Journal of Network Intelligence ©2023 ISSN 2414-8105 (Online)
Taiwan Ubiquitous Information Volume **8**, Number **4**, November **2023**

Real-time Energy Dispatching Strategy for Microgrid Based on Improved SD3 Algorithm

Tian-Xiao Hang

School of Computer Science and Mathematics
Fujian University of Technology
Xuefu South Road, Fuzhou City, Fujian Province, 350118, China
htxenergy@sina.com

Wen-Wu He*

School of Computer Science and Mathematics
Fujian Provincial Key Laboratory of Big Data Mining and Applications
Fujian University of Technology
Xuefu South Road, Fuzhou City, Fujian Province, 350118, China
hwwhbb@163.com

Bo-Yu Pei

School of Computer Science and Mathematics
Fujian University of Technology
Xuefu South Road, Fuzhou City, Fujian Province, 350118, China
priplex@163.com

Mei-Li Lin

School of Computer Science and Mathematics
Fujian University of Technology
Xuefu South Road, Fuzhou City, Fujian Province, 350118, China
meililin222@hotmail.com

Pei-Qiang Li

School of Electronic, Electrical Engineering and Physics
Fujian University of Technology
Xuefu South Road, Fuzhou City, Fujian Province, 350118, China
lpqcs@hotmail.com

*Corresponding author: Wen-Wu He

Received February 16, 2023, revised April 15, 2023, accepted June 5, 2023.

Abstract. *Deep reinforcement learning algorithms has attracted attention in the field of energy management in recent years, which can effectively interact with the environment and learn relevant policies to achieve specific goals. To eliminate Q-value estimation bias, in this paper we propose a real-time energy dispatching strategy for microgrid based on SD3 algorithm, and introduce prioritized experience replay and Huber loss, to develop an improved algorithm, called SD3 with loss-adjusted prioritized experience replay (SD3-LAP). A grid-connected microgrid that includes photovoltaic system, wind turbine, micro-turbine, energy storage system and electrical load is built to simulate the environment and validate the proposed methods. The problem of real-time energy dispatching for microgrid is recast as a Markov Decision Process, which aims to minimize the operating costs under the condition that the constraints of power balance and the upper and lower limits of SOC of the energy storage system are satisfied. Experimental results show that compared with baseline models, the energy dispatching strategies provide by SD3 algorithm and SD3-LAP algorithm are more effective and robust, and the mechanism of LAP is a general purposed plug-in that can effectively improve the performance of related reinforcement learning algorithm. The daily average operating cost of SD3-LAP algorithm and SD3 algorithm are respectively* 37.9% *and* 22.89% *lower than TD3 algorithm, while the daily average operating cost of of TD3 algorithm is* 21.9% *lower than DQN. In addition, compared with the original algorithms, namely SD3, TD3 and DQN, the LAP mechanism reduces respectively the daily average operating cost by* 19.5%, 11.7%, *and* 20.1%.

Keywords: Grid connected microgrid, Real-time energy dispatching, Deep reinforcement learning, SD3, SD3-LAP.

1. **Introduction.**

Due to its sustainability and environmental friendliness, the proportion of renewable energy power generation in the power system is increasing [1]. A microgrid is a power cluster composed of local loads, distributed generators, energy storage devices, etc, which can be connected to an external grid (EG) or run in an island mode [2]. As an effective carrier of distributed power generation, the microgrid can reduce its impact on the external power grid, improve the utilization rate of renewable energy and promote the nearby consumption of distributed renewable energy. The energy management system (EMS) is the control and decision-making center of the microgrid, which formulates reasonable operation plans to ensure the stability and efficiency [3]. In real-world scenarios, it is generally challenging to find the optimal energy management strategy due to the complexity of the operation environment, including the randomness of renewable energy output, the volatility of electricity load, the peak-valley nature of electricity prices, and other variability of related constraints. To meet the challenge and provide effective energy management strategy, a series of methods have been developed which mainly include mathematical programming methods [4, 5], model predictive control algorithms [6, 7], metaheuristic methods [8, 9, 10], etc. These methods use day-ahead scheduling or intra-day rolling optimization, and the effectiveness of the solution is highly dependent on the prediction accuracy of the time series model, which reduces their applicability for real-time energy dispatching.

Recently, deep reinforcement learning (DRL)-based algorithms have also been exploited for the optimal control of power systems [11]. By constructing the functional mapping relationship between the input state and the output action, DRL-based algorithms can provide rapid response to the change of operation environment and reduce the online optimization time. In addition, compared with traditional numerical optimization algorithms, DRL-based algorithms can learn the policy by interacting with the environment and do not need to build an explicit mathematical model. In particular, according to the action space, DRL-based algorithms can be divided into discrete and continuous domains.

As one of the typical work on discrete DRL-based algorithms, Alabdullah and Abido [12] developed a dispatching algorithm based on deep Q-network (DQN) to realize microgrid energy scheduling. To alleviate the overestimation problem suffered by the DQN-base algorithm, Liang et al. [13] further introduced the double deep Q-network (DDQN) and developed an improved algorithm by decoupling the selection and calculation of the target Q-values, to optimize the energy storage control strategy. Wang et al. [14] proposed the dueling deep Q-network (Dueling DQN) algorithm, an improved algorithm based on DQN, which decouples the optimal action-value function $Q^*(s,a)$ into the optimal state value function $V^*(s)$ and the optimal advantage function $A^*(s,a)$. Based on the improved Dueling DQN algorithm, Li et al. [15] proposed an energy management and optimization strategy for microgrid, which adopts a multi-parameter operation exploration mechanism to improve the possibility to find the optimal action. However, these discrete DRL-based algorithms can only deal with the discrete action space. Although the continuous action space can be discretized, too many actions will reduce the exploration efficiency of reinforcement learning and result in suboptimal solutions. Therefore, continuous continuous DRL-based algorithms are naturally more suitable for the optimal control of microgrid in real-world scenarios.

As for the continuous DRL-based algorithms, Guo et al. [16] adopted the proximal policy optimization (PPO) algorithm to solve the real-time microgrid energy management optimization problem in an uncertain environment and discusses the influence of different clipping rates on cumulative rewards. Fan et al. [17] combined the deep deterministic policy gradients (DDPG) algorithm with transfer learning and applied the resultant algorithm to different scenarios in microgrid. As pointed out by [18], however, the DDPG algorithm also has the drawback of overestimation. Cheng et al. [19] considered the action continuity and used the twin delayed deep deterministic policy gradient (TD3) algorithm to solve the entire life cycle optimization problem of the energy storage system. Based on the TD3 algorithm, Ye et al. [20] proposed a real-time demand response management strategy to cope well with the multi-source uncertainties and reduce the energy cost of residential household. TD3-based algorithms can alleviate the overestimation bias as mentioned above but it may lead to an underestimation bias [21].

In this paper, we revisit and improve the softmax deep double deterministic policy pradients (SD3) algorithm [22] to solve the optimal energy management problem in microgrid system. In the SD3 algorithm, Boltzmann softmax operator is used to smooth the optimization landscape and in the meanwhile, combined with the double actors and clipped double Q-learning to address the problem of estimation bias. DRL-based algorithms including SD3, however, ignoring the relative importance of different experience transition samples and used the method of uniform sampling method to select samples from the replay buffer. To fixe this, the idea of non-uniform sampling is exploited and the resultant algorithms are developed and studied. The main contributions of this paper are summarized as follows:

(1) A typical grid-connected microgrid including distributed generation units, energy storage systems, power loads and energy management systems is constructed, of which the optimal energy management problem is recast as a Markov decision process (MDP) to facilitate DRL-based energy dispatching.

(2) A real-time energy dispatching optimization algorithm based on the novel continuous SD3 algorithm is developed to minimize the operating cost of the target microgid, to alleviate the estimation bias suffered by existing algorithms used for energy dispatching.

(3) An improved algorithm SD3-LAP is further proposed to carry out a novel nonuniform experience transitions sampling to keep the diversity and in the meanwhile reduce

the gradient bias to stabilize its convergence, by exploiting the loss-adjusted prioritized experience replay (LAP) mechanism.

(4) A series of experiments over related DRL-based algorithms including the new developed ones are performed to verify the effectiveness and the practical utility of the methods studied in this paper.

The remainder of the paper is organized as follows. The target microgrid is constructed in Section 2. The MDP framework for microgrid energy dispatching is presented in Section 3. In Section 4, the SD3-LAP algorithm is proposed in detail to solve the involved energy dispatching problem. Experimental settings and related results are reported in Section 5, and the last Section gives some concluding remarks.

2. **Structure of the Target Microgrid.** In this paper, we consider a grid-connected microgrid with the structure as shown in Figure 1, which includes photovoltaic (PV) system, wind turbine (WT), micro-turbine (MT), energy storage system (ESS), energy management system (EMS) and electrical load components. Without loss of generality, in this study we assume that PV and WT are always in the working (open) state. The EMS controls the actions of charging and discharging of the ESS, the actions of purchasing and selling of the electricity with the EG, and the output actions of the MT. We aim to automatically learn the energy dispatch strategy to optimize microgrid operating costs with DRL-based algorithms.

