Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

# A home energy management approach using decoupling value and policy in reinforcement learning*

Luolin XIONG[1], Yang TANG[†‡1], Chensheng LIU[1], Shuai MAO[2], Ke MENG[3], Zhaoyang DONG[4], Feng QIAN[†‡1]

[1]*Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China*

[2]*Department of Electrical Engineering, Nantong University, Nantong 226019, China*

[3]*School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney NSW 2052, Australia*

[4]*School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore 639798, Singapore*

[†]E-mail: tangtany@gmail.com; fqian@ecust.edu.cn

Received Dec. 27, 2022; Revision accepted Apr. 9, 2023; Crosschecked May 8, 2023; Published online Aug. 10, 2023

**Abstract:** Considering the popularity of electric vehicles and the flexibility of household appliances, it is feasible to dispatch energy in home energy systems under dynamic electricity prices to optimize electricity cost and comfort residents. In this paper, a novel home energy management (HEM) approach is proposed based on a data-driven deep reinforcement learning method. First, to reveal the multiple uncertain factors affecting the charging behavior of electric vehicles (EVs), an improved mathematical model integrating driver's experience, unexpected events, and traffic conditions is introduced to describe the dynamic energy demand of EVs in home energy systems. Second, a decoupled advantage actor-critic (DA2C) algorithm is presented to enhance the energy optimization performance by alleviating the overfitting problem caused by the shared policy and value networks. Furthermore, separate networks for the policy and value functions ensure the generalization of the proposed method in unseen scenarios. Finally, comprehensive experiments are carried out to compare the proposed approach with existing methods, and the results show that the proposed method can optimize electricity cost and consider the residential comfort level in different scenarios.

**Key words:** Home energy system; Electric vehicle; Reinforcement learning; Generalization
https://doi.org/10.1631/FITEE.2200667 **CLC number:** TP181

## 1 Introduction

With recent advances in the field of electric vehicles (EVs) and intelligent household appliances, flexibility on the demand side is increasing (Parag and Sovacool, 2016; Wen et al., 2021). Meanwhile, fluctuating renewable energy resources bring uncertainties to the energy-supply side (Agnew and Dargusch, 2015; Luo et al., 2020), which have greatly driven the rapid growth of home energy management (HEM) techniques to adjust energy consumption according to real-time electricity prices and flexible renewable energy generation. Efficient HEM techniques are promising in optimizing electricity cost, comforting residents, and reducing carbon emissions, which are achieved by intelligently scheduling and controlling EVs and household appliances (Liu YB et al., 2016;

‡ Corresponding authors

* Project supported by the National Natural Science Foundation of China (Nos. 62293502, 62293500, 62293504, 62073138, and 62173147), the Fundamental Research Funds for the Central Universities, China (No. 222202317006), and the Nanyang Technological University Startup Grant and MOE Tier 1 (No. RG59/22)

ORCID: Luolin XIONG, https://orcid.org/0009-0001-0142-7933; Yang TANG, https://orcid.org/0000-0002-2750-8029; Feng QIAN, https://orcid.org/0000-0003-2781-332X

Xia et al., 2016; Hu et al., 2018; Mao et al., 2019).

Numerous investigations have been carried out into feasible strategies of HEM from the demand-side perspective (Li HP et al., 2020b; Luo et al., 2020; Kong et al., 2021). In the literature, HEM approaches are divided mainly into two categories: model-driven approaches and data-driven approaches (Xu et al., 2020; Xiong et al., 2022). Model-driven methods are superior in terms of reducing the computational burden and alleviating data dependence. For instance, a multi-objective mixed-integer nonlinear programming model has been developed to optimize energy use in a smart home (Anvari-Moghaddam et al., 2015). In this work, a trade-off between energy saving and a high comfort level was considered, and the running of the proposed model-driven algorithms can be finished within several seconds. A probabilistic optimization problem has been designed for HEM in a renewable-energy-based residential energy hub, where the two-point estimation method has been developed to model the uncertainty in the output power of solar panels, and less than 1 min was required for running the proposed HEM program (Rastegar et al., 2017). A new framework has been proposed with related analysis models to determine optimal demand response strategies, where resident occupancy behavior and forecasted renewable energy generation have been estimated (Baek et al., 2021). However, the scalability of these model-driven approaches in different scenarios is poor, and re-established models are a prerequisite for excellent performance in new scenarios. Moreover, the performance of these model-driven methods is dependent heavily on complex system models, the accuracy of which is hard to be guaranteed with multiple uncertainties (Shi et al., 2023).

To tackle these challenges encountered by model-driven approaches, data-driven approaches have gained particular research attention in HEM (Shirsat and Tang, 2021; Huang et al., 2022; Liu SG et al., 2022). A data-driven distributional optimization method has been proposed to guarantee robustness against the worst probability distribution of multiple uncertainties (Saberi et al., 2021). In this study, the authors have pointed out that probability distributions always have errors, even under the assumption that the accurate probability distributions of uncertainties are known. Furthermore, with the rapid development of artificial intelligence technology, the optimization problems of complex large-scale energy systems have been solved by intelligent dynamics optimization methods (Qian, 2019, 2021; Gao et al., 2022). As an emerging phenomenon of artificial intelligence techniques, reinforcement learning (RL) has attracted considerable attention due to its data-driven outstanding decision-making capability and superhuman performance in a number of challenging tasks (Wang HN et al., 2020; Wang YP et al., 2021; Tang et al., 2022; Wang JR et al., 2022). RL techniques are typically used in HEM, and there exist two kinds of approaches: value-based RL and policy-based RL (Xiong et al., 2021). Value-based RL algorithms estimate an action-value table or function and choose the action with the maximum value according to the table or function. Therefore, value-based algorithms can be used only for tasks with discrete actions. For example, to propose a secure demand response management scheme for HEM, Q-learning has been adopted to make optimal price decisions using Markov decision process (MDP) with the objective of reducing energy consumption, which has benefited both consumers and utility providers (Kumari and Tanwar, 2022).

For tasks with high-dimensional action space (such as control of mass appliances) or continuous action space (such as charging control of EVs), it is challenging for value-based algorithms to achieve excellent performance (Qin et al., 2021). Consequently, methods combining value- and policy-based algorithms, such as deep deterministic policy gradient (DDPG) (Zenginis et al., 2022) and advantage actor-critic (A2C) (Shuvo and Yilmaz, 2022), have been used more extensively in HEM. For instance, a novel privacy-preserving load control scheme based on the vectorized A2C algorithm has been developed to tackle high-dimensional action space and the partial observability of state for the residential microgrid, in which the microgrid operator manages a multitude of home appliances, including EVs and air conditioners (Qin et al., 2021). Due to the random nature of electricity prices, appliance demand, and user behavior, a novel reward shaping based actor-critic deep RL (DRL) algorithm has been presented to manage the residential energy consumption profile with limited information about the uncertain factors (Lu et al., 2023). In these methods, accurate estimation of the value function can instruct the HEM agent to learn the most valuable policies and contribute to

remarkable performance.

