

Deep Reinforcement Learning Based Bi-layer Optimal Scheduling for Microgrids Considering Flexible Load Control

Zitong Zhang, Jing Shi, *Senior Member, IEEE*, Wangwang Yang, Zhaofang Song, Zexu Chen, and Dengquan Lin

Abstract—In this paper, the bi-layer scheduling method for microgrids, based on deep reinforcement learning, is proposed to achieve economic and environmentally friendly operations. First, considering the uncertainty of renewable energy, the framework of day-ahead and intra-day scheduling is established, and the implementation scheme for both price-based and incentive-based demand response (DR) for the flexible load is determined. Then, comprehensively considering the operating characteristics of the microgrid in the day-ahead and intra-day time scales, a bi-layer scheduling model of the microgrid is established. In terms of algorithms, since day-ahead scheduling has no strict requirement for dispatching time, the particle swarm optimization (PSO) algorithm is used to optimize the time-of-use electricity price and distributed power output for the next day. Considering the environmental fluctuations and requirements for rapidity of intra-day online scheduling, the deep reinforcement learning (DRL) algorithm is adopted for optimization. Finally, based on the data from the actual microgrid, the rationality and effectiveness of the proposed scheduling method is verified. The results show that the proposed bi-layer scheduling based on the PSO and DRL algorithms achieves the optimization of scheduling cost and calculation speed, and is suitable for microgrid online scheduling.

Index Terms—Bi-layer optimal scheduling, demand response, deep reinforcement learning, microgrid scheduling.

I. INTRODUCTION

WITH the increasing elevation of environmental and energy problems, distributed renewable energy has attracted attention from all countries due to its environmentally friendly and sustainable characteristics [1]. The microgrid combines distributed power sources, load, energy storage devices and converters into a controllable unit [2], which provides an effective approach for the utilization of renewable energy. However, large-scale access of renewable energy to the microgrid will bring challenges to its operation and

dispatching [3], [4]. In order to reduce the impact of the uncertainty of renewable energy, the energy storage system (ESS) and demand response (DR) schemes are widely used in the microgrid [5]–[7]. With the continuous promotion of microgrid technology, how to properly manage the operation of distributed power sources, flexible load and energy storage systems in the microgrid to maximize the economic, technical and environmental benefits has become an important research topic [8].

Traditional mathematical programming (MP) methods, such as linear programming (LP), nonlinear programming (NLP), mixed integer linear programming (MILP) have been used to perform power scheduling tasks in microgrids for years [9]–[12], and they are easy to implement by using a solver, such as Cplex and Gurobi. However, these methods have difficulty in dealing with complex optimization problems with highly nonlinear objective functions and constraints. On the contrary, heuristic algorithms, such as the genetic algorithm (GA), particle swarm optimization (PSO) [13], [14], ant colony optimization (ACO) [15] and gravitational search algorithm (GSA) [16] are suitable for nonlinear and non-convex optimization problems, and they have been widely used in day-ahead scheduling of microgrids. Ref. [17] improves GA based on memory mechanism to minimize the operating cost of the microgrid. Ref. [18] develops the evolutionary PSO algorithm to optimize the combined cooling, heating and power system with renewable energy. Compared to the traditional MP methods, the group-based heuristic algorithms can greatly improve searching abilities.

The uncertainty of renewable energy output will affect the economy of dispatching, and there are two main methodologies to deal with this problem: proactive and reactive approaches [19]. Proactive approaches take all the possible scenarios into account to obtain a solution that is suitable for all scenarios, such as robust optimization [20] and chance-constrained programming [21]. Since all possibilities must be considered, this solution is relatively conservative. Reactive approaches, such as model predictive control (MPC), which continuously modify the scheduling plan based on updated forecast data to meet environmental changes, are more commonly used in current scheduling [22], [23]. Research indicates that the forecast error decreases with the shortening of the forecast time range [24]. The smaller the forecast error is, the more accurate the scheduling will be, provided that the

Manuscript received August 18, 2021; revised November 18, 2021; accepted December 29, 2021. Date of online publication May 6, 2022; date of current version September 7, 2022. This work was supported in part by National Key R&D Program of China under Grant 2021YFB3800200.

Z. Zhang, J. Shi (corresponding author, e-mail: shijing@mail.hust.edu.cn), W. Yang, Z. Song, Z. Chen, and D. Lin are with the State Key Laboratory of Advanced Electromagnetic Engineering and Technology, School of Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China.

DOI: 10.17775/CSEEJPES.2021.06120

calculation speed needs to be fast enough. The optimization methods adopted in the above literature are all model-based, which means that they all need to establish an accurate physical model and determine the parameters. Whereas, as the scale of the microgrid expands, the computational accuracy and efficiency will be greatly reduced. Therefore, the aforementioned two methods cannot meet the requirements of short-term scheduling for fast calculation and adaptability to uncertainty.

With the vigorous development of artificial intelligence technology, model-free deep reinforcement learning (DRL) algorithms have emerged. DRL is a type of data-driven intelligent algorithm, which possesses both perceptual ability of deep learning (DL) and the decision-making ability of reinforcement learning (RL), and is applicable for dynamic decision problems that are difficult to be accurately modeled. Moreover, due to the fast forward propagation speed of the neural network, DRL can achieve excellent performance in instant decision making. Up to now, DRL has been widely used in electric power related fields [25], [26]. For power system scheduling, an intra-day dynamic dispatch strategy for a power system based on the deep Q learning method is proposed in reference [27]. Reference [28] proposes a DRL approach for combined heat and power system economic dispatch, which obtains adaptability for different scenarios and decreases the computational complexity without affecting accuracy. For building energy management, DRL is utilized to realize online optimization of energy management in reference [29] to reduce the cost and the peak-to-valley differences of load. For DR scheduling, a DRL-based load scheduling algorithm which establishes a dueling deep Q-network model with a time-of-use pricing mechanism is developed in reference [30]. Reference [31] proposes a DRL algorithm for residential users, while accounting for uncertainties in the load demand, electricity price, users' privacy concerns, and power flow constraints. In the field of microgrid scheduling, reference [32] models the energy management of the microgrid as a Markov decision process (MDP) with the goal of minimizing daily operation costs, and uses the deep Q network to solve the problem. Whereas, the outputs of the deep Q network are discrete, which reduces the accuracy of the scheduling. A distributed energy management system that adopts the DRL algorithm is proposed in reference [33] to realize the interaction between different distributed power sources and the environment. In reference [34], the adjustment of thermostatically controlled load and price response load is considered in the microgrid energy management system, and 7 DRL algorithms are used for comparison and verification. However, these studies do not consider the following issues: 1) The impact of forecast accuracy of renewable energy generation on microgrid scheduling. 2) Appropriate strategy for collaborative scheduling of distributed energy and flexible load in microgrids. 3) Efficient DRL algorithms suitable for microgrid dispatching.

In this paper, a microgrid with renewable power generation, flexible load and ESS is studied to obtain its optimal scheduling plan. Due to the uncertainty of the environment, a bi-layer scheduling scheme, including day-ahead and intra-day

scheduling, is designed, aiming at reducing the influence of forecast error and making full use of the dispatchable resources in the microgrid. In the intra-day short-term scheduling, a DRL method is adopted to obtain the scheduling instruction. The contributions of this paper can be summarized as follows:

1) The framework of bi-layer scheduling for the microgrid is established. Because of the errors in the forecast data of the microgrid, bi-layer scheduling is needed. It can make full use of the day-ahead and intra-day forecast information, thereby effectively improving the scheduling accuracy and reduce the influence of uncertain factors.

2) A microgrid model considering two types of DR is built. Based on the analysis of the characteristics of dispatchable resources in the microgrid, the scheduling model of the microgrid is established, where price-based and incentive-based DR models are implemented.

3) A distributed proximal policy optimization (DPPO) algorithm, which is a DRL algorithm suitable for continuous problems, is adopted to optimize the intra-day scheduling of the microgrid. The comparison between the optimal results of the DPPO and MPC method is provided to demonstrate the effectiveness of the proposed scheduling method.

The remainder of this paper is organized as follows: Section II introduces the architecture of bi-layer optimal scheduling for the microgrid. Section III establishes the scheduling model of the microgrid. In Section IV, algorithms utilized in the two layers are discussed in detail, including the formulation of the microgrid model into the MDP and adopting the DPPO to solve this decision-making problem. A case study is presented in Section V. Section VI concludes this paper.

II. ARCHITECTURE OF BI-LAYER OPTIMAL SCHEDULING FOR MICROGRID

The typical topology of the microgrid is shown in Fig. 1. The microgrid is connected to the upper-level grid, and it contains a wind turbine (WT), photovoltaic (PV), micro turbine (MT), fuel cell (FC), energy storage system (ESS) and load.

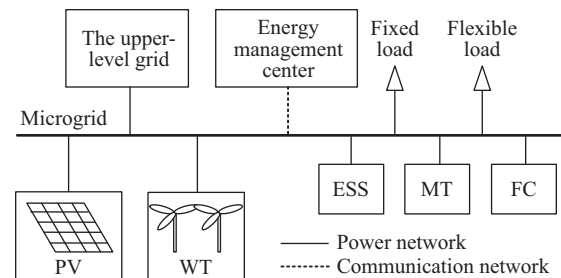


Fig. 1. Topology of the microgrid.

Generally, DR strategies can be divided into two types: price-based DR (PDR) and incentive-based DR (IDR). PDR refers to consumers adjusting their electricity consumption behavior according to the signals of price change. It primarily includes time of use pricing (TOU), real time pricing (RTP) and critical peak pricing (CPP), etc. IDR formulates certain policies, such as electricity price discounts or load shedding

compensation, to encourage the load adjustment [35]. Considering these two types of DR, the following strategies are formulated:

1) Setting TOU price. The purpose of setting TOU price is to encourage customers to shift their electricity consumption time from on-peak periods to off-peak periods, thereby reducing the peak-to-valley difference. The TOU price will be published one day before and updated once a day, and then the consumers will actively adjust their electricity consumption.

