Optimal Scheduling of Residential Heating, Ventilation and Air Conditioning Based on Deep Reinforcement Learning

Mingchao Xia, Fangjian Chen, Qifang Chen, Siwei Liu, Yuguang Song, and Te Wang

***Abstract*—Residential heating, ventilation and air conditioning (HVAC) provides important demand response resources for the new power system with high proportion of renewable energy. Residential HAVC scheduling strategies that adapt to real-time electricity price signals formulated by demand response program and ambient temperature can significantly reduce electricity costs while ensuring occupants' comfort. However, since the pricing process and weather conditions are affected by many factors, conventional model-based method is difficult to meet the scheduling requirements in complex environments. To solve this problem, we propose an adaptive optimal scheduling strategy for residential HVAC based on deep reinforcement learning (DRL) method. The scheduling problem can be regarded as a Markov decision process (MDP). The proposed method can adaptively learn the state transition probability to make economical decision under the tolerance violations. Specifically, the residential thermal parameters obtained by the least-squares parameter estimation (LSPE) can provide a basis for the state transition probability of MDP. Daily simulations are verified under the electricity prices and temperature data sets, and numerous experimental results demonstrate the effectiveness of the proposed method.**

***Index Terms*—Residential heating, ventilation and air conditioning (HVAC), scheduling, deep reinforcement learning, least-squares parameter estimation (LSPE).**

I. Introduction

DEVELOPING the clean energy and promoting the transformation of the energy structure vigorously are inevitable requirements for ensuring the sustainable development. The proportion of non-fossil energy such as wind power (WP) and photovoltaics (PVs) at the power supply side is increasing. However, the intermittent and uncertain outputs of WP and PV have made the planning, the control, and the balance of power grid more complicated [1]. The new power system with high proportion of renewable energies puts forward new requirements for system flexibility [2]-[4]. Due to the reduction of the proportion of traditional generating units, the standby resources at the generation side are scarce, and it is necessary to fully tap the flexibility potential of other resources. Demand response (DR) adopts the direct control or the price guidance to regulate the flexible load resources at the demand side, so as to absorb renewable energies and reduce the operation cost of power grid [5]-[8].

Heating, ventilation and air conditioning (HVAC) accounts for about 45% of average summer peak-day loads [9]. Moreover, due to the transferability of HVAC load, it is used as an important DR resource to provide flexible adjustment capabilities for the power grid [10]. Thus, on the premise of not having a significant impact on the thermal comfort of occupants, formulating the optimal scheduling strategies of HVAC participation in power grid DR has become a hot topic of concern to scholars. Reference [11] proposes an optimal strategy for participation of commercial HVAC systems in frequency regulation. It assumes that the aggregator has signed a contract with the power grid, so it adopts the method of direct control and does not take into account the impact of prices and environment on the model. Reference [12] proposes a direct load control algorithm of HVAC, which uses the temperature priority list to formulate the optimal scheduling plan of HVAC, and uses six different outdoor temperature baseline loads to simulate different weather conditions. Reference [13] proposes a bi-level decision model for HVAC to participate in power grid DR, where the upper layer formulates the optimal retail electricity prices, and the lower layer formulates the optimal HVAC scheduling strategy to save electricity costs. In the modeling process of [12] and [13], the typical curve or deterministic data is used as the ambient temperature, and the uncertainty caused by prediction errors is ignored, which may affect the optimal policy decision.

In order to make full use of HVAC flexibility resources to participate in DR programs and reduce costs, it is necessary to predict the uncertain information in advance to formulate the optimal control strategies that are more in line with the demand of the power grid and occupants. Model predictive control (MPC) strategies are used to cope with control prob-

Manuscript received: April 28, 2022; revised: July 5, 2022; accepted: September 22, 2022. Date of CrossCheck: September 22, 2022. Date of online publication: November 28, 2022.

This work was supported in part by the Fundamental Research Funds for the Central Universities (No. 2018JBZ004) and the National Natural Science Foundation of China (No. 52007004).

M. Xia, F. Chen, Q. Chen (corresponding author), Y. Song, and T. Wang are with the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China (e-mail: mchxia@bjtu.edu.cn; 20117016@bjtu.edu.cn; chenqf@bjtu.edu.cn; 16121517@bjtu.edu.cn; 19121497@bjtu.edu.cn).

S. Liu is with the State Grid State Power Economic Research Institute, Beijing, China (e-mail: liuswmax@163.com).

DOI: 10.35833/MPCE.2022.000249

lems with complexity and multi-variability of the environment. In [14], a coordinated control strategy of HVACs and electric vehicles (EVs) based on the Markov decision process (MDP) method is proposed to reduce the cost and accommodate the uncertainties in the PV supply. The time-of-use (TOU) electricity price is used to calculate the users' economy, and uncertainties such as outdoor temperature are predicted through historical data [15]. Reference [16] proposes a collaborative optimization operation method for multiple smart buildings to reduce total energy costs. A rolling horizon scheme is proposed to reduce the amount of forecasting information and computation cost. Compared with [14], the forecasting information required for decision-making in [16] is significantly reduced, but the results of the model forecast will still directly affect the strategic decision, and the forecasting information is often affected by numerous uncertain factors. Thus, they shall be properly addressed during the operation of HVAC systems.

Probabilistic solutions are used to cope with the uncertainty in the scheduling process. Fuzzy control, stochastic programming, and robust optimization are commonly used to solve such uncertain problems. Reference [17] uses fuzzy sets to describe the uncertainties in retail electricity prices and temperature. In [18], a solar thermal supplemental multi-energy heating system controlled by fuzzy controller is proposed. Monte Carlo methods are often used to generate uncertain features [19]. References [20] and [21] propose an optimal scheduling model for microgrid based on stochastic programming method and use Monte Carlo method for scenario generation. The Monte Carlo method requires a large number of simulations to fit the convergent parameters, and the huge computational load hinders the development of this technology. In [22], a chance-constrained optimization model is proposed to solve the uncertainty of real-time electricity prices and load in home energy management systems. Reference [23] establishes a bi-level optimal scheduling model for the community integrated energy system. The chance-constrained model is converted into a mixed-integer linear programming model and solved by the CPLEX solver. In [24], a two-stage stochastic programming model is proposed for the uncertain renewable energy output, power load, and price in multi-energy microgrids. In [25], the electricity price is considered as a robust optimization technique with a certain degree of confidence. References [26] and [27] propose a robust scheduling method for integrated energy systems considering economics, where the uncertainties of outdoor temperature and the thermal comfort of occupants are taken into account. Nevertheless, both the fuzzy rules of fuzzy control and the probability distribution of random variables in stochastic programming require certain prior knowledge of uncertainty. Estimating the uncertain information with subjective experience leads to a lack of confidence in many practical situations. Robust optimization will obtain relatively conservative results, which may result in a waste of resources in many cases.