However, these standard DRL algorithms usually use shared networks for the policy and value functions, which therefore limits the estimation accuracy of the value function (Raileanu and Fergus, 2021). Since the value function estimation requires more information than the policy function estimation, shared networks for the policy and value functions can easily lead to overfitting.

Motivated by the aforementioned phenomenon, in this work, we introduce a decoupled A2C (DA2C) algorithm with separate networks for the policy and value functions to mitigate the overfitting problem in HEM. Different from existing RL algorithms with shared networks for the value and policy functions, which suffer from poor generalization (Li HP et al., 2020a), the DA2C algorithm proposed in Raileanu and Fergus (2021) can not only achieve outstanding performance in reducing the electricity cost and comforting the residents but also show good generalization to new scenarios with different residents and different seasons. The main contributions of this paper are summarized as follows:

1. Multiple uncertainties affecting the dynamic charging behavior of individual EVs, such as driver's experience, unexpected events, and traffic conditions, are integrated into an improved mathematical model. Compared with the model proposed in Yan et al. (2021), the improved EV-charging model describes the energy demand of EVs more exactly.

2. The DA2C approach with separate networks for the policy and value functions is developed to schedule the charging operation of EVs and the energy consumption of household appliances in smart homes. Compared with existing RL algorithms used in HEM (Li HP et al., 2020a), the DA2C algorithm enhances the generalization by decoupling value and policy to alleviate the overfitting problem.

3. A set of experiments are conducted to verify the performance of the proposed data-driven intelligent optimization method. Compared with existing methods, our proposed method shows competitive performance in optimizing electricity cost and comforting the residents even in different scenarios.

## 2 Problem formulation

As shown in Fig. 1, this work aims at developing a data-driven intelligent optimization method based on RL to achieve optimal energy management of smart home systems, by appropriately scheduling the charging operation of EVs and the energy consumption of household appliances on the demand side. In this section, an improved mathematical model describing the dynamic energy demand of individual EVs, which integrates various uncertain factors, is

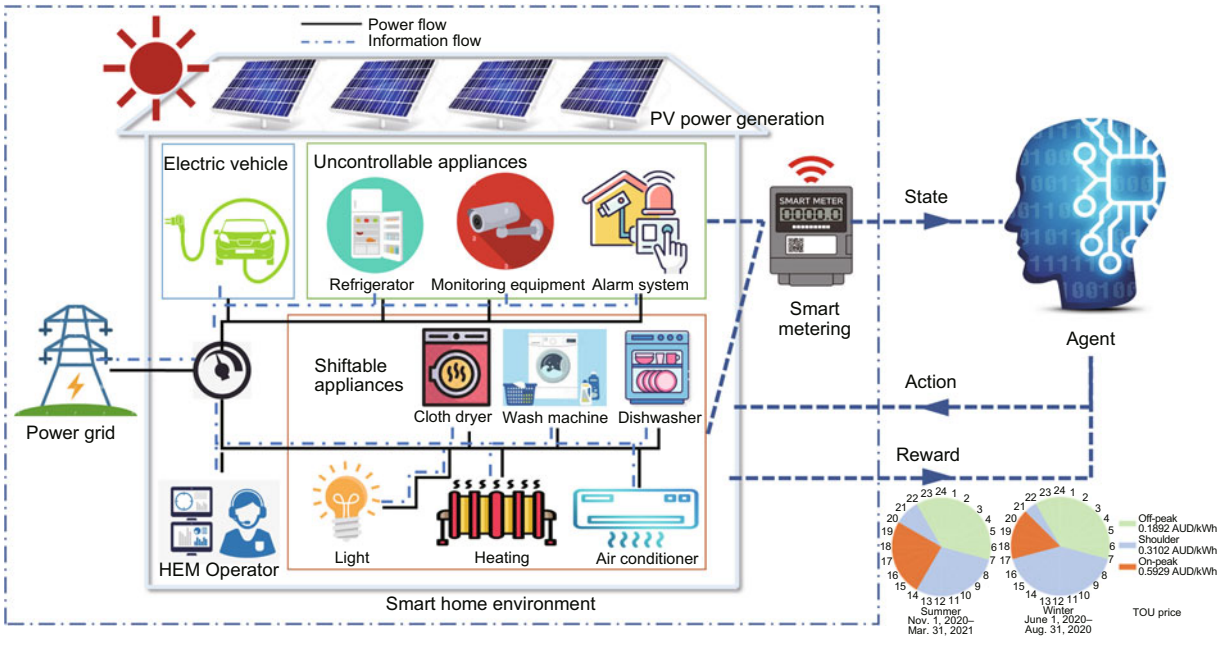


**Fig. 1 Reinforcement learning based home energy management system (PV: photovoltaic)**

proposed first. Furthermore, we formulate the HEM problem as an MDP, in which the charging operation of private EVs and the energy consumption of household appliances are regarded as the actions to be optimized.

### 2.1 Improved EV-charging model

In the HEM problem, the charging operation of private EVs is worthy of in-depth consideration owing to its complex characteristics compared with other household appliances (Zhang YA et al., 2022; Zhang YI et al., 2022). The charging operation depends directly on the energy demand of EVs' batteries. Consequently, we present an improved EV-charging model to describe the energy demand of EVs, which is comprehensively affected by a wide variety of uncertain factors, such as driver's experience, unexpected events, and traffic conditions.

**Remark 1** Specifically, experienced drivers may not charge frequently, compared to novices, to make sure that the state of charge (SoC) always remains at a high level, and they are more tolerant to the anxiety of power depletion. Unexpected events may prompt the EV to leave before the preset departure time, and sudden termination of charging will lead to a lack of energy for the subsequent trip. Besides, traffic conditions have a significant impact on the charging behavior, even if the drivers are experienced. When encountering congestion, accidents, extreme weather, special events, and other costly delays, driver's experience often fails, leading to increased anxiety. Consequently, an improved EV-charging model is presented with the aforementioned uncertain factors integrated.

