort during this test day.

5.2.3. Comparison with the benchmark methods

A comparison of the calculation performances of the different methods is shown in Table 4. It can be intuitively observed that the daily online runtime based on the MILP method is the longest, followed by that of MPC and then those of the other DRL methods. This is because both MILP and MPC require rolling calculations of mathematical programming problems with multiple variables. As the problem size increases, the computational complexity increases exponentially, which may make it impossible to meet the requirements of online operation. DRL-based methods can make scheduling decisions immediately using trained DNNs without requiring repeated calculations, which results in shorter response times. Among the DRL-based methods, the DQN has the shortest offline training time, followed by the proposed PSTER-TD3. This is because the DQN must train only one DNN, resulting in a higher training speed. However, the PSTER strategy in the proposed method helps it achieve the shortest online running time (0.342 seconds) of all methods. Therefore, the proposed method can meet the computational performance requirements of online operation better than the compared methods.

Table 5 lists the performance indicators of the different methods on the test set. In terms of the optimization accuracy, the MACE of the proposed method is 12.63%/16.67%/8.34%/23.96%/19.80%/5.21% lower than those of MILP/MPC/DQN/PPO/DDPG/TD3. Moreover, the proposed method is the only one that has an MTD of 0.00 ℃, indicating that only the proposed method effectively maintains the indoor temperature. The proposed method also achieves the smallest RANGE (0.15) and STD (0.048) among all methods, indicating that the variation number and degree of dispersion of the optimization error in the test process are the lowest. Therefore, the proposed PSTER-TD3 has better optimization accuracy and stability than the other methods. In addition, the PSTER strategy and performance improvement technology in the proposed method significantly improve the optimization accuracy and stability of the scheduling strategy learned by the agent.

Notably, compared to the MILP method, which relies on rigorous mathematical theory to obtain theoretically globally optimal solutions, the proposed method achieves the same MTD and achieves better results on MACE while also obtaining lower values of RANGE and STD. This indicates that the proposed method has superior accuracy and stability. MPC also achieves relatively good performance through rolling optimization and feedback correction within a finite time domain. However, similar to the MILP method, its optimization performance relies on prediction accuracy, making it susceptible to the influence of uncertain parameter prediction errors. As a result, both of these methods have inferior optimization accuracy and stability to the proposed method. Another interesting result is that PPO has good optimization stability, but its optimization accuracy is poor. This is because the environment exploration method of PPO based on policy improves the training stability while reducing the motivation of agents to attempt random actions, resulting in conservative suboptimal strategies with lower optimization accuracy. In addition, the DQN has the optimal performance among the DRL-based methods, which indicates that discretizing the action domain significantly reduces the optimization quality of the policy. The reward distributions for different DRL methods during testing are shown in Fig. 8. From the graph, it can be observed that the mean values of the RM electricity cost obtained by DQN/PPO/DDPG/TD3/the method proposed in this paper are 1.24/1.39/1.35/1.21/1.16, the mean values of thermal discomfort are 2.16/1.94/1.73/1.07/0, and the mean values of range anxiety are 9.93/6.48/5.26/3.04/0. Compared to DQN/PPO/DDPG/TD3, the method proposed in this paper has the following three advantages: (i) it reduced the electricity costs by approximately 6.45%/16.55%/14.07%/4.13%; (ii) it decreased thermal discomfort by 2.16/1.94/1.73/1.07; and (iii) it lowered range anxiety by 9.93/6.48/5.26/3.04. Notably, the thermal discomfort of the proposed PSTER-TD3 method is 0, indicating its effective maintenance of the indoor temperature within the threshold range, and it meets the thermal comfort requirements of households; the proposed method also results in the minimum range anxiety, indicating its ability to ensure sufficient battery power for EVs before departure. This means that the proposed method not only effectively reduces the energy cost of residential microgrids but also meets the thermal comfort requirements of households and reduces the range anxiety of EV usage. Consistent with Table 5, Fig. 8 also shows the superiority of the proposed method over other DRL-based methods in the RM optimization scenario. Compared to DDPG, the superior performance of PSTER-TD3 is primarily attributed to the following four features: (i) PSTER-TD3 employs the technique of truncated double Q-learning, effectively mitigating the bias of overestimation by the truncation technique of minimizing the double Q regression target, thus stabilizing the value function optimization. (ii) DDPG introduces noise during policy updates to aid exploration, but excessive noise can lead to instability. In PSTER-TD3, by incorporating a target policy smoothing regularization term, the Q-function along action changes is smoothed, and the stability and convergence of behavioral policy learning are enhanced. (iii) DDPG updates the network parameters at each time step, potentially causing excessive parameter updates and unstable training. By delaying and reducing the update frequency of target networks and actor networks, PSTER-TD3 further improves stability, making the training process smoother. Compared to TD3, the superior performance of PSTER-TD3 is due to its introduction of a priority sampling mechanism based on the sum-tree method. By assigning different priorities to samples in experience replay, important samples can be learned from historical experience more intensively, significantly improving the learning efficiency of the policy network and enhancing the algorithm's global search capability. Additionally, this sparse emphasis update learning method further reduces the bias of overestimation, facilitating stable policy optimization. Overall, the proposed PSTER-TD3 can