The uncertainty of electricity prices and environment is affected by many different factors, and the probability distribution of uncertain features must be obtained through a large number of complex calculations. It is difficult for the above model-based methods to guarantee a low computational complexity and a good performance at the same time. Model-free reinforcement learning (RL) is frequently employed to solve the decision-making problem of nonlinear systems. In [28], a Q-learning agent is enabled to control the HVAC system considering the energy consumption and temperature range. Reference [29] uses batch RL to realize residential DR of thermostatically controlled loads. However, the RL method is not suitable for the system with large state space, and the state update mechanism exhibits high computational cost [30]. Deep neural network uses powerful high-dimensional data feature extraction and complex mapping ability to approximate the value function, which can achieve the rapid environment perception. By combining the deep learning technology and RL technology, deep reinforcement learning (DRL) obtains significant success in many complex decision-making problems [31]. With specific respect to the HVAC scheduling problem, the scheduling frameworks based on deep Q network (DQN), advantage actor critic (A2C), and deep deterministic policy gradient (DDPG) are established for building energy [32]-[35]. Reference [32] proposes an HVAC airflow direction control method based on DQN, aiming at achieving uniform comfort of the indoor environment. Reference [33] designs an intelligent DQN agent to reduce energy usage and peak load. Reference [34] proposes an A2C method for HVAC system to minimize the energy consumption while maintaining the thermal comfort. In [35], the DDPG method is applied to optimize the continuous control of multi-zone HVAC. All the above research works have demonstrated the effectiveness of the DRL method in HVAC system. However, the building energy consumption is usually used as the cost function, and there is a lack of residential HVAC scheduling strategy adapted to time-varying electricity prices under the background of electricity marketization. Simultaneously, the occupants' comfort should always be taken into account.

Given this context, we propose an adaptive scheduling strategy for residential HVAC based on model-free DRL to cope with the dual uncertainty of time-varying electricity prices and ambient temperature. Specifically, DRL methods can be divided into value-based methods and policy-based methods, which are often used to solve control problems in discrete spaces and continuous spaces, respectively. Since the residential HVAC system is typically an on/off type unit [11], [36], the value-based dueling DQN algorithm is used to solve the above problem. The main contributions of this paper are summarized as follows.

1) The adaptive scheduling strategy for residential HVAC based on DRL including the definition of state, action, and reward is established. The state transition probability of the complex nonlinear system is learned by deep neural network, and the adaptive scheduling strategy is formulated. Deep neural network can deeply tap the historical electricity prices and the temperature information. The residential HVAC agent formulates the scheduling strategy according to the time-varying electricity prices, which can not only respond to the demand of the power grid, but also ensure the

good economy of occupants.

2) The LSPE method is used to obtain the thermal capacity and the thermal resistance parameters of the residence, which provides a definite basis for the state transition of the state dimension of HVAC operation temperature. In this paper, a large number of historical temperature measurements are used to estimate the heterogeneous thermal parameters of residence based on the HVAC state equation.

3) The proposed optimal scheduling strategy is validated on real datasets. The effectiveness of the proposed method in terms of economy and comfort is demonstrated by comparing with the MPC method, the uncontrolled methods, and other DRL methods.

II. Residence Thermal Parameter Estimation Based on LSPE

A. HVAC State Equation

The indoor temperature is an important state variable to characterize the dynamic operation characteristics of HVAC. For simplicity, the first-order thermal transfer function is utilized to model the dynamic indoor temperature of a building [37] - [40]. The residential HVAC in cooling mode can be modeled as:

$$T_{i,t+1}=a_iT_{i,t}+b_iT_{out,t}+g_iu_{i,t} \tag{1}$$

$$\begin{cases} a_i=1-\dfrac{1}{C_iR_i} \\ b_i=\dfrac{1}{C_iR_i} \\ g_i=-\dfrac{\eta P_{rated}}{C_i} \end{cases} \tag{2}$$

where $T_{i,t+1}$ and $T_{i,t}$ are the indoor temperatures of the i^{th} residence at time steps $t+1$ and t, respectively; $T_{out,t}$ is the ambient temperature at time step t; $u_{i,t}$ is the binary variable representing the HVAC on/off status of the i^{th} residence at time step t; a_i, b_i, and g_i are the coefficients of thermal function of the i^{th} residence; R_i and C_i are the thermal resistance and capacitance of the i^{th} residence, respectively; η is the cooling efficiency of the i^{th} residence; and P_{rated} is the rated power.

B. Thermal Parameter Estimation Based on LSPE

During the residential HVAC operation, the indoor temperature needs to be controlled within a certain range according to the comfort requirements of occupants, i.e., $T_{in}^{low}<T_{i,t+1}<T_{in}^{up}$. The thermal parameters of the residence affect the rate of change of indoor temperature. However, it is difficult to quantify the indoor building structure, facility layout, and occupants' behavior, and the thermal parameters cannot be measured directly. The LSPE is proposed to obtain the thermal resistance parameters and the thermal capacity parameters. The measurement equation of LSPE is expressed as:

$$\boldsymbol{Y}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{e} \tag{3}$$

$$\boldsymbol{J}=\sum_{t=1}^{T}\boldsymbol{e}_t^2=(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{x}) \tag{4}$$

$$\boldsymbol{x}=\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y} \tag{5}$$

where $\boldsymbol{x}=\left[a_i,b_i,g_i\right]$; $\boldsymbol{A}=\left[\boldsymbol{T}_i,\boldsymbol{T}_{out},\boldsymbol{u}_i\right]^{\mathrm{T}}$, $\boldsymbol{T}_i=\left[T_{i,1},\ldots,T_{i,t},\ldots,T_{i,T}\right]$, $\boldsymbol{T}_{out}=\left[T_{out,1},\ldots,T_{out,t},\ldots,T_{out,T}\right]$, $\boldsymbol{u}_i=\left[u_{i,1},\ldots,u_{i,t},\ldots,u_{i,T}\right]$; and $\boldsymbol{Y}=\left[T_{i,2},\ldots,T_{i,t+1},\ldots T_{i,T+1}\right]$.

Equation (4) can be obtained by minimizing the total square sum of errors of multiple groups of measurement results $\boldsymbol{e}_t$ as the objective function, and (5) can be obtained by solving (4).

The coefficients of thermal transfer function can be solved by substituting the measured data into (5), and then the thermal resistance and thermal capacity can be solved, which can provide a basis for the state transition probability matrix of MDP.

III. Adaptive Scheduling Strategy Based on Dueling DQN

A. MDP Problem Description

In this paper, an MDP with discrete time step is applied to formulate the real-time scheduling strategy for HVAC against the randomness of time-varying electricity prices and ambient temperature. The electricity prices fluctuate over time, which is different from fixed prices or TOU prices. The interval of price change is one hour. At time step t, we can observe the real-time prices, the indoor temperature, and the ambient temperature, and then we choose the switch action of HVAC. After the action is executed, it can be observed from (1) that the indoor temperature will change accordingly, which will affect the state at the next moment. The new system state can be observed and the new action can be chosen at time step $t+1$. In particular, the thermal parameters of the residence are used in the calculation of state transition. The method takes one day as the cycle to formulate the scheduling strategy of the residential HVAC. Its complexity is that the time-varying electricity prices and the ambient temperature state are difficult to be accurately predicted by the model-based method, and the policy decision is affected. DRL can solve this problem well.