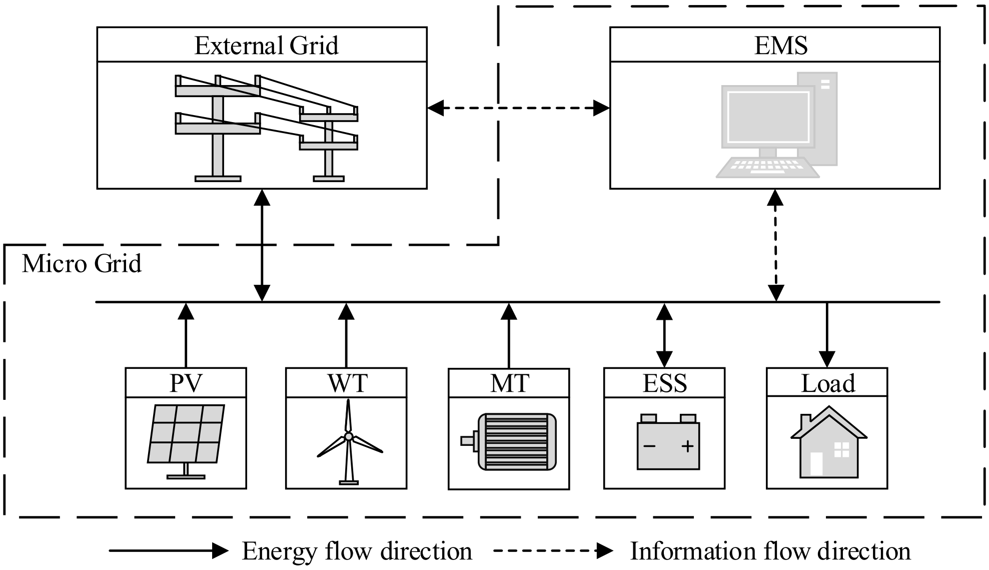


Figure 1. Structure of the target microgrid

2.1. **Component models in the microgrid.**

PV model: The power output of PV at time t is related to the irradiation intensity, the photovoltaic module area and the conversion efficiency, which can be formulated with

$$P_t^{\mathrm{PV}} = G_t A^{\mathrm{PV}} \eta^{\mathrm{PV}} / 1000 \tag{1}$$

where G_t is the solar radiation intensity (W/m^2) at time t, A^{PV} is the area of the PV modules (m^2) and η^{PV} is the conversion efficiency of the PV modules.

WT model: The power output of WT at time t can be regarded as a piecewise function, as shown below:

$$P_t^{\mathrm{WT}} = \begin{cases} 0 & v_t < v_{ci} \\ P_{\mathrm{r}}^{\mathrm{WT}} \frac{v_t - v_{\mathrm{ci}}}{v_{\mathrm{r}} - v_{\mathrm{ci}}} & v_{ci} \leq v_t < v_r \\ P_{\mathrm{r}}^{\mathrm{WT}} & v_r \leq v_t < v_{co} \\ 0 & v_t \geq v_{co} \end{cases} \tag{2}$$

where P_r^{WT} is the rated power of the WT (kw), v_{ci}, v_{co} and v_r are respectively the cut-in, the cut-off and the rated wind speed(m/s), v_t is the wind speed(m/s) at time t.

MT model: The MT is equipped to provide an additional power supply for the microgrid and to reduce the dependency of the microgrid on the EG. Its comprehensive costs include the operation management cost, the fuel cost and the environmental management cost, specifically

$$\begin{cases} c_t^{\mathrm{MT}} = c_t^{\mathrm{fuel}} + c_t^{\mathrm{env}} + c_t^{\mathrm{op}} \\ c_t^{\mathrm{fuel}} = \frac{c^{\mathrm{gas}}}{LHV} \frac{P_t^{\mathrm{MT}}}{\eta_t^{\mathrm{MT}}} \Delta t \\ c_t^{\mathrm{env}} = \sum\limits_{i=1}^{n} (p_i^{\mathrm{env}} \mathrm{u}_i P_t^{\mathrm{MT}}) \Delta t \\ c_t^{\mathrm{MT,op}} = p^{\mathrm{MT,op}} P_t^{\mathrm{MT}} \Delta t \end{cases} \tag{3}$$

$$\eta_t^{\mathrm{MT}} = 0.0752 \times \left(\frac{P_t^{\mathrm{MT}}}{65} \right)^3 - 0.3093 \times \left(\frac{P_t^{\mathrm{MT}}}{65} \right)^2 + 0.4174 \times \left(\frac{P_t^{\mathrm{MT}}}{65} \right) + 0.1069 \tag{4}$$

where c_t^{fuel}, c_t^{env} and c_t^{op} are respectively the fuel cost, the environmental management cost and the operation management cost (CNY), c^{gas} is the price of natural gas (CNY/m^3), P_t^{MT} is the output power of MT, LHV is the low calorific value of natural gas (kWh/m^3), η_t^{MT} is the power generation efficiency of the MT, p_i^{env} is the pollution-control cost of the i-th pollutant (CNY/kg), u_i is the emission coefficient of the i-th pollutant (kg/kWh) and $p^{\mathrm{MT,op}}$ is cost coefficient of the operation and maintenance of MT (CNY/kWh), Δt denotes the corresponding time duration. Note that the constants used in Equation (4) are drawn from [23].

ESS model: The ESS is equipped to maintain the power balance and improve the power quality of microgrid. The dynamic transition of the state of charge (SOC) of the energy storage system is formulated as follows:

$$SOC_{t+1} = SOC_t + \rho_t \frac{P_t^{\mathrm{cha}} \eta^{\mathrm{cha}}}{C_{\mathrm{N}}} \Delta t - (1 - \rho_t) \frac{P_t^{\mathrm{dis}}}{\eta^{\mathrm{dis}} C_{\mathrm{N}}} \Delta t \tag{5}$$

where P_t^{cha} and P_t^{dis} are respectively the charging and discharging power of the ESS at time t, η^{cha} and η^{dis} are correspondingly the charging and discharging efficiency, C_{N} is the rated capacity of the ESS, ρ_t is the indicator and $\rho_t = 1$ indicates that the ESS is in discharging state and $\rho_t = 0$ indicates that it is in the charging state.

As for the operation and management cost of the ESS, we formulate it with

$$c_t^{\mathrm{ESS}} = p^{\mathrm{ESS,op}} (P_t^{\mathrm{cha}} \eta^{\mathrm{cha}} + P_t^{\mathrm{dis}} / \eta^{\mathrm{dis}}) \Delta t \tag{6}$$

where c_t^{ESS} denotes the operation and management cost of the ESS at time t and $p^{\mathrm{ESS,op}}$ denotes correspondingly its cost coefficient.

3. **Problem Formulation.**

3.1. **Objective function.** As mentioned before, we aim to minimize the operating cost of the microgrid within a dispatching period. Specifically, the operating cost includes the transaction cost between the microgrid and the EG, the comprehensive cost of the MT, and the operation and management cost of the ESS, namely the objective function can be formulated as

$$C_T = \min\left(\sum_{t=0}^{T}(c_t^{\text{Grid}} + c_t^{\text{MT}} + c_t^{\text{ESS}})\right) \tag{7}$$

$$c_t^{\text{Grid}} = \beta_t p_t^{\text{sell}} P_t^{\text{Grid}} \Delta t + (\beta_t - 1) p_t^{\text{buy}} P_t^{\text{Grid}} \Delta t \tag{8}$$

where c_t^{Grid} is the transaction cost between the microgrid and the EG at time t, P_t^{Grid} is the exchanging power between the microgrid and the EG at time t, β_t is the indicator and $\beta_t = 1$ indicates that the microgrid is in selling electricity state and $\beta_t = 0$ indicates that it is in buying electricity state, p_t^{buy} and p_t^{sell} are respectively the buying and selling price (CNY/kWh) at time t.