2) Signing the IDR contracts. Signing the IDR contracts with the consumers in advance to determine the compensation price, interruption time and other contents. Then they can get compensation according to the actual scheduling situation.

The overall scheduling process is divided into two layers: day-ahead scheduling and intra-day rolling scheduling, as shown in Fig. 2. The forecast data from different time scales is input into the scheduling model of the microgrid. The objective functions and constraints are set to optimize the model. The detailed contents are as follows:

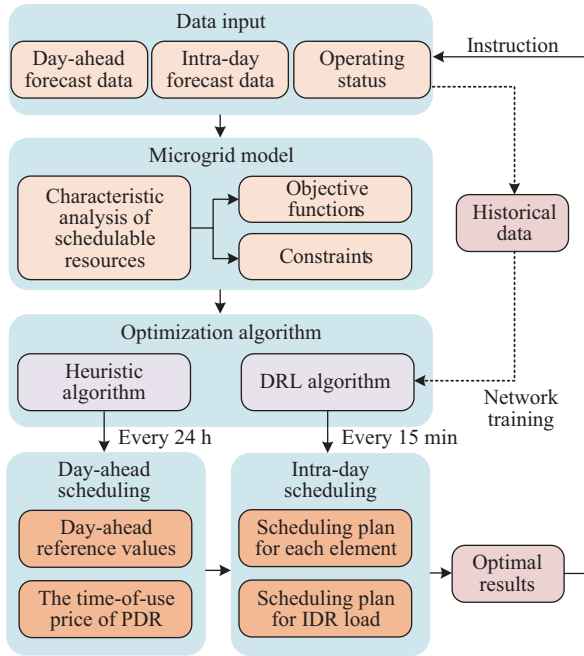


Fig. 2. The framework of bi-layer scheduling.

1) Day-ahead scheduling: carried out every 24 h with the resolution of 15 min. This layer uses day-ahead forecast information to optimize the model, and its scheduling results determine the TOU electricity price for PDR and the reference for intra-day scheduling. Day-ahead scheduling not only needs to optimize 96 points at a time, but also involves a large number of variables and nonlinear constraints. Thus, an efficient algorithm suitable for nonlinear problems will be appropriate.

2) Intra-day scheduling: carried out every 15 min with the resolution of 15 min. In this layer, the scheduling results of IDR and each micro source are obtained according to the day-ahead reference. Considering the uncertainty of forecast and the short scheduling time scale, an algorithm that can adapt to randomness and make fast decisions is needed.

III. SCHEDULING MODEL OF THE MICROGRID

A. Analysis of Dispatchable Resources

1) Micro Turbine

The power generation efficiency and fuel cost of the Capstone Company's C65 micro turbine can be expressed as:

$$\begin{cases} \eta_{MT} = 0.0753(P_{MT}/65)^3 - 0.3095(P_{MT}/65)^2 \\ \quad + 0.4174(P_{MT}/65) + 0.1068 \\ C_{MT} = C_n P_{MT} / LHV_{ng} \eta_{MT} \end{cases} \quad (1)$$

where C_n is the price of natural gas, yuan/m³. LHV_{ng} is the low calorific value of natural gas, kWh/m³, which is taken as 9.7 kWh/m³. P_{MT} is the net output power of the MT. η_{MT} is the unit's efficiency.

2) Fuel Cell

The power generation efficiency and fuel cost of the fuel cell can be expressed as:

$$\begin{cases} \eta_{FC} = -0.0023P_{FC} + 0.6735 \\ C_{FC} = C_{ng} P_{FC} / LHV_{ng} \eta_{FC} \end{cases} \quad (2)$$

where P_{FC} and η_{FC} are the output power and the efficiency of the fuel cell.

3) Energy Storage System

In microgrid projects, the battery is the most commonly used energy storage system. The operating cost of the battery is related to its cycle life [36]. Considering only the effect of discharge of depth D_{oD} on battery's cycle life N , the rain flow counting method is used to fit the discharge characteristics of the battery, which is defined as:

$$N = -3278D_{oD}^4 - 5D_{oD}^3 + 12823D_{oD}^2 - 14122D_{oD} + 5112 \quad (3)$$

According to the relationship between the battery's cycle life and the depth of discharge, the cost of the battery's life loss can be obtained by the throughput estimation method:

$$C_{ESS}^t = C_{init} \frac{u_{cha}^t P_{cha}^t + u_{dis}^t P_{dis}^t}{N \cdot 2E_{ESS}} \quad (4)$$

where C_{init} is the investment cost of the battery. u_{cha}^t and u_{dis}^t are integer variables of 0–1 representing the battery charge and discharge state at time t . P_{cha}^t and P_{dis}^t are the charge and discharge power at time t . E_{ESS} is the rated capacity.

4) Demand Response

a) *PDR*: The variation of load with the electricity price is usually described by the demand price elasticity coefficient. It represents the sensitivity of electricity consumption to the electricity price in different periods. The demand price elasticity coefficient can be expressed as [37]:

$$\varepsilon_{ij} = \frac{\Delta L_i / L_i^0}{\Delta p_j / p_j^0} \quad (5)$$

where L_i^0 and ΔL_i are the electricity consumption and its increment in the period i before and after the implementation of PDR. p_j^0 and Δp_j are the electricity sale price and its increment in the period j before and after the implementation of PDR. ε_{ij} is the influence of electricity price change in period i on demand in period j . When $i = j$, ε_{ij} represents

the self-elasticity coefficient, which is negative. When $i \neq j$, ε_{ij} represents the cross-elasticity coefficient, which is positive.

Considering the influence of the self-elasticity coefficient and the cross-elasticity coefficient, the load after PDR can be written as:

$$L_i = L_i^0 \left[1 + \varepsilon_{ii} \left(\frac{\Delta p_i}{p_0} \right) + \sum_{j=1}^{24} \varepsilon_{ij} \left(\frac{\Delta p_j}{p_0} \right) \right] \quad (6)$$

b) *IDR*: The cost of IDR is the load adjustment multiplied by the consumers' unit compensation, which can be given as:

$$C_{IDR}^t = \mu_C^t \cdot |\Delta P_{IDR}^t| \quad (7)$$

where ΔP_{IDR}^t is the load change in the period t and μ_C^t is the coefficient of IDR compensation cost in the period t .

B. Objective Functions

1) Day-ahead Scheduling

a) *Economic Cost*: The economic cost of day-ahead optimal scheduling includes the cost of micro-source power generation and the cost of power exchange between the microgrid and the upper grid. As WT and PV belong to the renewable energy power generation, their costs are ignored. The expression of the economic cost C_{ec} is:

$$C_{ec} = \sum_{t=1}^T (C_{MT}^t + C_{FC}^t + C_{ESS}^t + C_{grid}^t) \quad (8)$$

where T is the total time periods of day-ahead optimal scheduling with the value of 96. C_{MT}^t , C_{FC}^t , C_{ESS}^t are respectively the operating costs of the micro turbine, fuel cell and energy storage system in the period t . C_{grid}^t is the cost of the power exchange between the microgrid and the upper-level grid in the period t .

b) *Environmental Cost*: The power generation devices in the microgrid will affect the environment to varying degrees, thus incurring environmental cost. According to the outputs of the power sources, the emissions of various pollutants are obtained. Then the cost of impact on the environment can be calculated according to the discount standard and the amount of fines. Therefore, the environmental cost of the day-ahead dispatch C_{en} can be defined as:

$$C_{en} = \sum_{t=1}^T \sum_{j=1}^M Y_j^t D_j^t \quad (9)$$

where M is the type of pollutant (NO_X , CO_2 , CO , SO_2). Y_j^t is the amount of the j th pollutant and D_j^t is the converted cost of the j th pollutant in the period t .

Combining the above two costs, the objective function for the day-ahead scheduling can be written as:

$$\min F_{DA} = C_{ec} + C_{en} \quad (10)$$

2) Intra-day Scheduling

In the intra-day scheduling, to maintain the long-term

economy, the results of intra-day scheduling should be close to the day-ahead plan. Therefore, intra-day scheduling adds adjustment costs on the basis of economic and environmental costs. The adjustment cost reflects the difference between the two scheduling results of ESS, MT and FC, which can be described as:

$$\begin{cases} C_{pu} = \sum_{t=1}^T C_{\Delta SOC}^t + C_{\Delta MT}^t + C_{\Delta FC}^t \\ C_{\Delta SOC}^t = \alpha (S_{OC}^t - S_{OC,ref}^t)^2 \\ C_{\Delta MT}^t = \beta (P_{MT}^t - P_{MT,ref}^t)^2 \\ C_{\Delta FC}^t = \lambda (P_{FC}^t - P_{FC,ref}^t)^2 \end{cases} \quad (11)$$

where $C_{\Delta SOC}^t$, $C_{\Delta MT}^t$, $C_{\Delta FC}^t$ are the adjustment costs of S_{OC} , MT and FC respectively. α , β , λ are the coefficients of adjustment cost of S_{OC} , MT and FC, which are set as 1000, 0.1, and 0.1 in this paper. $S_{OC,ref}^t$, $P_{MT,ref}^t$ and $P_{FC,ref}^t$ are the reference values for the intra-day scheduling.