First, as shown in Eq. (1), the dynamics of the EV batteries are modeled with two modes: charging and discharging:

$$E_{t+1}^{\mathrm{SoC}}=\begin{cases}E_t^{\mathrm{SoC}}+\eta_{\mathrm{ch}}P_t^{\mathrm{EV}}\Delta t, & \text{if } P_t^{\mathrm{EV}}\geq 0,\\ E_t^{\mathrm{SoC}}+\eta_{\mathrm{dis}}P_t^{\mathrm{EV}}\Delta t, & \text{otherwise,}\end{cases} \tag{1}$$

where $E_t^{\mathrm{SoC}}$ denotes the current battery energy of EV at time slot $t$, $P_t^{\mathrm{EV}}$ represents the charging power if $P_t^{\mathrm{EV}}\geq 0$, and if $P_t^{\mathrm{EV}}<0$, it is the discharging power, and $\eta_{\mathrm{ch}}\in(0,1]$ and $\eta_{\mathrm{dis}}\in(0,1]$ are the charging and discharging efficiency coefficients respectively, which vary with heterogeneous battery performance of different EVs. Besides, the battery capacities $C$ of different EVs are varied. Therefore, the energy of different EV batteries is normalized as follows:

$$\mathrm{SoC}_t=\frac{E_t^{\mathrm{SoC}}}{C}, \tag{2}$$

where $\mathrm{SoC}_t$ means the SoC of the EV at time slot $t$. When EV battery is fully charged, $E_t^{\mathrm{SoC}}=C$ and $\mathrm{SoC}_t=1$. There are also some constraints on battery dynamics:

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

$$0\leq E_t^{\mathrm{SoC}}\leq E_{t,\max}^{\mathrm{SoC}}, \tag{4}$$

where $P_{t,\max}^{\mathrm{EV}}$ is the maximum charging power at time slot $t$, and $E_{t,\max}^{\mathrm{SoC}}$ is the maximum level of the battery energy, which is equal to the battery capacity $C$.

Next, the charging operation of individual EVs is determined by driver's anxiety about the SoC, which is further affected by multiple uncertainties including driver's experience, unexpected events, and traffic conditions. It is obvious that conservative drivers with less experience need to ensure a higher SoC level for departure. As the experience increases, drivers can estimate the energy required for the journey more accurately, and the anxiety level will decrease. In addition, it is worth noting that the traffic conditions may affect drivers' judgment of the energy demand for the journey, even for experienced drivers. For example, when traffic jams occur due to extreme weather or traffic accidents, more energy is needed for the same journey. Similarly, when the temperature is high (in summer) or while encountering cold weather conditions (in winter), the use of air conditioning in the EV will lead to more energy consumption. All of the above may cause inaccurate estimation of energy required, which in turn incurs an escalation of anxiety. Here, we describe driver's anxiety with the expected SoC $Z_{\mathrm{SoC}}(t)$ in Eq. (5), which integrates the aforementioned uncertainties comprehensively (Yan et al., 2021):

$$Z_{\mathrm{SoC}}(t)=\frac{k_1k_3\left(\mathrm{e}^{-k_2(t-t_{\mathrm{a}})/(t_{\mathrm{d}}-t_{\mathrm{a}})}-1\right)}{\mathrm{e}^{-k_2}-1}, \tag{5}$$

where $t_{\mathrm{a}}$ and $t_{\mathrm{d}}$ represent the arrival time and the departure time respectively, and $t\in(t_{\mathrm{a}},t_{\mathrm{d}}]$. $k_1\in[0,1]$, $k_2\in(-\infty,0)\cup(0,+\infty)$, and $k_3\in[0,1]$ are the shape parameters changing with driver's experience, unexpected events, and traffic conditions, respectively.

When $t=t_{\mathrm{d}}$, $Z_{\mathrm{SoC}}(t)=k_1k_3$ shows that the expected SoC is determined by parameters $k_1$ and $k_3$.

In other words, the value of $k_1k_3$ reflects the anxiety level of the driver at departure time $t_{\mathrm{d}}$, which is related to driver's experience and traffic conditions. Hence, driver's experience and traffic conditions can be characterized by parameters $k_1$ and $k_3$, respectively. When $k_1k_3=1$, the expected SoC $Z_{\mathrm{SoC}}(t)$ is determined by $k_2$, and these three kinds of uncertainties can be described by the combination of $k_1$, $k_2$, and $k_3$. A higher SoC during the charging duration corresponds to larger $k_1$, $k_2$, and $k_3$ values. Hence, drivers can set the values of $k_1$, $k_2$, and $k_3$ according to their actual situation.

### 2.2 MDP formulation

With the improved EV-charging model, the scheduling problem of home energy is then formulated as an MDP containing four elements $<S, A, R, P>$, where $S$, $A$, $R$, and $P$ represent the state set, action set, reward function, and state transition probability, respectively. Detailed definitions for these four elements in HEM are presented as follows:

1. State: The state $s_t$ contains the energy consumption information of the smart home at time slot $t$:

$$s_t=\{\lambda_t, P_t^{\mathrm{PV}}, P_t^{\mathrm{UL}}, P_t^{\mathrm{SH}}, E_t^{\mathrm{SoC}}\}, \tag{6}$$

where $\lambda_t$ is the time-of-use electricity price, and $P_t^{\mathrm{PV}}$ represents solar photovoltaic (PV) power generation. Moreover, $P_t^{\mathrm{UL}}$ and $P_t^{\mathrm{SH}}$ are the energy consumptions of uncontrollable appliances and shiftable appliances, respectively, at time slot $t$.

2. Action: The actions are defined as follows:

$$a_t=\{\Delta P_t^{\mathrm{SH}}, P_t^{\mathrm{EV}}\}, \tag{7}$$

where $\Delta P_t^{\mathrm{SH}}$ denotes the adjustment of energy consumption of these shiftable appliances; $\Delta P_t^{\mathrm{SH}}>0$ means that the energy consumption increases at time slot $t$, and vice versa.

3. Reward: To achieve the objectives of minimizing the electricity cost and comforting the residents, the reward function integrates the electricity cost and the residential discomfort through weighted coefficients $\beta_1$, $\beta_2$, and $\beta_3$. Residents' discomfort is quantified by the dispatching energy consumption of shiftable appliances and the anxiety about the EV's SoC level.

$$\begin{aligned} r_t=&-\beta_1\lambda_t P_t^{\mathrm{G}}-\beta_2\mid\Delta P_t^{\mathrm{SH}}\mid\Delta t\\ &-\beta_3\max\left(Z_{\mathrm{SoC}}(t)-E_t^{\mathrm{SoC}},0\right), \end{aligned} \tag{8}$$

where $P_t^{\mathrm{G}}=P_t^{\mathrm{UL}}+P_t^{\mathrm{SH}}+\Delta P_t^{\mathrm{SH}}+P_t^{\mathrm{EV}}-P_t^{\mathrm{PV}}$ is the electricity bought from the main grid calculated according to the power balance in home energy systems, $\beta_1$ is the weight of electricity cost to the total cost, $\beta_2$ and $\beta_3$ are discomfort coefficients and vary flexibly with different discomfort feedback of the residents, and $E_t^{\mathrm{SoC}}$ is the actual SoC level of the EV battery. The objective function is to minimize the following discounted cumulative reward for the resident in time period 0–$T$:

$$C_T=-\sum_{t=0}^{T}\gamma^t r_t, \tag{9}$$

where $\gamma\in[0,1]$ is the discount factor for future decisions.