3.2. **Restraint conditions.** The above objective function needs to be achieved under a series of constraints, including the power balance of the whole system, the charging and discharging power constraint of the ESS, the upper and lower limits of the SOC of ESS, the output power constraint of the MT, and the upper and lower limits of selling power of the microgrid, which can be specifically formulated with

$$\begin{cases} P_t^{\text{PV}} + P_t^{\text{WT}} + P_t^{\text{ESS}} + P_t^{\text{MT}} + P_t^{\text{Grid}} - P_t^{\text{Load}} = 0 \\ -P_{\max}^{\text{ESS}} \le P_t^{\text{ESS}} \le P_{\max}^{\text{ESS}} \\ SOC_{\min} \le SOC_t \le SOC_{\max} \\ P_{\min}^{\text{MT}} \le P_t^{\text{MT}} \le P_{\max}^{\text{MT}} \\ P_{\min}^{\text{Grid,sell}} \le P_t^{\text{Grid,sell}} \le P_{\max}^{\text{Grid,sell}} \end{cases} \tag{9}$$

where $-P_{\max}^{\text{ESS}}$ and $P_{\max}^{\text{ESS}}$ are respectively the maximum charging power and discharging power of the ESS, $SOC_{\max}$ and $SOC_{\min}$ are respectively the upper and lower limits of the SOC, $P_{\min}^{\text{MT}}$ and $P_{\min}^{\text{MT}}$ are respectively the minimum and maximum output power of the MT, $P_t^{\text{Grid,sell}}$ is the transmission power from the microgrid to the EG at time t, $P_{\min}^{\text{Grid,sell}}$ and $P_{\max}^{\text{Grid,sell}}$ are the minimum and maximum of transmission power.

3.3. **MDP modeling.** In this section, we recast the aforementioned energy management problem of the microgrid as an MDP problem to facilitate energy dispatching strategy solving based on reinforcement learning. An MDP problem generally consists of five tuples, namely the state space, the action space, the state transition function, the reward function and the discount factor.

State space: The state space S includes all the relevant states of the grid-connected microgrid, and at time t it can be formulated as

$$s_t = \{SOC_t, P_t^{\text{RG}}, P_t^{\text{Load}}, p_t^{\text{sell}}, p_t^{\text{buy}}, t\} \tag{10}$$

where $P_t^{\text{RG}} = P_t^{\text{PV}} + P_t^{\text{WT}}$ which denotes the generation power of renewable energy at time t.

Action space: The action space A includes the outpue of ESS and the output of MT, and at time t it can be formulated as

$$a_t = \{P_t^{\text{ESS}}, P_t^{\text{MT}}\} \tag{11}$$

where P_t^{ESS} and P_t^{MT} denotes respectively the output of ESS and the output of the MT at time t. Note that, due to the constrain condition of power balance of the system, the exchanging power between the microgrid and the EG does not need to be explicitly expressed.

Reward function: Recall that, we aim to minimize the operation cost of the microgrid of interest under a series of constraint conditions. Accordingly, the reward function including two parts. The first one responds to the operating cost and the second one responds to the constraint conditions. In particular, to encourage the action that meets the constraint conditions and punish the one who violates the conditions, two concrete terms are constructed respectively for the second part. Specifically, we formulate the reward function with

$$\begin{cases} r_t = -\omega_1 r_t^{\text{cost}} - \omega_2 r_t^{\text{pen}} + r_t^{\text{ex}} \\ r_t^{\cos t} = c_t^{\text{Grid}} + c_t^{\text{MT}} + c_t^{\text{ESS}} \\ r_t^{\text{pen}} = \ln(\upsilon_1 + pen_t^{\text{SOC}}) + \ln(\upsilon_2 + pen_t^{\text{Grid}}) \end{cases} \tag{12}$$

where r_t^{cost}, r_t^{pen} and r_t^{ex} are respectively the operating cost function, the penalty function (to punish the actions who violate the constraint conditions) and the extra reward (to encourage the actions who meet the constraint conditions) at time t, ω_1 and ω_2 are two weight coefficients to make a tradeoff among the three terms, pen_t^{SOC} and pen_t^{Grid} are respectively the spillage at time t that violates the SOC constraints and the one that violates the transmission power constraints. Note that, here, υ_1 and υ_1 are two constants used to keep the penalty term working well in practice, and the extra reward r_t^{ex} is set to be a pre-specified constant to keep the reword function being simple but effective.

Th design of Equation (12) encourages the EMS to take valid actions and accelerate the model convergence. As for the penalty function design, we utilize the logarithm operation and introduce two constants to keep it working in a robust manner. Actually, when the penalty value is too large, the microgrid dispatch strategy tends to be conservative, for example, the ESS does not perform any charging or discharging operation to keep the SOC of the system within the constraint; when the penalty value is too small, the microgrid dispatching strategy tends to violate frequently the constraints to pursue higher returns, for example, the MT chooses to generate electricity more frequently to get more revenue by selling the electricity which however, tends to lead abandoning the surplus electricity.

4. Real-time Energy Dispatching Based on Improved SD3 Algorithm.

4.1. Basics.

In the case considered in this paper, DRL-based algorithm continuously interacts with the microgrid environment during the training stage to obtain the feedback information and the optimization strategies. It aims to find the optimal energy scheduling strategy π^* to maximize the expected cumulative return $J(\pi)$ (equivalently, minimize the operation cost) within the specified scheduling period, namely,

$$\begin{cases} J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=1}^{T} \gamma^t r_t\right] \\ \pi^* = \arg\max_{\pi} J(\pi) \end{cases} \tag{13}$$

where $\tau = \{s_1, a_1, r_1, s_2, a_2, r_2, ..., s_T, a_T, r_T\}$ is the trajectory composed of a sequence of states, actions, and rewards, $\tau \sim \pi$ is the trajectory τ obtained by policy π, γ denotes the discount factor and T is the scheduling period.

4.2. SD3 algorithm.

The improvement of the SD3 over the TD3 algorithm lies in the introduction of the double actors and the softmax operator. Compared with the single actor, the double actors enable the agent to visit more valuable states and enhance the exploration capability of the agent [24]. The softmax operator mitigates the estimation bias by normalizing the Q-value and reduces the possibility of the algorithm falling into local optimum.

The actor network is used to learn the state-action mapping relationship which updates its parameters through policy gradient. In particular, the policy gradient is defined as follows,

$$\nabla_{\phi_i} J \approx \frac{1}{N} \sum\nolimits_j [\nabla_{\phi_i}(\pi(s_j;\phi_i)) \nabla_a Q_i(s_j, a_j; \theta_i)|_{a_j=\pi(s_j;\phi_i)}], i = 1, 2 \tag{14}$$

where $\nabla_{\phi_i}(\pi(s_j;\phi_i))$ denotes the gradient of the actor network, $\nabla_a Q_i(s_j, a_j; \theta_i)$ denotes the gradient of the critic network, N denotes the sampling batch size.

The critic network is used to evaluate the strategy proposed by the actor network to improve its decision-making ability, and it updates its parameters by minimizing the following MSE loss,

$$\delta_i(j) = Q_i(s_j, a_j; \theta_i) - y_i \tag{15}$$

$$L_{\mathrm{SD3}}(\delta_i(j)) = \frac{1}{N} \sum\nolimits_j (\delta_i(j))^2 \tag{16}$$

where j is the transition $\{s_j, a_j, r_j, s_{j+1}\}$, $Q_i(s_j, a_j; \theta_i)$ is the Q-value, y_i is the target Q-value, $\delta_i(j)$ is the temporal-difference (TD) error, indicating the difference between the Q-value and the target Q-value.

The SD3 algorithm approximates the target Q-value through the softmax operator. Firstly, It selects K truncated Gaussian noises to perturb a_{j+1}, and then selects the smaller Q-value from the two target critic networks, as shown below:

$$a_{j+1} \leftarrow \pi(s_{j+1}, \phi_i^-) + \mathrm{clip}(\varepsilon, -\mathrm{c}, \mathrm{c}),\ \varepsilon \sim \mathcal{N}(0, \sigma) \tag{17}$$

$$\hat{Q}(s_{j+1}, a_{j+1}) = \min_{i=1,2} \left(Q_i(s_{j+1}, a_{j+1}; \theta_i^-)\right) \tag{18}$$

where ε is the disturbance factor sampled from a Gaussian.

Then, in order to make the calculation of the softmax operator feasible in continuous space, the method of importance sampling is utilized to obtain the unbiased estimation. In particular,

$$\mathrm{softmax}_\beta\left(\hat{Q}(s_{j+1}, \cdot)\right) \leftarrow \mathbb{E}_{a_{j+1}\sim p}\left[\frac{\left[\exp(\beta\hat{Q}(s_{j+1}, a_{j+1}))\hat{Q}(s_{j+1}, a_{j+1})\right]}{p(a_{j+1})}\right] / \mathbb{E}_{\hat{a}_{j+1}\sim p}\left[\frac{\exp(\beta\hat{Q}(s_{j+1}, a_{j+1}))}{p(a_{j+1})}\right] \tag{19}$$

where $p(a_{j+1})$ is the probability density function of the Gaussian, β is the parameter of the softmax operator, which is used to control the deviation of Q-value estimation.

Finally, we approximate the target Q-value with $\mathrm{Softmax}_\beta\left(\hat{Q}_j(s_{j+1}, \cdot)\right)$. Specifically,

$$y_i \leftarrow r + \gamma(1 - d)\mathrm{Softmax}_\beta(\hat{Q}(s_{j+1}, \cdot)) \tag{20}$$

where d is a binary variable, indicating whether the state is terminated.