B. DRL Method

The model-free DRL method is often used to solve the policy decision problems of nonlinear complex systems. Based on a large number of historical data, the DRL method can capture the law of real-time prices and ambient temperature fluctuation well to guide the agent to make decision.

1) State: the state parameter of the problem consists of four parts, i.e., the time step of the day t, indoor temperature, ambient temperature, and electricity prices P_t. The uncertainty of the state variables mainly comes from ambient temperature and electricity prices. The ambient temperature and electricity prices in the past 24 hours are used as the state input for training in the reinforcement learning. Therefore, the state vector at time step t is $\boldsymbol{s}_t=\left[t,T_{i,t},T_{out,t},T_{out,t-1},\ldots,T_{out,t-23},P_t,P_{t-1},\ldots,P_{t-23}\right]$, which contains 50 dimensions. In particular, it is necessary to maintain the indoor temperature $T_{i,t}$ within the range set by occupants,

i.e., $T_{i,t}^{low} \le T_{i,t} \le T_{i,t}^{up}$.

2) Action: the action a_t represents the switch state of residential HVAC u_t at time step t. When the indoor temperature is close to the upper limit and the electricity price is low, the residential HVAC agent automatically chooses the "on" status. When the indoor temperature is close to the lower limit and the electricity price is high, the HVAC agent automatically chooses the "off" status.

3) Reward and return: the reward r_t consists of two parts, i.e., power purchase cost and boundary crossing penalty. To ensure less cost, the reward r_t is defined as:

$$r_t = -l_1 E_{CO} - l_2 C_{OF} \tag{6}$$

where $E_{CO} = P_t u_t$ is the power purchase cost; $C_{OF} = \max\left(0, T_{i,t} - T_{i,t}^{up}\right) + \max\left(0, T_{i,t}^{low} - T_{i,t}\right)$ is the boundary crossing cost of indoor temperature; and l_1 and l_2 are the weights of economy and comfort, respectively. Let $l = l_1/l_2$, r_t can be expressed as:

$$r_t = -lP_t u_t - \left[\max\left(0, T_{i,t} - T_{i,t}^{up}\right) + \max\left(0, T_{i,t}^{low} - T_{i,t}\right)\right] \tag{7}$$

The reward decreases as the electricity price rises. When the indoor temperature crosses the bounds, the agent will make the corresponding punishment. The more boundary is crossed, the greater the punishment will be. The parameter l is the coefficient to balance occupants' demands for comfort and economy.

The return U_t is defined as the cumulative discounted rewards:

$$U_t = r_t + \gamma r_{t+1} + \ldots + \gamma^{T-t} r_{i,T} \tag{8}$$

where $\gamma \in [0,1]$ is the discount factor that is used to trade off the importance between immediate and future rewards.

4) State transition: $P_{ss'}^{a}$ is the probability that the state $\boldsymbol{s}_t$ will transfer to the state $\boldsymbol{s}_{t+1}$ after taking the action a_t. We select the state function as the time step of the day, the indoor temperature, the historical ambient temperature, and the historical electricity prices, where the state transition is influenced by the uncertain factors. Specifically, the state transition for indoor temperature $T_{i,t}$ is controlled by action and it can be expressed by the deterministic formula (1). The state transition of ambient temperature and electricity prices is random. It is difficult for model-based method to find an accurate probability distribution to describe the state transition. To solve this problem, an improved reinforcement learning method is proposed to learn the state transition, as shown in Section III-C.

C. Dueling DQN

Dueling DQN is an improved DQN method. The advantage function is defined to evaluate the performance of the action a_t in the current state $\boldsymbol{s}_t$. The optimal advantage function can guide the agent to make the action, and then solve the policy decision problem of nonlinear complex system. The optimal advantage function is defined as:

$$A^*(\boldsymbol{s}, a) = Q^*(\boldsymbol{s}, a) - V^*(\boldsymbol{s}) \tag{9}$$

where $Q^*(\boldsymbol{s}, a)$ is the optimal action value function, which can evaluate the quality of the action a taken in the state $\boldsymbol{s}$; and $V^*(\boldsymbol{s})$ is the optimal state value function, which can evaluate the quality of the state $\boldsymbol{s}$. The optimal state value function is regarded as the baseline. Then, the optimal advantage function $A^*(\boldsymbol{s}, a)$ is the advantage of action a over baseline.

Neural networks are used to approximate $A^*(\boldsymbol{s}, a)$, $V^*(\boldsymbol{s})$, and $Q^*(\boldsymbol{s}, a)$, and the following formula can be obtained:

$$Q\left(\boldsymbol{s}, a; w^A, w^V\right) = V\left(\boldsymbol{s}; w^V\right) + A\left(\boldsymbol{s}, a; w^A\right) \tag{10}$$

where $Q\left(\boldsymbol{s}, a; w^A, w^V\right)$ is called the dueling network; w^V denotes the network parameters of $V\left(\boldsymbol{s}; w^V\right)$; and w^A denotes the network parameters of $A\left(\boldsymbol{s}, a; w^A\right)$.

However, the state value and advantage value cannot be uniquely estimated by learning Q function. To solve this problem, [41] improves the stability by adding the advantage mean as baseline to the estimation.

$$Q\left(\boldsymbol{s}, a; w^A, w^V\right) = V\left(\boldsymbol{s}; w^V\right) + A\left(\boldsymbol{s}, a; w^A\right) - \underset{a}{mean}\, A\left(\boldsymbol{s}, a; w^A\right) \tag{11}$$

where $\underset{a}{mean}\, A(\boldsymbol{s}, a; w^A)$ is the average of the advantage function. The dueling DQN architecture is shown in the Fig. 1. The neural network is used to extract features of the current state, and the features are mapped to the advantage value and state value, respectively. Therefore, parameters w^V and w^A overlap partially.

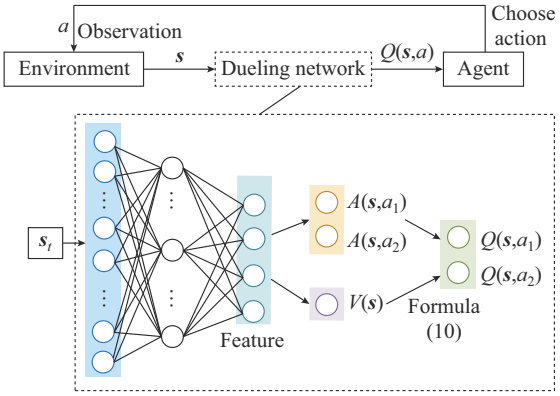

Fig. 1. Dueling DQN architecture.