Considering IDR, the economic cost C_{ec} becomes:

$$C_{ec} = \sum_{t=1}^T (C_{MT}^t + C_{FC}^t + C_{ESS}^t + C_{grid}^t + C_{IDR}^t) \quad (12)$$

Finally, the objective function of intra-day scheduling can be given as:

$$\min F_{ID} = C_{ec} + C_{en} + C_{pu} \quad (13)$$

C. Constraints

The variables in the microgrid dispatching model should meet certain constraints [38], as shown as follows.

1) Power Balance

The day-ahead power balance constraint is:

$$P_{WT}^t + P_{PV}^t + P_{MT}^t + P_{FC}^t + P_{grid}^t = u_{cha}^t P_{cha}^t + u_{dis}^t P_{dis}^t + L^t \quad (14)$$

Since IDR is taken into account during the intra-day scheduling, the power balance constraint becomes:

$$P_{WT}^t + P_{PV}^t + P_{MT}^t + P_{FC}^t + P_{grid}^t = u_{cha}^t P_{cha}^t + u_{dis}^t P_{dis}^t + P_{L,PDR}^t + L^t + \Delta P_{IDR}^t \quad (15)$$

2) Tie Line Power

The maximum power of the tie line shall not exceed its capacity limit or the supply and demand agreement, so the tie line power constraint can be given as:

$$-P_{out,max} \leq P_{grid}^t \leq P_{in,max} \quad (16)$$

where $P_{in,max}$, $P_{out,max}$ are the upper limits of power purchase and power sale from the upper-level grid.

3) Micro Source

The output power of the micro turbine and fuel cell should be kept within a certain range, and the climbing constraints should also be met. The constraints of MT and FC are defined as:

$$\begin{cases} P_{MT,min} \leq P_{MT}^t \leq P_{MT,max} \\ P_{FC,min} \leq P_{FC}^t \leq P_{FC,max} \\ C_{down,MT} \leq P_{MT}^{t+1} - P_{MT}^t \leq C_{up,MT} \\ C_{down,FC} \leq P_{FC}^{t+1} - P_{FC}^t \leq C_{up,FC} \end{cases} \quad (17)$$

where $P_{FC,min}$, $P_{MT,max}$, $P_{FC,max}$ are the lower and upper limits of the output of MT and FC. $C_{down,MT}$, $C_{down,FC}$, $C_{up,MT}$, $C_{up,FC}$ are the minimum and maximum climbing power of MT and FC.

4) Energy Storage System

The state of charge (SOC) of the battery is an important decision variable that can prevent the battery from over-charging or over-discharging. S_{OC} can be expressed by the following formula:

$$S_{OC}^{t+1} = S_{OC}^t + \left(\frac{u_{cha}^t P_{cha}^t + u_{dis}^t P_{dis}^t}{E_{ESS}} \right) \Delta t \quad (18)$$

The constraints of ESS can be written as:

$$\begin{cases} u_{cha}^t \cdot u_{dis}^t = 0 \\ 0 \leq P_{cha}^t \leq P_{cha,max} \\ P_{dis,min} \leq P_{dis}^t \leq 0 \\ S_{OC,min} \leq S_{OC}^t \leq S_{OC,max} \end{cases} \quad (19)$$

5) PDR

TOU electricity prices are set according to on-peak, normal and off-peak periods. The prices in different periods are subtracted from the original electricity price to obtain Δp in (6).

If the ratio of the peak price to the valley price is too small, the response may be insufficient; if it is too large, the peak and valley may be reversed. In order to achieve effective peak-cutting and valley-filling, the following constraints are required:

$$r_{min} \leq \frac{p_{peak}}{p_{valley}} \leq r_{max} \quad (20)$$

where p_{peak} and p_{valley} are on-peak electricity price and off-peak electricity price.

6) IDR

During the scheduling, the load adjustment cannot exceed the maximum adjustable value in the period t :

$$0 \leq \Delta P_{IDR}^t \leq \Delta P_{IDR,max}^t \quad (21)$$

In addition, it is necessary to ensure that the accumulated amount of load adjustment is within the acceptable range. Therefore, to reasonably set the constraint of IDR, the state of load can be regarded as a small-capacity ESS. Set the state of load (SOL) as S_{OL} . S_{OL} indicates the relative cumulative adjustment of load, which is similar to S_{OC} . The change process of S_{OL} is shown in Fig. 3.

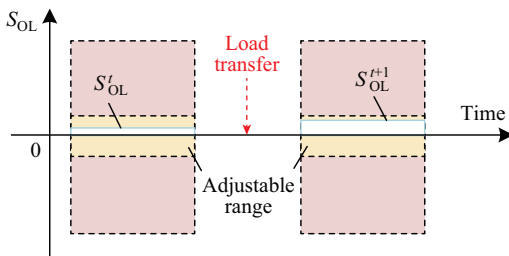


Fig. 3. The change process of S_{OL} .

The relationship between S_{OL} and load adjustment and the constraint of S_{OL} can be written as (22).

$$\begin{cases} S_{OL}^{t+1} = S_{OL}^t + \Delta P_{IDR}^t \Delta t / P_{L,av} \\ S_{OL,min} \leq S_{OL}^t \leq S_{OL,max} \end{cases} \quad (22)$$

where $P_{L,av}$ is the average load of a scheduling period. $S_{OL,min}$ and $S_{OL,max}$ are the lower and upper limits of S_{OL} .

IV. OPTIMAL ALGORITHMS FOR BI-LAYER SCHEDULING

In the above two layers of scheduling, the input data and scheduling objectives are different. The problem is a signomial geometric programming (SGP) due to the efficiency calculation formulas of MT and FC in equations (1) and (2). So the common solvers for MILP, such as Cplex and Gurobi, are no longer applicable. In order to ensure the efficiency of the solution and obtain better results, two different methods are used to solve the model.

(1) Day-ahead Scheduling: PSO Algorithm.

In day-ahead scheduling, 96 points of 24 h need to be optimized over time. If the MP method is adopted, the model will be difficult to converge, and the calculation time will explode. Compared with the MP method, the heuristic algorithm is more efficient and has been widely used in day-ahead scheduling of the microgrid. To optimize the nonlinear day-ahead model in acceptable computation time, PSO with a simple operation and fast convergence is selected.

(2) Intra-day Scheduling: DRL Algorithm.

In intra-day scheduling, the calculation speed is required to be fast enough to allow time for immediate instruction, and the algorithm should be able to accommodate fluctuations in the environment. Commonly used model-based algorithms often employ MPC methods to ensure long-term economy. However, with this method, the number of variables is multiplied and the calculation time is further increased. In contrast, the model-free algorithm will be more applicable. The DRL algorithm, which combines model-free reinforcement learning and perceptual deep learning, is an algorithm suitable for complex continuous decision-making problems. It can not only make decisions within seconds but also deal with environmental uncertainty.

A. Algorithm for Day-ahead Optimal Scheduling

The PSO algorithm originates from the study of predation behaviors of birds and fish, and its process is shown in Fig. 4.

First, a group of particles that meet the constraints are generated. Then, the fitness of each particle is calculated, and the optimal positions of the individual and the group are obtained by comparison. According to the optimal positions, the particles' information is continuously updated until the iteration is completed. The update process of the velocities and positions of all particles can be defined as [39]:

$$\begin{cases} v_i^{k+1} = \omega v_i^k + c_1 r_1 (P_{besti} - x_i^k) + c_2 r_2 (G_{besti} - x_i^k) \\ x_i^{k+1} = x_i^k + v_i^{k+1} \end{cases} \quad (23)$$

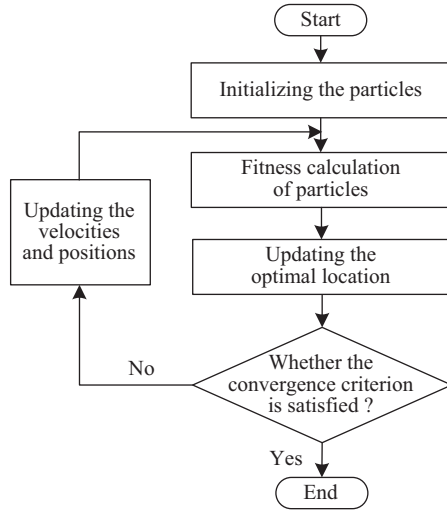


Fig. 4. Flow chart of PSO algorithm.

where v_i^k represents the speed of the particle i in the k th flight. c_1 and c_2 are the learning factors. r_1 and r_2 are random numbers between 0 and 1. ω is the inertia weight coefficient.

In the PSO algorithm, the linearly decreasing inertia weight is adopted, which is updated according to:

$$\omega = \omega_{\max} - k(\omega_{\max} - \omega_{\min})/iter_{\max} \quad (24)$$

where $\omega_{\max}$ and $\omega_{\min}$ are the maximum weight and the minimum weight. $iter_{\max}$ is the maximum number of iterations.

Due to the equality constraint of the power balance, only the outputs of MT, FC, and ESS at 96 points and the TOU price of the next day are set as variables. Then, the tie-line power at each point can be determined by the equality constraint. Thus, the number of variables is 290. To meet the boundary constraint of tie line power, the method of the penalty function is adopted. The method is to multiply the part of the variable beyond the constraint by the penalty coefficient and add it to the original objective function. Finally, the constrained optimization (25) is transformed into an unconstrained optimization (26).

$$\begin{cases} \min f(x) \\ g(x) \leq a \end{cases} \quad (25)$$

$$\min F(x) = f(x) + k[\max(a, g(x)) - a]^2 \quad (26)$$

B. Algorithm for Intra-day Optimal Scheduling

1) Reinforcement Learning

Reinforcement learning (RL) is a type of learning mechanism where the agent continuously interacts with the environment. The learning process is usually described as a Markov decision process (MDP). A typical MDP generally consists of a tuple as follows: $\{S, A, R, \gamma\}$. Among them, S and A represent the state space and action space of the agent. R is used to evaluate the reward of the current action in the current state. γ is the discount factor, which reflects the influence of future rewards on the current action selection [40].