In order to reduce the error accumulation and improve the stability of the algorithm, target networks are utilized which update their parameters through the following soft update,

$$\begin{cases} \theta_i^- \leftarrow \zeta\theta_i + (1 - \zeta)\theta_i^- \\ \phi_i^- \leftarrow \zeta\phi_i + (1 - \zeta)\phi_i^- \end{cases} \tag{21}$$

where θ_i^- denotes the learnable parameters of the target actor network, ϕ_i^- denotes the ones of the target critic network and ζ denotes the soft update coefficient.

4.3. **SD3-LAP algorithm.** The SD3 algorithm utilizes the experience replay (ER) mechanism which are commonly used by the off-policy DRL-based algorithms. The ER mechanism samples uniformly experience transitions from the replay buffer with efficiency, but it ignores the relative importance of different experience transitions. To this end, prioritized experience replay [25] (PER) mechanism evaluates the importance of experience samples based on the absolute value of the TD error $|\delta_i(j)|$ and gives them different priorities to increase the sampling probability of important experiences.

Specifically, one can define the priority of the j-th transition corresponding to the i-th critic network as $p_i(j) = |\delta_i(j)| + \varepsilon$ and calculate the sampling probability with

$$P_i(j) = \frac{p_i^{\alpha}(j)}{\sum_k p_i^{\alpha}(k)} \tag{22}$$

where α is used to control the influence of the priority value on the sampling probability.

Using directly the aforementioned sampling probability tends to oversample the experience transitions with high priority and undersample the ones with low priority, and leads to suboptimal strategies. As pointed out in [26], the combination of priority sampling and MSE loss will further amplify the gradients of the training loss with respect to the model parameters when the TD error is too large, and make the convergence of the algorithm unstable. To this end, we combine the SD3 algorithm with the LAP [26] mechanism, which replaces the MSE loss with a more robust one, namely, the Huber loss, and modifies the sampling probability. We call the resultant algorithm SD3-LAP algorithm. Specifically, the Huber loss function and the sampling probability are as follows:

$$L_{\text{Huber}}(\delta_i(j)) = \begin{cases} \frac{1}{2}\delta_i(j)^2 & \text{if } |\delta_i(j)| \leq 1 \\ |\delta_i(j)| & \text{otherwise} \end{cases} \tag{23}$$

$$P_i(j) = \frac{\max(|\delta_i(j)|^{\alpha}, 1)}{\sum_k \max(|\delta_i(j)|^{\alpha}, 1)} \tag{24}$$

It can be seen that when $|\delta_i(j)| \leq 1$, the Huber loss reduces to the MSE loss and it becomes the L_1 loss otherwise. The LAP mechanism takes the advantages both of the uniform sampling and the priority sampling, and assigns the priority of samples with the truncated absolute value of TD error, so that the experience transitions with low priority can be uniformly sampled to ensure the sampling diversity and in the meanwhile reduce the gradient bias to stabilize the convergence of the resultant algorithm.

As fot the loss function of the i-th critic network used in the SD3-LAP algorithm, we formulate it as follows:

$$L_{\text{LAP}}(\delta_i(j)) = \frac{1}{N}\sum\nolimits_j L_{\text{Huber}}(\delta_i(j)) \tag{25}$$

Figure 2 shows the framework of the proposed SD3-LAP algorithm for real-time microgrid energy dispatching and Algorithm 1 details the training process of the SD3-LAP.

5. Case Studies.

5.1. **Experimental settings.** The radiation intensity data and the wind speed data involved in the example are drawn from the reference [27] and they are transformed into the renewable energy power data with Equation (1) and Equation (2). The load data are taken from the reference [28]. Figure A1(a) shows the renewable energy and the load, where the solid line represents the average values while the shaded part covers the variation range from the minimum values to the maximum values. The electricity prices are taken from the reference [23]. In [29], the buying and selling prices are originally the same, we set the buying price to be 1.2 times of the selling price to simulate the acts of purchase

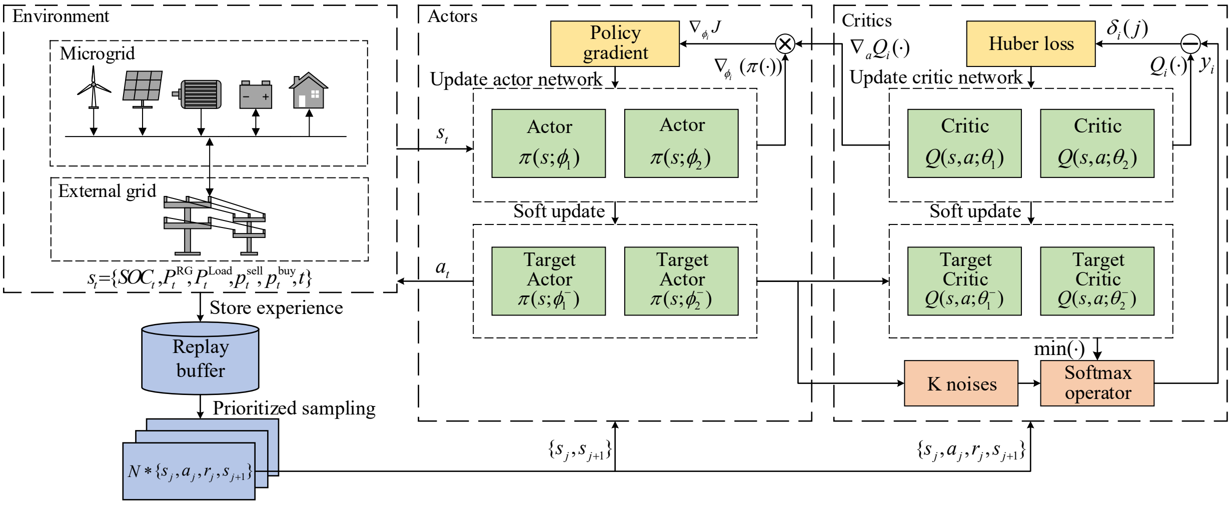


FIGURE 2. The framework of the proposed SD3-LAP algorithm

and sale. The resultant data is shown in Figure A1(b). The operation and maintenance cost along with the pollution control cost of MT and the emission coefficients of ESS are all drawn from the reference [23]. The microgrid operating parameters are shown in Table A1. We combine the above data into a 360-day microgrid dataset and the time interval is 1 hour. From each month we select 3 days of data to construct the testing set and the rest is used as the training set.

To verify the effectiveness of the methods studied in this paper, DRL-based algorithms such as DQN, TD3 along with SD3 are used as the baselines. As we know DQN and TD3 have been used in literatures as the algorithms to find the optimal strategy for energy dispatching and in the experiment , the two are combined with the LAP mechanism as well for ablation analyses and the resultant algorithms are called respectively DQN-LAP and TD3-LAP. As for the two discrete algorithms DQN and DQN-LAP, their action spaces are discretized into $4 \times 7 = 28$ action combinations, namely $P_t^{\text{MT}} = \{0, 10, 20, 30\}$ and $P_t^{\text{ESS}} = \{-30, -20, -10, 0, 10, 20, 30\}$. The specific structure of actor-critic framework used in the algorithms based on TD3 and SD3 is shown in Figure 3. The input of the actor network is a 6-dimensional state vector of $s_t = \{SOC_t, P_t^{\text{RG}}, P_t^{\text{Load}}, p_t^{\text{sell}}, p_t^{\text{buy}}, t\}$ and the output is a 2-dimensional action vector of $a_t = \{P_t^{\text{ESS}}, P_t^{\text{MT}}\}$. The input of the critic network is a 2-dimensional state-action vector of $\{s_t, a_t\}$, and the output is a state-action value $Q^\pi(s_t, a_t)$. A 2-layer of fully connected network is adopted which includes 128 and 64 neurons respectively and uses ReLU as the activation function.

Figure 4 shows the procedure to solve the optimal energy management problem with DRL-based algotithms, including the process of offline training and of online testing. The offline training includes E^{train} episodes and in each episode the DRL algorithm randomly selects one day from the training set to perform network training. When the offline training is completed, the trained parameters will be saved for future online testing. During online testing, the test data are fed into the DRL algorithm with trained model parameters and outputs energy scheduling strategies. In the experiments, the learning rate of the actor network and the critic network $lr = 0.001$, the discount factor $\gamma = 0.99$, the sampling batch size $N = 128$ and the memory size $S = 1024$. Models use the Adam optimizer, and the training episode number $E^{train} = 5000$.