D. Policy Update of Dueling Network

The temporal-difference (TD) [42] is the core prediction method for the policy update of DRL:

$$y_t = r_t + \gamma \max_a Q\left(\boldsymbol{s}_{t+1}, a; \boldsymbol{w}\right) \tag{12}$$

$$\delta_t = Q\left(\boldsymbol{s}_t, a_t; \boldsymbol{w}\right) - y_t \tag{13}$$

where y_t is called the TD target, which includes the reward at time step t and the maximum estimation of Q function at time step $t+1$; δ_t is called the TD error; $Q\left(\boldsymbol{s}_t, a_t; \boldsymbol{w}\right)$ is the estimation of Q function at time step t; and $\boldsymbol{w} = \left[w^A, w^V\right]$ denotes the parameters of dueling network. The TD algorithm uses subsequent Q function estimation to update the current Q function estimation, which is a bootstrapping method [43]. Several techniques for improving the algorithm convergence speed and the training stability are introduced.

1) Fixed Q Target

To prevent the overestimation of Q value, a target network is used to calculate the TD target value [44].

$$\hat{y}_t = r_t + \gamma \max_a \hat{Q}\left(\boldsymbol{s}_{t+1}, a; \hat{\boldsymbol{w}}\right) \tag{14}$$

where $\hat{Q}\left(\boldsymbol{s}, a; \hat{\boldsymbol{w}}\right)$ is called the target network, which has the same network structure as $Q\left(\boldsymbol{s}, a; \boldsymbol{w}\right)$. During the training process, the parameters $\boldsymbol{w}$ are updated every step, while the parameters $\hat{\boldsymbol{w}}$ of target network are updated every C steps.

2) Experience Replay

At time step t, recent $\mathcal{N}$ transitions $\left(\boldsymbol{s}_i, a_i, r_i, \boldsymbol{s}_{i+1}\right)$ can be stored in a replay buffer $\mathcal{N}$ to prevent the waste of experience. At each step of training, M transitions are chosen from replay buffer $\mathcal{N}$ to form a mini-batch D. The parameters $\boldsymbol{w}$ are updated by gradient descent based on D to minimize the loss function [45]:

$$L(\boldsymbol{w}) = \frac{1}{M} \sum_{\left(\boldsymbol{s}_i, a_i, r_i, \boldsymbol{s}_{i+1}\right) \in D} \left(\hat{y}_i - Q\left(\boldsymbol{s}_i, a_i; \boldsymbol{w}\right)\right)^2 \tag{15}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - LR \cdot \nabla_{\boldsymbol{w}} L(\boldsymbol{w}) \tag{16}$$

where $L(\boldsymbol{w})$ is the loss function; and LR is the learning rate. Randomly chosen D for updating parameters can cut off the correlation between experiences and ensure better performance.

3) Decayed-ε-greedy

The greedy method tends to make the exploration problem fall into a local optimal solution. The parameter ε is used to balance the exploitation and exploration. In the training process of this paper, ε gradually decays as the number of iterations increases. At the early stage of the iteration, we encourage the exploration and focus on greedy at the later stage to ensure stable convergence of the algorithm.

$$\varepsilon = \max\left(1 - \frac{n_{epo}}{N_{epo}}, \varepsilon_{\min}\right) \tag{17}$$

where $\varepsilon_{\min}$ is the minimum exploration rate; $n_{epo} = Epo \cdot \left(t - t_{ori}\right)$ is the number of current experiences, and Epo is the number of current iterations; and $N_{epo} = Epo_{\max} \cdot \left(t_{end} - t_{ori}\right)$ is the number of total experiences, $Epo_{\max}$ is the number of total iterations, t_{ori} is the initial time step, and t_{end} is the end time step.

The performance of dueling DQN will be greatly improved by the above techniques. The policy update process of dueling DQN is shown in Algorithm 1, and the specific implementation process of optimal HVAC scheduling using dueling DQN is shown in Fig. 2.

Algorithm 1: policy update process of dueling DQN

Initialize network $Q\left(\boldsymbol{s}, a; \boldsymbol{w}\right)$ with random parameters $\boldsymbol{w}$
Initialize target network $\hat{Q}\left(\boldsymbol{s}, a; \hat{\boldsymbol{w}}\right)$ with random parameters $\hat{\boldsymbol{w}} \leftarrow \boldsymbol{w}$
for Epo=1:$Epo_{\max}$ **do**
  Obtain the initial state $\boldsymbol{s}_{t_{ori}}$
  for $t = t_{ori}$:t_{end} **do**
    Update ε according to (17)
    Select action a_t based on decayed-ε-greedy search
    Execute a_t, observe reward r_t, update environment, and observe new state $\boldsymbol{s}_{t+1}$
    Store transition $\left(\boldsymbol{s}_t, a_t, r_t, \boldsymbol{s}_{t+1}\right)$ in replay buffer
    Randomly choose M transitions from $\mathcal{N}$ to form a mini-batch D
    Calculate target value $\hat{y}_t$ according to (14)
    Calculate loss function $L(\boldsymbol{w})$ according to (15)
    Update parameters $\boldsymbol{w}$ according to (16)
    Reset $\hat{\boldsymbol{w}} \leftarrow \boldsymbol{w}$ every C steps to update target network
  end for
end for