The diagram of MDP is shown in Fig. 5(a).

First, the agent observes the current environment s_t . Then, according to certain strategies, the agent performs the action

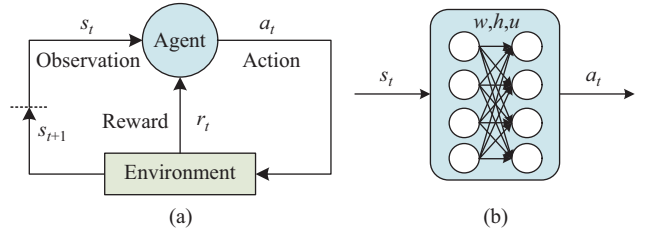


Fig. 5. Diagram of MDP (a) original MDP (b) MDP of DRL.

a_t and transfers the environment to the next state s_{t+1} . After that, the environment feeds back a reward r_t to the agent so that the agent can update its strategy. If the environment becomes complicated, it will be difficult to determine the agent's decision function. The deep neural network has strong fitting capabilities, so the decision-making process of RL can be replaced by a deep neural network, as shown in Fig. 5(b). In this way, DRL is formed.

2) The Reinforcement Learning Model of the Microgrid

In this paper, the dispatching center of the microgrid is the RL agent. To realize the optimal scheduling of the intra-day model, the mathematical model of the microgrid should be defined as the RL model $\{S, A, R, \gamma\}$ according to the Markov decision process.

a) *State Space*: The state space refers to the observed values and required data from the dispatching center. The factors that will affect the scheduling should be considered as much as possible, such as the forecast values of renewable energy power generation and load in the next 4 hours (16 points), as described in (27).

$$s_{f,16} = \{P_{WT}(t), P_{PV}(t), P_L(t)\} \quad (27)$$

In addition, in order to add ESS, IDR and climbing constraints, the operation state of the microgrid at the last point of time should also be considered, including S_{OC} , S_{OL} and the actual output of each micro source. Thus, the operation state s_s can be given as:

$$s_s = \{SOC(t-1), SOL(t-1), P_{MT}(t-1), P_{FC}(t-1)\} \quad (28)$$

Furthermore, the day-ahead scheduling results of ESS, MT and FC are used as references for intra-day scheduling:

$$s_{ref} = \{P_{ESS,ref}, P_{MT,ref}, P_{FC,ref}\} \quad (29)$$

In order to reduce the dimension of the state space, all time-related constants, such as electricity prices, are represented by the state of the time period. Therefore, the state space of intra-day scheduling can be expressed as a collection of forecast values, operating status, day-ahead reference and time period:

$$s(t) = \{s_{f,16}, s_s, s_{ref}, t\} \quad (30)$$

b) *Action Space*: Action space is the decision variable of the model. After observing the environment state, the agent selects an action in the action space A according to the strategy set. The output of each device and the load participating in IDR constitute the action space of the microgrid, which can be written as:

$$a_{ID}(t) = \{a_{MT}(t), a_{FC}(t), a_{ESS}(t), a_{grid}(t), a_{IDR}(t)\} \quad (31)$$

c) *Reward Function and Discount Factor*: The instant reward is obtained from the objective function of (13). On this basis, a penalty function similar to the day-ahead scheduling is defined, which can be given as:

$$D_p = k[\max(a, g(x)) - a]^2 \quad (32)$$

where D_p is the penalty function for the constraint of the tie line power.

The goal of reinforcement learning is to maximize the reward. Therefore, the instant reward $r(t)$ is defined as the negative value of the total cost:

$$r(t) = -(F_{ID} + D_p) \quad (33)$$

The discount factor indicates the importance of the future reward to the current reward. For economic dispatch, the decision at this point will have important impact on the future. Thus, γ should be a large value, which is set to 0.98.

Finally, the objective and constraints have all been considered in the MDP model, as shown in Fig. 6.

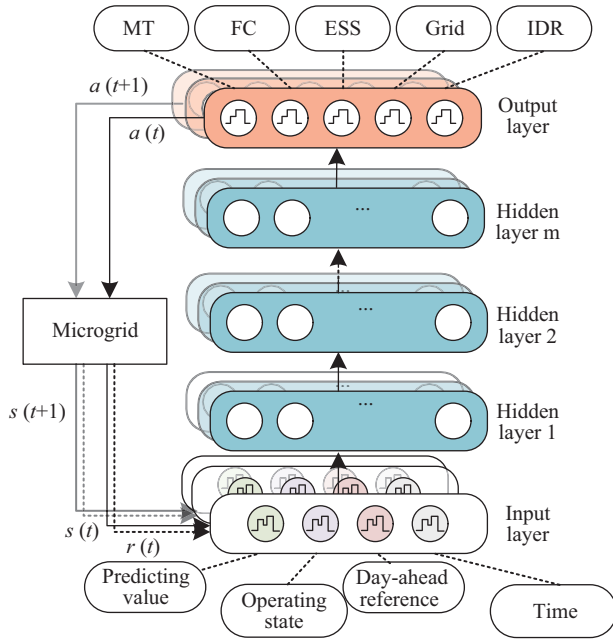


Fig. 6. The DRL model of the microgrid.

In the DRL model, the state space is the input and the action space is the output. According to the action, the microgrid returns the instant reward and new latest state. Then, the network updates in the direction of maximum reward.

3) Distributed Proximal Policy Optimization

a) *Proximal Policy Optimization*: Reinforcement learning can be divided into value-based learning and policy-based learning. Among them, the output of the value-based learning is the reward value of the action. It selects action with the highest value and it is suitable for discrete action. The common value-based methods are Q-learning and Sarsa. Policy-based learning directly obtains the probabilities of the next actions and selects the action according to the probability, which is suitable for discontinuous and continuous actions. The common policy-based method is policy gradients. In the

microgrid model, the action space and state space are both continuous, so the policy-based method is selected.

The policy-based method often uses the advantage function to construct the policy gradient. The advantage function represents the effect of the selected action a in the state s , which is calculated from (34). If the advantage function is positive, it means that the current action has a positive effect on the optimization goal, so the probability of its occurrence should be increased; otherwise, the probability should be reduced.

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s) \quad (34)$$

where $V_\pi(s)$ is the expected value of the return of the policy in state s , and $Q_\pi(s, a)$ is the expected value of the return of the policy after taking action a in state s , which can be defined as:

$$\begin{cases} V_\pi(s) = E_\pi[R(\tau)|s] \\ Q_\pi(s, a) = E_\pi[R(\tau)|s, a] \end{cases} \quad (35)$$

In most of the policy gradient methods, it is difficult to choose the appropriate update step size, resulting in the instability of training. Schulman [41] proposes trust region policy optimization (TRPO) to deal with the training of random strategy and it guarantees the stability of the policy. However, the efficiency and scalability of TRPO still needs to be improved. Therefore, proximal policy optimization (PPO) is proposed by OpenAI [42]. PPO can be regarded as an approximation of TRPO, which introduces a first-order approximation algorithm to improve efficiency and reliability. In practice, PPO has a good effect in dealing with the problem of high-dimensional continuous control. In addition, to maximize improvement without the risk of performance collapse, PPO uses a probability ratio between the current and old policies, which is defined as:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \quad (36)$$

In the policy gradient algorithm, the objective function of the network's parameter θ is:

$$L^{PG}(\theta) = \hat{E}_t[r_t(\theta)\hat{A}] \quad (37)$$

In PPO, the objective function is changed to:

$$L^{CLIP}(\theta) = \hat{E}_t \left[\min \left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t \right) \right] \quad (38)$$

where $\text{clip}(\cdot)$ is the clipping function and ε is its parameter. The clip function is the truncation function, which limits the probability ratio to $[1 - \varepsilon, 1 + \varepsilon]$.

b) *Distributed Proximal Policy Optimization*: To achieve good performance in random scenarios, a surrogate must be able to explore in as many different environments as possible. Therefore, distributed proximal policy optimization (DPPO) is adopted. DPPO is a multi-threaded, parallel version of PPO, where multiple parallel agents (workers) collect data in different environments and share the same global learner. Global learner updates the network through the experience collected by different workers. During the update process, the workers will stop interacting with the environment. After

the update, each worker continues to use the latest policy of the global network to explore the environment. Parallel agents obtain each other's learning information from the global network, thereby breaking the interdependence of gradient updates and improving the learning efficiency.

The pseudocodes of the global network and the worker of DPPO are shown in Algorithms 1 and 2, respectively.

c) *Training Process of the Networks*: Before adopting the bi-layer scheduling, the DPPO network needs to be trained. The training process is shown in Fig. 7.

Because of the large number of training data samples, the scenarios are reduced first to improve training efficiency. Then, the day-ahead reference is obtained by optimizing the day-ahead data of the reduced scene through the PSO algorithm. The reference and intra-day forecast data are input into DPPO

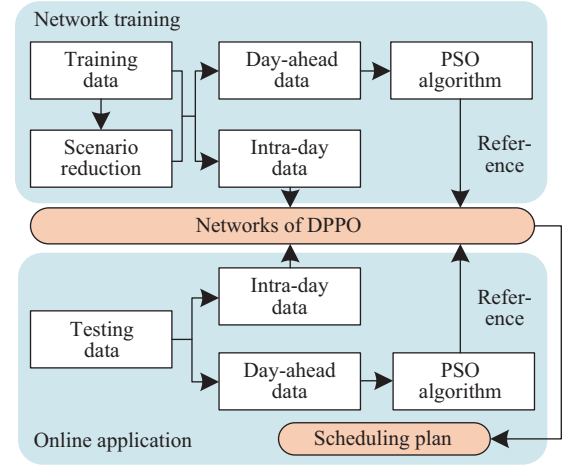


Fig. 7. The training process of the networks of DPPO.