4. Transition: The system state transition reveals the change of system state $s_t$ according to action $a_t$. The transition of $E_t^{\mathrm{SoC}}$ can be depicted by the dynamics of EV batteries shown in Eq. (1). Besides, electricity prices, PV power generation, and power consumption of uncontrolled appliances fluctuate stochastically, which also constitute the transition of the system states.

**Remark 2** Note that time slot $t$ represents an interval of half an hour. Within half an hour, the prices, PV power generation, and energy consumption in state $s_t$ are assumed to be constant. The actions also remain constant in half an hour.

## 3 Proposed approach

In this section, the detailed DA2C algorithm is presented, in which the value network and the policy network are decoupled to meet differentiated information demand for value function estimation and policy function estimation.

It has been already proved that the accurate estimation of the value function requires more information than learning of the optimal policy (Raileanu and Fergus, 2021). The widely used shared networks for the policy and value functions can cause overfitting, which also results in poor generalization performance of RL. However, in the HEM problem, the RL agent encounters various scenarios due to the change in power consumption habits caused by different residents and the change in PV power generation caused by seasonal switching. The generalization performance is critical for the scheduling schemes. An intuitive approach is to provide two completely

independent networks, representing the value network and the policy network, as shown in Fig. 2.

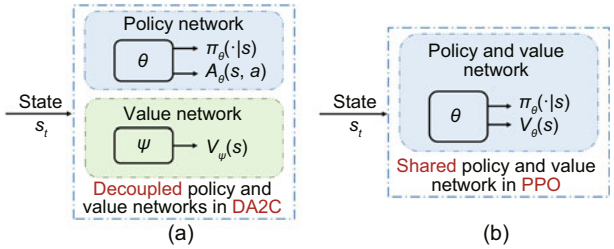


**Fig. 2 Comparison of decoupled architecture in DA2C (a) and shared architecture in PPO (b) (DA2C: decoupled advantage actor-critic; PPO: proximal policy optimization)**

Nevertheless, due to the inaccessibility of the policy network to the gradients from the value function, the agent struggles to learn the optimal policy effectively and suffers from worse training performance compared with approaches using shared network (Cobbe et al., 2021). In fact, the learning process of the policy network rests largely on gradients from the value function for the reason that gradients from the policy function are drastically sparse and show high variance (Raileanu and Fergus, 2021). On the contrary, gradients from the value function are denser and less noisy. Accordingly, an auxiliary loss consisting of the advantage is introduced to guide the training of the policy network, which is a relative measure of the action value. Moreover, it is worth emphasizing that networks of the RL agent should be relatively shallow to ensure fast fitting in the HEM problem compared with deeper networks in the field of image processing. Besides, larger networks are not accepted because excessive redundant parameters may aggravate overfitting (Ota et al., 2020).

In detail, we use two separate networks to represent the policy function and the value function, as depicted in Fig. 3. The policy network is parameterized with $\theta$, which is trained to learn the optimal scheduling policy $\pi_\theta(s_t)$ and the advantage function $A_\theta(s_t, a_t)$. The value network is parameterized with $\psi$ and is adopted to predict the value function $V_\psi(s_t)$. Then, the objective of the policy network and the loss of the value network are depicted as follows:

The objective of the policy network is represented by the following expression:

$$J_{\mathrm{DA2C}}(\theta) = J_\pi(\theta) + \rho_{\mathrm{e}} E_\pi(\theta) - \rho_{\mathrm{a}} L_A(\theta), \qquad (10)$$

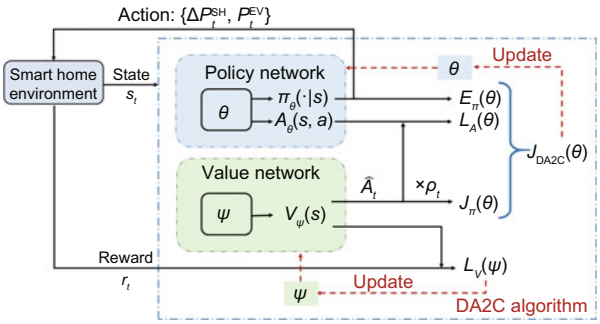


**Fig. 3 Decoupled advantage actor-critic (DA2C) algorithm**

where $J_\pi(\theta)$ is the policy gradient objective, as shown in Eq. (11). It is the same as the objective in the classic proximal policy optimization (PPO) algorithm (Schulman et al., 2017). $E_\pi(\theta)$ is the entropy used to encourage exploration, $L_A(\theta)$ is the auxiliary advantage loss, providing a useful gradient for training the policy network, and $\rho_{\mathrm{e}}$ and $\rho_{\mathrm{a}}$ represent the weights corresponding to the importance of the entropy $E_\pi(\theta)$ and the loss $L_A(\theta)$, respectively.

Specifically, the policy gradient objective is defined as follows:

$$\begin{aligned} J_\pi(\theta) = &\hat{E}_t\Big[\min\Big(\rho_t(\theta)\hat{A}_t, \mathrm{Clip}(\rho_t(\theta), \\ &1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\Big], \end{aligned} \qquad (11)$$

where $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}$, $\hat{A}_t = \sum_{k=t}^{T} \gamma^{k-t}\delta_k$ is the advantage function estimated with $\delta_t = r_t + \gamma V_\psi(s_{t+1}) - V_\psi(s_t)$ computed from the value network.

The auxiliary advantage loss is presented as

$$L_A(\theta) = \hat{E}_t\big(A_\theta(s_t, a_t) - \hat{A}_t\big)^2, \qquad (12)$$

where $A_\theta(s_t, a_t)$ is one of the policy network's outputs.

The loss of the value network is represented by the following expression:

$$L_V(\psi) = \hat{E}_t\big(V_\psi(s_t) - \hat{V}_t\big)^2, \qquad (13)$$

where $V_\psi(s_t)$ is the output of the value network, and $\hat{V}_t = \sum_{k=t}^{T} \gamma^{k-t} r_k$ is the total discounted reward obtained during the corresponding episodes after time slot $t$.

Since two separate networks are used to learn the policy and value functions, the learning process of each network can be decoupled and different updating frequencies are allowed. It has been found

that the value network allows for larger amounts of sample reuse than the policy network (Cobbe et al., 2021). Hence, it is worth exploring whether updating the value network for every $n$ updates of the policy network contributes to better performance and training stability. The pseudo-code of the proposed DA2C is displayed in Algorithm 1.