5.2. **Training process.** To evaluate the stability of convergence of the revolved algorithms, 10 different random seeds are used and each algorithm is run 10 times. Figure 5

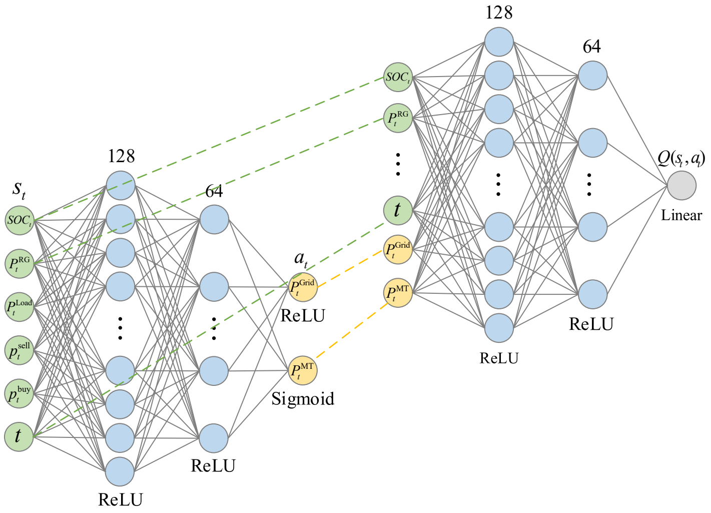


FIGURE 3. The structures of actor-critic network

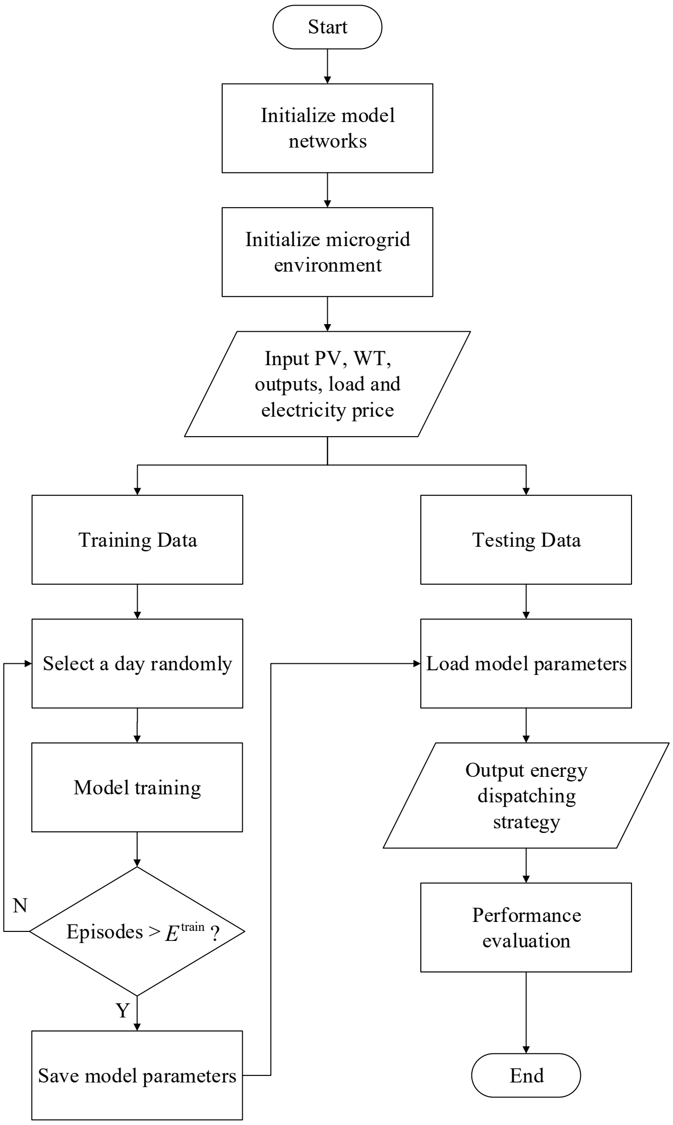


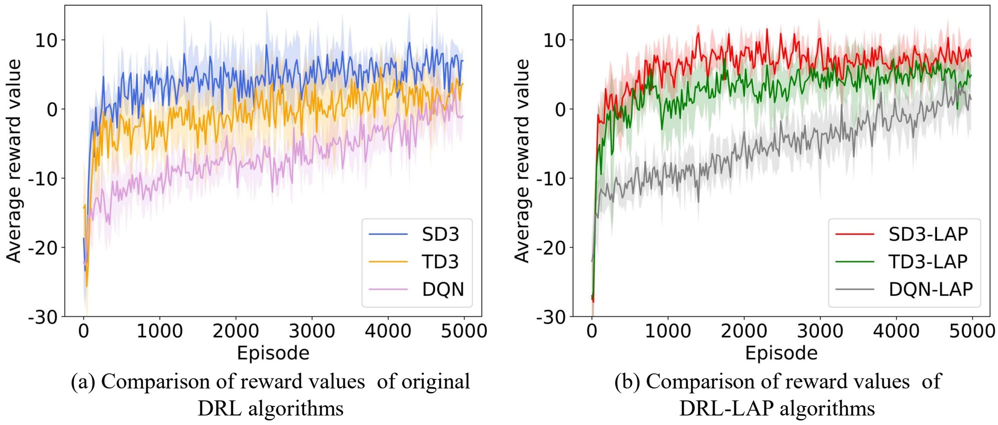
FIGURE 4. Flow chart of energy dispatching based on DRL-based algotithms

Algorithm 1 Pseudocode for SD3-LAP

```
Initialize microgrid environment.
Initialize actor network parameters $\phi_1$, $\phi_2$, and critic network parameters $\theta_1$, $\theta_2$.
Initialize target actor network parameters $\phi_1^- \leftarrow \phi_1$, $\phi_2^- \leftarrow \phi_2$, and target critic network parameters $\theta_1^- \leftarrow \theta_1$, $\theta_2^- \leftarrow \theta_2$.
Initialize replay buffer $R$ with size $S$, training episodes $E^{train}$, and other hyperparameters.
for $t = 1$ to $E^{train}$ do
    Select microgrid data of a day randomly from the training set.
    for $episode = 1$ to $T$ do
        Observe state $s_t$, and select action $a_t$ by actor network $\pi_1$ and $\pi_2$.
        Execute action $a_t$, calculate reward $r_t$ by (13), and transit to next state $s_{t+1}$.
        Store transition $\{s_t, a_t, r_t, s_{t+1}\}$ in $R$.
        if $t > S$ then
            for $i = 1, 2$ do
                for $j = 1$ to $N$ do
                    Sample transition $j$ with probability $P_i(j)$ by (24).
                    Sample $K$ noises $\varepsilon \sim \mathcal{N}(0, \sigma)$ and obtain $a_{j+1}$ by (17).
                    Calculate $\mathrm{softmax}_\beta\left(\hat{Q}(s', \cdot)\right)$ according to (18) and (19).
                    Approximate target Q-value $y_i$ by (20).
                    Compute TD-error $\delta_i(j)$ by (16).
                    Update the priority $p_i(j)$ of transition $j$ by $|\delta_i(j)| + \varepsilon$.
                end for
                Update the critic parameters $\theta_i$ using Huber loss by (25).
                Update the actor parameters $\phi_i$ using policy gradient by (14).
                Update the target networks $\phi_1^-$, $\phi_2^-$, $\theta_1^-$ and $\theta_2^-$ by (21).
            end for
        end if
    end for
end for
```

shows the curves of the average reward value over every 20 episodes of related algorithms, where the solid line and the shaded ones represent the mean and the standard deviation. In the early stage, the EMS randomly selects actions to explore fully the working environment and there is no substantive training. Therefore, the reward values in this stage are small. When the replay buffer is full, the model starts training and the reward value keeps increasing, and finally it converges to a certain range. It is worth noting that, from Figure 5, the LAP mechanism can generally improve the rewards of the DRL algorithms using it, and SD3-LAP algorithm outperforms all its competitors, which indicate that the proposed algorithm and the LAP mechanism is feasible and effective for microgrid energy dispatching.

Figure 6 shows the curves of average violation of constraint power over every 20 episodes of the revolved algorithms, where the meaning of the solid line and the shaded ones are the same as those in Figure 5. In the early stage, all the algorithms violate frequently the constraints. With the increasing of training episode, their average violation values gradually decrease and become relatively stable. It can be observed from Figure 6 that, compared with other DRL-based algorithms, the SD3 algorithm has the lowest violation values in most cases, which suggests that SD3-based dispatching algorithm has a stronger "risk aversion" ability. It is more encouraging that, the LAP mechanism can further reduce

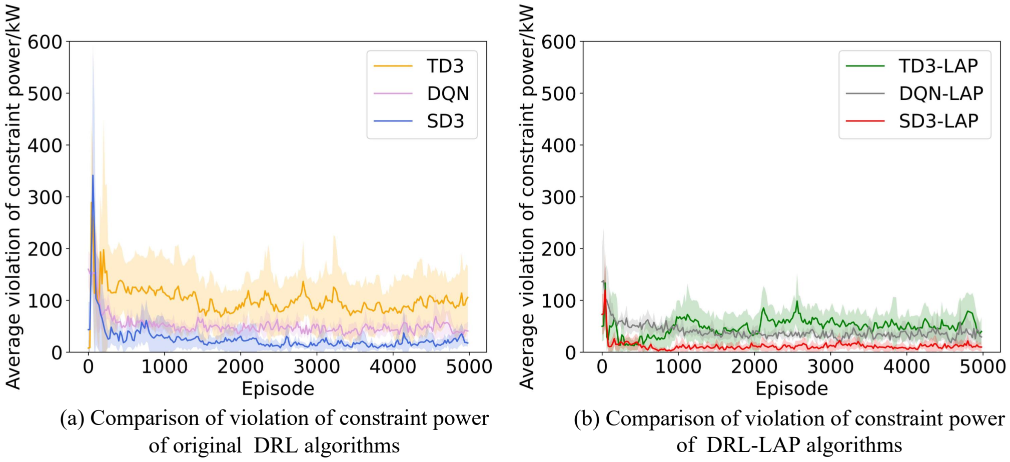


(a) Comparison of reward values of original DRL algorithms

(b) Comparison of reward values of DRL-LAP algorithms

FIGURE 5. Curves of average reward value

the violations and keep the algorithms using it work in a well-mannered but effective manner. Overall, the SD3-LAP algorithm performs the best, SD3 algorithm performs the second, and the LAP mechanism can further provide additional general improvement.