Algorithm 1: DPPO-Global network

```

1 Initialize the parameters of critic network and actor network
2 for iteration = 1, 2, ..., iter_max do
3   Wait until transitions of all the workers are sent to the global replay buffer R
4   Update the policy:  $\pi_{\theta_{old}} \leftarrow \pi_{\theta}$ 
5   for i = 1, 2, ..., step_update do
6     Collect a mini-batch of transitions  $\{s_t, a_t, r_t, s_{t+1}\}$ 
7     Estimate  $Q_{\pi}(s_t, a_t)$ ,  $V_{\pi}(s_t)$ ,  $A_{\pi}(s_t, a_t)$  and  $L^{CLIP}(\theta)$ 
8     Update the parameters of critic network and actor network
9   end
10  Start worker threads
11 end

```

Algorithm 2: DPPO-Worker

```

1 Initialize the global replay buffer R
2 for iteration = 1, 2, ..., iter_max do
3   Obtain the forecast data and operation status of the microgrid
4   Input the initial state  $s_0 = \{s_{f16}, s_s, s_{ref}, t|t=0\}$ 
5   for t = 1, 2, ..., T do
6     Choose an action  $a_{ID}(t) = \{a_{t,MT}, a_{t,FC}, a_{t,ESS}, a_{t,grid}, a_{t,IDR}\}$ 
7     Get the reward  $r_t$  and the updated state  $s_{t+1}$ 
8     Store transition  $\{s_t, a_t, r_t, s_{t+1}\}$  in the local replay buffer
9     if N transitions are stored then
10      Estimate  $Q_{\pi}(s_t, a_t)$ ,  $V_{\pi}(s_t)$  and  $A_{\pi}(s_t, a_t)$ 
11      Send all the transitions and their  $A_{\pi}(s_t, a_t)$  to the global buffer R
12      Clear the local replay buffer until the global network is updated
13    end
14  end
15 end

```

to train the network. In the online application, the testing data is processed to acquire the input of the network, and the trained network will directly output the final scheduling plan.

Due to the training of reduced historical scenarios, the DRL algorithm has learned the environment under different conditions as far as possible, and has strong adaptability to environmental changes within a certain range. Although the environment changes, the algorithm can also sense the environmental state of the microgrid and directly make accurate decisions. In addition, the day-ahead forecast data, intra-day forecast data and actual operating data in the microgrid database are constantly updated with the progress of forecasting and scheduling, and the DRL network will be updated by regular data sampling. It means that the algorithm can adapt to changes in the environment even if there are major changes, such as upgrading and transformation of the microgrid.

V. CASE STUDY

A. Parameters and Data Processing

The scheduling method proposed is verified in the microgrid system as shown in Fig. 1. Set the base value to 500 kVA, the total active power of the load is 1.8 p.u., and the total active output of WT and PV are 0.4 p.u. and 0.5 p.u. Table I shows the parameters of the distributed power sources in the microgrid. The division of on-peak, normal and off-peak periods and the price of the power exchange between the microgrid and large grid are shown in Table II. The treatment cost and emission coefficient of pollutants are shown in Table III. The price elasticity matrix of demand is set according to reference [43].

The forecast and measured data were all taken from the data of the Belgian power grid every 15 minutes in 2020 [44]. The data shows that the intra-day forecast error is about 5% smaller than the day-ahead forecast error. Fig. 8 shows the day-ahead and intra-day forecast data of wind power, photovoltaic power generation and load for 153 days from June to November 2020.

The first step is to reduce the scenarios in the collected data. To further improve the efficiency and simplify the data, this paper first subtracts the renewable energy generation power

TABLE I
PARAMETERS OF DISTRIBUTED POWER SOURCES IN THE MICROGRID

Title	Size	Quantity	Total power
WT	10 kW	20	200 kW
PV	200 W	1250	250 kW
MT	65 kW	5	325 kW
FC	40 kW	5	200 kW
BA	25 kW	5	500 kWh

TABLE II
DIVISION OF TIME PERIODS AND PRICE OF POWER EXCHANGE

Period	Time	Electricity price (yuan/kWh)	
		Purchase	Sell
On-peak	10:00–14:00;	1.13	1.02
	18:00–22:00		
Normal	8:00–10:00;	0.69	0.51
	14:00–18:00;		
Off-peak	22:00–24:00	0.35	0.29
	0:00–8:00		

TABLE III
THE TREATMENT COST AND EMISSION COEFFICIENT OF POLLUTANTS

Pollutant	Treatment cost (yuan/kg)	Emission coefficient (g/kWh)				
		MT	FC	Grid	WT	PV
NO _x	8.000	0.619	0.010	1.6	0	0
CO ₂	0.023	184	489	889.0	0	0
SO ₂	6.000	0.001	0.003	1.8	0	0

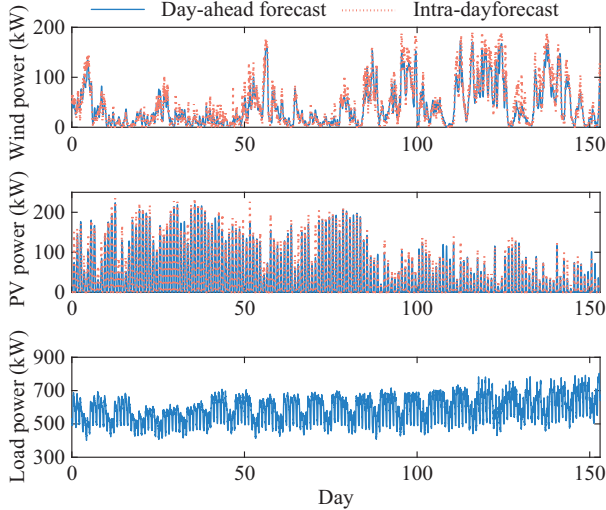


Fig. 8. The historical data from June to November 2020.

from the load power to obtain the power shortage, and then the K-means clustering is used to reduce the power shortage scenarios. K-means clustering [45] is a common scenario clustering method, which takes the mean value of the samples in each cluster as the clustering center. Fig. 9 shows the power shortage scenarios before and after the reduction. The scenes are divided into three centers by the K-means method. It can be seen that the power shortages of the three scenarios are primarily different at noon, because the daily photovoltaic output varies the most at noon. If the illumination intensity is high, the power shortage at noon will be relatively small.

B. Result Analysis

The typical 3 scenarios are used as the training set to train

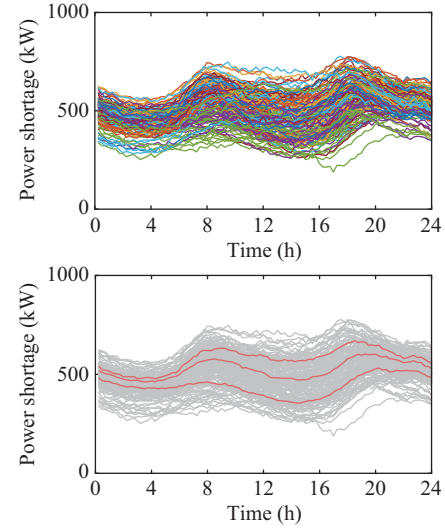


Fig. 9. Power shortage scenarios before and after the reduction.

the DPPO network. Then, 30 scenarios are randomly selected from the collected data as the test set for method verification.

1) Optimal Results of the Day-ahead Scheduling

In the PSO algorithm, the number of particles is set to 150 and the number of iterations is set to 500. Suppose the inertia weight coefficient $\omega_{\max} = 0.9$, $\omega_{\min} = 0.4$ and the learning factor $c_1 = c_2 = 1.49$. The PSO algorithm is implemented on Matlab 2017a. According to the proportion of renewable energy power generation, two scenarios are selected to display the results.

Figures 10 and 11 show the results of day-ahead scheduling optimization under the conditions of sufficient renewable energy output (scenario 19) and insufficient renewable energy output (scenario 9). It can be seen that the implementation of time-of-use electricity prices can reduce the peak-to-valley load difference to a certain extent. In Fig. 10, there is a high proportion of renewable energy output in scenario 19. Due to the high cost, MT is only turned on during the on-peak period. In addition, the microgrid purchases electricity during the off-peak period and sells electricity during the on-peak period, realizing the economy of dispatch. As shown in Fig. 11, the output of renewable energy is almost zero. MT remains on during both on-peak and normal periods. As MT and FC are insufficient to meet the power supply demand, the microgrid continuously purchases electricity from the upper-level grid to ensure the reliability of the power supply. According to the curve of S_{OC} in Fig. 10(b) and Fig. 11(b), the ESS in both cases discharges between 10:00 to 14:00 during the on-peak period, and it is more fully utilized when renewable energy generation is insufficient.