**Algorithm 1** Decoupled advantage actor-critic (DA2C)

**Initialize:** parameters $\theta$ of the policy network and $\psi$ of the value network
**Input:** $s_t=\{\lambda_t, P_t^{\mathrm{PV}}, P_t^{\mathrm{UL}}, P_t^{\mathrm{SH}}, E_t^{\mathrm{SoC}}\}$
**Onput:** $a_t=\{\Delta P_t^{\mathrm{SH}}, P_t^{\mathrm{EV}}\}$
**while** $i<N_{\mathrm{ep}}$ **do**
  Collect $D=\{(s_t,a_t,r_t,s_{t+1})\}_{t+1}^{T}$ using $\pi(\theta)$
  Compute the value and advantage targets $\hat{V}_t$ and $\hat{A}_t$ for all state $s_t$
  **while** $j_\pi<N_\pi$ **do**
    $L_A(\theta)=\hat{E}_t\left(A_\theta(s_t,a_t)-\hat{A}_t\right)^2$
    $J_{\mathrm{DA2C}}(\theta)=J_\pi(\theta)+\rho_{\mathrm{e}}E_\pi(\theta)-\rho_{\mathrm{a}}L_A(\theta)$
    $\theta\leftarrow\mathrm{argmax}_\theta J_{\mathrm{DA2C}}$
    $j_\pi+1$
  **end while**
  **if** $i\%n==0$ **then**
    **while** $j_V<N_V$ **do**
      $L_V(\psi)=\hat{E}_t\left(V_\psi(s_t)-\hat{V}_t\right)^2$
      $\psi\leftarrow\mathrm{argmin}_\psi L_V$
      $j_V+1$
    **end while**
  **end if**
  $i+1$
**end while**

## 4 Experiments and results

The effectiveness of the proposed DA2C approach in HEM is verified in scheduling tasks of various residents with different electricity preferences. Moreover, the generalization to unseen scenarios of the learning algorithm is confirmed. In this section, the experimental setup is described first. Then, the optimization results and generalization performance compared with other approaches are verified, and the corresponding discussion is provided. Finally, the analysis of the hyperparameters is shown, including the value of $n$ and the trade-off between electricity cost and residential comfort level.

### 4.1 Experimental setup

For the evaluation of the proposed HEM scheme, the dataset and the hyperparameters are first presented in this part. We consider the energy scheduling of intelligent appliances and EVs during one day, with 30 min as one time slot, such that the time horizon $T$ is 48. The dataset consists of four randomly selected residents in the Australian grid, who are billed on a domestic tariff and have a gross metered solar system installed from July 1, 2010, to June 30, 2013 (Ausgrid, 2014). The real-world power consumption data of household appliances and PV power generation data are also obtained from the above dataset. The time-of-use electricity prices have been obtained from Zhou (2020), which are different in the off-peak, shoulder, and on-peak periods, and the division of time periods is different in summer and winter. Other parameters concerning the RL algorithm are reported in Table 1. Both the policy network and the value network consist of three fully connected layers, each with 100 neurons. It is worth noting that 10 trials are implemented for each of the experiments, and the experiments are implemented on Python 3.6.13 with the machine learning package PyTorch 1.8.1.

**Table 1 List of hyperparameters used in the proposed DA2C approach**

| Hyperparameter | Value |
|---|---|
| $\gamma$ | 0.999 |
| $\rho_{\mathrm{e}}$ | 0.01 |
| $\rho_{\mathrm{a}}$ | 0.25 |
| $N_\pi$ | 1 |
| $N_V$ | 9 |
| $n$ | 1 |
| $N_{\mathrm{ep}}$ | 500 |
| Learning rate | 5e-4 |
| Optimizer | Adam |

### 4.2 Performance comparison

In this case, an HEM scheme for different residents with different electricity preferences is designed to verify the generalization performance of the proposed approach under dynamic electricity prices in different seasons. We first divide the scheduling tasks of different residents into the training and test tasks. Specifically, the data from residents 1–3 are used for training, while the data from resident 4 are used for test. Besides, the generalization performance is verified with cross-seasonal tasks. The scheduling tasks during spring, summer, and autumn are viewed as training tasks, and scheduling in winter

is the test task. The proposed approach is compared with several benchmarks, such as PPO (Li HP et al., 2020a), invariant decoupled advantage actor-critic (IDAAC) (Raileanu and Fergus, 2021), and attention-based partially decoupled actor-critic (APDAC) (Nafi et al., 2022). Besides, to verify the efficacy of the state and reward definitions in the proposed MDP, we compare the proposed MDP with different MDP formulations (Xu et al., 2020).

The method is first tested over resident 4 when the agent is trained over residents 1–3. The generalization performance can be verified due to the differences in power consumption mode and time among different residents. Table 2 shows the training and test performances of different approaches. DA2C outperforms all the methods on both training and test tasks. Fig. 4 illustrates the training and test performances on tasks from different residents, where the average (dark line) and the standard deviation (shaded area) are shown. We show comparisons between the RL algorithms: PPO, IDAAC, and APDAC. Our approach, namely DA2C, shows superior results on the training level, relative to the other three methods. In addition, DA2C achieves comparable performance on the test level for the scheduling tasks of resident 4, which further illustrates the superiority of the decoupled policy and value networks in enhancing optimization and generalization performances.

Moreover, we define the MDP formulation in Xu et al. (2020) as MDP_1, whose system states, actions, and reward functions are different from those of ours. Obviously, the averaged cost we compared is directly related to the reward function, so we define MDP_2 with the same reward function as our MDP, the same system states and actions as those in Xu et al. (2020). Table 2 demonstrates the efficacy of our MDP. Comparing the results of MDP and MDP_2, it can be found that the system state and action definitions are reasonable and effective since lower cost is achieved with the same algorithm. Results of MDP_1 are better than those of MDP, because the dissatisfaction cost in MDP_1 is related only to the upper bound of energy consumption. In our MDP, the dissatisfaction cost is more comprehensive, integrating uncertainties such as driver's experience, unexpected events, and traffic conditions, which also leads to higher costs. As shown in Fig. 5, our proposed method achieves the best performance in most problems, except for the training process of MDP_1, which also verifies the effectiveness of our method in dealing with generalizable HEM problems.

Besides, we compare the total cost during different seasons, and the agent is tested over winter while being trained over spring, summer, and autumn. The appliances in home energy systems are operated differently in different seasons, such as the heating, ventilation, and air conditioning (HVAC) systems operating in the cooling mode in summer and the heating mode in winter. Table 3 shows the training and test performances of different approaches and reveals that DA2C outperforms the other methods on

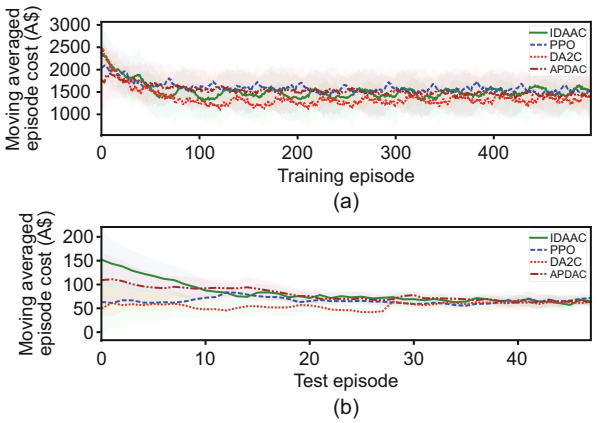


**Fig. 4 Comparison of different methods in terms of training (a) and test (b) performances on tasks from different residents**