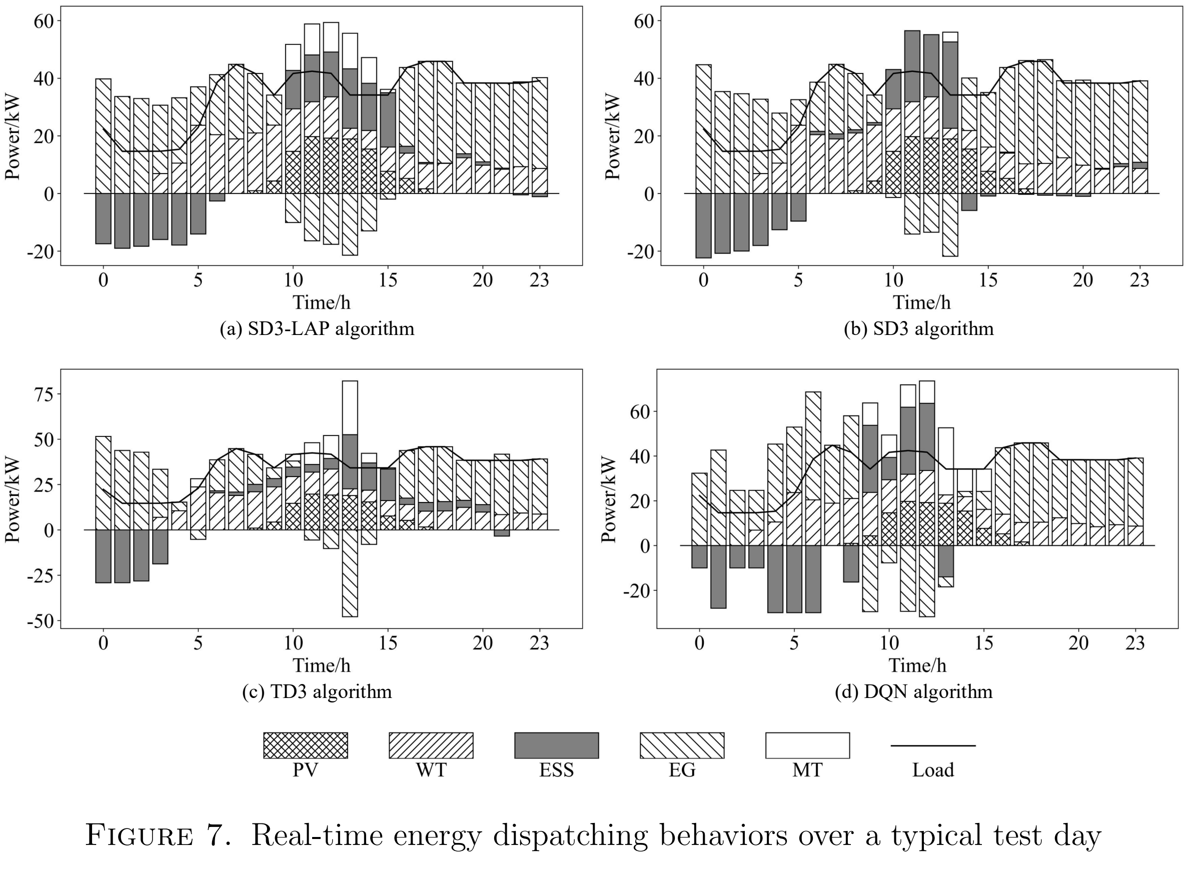
(a) Comparison of violation of constraint power of original DRL algorithms

(b) Comparison of violation of constraint power of DRL-LAP algorithms

FIGURE 6. Curves of average violation of constrain power

5.3. **Testing process.** In this subsection, we analyze and discuss the effectiveness of energy dispatching strategies offered by revolved DRL-based algorithms. Following the common way used by related references, we mainly focus on the best performance of each algorithm out of 10 runs as discussed in the last subsection.

Energy dispatching behaviors: Figure 7 shows the best performances of real-time energy dispatching of revolved algorithms over a typical test day. For the ESS, the positive power indicates that it is in a discharging state and the negative power corresponds to a charging state. For the EG, the positive power indicates that the microgrid purchases electricity from the EG and the negative one means that the microgrid sales electricity to the EG. It can be seen from Figure 7 that, the real-time decision-making offered by the SD3-LAP algorithm basically follows the changing of real-time electricity price. Specifically, during the off-peak period of $0:00 \sim 7:00$, it is not sunny enough and the PV is inactive, but it is windy and the WT works; the MT is inactive since the comprehensive costs of the MT during this period (obtainable by calculating Equation (3) and (4)) is higher than the purchase price (as can be found in Figure A1); the microgrid correspondingly purchases electricity from the EG to meet its load demand, and in the meanwhile the ESS is in charging state to store energy for future demand. During the mid-peak period of $8:00 \sim 10:00$, it is sunny enough and the PV begins to generate electric power and the WT still works on; the MT is inactive since its comprehensive costs is still higher than the purchase price; the ESS stops working when its SOC is at the maximum and the microgrid continues to purchase electricity from the EG to meet its load demand. During the peak period of $10:00 \sim 16:00$, both the PV and the WT work on with the good weather condition; the purchase price is higher than the comprehensive cost of MT and it starts to work; the ESS is in discharging state to release energy stored in early stage, and the whole output of the microgrid is higher than its load demand and it can sale the surplus energy to the EG, to realize arbitrage. During the off-peak period of $16:00 \sim 23:00$, the WT works on and the PV works over $16:00 \sim 18:00$; the MT is inactive since the purchase price is low; the ESS is in discharging state over $16:00 \sim 22:00$ and in charging state over $22:00 \sim 24:00$, corresponding to the changing of purchase price; the microgrid mainly purchases electricity to meet the load demand.



(a) SD3-LAP algorithm

(b) SD3 algorithm

(c) TD3 algorithm

(d) DQN algorithm

FIGURE 7. Real-time energy dispatching behaviors over a typical test day

As discussed above, the SD3-LAP algorithm can adapt itself well to implement effective energy dispatching with the dynamic changing of working environment. In contrast, other algorithms perform not so well and during some periods, their dispatching decisions are not quite reasonable. Actually, from Figure 7(b-d) one can find negative examples such as: SD3-based strategy purchases electricity from the EG to meet the load demand of microgrid during the peak period of $14:00 \sim 15:00$, and in the meanwhile the ESS is in charging state rather than in discharging state to release energy for off-peak power consumption; TD3-base strategy activates the MT to generate power during the period of $5:00 \sim 6:00$ but the comprehensive cost of MT in this period is higher than the purchase price, and at the same time sells electricity to the EG but the sales price in this period is at the valley; DQN-based strategy activates the MT during the mid-peak period of $9:00 \sim 10:00$ while the purchase price is lower than its generating cost, and sells electricity to the EG by releasing the energy stored by the ESS which obviously misses the peak period of selling price.

Operating cost analysis: Now we turn to check the operation costs of target microgrid corresponding to different algorithms. To this end, Figure 8 shows their optimal cumulated operating costs out of 10 runs over 36 test days. As expected, it can be found that the cumulated operating cost of the SD3-LAP algorithm has always been the lowest, and the one of SD3 algorithm is the second lowest. In addition, one can also find that the LAP mechanism can provide general cost reduction for all the algorithms using it. Actually, the daily average operating cost of SD3-LAP is 37.2 CNY which is respectively 37.9% and 51.5% lower than those of TD3 algorithm and of DQN algorithm, and compared with the original algorithms, namely SD3, TD3 and DQN, the LAP mechanism reduces respectively the daily average operating cost by 19.5%, 11.7%, and 20.1%.