Similar to PSO, GSA [16] is derived from the swarm intelligence optimization algorithm generated by the simulation of universal gravitation. For comparative verification, GSA is also used to optimize the model and obtain the day-ahead plan of 30 scenarios. Fig. 12. shows the performance of the PSO algorithm and GSA. As shown in Fig. 12(a), when the same scenario is optimized, GSA reaches convergence earlier, and the fitness function does not significantly change compared

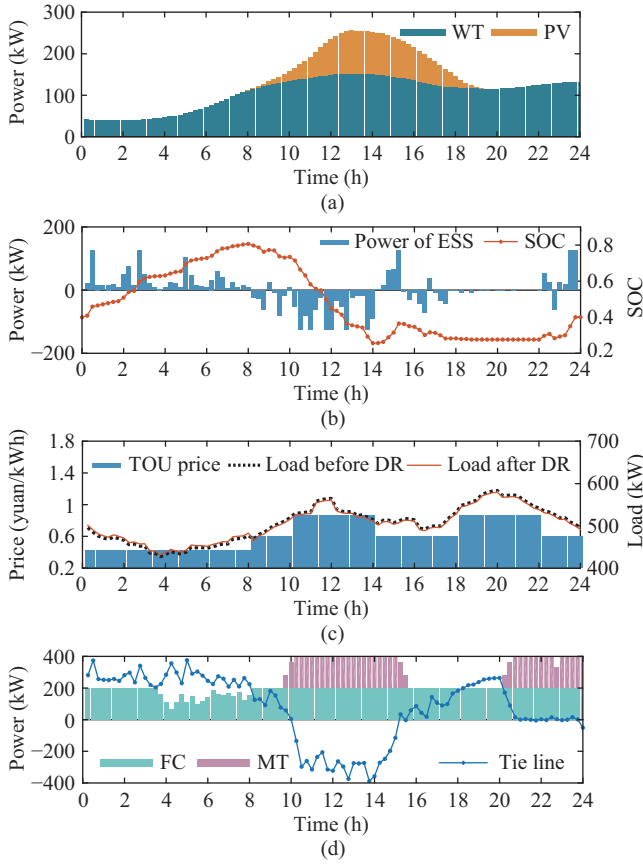


Fig. 10. The day-ahead optimal results of scenario 19 (a) Renewable energy; (b) ESS; (c) TOU price and load; (d) FC, MT and tie line.

with its initial value, which means that GSA has fallen into a premature state. Fig. 12(b) shows that under the same number of iterations, GSA requires a longer time than PSO. The reason is that GSA needs to calculate the mass, gravity and acceleration of particles before updating their positions, making the iterative process more complicated. As shown in Fig. 12(c), the final optimization result of PSO is better than that of GSA, indicating that GSA has stronger optimization ability. According to the principles of the two algorithms, in the search process, GSA can only perceive the current position of the particles without the function of group memory, so it is difficult to obtain the global optimum. PSO attaches great importance to the sharing of group information, so it has stronger search capabilities. Therefore, PSO is more suitable for the day-ahead scheduling model in this paper.

2) Optimal Results of the Intra-day Scheduling

In the DPPO network, the number of hidden layers in both the actor network and the critic network is 3. The ReLU activation function is used in all hidden layers. Adam optimizer is used to update the weights of the all the networks. Other parameters are listed in Table IV. Based on the above parameters, using Python 3.6 and framework of TensorFlow 1.18, the intra-day scheduling is implemented on a 64-bit Windows-based computer with 16 GB of RAM and AMD 6-Core processor clocking at 3.60 GHz.

The results of intra-day scheduling of two chosen scenarios are shown in Figs. 13 and 14. Due to the forecast error,

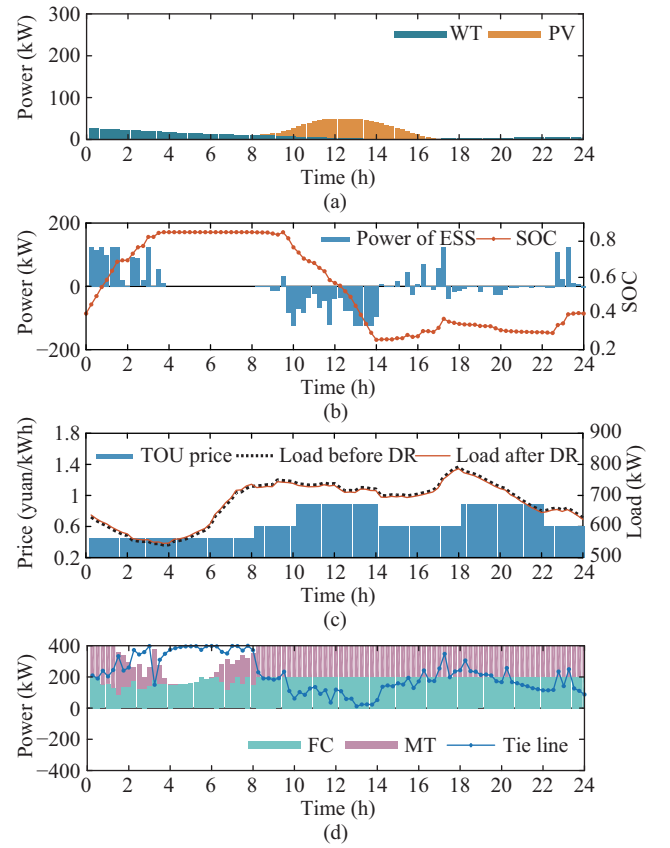


Fig. 11. The day-ahead optimal results of scenario 9 (a) Renewable energy; (b) ESS; (c) TOU price and load; (d) FC, MT and tie line.

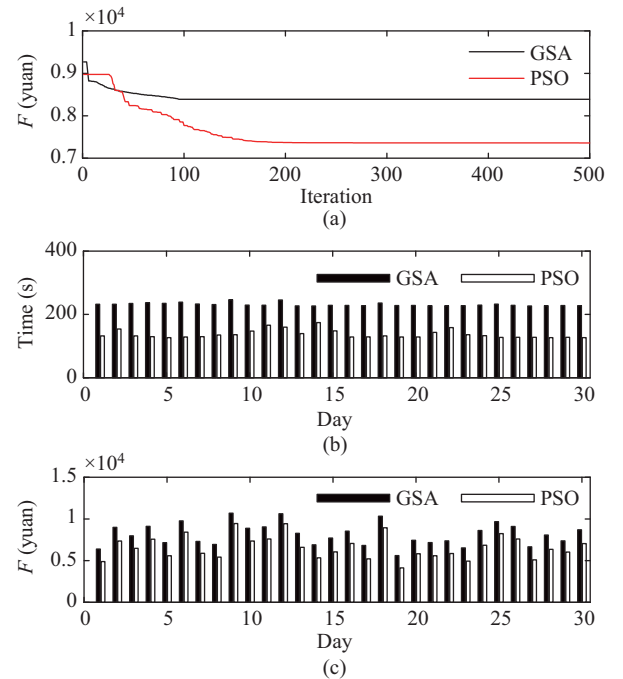
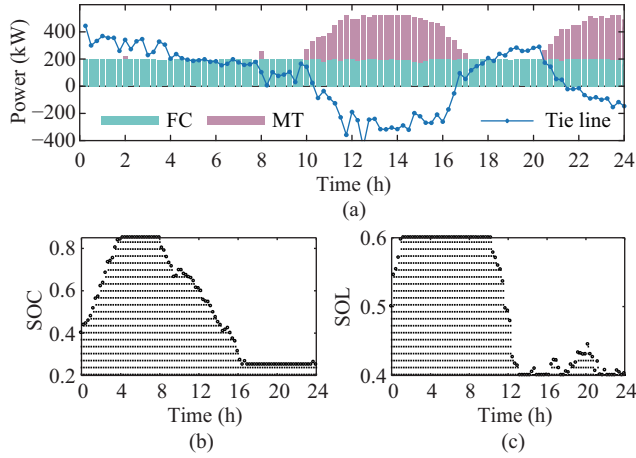
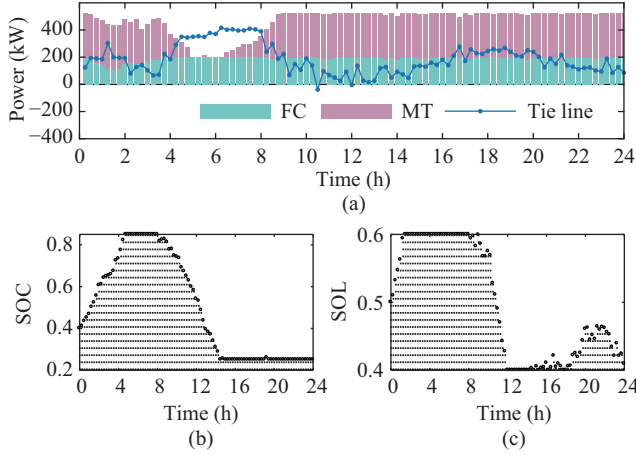


Fig. 12. The performance of PSO and GSA (a) Convergence curves of scenario 1; (b) Calculation time; (c) Results of the objective function.

although the day-ahead scheduling plan is taken as reference, the results of intra-day scheduling are not completely similar to the day-ahead results. During the scheduling, DRL perceives

TABLE IV
PARAMETERS OF DPPO

Hyper-parameters	Value
Learning rate of actor α_π	0.00001
Learning rate of critic α_v	0.0001
Number of neurons in the hidden layer of the actor network	50, 30, 30
Number of neurons in the hidden layer of the critic network	300, 100, 100
Capacity of experience replay pool	64
Number of threads	4
Clip parameter ε	0.2

Fig. 13. The intra-day optimal results of scenario 19 (a) FC, MT and tie line; (b) S_{OC} ; (c) S_{OC} .Fig. 14. The intra-day optimal results of scenario 9 (a) FC, MT and tie line; (b) S_{OC} ; (c) S_{OC} .

the difference between the latest data and the previous plan, then outputs the final optimization results. Because of the penalty constraints, the trend of S_{OC} is close to the results of day-ahead scheduling. The SOL curve reflects the load adjustment. It can be seen that SOL rapidly decreases to the lower limit from 8:00 to 12:00, which means the load is reduced. By reducing the peak load through IDR, the peak-valley gap can be narrowed, thus saving scheduling costs.