Table 6
Performance indicators of different methods under different prediction errors

Methods	±10% *PE*				±15% *PE*				±20% *PE*			
	MACE(%)	*MTD*(℃)	*RANGE*	*STD*	*MACE*(%)	*MTD*(℃)	*RANGE*	*STD*	*MACE*(%)	*MTD*(℃)	*RANGE*	*STD*
MPC	48.96	5.46	0.31	0.124	54.17	6.12	0.34	0.139	61.46	6.84	0.38	0.157
DQN	40.63	12.12	0.46	0.175	45.83	12.77	0.49	0.197	52.08	13.42	0.53	0.223
PPO	55.21	10.61	0.25	0.069	59.38	10.98	0.27	0.079	63.54	11.37	0.29	0.092
DDPG	54.17	9.94	0.37	0.147	61.46	10.61	0.41	0.165	69.79	10.81	0.45	0.186
TD3	35.42	6.38	0.31	0.092	39.58	6.93	0.33	0.101	45.83	7.43	0.36	0.115
Proposed	27.08	0.75	0.18	0.053	30.21	1.09	0.19	0.060	34.38	1.46	0.21	0.071

effectively address the online optimization operation problems of the RM, and it has better optimization accuracy and stability than the compared methods.

5.3. Algorithmic robustness analysis

To verify the robustness of the proposed method against environmental changes, we compare and analyze the performance of different online optimization methods under an additional prediction error (PE) in the test dataset. The size of the PE is sequentially set to ± 10%, ± 15%, and ± 20%.

The comparison of the performance indicators of the different methods for a gradually increasing PE is shown in Table 6. As the PE increases, the optimization accuracy and stability of each method are negatively affected to varying degrees, and their optimization accuracy indicators (MACE and MTD) and optimization stability indicators (RANGE and STD) both increase. This is because the comparison methods encounter more interference states with high bias under an increasing PE, which increases the difficulty of precise scheduling. The proposed method exhibits better performance than the other methods in the three different situations. Although its various performance indicators slightly increase under the influence of PE, no significant mutation occurs, and they remain relatively low. This indicates that the proposed scheduling strategy can adapt to uncertain environments and achieve more stable scheduling results than the other methods can. In addition, Fig. 9 compares the test rewards of various DRL methods under different prediction errors. Consistent with Table 6, this graph verifies the advantages of the proposed method over the other DRL-based methods in resisting environmental uncertainty. Although the proposed method results in certain thermal discomfort and range anxiety punishments as the PE is increased, it still keeps them at a lower value than the other methods. Even under the condition of ± 20% PE, the thermal discomfort and range anxiety of the proposed method are only 0.29 and 0.35, respectively, which are significantly lower than those of the other methods. The above results indicate that compared to the other methods, the proposed method can achieve more stable scheduling results when uncertainty changes occur in the RM environment and is more robust in resisting uncertainty. This is primarily because the proposed method does not rely on predictive inputs or the modeling of environmental transition probabilities. On the basis of learning a large amount of historical environmental data, this method adopts truncated double Q-networks, target policy smoothing and delayed policy updates, as well as the prioritized sum tree experience replay mechanism, which give it good robustness and generalizability. Specifically, (i) the proposed method is capable of capturing data features and adaptively updating network parameters by learning from a large amount of historical environmental data, giving it excellent generalizability to environmental changes. (ii) The design of twin critics within the actor-critic framework enhances the accuracy and robustness of the state value evaluation. Through the mutual supervision and correction of the two critic networks, the risk of overfitting is reduced, and the stability and accuracy of Q-value estimation are improved. This provides reliable guidance for the policy network, thereby enhancing its robustness. (iii) By introducing a regularization term into the output of the target policy, the proposed method smooths the target policy actions, ensuring the continuity of the policy between different states and the stability of policy updates. Moreover, by reducing the update frequency of the actor and target network parameters, it is possible to mitigate the instability caused by continuously changing targets, further enhancing the stability of learning and control and thus strengthening robustness to environmental uncertainty. (iv) Through the efficient sampling mechanism of the prioritized sum-tree structure, the proposed method automatically identifies and focuses on learning from high-value experience samples with larger TD errors, allowing the training distribution to adaptively concentrate on state areas with significant environmental changes, thereby enhancing generalizability. (v) By employing importance sampling based on the prioritized level and reweighting the loss function based on importance weights, critical experiences with high errors can be effectively reused, reinforcing the agent's control ability and enhancing the robustness of the policy.