**Table 2 Comparison of different methods and different Markov decision processes (MDPs) in terms of training and test performances on tasks from different residents**

| Method | Training cost ($\times 10^2$ A$) | | | Test cost (A$) | | |
|---|---|---|---|---|---|---|
| | MDP | MDP_1 | MDP_2 | MDP | MDP_1 | MDP_2 |
| PPO | 15.45±7.08 | **9.31±6.40** | 17.98±10.53 | 59.87±24.65 | 60.67±34.72 | 65.56±21.27 |
| IDAAC | 14.93±9.35 | 10.01±7.80 | 14.12 ± 7.67 | 62.57±25.38 | 51.70±23.76 | **58.92±22.38** |
| APDAC | 14.58±8.86 | 13.51±8.32 | 15.17 ± 6.63 | 64.13±21.25 | 43.33±18.01 | 68.32±23.20 |
| DA2C | **13.31±7.40** | 9.90±4.88 | **13.90±7.72** | **58.33±25.72** | **42.76±18.80** | 60.87±20.64 |

Results in bold and cells colored gray denote the best and the second best, respectively

both training and test tasks. The average (dark line) and standard deviation (shaded area) are depicted in Fig. 6 to illustrate the outstanding performance on tasks from different seasons.

### 4.3 Analysis of hyperparameters

In this part, we first search for a hyperparameter of the number of value updates, after which we perform a policy update $n \in \{1, 8, 32\}$. The results are presented in Fig. 7, and $n = 1$ is found to be the best hyperparameter. Therefore, we use this value to obtain the results reported in the paper.

To further demonstrate the trade-off between electricity cost and residential comfort, we change the values of $\beta_1$, $\beta_2$, and $\beta_3$ to show the preference of reducing electricity cost and comforting residents. Higher $\beta_2$ and $\beta_3$ means that the residents prefer comforting themselves. In contrast, higher $\beta_1$ reveals higher cost sensitivity and higher tolerance for discomfort. The numerical results obtained from residents with different preferences are shown in Table 4. It is intuitive that residents with higher cost sensitivity have lower total cost and those with higher comfort requirements need to pay more fees. Fig. 8 reflects the fact that the cost reduces along with the increase of $\beta_1$. Besides, when the residents prefer comforting themselves, as shown by the blue and red dotted lines in Fig. 8, the total cost even increases with the increase of the episodes, which is the cost that must be paid to improve the comfort level.

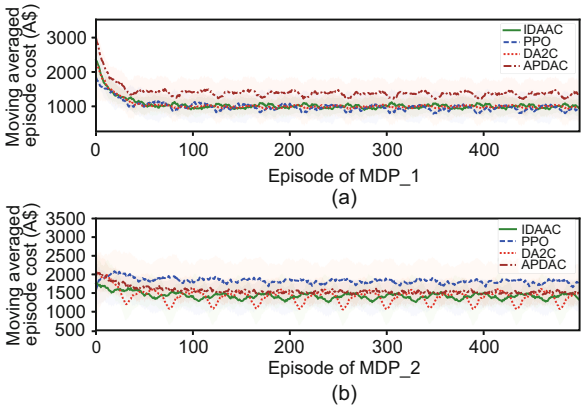


**Fig. 5 Comparison of different methods with different MDPs in terms of different residents: (a) MDP_1; (b) MDP_2**

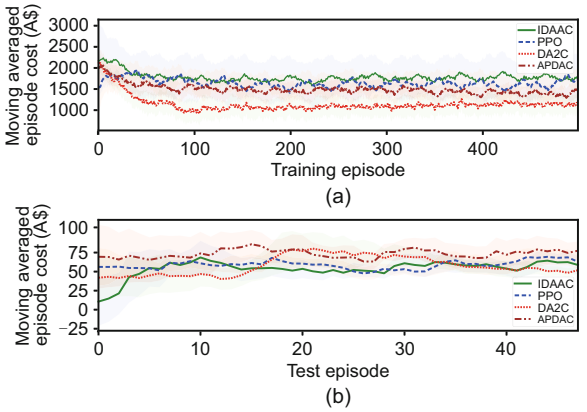


**Fig. 6 Comparison of different methods in terms of different seasons: (a) training; (b) test**

## 5 Conclusions

In this paper, a data-driven intelligent optimization approach is proposed to achieve optimal energy management of smart home systems, by appropriately scheduling the charging operation of EVs

**Table 3 Comparison of different methods in terms of different seasons**

| Method | Training with our MDP ($\times 10^2$ A$) | Test with our MDP (A$) |
|---|---|---|
| PPO | 16.26±5.40 | 57.76±20.36 |
| IDAAC | 17.67±9.36 | 56.07±21.97 |
| APDAC | 14.26±4.19 | 71.35±29.36 |
| DA2C | **11.27 ± 5.00** | **55.09 ± 17.29** |

Best results are in bold

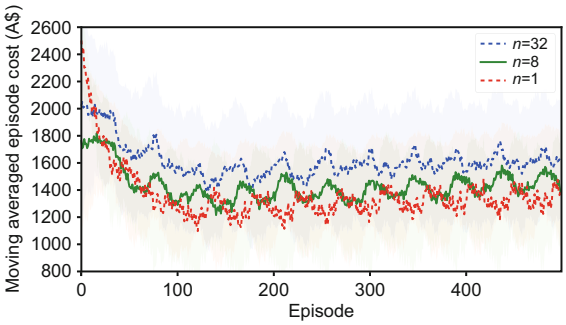


**Fig. 7 Hyperparameter search $n \in \{1,8,32\}$**

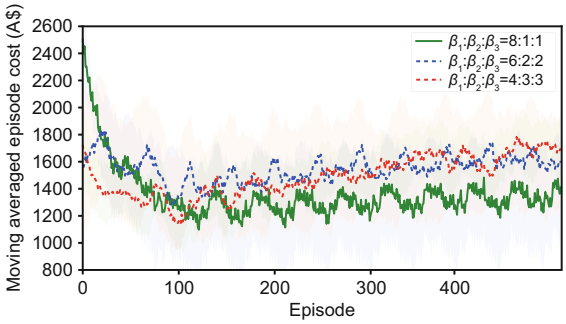


**Fig. 8 Trade-off between electricity cost and residential comfort level (References to color refer to the online version of this figure)**

**Table 4 Trade-off between electricity cost and residential comfort level**

| $\beta_1 : \beta_2 : \beta_3$ | Cost ($\times 10^2$A$) |
|---|---|
| 8 : 1 : 1 | 13.31±7.40 |
| 6 : 2 : 2 | 15.89±7.43 |
| 4 : 3 : 3 | 16.32±3.96 |

and the energy consumption of household appliances. First, an improved mathematical model is designed to quantify EVs' energy demand more practically and completely, where driver's experience, unexpected events, and traffic conditions are integrated. Second, a novel RL-based DA2C approach is developed to alleviate the overfitting problem and enhance generalization performance by decoupling the policy and value networks. Moreover, a set of experiments carried out on practical data from the Australian grid verify the performance and generalization ability of the proposed approach for the HEM problem. In our future work, we plan to extend the approach to more complex large-scale energy systems with multiple uncertainties (Liu ZT et al., 2022, 2023; Zhang HF et al., 2022), such as smart home systems with HVAC and combined cooling heating and power systems. Moreover, privacy preservation is an interesting topic in energy systems, which not only guarantees the privacy of residential information but also contributes to cyber-physical security (Li JH, 2018; Mao et al., 2021; Wang AJ et al., 2022).