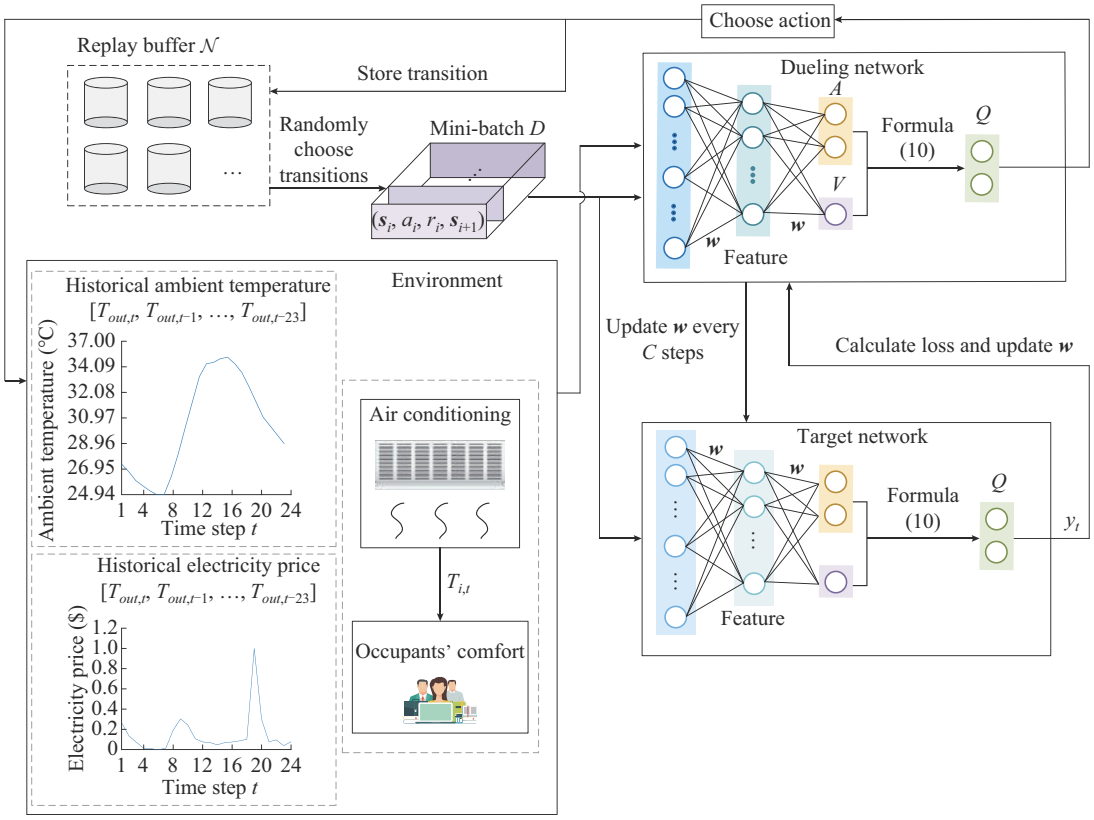

Fig. 2. Implementation process of optimal scheduling for HVAC using dueling DQN.

IV. Case Studies

In this section, the proposed method is tested on the residential HVAC operations for a single residence. Simulations are performed on a laptop with AMD Ryzen 7 4800H 2.90 GHz CPU, and 16 GB RAM. The coding work is carried out in Python 3.6 with PyTorch 1.10.1 and MATLAB 2018b.

A. Experiment Setup

The performance of the proposed method is evaluated in the following scenario. We collect hourly electricity prices from Australia Energy Market Operator (AEMO). The hourly ambient temperature and indoor temperature are acquired from TRNSYS 18.0. The electricity prices and ambient temperature in seven days are shown in Fig. 3. For the state transition model of indoor temperature, we set P_{rated}=5.6 kW, $\eta = 0.9$, and then R and C can be evaluated according to (5). Moreover, for the thermal comfort constraint of the occupants, we set the temperature range as $T_i^{up} = 26$ ℃ and $T_i^{low} = 22$ ℃. We collect the data of July and August in 2015 and 2016 as the train data and the data of July and August in 2017 as the test data to evaluate the improvement by the proposed method. The simulation cycle is set to be 09: 00 to 21:00.

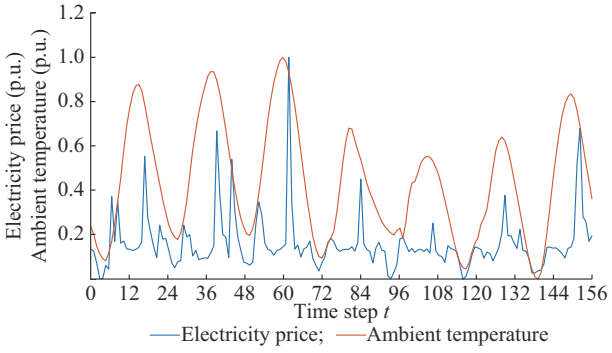

Fig. 3. Electricity prices and ambient temperature in seven days of normalized historical data.

As shown in Fig. 2, the dueling DQN is proposed to formulate the optimal scheduling strategy of residential HVAC. The network structure of dueling DQN can be divided into two stages: feature extraction and mapping. The feature extraction stage of dueling network includes an input layer with 50 neurons and 3 hidden layers with 256, 128, and 64 neurons, respectively. At the mapping stage, the features can be mapped to 2 neurons (advantage value) and 1 neuron (state value), respectively, and then the output layer with 2 neurons (Q value) can be calculated. And we use the rectified linear unit (ReLU) as the activation function to each layer. The other hyperparameters during training are as follows: the capacity of replay buffer $\mathcal{N}$ is 1×10^4, the batch size of the sampled transition M for training is 32, and the update frequency of target network is 1000. The Adam optimizer is used for learning the neural network with the learning rate $LR = 1 \times 10^{-5}$, the discount factor $\gamma = 0.99$, the minimum exploration rate $\varepsilon_{\min} = 0.01$, and the maximum number of iterations is 2000.

B. Performance Evaluation

To show the effectiveness of the proposed method, the optimal scheduling strategy for residential HVAC based on dueling DQN is compared with several benchmark algorithms.

1) Without considering the fluctuation of electricity prices, the residential HVAC can be controlled automatically according to the comfort temperature threshold set by occupants (labeled as "Uncontrolled"). In the "Uncontrolled" mode, the initial state of HVAC is assumed as off-state. As the ambient temperature rises, the indoor temperature rises accordingly. When the indoor temperature crosses the temperature upper limit, the residential HVAC stays on-state until the indoor temperature drops to the temperature lower limit, and then it enters the off-state again.

2) The optimal scheduling is formulated based on the predicted time-varying electricity prices and ambient temperature (labeled as "MPC"). The real electricity prices and ambient temperature data are randomly added to the bias within ±10% as the result of the model prediction, and the optimal scheduling strategy is formulated based on the result.

3) The future time-varying electricity prices and ambient temperature are assumed to be known and the optimal scheduling strategy is formulated on the real dataset (labeled as "Optimal"). This method is only used for the comparative case, which does not exist in the real world.

4) The method proposed in this paper is labeled as "Dueling DQN". Moreover, two other DRL methods are chosen for comparative case to enhance persuasion.

5) The classical DQN algorithm based on value function is labeled as "DQN".

6) The proximal policy optimization (PPO) based on policy function is labeled as "PPO". The neural network structure and parameter settings of DQN and PPO algorithms are the same as the proposed method.

The evolution of the accumulative reward during training process of dueling DQN, classic DQN, and PPO algorithms is shown in Fig. 4. As the number of iterations increases, the cumulative reward increases quickly and tends to converge. The dueling DQN and classic DQN algorithms converge after 400 iterations, and the PPO algorithm converges after 1000 iterations. The convergence time of the three algorithms is 96 min, 85 min, and 181 min, respectively. It can be observed from Fig. 4 that the PPO algorithm based on the AC framework has the slowest convergence speed and reward, which may have more advantages in coping with continuous control problems. The value-based dueling DQN and classic DQN algorithms have similar convergence time, while the dueling DQN algorithm has higher reward.

We test the performance of the benchmark algorithm by comparing the electricity cost on different dates in the test set. The data of July and August in 2017 are selected as the test set, since the policy proposed in this paper needs to use the electricity prices of the previous day in the benchmark algorithm. The electricity cost from July 2 to August 31 is selected as the comparative objects. Figure 5 reflects the daily electricity cost in the test set. It can be observed that the daily electricity cost is well controlled under the optimal scheduling of the proposed method. The accumulative electricity cost of the 61 test days is shown in Fig. 6. As labeled in this figure, the accu-

mulative costs of the "Uncontrolled", "MPC", "Optimal", "Dueling DQN", "DQN", and "PPO" algorithms are $1195.91, $991.18, $819.14, $836.10, $865.62, and $914.75, respectively. The proposed "Dueling DQN", "DQN", "PPO", and MPC algorithms learn the trends of time-varying electricity prices and ambient temperature, and then formulate the scheduling strategy, which saves 30.09%, 27.61%, 23.51%, and 17.12% of electricity costs, respectively.