Figures 15 shows the cumulative reward and loss functions of DPPO. As can be seen from the reward (after scaling down) in Fig. 15, with the progress of training, the performance of the

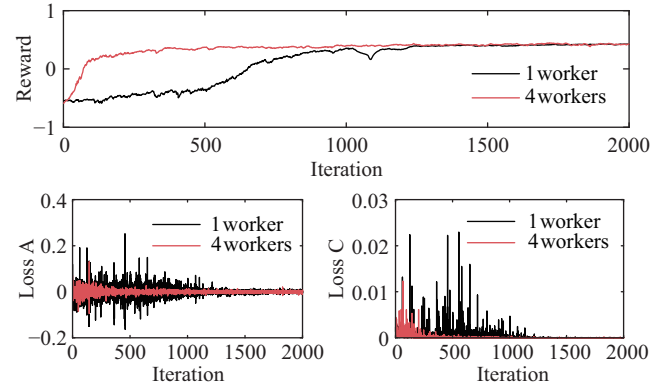


Fig. 15. Cumulative reward and loss functions of DPPO.

model becomes better and better. There are large oscillations of the loss function caused by the environmental exploration and random actions, but both the loss functions are less than 10^{-2} most of the time, indicating the rapid convergence of DPPO. By comparing the training effect of a single-thread PPO and multi-thread DPPO, it can be found that the multi-thread DPPO has faster convergence speed and higher learning efficiency.

3) Comparative Verification

To verify the effectiveness and economy of the proposed scheme, the scheduling cost of 30 days under the following four cases are separately calculated:

Case 1: Day-ahead scheduling optimized by PSO.

Case 2: Day-ahead scheduling optimized by PSO and intra-day scheduling optimized by MPC (with DR). Fmincon solver is used to solve the problem.

Case 3: Day-ahead scheduling optimized by PSO and intra-day scheduling optimized by DPPO (without DR).

Case 4: Day-ahead scheduling optimized by PSO and intra-day scheduling optimized by DPPO (with DR).

The scheduling costs are shown in Table V.

Table V shows that the cost of bi-layer scheduling is about 2% less than single-layer scheduling. This is because the day-ahead forecast errors requires automatic generation control units to participate in power generation adjustment, which is expensive. Comparing case 3 with case 4, it can be concluded that DR strategy significantly reduces the power generation cost. Comparing case 2 and case 4, it can be perceived that it is more economical to use DPPO than fmincon solver. The calculation of MPC needs the forecast data of each point in the rolling cycle, so some points with large forecast errors will be considered, which reduces the accuracy of the model. However, DRL algorithms do not use a certain model. It takes only the most recent forecast value as the input state, which improves its adaptability to uncertainty. Moreover, the influence of historical rewards on instant rewards is considered when training the algorithm. Therefore, the optimization results of DPPO also achieve long-term economics. Compared with the method of MPC, the cost is reduced by about 2%.

The calculation time of the MPC and DRL methods is shown in Fig. 16. The time required to calculate each point with the MPC method ranges from seconds to minutes.

TABLE V
THE SCHEDULING COSTS OF 30 DAYS UNDER DIFFERENT CASES

Day	Cost(yuan)			
	Day-ahead scheduling		Bi-layer scheduling	
	Case 1	Case 2	Case 3	Case 4
Day1	7046.1	5464.1	5342.4	5137.4
Day2	9741.3	8024.1	8322.8	7871.3
Day3	8498.6	6709.3	6706.5	6430.9
Day4	9803.5	8082	8221.5	7857.4
Day5	7703.1	6204.5	6124.3	5821.1
Day6	10491.7	8564.8	8595	8406.7
Day7	7776.3	6044.7	6059.1	5781
Day8	7529.7	5846.1	5975.2	5647.6
Day9	11594.5	9730.9	9782.6	9591.6
Day10	9409.3	7650.6	7762.4	7507.5
Day11	9723.5	8136.9	8394.4	8071
Day12	11585.8	9695.4	9701.1	9576.8
Day13	8739.7	7181.9	7066.8	6920.5
Day14	7466.6	5901.4	5983.9	5680
Day15	8192.2	6305.4	6280.5	6014.9
Day16	8978.4	7390.6	7786.3	7361.4
Day17	7406.3	6129.7	6042.3	5738.2
Day18	11212.7	9233.3	9328.1	9105.5
Day19	6357.3	4886.3	4862.9	4654.5
Day20	8141	6724.2	6420	6261.9
Day21	7727.8	6133.3	6242.5	5927.9
Day22	7939.3	6332.2	6422.7	6002.3
Day23	7135.9	5418.2	5445.6	5301.2
Day24	9098.3	7361.4	7810.4	7508.8
Day25	10359.1	8642.2	8873.3	8603.9
Day26	10018.4	8205.1	8297.5	8096
Day27	7303.7	5796.6	5811.8	5486
Day28	8707.7	6804.3	6777.8	6506.3
Day29	7965.5	6431.1	6483.1	6226.6
Day30	9287.2	6058.4	7942.1	7725
Total	262933.4	211080.7	214855.7	206802.8

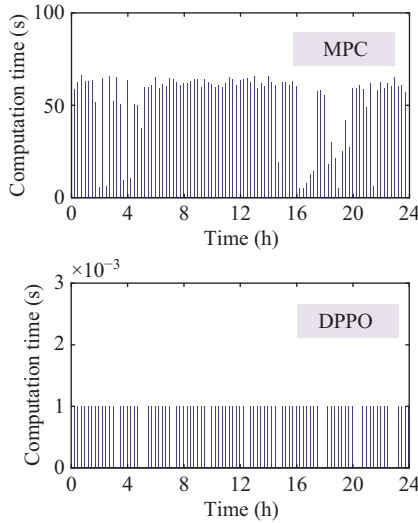


Fig. 16. The calculation time of Fmincon and DPPO algorithm.

Whereas, when using the trained DPPO for optimization, it only takes milliseconds. Since MPC is a model-based optimization method, all points in the time scale need to be optimized each time, which increases the dimension of the model. Being different, DPPO is a model-free algorithm. Although it takes some time to train the network in the early stages, the scheduling instructions can be obtained quickly through the trained network during online scheduling, so the speed of DPPO is faster.

Compared with the method of MPC, using DPPO to optimize the model reduces the cost by about 2%, and significantly shortens the calculation time. Therefore, for the proposed bi-layer scheduling scheme, the DPPO algorithm can realize fast and economical dynamic scheduling. In fact, if the period of intra-day scheduling is further shortened, DPPO will obtain better results with the improvement of forecast accuracy.

VI. CONCLUSION

In this paper, considering the uncertainty of renewable energy and flexible load control at different time scales, a bi-layer optimal dispatching method for microgrids is proposed. The scheduling is divided into two layers: day-ahead scheduling and intra-day scheduling. Taking into account the different requirements for calculation time and accuracy of the two layers, the day-ahead scheduling model is solved by the PSO algorithm, and the intraday scheduling model is solved by the DPPO algorithm of DRL. Based on the data from the Belgian power grid, the proposed dispatch plan is verified.

1) The cost of the proposed bi-layer scheduling is 20% lower than that of day-ahead single-layer scheduling, and the DR strategy can further reduce scheduling cost.

2) For the day-ahead scheduling model, the PSO algorithm has stronger search capability and computational efficiency than the GSA.

3) Compared with the MPC method, DPPO can reduce the scheduling cost by about 2%, and the scheduling time can be reduced from minutes to milliseconds.

4) The proposed scheduling scheme is adaptable for environmental fluctuations, and the uncertainty of renewable energy has little effect on the scheduling schemes.

In a future study, the time resolution of the scheduling model can be further reduced, such as 1 minute or 5 minutes, so that the model can be applied to the real-time scheduling of the microgrid.

REFERENCES

- [1] A. Chaouachi, R. M. Kamel, R. Andoulsi, and K. Nagasaka, "Multi-objective intelligent energy management for a microgrid," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 4, pp. 1688–1699, Apr. 2013.
- [2] T. Strasser, F. Andr n, J. Kathan, C. Cecati, C. Buccella, P. Siano, P. Leit o, G. Zhabelova, V. Vyatkin, P. Vrba, and V. Maf k, "A review of architectures and concepts for intelligence in future electric energy systems," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 4, pp. 2424–2438, Apr. 2015.
- [3] G. R. Ren, J. F. Liu, J. Wan, Y. F. Guo, and D. R. Yu, "Overview of wind power intermittency: impacts, measurements, and mitigation solutions," *Applied Energy*, vol. 204, pp. 47–65, Oct. 2017.
- [4] E. O. Arwa and K. A. Folly, "Reinforcement learning techniques for optimal power control in grid-connected microgrids: a comprehensive review," *IEEE Access*, vol. 8, pp. 208992–209007, Nov. 2020.
- [5] M. Parsa Moghaddam, A. Abdollahi, and M. Rashidinejad, "Flexible demand response programs modeling in competitive electricity markets," *Applied Energy*, vol. 88, no. 9, pp. 3257–3269, Sep. 2011.
- [6] D. S. Callaway and I. A. Hiskens, "Achieving controllability of electric loads," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 184–199, Jan. 2011.
- [7] I. R. S. da Silva, J. E. A. de Alencar, and R. de Andrade Lira Rab lo, "A preference-based multi-objective demand response mechanism," in *Proceedings of 2020 IEEE Congress on Evolutionary Computation*, 2020, pp. 1–8.
- [8] B. Roberts, "Capturing grid power," *IEEE Power and Energy Magazine*, vol. 7, no. 4, pp. 32–41, Jul./Aug. 2009.