## Contributors

Luolin XIONG designed the research. Luolin XIONG, Yang TANG, and Chensheng LIU proposed the methods. Luolin XIONG conducted the experiments. Ke MENG and Zhaoyang DONG processed the data. Luolin XIONG and Yang TANG participated in the visualization. Luolin XIONG drafted the paper. Yang TANG and Shuai MAO helped organize the paper. Yang TANG, Chensheng LIU, Shuai MAO, and Feng QIAN revised and finalized the paper.

## Compliance with ethics guidelines

Yang TANG is a guest editor of this special feature, and he was not involved with the peer review process of this manuscript. Luolin XIONG, Yang TANG, Chensheng LIU, Shuai MAO, Ke MENG, Zhaoyang DONG, and Feng QIAN declare that they have no conflict of interest.

## Data availability

The authors confirm that the data supporting the findings of this study are available within the article.

## References

Agnew S, Dargusch P, 2015. Effect of residential solar and storage on centralized electricity supply systems. *Nat Climate Change*, 5(4):315-318. https://doi.org/10.1038/nclimate2523

Anvari-Moghaddam A, Monsef H, Rahimi-Kian A, 2015. Optimal smart home energy management considering energy saving and a comfortable lifestyle. *IEEE Trans Smart Grid*, 6(1):324-332. https://doi.org/10.1109/TSG.2014.2349352

Ausgrid, 2014. Solar Home Electricity Data. http://www.ipart.nsw.gov.au [Accessed on Nov. 30, 2022].

Baek K, Lee E, Kim J, 2021. Resident behavior detection model for environment responsive demand response. *IEEE Trans Smart Grid*, 12(5):3980-3989. https://doi.org/10.1109/TSG.2021.3074955

Cobbe K, Hilton J, Klimov O, et al., 2021. Phasic policy gradient. Proc 38th Int Conf on Machine Learning, p.2020-2027.

Gao HJ, Li ZK, Yu XH, et al., 2022. Hierarchical multiobjective heuristic for PCB assembly optimization in a beam-head surface mounter. *IEEE Trans Cybern*, 52(7):6911-6924. https://doi.org/10.1109/TCYB.2020.3040788

Hu KY, Li WJ, Wang LD, et al., 2018. Energy management for multi-microgrid system based on model predictive control. *Front Inform Technol Electron Eng*, 19(11):1340-1351. https://doi.org/10.1631/FITEE.1601826

Huang G, Wu F, Guo CX, 2022. Smart grid dispatch powered by deep learning: a survey. *Front Inform Technol Electron Eng*, 23(5):763-776. https://doi.org/10.1631/FITEE.2000719

Kong WC, Luo FJ, Jia YW, et al., 2021. Benefits of home energy storage utilization: an Australian case study of demand charge practices in residential sector. *IEEE Trans Smart Grid*, 12(4):3086-3096. https://doi.org/10.1109/TSG.2021.3054126

Kumari A, Tanwar S, 2022. A reinforcement-learning-based secure demand response scheme for smart grid system. *IEEE Internet Things J*, 9(3):2180-2191. https://doi.org/10.1109/JIOT.2021.3090305

Li HP, Wan ZQ, He HB, 2020a. A deep reinforcement learning based approach for home energy management system. Proc IEEE Power & Energy Society Innovative Smart Grid Technologies Conf, p.1-5. https://doi.org/10.1109/ISGT45199.2020.9087647

Li HP, Wan ZQ, He HB, 2020b. Real-time residential demand response. *IEEE Trans Smart Grid*, 11(5):4144-4154. https://doi.org/10.1109/TSG.2020.2978061

Li JH, 2018. Cyber security meets artificial intelligence: a survey. *Front Inform Technol Electron Eng*, 19(12):1462-1474. https://doi.org/10.1631/FITEE.1800573

Liu SG, Zheng SZ, Zhang WB, et al., 2022. A power resource dispatching framework with a privacy protection function in the power Internet of Things. *Front Inform Technol Electron Eng*, 23(9):1354-1368. https://doi.org/10.1631/FITEE.2100518

Liu YB, Liu JY, Taylor G, et al., 2016. Situational awareness architecture for smart grids developed in accordance with dispatcher's thought process: a review. *Front Inform Technol Electron Eng*, 17(11):1107-1121. https://doi.org/10.1631/FITEE.1601516

Liu ZT, Lin WY, Yu XH, et al., 2022. Approximation-free robust synchronization control for dual-linear-motors-driven systems with uncertainties and disturbances. *IEEE Trans Ind Electron*, 69(10):10500-10509. https://doi.org/10.1109/TIE.2021.3137619

Liu ZT, Gao HJ, Yu XH, et al., 2023. B-spline wavelet neural network-based adaptive control for linear motor-driven systems via a novel gradient descent algorithm. *IEEE Trans Ind Electron*, early access. https://doi.org/10.1109/TIE.2023.3260318

Lu RZ, Jiang ZY, Wu HM, et al., 2023. Reward shaping-based actor-critic deep reinforcement learning for residential energy management. *IEEE Trans Ind Inform*, 19(3):2662-2673. https://doi.org/10.1109/TII.2022.3183802

Luo FJ, Kong WC, Ranzi G, et al., 2020. Optimal home energy management system with demand charge tariff and appliance operational dependencies. *IEEE Trans Smart Grid*, 11(1):4-14. https://doi.org/10.1109/TSG.2019.2915679

Mao S, Wang B, Tang Y, et al., 2019. Opportunities and challenges of artificial intelligence for green manufacturing in the process industry. *Engineering*, 5(6):995-1002. https://doi.org/10.1016/j.eng.2019.08.013

Mao S, Tang Y, Dong ZW, et al., 2021. A privacy preserving distributed optimization algorithm for economic dispatch over time-varying directed networks. *IEEE Trans Ind Inform*, 17(3):1689-1701. https://doi.org/10.1109/TII.2020.2996198

Nafi NM, Glasscock C, Hsu W, 2022. Attention-based partial decoupling of policy and value for generalization in reinforcement learning. Proc 21st IEEE Int Conf on Machine Learning and Applications, p.15-22. https://doi.org/10.1109/ICMLA55696.2022.00011

Ota K, Oiki T, Jha D, et al., 2020. Can increasing input dimensionality improve deep reinforcement learning? Proc 37th Int Conf on Machine Learning, p.7424-7433.