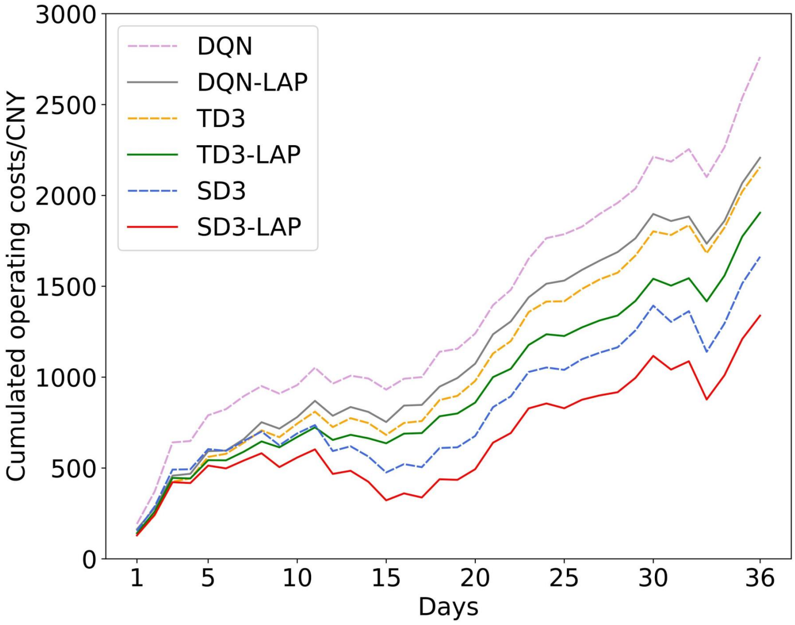


Figure 8. Optimal cumulated operating costs of revolved algorithms

Table 1 shows the cumulated operation costs of the involved algorithms which are run 10 times, where the minimum value, maximum value, average value and the standard deviation of the operating cost over 36 test days of each algorithm are presented. From Table 1 we can find that, the SD3-LAP algorithm outperforms all its competitors with the three values mentioned above and its standard deviation is also the smallest one, which indicates that SD3-LAP algorithm adapts itself well to the randomness of renewable energy output and electrical load. Once again, we can see clearly the universal reduction offered by the LAP mechanism. Table 2 shows the average values of cumulated abandoned-electricity (CAE for short) of the algorithms of interest over 10 runs. It can be found that the means of CAE of SD3-LAP algorithm and SD3 algorithm are significantly lower than those of other algorithms, showing that the methods studied in this paper are effective.

Energy consumption analysis: Here, considering the fact that the SD3-LAP algorithm performs the best, we take it as the example to show energy consumption of the target microgrid. Specifically, Figure 9 shows the energy consumption of SD3-LAP

Table 1. Cumulated operating cost of revolved algorithms

Algorithm	Cumulated operating cost (CNY)			
	min	max	mean	std
SD3-LAP	1338.3	1935.7	1600.72	212.1
SD3	1662.5	2435.8	2118.7	259.2
TD3-LAP	1904.6	2637.1	2235.6	285.0
TD3	2156.1	2999.1	2616.3	331.0
DQN-LAP	2206.3	3495.8	2653.7	460.1
DQN	2761.5	3822.3	3302.9	483.3

Table 2. Averaged cumulated abandoned-electricity of revolved algorithms

Algorithm	Mean value of CAE (kWh)
SD3-LAP	172.5
SD3	185.9
TD3-LAP	456.3
TD3	522.6
DQN-LAP	328.9
DQN	401.3

algorithm over $10:00 \sim 11:00$ of a typical test day. The microgrid and the EG are interconnected through the transmission lines. Notice that, the electricity price of EG is at the peak and the cost of MT is lower than it. Therefore, it is reasonable for the microgrid to activate the MT and let the ESS be in the discharging state (the PV and the WT are always on and their working states depend on the weather conditions). Actually, Figure 9 validates well the effectiveness of the dispatching strategy offered by SD3-LAP algorithm. From this figure, one can clearly see that the MT is on, the renewable energy sources offer their maximum output, and the ESS is in discharging state to release the energy stored during the off-peak tariff period. In addition, as one may have noticed that, there is no abandoned electricity at this time frame.

Ablation analysis: To further check the contribution of LAP mechanism we also carry out a series of experiments on the variants of SD3 algorithm such as the one with the PER mechanism and the one with the Hubber loss. Specifically, Table 3 shows their operating costs over the testing set where the results of SD3 algorithm and SD3-LAP algorithm are included as well for readability. It can be seen that, the PER mechanism can reduce the cumulated average operation cost, but its standard deviation is larger than that of SD3 algorithm for the reason as discussed before that the combination of PER mechanism and MSE loss tends to cause gradient bias and make the revolved algorithm unstable. On the contrary, one can find that, by replacing directly the MSE loss used by the original SD3 algorithm with the Huber loss, both the mean and the standard deviation of the cumulative operating cost are reduced, suggesting that, the Huber loss is a better

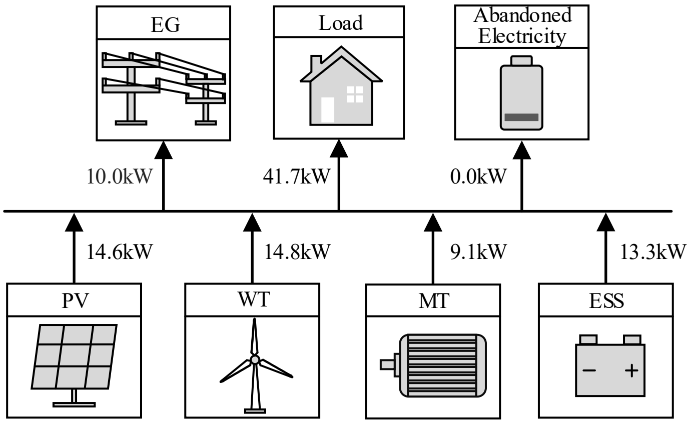


Figure 9. A snapshot of energy consumption of SD3-LAP

choice for robust performance. Once again, the SD3-LAP algorithm that combines the non-uniform sampling with the Hubber loss outperforms its competitors and validates well is practical utility.

Table 3. Analysis of ablation results

Algorithm	Cumulated operating cost(CNY)			
	min	max	mean	std
SD3	1662.5	2435.8	2118.7	259.2
+PER	1510.6	2588.6	2106.9	355.5
+Huber	1666.2	2364.3	2041.9	252.6
+LAP	1338.3	1935.7	1600.72	212.1

6. **Conclusion.** In this paper we consider the real-time energy dispatching problem of a microgrid and recast it as an MDP to facilitate DRL-based solving. A specially designed microgird that include typical components is construct to provide a general and practical example for the validation of the effectiveness of the revolved methods. Recently developed SD3 algorithm is utilized to perform energy dispatching for the target microgrid to alleviate the estimation bias suffered by existing DRL-based energy dispatching algorithms. It is further integrated with the LAP mechanism to carry out a novel non-uniform experience transitions sampling, to keep the diversity and in the meanwhile reduce the gradient bias to stabilize its convergence. Several representative DRL-based algorithms and the one proposed in this paper are utilized in the target microgrid to validate their utility. Experimental analysis shows that, compared with existing energy dispatching algorithms, the proposed SD3-LAP algorithm performs the best and the SD3 algorithm performs the second and their dispatching behaviors are more effective and robust. Moreover, the LAP mechanism shows general improvements for DRL-based algorithms using it, which provides an interesting direction for future work on energy dispatching.

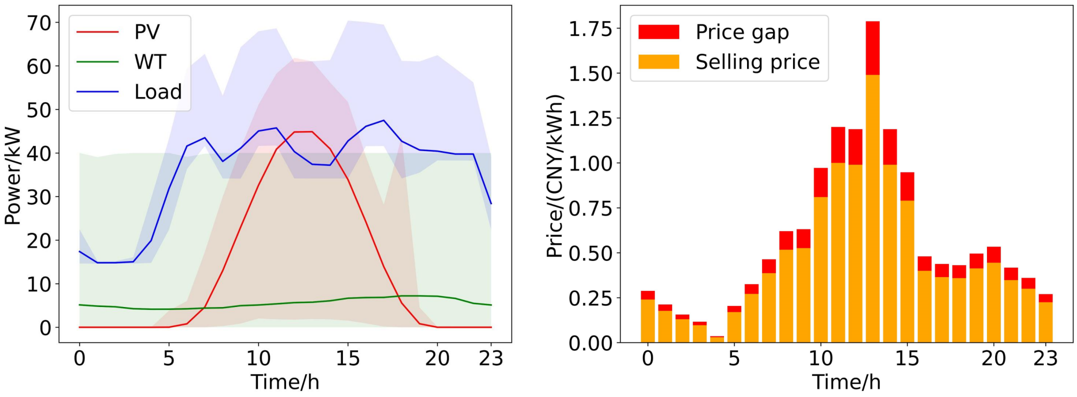
The secure transmission of information between different entities in a microgrid [30] is out of the scope of this research. It is interesting to develop an effective method to combine adversarial training [31] with DRL-based algorithms to enhance the security and robustness, and it will be included in our near future work.