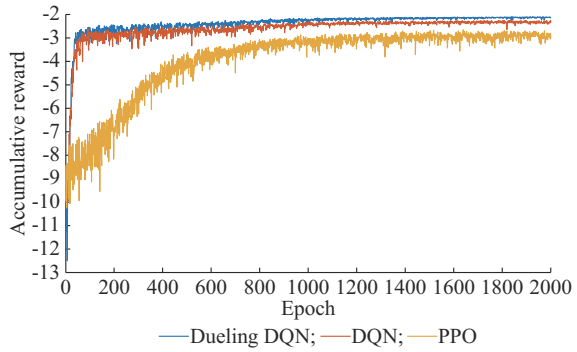

Fig. 4. Accumulative reward during training process.

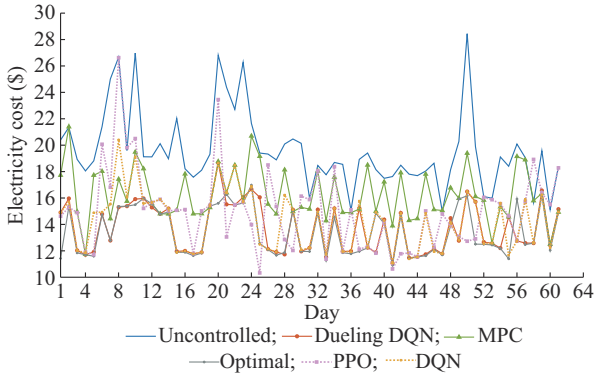

Fig. 5. Daily electricity cost.

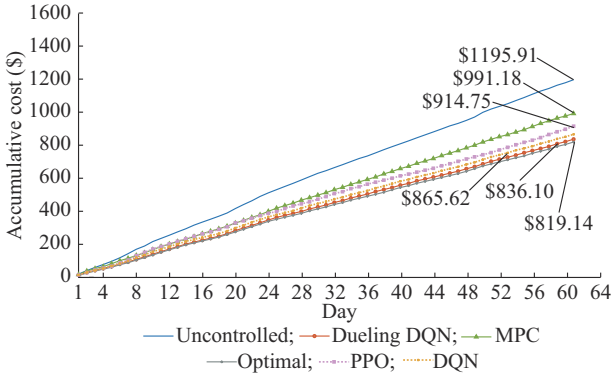

Fig. 6. Accumulative electricity cost.

Figure 7 shows the detailed response of scheduling based on the proposed method in the seven days of test set. The gray area is not the concerned time period of this paper. Figure 7(a) shows the normalized time-varying electricity prices and the switching status of residential HVAC. It can be observed that the proposed method can effectively avoid turning on the residential HVAC during the peak time of electricity prices, which saves more electricity costs for occupants. Figure 7(b) shows the response of indoor temperature, which is basically controlled within the set point range.

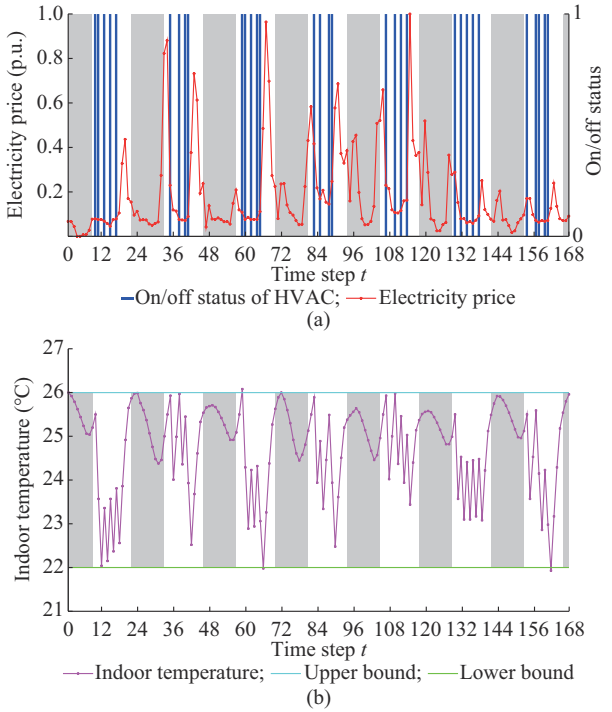

Fig. 7. Detailed response of scheduling based on proposed method. (a) Optimal scheduling results. (b) Indoor temperature response.

The average electricity cost on the test day is shown in Table I, which can be calculated by $E_{pr}=\frac{1}{N}\sum_{t=10}^{21}\frac{1}{4}u_t P_{rated} P_t$. Except for the ideal "Optimal" algorithm, the proposed method has the lowest average electricity cost. However, the electricity cost is not the only criterion to measure the occupants' satisfaction, the comfort is also an important factor. The number of violations is also compared in Table I. It can be observed that the number of violations of the proposed "Dueling DQN" algorithm is the least among the benchmark algorithms. The "Uncontrolled" algorithm is a disadvantaged method, whose switch state changes only when the indoor temperature crosses the boundary of the upper or the lower temperature limit. Specifically, the violation degree is shown in Fig. 8.

TABLE I
AVERAGE ELECTRICITY COST AND TOTAL NUMBER OF VIOLATIONS ON TEST DAY

Algorithm	Average electricity cost ($)	Total number of violations
Uncontrolled	19.61	183
MPC	16.24	100
Optimal	13.42	0
Dueling DQN	13.71	64
DQN	14.19	77
PPO	15.00	93

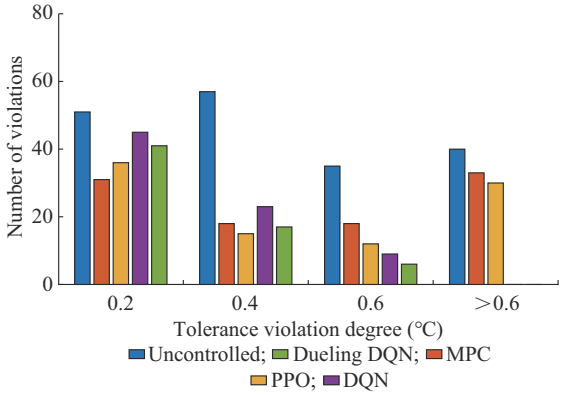

Fig. 8. Violation degree in test set.

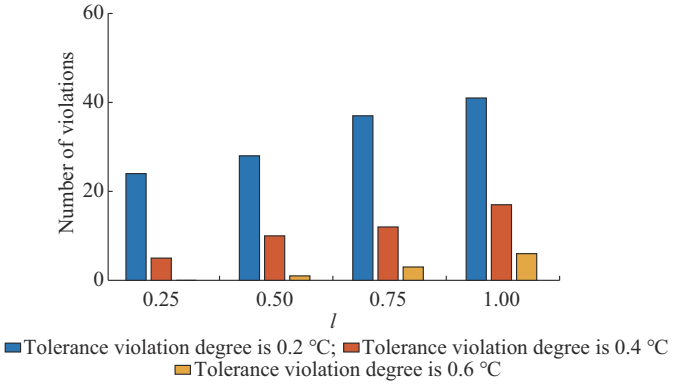

Fig. 9. Violations with different l.