6. Conclusion

In this paper, a new energy management method based on PSTER-TD3 is proposed for grid-connected RMs. First, considering household thermal discomfort and EV range anxiety, the sequential decision-making problem of RM energy management is described as an MDP with the objective of meeting households' comfort needs and minimizing residential energy costs. Then, the PSTER-TD3 method is proposed to determine the optimal scheduling strategy to achieve this goal. This method integrates a priority experience replay strategy based on a sum tree structure into TD3, enabling agents to learn the optimal strategy in complex environments. In particular, this method combines TD3 with the PSTER strategy proposed in this paper. The proposed PSTER strategy prioritizes the sampling of high TD-error experience related to training with lower time complexity, which further improves the speed and quality of energy management strategy learning. The case study results based on real-world data show that compared with other DRL-based methods and MPC-based methods, the proposed method can effectively meet the needs of households for thermal comfort and range anxiety while reducing RM energy costs. Moreover, the proposed method has stronger robustness when resisting uncertain changes in the RM environment.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 52107108 and in part by the Natural Science Foundation of Hubei Province under Grant 2021CFB163.

References

[1] Zhang Y, Qian W, Ye Y, Li Y, Tang Y, Long Y, et al. A novel non-intrusive load monitoring method based on ResNet-seq2seq networks for energy disaggregation of distributed energy resources integrated with residential houses. Appl Energy 2023;349:121703.

[2] Zhang S, Jia R, Pan H, Cao Y. A safe reinforcement learning-based charging strategy for electric vehicles in residential microgrid. Appl Energy 2023;348:121490.

[3] Manojkumar R, Kumar C, Ganguly S, Gooi HB, Mekhilef S, Catalão JPS. Rule-Based Peak Shaving Using Master-Slave Level Optimization in a Diesel Generator Supplied Microgrid. IEEE Trans Power Syst 2023;38(3):2177-88. http://dx.doi.org/10.1109/TPWRS.2022.3187069.

[4] Yan D, Li T, Ma C, Lai LL, Tsang KF. Cost effective energy management of home energy system with photovoltaic-battery and electric vehicle. In: Proceedings of the IECON 2020 the 46th annual conference. IEEE Ind Electron Soc; 2020, p. 3611-16. http://dx.doi.org/10.1109/IECON43393.2020.9255317.

[5] Zhang B, Hu W, Ghias AM., Xu X, Chen Z. Two-timescale autonomous energy management strategy based on multi-agent deep reinforcement learning approach for residential multicarrier energy system. Appl Energy 2023;351:121777.

[6] Chang X, Xu Y, Sun H. A Distributed Online Learning Approach for Energy Management With Communication Noises. IEEE Trans Sustain Energy 2022;13(1):551-66. http://dx.doi.org/10.1109/TSTE.2021.3119657.

[7] Morsali R, Thirunavukkarasu GS, Seyedmahmoudian M, Stojcevski A, Kowalczyk R. A relaxed constrained decentralised demand side management system of a community-based residential microgrid with realistic appliance models. Appl Energy 2020;277:115626.