- [9] Y. Riffonneau, S. Bacha, F. Barruel, and S. Ploix, "Optimal power flow management for grid connected PV systems with batteries," *IEEE Transactions on Sustainable Energy*, vol. 2, no. 3, pp. 309–320, Jul. 2011.
- [10] A. Parisio, E. Rikos, and L. Glielmo, "A model predictive control approach to microgrid operation optimization," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 5, pp. 1813–1827, Sep. 2014.
- [11] R. Palma-Behnke, C. Benavides, F. Lanas, B. Severino, L. Reyes, J. Llanos, and D. Sáez, "A microgrid energy management system based on the rolling horizon strategy," *IEEE Transactions on Smart Grid*, vol. 4, no. 2, pp. 996–1006, Jun. 2013.
- [12] D. E. Olivares, C. A. Cañizares, and M. Kazerani, "A centralized energy management system for isolated microgrids," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1864–1875, Jul. 2014.
- [13] S. A. Pourmousavi, M. H. Nehrir, C. M. Colson, and C. S. Wang, "Real-time energy management of a stand-alone hybrid wind-microturbine energy system using particle swarm optimization," *IEEE Transactions on Sustainable Energy*, vol. 1, no. 3, pp. 193–201, Oct. 2010.
- [14] S. A. Alavi, A. Ahmadian, and M. Aliakbar-Golkar, "Optimal probabilistic energy management in a typical micro-grid based-on robust optimization and point estimate method," *Energy Conversion and Management*, vol. 95, pp. 314–325, May 2015.
- [15] M. Marzband, E. Yousefnejad, A. Sumper, and J. L. Domínguez-García, "Real time experimental implementation of optimum energy management system in standalone Microgrid by using multi-layer ant colony optimization," *International Journal of Electrical Power & Energy Systems*, vol. 75, pp. 265–274, Feb. 2016.
- [16] M. Marzband, M. Ghadimi, A. Sumper, and J. L. Domínguez-García, "Experimental validation of a real-time energy management system using multi-period gravitational search algorithm for microgrids in islanded mode," *Applied Energy*, vol. 128, pp. 164–174, Sep. 2014.
- [17] A. Askarzadeh, "A memory-based genetic algorithm for optimization of power generation in a microgrid," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 3, pp. 1081–1089, Jul. 2018.
- [18] A. Lorestani and M. M. Ardehali, "Optimal integration of renewable energy sources for autonomous tri-generation combined cooling, heating and power system based on evolutionary particle swarm optimization algorithm," *Energy*, vol. 145, pp. 839–855, Feb. 2018.
- [19] Z. K. Li and M. Ierapetritou, "Process scheduling under uncertainty: review and challenges," *Computers & Chemical Engineering*, vol. 32, no. 4–5, pp. 715–727, Apr. 2008.
- [20] J. L. Yi, P. F. Lyons, P. J. Davison, P. F. Wang, and P. C. Taylor, "Robust scheduling scheme for energy storage to facilitate high penetration of renewables," *IEEE Transactions on Sustainable Energy*, vol. 7, no. 2, pp. 797–807, Apr. 2016.
- [21] H. Y. Wu, M. Shahidehpour, Z. Y. Li, and W. Tian, "Chance-constrained day-ahead scheduling in stochastic power system operation," *IEEE Transactions on Power Systems*, vol. 29, no. 4, pp. 1583–1591, Jul. 2014.
- [22] Z. J. Bao, Q. Zhou, Z. H. Yang, Q. Yang, L. Z. Xu, and T. Wu, "A multi time-scale and multi energy-type coordinated microgrid scheduling solution—Part I: model and methodology," *IEEE Transactions on Power Systems*, vol. 30, no. 5, pp. 2257–2266, Sep. 2015.
- [23] F. Arasteh and G. H. Riahy, "MPC-based approach for online demand side and storage system management in market based wind integrated power systems," *International Journal of Electrical Power & Energy Systems*, vol. 106, pp. 124–137, 2019.
- [24] S. Tewari, C. J. Geyer, and N. Mohan, "A statistical model for wind power forecast error and its application to the estimation of penalties in liberalized markets," *IEEE Transactions on Power Systems*, vol. 26, no. 4, pp. 2031–2039, Nov. 2011.
- [25] Z. Zhang, D. Zhang and R. C. Qiu, "Deep reinforcement learning for power system applications: An overview," *CSEE Journal of Power and Energy Systems*, vol. 6, no. 1, pp. 213–225, Mar. 2020.
- [26] M. Khodayar, G. Liu, J. Wang and M. E. Khodayar, "Deep learning in power systems research: A review," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 2, pp. 209–220, Mar. 2021.
- [27] H. Tang, S. P. Wang, K. J. Chang, and J. Y. Guan, "Intra-day dynamic optimal dispatch for power system based on deep Q-learning," *IEEE Transactions on Electrical and Electronic Engineering*, vol. 16, no. 7, pp. 954–964, Jul. 2021.
- [28] S. Y. Zhou, Z. J. Hu, W. Gu, M. Jiang, M. Chen, Q. T. Hong, and C. Booth, "Combined heat and power system intelligent economic dispatch: a deep reinforcement learning approach," *International Journal of Electrical Power & Energy Systems*, vol. 120, pp. 106016, Sep. 2020.
- [29] E. Mocanu, D. C. Mocanu, P. H. Nguyen, A. Liotta, M. E. Webber, M. Gibescu, and J. G. Sloopweg, "On-line building energy optimization using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.
- [30] B. Wang, Y. Li, W. Y. Ming, and S. R. Wang, "Deep reinforcement learning method for demand response management of interruptible load," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3146–3155, Jul. 2020.
- [31] S. Bahrami, Y. C. Chen, and V. W. S. Wong, "Deep reinforcement learning for demand response in distribution networks," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1496–1506, Mar. 2021.
- [32] Y. Ji, J. H. Wang, J. C. Xu, X. K. Fang, and H. G. Zhang, "Real-time energy management of a microgrid using deep reinforcement learning," *Energies*, vol. 12, no. 12, pp. 2291, Jun. 2019.
- [33] E. Foruzan, L. K. Soh, and S. Asgarpour, "Reinforcement learning approach for optimal distributed energy management in a microgrid," *IEEE Transactions on Power Systems*, vol. 33, no. 5, pp. 5749–5758, Sep. 2018.
- [34] T. A. Nakabi and P. Toivanen, "Deep reinforcement learning for energy management in a microgrid with flexible demand," *Sustainable Energy, Grids and Networks*, vol. 25, pp. 100413, Mar. 2021.
- [35] J. Valenzuela, P. R. Thimmapuram, and J. Kim, "Modeling and simulation of consumer response to dynamic pricing with enabled technologies," *Applied Energy*, vol. 96, pp. 122–132, Aug. 2012.
- [36] B. Zhao, X. S. Zhang, J. Chen, C. S. Wang, and L. Guo, "Operation optimization of standalone microgrids considering lifetime characteristics of battery energy storage system," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 4, pp. 934–943, Oct. 2013.
- [37] D. S. Kirschen, G. Strbac, P. Cumperayot, and D. de Paiva Mendes, "Factoring the elasticity of demand in electricity prices," *IEEE Transactions on Power Systems*, vol. 15, no. 2, pp. 612–617, May 2000.
- [38] R. X. Yin, E. C. Kara, Y. P. Li, N. DeForest, K. Wang, T. Y. Yong, and M. Stadler, "Quantifying flexibility of commercial and residential loads for demand response using setpoint changes," *Applied Energy*, vol. 177, pp. 149–164, Sep. 2016.
- [39] A. Hossain, H. R. Pota, S. Squartini, and A. F. Abdou, "Modified PSO algorithm for real-time energy management in grid-connected microgrids," *Renewable Energy*, vol. 136, pp. 746–757, Jun. 2019.
- [40] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proceedings of the 4th International Conference on Learning Representations*, 2015.
- [41] J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015.
- [42] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, Proximal policy optimization algorithms, 2017, [Online] Available: <https://arxiv.org/abs/1707.06347>
- [43] S. M. Zhang, J. Q. Rong, and B. Y. Wang, "An optimal scheduling scheme for smart home electricity considering demand response and privacy protection," *International Journal of Electrical Power & Energy Systems*, vol. 132, pp. 107159, Nov. 2021.
- [44] Elia Group. Transparency on grid data. [Online]. Available: <http://www.elia.be/en/grid-data>.
- [45] Y. Zeng, C. Li, and H. M. Wang, "Scenario-set-based economic dispatch of power system with wind power and energy storage system," *IEEE Access*, vol. 8, pp. 109105–109119, Jun. 2020.



Zitong Zhang received a B.S. degree in Electrical Engineering from the School of Electrical Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2020. Her main research interests include multi-time scale economic dispatch, deep learning and its application in smart grids and grid connected operation of energy storage systems.



Jing Shi (M'10–SM'17) received Ph.D. degrees from the School of Electrical Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2008. Her main research interests include application of energy storage in power systems, and applied superconductivity in power systems.



Zhaofang Song is currently pursuing a M.S. degree in Electrical Engineering at Huazhong University of Science and Technology, Wuhan, China. His current research interests include load forecasting and electric vehicles.



Wangwang Yang received a B.S. degree in Electrical Engineering from the School of Electrical Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2020. His main research interests include assessment of demand response potential, data-based load characteristics and load forecasting.



Dengquan Lin received a B.S. degree in Electrical Engineering from the School of Electrical Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2021. His main research interests include the applications of energy storage systems.



Zexu Chen received a B.S. degree in Electrical Engineering from the School of Electrical Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2019. His main research interests include the characteristics of temperature-controlled loads participating in demand response and its control method.