The abscissa in Fig. 8 is the violation degree, which represents the degree of indoor temperature crossing the boundary. It can be observed in Fig. 8 that the "Uncontrolled", "MPC", and "PPO" algorithms are relatively evenly distributed in the four intervals, which still have some distributions in the interval, where the violation exceeds 0.6 ℃. The violation degree distribution of the proposed method and the "DQN" algorithm is similar, while the violations always keep within 0.6 ℃ and most violations are distributed in $[0, 0.2]$℃. Obviously, the proposed method performs better than the "DQN" algorithm. In summary, the proposed method has advantages in both the number of violations and the violation degree. The occupant violation sensitivity of the proposed method is detailed as follows.

C. Violation Sensitivity

The violation sensitivity refers to the occupants' tolerance to the indoor temperature crossing the boundary. We believe that the occupants' satisfaction consists of two aspects: economy and comfort, but it is often difficult to guarantee both at the same time. If the occupants are highly sensitive to the comfort, the agent needs to ensure the indoor temperature limit firstly and ignore some economic requirements to meet the occupants' comfort requirements. On the contrary, if occupants are highly sensitive to the economic, the agent gives priority to maintain "off" status at high electricity prices. And the comfort requirements will be ignored sometimes. The coefficient l in (7) can be used as an adjustable parameter to balance the proportion of occupants' demand.

The violations with different coefficients l are analyzed in Fig. 9. The accumulative costs and total number of violations with different coefficients l are shown in Table II.

TABLE II
PERFORMANCE COMPARISON IN DIFFERENT l

l	Accumulative cost ($)	Total number of violations
0.25	926.06	29
0.50	884.62	39
0.75	864.31	52
1.00	836.10	64

Combined with Fig. 9 and Table II, it can be observed that a larger l will result in a larger number and a greater degree of violations, which means better economy. If occupants are more interested in economy, a higher coefficient l can be set. And if they pay more attention to comfort, a lower coefficient l can be set.

V. CONCLUSION

In this paper, we introduce an adaptive scheduling strategy of residential HVAC based on dueling DQN, which can make the optimal scheduling strategy of HVAC according to the time-varying electricity prices and the uncertain ambient temperature. By establishing the reward mechanism according to the real-time prices and the indoor state information, the residential HVAC agent can learn the trend of the electricity prices and the temperature changes through the historical data, then make optimal decisions. In this process, the residence thermal parameters obtained by LSPE method provide the basis for the temperature state transition matrix. The data of July and August in 2017 are taken as the test set to analyze the performance of the proposed method on electricity cost and tolerance violations. Compared with the "Uncontrolled", "MPC", "Optimal", "DQN" and "PPO" algorithms, the advantages of the proposed method are verified. More advanced artificial intelligence methods can also be applied within the same framework in the future.

REFERENCES

[1] P. Samadi, V. W. S. Wong, and R. Schober, "Load scheduling and power trading in systems with high penetration of renewable energy resources," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 1802-1812, Jul. 2016.

[2] X. Zhu, M. Xia, and H. Chiang, "Coordinated sectional droop charging control for EV aggregator enhancing frequency stability of microgrid with high penetration of renewable energy sources," *Applied Energy*, vol. 210, pp. 936-943, Jan. 2018.

[3] S. Lin, F. Li, E. Tian *et al.*, "Clustering load profiles for demand response applications," *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1599-1607, Mar. 2019.

[4] C. Si, S. Xu, C. Wan *et al.*, "Electric load clustering in smart grid: methodologies, applications, and future trends," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 2, pp. 237-252, Mar. 2021.

[5] Y. Huang, Z. Lin, X. Liu *et al.*, "Bi-level coordinated planning of active distribution network considering demand response resources and severely restricted scenarios," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 5, pp. 1088-1100, Sept. 2021.