Parag Y, Sovacool BK, 2016. Electricity market design for the prosumer era. *Nat Energy*, 1(4):16032. https://doi.org/10.1038/nenergy.2016.32

Qian F, 2019. Smart process manufacturing systems: deep integration of artificial intelligence and process manufacturing. *Engineering*, 5(6):981. https://doi.org/10.1016/j.eng.2019.10.002

Qian F, 2021. Editorial for special issue "artificial intelligence energizes process manufacturing". *Engineering*, 7(9):1193-1194. https://doi.org/10.1016/j.eng.2021.08.003

Qin ZM, Liu D, Hua HC, et al., 2021. Privacy preserving load control of residential microgrid via deep reinforcement learning. *IEEE Trans Smart Grid*, 12(5):4079-4089. https://doi.org/10.1109/TSG.2021.3088290

Raileanu R, Fergus R, 2021. Decoupling value and policy for generalization in reinforcement learning. Proc 38th Int Conf on Machine Learning, p.8787-8798.

Rastegar M, Fotuhi-Firuzabad M, Zareipour H, et al., 2017. A probabilistic energy management scheme for renewable-based residential energy hubs. *IEEE Trans Smart Grid*, 8(5):2217-2227. https://doi.org/10.1109/TSG.2016.2518920

Saberi H, Zhang C, Dong ZY, 2021. Data-driven distributionally robust hierarchical coordination for home energy management. *IEEE Trans Smart Grid*, 12(5):4090-4101. https://doi.org/10.1109/TSG.2021.3088433

Schulman J, Wolski F, Dhariwal P, et al., 2017. Proximal policy optimization algorithms. https://arxiv.org/abs/1707.06347

Shi PW, Sun WC, Yang XB, et al., 2023. Master-slave synchronous control of dual-drive gantry stage with cogging force compensation. *IEEE Trans Syst Man Cybern Syst*, 53(1):216-225. https://doi.org/10.1109/TSMC.2022.3176952

Shirsat A, Tang WY, 2021. Data-driven stochastic model predictive control for DC-coupled residential PV-storage systems. *IEEE Trans Energy Convers*, 36(2):1435-1448. https://doi.org/10.1109/TEC.2021.3061360

Shuvo SS, Yilmaz Y, 2022. Home energy recommendation system (HERS): a deep reinforcement learning method based on residents' feedback and activity. *IEEE Trans Smart Grid*, 13(4):2812-2821. https://doi.org/10.1109/TSG.2022.3158814

Tang Y, Zhao C, Wang J, et al., 2022. Perception and navigation in autonomous systems in the era of learning: a survey. *IEEE Trans Neur Netw Learn Syst*, early access. https://doi.org/10.1109/TNNLS.2022.3167688

Wang AJ, Liu WP, Dong T, et al., 2022. DisEHPPC: enabling heterogeneous privacy-preserving consensus-based scheme for economic dispatch in smart grids. *IEEE Trans Cybern*, 52(6):5124-5135. https://doi.org/10.1109/TCYB.2020.3027572

Wang HN, Liu N, Zhang YY, et al., 2020. Deep reinforcement learning: a survey. *Front Inform Technol Electron Eng*, 21(12):1726-1744. https://doi.org/10.1631/FITEE.1900533

Wang JR, Hong YT, Wang JL, et al., 2022. Cooperative and competitive multi-agent systems: from optimization to games. *IEEE/CAA J Autom Sin*, 9(5):763-783. https://doi.org/10.1109/JAS.2022.105506

Wang YP, Zheng KX, Tian DX, et al., 2021. Pre-training with asynchronous supervised learning for reinforcement learning based autonomous driving. *Front Inform Technol Electron Eng*, 22(5):673-686. https://doi.org/10.1631/FITEE.1900637

Wen GH, Yu XH, Liu ZW, 2021. Recent progress on the study of distributed economic dispatch in smart grid: an overview. *Front Inform Technol Electron Eng*, 22(1):25-39. https://doi.org/10.1631/FITEE.2000205

Xia YH, Liu JY, Huang ZW, et al., 2016. Carbon emission impact on the operation of virtual power plant with combined heat and power system. *Front Inform Technol Electron Eng*, 17(5):479-488. https://doi.org/10.1631/FITEE.1500467

Xiong LL, Mao S, Tang Y, et al., 2021. Reinforcement learning based integrated energy system management: a survey. *Acta Autom Sin*, 47(10):2321-2340 (in Chinese). https://doi.org/10.16383/j.aas.c210166

Xiong LL, Tang Y, Mao S, et al., 2022. A two-level energy management strategy for multi-microgrid systems with interval prediction and reinforcement learning. *IEEE Trans Circ Syst I Regul Pap*, 69(4):1788-1799. https://doi.org/10.1109/TCSI.2022.3141229

Xu X, Jia YW, Xu Y, et al., 2020. A multi-agent reinforcement learning-based data-driven method for home energy management. *IEEE Trans Smart Grid*, 11(4):3201-3211. https://doi.org/10.1109/TSG.2020.2971427

Yan LF, Chen X, Zhou JY, et al., 2021. Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors. *IEEE Trans Smart Grid*, 12(6):5124-5134. https://doi.org/10.1109/TSG.2021.3098298

Zenginis I, Vardakas J, Koltsaklis NE, et al., 2022. Smart home's energy management through a clustering-based reinforcement learning approach. *IEEE Internet Things J*, 9(17):16363-16371. https://doi.org/10.1109/JIOT.2022.3152586

Zhang HF, Yue D, Dou CX, et al., 2022. Two-layered hierarchical optimization strategy with distributed potential game for interconnected hybrid energy systems. *IEEE Trans Cybern*, early access. https://doi.org/10.1109/TCYB.2022.3142035

Zhang YA, Yang QY, An D, et al., 2022. Multistep multiagent reinforcement learning for optimal energy schedule strategy of charging stations in smart grid. *IEEE Trans Cybern*, 53(7):4292-4305. https://doi.org/10.1109/TCYB.2022.3165074

Zhang YI, Ai ZY, Chen JC, et al., 2022. Energy-saving optimization and control of autonomous electric vehicles with considering multiconstraints. *IEEE Trans Cybern*, 52(10):10869-10881. https://doi.org/10.1109/TCYB.2021.3069674

Zhou SP, 2020. Summary of Time of Use Electricity Price Policy at Home and Abroad (in Chinese). https://shoudian.bjx.com.cn/html/20200807/1095247.shtml [Accessed on Nov. 30, 2022].