[8] Houben N, Cosic A, Stadler M, Mansoor M, Zellinger M, Auer H, Ajanovic A, Haas R. Optimal dispatch of a multi-energy system

microgrid under uncertainty: A renewable energy community in Austria. Appl Energy 2023;337:120913.
[9] Nikmehr N, Zhang P, Bragin MA. Quantum Distributed Unit Commitment: An Application in Microgrids. IEEE Trans Power Syst 2022;37(5):3592-603. http://dx.doi.org/10.1109/TPWRS.2022.3141794.
[10] Yu L, Jiang T, Zou Y. Online energy management for a sustainable smart home with an HVAC load and random occupancy. IEEE Trans Smart Grid 2019;10(2):1646-59. http://dx.doi.org/10.1109/TSG.2017.2775209.
[11] Yu L, Xie D, Huang C, Jiang T， Zou Y. Energy optimization of HVAC Systems in commercial buildings considering indoor air quality management. IEEE Trans Smart Grid 2019;10(5):5103-13. http://dx.doi.org/10.1109/TSG.2018.2875727.
[12] Wang C, Tian T, Xu Z, Cheng S, Liu S， Chen R. Optimal Management for Grid-Connected Three/Single-Phase Hybrid Multimicrogrids. IEEE Trans Sustain. Energy 2020;11(3):1870-82. http://dx.doi.org/10.1109/TSTE.2019.2945924.
[13] Zhao L, Yang T, Li W, Zomaya AY. Deep reinforcement learning-based joint load scheduling for household multi-energy system. Appl Energy 2022;324:119346.
[14] Mocanu E， Mocanu DC, Nguyen PH, Liotta A, Webbe ME, Gibescu M, et al. On-line building energy optimization using deep reinforcement learning. IEEE Trans Smart Grid 2019;10(4):3698–708. http://dx.doi.org/10.1109/TSG.2018.2834219.
[15] Song H, Liu Y, Zhao J, Liu J, Wu G. Prioritized replay dueling DDQN based grid-edge control of community energy storage system. IEEE Trans Smart Grid 2021;12,(6):4950–61. http://dx.doi.org/10.1109/TSG.2021.3099133.
[16] Wang B, Li Y, Ming W, Wang S. Deep reinforcement learning method for demand response management of interruptible load. IEEE Trans Smart Grid 2020;11(4):3146-55. http://dx.doi.org/10.1109/TSG.2020.2967430.
[17] Lillicrap T P, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y. Continuous control with deep reinforcement learning. Proc Int Conf Learn Represent; 2016, p. 834-43. https://doi.org/10.48550/arXiv.1509.02971
[18] Li H, Wan Z, He H. Real-time residential demand response. IEEE Trans Smart Grid 2020;11(5):4144–54. https://doi.org/10.1109/TSG.2020.2978061.
[19] Guo C, Wang X, Zheng Y, Zhang F. Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning. Energy 2022;238:121873. https://doi.org/10.1016/j.energy.2021.121873.
[20] Yu L, Xie W, Xie D, Zou Y, Zhang D, Sun Z, et al. Deep Reinforcement Learning for Smart Home Energy Management. IEEE Internet Things J 2020;7(4):2751–62. https://doi.org/10.1109/JIOT.2019.2957289.
[21] Qin Z, Liu D, Hua H, Cao J. Privacy preserving load control of residential microgrid via deep reinforcement learning. IEEE Trans Smart Grid 2021;12(5):4079-89. https://doi.org/10.1109/TSG.2021.3088290
[22] Wan Z, Li H, He H, Prokhorov D. Model-free real-time EV charging scheduling based on deep reinforcement learning. IEEE Trans Smart Grid 2019;10(5):5246-57. https://doi.org/10.1109/TSG.2018.2879572.
[23] Yu L, Jiang T, Zou Y. Online energy management for a sustainable smart home with an HVAC load and random occupancy. IEEE Trans Smart Grid 2019;10(2):1646–59. https://doi.org/10.1109/TSG.2017.2775209.
[24] Li H, Wan Z, He H. Constrained EV charging scheduling based on safe deep reinforcement learning. IEEE Trans Smart Grid 2020;11(3):2427-39. https://doi.org/10.1109/TSG.2019.2955437.
[25] Pecan Street Database. [Online]. Available: http://www.pecanstreet.org/.
[26] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc of the 3rd Int Conf Learn Represent (ICLR) 2015:1–15. https://doi.org/10.48550/arXiv.1412.6980.
[27] Liu J, Zhong W, Jiao L. A multiagent evolutionary algorithm for constraint satisfaction problems. IEEE Trans Syst Man and Cybern B 2006;36(1):54-73. https://doi.org/10.1109/TSMCB.2005.852980.
[28] Wu D, Xu S, Kong F. Convergence analysis and improvement of the chicken swarm optimization algorithm. IEEE Access 2016;4:9400-12. https://doi.org/10.1109/ACCESS.2016.2604738.