Acknowledgment. This work is supported by the Natural Science Foundation of Fujian Province under Grant 2020J01891 and Grant 2019J05113.

Appendix A.

Table A1. Microgrid operating parameters

Parameter	Value	Parameter	Value
$A/(\mathrm{W/m^2})$	300	η^{dis}	0.95
η^{PV}	0.2	$C_{\mathrm{N}}/(\mathrm{kWh})$	200
$v_{ci}/(\mathrm{m/s})$	3	$p^{\mathrm{ESS,op}}/\mathrm{CNY}$	0.045
$v_{co}/(\mathrm{m/s})$	25	$P_{\min}^{\mathrm{MT}}/\mathrm{kW}$	0
$v_r/(\mathrm{m/s})$	14	$P_{\max}^{\mathrm{MT}}/\mathrm{kW}$	30
$P_r^{\mathrm{WT}}/\mathrm{kW}$	40	$P_{\min}^{\mathrm{Grid,sell}}/\mathrm{kW}$	0
$c^{\mathrm{gas}}/(\mathrm{CNY/m^3})$	1.4	$P_{\max}^{\mathrm{Grid,sell}}/\mathrm{kW}$	60
$LHV/(\mathrm{kWh/m^3})$	9.7	SOC_{min}	0.3
$p^{\mathrm{MT,op}}/(\mathrm{CNY/kWh})$	0.128	SOC_{max}	0.9
T	24	SOC_{ini}	0.4
η^{cha}	0.95	$P_{\max}^{\mathrm{ESS}}/\mathrm{kW}$	30



(a) Renewable energy power and load power

(b) Peak-valley electricity price

Figure A1. Related data

REFERENCES

[1] R. H. M. Zargar, and M. H. Y. Moghaddam , "Development of a Markov-chain-based solar generation model for smart microgrid energy management system," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 2, pp. 736–745, 2019.

[2] I. Patrao, E. Figueres, G. Garcerá and R. González-Medina, "Microgrid architectures for low voltage distributed generation," *Renewable and Sustainable Energy Reviews*, vol. 43, pp. 415–424, 2015.

[3] Q.-Y. Jiang, M.-D. Xue, and G.-C. Geng, "Energy management of microgrid in grid-connected and stand-alone modes," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 3380–3389, 2013.

[4] O. Pisacane, M. Severini, M. Fagiani, and S. Squartini, "Collaborative energy management in a micro-grid by multi-objective mathematical programming," *Energy and Buildings*, vol. 203, pp. 109432, 2019.

[5] P. L. Querini, U. Manassero, E. Fernádez, and O. Chiotti, "A two-level model to define the energy procurement contract and daily operation schedule of microgrids," *Sustainable Energy, Grids and Networks*, vol. 26, pp. 100459, 2021.

[6] S. Batiyah, R. Sharma, S. Abdelwahed, and N. Zohrabi, "An MPC-based power management of standalone DC microgrid with energy storage," *International Journal of Electrical Power & Energy Systems*, vol. 120, pp. 105949, 2020.

[7] Y. Zhang, F.-L. Meng, R. Wang, W.-L. Zhu, and X.-J. Zeng, "A stochastic MPC based approach to integrated energy management in microgrids," *Sustainable Cities and Society*, vol. 41, pp. 349-362, 2018.

[8] M. 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, 2019.

[9] P. C. Sahu, R. C. Prusty, and S. Panda, "Improved-GWO designed FO based type-II fuzzy controller for frequency awareness of an AC microgrid under plug in electric vehicle," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 1879–1896, 2021.

[10] A. Fathy, T. M. Alanazi, H. Rezk, and D. Yousri, "Optimal energy management of micro-grid using sparrow search algorithm," *Energy Reports*, vol. 8, pp. 758–773, 2022.

[11] J.-J. Yang, M. Yang, M.-X. Wang, P.-J. Du, and Y.-X. Yu, "A deep reinforcement learning method for managing wind farm uncertainties through energy storage system control and external reserve purchasing," *International Journal of Electrical Power & Energy Systems*, vol. 119, pp. 105928, 2020.

[12] M.H. Alabdullah, M.A. Abido, "Microgrid energy management using deep Q-network reinforcement learning," *Alexandria Engineering Journal*, vol. 61, no. 11, pp. 9069–9078, 2022.

[13] H. Liang, H.-X. Li, H.-Y. Zhang, Z.-H. Hu, Z.-M. Qing, and J.-W. Cao, "Research on Control Strategy of Microgrid Energy Storage System based on Deep Reinforcement Learning," *Power System Technology*, vol. 45, no. 10, pp. 3869–3877, 2021.

[14] Z.-Y. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *33rd International Conference on Machine Learning (ICML2016)*, PMLR, 2016, pp. 1995–2003.

[15] H.-T. Li, B.-C. Shen, Y.-H. Yang, W. Pei, X. Lyu, Y.-T. Han, "Energy Management and Optimization Strategy for Microgrid Based on Improved Dueling Deep Q Network Algorithm," *Automation of Electric Power Systems*, vol. 46, no. 7, pp. 42–49, 2022.

[16] C.-Y. Guo, X. Wang, Y.-H. Zheng, and F. Zhang, "Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning," *Energy*, vol. 238, pp. 121873, 2022.

[17] L.-Q. Fan, J. Zhang, Y. He, Y. Liu, T. Hu, and H. Zhang, "Optimal scheduling of microgrid based on deep deterministic policy gradient and transfer learning," *Energies*, vol. 14, no. 3, pp. 584, 2021.

[18] S. Fujimoto, H. Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," in *35th International Conference on Machine Learning (ICML 2018)*, PMLR, 2018, pp. 1587–1596.

[19] T.-X. Cheng, X.-Y. Xu, Z. Yan, and Y.-M Zhu, "Optimal operation based on deep reinforcement learning for energy storage system in photovoltaic-storage charging station," *Electric Power Automation Equipment*, vol. 41, no. 10, pp. 90–98, 2021.

[20] Y.-J. Ye, D.-W. Qiu, H.-Y. Wang, Y. Tang, and G. Strbac. "Real-time autonomous residential demand response management based on twin delayed deep deterministic policy gradient learning," *Energies*, vol. 14, no. 3, pp. 531, 2021.

[21] K. Ciosek, Q. Vuong, R. Loftin, and K. Hofmann. Better exploration with optimistic actor critic. in *Neural Information Processing Systems (NeurIPS 2019)*, 2019, pp. 1785–1796.

[22] L. Pan, Q.-P. Cai, L.-B. Huang. "Softmax deep double deterministic policy gradients," in *Neural Information Processing Systems (NIPS 2020)*, 2020, vol. 33, pp. 11767–11777.

[23] X.-T. Hu, T.-Q. Liu, C. He, S. Liu, and Y.-K. Liu, "Multiobjective optimal operation of microgrid considering the battery loss characteristics," *Proceedings of the CSEE*, vol. 36, no. 10, pp. 2674–2681, 2016.

[24] J.-F. Lyu, X.-T. Ma, J.-P. Yan and X. Li, "Efficient continuous control with double actors and regularized critics," *Association for the Advancement of Artificial Intelligence*, vol. 36, no. 7, pp. 7655–7663, 2022.

[25] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. "Prioritized experience replay," in *International Conference on Learning Representations (ICLR 2016)*, 2016.

[26] S. Fujimoto, D. Meger, and D. Precup, "An equivalence between loss functions and non-uniform sampling in experience replay," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020, vol. 33, pp. 14219–14230.

[27] J.-J. Yang, Z.-Y. Sun, W.-Q. Hu, and L. Steinmeister, “Joint control of manufacturing and onsite microgrid system via novel neural-network integrated reinforcement learning algorithms,” *Applied Energy*, vol. 315, pp. 118982, 2022.

[28] G. Henri, T. Levent, A. Halev, R. Alami, and P. Cordier, “pymgrid: An Open-Source Python Microgrid Simulator for Applied Artificial Intelligence Research,” in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[29] M. Ding, Y.-Y. Zhang, M.-Q. Mao, X.-P. Liu, and N.-Z. Xu, “Economic Operation Optimization for Microgrids Including Na/S Battery Storage,” *Proceedings of the CSEE*, vol. 31, no. 04, pp. 7–14, 2011.

[30] T.-Y. Wu, Y.-Q. Lee, C.-M. Chen, Y. Tian, and N. A. AI-Nabhan, “An enhanced pairing-based authentication scheme for smart grid communications,” *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2021.

[31] K. Wang, F.-J. Li, C.-M. Chen, M. M. Hassan, J.-Y. Long, and N. Kumar, “Interpreting adversarial examples and robustness for deep learning-based auto-driving systems,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9755–9764, 2021.