[6] R. Deng, Z. Yang, M.-Y. Chow *et al*., "A survey on demand response in smart grids: mathematical models and approaches," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 3, pp. 570-582, Jun. 2015.
[7] M. Xia, Y. Song, and Q. Chen, "Hierarchical control of thermostatically controlled loads oriented smart buildings," *Applied Energy*, vol. 254, p. 113493, Nov. 2019.
[8] E. Loukarakis, C. J. Dent, and J. W. Bialek, "Decentralized multi-period economic dispatch for real-time flexible demand management," *IEEE Transactions on Power Systems*, vol. 31, no. 1, pp. 672-684, Jan. 2016.
[9] R. Mowris and E. Jones, "Peak demand and energy savings from properly sized and matched air conditioners," *ACEEE Summer Study on Energy Efficiency in Buildings*, no. 2008, pp. 196-208, Jan. 2008.
[10] L. Zhang, N. Good, and P. Mancarella, "Building-to-grid flexibility: modelling and assessment metrics for residential demand response from heat pump aggregations," *Applied Energy*, vol. 233-234, pp. 709-723, Jan. 2019.
[11] H. Liu, H. Xie, H. Luo *et al*., "Optimal strategy for participation of commercial HVAC systems in frequency regulation," *IEEE Internet of Things Journal*, vol. 8, no. 23, pp. 17100-17110, Dec. 2021.
[12] N. Lu, "An evaluation of the HVAC load potential for providing load balancing service," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1263-1270, Sept. 2012.
[13] A.-Y. Yoon, Y.-J. Kim, and S.-I. Moon, "Optimal retail pricing for demand response of HVAC systems in commercial buildings considering distribution network voltages," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5492-5505, Sept. 2019.
[14] H. Zhao, Z. Xu, J. Wu *et al*., "Optimal coordination of EVs and HVAC systems with uncertain renewable supply," in *Proceedings of 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, Vancouver, Canada, Aug. 2019, pp. 733-738.
[15] O. Erdinc, A. Tascikaraoglu, N. G. Paterakis *et al*., "End-user comfort oriented day-ahead planning for responsive residential HVAC demand aggregation considering weather forecasts," *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 362-372, Jan. 2017.
[16] J. A. Pinzon, P. P. Vergara, L. C. P. da Silva *et al*., "Optimal management of energy consumption and comfort for smart buildings operating in a microgrid," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3236-3247, May 2019.
[17] D. T. Nguyen and L. B. Le, "Optimal bidding strategy for microgrids considering renewable energy and building thermal dynamics," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1608-1620, Jul. 2014.
[18] R. Ma, J. Li, F. Jin *et al*., "Fuzzy theory-based energy-saving control of solar thermal supplemental multi-energy heating system," *Energy Reports*, vol. 8, pp. 636-646, Sept. 2022.
[19] R. C. Rodrigues, "Modeling urban traffic noise dependence on energy, assisted with Monte Carlo simulation," *Energy Reports*, vol. 8, pp. 583-588, Jun. 2022.
[20] E. Grover-Silva, M. Heleno, S. Mashayekh *et al*., "A stochastic optimal power flow for scheduling flexible resources in microgrids operation," *Applied Energy*, vol. 229, pp. 201-208, Nov. 2018.
[21] H. Shuai, J. Fang, X. Ai *et al*., "Stochastic optimization of economic dispatch for microgrid based on approximate dynamic programming," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2440-2452, May 2019.
[22] Y. Huang, L. Wang, W. Guo *et al*., "Chance constrained optimization in a home energy management system," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 252-260, Jan. 2018.
[23] Y. Li, M. Han, Z. Yang *et al*., "Coordinating flexible demand response and renewable uncertainties for scheduling of community integrated energy systems with an electric vehicle charging station: a bi-level approach," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 4, pp. 2321-2331, Oct. 2021.
[24] Z. Li and Y. Xu, "Temporally-coordinated optimal operation of a multi-energy microgrid under diverse uncertainties," *Applied Energy*, vol. 240, pp. 719-729, Apr. 2019.
[25] Z. Wu, S. Zhou, J. Li *et al*., "Real-time scheduling of residential appliances via conditional risk-at-value," *IEEE Transactions on Smart Grid*, vol. 5, no. 3, pp. 1282-1291, May 2014.
[26] S. Lu, W. Gu, S. Zhou *et al*., "Adaptive robust dispatch of integrated energy system considering uncertainties of electricity and outdoor temperature," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 7, pp. 4691-4702, Jul. 2020.
[27] S. Lu, W. Gu, K. Meng *et al*., "Economic dispatch of integrated energy systems with robust thermal comfort management," *IEEE Transactions on Sustainable Energy*, vol. 12, no. 1, pp. 222-233, Jan. 2021.
[28] E. Barrett and S. P. Linder, "Autonomous HVAC control, a reinforcement learning approach," *Lecture Notes in Computer Science*, vol. 2015, pp. 3-19, Jan. 2015.
[29] F. Ruelens, B. J. Claessens, S. Vandael *et al*., "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2149-2159, Sept. 2017.
[30] Y. R. Yoon and H. J. Moon, "Performance based thermal comfort control (PTCC) using deep reinforcement learning for space cooling," *Energy and Buildings*, vol. 203, p. 109420, Nov. 2019.
[31] C. Guo, X. Wang, Y. Zheng *et al*., "Optimal energy management of multi-microgrids connected to distribution system based on deep reinforcement learning," *International Journal of Electrical Power & Energy Systems*, vol. 131, p. 107048, Oct. 2021.
[32] Y. Sakuma and H. Nishi, "Airflow direction control of air conditioners using deep reinforcement learning," in *Proceedings of 2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, Seville, Spain, May 2020, pp. 61-68.
[33] A. Mathew, A. Roy, and J. Mathew, "Intelligent residential energy management system using deep reinforcement learning," *IEEE Systems Journal*, vol. 14, pp. 5362-5372, Jun. 2020.
[34] Y. Wang, K. Velswamy, and B. Huang, "A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems," *Processes*, vol. 5, pp. 46-63, Sept. 2017.
[35] Y. Du, H. Zandi, O. Kotevska *et al*., "Intelligent multi-zone residential HVAC control strategy based on deep reinforcement learning," *Applied Energy*, vol. 281, p. 116117, Jan. 2021.
[36] X. Wu, J. He, Y. Xu *et al*., "Hierarchical control of residential HVAC units for primary frequency regulation," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3844-3856, Oct. 2018.
[37] D. S. Callaway, "Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy," *Energy Conversion and Management*, vol. 50, pp. 1389-1400, May 2009.
[38] Q. Shi, C. Chen, A. Mammoli *et al*., "Estimating the profile of incentive-based demand response (IBDR) by integrating technical models and social-behavioral factors," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 171-183, Jan. 2020.
[39] X. Wang, F. Li, J. Dong *et al*., "Tri-level scheduling model considering residential demand flexibility of aggregated HVACs and EVs under distribution LMP," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 3990-4002, Sept. 2021.
[40] X. Chen, Q. Hu, Q. Shi *et al*., "Residential HVAC aggregation based on risk-averse multi-armed bandit learning for secondary frequency regulation," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1160-1167, Nov. 2020.
[41] H. Qiu and F. Liu, "A state representation dueling network for deep reinforcement learning," in *Proceedings of 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, Baltimore, USA, Nov. 2020, pp. 669-674.
[42] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3, pp. 9-44, Feb. 1988.
[43] R. S. Sutton and A. G. Barto, "Reinforcement learning: an introduction," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 1054-1054, Sept. 1998.
[44] M. Volodymyr, K. Koray, S. David *et al*., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529-533, Feb. 2015.
[45] B. Hallam, D. Floreano, J. Meyer *et al*., "Evolving integrated controllers for autonomous learning robots using dynamic neural networks," in *Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*. Cambridge: MIT Press, 2002.

Mingchao Xia received the B.S. and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1998 and in 2003, respectively. He is currently a Professor with the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China. His current research interests include smart power distribution system control and optimization, power electronics in power distribution, and flexible load control.

Fangjian Chen received the B.S. degree from the Shandong Agricultural University, Taian, China, in 2018. She is currently pursuing the Ph.D. degree with the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China. Her research interests include deep learning, deep reinforcement learning, demand response, and flexible load control.

Qifang Chen received the B.S. and M.S. degrees in communication engineering an electric engineering from Xiangtan University, Xiangtan, China, in 2010 and 2013, respectively, and the Ph.D. degree in electrical and electronic engineering, North China Electric Power University, Beijing, China, in 2017. He is currently an Associate Professor with the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China. His research interests include demand side management, integrated energy system, and vehicle to grid (V2G).

Siwei Liu received the B.S. and Ph.D degrees in power system and automation from North China Electric Power University, Beijing, China, in 2010 and 2016, respectively. He is currently the Deputy Chief of Planning Division II in the Transmission Power Grid Planning Center, State Grid Economic and Technological Research Institute, Co., Ltd., Beijing, China. He also serves as Deputy Secretary in IEEE PES (China) Energy Storage Planning and Marketing Committee. His research interests include power grid planning, renewable energy integration and accommodation, and energy storage planning.

Yuguang Song received the B.S. degree from the Henan Polytechnic University, Jiaozuo, China, in 2016, and the M.S. degree from the School of Electrical Engineering, Beijing Jiaotong University, Beijing, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering, Beijing Jiaotong University. His research interests include demand response, smart grid, and load modeling.

Te Wang received the B.S. degree in electrical engineering from China University of Mining & Technology, Beijing, China, in 2019. He is currently working toward the M.S. degree in electrical engineering at Beijing Jiaotong University, Beijing, China. His research interests include demand side energy management and vehicle to grid (V2G).